Stat.BD

CLASS NOTES • CHAPTER 2

Population, Samples, Variables & Data

What one row is, what one column is, and what job each column does

Applied Biostatistics & Research Methodology for Clinical Research
Stat.BD mentorship programme

Mentor: Muhtasim Munif Fahim, MSc

Chapter 2

Population, Samples, Variables & Data

Every study, however it was designed, ends as a rectangle of numbers. This chapter is about naming the parts of that rectangle precisely enough to argue about them.

Chapter 1 drew a line between description, which is a statement about the people who were measured, and inference, which is a statement about the people who were not. That line is the foundation of everything that follows, and it has a gap in it. It says nothing about *who* the second group is, *what* was measured on the first, or *which* measurement is doing the work.

This chapter closes that gap, and it does so in a particular order. It begins with the shape a dataset takes, because that shape decides what every later question can mean. It then names the populations a study moves between, and the two families of quantity that live in them. It classifies the columns of the rectangle by what kind of thing each one is, and then by what job each one has been given. The last idea is the one the rest of the chapter exists to support: the job a column does is not a property of the column. It is assigned by the question, and changing the question changes the job without changing a single number.

What this chapter establishes

1. That a study's data form a rectangle whose rows are units of observation and whose columns are variables, and that naming both is the first step in reading any paper.
2. That "the population" in a research report is at least four different things, and that the distance between them is where generalisability is won or lost.
3. That a parameter describes a population and is never observed, while a statistic describes a sample and is always calculated.
4. That variables divide into categorical and numerical kinds, that the kind constrains what may sensibly be done with the column, and that the same underlying measurement often appears in a paper in two different kinds.
5. That a column may be measured, reported or constructed, and that a constructed column carries every decision made in constructing it.
6. That outcome, exposure, predictor and covariate are roles assigned by a research question rather than properties of columns.
7. That an association is a feature of a dataset and causation is a feature of the world, that four distinct explanations can produce an association, and that no dataset by itself distinguishes between them.

Notation

The convention below is close to universal in medical statistics. Greek letters denote quantities belonging to a population, which are fixed and never observed. Roman letters denote quantities calculated from a sample, which are known and differ from one sample to the next. A circumflex, read as ”hat”, marks a quantity estimated from data.

Symbol	Name	What it refers to
N	population size	The number of units in the population. Usually unknown, and often not a finite number in any practical sense.
n	sample size	The number of rows in the dataset. Always known.
μ	population mean	The mean of a numerical variable across the whole population.
$\bar{x}$	sample mean	The mean of that variable among the units actually measured.
σ	population standard deviation	The spread of the variable in the population.
s	sample standard deviation	The spread among the units measured.
π	population proportion	The fraction of the population in some category.
$\hat{p}$	sample proportion	The fraction of the sample in that category.
x_i	one observation	The value of a variable for the ith unit in the sample.
a	a count	The number of sampled units falling in a given category.

Contents

1 The shape of data

A clinical study can be designed in many ways. It can follow people forward for a decade, interview them once in their homes, randomise them to two treatments, or trawl a hospital's records for cases that already happened. Whatever the design, the thing that arrives at the analysis is the same object: a rectangle. Down the side runs one line for each unit that was studied. Across the top runs one heading for each thing that was recorded about every unit. Every question that can be asked of a study is, in the end, a question about that rectangle.

Most clinicians have kept one for years without giving it a name. A ward register has one line per patient and the same columns filled in for everybody: bed number, age, diagnosis, admission date, outcome. It is a dataset, and the habits that make a register useful are the habits that make a dataset analysable.

Intuition

A dataset is a table in which every row is one of the things being studied and every column is one thing recorded about all of them. Nothing more complicated than that, and the discipline is in refusing to let either rule bend.

Definition 1.1 • Unit of observation

The *unit of observation* is the thing that one row of the dataset describes. The *unit of analysis* is the thing that the research question is about. They are frequently the same, and when they are not, the difference matters more than almost anything else in the study.

Variable
One thing, measured or recorded the same way, on every unit in the study. It is a column, not a value.

A *variable* is a column. The word is worth defending against everyday usage, because a variable is not a number: it is the whole column of numbers, and the thing it names is the quantity that varies from row to row. Age is a variable; forty-seven is a value that age took in one row. This distinction looks pedantic on the page and stops being pedantic the moment someone reports "the mean of the variable" and means something else by it.

one column is one variable
the same thing measured on everybody

one row is one participant
the unit of observation

ID	Age	Sex	Waist (cm)	SBP	Hypertension
001	52	M	88.0	138	No
002	67	F	84.5	152	Yes
003	41	M	72.3	118	No
004	59	F	91.2	146	Yes

this whole object is "the rectangle"

Illustrative rows, in the shape of the Sylhet survey dataset. Whatever the design, this is what reaches the analysis. Before anything else can be said about a study, two questions have answers: *what is one row?* and *what is one column?*

Figure 1: A study's data, in the shape every study eventually takes. The rows are illustrative and are not participants from any published study.

Two questions follow from Figure 1, and they are the first two questions worth asking of any paper. What is one row? And what is one column?

The first question sounds trivial and is not. A study of cataract surgery might report results on two hundred eyes belonging to one hundred and twenty patients. Its rectangle has two hundred rows and it does not have two hundred independent patients. The two eyes of one person are more alike than two eyes drawn at random from the clinic. Any method that treats them as unrelated will overstate how much the study knows. The same arithmetic problem arises with five hundred outpatient visits made by ninety children, with repeated blood pressure readings on one person, and with teeth, joints, lesions and pregnancies. The rectangle makes the problem visible. Abstracts usually do not.

Watch out

When a paper's sample size is larger than the number of people in the study, find out what the extra rows are before reading anything else. The design has made a promise about independence that the analysis has to keep.

The second question, what is one column, has a stricter answer than it appears to. A column is one thing, recorded the same way, on everybody. If blood pressure was measured with a calibrated machine for half the participants and taken from their recollection for the other half, those are not one column with some noise in it. They are two different measurements that have been stacked into one place, and the stacking is a decision that ought to be defended in the Methods section.

2 Four populations, not one

Chapter 1 made the point that a population is a decision the researcher makes rather than a fact about the world. That statement is true and it is not yet usable, because research reports use the word "population" for at least four different groups of people, and the differences between them are exactly where a study's claims are strongest or weakest.

Definition 2.1 • Target, source, study population and sample

The *target population* is the group the answer is meant to apply to. The *source population* is the group that could actually have been sampled, usually defined by whatever list or register the researchers had access to. The *study population* is the group that was approached. The *sample* is the set of units that produced data.

These four are nested, each contained in the one before it, and something is lost at every step. What is lost between target and source is a matter of who could have been reached at all. What is lost between approached and sample is a matter of who agreed, who was found, and who was still alive.

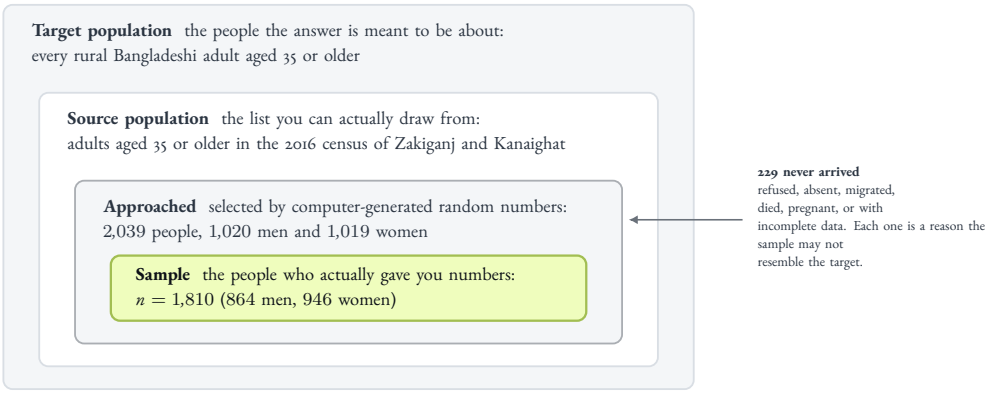


Figure 2: The four populations of a cross-sectional survey of hypertension in rural Sylhet, with the counts the study reported at each stage. Numbers from Khanam and colleagues, 2019 (reference 1.).

Example 2.1 • Naming the four populations of a real survey

A cross-sectional survey conducted between August 2017 and January 2018 measured blood pressure in two subdistricts of Sylhet district, Bangladesh. Its four populations can be read off the Methods section.

Target. Adults aged 35 years and older in rural Bangladesh. This is what the study is offered as evidence about.

Source. Adults aged 35 and older recorded in the 2016 census of the Zakiganj and Kanaighat subdistricts, where the research group had maintained a field site for about two decades. This is the list that could be sampled.

Study population. The 2,039 people selected from that list by computer-generated random numbers: 1,020 men and 1,019 women.

Sample. The 1,810 people who completed the survey: 864 men and 946 women.

The gap between target and source is the one that limits the study most. Sylhet was, in the national survey the authors cite, the division with the lowest recorded prevalence of hypertension among the country's eight. A result from two of its subdistricts is not a result about Dhaka, and the paper does not claim otherwise.

2.1 Inclusion, exclusion, and the frame

Two mechanisms turn a target population into a source population, and they are different kinds of decision.

Definition 2.2 • Inclusion and exclusion criteria

An *inclusion criterion* states what a person must be to enter the study. An *exclusion criterion* disqualifies somebody who meets the inclusion criteria, for a separate reason that has to be given.

In the Sylhet survey the inclusion criteria were an age of 35 years or more, residence in one of the two subdistricts, and presence on the 2016 census list. The exclusion criterion was pregnancy, and 14 women were excluded on that ground.

The exclusion is a good one. Pregnancy alters blood pressure, so retaining pregnant women would have blurred every estimate. It also means the study says nothing about hypertension in pregnancy, which is a real clinical question that these data cannot address. Every criterion narrows the population the answer applies to, and that cost is paid whether or not anybody notices it.

A third mechanism operates without anybody deciding anything. The sampling frame here was a census taken in 2016, and fieldwork ran from August 2017. Anybody who moved into the two subdistricts after the census was therefore never eligible to be drawn, however long they had lived there by the time a fieldworker arrived.

Sampling frame
The list from which the sample is actually drawn. The source population as it exists on paper.

Watch out

Those people are not non-responders. Non-response is countable, and this study counts it honestly. A gap between the target population and the sampling frame is not countable from inside the study at all, because the study's only picture of the population is the frame itself. The inference about recent arrivals is ours; the paper states the census year and the fieldwork dates and does not discuss it. Sampling frames and their failures are the subject of Chapter 16.

2.2 The people who did not arrive

Between the study population and the sample, 229 people were lost. Their reasons are recorded, and reading them is more instructive than the headline result.

Reason	Number	Why it might matter
Refused	36	Refusal is rarely random.
Absent	56	The frequently absent may differ in occupation and income.
Migrated out	76	Migration is related to age, wealth and health.
Died	42	Death is not unrelated to blood pressure.
Pregnant	14	Excluded by design, which is a decision rather than a loss.
Incomplete data	5	Usually harmless at this size.
Total	**229**	

These six totals are ours. The study reports its losses separately for men and women; the figures here are those two lists added together.

The question to put to each line is the same: could the reason a person is missing be related to the thing the study is measuring? For refusals the honest answer is that nobody knows. For the forty-two deaths the answer is almost certainly yes, because blood pressure and mortality are related, and the people who died before the survey reached them were on average older and sicker than those who did not.

Watch out

Non-response is not administrative housekeeping. It is the first place where a sample stops resembling the population it was drawn from, and unlike most threats to a study's validity it is usually visible in the paper. The reasons are worth reading every time.

How far this kind of loss actually distorts a result, and what can be done about it, belongs to the treatment of bias in Chapter 15 and of sampling and generalisability in Chapter 16. What belongs here is the habit of looking.

3 Parameters and statistics

A study exists because somebody wants to know something about a population and cannot measure it. What they want to know is a number: the proportion of rural adults with hypertension, the mean systolic pressure in that group, the spread of pressures around that mean. Those numbers exist. They are simply not available.

What is available is the same set of numbers computed on the sample. The distinction between the two families is the single most useful piece of notation in medical statistics, and it is carried entirely by which alphabet a symbol comes from.

Definition 3.1 • Parameter and statistic

A *parameter* is a numerical summary of a population. It is a fixed quantity and it is never observed. A *statistic* is the same kind of numerical summary computed from a sample. It is always known, and it takes a different value in every sample that could have been drawn.

Intuition

The parameter is the patient's true blood pressure: the quantity that would be known if it could be measured perfectly and forever. The statistic is what the cuff read this morning. Treatment decisions are made on the second while the thing anybody actually cares about is the first.

Estimate
A statistic used as a stand-in for the parameter it corresponds to. The hat notation, as in $\hat{p}$, marks one.

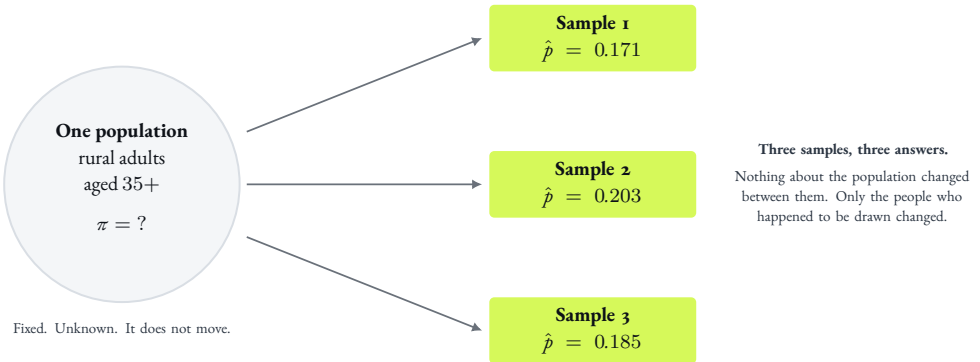
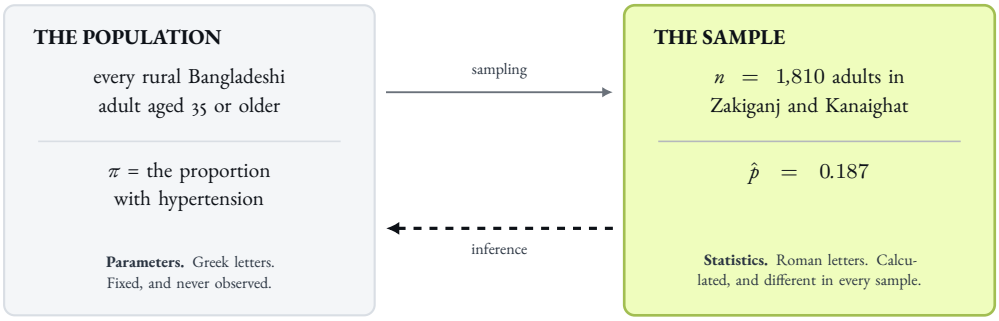


Figure 3: One population, three samples, three different answers. The parameter did not move; only the people who happened to be drawn did. The sample values are illustrative.

Figure 4 carries an asymmetry worth stating plainly. The right-hand box contains everything that was measured. The left-hand box contains everything the study is actually about. Only one of the two arrows between them can be wrong: sampling is a procedure that either was



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 4: The two worlds a study moves between, and the two arrows that connect them. Sampling is a procedure; inference is a claim.

or was not carried out as described, whereas inference is a claim about people nobody met.

3.1 Two summaries, stated

Two statistics account for most of what a descriptive table reports. Both are stated here so that the notation has somewhere to land. Neither is derived, and choosing between summaries is the subject of Chapters 3 and 4 rather than this one.

For a numerical variable, the sample mean is the total divided by the count:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.1}$$

- $\bar{x}$ the sample mean, a statistic
- n the number of units in the sample
- x_i the value of the variable for the ith unit
- $\sum_{i=1}^{n}$ an instruction to add up the values for units 1 through n

For a categorical variable, the sample proportion in a category is the count in that category divided by the total:

$$\hat{p} = \frac{a}{n} \tag{3.2}$$

- $\hat{p}$ the sample proportion, a statistic estimating π
- a the number of sampled units in the category of interest
- n the number of units in the sample

Example 3.1 • A proportion, from a real survey

In the Sylhet survey, 339 of the 1,810 adults surveyed met the study's definition of hypertension: 162 of 864 men and 177 of 946 women. Substituting into equation (3.2),

$$\hat{p} = \frac{339}{1{,}810} = 0.187.$$

The study reports this as a prevalence of 18.7%, with a 95% confidence interval running

Example 3.1 • *continued*

from 17.0 to 20.6.

Three things in that sentence belong to different categories. The counts 339 and 1,810 are facts about the sample. The figure 0.187 is $\hat{p}$, a statistic. The quantity π, the proportion of all rural Bangladeshi adults aged 35 and over with hypertension, is a parameter: it has a value, the study does not contain it, and no study ever will.

Provenance. The counts and the percentage are both reported in the paper. The division shown here is ours, and is included only to make the substitution visible.

The interval attached to 18.7% is a statement about how precisely the sample pins down the parameter. What it does and does not mean is the subject of Chapter 8, and Chapter 7 supplies the machinery it is built from. Guessing at its meaning in the meantime is how the most common misconception in medical statistics gets established.

Watch out

A statistic is not a smaller, less impressive version of a parameter. It is a different kind of object. It has a value that can be checked by recomputing it, and its value would have been different had a different sample been drawn.

4 What kind of thing is a column?

Every column in a dataset records something, but not all columns record the same kind of thing, and the kind decides what can sensibly be done with the column. Averaging blood group produces a number and the number means nothing. Counting how many participants had a waist circumference of exactly 76.4 centimetres produces an answer that is almost always one. The classification below is not a piece of taxonomy for its own sake. It is the first filter in every decision about how to summarise, display and analyse a variable.

The first split is whether the column names a group or measures a quantity.

Definition 4.1 • Categorical and numerical variables

A *categorical* variable places each unit into one of a set of groups. It is *binary* when there are two groups, *nominal* when there are more than two and they have no natural order, and *ordinal* when the groups have an order but the distances between them are neither equal nor known.

A *numerical* variable records a quantity. It is *discrete* when it can only take separated values, typically because it is counted, and *continuous* when any value in a range is possible, typically because it is measured.

Ordinal is the kind that repays attention, because it is the one most often handled as though it were something else. Wealth tertile has three categories in a genuine order: lowest, middle, highest. What it does not have is a distance. The gap between the lowest and middle tertile

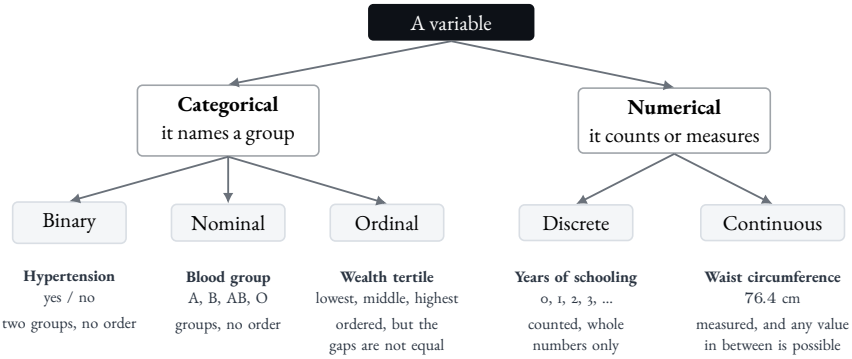


Figure 5: The five kinds of variable, with a clinical instance of each. All but blood group are columns of the Sylhet survey.

is not the same quantity as the gap between middle and highest, and it is not clear what it would mean to say it was. Coding them 1, 2 and 3 and taking an average produces a number, and that number is not the average wealth of anything.

Example 4.1 • Classifying the columns of a real survey

Eight of the columns of the Sylhet survey, with the kind each one belongs to.

Column	Kind	Note
Sex	Binary	Two groups, no order.
Current smoker	Binary	Recorded as reported by the participant.
Hypertension	Binary	Not measured. Constructed from other columns.
Wealth tertile	Ordinal	Ordered; the gaps are neither equal nor known.
Age, in completed years	Discrete	Counted in whole years, and treated as continuous.
Years of schooling	Discrete	Counted, and frequently reported as ordered bands instead.
Fruit and vegetable servings	Discrete	Counted per day, as reported by the participant.
Waist circumference	Continuous	Measured with a tape, to one decimal place.

Two of these are worth arguing about. Age in completed years is strictly discrete and is treated as continuous by everyone, at no practical cost, because the values are numerous and closely spaced. Years of schooling is discrete in the same way but is often used as though it were ordinal, in bands of none, one to five, and six or more, which is what this study did.

Watch out

The boundary between discrete and continuous is softer than the diagram suggests, and little turns on it. The boundary between categorical and numerical is not soft, and a great deal turns on it. A column that names a group has no mean.

4.1 The boundary that is genuinely arguable

The tree of Section 4 is a tool for choosing a summary rather than a taxonomy to be defended, and one boundary in it is genuinely contested.

Consider a pain score out of ten. The case for treating it as ordinal is that the gap between 2 and 3 is not the same amount of pain as the gap between 8 and 9: nobody calibrated the scale, so averaging it averages unequal things. The case for treating it as discrete is that it is a whole number on a fixed range, that everybody uses the same anchors, and that treating it as numerical permits methods which are far more informative.

The point

What settles it is not the column but the question, and how much argument the author is willing to have. Reporting a median pain score of 6 is unarguable. Reporting a mean of 6.3 is a claim that the scale has equal intervals, and a reviewer may ask for that claim to be defended.

The same argument applies to any Likert-type scale, and it reaches the same place. Chapters 3 and 4 take up how a categorical column and a numerical column are each summarised, which is why the boundary matters at all.

4.2 The same measurement, in two kinds

A descriptive table often contains the same underlying measurement twice: once as the number that was recorded, and once as a set of categories somebody sorted that number into. Reading a table without noticing this is common, and it hides a methodological decision in plain sight.

	As measured	As categorised
Body mass index	20.3 kg/m^2, median	Underweight / normal / overweight
Waist circumference	77.0 cm, median	Low risk / high risk
Years of schooling	2 years, median	None / 1 to 5 / 6 or more
Age	48 years, median	35–44 / 45–54 / 55–64 / 65+

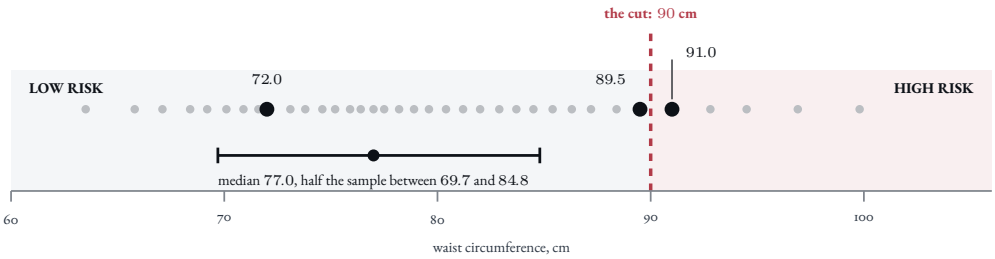
Medians are for the whole sample of 1,810. Khanam and colleagues, 2019, Table 1.

The right-hand column did not come from the participants. It came from the researchers, who chose where to place each boundary. Sometimes the choice follows an external standard, as the body mass index bands here follow the World Health Organization's. Sometimes it follows convention, and sometimes it follows the data. In all three cases it is a decision that belongs in the Methods section, and it is not always there.

4.3 What a threshold costs

Cutting a measured quantity into categories is a trade, and it is worth being precise about what is given and what is received.

Waist circumference in the Sylhet survey was measured in centimetres. The median was 77.0 cm, with half the sample lying between 69.7 and 84.8 cm. It was then reduced to two categories, at 90 cm for men and 80 cm for women. The headline association the study reports is for the two categories and not for the centimetres: among men, an adjusted odds ratio of 4.0, with a 95% confidence interval from 2.5 to 6.4.



72.0 **and** 89.5**: the same.**
17.5 cm apart, and the categorised column cannot tell them apart at all.

89.5 **and** 91.0**: different.**
1.5 cm apart, which is inside the variation between two careful tape measurements.

The reported odds ratio of 4.0 is for the two boxes, not for the centimetres. The cut buys a rule a health worker can apply with a tape measure and no calculator, and it pays for that with every distinction inside each box.

Figure 6: The waist column as it was measured, with the cut drawn on it. The median and interquartile range are as reported; the individual dot positions are illustrative.

Intuition

Cutting a measured number into two boxes discards every distinction inside each box and invents one at the boundary. A man at 72 cm and a man at 89 cm become identical. A man at 89 cm and a man at 91 cm become different kinds of person. Neither statement is true of the men.

What is received in exchange is real. A threshold can be applied by a health worker with a tape measure and no calculator, it produces a rule that can be written into a screening programme, and the authors of this study say directly that ease of measurement was part of why they preferred waist circumference to body mass index. A study designed to inform a programme has good reason to report the quantity the programme will actually use.

So dichotomisation is a defensible choice rather than an error. What turns it into a problem is silence: a reader who meets an odds ratio of 4.0 and assumes it refers to centimetres has been misled, whether or not anybody intended it. The general cost, which is usually a loss of precision, belongs with the treatment of precision and sample size in Chapter 17.

Watch out

Whenever a result is reported for a categorised version of a measured variable, the effect refers to the categories. Moving from one side of the cut to the other is what the number describes, not moving by one unit of the original measurement.

4.4 Where a column comes from

Columns arrive in a dataset by three different routes, and the route determines what can go wrong with them.

Measured. An instrument produced the value. Waist circumference came from a tape. Blood pressure came from a calibrated digital machine, and the protocol specifies three readings about ten minutes apart, with the average of the last two entering the dataset.

Reported. The participant said so. Smoking status and daily servings of fruit and vegetables are of this kind. So is much of what any questionnaire collects.

Constructed. The researchers built the column out of other columns by applying a rule.

The blood pressure protocol deserves a second look, because it contains an idea this course returns to. Three readings were taken and the first was discarded. That is an admission that one reading is not the quantity of interest but one look at it, and that looks vary. Chapter 7 gives that variation a name and a size.

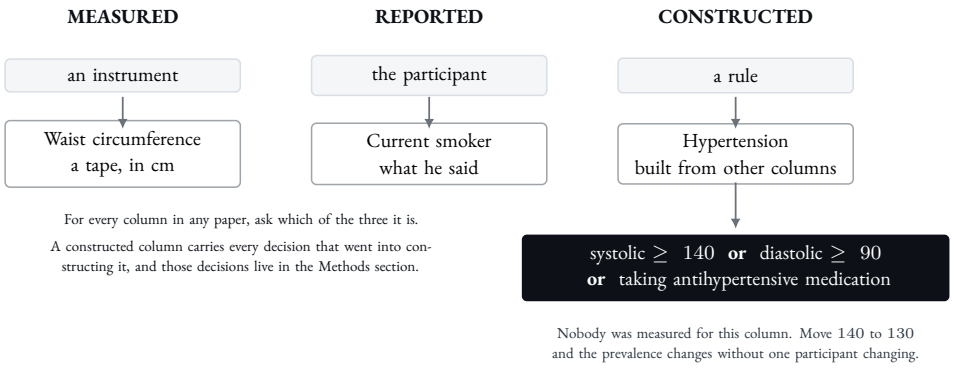


Figure 7: The three routes a column can take into a dataset. The third carries every decision made in constructing it.

Definition 4.2 • Operational definition

An *operational definition* is the exact rule by which a study decides what value a variable takes for a given unit. It is what makes a clinical concept into a column.

Example 4.2 • Two constructed columns

Neither of the two columns below was measured on anybody. Each is a rule.

Hypertension. A participant was recorded as hypertensive if systolic pressure was at least 140 mm Hg, *or* diastolic pressure was at least 90 mm Hg, *or* the participant reported taking antihypertensive medication.

Three consequences follow immediately. A participant whose treatment is working well, reading 128/82 on medication, is counted as hypertensive, which is clinically correct and is not what the words "blood pressure of at least 140" would suggest on their own. Changing the threshold from 140 to 130 would change the reported prevalence without a single participant changing. And the column depends on self-reported med-

Example 4.2 • *continued*

ication use, so it inherits whatever errors that reporting contains.

Wealth tertile. Housing type, drinking water source, toilet type, availability of electricity and a list of household possessions were combined into a single score, and the households were then divided into three equal groups by that score.

This column is constructed twice over. The score is built from five inputs by a statistical procedure, and the tertiles are cut from the score. It also measures a property of a household rather than of a person, which Section 7 returns to.

The habit worth building

For every column in any paper, ask which of the three routes it took. A constructed column carries every decision that went into constructing it, and those decisions are often the most consequential thing in the Methods section.

5 The jobs a column can do

Knowing what kind of thing a column is settles how it may be summarised. It does not say what the column is *for*. A study of hypertension in rural Sylhet recorded age, sex, education, wealth, body mass index, waist circumference, smoking, smokeless tobacco use, diet and physical activity. All of them are variables and they are not all doing the same work.

Definition 5.1 • Outcome, exposure, predictor, covariate

The *outcome* is the variable the study is trying to explain or predict. The *exposure* is the variable whose effect on the outcome is being asked about. A *predictor* is a variable used to forecast the outcome, with no claim that it produces the outcome. A *covariate* is a variable measured and accounted for so that the relationship between exposure and outcome can be seen more clearly.

Covariate
A variable accounted for in an analysis. Whether it is a *confounder* is a separate, causal claim, and is the subject of Chapter 15.

The word covariate is deliberately modest. It says that a variable was measured and taken account of. It does not say why, and in particular it does not say that the variable distorts the relationship under study. That stronger claim has its own word, confounding, and its own chapter.

Exposure and predictor look interchangeable and are not, and the difference is the hinge of this chapter.

Two different questions

Prediction asks: given what can be seen about this patient, what is likely to happen? Causation asks: if this were changed, would the outcome change?

A variable can be an excellent predictor while causing nothing at all. Yellowed fingers predict lung cancer usefully and cause none of it. Prediction identifies whom to watch.

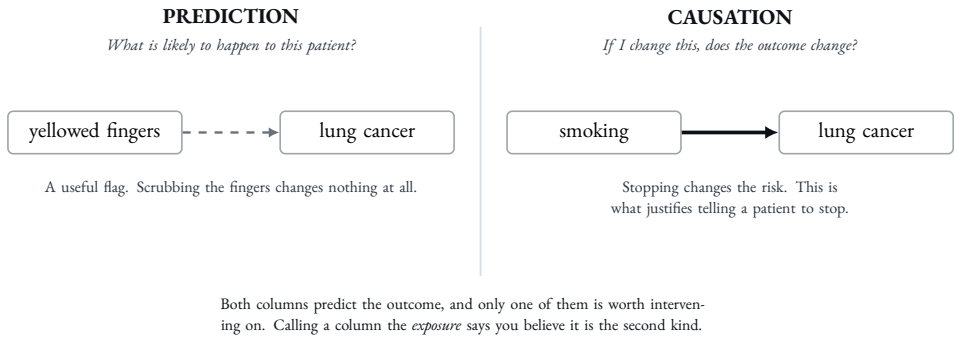


Figure 8: Two different questions asked of the same kind of data. Both columns predict the outcome; only one is worth intervening on.

> Only causation justifies an intervention.

5.1 The role is assigned by the question

Roles are not properties of columns. Nothing about a column of waist measurements marks it as an exposure, and nothing about a column of ages marks it as a covariate. The role is conferred by the question being asked, and asking a different question of the same dataset reassigns the roles without changing a single value.

The Sylhet study does exactly this in its own analysis, rather than in any example invented for teaching. Its authors found that body mass index and waist circumference were strongly related to each other, reporting a correlation of 0.68, and so entered waist circumference into their main model and left body mass index out. They then fitted a second model the other way round, with body mass index in place of waist circumference. In the first model waist circumference is the exposure under study. In the second it is absent, and body mass index carries the role.

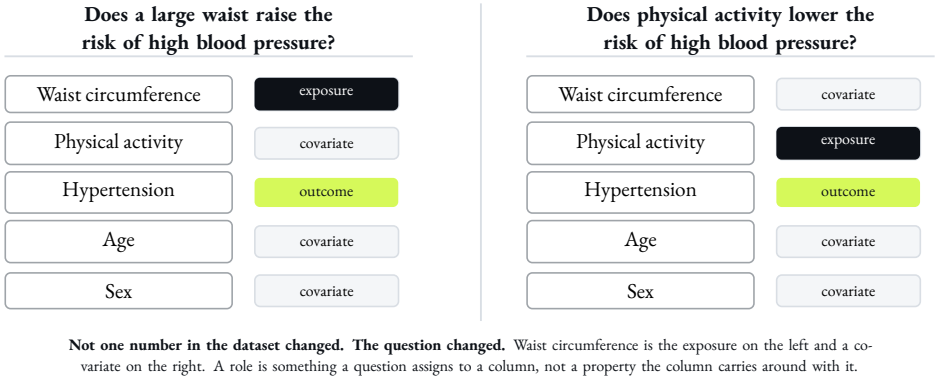


Figure 9: One dataset, two research questions, and the roles that change between them. Both analyses are the study's own.

Example 5.1 • The same column, two jobs

Consider the two questions in Figure 9, asked of the same rectangle of data.

Does a large waist raise the risk of high blood pressure? Waist circumference is the exposure. Hypertension is the outcome. Age, sex, wealth, smoking and physical activity are covariates.

Does physical activity lower the risk of high blood pressure? Physical activity is now the exposure. Hypertension is still the outcome. Waist circumference has become a covariate, and is now one of the things accounted for rather than the thing under study.

Not one number in the dataset differs between the two analyses.

Even the outcome is not permanently fixed. Hypertension is the outcome throughout this study, and in a study of stroke it would be an exposure.

The point

A role is something a question assigns to a column, not a property the column carries around with it. Before a variable can be called an exposure, the question has to be stated, because ”exposure” already contains a claim about which way the influence is supposed to run.

6 Association and causation

The reason roles matter is that one of them, exposure, is a causal word. Calling a column the exposure asserts that intervening on it would move the outcome. That assertion is not something a dataset can supply.

Definition 6.1 • Association and causation

Two variables are *associated* when knowing the value of one tells you something about the likely value of the other. This is a property of a dataset and it can be checked by computation.

One variable *causes* another when intervening to change the first would change the second. This is a property of the world. No dataset contains it.

The familiar slogan that correlation is not causation is true and teaches very little, partly because it is worn smooth by repetition and partly because it uses the wrong word. Correlation is not a synonym for association. It is one specific measure of one specific kind of association: two numerical variables, moving together in something approximating a straight line. The relationship between hypertension and wealth tertile in the Sylhet data is a perfectly good association and there is no correlation coefficient anywhere near it.

Intuition

Every correlation is an association. Most associations are not correlations. The useful distinction is not between correlation and causation but between what is in the data and what is in the world.

The correlation coefficient itself, in its Pearson and Spearman forms, is the subject of Chapter 24. What matters here is the logical gap, not the arithmetic.

6.1 Four explanations for one finding

An association is a finding. It is not yet an explanation, and there are four distinct explanations available. The data alone cannot distinguish between them.

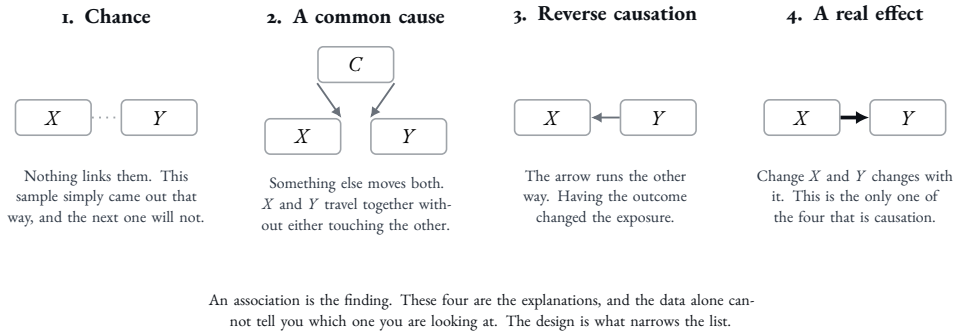


Figure 10: The four ways two variables come to move together. Only the last is causation.

Chance. The association is a feature of this sample and not of the population. Smaller studies produce this more often, which is why a single surprising result is a question rather than a conclusion.

A common cause. Some third variable influences both, so the two move together without either touching the other. This is confounding, and Chapter 15 is devoted to it.

Reverse causation. The influence runs the other way. Having the outcome changed the exposure.

A real effect. Changing the first variable would change the second.

Only the design of a study, not the size of its dataset, narrows this list. Chapters 12 and 13 are about which designs narrow it and by how much. What can be said now is that randomisation, when it is available, is the only common device that addresses the second explanation for variables nobody thought to measure, which is why Chapter 1 gave it the weight it did.

Example 6.1 • An association nobody believes

Among the men in the Sylhet survey, current smokers had *lower* odds of hypertension than non-smokers.

	Odds ratio	95% confidence interval
Unadjusted	0.5	0.4 to 0.7
Adjusted	0.7	0.5 to 1.0

Example 6.1 • *continued*

An odds ratio below one indicates lower odds in smokers. The unadjusted figure is well below one and its interval excludes one. Nobody proposes that smoking prevents hypertension.

The study's authors raise reverse causation by name: people who already knew they had high blood pressure might have changed their habits, which would put the non-smokers disproportionately among the hypertensive. They then argue against their own explanation, on the grounds that about half of those found to have hypertension in this survey did not know they had it and so had no diagnosis to react to.

That sequence is worth more than the result. A specific alternative explanation is named, and then evidence is brought against it. It is what an honest discussion section looks like.

Adjustment moved the estimate from 0.5 to 0.7 and widened the interval until it touched one. This is worth noticing in both directions: accounting for age, wealth and the rest explained part of the apparent protection, and it could only account for variables that were measured.

The study is cross-sectional, and its authors state that this design cannot establish whether the observed factors cause hypertension. That limitation is not a flaw in this study. It is a property of the design, and Chapter 12 is where designs are matched to the questions they can answer.

Watch out

Observational studies routinely report associations in causal language: a factor is "protective", an exposure "leads to" an outcome, a behaviour "reduces risk". The verb is doing work the design cannot support. Reading for those verbs, and asking which of the four explanations has actually been ruled out, is most of what critical appraisal consists of.

7 What one row is, revisited

Section 1 asked what one row of a dataset is and answered that it is the unit of observation. The question is worth returning to with the rest of the chapter's vocabulary available, because the answer is not always a patient, and because the unit of observation and the unit of analysis can differ.

Three cases, in increasing distance from the familiar one.

In the Sylhet survey each row is a person, and blood pressure, waist circumference and smoking all belong to that person. Wealth does not. The wealth score was built from housing, water supply, sanitation, electricity and possessions, all of which are properties of a household. Several people in the sample may share one household and therefore one wealth value. The row is a person; one of its columns describes something larger.

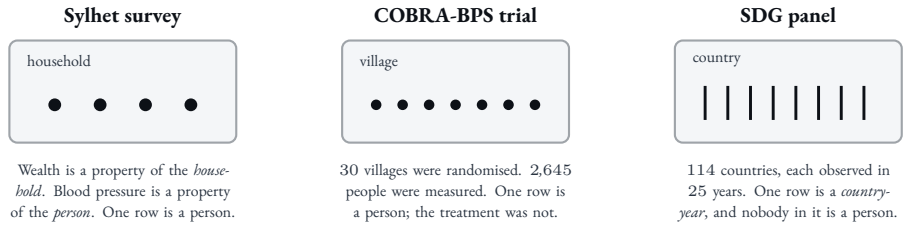


Figure 11: Three studies with three different units of observation. Only one of them has a person in the row.

In the COBRA-BPS trial, discussed in Chapter 1, thirty communities were randomly assigned to the intervention or to usual care, and 2,645 adults with hypertension were enrolled and measured. The treatment was given to a village. The blood pressure belongs to a person. Two people in the same village received the same assignment and share whatever else their village has in common, so their outcomes are not independent in the way that two people drawn at random from the country would be. The design has a name, cluster randomisation, and an analysis that ignores the clustering will claim more precision than the study earned. Chapters 13 and 26 return to it.

In a study of national development indicators, the rows are neither people nor villages. A recent analysis of interdependence between the Sustainable Development Goals used a balanced panel of 114 countries observed annually from 2000 to 2024, so one row is a country in a year, and there are no individuals in the dataset at all. The same study illustrates the role idea of Section 5 in an unusually direct way: it classifies goals as ”drivers” and ”receivers” according to the direction of the linkage being examined, and a goal that is a driver in one linkage is a receiver in another.

That analysis is a preprint and is under review at *Scientific Reports*: reference 4.. It appears here only as an example of a dataset whose rows are not people.

The point

Say what one row is before saying anything else about a study. The answer decides what the sample size means, whether the rows can be treated as independent, and which of the columns describe the row rather than something it belongs to.

Watch out

A statistical method with the word ”causality” in its name is not thereby a method for establishing causation. Several tests in wide use assess whether one series helps forecast another, which is a statement about prediction in time. Chapter 15 sets out what establishing causation actually requires.

7.1 What the table is called

The table this chapter has been building towards has a professional name. A *data dictionary*, sometimes called a codebook, is the document that says what every column of a dataset contains. Every dataset that can be reused has one, and the ones that cannot are usually the ones that do not.

Beyond the five columns already described, a working data dictionary records the variable name exactly as it is spelled in the file, the permitted values and the code used for each, and the code used for missing.

Watch out

The last of those causes real damage. A missing value coded as 9, or 99, or -1, in a column where those are also possible real values, will be analysed as data. Nobody notices until a mean parity of 9.4 appears in a table.

The point

Write the data dictionary before collecting anything. Written afterwards it is documentation. Written first it is a protocol, and it forces every operational definition to be settled while changing one is still cheap.

8 Chapter summary

1. Every study, whatever its design, arrives at the analysis as a rectangle. The rows are units of observation and the columns are variables.
2. The two questions to ask first of any study are what one row is and what one column is. A sample size larger than the number of people in the study means the rows are not people.
3. "The population" names at least four different groups: target, source, study population and sample. Something is lost at every step, and the loss between approached and sampled is usually itemised in the paper.
4. A parameter summarises a population, is fixed, and is never observed. A statistic summarises a sample, is always calculable, and takes a different value in every sample. Greek letters mark the first, roman letters the second, and a hat marks an estimate.
5. Variables are categorical or numerical. Categorical variables are binary, nominal or ordinal; numerical variables are discrete or continuous. The kind constrains what may sensibly be done with the column.
6. Ordinal categories have an order but no known distance between them.
7. The same underlying measurement frequently appears twice in one table, once as recorded and once as categorised. The categorised version reflects a decision made by the researchers.
8. Cutting a measured variable at a threshold discards distinctions within each category and creates one at the boundary. It buys usability, and any effect reported afterwards refers to the categories rather than to the original units.
9. Columns are measured, reported or constructed. A constructed column carries every decision made in constructing it, and its operational definition is the rule that produced it.
10. Outcome, exposure, predictor and covariate are roles. They are assigned by the research question and are not properties of columns.

11. Asking a different question of the same data reassigns the roles without changing any value. Exposure is a causal word; predictor is not.
12. An association is a feature of a dataset. Causation is a feature of the world. Correlation is one narrow measure of one narrow kind of association.
13. Four explanations produce an association: chance, a common cause, reverse causation, and a real effect. Design, not sample size, narrows the list.
14. The unit of observation is not always a person, and some columns describe something larger than the row that carries them.

9 Key terms

Association A relationship in a dataset such that knowing one variable tells you something about another. See Definition 6.1.

Binary A categorical variable with two categories.

Categorical variable A variable that places each unit into a group.

Causation A relationship such that intervening on one variable would change another. Not contained in any dataset.

Continuous A numerical variable that can take any value in a range.

Correlation One measure of association between two numerical variables that move together approximately linearly. Chapter 24.

Covariate A variable measured and accounted for so that the relationship between exposure and outcome can be seen more clearly.

Discrete A numerical variable that can take only separated values, typically because it is counted.

Estimate A statistic used to stand in for a parameter. Marked with a hat.

Exposure The variable whose effect on the outcome is being asked about. The word carries a causal claim.

Nominal A categorical variable with more than two unordered categories.

Numerical variable A variable that records a quantity.

Operational definition The exact rule by which a study assigns a value to a variable. See Definition 4.2.

Ordinal A categorical variable whose categories have an order but no known distance between them.

Outcome The variable a study is trying to explain or predict.

Parameter A numerical summary of a population. Fixed, and never observed.

Predictor A variable used to forecast the outcome, with no claim that it produces the outcome.

Reverse causation The case in which the outcome influenced the exposure rather than the other way round.

Sample The units that produced data.

Source population The group that could actually have been sampled.

Statistic A numerical summary of a sample. Always calculable, and different in every sample.

Study population The group that was approached.

Target population The group the answer is meant to apply to.

Unit of analysis The thing the research question is about.

Unit of observation The thing one row of the dataset describes. See Definition 1.1.

Variable One thing recorded the same way on every unit. A column.

N, n Population and sample size.

μ, $\bar{x}$ Population and sample mean.

σ, s Population and sample standard deviation.

π, $\hat{p}$ Population and sample proportion.

10 Exercises

The exercises below are built on a single published study: a cross-sectional survey of hypertension among adults aged 35 and over in two rural subdistricts of Sylhet district, Bangladesh (reference 1.). Its numbers are re-typeset here as matters of fact; its text is not reproduced. Readers who want the full paper should obtain it from the publisher, where it is free to read.

THE STUDY, IN BRIEF

Design. Population-based cross-sectional survey, August 2017 to January 2018, in the Zakiganj and Kanaighat subdistricts of Sylhet district, Bangladesh.

Participants. Adults aged 35 years and older, drawn at random from a 2016 census of the study area. Pregnant women excluded. Of 2,039 people approached, 1,810 completed: 864 men and 946 women.

Measurement. An adapted WHO STEPS instrument administered at home by trained community health workers, followed by physical measurements. Blood pressure was taken three times at roughly ten-minute intervals with a calibrated digital machine, and the average of the last two readings was used.

Definition of the outcome. Systolic pressure of at least 140 mm Hg, or diastolic pressure of at least 90 mm Hg, or currently taking antihypertensive medication.

Principal findings. Prevalence of hypertension 18.7% overall (95% CI 17.0 to 20.6); 18.8% in men and 18.7% in women. Among those found to be hypertensive, 44.5% were newly identified. High waist circumference was associated with hypertension in both sexes: adjusted odds ratio 4.0 (95% CI 2.5 to 6.4) in men and 2.8 (95% CI 2.0 to 4.1) in women.

Part A. Rows and columns

1. State what one row of this study's dataset represents, and how many rows the dataset has.

2. A different study of diabetic retinopathy reports findings on 240 eyes belonging to 130 patients. What is the unit of observation, and why should a reader be cautious about a claim based on a sample size of 240?

3. The wealth score in the Sylhet survey was built from the type of housing, the source of drinking water, the type of toilet, the availability of electricity, and a list of household possessions. Whose property is that score: the participant's, or something else's? What follows for two participants living in the same house?

Part B. Populations

4. Name the target population, the source population, the study population and the sample for this survey.

5. The authors note that in an earlier national survey, Sylhet division had the lowest recorded prevalence of hypertension of the country's eight divisions. Which of the four populations does that fact bear on, and what does it imply about the study's conclusions?

6. Of the 229 people who did not complete the survey, 42 had died and 76 had migrated out of the area. For each of those two reasons, state whether it is plausibly related to blood pressure, and say which direction the resulting distortion would run in.

Part C. Parameters and statistics

7. For each quantity, state whether it is a parameter or a statistic: 1,810; 18.7%; the proportion of all rural Bangladeshi adults aged 35 and over who have hypertension; 77.0 cm.

8. Among the 946 women in the sample, 177 met the study's definition of hypertension. Using equation (3.2), calculate $\hat{p}$ for women and express it as a percentage to one decimal place.

9. The study reports a prevalence of 18.7% with a 95% confidence interval of 17.0 to 20.6. Explain why an interval is attached to this figure but no interval is attached to the count of 339.

Part D. Kinds of variable

10. Classify each of the following columns as binary, nominal, ordinal, discrete or continuous: sex; wealth tertile; systolic blood pressure in mm Hg; number of servings of fruit and vegetables per day; subdistrict of residence; hypertension.

11. Body mass index appears in the study's descriptive table both as a median of 20.3 kg/m^2 and as three categories. Explain in one sentence why these are not two variables.

12. Waist circumference was cut at 90 cm for men. Consider three men with measurements of 72.0, 89.5 and 91.0 cm. Which two are treated as alike by the categorised variable, and which two are treated as different? Comment on whether that grouping reflects anything about the three men.

13. A man taking antihypertensive medication is measured at 128/82 mm Hg. According to this study's operational definition, is he counted as hypertensive? State the part of the definition that decides it.

Part E. Roles

14. For the question *does a high waist circumference raise the risk of hypertension?*, assign a role to each of these columns: waist circumference; hypertension; age; physical activity.

15. Now assign roles to the same four columns for the question *does physical activity lower the risk of hypertension?* State which columns changed role and which did not.

16. A hospital wishes to flag patients at high risk of readmission so that a nurse can telephone them. A model finds that patients who arrive by ambulance are much more likely to be readmitted. Should arrival by ambulance be called a predictor or an exposure? Would it make sense to reduce readmissions by discouraging ambulance use?

Part F. Association and causation

17. Among men in this survey, current smokers had an unadjusted odds ratio for hypertension of 0.5 (95% CI 0.4 to 0.7). State what that number says about the dataset, and what it does not say about smoking.

18. List the four explanations that can produce an association, and say which one the study's authors raise by name for the smoking result.

19. Adjustment moved the smoking odds ratio from 0.5 to 0.7, with the confidence interval widening to 0.5 to 1.0. State one thing this tells you and one thing adjustment cannot do.

20. The following sentence is invented, not quoted: *"Physical activity reduced the risk of hypertension among men in this population."* Identify the word that makes a causal claim, rewrite the sentence so that it says only what a cross-sectional study can support, and name the study design that would be needed to support the original wording.

11 Solutions

Part A

1. One row is one adult aged 35 or older who completed the survey. There are 1,810 rows, of which 864 describe men and 946 describe women.

2. The unit of observation is an eye, not a patient. The dataset has 240 rows and 130 people, so on average each person contributes almost two rows. The two eyes of one person are more alike than two eyes taken from different people, because they share that person's diabetes, blood pressure, age and treatment. An analysis treating 240 eyes as 240 independent observations credits the study with more information than it has, and will report intervals that are too narrow. The unit of observation is the eye; the unit of analysis is arguably the patient.

3. The wealth score is a property of the household, not of the participant. Two participants living in the same house necessarily receive the same wealth value, so that column does not vary independently from row to row. This is the same structural point as Exercise 2: a column can describe something larger than the row that carries it.

Part B

4. Target: adults aged 35 and older in rural Bangladesh. Source: adults aged 35 and older listed in the 2016 census of the Zakiganj and Kanaighat subdistricts. Study population: the 2,039 people selected at random and approached. Sample: the 1,810 who completed the survey.

5. It bears on the distance between the target population and the source population. If Sylhet has the lowest prevalence of the eight divisions, then a prevalence estimated in two of its subdistricts is likely to understate the figure for rural Bangladesh as a whole. The estimate is a good description of the sample and a questionable estimate for the target, which is a limitation of external validity rather than an error in the study.

6. Death is clearly related to blood pressure: hypertension raises cardiovascular mortality, so the 42 people who died before being surveyed were more likely than average to have been hypertensive. Removing them tends to make the surveyed group healthier than the population it was drawn from, so the reported prevalence is likely to be an underestimate.

Migration is related to age, employment and wealth, and probably to health. Its direction is harder to argue. Migrants out of a rural area are typically younger and more economically active, and younger people in this sample had substantially lower odds of hypertension, so losing them would tend to push the reported prevalence up. The honest answer is that the direction is plausible rather than established.

Part C

7. 1,810 is a statistic: it is a count taken from the sample, and a different survey would have produced a different number. 18.7% is a statistic. The proportion of all rural Bangladeshi adults aged 35 and over with hypertension is a parameter: it has a value and nobody has observed it. 77.0 cm, the median waist circumference in the sample, is a statistic.

8. Using equation (3.2) with $a = 177$ and $n = 946$:

$$\hat{p} = \frac{177}{946} = 0.187,$$

which is 18.7% to one decimal place. This matches the figure the study reports for women. The division is the reader's; the counts and the percentage are both in the paper.

9. The count 339 is a fact about the sample and involves no uncertainty: exactly that many people in this dataset met the definition, and recounting would give the same answer. The percentage 18.7% is being used for something more ambitious. It is an estimate of π, the corresponding proportion in a population that was not measured, and a different sample would have produced a different estimate. The interval expresses how far the estimate might reasonably sit from the parameter. Precisely what "95%" means is the subject of Chapter 8.

Part D

10. Sex is binary. Wealth tertile is ordinal: three categories in a genuine order with no known distance between them. Systolic blood pressure in mm Hg is continuous. Number of servings per day is discrete, being a count. Subdistrict of residence is nominal, since Zakiganj and Kanaighat have no natural order; with only two categories it is also binary, and either answer is defensible provided the reason is given. Hypertension is binary, and it is constructed rather than measured.

11. They are one variable presented in two forms: the same measurement on the same people, once as the number recorded and once sorted into categories chosen by the researchers.

12. The categorised variable treats the men at 72.0 and 89.5 cm as alike, since both fall below the 90 cm threshold, and treats the men at 89.5 and 91.0 cm as different. Neither grouping reflects the measurements. The two men classed together differ by 17.5 cm; the two men classed apart differ by 1.5 cm, which is within the range of variation one would expect between two careful applications of a tape measure. This is what a threshold costs, and it is the trade discussed in Section 4.3.

13. Yes, he is counted as hypertensive. The operational definition has three limbs joined by "or", and the third is currently taking antihypertensive medication. His readings of 128 and 82 satisfy neither of the first two limbs, but the medication limb is enough on its own. This is clinically sensible, since a treated patient with controlled pressure still has the condition, and it is not what the phrase "blood pressure of at least 140" would suggest by itself.

Part E

14. Waist circumference is the exposure. Hypertension is the outcome. Age is a covariate. Physical activity is a covariate.

15. Physical activity is now the exposure. Hypertension is still the outcome. Age is still a covariate. Waist circumference has changed from exposure to covariate.

Two columns changed role and two did not, and no value in the dataset changed. This is the point of Section 5.1.

16. It should be called a predictor. Arriving by ambulance is a marker of how unwell and how poorly supported a patient was at the point of admission, and those are the things that raise the risk of readmission. Discouraging ambulance use would do nothing to the underlying risk and would delay care; it would remove the marker and leave the condition. This is the distinction of Section 5: a variable can identify whom to watch without being something worth changing.

Part F

17. It says that in this dataset, among these men, the odds of meeting the study's definition of hypertension were about half as large among current smokers as among non-smokers, and that an association this size is unlikely to be produced by sampling variation alone in a sample of this size. It says nothing about what would happen to a man's blood pressure if he started or stopped smoking. The first is a statement about a rectangle of data; the second would be a statement about the world.

18. Chance, a common cause, reverse causation, and a real effect. The authors raise reverse causation by name, suggesting that people already aware of high blood pressure may have changed their smoking habits. They then argue against it, because about half of those found to be hypertensive in this survey were unaware of the condition and so had nothing to react to.

19. It tells you that part of the apparent protection was explained by the other variables in the model: once age, wealth and the rest were accounted for, the estimate moved towards one and the interval came to include it. What adjustment cannot do is account for anything that was not measured. It is a partial repair whose reach is limited to the columns present in the dataset, which is why randomisation, which balances unmeasured variables too, occupies the position it does.

20. The word is "reduced", which asserts that a change in physical activity produced a change in risk. A defensible rewriting is: *among men in this survey, higher reported physical activity was associated with lower odds of hypertension.* That states what was found in the data and claims nothing about intervening.

Supporting the original wording would require a design that establishes the order of events and controls what determines the exposure. A randomised trial of a physical activity intervention would do it. A prospective cohort measuring activity in people free of hypertension and following them to see who develops it would be weaker but would at least settle the temporal order, which a cross-sectional survey cannot.

12 References and further reading

1. 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. Published under CC BY-NC; its figures are re-typeset in this chapter as facts and none of its text is reproduced.
2. Jafar TH, Gandhi M, de Silva HA, Jehan I, Naheed A, Finkelstein EA, Turner EL, Morisky D, Kasturiratne A, Khan AH, Clemens JD, Ebrahim S, Assam PN, Feng L. A community-based intervention for managing hypertension in rural South Asia. *N Engl J Med* 2020;382(8):717–726. PMID 32074419. doi:10.1056/NEJMoa1911965.
3. Khair Z, Rahman MM, Kazawa K, Jahan Y, Faruque ASG, Chisti MJ, Moriyama M. Health education improves referral compliance of persons with probable diabetic retinopathy: a randomized controlled trial. *PLoS ONE* 2020;15(11):e0242047. PMID 33180863. PMC7660573. doi:10.1371/journal.pone.0242047.
4. Fahim MMM, Imran MJH, Molla MN, Debnath L, Shil T, Pranto EB, Likhon MMR, Saad MSS, Karim MR. Drivers, receivers, and dynamic linkages: the directed structure of SDG interdependence, 2000–2024. arXiv:2601.20875 [stat.AP], 2026. **A preprint, under review at *Scientific Reports***, and is cited here only as an example of a dataset whose rows are not individuals.
5. Daniel WW. *Biostatistics: A Foundation for Analysis in the Health Sciences*, 9th ed. Hoboken: Wiley. Sections 1.2 to 1.4, pp. 3–13, on population, sample and measurement scales. A textbook, so it carries no PMID or DOI.
6. World Health Organization. *STEPwise Approach to NCD Risk Factor Surveillance (STEPS)*. The survey instrument used by reference 1.. A methodological resource rather than a journal article, so it carries no PMID.

13 For students in the mentorship programme

Everything above this section stands on its own. This section does not.

Before the next session

Build the variable table for your own study. One row per variable, and five columns:

1. Name.
2. Role: outcome, exposure, predictor or covariate.
3. Kind: binary, nominal, ordinal, discrete or continuous.
4. How a value gets into it: the procedure that measured it, or, if it is constructed, the rule that constructs it.
5. Unit, or the levels if it has no unit. A binary column has levels rather than a unit, and writing "yes / no" there is right, not a gap.

Then write down a second research question that could be asked of the same data, and fill in the Role column again for it. Whatever changes between the two versions is the substance of this chapter. Bring both.

Name	Role	Kind	How a value gets into it	Unit or levels
Hypertension	outcome	binary	$\geq$ 140/90, or on medication	yes / no
Waist circumference	exposure	continuous	tape, standing, to 0.1 cm	cm
Age	covariate	discrete	reported, completed years	years
Wealth tertile	covariate	ordinal	5 household items, scored, cut in thirds	tertile
Physical activity	covariate	ordinal	STEPS questionnaire, MET-min per week, *then cut* into 3 bands	3 bands

Five columns, one row per variable. This is the first page of a protocol, and it is what the session is for.
Kind is a fact about the column. Role is a fact about the question. Only Role moves.
Fill Role in twice, for two different questions, and whatever changes between them is the whole of today.
Physical activity is the row worth a second look: it was recorded as a number and reported as bands, and both belong in the table.

Figure 12: The same table, filled in for the Sylhet survey, as a worked standard.

Reading

Reference 1. is free to read at the publisher and on PubMed Central. Read its Methods section rather than its abstract, and specifically the paragraphs describing how blood pressure was measured and how hypertension, body mass index, waist circumference and the wealth index were defined. Those paragraphs are the operational definitions of Section 4.4.

Questions this chapter deliberately left open

Question	Settled in
How should a categorical variable actually be summarised?	Chapter 3
Which measure of centre and spread suits a numerical variable?	Chapter 4
What does the interval around 18.7% mean?	Chapter 8
How far does an estimate move from sample to sample?	Chapter 7
What is the difference between a confounder, a mediator and an effect modifier?	Chapter 15
How much precision does cutting a variable at a threshold cost?	Chapter 17
Which designs can rule out which of the four explanations?	Chapters 12 and 13
How is a correlation coefficient computed and read?	Chapter 24
What does an adjusted odds ratio actually adjust?	Chapters 25 and 26
Why does randomising villages rather than people change the analysis?	Chapters 13 and 26

Next

Session 3: Describing Categorical Data. Counts, proportions and percentages, the denominators that make them mean something, and two-way tables. It takes the categorical half of this chapter's classification and does the work this chapter deliberately left undone.