Stat.BD

CLASS NOTES • CHAPTER 3

Describing Categorical Data

Counts, and the four things people do to them

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

Mentor: Muhtasim Munif Fahim, MSc

Chapter 3

Describing Categorical Data

A categorical column cannot be averaged. Counting is not a preliminary to describing it; counting is the whole of describing it, and almost everything that goes wrong goes wrong in the denominator.

Chapter 2 classified the columns of a dataset by what kind of thing each one holds. Some name a group and some measure a quantity, and the difference decides what may be done with them. This chapter takes the first family, the columns that name a group, and works out what a complete description of one of them looks like.

That description is shorter than it sounds. A categorical column has no mean, no median and no spread, so the only honest summary of it is how many observations fell into each category and out of how many. Everything else in this chapter is arithmetic on those counts, or a warning about what that arithmetic conceals.

The warnings occupy more space than the arithmetic, and deliberately so. The calculation is division. The difficulty is that a printed percentage has thrown away the two numbers it came from, and a reader who cannot recover them is not reading a result but accepting one.

What this chapter establishes

1. That the frequency distribution, a list of counts against categories with a total that matches n, is the complete description of a categorical column, and that proportions and percentages add no information to it.
2. That ratio, proportion, percentage and rate are four different objects, and that only one of them has time in its denominator.
3. That prevalence is a proportion that has to be reported with a population, a place, a period and a case definition, and that dropping any of the four changes the number.
4. That the denominator of a percentage is a choice which the writer makes and the reader must recover, and that a single published table may use more than one.
5. That every cell of a two-way table yields three correct percentages answering three different questions, and that the one to print is the one that makes the groups being compared each sum to a hundred.
6. That a percentage printed alone conceals its own fragility, and that the count and the denominator printed beside it restore what it hid.

Notation

n	the number of observations in the sample, or in the group under discussion
a	a count: the number of observations falling into one category
$\hat{p}$	a sample proportion, a/n; the hat marks it as estimated from data
π	the corresponding population proportion, never observed
k	the number of categories a variable can take

Chapter 2 introduced π and $\hat{p}$ as a parameter and a statistic. This chapter works almost entirely with $\hat{p}$, because a description is a statement about the people who were measured. How far $\hat{p}$ may sit from π is the subject of Chapters 7 and 8.

1 The frequency distribution

A categorical column holds, for each observation, the name of the group that observation belongs to. Nothing can be added, subtracted or averaged. The only operation available is counting how many times each name occurs, and the result of that counting is the frequency distribution.

Definition 1.1 • Frequency distribution

For a categorical variable with k categories, the *frequency distribution* is the list of counts $a_1, a_2, \ldots, a_k$, where a_j is the number of observations falling into category j. The counts satisfy

$$a_1 + a_2 + \cdots + a_k = n,$$

where n is the number of observations described. A frequency distribution whose counts do not sum to n is incomplete, and the shortfall is itself information.

Frequency
A count. The number of observations in a category. Sometimes called the absolute frequency, to distinguish it from the relative frequency.

That last sentence is the practical content of the definition. Checking that a published table adds up takes five seconds and finds real errors, and where it does not add up the missing observations are usually people who did not answer, which is a fact about the study rather than a fact about the arithmetic.

Intuition

A frequency distribution is a tally. It is what a ward clerk produces without being taught any statistics: a list of categories down the left, a stroke for each patient, and a total at the bottom that has to match the register.

1.1 Relative frequency: proportion and percentage

The counts answer how many. Dividing each count by the total answers what share, and makes two studies of different sizes comparable.

$$\hat{p}_j = \frac{a_j}{n} \tag{1.1}$$

$\hat{p}_j$ the sample proportion in category j, a number between 0 and 1
a_j the count in category j
n the total number of observations, the *denominator*

$$\text{percentage in category } j \;=\; 100 \times \hat{p}_j \tag{1.2}$$

A percentage is a proportion expressed in hundredths. It is the same quantity in different units, exactly as a length in centimetres is the same length in metres, and no information is created or lost in the conversion. Because the proportions are the counts divided by a common total, they sum to 1 and the percentages sum to 100, subject to rounding, which Section 8.2 takes up.

Years of schooling	Count	Proportion	Percentage
No education	333	$333/1810 = 0.184$	18.4%
1 to 5 years	1,126	$1126/1810 = 0.622$	62.2%
6 years or more	351	$351/1810 = 0.194$	19.4%
Total	**1,810**	1.000	100.0%

This is what was collected. Somebody asked 1,810 people and wrote down a tally. Everything to the right is arithmetic.

Nothing new is measured here. A proportion is the count divided by the total; a percentage is that proportion multiplied by a hundred. Three ways of writing one fact.

Figure 1: One categorical column of a published survey, shown as counts, as proportions and as percentages. Only the left-hand column was collected; the rest is arithmetic. Counts and printed percentages from reference 1..

Example 1.1 • Education in a rural Bangladeshi survey

A cross-sectional survey of 1,810 adults aged 35 and over in two rural subdistricts of Sylhet district recorded years of schooling in three bands. The counts were 333 with no education, 1,126 with one to five years, and 351 with six years or more.

The counts sum to 1,810, matching the number of participants, so no category is missing. Applying Equation 1.1 to the first band,

$$\hat{p} \;=\; \frac{333}{1810} \;=\; 0.184,$$

and Equation 1.2 turns this into 18.4%. The other two bands give 62.2% and 19.4%. The paper prints the counts and the percentages; the proportions in between are shown here to make the step visible.

The striking feature of this distribution is not the extremes but the middle. The modal category holds 62.2% of the sample, and it is one to five years of schooling. Any summary that does not preserve that fact has lost the most important thing the column had to say.

Watch out

Rounding a set of percentages to one decimal place can produce a column that sums to 99.9 or 100.1. This is arithmetic, not error, and it is discussed in Section 8.2. What it means in practice is that the counts are the data and the percentages are a rendering

of them, so where the two disagree the counts are authoritative.

1.2 What else a categorical column has

The opening claim of this chapter, that a categorical column has no mean and no spread, is exactly true for a nominal variable and slightly too strong for an ordinal one. Two summaries survive, and it is worth being precise about which.

Definition 1.2 • Mode

The *mode* of a categorical variable is the category with the largest count. It is defined for every categorical variable, nominal or ordinal, and a distribution may have more than one if two categories tie.

Modal category
The category with the largest count. The only measure of central tendency available to a nominal variable.

The mode is the only measure of centre a nominal variable admits. Blood group has no mean and no middle, because there is no order in which A, B, AB and O could be placed that is not arbitrary, but one of the four occurs most often and naming it is informative.

For an ordinal variable the order is real, so a middle observation exists and the median is available. What remains unavailable is the mean, because the gaps between adjacent categories are not equal and averaging them would treat those unequal gaps as though they were the same quantity.

In practice

Wealth in the Sylhet survey was reported in tertiles, with 610, 611 and 589 participants in the lowest, middle and highest. The distribution is almost flat by construction, since tertiles are defined to split a population into thirds, so the mode is barely informative here.

Education is more useful. Its modal category is one to five years of schooling, holding 1,126 of 1,810 participants. Since education is ordinal, the median is also available: the participant in the middle of the ordered list falls in that same band. A mean number of years cannot be computed from the banded column at all, because the bands do not say how many years each participant had.

The point

Nominal: the mode, and nothing else. Ordinal: the mode and the median, but never the mean. The moment a mean appears for a categorical column, either the column was numerical all along or something has gone wrong.

The median of an ordinal variable, and every summary available to a genuinely numerical column, are the subject of Chapter 4.

1.3 Cumulative frequency

For an ordinal variable, the counts can be accumulated down the table, giving the number of observations at or below each level. The result answers a different and often more useful

question than the individual counts do.

Example 1.2 • Schooling, at or below

The three education counts are 333, 1,126 and 351, or 18.4%, 62.2% and 19.4%. Accumulating the percentages gives 18.4, then $18.4 + 62.2 = 80.6$, then 100.0.

Read as a sentence, the middle figure says that four in five adults in this population had five years of schooling or fewer. That number appears nowhere in the paper and follows from it in one addition, which is ours.

For planning a health education programme it is more useful than any of the three individual percentages, because the question a programme faces is how many people are at or below a level of literacy, not how many are in a particular band.

Watch out

Cumulating requires an order, so it applies to ordinal variables and not to nominal ones. Accumulating blood groups down a table produces numbers, and they mean nothing, because the order they were printed in was arbitrary.

2 Four words that are not synonyms

Ratio, proportion, percentage and rate are used interchangeably in speech and are four different objects in print. The distinctions are not pedantry: a reviewer will mark the last of them, and a reader who takes one for another will misestimate a burden of disease.

Definition 2.1 • Ratio, proportion, percentage, rate

A *ratio* compares two counts to each other, $a : b$, where the two need not be parts of a common whole and the value has no upper bound.

A *proportion* divides a count by the total of which it is a part, a/n, and therefore lies between 0 and 1.

A *percentage* is a proportion multiplied by a hundred.

A *rate* divides a count by the amount of *time* over which the observations were at risk, so its units include a reciprocal time, as in cases per thousand person-years.

The word most often misapplied is rate. A cross-sectional survey measures everybody once, so no participant contributes any person-time and no rate is available from it. The figure such a study produces is a proportion, however frequently it is called a rate in the discussion section. The distinction matters because a rate and a proportion respond differently to the length of follow-up: a proportion of deaths rises simply because a study ran longer, whereas a rate does not.

Person-time
The sum, over all participants, of the time each was under observation and at risk. The denominator of a true rate. Chapter 23.

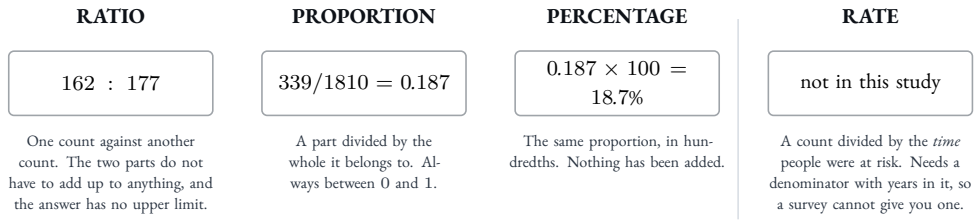

Figure 2: The first three are all present in the Sylhet survey. The fourth cannot be computed from a cross-sectional study at all, because nobody was followed for any length of time.

In practice

A hospital reports that 12% of its admissions were readmitted. That is a proportion, and it is not a readmission rate, because it says nothing about the window over which readmission was counted. The same patients observed for a year rather than thirty days would give a larger figure without anything about the hospital having changed.

2.1 The denominator that counts time

Of the four words, only *rate* requires a denominator that is not a count of people. A rate counts observation: how much watching was done, and for how long.

Definition 2.2 • Incidence rate

$$\text{incidence rate} = \frac{\text{number of new cases}}{\text{person-time at risk}} \tag{2.1}$$

person-time the sum, over everyone observed, of how long each was observed while still at risk. Two hundred patients followed for six months contribute one hundred person-years, and so does one patient followed for a hundred years

at risk observation stops when the event happens, when the participant is lost, or when the study ends, whichever comes first

Infections per 1,000 catheter-days is an incidence rate of exactly this shape, and it is why two wards reporting the same number of infections cannot be compared until the catheter-days behind each are known.

A cross-sectional survey has no person-time in it. It measures who has the condition at one moment, which is a prevalence, and no amount of rewording turns it into a rate. Choosing between a proportion, an odds and a rate for a given question is the subject of Chapter 23.

3 Prevalence

Prevalence is the proportion of a defined population that has a condition at a defined time. It is arithmetically ordinary, being a proportion like any other, and it is reported badly more often than almost any other summary in clinical research, because the definition that makes it meaningful lives outside the arithmetic.

$$\text{prevalence} = \frac{\text{number with the condition}}{\text{number in the defined population}} \tag{3.1}$$

numerator those meeting the case definition, which must be stated
denominator everybody in the population as defined, whether or not they were found to have the condition

Example 3.1 • Hypertension in rural Sylhet

Of the 1,810 adults surveyed, 339 met the study's definition of hypertension. By Equation 3.1 the prevalence is

$$\frac{339}{1810} = 0.187, \quad \text{that is } 18.7\%.$$

The paper reports 18.7% with a 95% confidence interval of 17.0 to 20.6. The division shown here is ours; the paper prints both the count and the percentage.

Four qualifications belong to that number and none of them is optional:

The population. Adults aged 35 years and over. The figure would be substantially lower in a population that included younger adults.

The place. Two rural subdistricts, Zakiganj and Kanaighat, in Sylhet district. Not Sylhet city, not Dhaka, and not Bangladesh.

The period. August 2017 to January 2018.

The case definition. Systolic pressure of at least 140 mm Hg, or diastolic of at least 90, or currently taking antihypertensive medication. Moving the systolic threshold to 130 raises the prevalence without one participant's blood pressure changing.

The point

A prevalence quoted without its population, place, period and case definition is not a smaller version of the result. It is a different claim, and usually a false one.

The case definition is a constructed column in the sense of Chapter 2: no participant was measured for hypertension. A rule was applied to two measured columns and one reported one.

3.1 Comparing a prevalence with another study's

Two prevalences from two studies are not two measurements of the same quantity unless four things match.

Who was counted. An age floor of 35 and an age floor of 18 describe different populations, and hypertension is strongly age-related.

What counted as a case. The Sylhet definition counts anybody taking antihypertensive medication, whatever their reading, so a perfectly controlled patient at 128/82 is a case. A

study defining hypertension purely on the reading will report a lower figure, and none of that difference is about the two populations.

Where, and when.

How it was measured. A single reading taken at one visit and the average of the second and third of three readings are not the same measurement, and the first tends to run higher.

The point

Where any of the four differ, the gap between two reported prevalences is partly a gap between two protocols. This is the commonest weakness in a discussion section, and it is the one a reviewer will find.

4 When a comparison is not a fair one

Section 3.1 listed four things that must match before two prevalences can be compared. There is a fifth, and unlike the other four it can be repaired arithmetically rather than only noted: the two populations may differ in their composition. A population with more old people in it will show more hypertension whatever its underlying health, because hypertension rises steeply with age. The comparison is then partly a comparison of age structures.

Two tools follow. The first detects the problem and needs nothing but the table. The second corrects for it, and needs the outcome broken down by the variable causing the trouble.

4.1 Stratification, and a comparison that reverses

A crude percentage computed over a whole group, and the same percentage computed separately inside each stratum of that group, can point in opposite directions. The arithmetic is not wrong in either case. They are answers to different questions.

Example 4.1 • Two hospitals, constructed

The figures below are **constructed**, not taken from a study. They are chosen so that the reversal is exact and every total reconciles.

Survival to discharge	**Hospital A**	**Hospital B**
Mild cases	$99/100 = 99.0\%$	$882/900 = 98.0\%$
Severe cases	$720/900 = 80.0\%$	$78/100 = 78.0\%$
Everybody	$\mathbf{819/1000 = 81.9\%}$	$\mathbf{960/1000 = 96.0\%}$

Hospital B has much the better overall survival. Hospital A has better survival among mild cases and better survival among severe cases. Both statements are true.

The explanation is in the denominators rather than in the percentages. Hospital A treats 900 severe cases and Hospital B treats 100. The crude figure is dominated by

Example 4.1 • *continued*

how many severe cases each hospital saw, which is a fact about referral rather than about care.

Intuition

A crude percentage over a mixed group is a weighted average of the group's parts, and the weights are how many of each part there were. Change the mix and the crude figure moves, even if nothing about any part has changed.

This pattern is known as Simpson's paradox. The name is unhelpful: nothing is paradoxical and no arithmetic is wrong. It is a consequence of averaging over groups of unequal size.

Watch out

A reversal establishes that the crude comparison is unfair. It does not establish what a fair comparison would show. Stratifying by severity corrects for severity and leaves everything else uncorrected, and there may be a further variable that reverses the result again. Which variables must be accounted for, and what it means to say one explains an association, is the subject of Chapter 15.

4.2 Direct standardisation

Where the composition causing the trouble is known and the outcome is reported within each of its categories, the comparison can be repaired. Every population being compared is re-scored on one common composition, called the standard population. Only the rates are allowed to differ.

Definition 4.1 • Directly standardised rate

For a population whose rate in stratum i is p_i, and a standard population in which stratum i holds a share w_i of the whole,

$$p_{\text{std}} = \sum_i w_i p_i, \qquad \text{where} \quad \sum_i w_i = 1. \tag{4.1}$$

p_i the rate observed in stratum i of the population being standardised, for example the prevalence of hypertension among those aged 45 to 54

w_i the share of the *standard* population falling in stratum i, the same value for every population being compared

p_{std} the rate the population would have shown if its composition had been the standard one

The whole mechanism sits in the requirement that w_i is the same for everybody being compared. Nothing else about the calculation matters as much. It also follows that p_{std} is not a property of the population alone: change the standard and the number changes. A standardised rate is therefore never reported without naming its standard, and two standardised rates from different standards are not comparable.

Example 4.2 • Did hypertension rise, or did the population age?

A national analysis of two waves of the Bangladesh Demographic and Health Survey reports that hypertension among adults aged 35 and over rose from 25.84% in 2011 to 39.40% in 2017–18. Bangladesh also aged over that period, so the obvious objection is that the rise is a change in the population rather than in its health.

The same source reports prevalence within each age band, and the age distribution of the two waves combined. Taking that combined distribution as the standard:

Age band	w_i (%)	2011 p_i	2018 p_i
35–44	36.01	17.46	28.45
45–54	27.44	24.90	38.48
55–64	18.66	30.05	46.75
65–74	11.26	39.30	51.46
75+	6.63	41.66	59.85

Applying equation (4.1) to the 2011 column, the youngest band contributes $0.3601 \times 17.46 = 6.29$ points, the next $0.2744 \times 24.90 = 6.83$, and so on. The five contributions sum to 25.91%. The same sum on the 2018 column gives 39.29%.

Adults 35+	2011	2018
Crude	25.84%	39.40%
Age-standardised	25.91%	39.29%
Ratio, crude	1.52	
Ratio, standardised	1.52	

The adjustment moves each figure by less than a tenth of a percentage point and leaves the ratio unchanged. The rise is not an artefact of an ageing population.

The standardised figures and both ratios are computed here; the source reports the crude figures, the age bands and the weights.

The point

An adjustment that changes nothing is a result and belongs in the paper. It answers an objection a reader would otherwise be left holding, and the objection was reasonable.

4.3 Where the adjustment does bite

The two waves above have similar age structures, which is why the correction was small. The tool earns its place when the populations genuinely differ, and a comparison the course has already met is one such case.

Example 4.3 • Rural Sylhet against the nation

The Sylhet survey of Section 3 reports a hypertension prevalence of 18.7% among adults aged 35 and over. The national analysis reports 39.40% for the same age range in an overlapping period, and 39.08% for Sylhet division. The Sylhet figure is less than half the national one, and both studies are peer reviewed.

Is the Sylhet sample younger? Its age distribution is 34.1, 30.3, 16.9 and 18.7 per cent across the bands 35–44, 45–54, 55–64 and 65 and over. Combining the national 65–74 and 75+ rates by their weights gives a single 65+ rate of 54.57% for 2018. Applying the national age-specific rates to the Sylhet age structure:

$$0.341(28.45) + 0.303(38.48) + 0.169(46.75) + 0.187(54.57) \;=\; 39.5\%.$$

That is what rural Sylhet would have shown if it followed the national pattern. It observed 18.7%, a ratio of observed to expected of 0.47.

Age structure therefore explains essentially none of the gap. What remains are the four differences of Section 3.1: two rural subdistricts against a whole nation, a case definition that counts anyone on medication, a blood-pressure protocol averaging the second and third readings, and a sampling frame drawn from a 2016 census list.

Every figure in this example except the two published prevalences and the age distributions is computed here.

The point

Ruling one explanation out without being able to choose among the rest is a complete and honest result. The temptation is to decide which study is right; the evidence here does not support deciding, and saying so is the correct end point.

5 Decisions that shape a categorical column

Three decisions stand between a column of raw answers and a printed table. None of them is arithmetic, and each changes what the table says.

5.1 The order of the categories

For an ordinal variable the order is a property of the data and must be preserved. Sorting education bands by size destroys the only structure the column has, and makes it impossible to see whether the distribution rises or falls across the bands.

For a nominal variable there is no natural order and the writer chooses one. Commonest first serves most readers; alphabetical is defensible and helps when a reader is looking for a particular category. What matters is that the same order is used in every table in the paper, so that a reader who has learnt the layout once does not have to relearn it.

5.2 Missing values and the denominator

Missing is not a category and it is not zero. A participant who did not answer a question about smoking is neither a smoker nor a non-smoker, and placing them in either group invents data.

The decision that has to be made and stated is the denominator. If eleven of 1,810 participants did not answer, the percentage of smokers may be computed out of 1,810 or out of 1,799. Both are defensible. Only one of them is usually stated.

Watch out

Where a paper's categories sum to less than n, the difference is the missing observations, and the reader has learnt something the text may not say. Where the categories sum exactly to n but n differs from the sample size given in the Methods, the missing observations have been dropped silently.

Example 5.1 • The same count, two denominators

A survey of 500 people asks about smoking. 470 answer and 30 do not; of those answering, 141 say yes.

$$\frac{141}{470} = 30.0\% \qquad \text{and} \qquad \frac{141}{500} = 28.2\%.$$

Both are defensible and they answer different questions. *Of those who answered* is the honest description of what was measured. *Of everybody* is the right denominator for planning a service, because the thirty who did not answer still exist.

What is not defensible is reporting 30.0% under a heading that says $N = 500$. That is not a choice between two denominators; it is one denominator printed over another, and the gap of two percentage points is small enough that nothing looks wrong. All four numbers are constructed here.

Whether the missingness is itself informative, and what may be done about it, is the subject of Chapters 14 and 15. This chapter goes no further than requiring that the denominator be stated.

5.3 Ordinal columns, and what they earn

An ordinal column has an order without a distance. Four of the Sylhet columns are of this kind: schooling, wealth tertile, physical activity and servings per day. ASA grade, tumour stage and a pain score out of ten are the same object in clinical dress: ASA III is worse than ASA II, nobody can say by how much, and the average of a II and a IV is not a III.

What the order earns is a middle category and a cumulative reading, in either direction.

Schooling	%	cumulative %
None	18.4	18.4
1–5 years	62.2	**80.6**
≥ 6 years	19.4	100.0

Four adults in five had five years of schooling or less; equivalently, one in five had six years or more. Both sentences need the order to exist and neither is available for a column of bare names. The cumulative figure 80.6 is computed here, as $18.4 + 62.2$.

What the order does not earn is a mean. There is no arithmetic on the labels, and taking one assumes the gaps between categories are equal. That assumption is usually false and is almost never stated. How scores of this kind are constructed and whether they may be treated as numerical is the subject of Chapter 14.

5.4 Collapsing categories

Combining categories reduces a table and makes a comparison simpler. It is the same trade Chapter 2 described for a threshold applied to a measured column: what is bought is a rule somebody can act on, and what is paid is every distinction inside the new group.

Example 5.2 • Any schooling against none

The Sylhet survey's three education bands can be collapsed into two by merging the upper two, giving 333 (18.4%) with no education and 1,477 (81.6%) with some. The count of 1,477 and the percentage 81.6% are ours; the paper prints the three bands.

The collapsed table is easier to read and it conceals the finding. In the original distribution the largest group by a wide margin, 62.2% of the sample, had between one and five years of schooling. Collapsed, those participants are counted as educated alongside people with secondary schooling and above. A reader of the two-category table would conclude that four in five adults in this population were educated, which is true only under a definition of educated that includes a single year of primary school.

The reverse case is available in the same table. Body mass index was reported in three categories: 29.3% underweight, 56.7% normal, and 14.0% overweight or obese. Collapsing to normal against not normal would merge two opposite clinical problems into one number. Nobody made that collapse, but the table would have permitted it.

The point

Collapse categories when the groups being merged would be managed the same way. Do not collapse because a table has too many rows.

6 The denominator

A percentage is a pair of numbers reported as one. The numerator is usually easy to identify. The denominator is a choice the writer made, and recovering it is the central skill of reading a descriptive table.

Definition 6.1 • Denominator

The *denominator* of a percentage is the set of observations the numerator is being expressed as a share of. It is not necessarily the sample size, and a single table may use several.

6.1 A table with two denominators

Published tables change denominator, legitimately, and say so in a row header rather than in a warning.

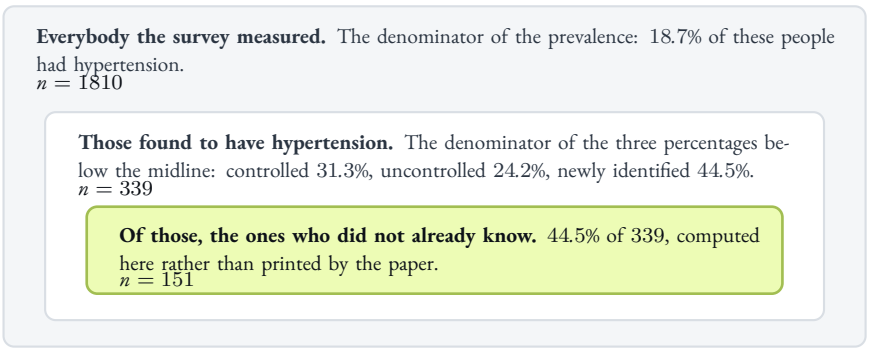

Figure 3: Three nested groups, and two denominators, inside one published table. Counts and percentages from reference 1.; the count of 151 is ours.

The upper half of the table describes everybody: normal blood pressure, prehypertension and hypertension, as percentages of 1,810. Below that, the table turns to the people found to have hypertension and describes them: controlled, uncontrolled, and newly identified, as percentages of 339. Both halves are correct and the change of denominator is announced by a row that gives the new n.

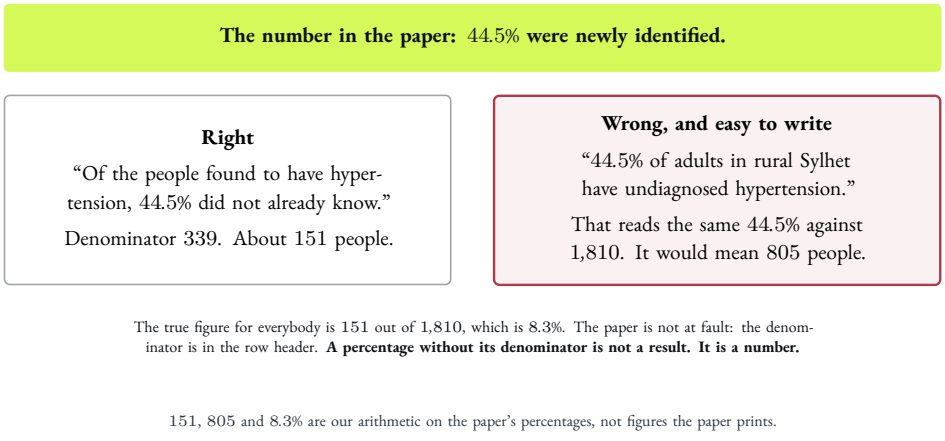

Figure 4: The same published percentage, written into two sentences. One is what the paper reports; the other is what a reader in a hurry produces. The counts 151 and 805 and the share 8.3% are ours.

Example 6.1 • What 44.5% is a percentage of

The survey reports that 44.5% of those found to have hypertension were newly identified, meaning they were not aware of the diagnosis before the survey reached them.

Read against the correct denominator of 339, that is about 151 people. Read, wrongly, against the sample size of 1,810, it would be about 805 people. Both counts are ours; the paper prints only the percentage and the two possible denominators.

Expressed as a share of everybody surveyed, the undiagnosed are $151/1810 = 8.3\%$. So two sentences are available and both are true:

Of the people who have hypertension in this population, 44.5% *do not know it.*

8.3% *of all adults aged* 35 *and over in this population have hypertension and do not know it.*

They differ by a factor of more than five, and they are for different readers. The first is a statement about the disease and argues for measuring blood pressure opportunistically rather than waiting for symptoms. The second is a statement about the population and is the one a health service would use to estimate how many cuffs, staff and tablets a screening programme requires.

The point

Choose the denominator that matches the decision the reader has to make, and then name it inside the sentence. *Of those with hypertension,* 44.5% costs four words and cannot be misread.

6.2 Where denominators go missing

Four places account for most of it.

A subgroup. Percentages computed within one sex or one age band, printed in a table whose heading gives the total sample size.

A filter question. A question asked only of those who answered an earlier one affirmatively changes n for every row after it.

Missing values. Dropped row by row, so that the denominator varies down the table and nothing announces it.

An abstract. Word limits remove the counts and leave the percentages, so the figure most widely read is the one least able to be checked.

The point

For every percentage read, write down the two numbers it came from. Where they cannot be recovered from the paper, that is a finding about the paper rather than a failure of the reader.

6.3 Small denominators

A percentage carries no information about how many observations produced it, and so a fragile figure and a solid one are printed identically.

Example 6.2 • Control among hypertensive men

Among the 162 men found to have hypertension, 23.5% had their blood pressure controlled. That percentage and that denominator are the paper's.

Applying the percentage to the denominator gives about 38 men; this count is ours, and the paper does not print it. Had four of those men been classified the other way, the figure would read 26.0% instead. A shift of four people in a survey of 1,810 moves the printed number by two and a half percentage points.

On the same page sits the prevalence of 18.7%, computed from 1,810 people. The two percentages are printed to the same precision and look equally settled. They are not, and nothing in the notation says so.

Watch out

This is not a claim that the paper is in error, and it is not a confidence interval. How much uncertainty attaches to an estimate, and how to express it in print, are the subjects of Chapters 7 and 8. The point here is narrower and is about reporting: a bare percentage conceals its own fragility, and a count printed beside it does not.

7 Two-way tables

Crossing two categorical columns produces a table of counts, one for each combination of categories. It is the first structure in this course that supports a comparison between groups, which is what a clinical question almost always is.

Definition 7.1 • Two-way table

A *two-way table*, or contingency table, cross-classifies n observations by two categorical variables with r and c categories, giving $r \times c$ cell counts. The row and column totals are called the *marginal* totals, and the sum of all cells is n.

Marginal total
A row or column total of a two-way table. It reproduces the frequency distribution of one variable, ignoring the other.

Only the cell counts are data. The margins, the grand total and every percentage in a printed contingency table can be reconstructed from them.

7.1 Three percentages for every cell

A single cell count can be divided by its row total, by its column total, or by the grand total. All three are arithmetically correct and they answer three different questions.

	Hypertension	No	
Men	162	702	864
Women	177	769	946
Total	**339**	**1471**	**1810**

Row percentage

162 ÷ 864, the shaded row.
Of the men, this share.

	Hypertension	No	
Men	162	702	864
Women	177	769	946
Total	**339**	**1471**	**1810**

Column percentage

162 ÷ 339, the shaded column.
Of those with it, this share are men.

	Hypertension	No	
Men	162	702	864
Women	177	769	946
Total	**339**	**1471**	**1810**

Percentage of **everybody**

162 ÷ 1810, everybody.
Of the whole survey, this share.

Figure 5: One cell, 162, divided three ways. The shaded group in each panel is the denominator. Counts 162, 177, 864, 946 and 1,810 are from reference 1.; the remaining cells, the margins and all three percentages are ours.

Example 7.1 • Hypertension by sex

Of 864 men, 162 were hypertensive; of 946 women, 177 were. The remaining two cells follow by subtraction: 702 men and 769 women were not. These four counts, and the margins built from them, are ours; the paper prints the two numerators, the two sex totals and the total of 339.

The cell containing 162 yields

$$\frac{162}{864} = 18.8\%, \qquad \frac{162}{339} = 47.8\%, \qquad \frac{162}{1810} = 9.0\%.$$

As sentences:

18.8% *of men had hypertension.* A row percentage, dividing by the row total.

47.8% *of the people with hypertension were men.* A column percentage, dividing by the column total.

9.0% *of the whole survey were men with hypertension.* A total percentage, dividing by n.

Only the first answers a clinical question about risk. The second describes the composition of a caseload, which is a service question rather than a clinical one. The third is rarely what anyone wants and is what a spreadsheet produces by default.

7.2 Tables larger than two by two

Nothing in Section 7.1 depended on both variables having two categories. Crossing a variable with r categories against one with c gives $r \times c$ cells, and each cell still yields a row percentage, a column percentage and a total percentage.

What changes is the amount of arithmetic and, more importantly, the amount of reading. A 4×3 table has twelve cells and thirty-six available percentages, of which a paper will print twelve. Which twelve were chosen is a decision the reader has to recover, and in a large table the footnote carrying that decision is doing a great deal of work.

In practice

Crossing the four age bands of the Sylhet survey against hypertension would give eight cells. Printing the row percentages gives the prevalence within each age band, which is

what a reader wants, and the four figures each stand on their own denominator. Printing the column percentages instead would give the age composition of the hypertensive group, which is a different and much less useful statement, and the four figures would sum to a hundred between them.

Marginal distribution
The frequency distribution of one variable of a two-way table, read off its margins, ignoring the other variable.

The margins of a two-way table are worth naming separately. The row totals reproduce the frequency distribution of the row variable on its own, and the column totals do the same for the column variable. These are the *marginal distributions*, and they are exactly the one-way tables of Section 1. A two-way table therefore contains both one-way tables inside it, which is why the margins are printed rather than left to the reader.

Watch out

A large contingency table invites the reader to hunt for the biggest number. Resist it. In a 4×3 table of counts the largest cell is usually just the combination of the two commonest categories, and says nothing about whether the two variables are related at all.

7.3 Percentages inside subgroups

Most descriptive tables in clinical research are stratified, and every percentage in a stratified table is a share of the number in its own column heading.

Example 7.2 • Three denominators in one row

The Sylhet survey reports current smoking as 546 (63.2%) of 864 men, 36 (3.8%) of 946 women, and 582 (32.2%) of all 1,810 participants. The total of 582 is ours; the rest is the paper's.

The combined figure of 32.2% does not sit midway between 63.2 and 3.8, which would be about 33.5. It is pulled towards the women, because there are more of them. A combined percentage is a weighted average of the subgroup percentages, and the weights are the subgroup sizes. Chapter 4 does that arithmetic.

7.4 A table you already read this way

The three percentages of Section 7.1 are not an academic construction. Every clinician computes two of them daily under other names.

Example 7.3 • Sensitivity is a column percentage

Consider a table of 1,000 people, constructed here with round numbers: of the 100 with the disease, the test is positive in 80; of the 900 without it, the test is positive in 90. The test is therefore positive in 170 people altogether.

Sensitivity is $80/100 = 80\%$: of those with the disease, the share the test caught. That is a column percentage, dividing the cell by its column total.

Positive predictive value is $80/170 = 47\%$: of those the test flagged, the share who really have the disease. That is a row percentage, dividing the same cell by its row total.

Example 7.3 • *continued*

One cell, 80, and two familiar quantities that differ by more than thirty percentage points because they use different denominators. Every argument about whether a screening test is worth running is an argument about which of the two is being quoted.

Watch out

This section is about which way a percentage runs, not about diagnostic accuracy, which is the subject of Chapter 27. The reason the predictive value is so much lower than the sensitivity here is that only one in ten of those tested has the disease, so most positive results come from the nine hundred who do not.

7.5 Which percentage to print

The point

Percentage across the exposure, so that each group being compared sums to 100%. Then the two figures are shares of their own groups and are comparable with each other.

In the example, men 18.8% against women 18.7% compares like with like: each is a share of that sex. Comparing 47.8% against 52.2% compares the two column percentages, which sum to 100 between them, and so reports only that there were more women than men in the sample. It looks like a finding about hypertension and is a finding about the sex ratio of Sylhet.

Intuition

Look at which direction adds to a hundred. If the two numbers being compared come from a set that sums to 100, they are shares of a single group, and comparing them says nothing about the two groups the reader has in mind.

On this evidence, men and women in this survey had the same prevalence of hypertension: 18.8% and 18.7%. Whether a difference of that size would be worth attention at all, and how to ask, is Chapter 21.

7.6 The footnote that decides the sentence

The survey reports that 63.2% of men were current smokers, against 3.8% of women, and a footnote records that these are column percentages. Read correctly, 63.2% is 546 divided by the 864 men.

Read as a row percentage, the same figure would be taken to mean that 63.2% of smokers were men. That quantity exists and it is not 63.2%: there were 582 smokers in total, of whom 546 were men, which is 93.8%. Both the 582 and the 93.8% are ours.

The two readings differ by thirty percentage points, and the only thing distinguishing them is a footnote set smaller than the table it explains.

Current smoker	Men ($n = 864$)	Women ($n = 946$)	All
No	318 (36.8%)	910 (96.2%)	1,228
Yes	546 (63.2%)	36 (3.8%)	582

Footnote in the paper: *column percentage.*

What 63.2% means

Of the 864 men, this share smoked.
546 ÷ 864.

What it does not mean

"63.2% of smokers are men."
That figure is 546 ÷ 582 = 93.8%.

Same four counts, two sentences, and one of them is out by thirty percentage points. The only thing that tells you which is right is a footnote in six point type.

Figure 6: A real row from a published table, with the footnote that determines how it must be read. Counts and column percentages from reference 1.; the total of 582 smokers, the 93.8% and the incorrect sentence are ours.

The point

When building a table, state the direction of the percentages in the column heading rather than in a footnote. *n (% of men)* cannot be misread.

7.7 Building one row, end to end

Six steps take a column of raw answers to a printed row. Only two of them are arithmetic.

Step	Decision, and where it is recorded
1. Name the variable	Exactly as it will appear. *Current smoker*, not *smoking*.
2. List the categories	In their own order if ordinal. Include every one.
3. Count	And check the total against n.
4. Fix the denominator	Everybody, or only those who answered. Say which.
5. Compute	One decimal place.
6. Write the heading	*n (%)*, with N and the direction stated.

The point

Four of the six steps involve no calculation. Most of what makes a descriptive table right or wrong is decisions, and those are the ones a reader has to reconstruct from the finished table. Step one is the operational definition of Section 5 arriving in a column heading.

7.8 Reading a descriptive table

The order below is deliberately the reverse of what most readers do.

1. **The title.** Who is in this table, and who has been excluded.
2. **The column headings.** What is n for each column? There is often more than one.
3. **The footnotes.** Which way do the percentages run, and what do the symbols mark?
4. **Only then, a number.**

Most readers begin with the numbers and consult the headings only when a number surprises them. This is backwards, because the number that does not surprise is the one that will be quoted, and quoted wrongly.

8 What a percentage does not carry

8.1 Auditing a table before believing it

Four checks establish whether a published table is internally consistent. None needs the raw data, and together they take about two minutes.

1. Every count column sums to its stated n.
2. Every percentage reproduces from its own count and denominator.
3. A count and its percentage imply one denominator. It should be findable, and it should be the one the heading claims.
4. Anything ordered runs in the direction it should, or the paper explains why not.

The Sylhet table passes all four. Its servings column reads $49.1 + 38.6 + 12.4 = 100.1$, which is not a failure: three percentages each rounded up by less than half a point must do this, and Section 8.2 sets out why.

When check four fires

The national analysis of Section 4.2 reports 2011 hypertension prevalence of 41.09% among the overweight and 28.0% among the obese. The order runs the wrong way and the paper does not comment. The printed margins do not permit an explanation, so this is not a finding and not an error: it is a place where a careful reader goes to the methods. Recording that the question was asked and could not be answered is the honest outcome.

The point

The justification for this is the one that applies to checking a drug chart before signing it. It is not that an error is expected. It is that signing means somebody looked.

8.2 Rounding

A column of percentages rounded to one decimal place need not sum to exactly 100. In the servings-of-fruit-and-vegetables row of the Sylhet survey, the unrounded percentages are 49.06, 38.56 and 12.38, which sum to 100.00 exactly. Rounded to one place, each moves up, and the printed column reads 49.1, 38.6 and 12.4, summing to 100.1.

No error has occurred. Many journals ask for a footnote explaining the discrepancy and it is polite to supply one, but the arithmetic is sound.

Servings per day	Count	Exact percentage	As printed
Fewer than 2	888	49.06…	49.1
2 to 4	698	38.56…	38.6
5 or more	224	12.38…	12.4
Total	**1,810**	100.00	100.1

Every one of the three was rounded *up*, so the printed column is a tenth heavier than the truth. This is not an error and it needs no footnote. What it does mean is that **the counts are the data and the percentages are a rendering of it**, so when the two disagree, go back to the counts.

The exact percentages are ours. The paper prints the counts and the rounded column.

Figure 7: Three percentages, each rounded up by less than a tenth, producing a column that sums to 100.1. Counts and printed percentages from reference 1.; the unrounded percentages are ours.

The point

The counts are the data and the percentages are a rendering of them. Where the two disagree, return to the counts. This is the reason the convention is to print both.

8.3 Precision that is not there

One decimal place is the conventional precision for a percentage in a clinical paper, and the convention is about honesty rather than style. A sample of 1,810 people changes by 0.055 percentage points when one person moves from one category to another. Reporting 18.73% implies a resolution finer than one person, which the data cannot support.

For small samples even one decimal place overstates the case. In a group of 40, one person is 2.5 percentage points, and a figure such as 27.5% suggests a precision that a single reclassification would destroy.

8.4 What travels with a percentage

The point

Report the count, the denominator and the percentage together. Any two of the three determine the missing one, and printing only the percentage is the single choice that loses information permanently.

In practice

The Sylhet survey reports that 87.9% of participants used smokeless tobacco, which is 1,591 of 1,810. The percentage alone conveys the scale of the finding. It does not convey how many people were asked, that the answer was self-reported, or that a stigmatised behaviour may be reported differently by men and by women. The first of those is restored by printing the denominator; the others are matters of measurement, and belong to Chapter 14.

8.5 What a percentage cannot distinguish

	With the condition	Total	Percentage
A village survey	9	48	18.8%
The Sylhet survey	162	864	18.8%
A national programme	9,411	50,059	18.8%

The middle row is real: 162 of the 864 men in the Sylhet survey. The first and third rows are constructed here to make the point and are not from any study.

The three do not deserve equal confidence, and the percentage cannot tell them apart, because the percentage has discarded the one quantity that distinguishes them. Restoring the counts restores the distinction without any further machinery.

How much more confidence the third row deserves, and how to express that in print, is the subject of Chapters 7 and 8. This chapter goes as far as establishing that the question exists and that a bare percentage suppresses it.

9 Chapter summary

1. A categorical column is described completely by its frequency distribution: the count in each category, with a total that matches n. Checking that total is the first thing to do to any published table.
2. A proportion is a count divided by the total it belongs to; a percentage is that proportion in hundredths. Neither adds information to the counts.
3. A nominal column admits the mode and nothing else. An ordinal column admits the mode and the median, but never the mean.
4. Counts can be accumulated down an ordinal table, answering how many observations lie at or below each level. Nominal counts cannot.
5. Two prevalences are comparable only if the population, the case definition, the period and the measurement all match.
6. A ratio compares two counts and has no upper bound. A rate divides by person-time. A cross-sectional survey can produce a proportion but never a rate.
7. Prevalence is a proportion reported with a population, a place, a period and a case definition. Dropping any of the four changes the claim.
8. The order of categories is data for an ordinal variable and a choice for a nominal one. Missing is neither a category nor zero, and the denominator chosen must be stated.
9. Collapsing categories buys simplicity and pays with every distinction inside the merged group. Collapse when the groups would be managed the same way.
10. The denominator of a percentage is a choice, and a single published table may use more than one. Recovering it is the reader's work.
11. Every cell of a two-way table gives three correct percentages. Print the one that makes the groups being compared each sum to 100.
12. Percentages are rendered from counts and may sum to 99.9 or 100.1 after rounding.

The counts remain authoritative.

13. Count, denominator and percentage travel together. A percentage alone conceals both the size of the study and its own fragility.

10 Key terms

Term	Meaning
Frequency distribution	The counts of observations falling in each category of a categorical variable, summing to n.
Frequency	A count. Sometimes called absolute frequency.
Relative frequency	A count divided by the total; a proportion.
Mode	The category with the largest count. Defined for every categorical variable.
Modal category	The same thing, named as a category rather than as a summary.
Proportion, $\hat{p}$	A count divided by the total it belongs to. Between 0 and 1.
Percentage	A proportion multiplied by a hundred.
Ratio	One count divided by another, where neither need be part of the other. No upper bound.
Rate	A count divided by person-time. Not obtainable from a cross-sectional survey.
Prevalence	The proportion of a defined population with a condition at a defined time.
Denominator	The set of observations a numerator is expressed as a share of. Not necessarily n.
Two-way table	A cross-classification of n observations by two categorical variables. Also contingency table.
Marginal total	A row or column total of a two-way table.
Marginal distribution	The frequency distribution of one variable of a two-way table, read off its margins.
Row percentage	A cell divided by its row total.
Column percentage	A cell divided by its column total.
Total percentage	A cell divided by the grand total.
Cumulative frequency	The number, or share, of observations at or below a given level of an ordinal variable.
Collapsing	Merging two or more categories into one.

Symbols introduced in this chapter: a for a count, k for the number of categories, and r and c for the numbers of rows and columns of a two-way table. n, $\hat{p}$ and π carry the meanings given in Chapter 2.

11 Exercises

The exercises below are built on the same published study as Chapter 2: 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.

THE NUMBERS YOU WILL NEED

Sample. 1,810 adults: 864 men and 946 women.

Education, whole sample. No education 333; one to five years 1,126; six years or more 351.

Current smoking, column percentages. Men: no 318 (36.8%), yes 546 (63.2%). Women: no 910 (96.2%), yes 36 (3.8%).

Smokeless tobacco, whole sample. Users 1,591 (87.9%).

Servings of fruit and vegetables per day. Fewer than two 888 (49.1%); two to four 698 (38.6%); five or more 224 (12.4%).

Body mass index. Underweight 531 (29.3%); normal 1,026 (56.7%); overweight or obese 253 (14.0%).

Blood pressure. Hypertension 18.7% overall (95% CI 17.0 to 20.6); 18.8% of men and 18.7% of women; 339 people in total, 162 men and 177 women. Among those 339: controlled 31.3%, uncontrolled 24.2%, newly identified 44.5%. Among the 162 hypertensive men, controlled 23.5%.

Part A. Frequency distributions

1. Verify that the three education counts form a complete frequency distribution, and state the check you performed.

2. Compute the percentage in each education band to one decimal place. Do they sum to 100.0?

3. A different survey of 500 people reports three categories of occupation with counts 210, 180 and 95. What can be said about this table?

Part B. The four words

4. Give the modal category of the education distribution and of the body mass index distribution. For which of the two is a median also available, and why?

5. Express the number of hypertensive men against the number of hypertensive women as a ratio, and then as a proportion of all hypertensive participants. Explain why the two answers are different kinds of quantity.

6. A hospital reports a "surgical site infection rate of 4.2%". Is this a rate? What would have to be reported instead for it to be one?

Part C. Prevalence

7. Write the survey's prevalence of hypertension as a single sentence that a reader could not misinterpret. Include everything that must accompany the number.

8. The case definition used a systolic threshold of 140 mm Hg. Without any calculation, state which way the reported prevalence would move if the threshold were 130, and why.

9. A colleague cites this study as evidence that "hypertension affects 18.7% of Bangladeshis". Identify three separate things wrong with that sentence.

Part D. Denominators

10. The survey reports that 44.5% of hypertensive participants were newly identified. How many people is that, and what is the corresponding percentage of the whole sample? Show both divisions.

11. You are writing two documents: a note for clinicians in a district hospital, and a budget submission for a screening programme. Which of the two figures from the previous exercise belongs in each, and why?

12. Among the 162 hypertensive men, 23.5% had controlled blood pressure. Convert this to a count. Then state what the percentage would become if four of those men had been classified the other way.

13. Take any published abstract in your own field. For its first three percentages, write down the numerator and denominator of each. Report how many of the three you could recover.

Part E. Two-way tables

14. Build the two-way table of sex against hypertension from the numbers in the box above. Give all four cell counts, both sets of marginal totals and the grand total.

15. For the cell containing hypertensive men, compute the row percentage, the column percentage and the total percentage. Write each as a sentence.

16. Which of your three sentences answers the question ”are men at higher risk of hypertension than women in this population?” What is the answer?

17. The survey reports that 63.2% of men were current smokers. A reader writes: ”63.2% of smokers in this survey were men.” Compute the figure the reader actually described, and state how far wrong the sentence is.

Part F. Reporting and appraisal

18. The three servings-of-fruit-and-vegetables percentages sum to 100.1. Recompute them from the counts to two decimal places and explain the discrepancy.

19. Collapse the education distribution into ”no education” against ”any education”. Give the counts and percentages, then state in one sentence what the collapsed table conceals.

20. Would collapsing the body mass index categories into ”normal” against ”not normal” be defensible in this population? Justify your answer using the reported figures.

21. A table reports ”32% of participants improved” with no other figures. List three things a reader cannot determine, and write the one sentence that would fix all three.

12 Solutions

Part A

1. $333 + 1126 + 351 = 1810$, which matches the number of participants. The check is that the counts in a frequency distribution must sum to n; if they do not, a category is missing or observations have been dropped.

2. $333/1810 = 18.4\%$, $1126/1810 = 62.2\%$ and $351/1810 = 19.4\%$. These sum to 100.0. Not every such column will, because rounding can carry the total to 99.9 or 100.1.

3. The counts sum to 485, not 500, so fifteen observations are unaccounted for. Either a category has been omitted from the table or fifteen participants have missing occupation data. The table cannot be interpreted until which of these applies is known, because the denominator of the reported percentages depends on it.

Part B

4. The modal education category is one to five years of schooling, with 1,126 participants. The modal body mass index category is normal, with 1,026.

A median is available for both, because both are ordinal: education runs from none upwards and body mass index from underweight upwards, so a middle observation is well defined in each. What is not available for either is a mean, because the gaps between adjacent bands are not equal quantities. Had body mass index been reported as the measured value in kg/m² rather than in bands, a mean would have been available, which is the subject of Chapter 4.

5. The ratio of hypertensive men to hypertensive women is $162 : 177$, or about 0.92 to 1. As a proportion of all hypertensive participants, men are $162/339 = 47.8\%$.

The ratio compares two counts to each other and neither is part of the other, so it has no upper bound and does not have to lie below 1. The proportion divides one count by a whole that contains it, so it necessarily lies between 0 and 1.

6. It is a proportion, not a rate. It presumably means that 4.2% of operations were followed by an infection, which is a count of infections divided by a count of operations.

For it to be a rate the denominator would have to be person-time: infections per 1,000 patient-days of post-operative follow-up, say, together with the length of the follow-up window. Without a stated window, the figure rises simply by watching patients for longer.

Part C

7. An acceptable sentence: *Among adults aged* 35 *years and over living in the Zakiganj and Kanaighat subdistricts of rural Sylhet district, surveyed between August* 2017 *and January* 2018*,* 18.7% *(*339 *of* 1,810*;* 95% *CI* 17.0 *to* 20.6*) had hypertension, defined as systolic pressure of at least* 140 *mm Hg, or diastolic of at least* 90*, or current antihypertensive medication.*

The four elements that must be present are the population including the age floor, the place, the period and the case definition. The count and denominator are added because the chapter requires them.

8. The prevalence would rise. Lowering the threshold moves people who were previously below the cut into the hypertensive group, and nobody moves the other way. No participant's blood pressure changes; only the rule that classifies it does.

9. Three of several: the study covered only adults aged 35 and over, so it says nothing about younger Bangladeshis and the figure for the whole population would be lower; it was conducted in two rural subdistricts of one district, so it cannot speak for Bangladesh and in particular not for urban Bangladesh; and it applies to a specific case definition and a specific period, either of which would change the number.

Part D

10. The denominator is the 339 participants found to have hypertension, so $0.445 \times 339 \approx 151$ people.

As a percentage of the whole sample, $151/1810 = 8.3\%$.

Both figures are computed here; the paper prints the 44.5% and the 339 but neither the count nor the 8.3%.

11. The clinical note takes 44.5%. Its denominator is the people who have the condition, and the sentence it supports is that hypertension is frequently silent, which argues for measuring blood pressure opportunistically rather than waiting for a complaint.

The budget submission takes 8.3%. Its denominator is the population the programme would serve, so it translates directly into how many people a screening programme would find, and therefore into staff, equipment and drugs.

12. $0.235 \times 162 \approx 38$ men. If four had been classified the other way the count would be 42, and $42/162 = 25.9\%$, or 26.0% with the arithmetic carried to one more place before rounding.

Either way, moving four people in a survey of 1,810 shifts the printed figure by roughly two and a half percentage points, while the neighbouring prevalence of 18.7%, resting on the full sample, would barely move at all. Nothing in the notation distinguishes them.

13. No single answer. The expected finding is that at least one of the three denominators cannot be recovered from the abstract alone, because word limits remove the counts. Where that happens, the correct response is to note it as a limitation of the reporting rather than to assume the sample size applies.

Part E

14.

	Hypertension	No hypertension	Total
Men	162	702	864
Women	177	769	946
Total	339	1,471	1,810

The two right-hand cells are obtained by subtraction: $864 - 162 = 702$ and $946 - 177 = 769$. The column total 1,471 is $1{,}810 - 339$.

15. Row percentage $162/864 = 18.8\%$: 18.8% *of men had hypertension.*

Column percentage $162/339 = 47.8\%$: 47.8% *of the people with hypertension were men.*

Total percentage $162/1810 = 9.0\%$: 9.0% *of the whole survey were men with hypertension.*

16. The row percentage, because it expresses the cell as a share of the group being compared. Set against the corresponding figure for women, $177/946 = 18.7\%$, the answer is that men were not at higher risk in this survey: the two prevalences are effectively the same.

The column percentages, 47.8% and 52.2%, sum to 100 between them and therefore describe the composition of the hypertensive group rather than the risk in either sex. They would differ even if the two prevalences were identical, simply because there were more women in the sample.

17. The total number of smokers is $546 + 36 = 582$, of whom 546 were men, giving $546/582 = 93.8\%$.

The reader's sentence therefore understates the figure by about thirty percentage points. The published 63.2% is a column percentage, 546 of the 864 men; the reader read it as a row percentage. The footnote is the only thing in the table that distinguishes the two.

Part F

18. $888/1810 = 49.06\%$, $698/1810 = 38.56\%$ and $224/1810 = 12.38\%$, summing to 100.00. Rounded to one decimal place each moves up, giving 49.1, 38.6 and 12.4, which sum to 100.1.

No error has occurred. The counts sum correctly to 1,810 and are the authoritative version; the percentages are a rendering of them at a coarser precision.

19. No education 333 (18.4%); any education $1{,}126 + 351 = 1{,}477$ (81.6%).

The collapsed table conceals that the largest group in the population, 62.2% of the sample, had only one to five years of schooling, and counts those participants as educated alongside people with secondary schooling and above.

20. It would not be defensible. In this population 29.3% were underweight and 14.0% overweight or obese, so the "not normal" group would be roughly two parts undernutrition to one part overnutrition. These are opposite clinical problems requiring opposite interventions, and merging them produces a number that supports no decision. The test in Section 5.4 fails: the merged groups would not be managed in the same way.

21. A reader cannot determine how many participants there were, how many of them improved, or what "improved" means as a definition. The first two are the numerator and denominator; the third is the case definition, in the sense of Section 3.

A sentence that fixes all three: 32% *of participants (*48 *of* 150*) improved, defined as a reduction of at least two points on the score at twelve weeks.*

13 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. Chowdhury MAB, Islam M, Rahman J, Uddin MT, Haque MR, Uddin MJ. Changes in prevalence and risk factors of hypertension among adults in Bangladesh: an analysis of two waves of nationally representative surveys. *PLoS One* 2021;16(12):e0259507. PMID 34855768. PMC8638884. doi:10.1371/journal.pone.0259507. Published under CC BY 4.0, so both its figures and its text may be used with attribution. The source of every age-specific prevalence in Section 4.
3. Altman DG, Bland JM. Presentation of numerical data. *BMJ* 1996;312(7030):572. PMID 8595293. PMC2350327. doi:10.1136/bmj.312.7030.572. One page, free on PMC, and worth reading before building a first table.
4. Daniel WW. *Biostatistics: A Foundation for Analysis in the Health Sciences*, 9th ed. Wiley. Section 2.3, on the frequency distribution and relative frequencies.

Figures computed in this chapter rather than reported by the paper. The count 151 and the share 8.3%; the count 805; the cells 702, 769 and 1,471 and every marginal total of the two-way table; the percentages 47.8 and 9.0; the total of 582 smokers and the share 93.8%; the count 38 and the figure 26.0%; the collapsed education figures 1,477 and 81.6%; the unrounded percentages 49.06, 38.56 and 12.38; and both invented rows of the three-studies table in Section 8.5.

From Section 4, computed here rather than reported by reference 2.: the standardised prevalences 25.91% and 39.29% and both ratios of 1.52; the combined 65+ rates 40.17% and 54.57%; the expected prevalence 39.5% for rural Sylhet and the observed-to-expected ratio 0.47. The two hospitals of Example 4.1 are constructed entirely and are not a study. The cumulative schooling figure 80.6 of Section 5.3 is ours.

14 For students in the mentorship programme

Everything above stands on its own. This section is the only part of the chapter that assumes a session.

Before the next session

One. Build the frequency table for every categorical column in your own study. Count, denominator and percentage, with *N* in the column heading.

Two. Audit somebody else's. Take Table 1 of any paper in your field and run the four checks of Section 8.1 on it. Expect to find at least one percentage whose two numbers you cannot recover; bring that one.

Three. Find a paper in your field reporting an age-standardised rate. Say which standard population it used, and whether it said so.

Reading

Altman and Bland (reference 3.) is one page and free. Daniel section 2.3 (reference 4.) covers the frequency distribution formally.

Questions this chapter deliberately left open

Question	Where it is settled
Is the difference between 18.8% and 18.7% real, or could it be chance?	Chapter 21, chi-square methods
How far might 18.7% be from the true population figure?	Chapter 7, standard error and precision
How is that uncertainty written down?	Chapter 8, confidence intervals
Can 87.9% self-reported be believed?	Chapter 14, measurement, reliability and validity
How should any of this be drawn rather than tabulated?	Chapter 6, tables and graphs
When is a proportion better expressed as an odds or a risk?	Chapter 23, epidemiological effect measures
Having found that a comparison reverses on stratification, how is the third variable actually handled?	Chapter 15, bias and confounding, and Chapters 25 and 26 for more than one at a time
Which of the two conflicting Sylhet prevalences should be believed?	Left open deliberately. Chapter 16 covers sampling and generalisability; the evidence in this chapter does not settle it

Next

Chapter 4, Describing Numerical Data: the columns that could not be counted, and why the study behind this chapter reported a median for every one of them.