



CLASS NOTES • CHAPTER 2

Describing Data

Tables, averages, spread and charts, before any test

Statistics & Data Analysis Fundamentals
Stat.BD

Mentor: Muhtasim Munif Fahim, MSc

CHAPTER 2

Describing Data

Before any test, look. This chapter turns a column of answers into a table, a centre, a spread and a chart, and shows what each one hides.

Row 37 of the class dataset records a student who studies 25.0 hours on a typical class day. A day has 24 hours. The form said 2.5, and the typist missed the decimal point. The error is obvious once someone looks at the row, and invisible to anyone who only looks at the average. How far does one wrong key move the class average of 120 students? How far does it move the middle value, and how far the spread? Most people guess that the average moves a great deal, because 25 is an enormous number. The answer, given in section 3.9, is that the average barely moves and the spread nearly doubles. The difference between those two behaviours is most of what this chapter teaches.

Describing data is the first of the three jobs of statistics set out in Chapter 1, and the one most often skipped. A student with a new dataset wants a p-value; a careful analyst wants to see the data first. Every summary answers a narrower question than it seems to, and every chart can be drawn to flatter. The chapter proceeds in the order an analyst works. It starts with tables for categories, where the whole difficulty lies in choosing the denominator. It then finds the centre of a numerical variable, measures the spread around that centre, and deals with values that stand out. It turns next to charts: which one suits which variables, and how charts are made to mislead. It ends by reading a table in a published paper with the same eyes, and finds a count that cannot be right.

The data are the class survey described in Chapter 1. Until the real answers are coded, every figure drawn from it comes from a **simulated** stand-in of 120 students generated by a program from a fixed seed. They are not real students. A few examples are constructed on purpose, with made-up numbers, and are labelled as constructed where they appear.

What this chapter establishes

1. How to build a frequency table and a cross-tabulation, and how to choose between row, column and total percentages from the question being asked.
2. When a cumulative percentage means something, and when it is noise.
3. How to compute a mean, a median and a mode, and which one to report for each scale of measurement and each shape of distribution.
4. Why the mean of a 0/1 column is a proportion, and how a weighted mean combines groups of different sizes.
5. How to compute a standard deviation by hand, and why it divides by $n - 1$.
6. How to find quartiles and the interquartile range, draw a box plot, and flag unusual values with fences.

7. What to do with a value that stands out, and why a flag is not a verdict.
8. How a change of units moves the summaries, and when a coefficient of variation or a skewness figure is meaningful.
9. How to choose a chart from the types of the variables, and how to recognise a chart built to mislead.
10. How to check a published table against its own totals.

Notation

n	the number of values, usually the number of people who answered
x, x_i	a value, and the value of the i -th person
$\sum x$	the sum of all the values; $\sum$ is read “sum of”
$\bar{x}$	the sample mean, read “x-bar”
$x - \bar{x}$	a deviation: how far one value sits from the mean
s^2, s	the sample variance and the sample standard deviation (SD)
Q_1, Q_3	the lower and upper quartiles, the 25th and 75th percentiles
IQR	the interquartile range, $Q_3 - Q_1$
w	a weight: a group’s size, or a course’s credits
CV	the coefficient of variation, the SD as a percentage of the mean
$\hat{p}$	a sample proportion

Contents

What this chapter establishes	1
Notation	2
1 Tables for categories	5
1.1 The frequency table	5
1.2 Valid and cumulative percentages	6
1.3 Grouping numbers into classes	7
1.4 The cross-tabulation	7
1.5 Row, column and total percentages	8
2 The centre	9
2.1 The mean	9
2.2 The median	9
2.3 The mode	10
2.4 The mean as a balance point	10
2.5 Shape: symmetric and skewed	10
2.6 Which centre to report	11
2.7 The mean of a 0/1 column is a proportion	12
2.8 The weighted mean	13
3 The spread	14
3.1 Same average, different classes	14
3.2 The range	14
3.3 The standard deviation	14
3.4 Why divide by $n - 1$	15
3.5 The two-SD rule of thumb	16
3.6 Percentiles and quartiles	16
3.7 The box plot	17
3.8 Fences: flagging values that stand out	17
3.9 The answer to the 25-hour question	18
3.10 The undeclared 99	19
3.11 What to do with a value that stands out	20

3.12	Units, and the coefficient of variation	20
3.13	Skewness as a number	21
3.14	Reporting a descriptive table	21
4	Charts	22
4.1	Choosing a chart from the variables	22
4.2	The histogram	22
4.3	Comparing groups	23
4.4	The pie chart	24
4.5	The scatter plot	25
4.6	Anscombe's quartet	25
5	Charts that mislead	26
5.1	A bar axis that does not start at zero	26
5.2	Icons that grow in two directions	26
5.3	Choosing the window	27
5.4	A percentage without its count	27
5.5	Six questions for any chart	27
6	Describing one variable: the procedure	28
7	Reading a table in a published paper	28
7.1	Islam 2020, Table 2	28
7.2	Three checks catch one misprint	29
8	Chapter summary	30
9	Key terms	31
10	Exercises	33
11	Solutions	37
12	Our arithmetic and constructed numbers	41
13	References and further reading	42
14	For students in the Stat.BD course	42

1 Tables for categories

1.1 The frequency table

The simplest summary in statistics is a count. A frequency table lists every category of a variable with the number of people who gave that answer, and usually the percentage. It answers the question “how many, and what share?” for a categorical variable, and it is the first thing an analyst produces for every such column in a new dataset. Table 1 is the frequency table for faculty in the simulated class. Each percentage is the count divided by the total and multiplied by a hundred. For Arts & Humanities that is $29 \div 120 \times 100 = 24.2$. The table has one row per category, and every student appears in exactly one row, which is Chapter 1’s rule of mutually exclusive and exhaustive options at work.

Frequency table

a table listing each category of a variable with its count and percentage

Table 1: Faculty in the **simulated** class data. The faculty grouping is the codebook’s.

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

$$\text{percentage} = \frac{\text{count}}{\text{total}} \times 100 \quad (1.1)$$

count the number of people in the category

total the number of people the percentage is *of*: the denominator

A total that is not 100

The percentages in Table 1 add to 100.1. Each was rounded to one decimal, and rounded parts need not add to exactly 100. The honest response is to leave the total as it is and, if a reader might be puzzled, to say in a footnote that percentages are rounded. Changing one entry so that the column adds to 100 is falsifying a table, however small the change.

A bar chart draws the same table (Figure 1). The eye compares lengths far better than it compares numbers in a column, so the chart shows at a glance what the table states. Three rules govern it. The bars have gaps between them, because the categories are separate. The axis starts at zero, because the *length* of a bar is its message; section 5 shows what happens when that rule is broken. And for nominal data, which have no natural order, the bars can be sorted by size to help the eye. For an ordinal variable such as year of study the bars follow the order of the categories instead.

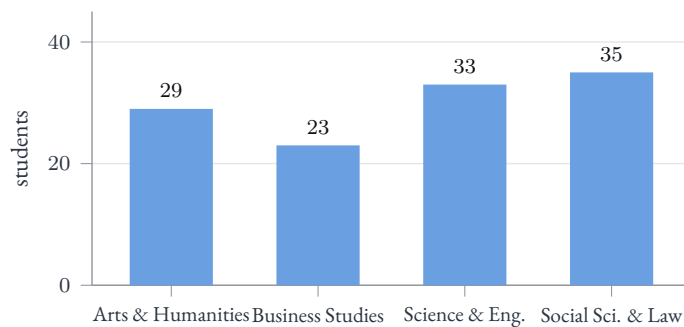


Figure 1: The faculty table drawn as a bar chart. **Simulated** class data, $n = 120$.

1.2 Valid and cumulative percentages

Not every student has a CGPA. The 15 first-semester students in the simulated class have no result yet, coded 9 in the codebook. A frequency table for CGPA band therefore has a choice to make about its denominator, and the table that SPSS prints, which Session 7 produces, shows both choices side by side (Table 2). The column headed “% of all” divides by all 120 students. The column headed “valid %” divides only by the 105 students who have a result. The question decides which is right. “What share of students with a result have a CGPA of 3.00 or above?” uses valid percentages, because a student with no result cannot be counted in either direction.

Valid percentage

a percentage computed only among those with an answer, leaving out missing and not-applicable cases

Table 2: CGPA band with three percentage columns. **Simulated** class data.

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

The last column is a cumulative percentage: the valid column added down the page. It reads as “this share of students is in this band or below”. Here 45.7% of students with a result are below 3.00. A running total needs the categories to have an order, because “or below” must mean something. CGPA bands run from low to high, so the cumulative column is informative. Faculty has no order, and a cumulative column for faculty, which software will print without complaint, answers no question anyone asked: that Arts and Business together make up 43.4% of the class is correct arithmetic about an arbitrary pairing. This is Chapter 1’s scale of measurement paying its first dividend. The scale decides which columns of a table mean anything.

Cumulative percentage

a running total of the valid percentages, from the lowest category up

Intuition

A percentage is always a percentage *of* something. Before reading one, find the denominator. Before computing a running total, check that the categories have an order.

1.3 Grouping numbers into classes

A numerical variable becomes a frequency table once its range is cut into class intervals. Table 3 does this for self-study hours in the corrected file. Three rules make such a table fair. The classes have equal widths, so that the counts compare like with like. The boundaries are written “2 to under 3”, so that a student who reports exactly 3 hours has one home and not two; that is the mutually exclusive rule of Chapter 1 again. And there are enough classes to show a shape, usually between five and ten. The cumulative column answers questions of the form “how many study less than...”: 61.7% of the class study under three hours a day, and 94.2% under five. This table is the skeleton of a histogram, which section 4.2 draws.

Class interval

one of the equal-width ranges into which a numerical variable is cut for a frequency table or histogram

Table 3: Self-study hours on a typical class day, grouped into one-hour classes. **Simulated** class data, corrected file, $n = 120$.

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

1.4 The cross-tabulation

A cross-tabulation counts two categorical variables at once. Table 4 crosses residence with whether a student does private tuition or a paid job. Every student sits in exactly one of the six inner cells. The row totals, down the right-hand side, are the frequency table for residence; the column totals, along the bottom, are the frequency table for tuition. These totals are called the margins. The inner cells are new information: they show how the two variables go together. In this simulated class 17 hall students do tuition and 10 family-home students do. That comparison is not yet fair, because the two groups are of different sizes. Making it fair is a matter of choosing the right percentage.

Cross-tabulation

a table counting two categorical variables at once, one cell for each combination of categories

Margins

the row and column totals of a cross-tabulation, which reproduce the two single-variable tables

Table 4: Residence by tuition or a paid job. **Simulated** class data, $n = 120$.

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



1.5 Row, column and total percentages

The same cell can be turned into three different percentages, and each answers a different question. Take the 17 hall students who do tuition. Divided by the 47 hall students, they are a row percentage, 36.2%: of hall students, how many work? Divided by the 59 students who do tuition, they are a column percentage, 28.8%: of those who work, how many live in a hall? Divided by all 120, they are a total percentage, 14.2%: of everyone, how many are hall students who work? Table 5 sets out all three for the “yes” column. The same students become three different numbers, and the most common error in reading a cross-tabulation is to take one kind for another.

Row percentage

a cell divided by its row total:
the share within the group
the row defines

Table 5: The “yes” cells of Table 4 as three kinds of percentage. **Simulated** class data.

Does tuition	Row % of that residence	Column % of all who do tuition	Total % of all 120
Hall	36.2	28.8	14.2
Mess	74.4	54.2	26.7
Family home	33.3	16.9	8.3

A quick check tells the three apart. Column percentages add to 100 down the column, since every student who does tuition lives somewhere. Row percentages do not add to 100 down the column: each is a share of a different group. Students who expect every column of a table to add to 100 are misled by exactly this.

Definition 1.1 • Row, column and total percentages

In a cross-tabulation, a **row percentage** divides a cell by its row total, a **column percentage** divides it by its column total, and a **total percentage** divides it by the grand total. To compare groups, divide within each group: the variable that defines the groups supplies the denominator.

Example 1.1 • Three questions, three percentages

Are mess students more likely to do tuition than hall students? This compares residence groups, so divide within each residence: row percentages, 74.4% against 36.2%. In this simulated class about three in four mess students do tuition or a job, against just over one in three hall students.

Where do the students who do tuition live? This describes one group, the tuition group, so divide within it: column percentages. 54.2% of them live in a mess, so the statement “most students who do tuition live in a mess” is true here, narrowly.

What share of the class is a mess student who does tuition? This is about everyone, so divide by everyone: a total percentage, 26.7%.

Whether a gap as large as the first could arise by chance is a question for the chi-square test of Chapter 5. This chapter only describes it.

A raw count is not a comparison

The counts suggest that hall students work far more often than family-home students: 17 against 10. But there are 47 hall students and only 30 family-home students. The row percentages, 36.2% and 33.3%, are nearly the same. A count reflects the size of the group as much as the habit being counted, so compute the row percentages before writing any sentence that compares groups.

2 The centre

2.1 The mean

The mean is the familiar average: add the values and divide by how many there are. Its most useful reading is as a fair share. If the total were shared out equally, each person would get the mean. The first seven self-study values in the class data are 2.5, 2.0, 3.0, 3.5, 3.0, 2.5, 2.0 hours. They add to 18.5, and shared among seven students that is 2.64 hours each. The fair-share reading also explains the mean's great weakness. One very large value is shared out along with the rest, so it pulls the mean towards itself. Nothing in the arithmetic protects the mean from a single wrong entry.

Mean

the sum of the values divided by how many there are; the arithmetic average

$$\bar{x} = \frac{\sum x}{n} \quad (2.1)$$

$\bar{x}$ the sample mean

$\sum x$ the sum of all n values

n the number of values

2.2 The median

The median is the middle value once the values are sorted. Sorted, the seven values are 2.0, 2.0, 2.5, 2.5, 3.0, 3.0, 3.5. The fourth of seven is the middle, so the median is 2.50. With an even number of values there are two middle values, and the median is their average. The step people skip is the sort: taking the middle of an unsorted list gives a number with no meaning.

Median

the middle value once the values are sorted; half lie below it and half above

Example 2.1 • What one more student does to each centre

Suppose an eighth student is added who studies 6.5 hours, a value made up for this example. The sorted list becomes 2.0, 2.0, 2.5, 2.5, 3.0, 3.0, 3.5, 6.5. There are now two middle values, the fourth and fifth, so the median is their average. The mean is the new total divided by eight, by equation (2.1).

	Seven students	With the eighth
Mean	2.64	3.125
Median	2.50	2.75

Example 2.1 • continued

One large value moved the mean almost twice as far as the median. The median registered only that the new value lies above the middle, not how far above. Had the eighth student reported 16 hours, the median would have been exactly the same, while the mean would have risen much further.

2.3 The mode

The mode is the most common value. Among the seven students, 2.0, 2.5 and 3.0 each appear twice, so there are three modes. The mode need not be unique, and in a small sample of numbers it is often useless. Across all 120 students the commonest self-study time is 2.0 hours. The mode earns its place with categories. For a nominal variable it is the only centre there is: the modal faculty is Social Sciences & Law, with 35 students, and there is no such thing as a mean faculty. A histogram with two separate humps is called bimodal, and it is often a sign that two different groups have been mixed in one column.

Mode

the most common value or category

2.4 The mean as a balance point

Figure 2 places the seven values on a ruler as equal weights. The point at which the ruler balances is the mean. Each value's deviation is its signed distance from the mean: negative on the left, positive on the right. For the seven students the deviations are -0.14 , -0.64 , 0.36 , 0.86 , 0.36 , -0.14 and -0.64 . Before rounding they add to exactly zero, and they always do. That is a property of the mean and of no other number.

Deviation

how far a value sits from the mean, with a sign: $x - \bar{x}$

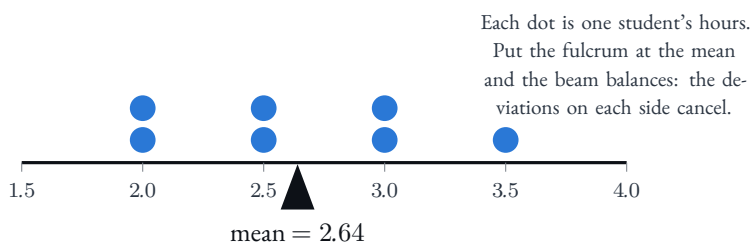


Figure 2: The seven values as weights on a ruler. The ruler balances at the mean. **Simulated** rows 1 to 7.

Two consequences follow, and both return later in the chapter. A value far out on the ruler has a long lever, so it moves the balance point a great deal; that is why the mean follows a long tail. And because the deviations always sum to zero, their plain average is always zero and says nothing about spread; that is why the standard deviation squares them first.

2.5 Shape: symmetric and skewed

Screen time in the class data has a long right tail (Figure 3). Nobody uses a phone for less than zero hours, but the longest reported day is 11.6 hours. Those few pull the mean, 4.81 hours, to the right of the median, 4.40. The median sits in the thick of the data and describes a typical student; the mean has been drawn towards the tail. Three students answered “don’t know”, so every screen-time figure is based on 117 students, not 120.



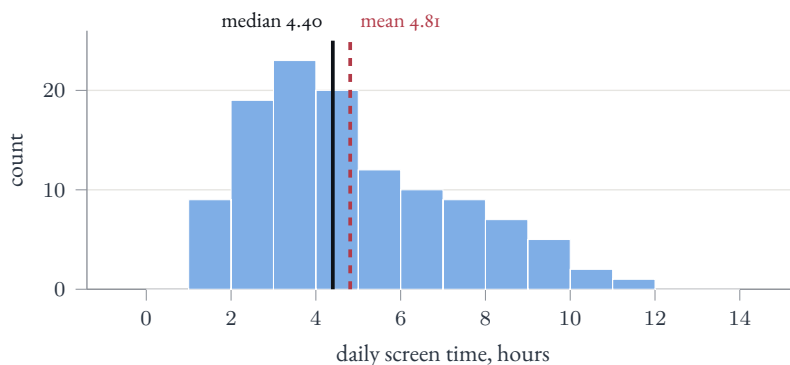


Figure 3: Daily screen time, with the mean and median marked. **Simulated** class data, $n = 117$.

A distribution is skewed when one tail is longer than the other, and it is named after the long tail. A long right tail is right skew, also called positive skew. Income, the number of friends on Facebook and the time taken to finish an exam are all right-skewed. When a distribution is symmetric, the mean and median coincide (Figure 4). When it is skewed, they separate, and the gap between them is a quick reading of the skew. A memory aid: the mean is pulled towards the tail, and the tail names the skew.

Skewed

having one tail longer than the other; named after the long tail

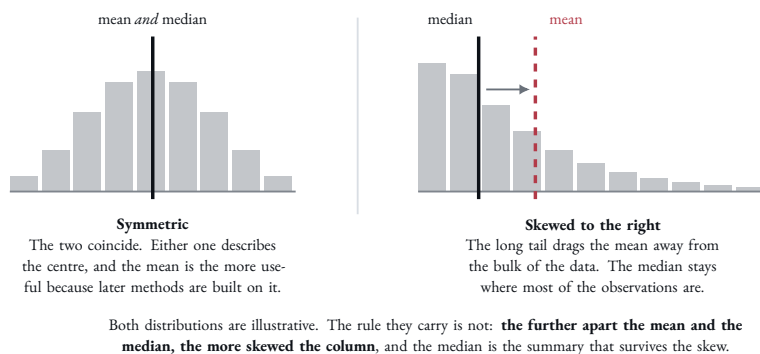


Figure 4: Symmetric and skewed shapes, and where the mean and median fall in each.

2.6 Which centre to report

The choice follows from two questions asked in order: what is the scale of measurement, and what is the shape? Table 6 sets out the answers. For nominal data only the mode makes sense. For ordinal data a middle exists, so the median works, but the gaps between categories are not known to be equal, so a mean is shaky. For numerical data the shape decides. A roughly symmetric variable gets the mean, which uses every value and on which the methods of Chapters 3 to 6 are built. A skewed variable, or one with extreme values, gets the median.

Example 2.2 • The median of an ordinal variable

The median CGPA band is found from the cumulative column of Table 2. Go down it until it first passes 50. It is 45.7 at the end of the second band and 89.5 at the end of the third, so the middle student with a result sits in the third band. The median CGPA band is 3.00–3.49. No arithmetic on the band codes is needed, and none would

Table 6: Choosing a centre from the scale and the shape.

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

Example 2.2 • continued

be meaningful.

Means of single Likert items

Many published papers report the mean of a single agree–disagree item. The practice is common and debated. For a single item, report the median and the distribution of answers. A total over several items, such as the class survey’s stress score, is a different matter, and Chapter 6 discusses when a mean of such a total is defensible.

2.7 The mean of a 0/1 column is a proportion

Code a yes/no variable as 0 for no and 1 for yes, and add the column. The sum counts the yeses, 59 for tuition in the class data. Divide by 120 and the result, 0.492, is the proportion who do tuition, the same 49.2% that the frequency table reports. A proportion is a mean in disguise. The observation looks small and has a long reach. Everything later chapters build for means, including the margin of error in Chapter 3, carries over to proportions because of it, and the logistic regression of Chapter 6 depends on 0/1 coding.

$$\hat{p} = \bar{x} = \frac{\text{number of is}}{n} \quad \text{when every } x \text{ is 0 or 1} \quad (2.2)$$

- $\hat{p}$ the sample proportion of yeses
- $\bar{x}$ the mean of the 0/1 column
- n the number of people with an answer

Only 0 and 1 will do

Some software codes a two-category variable 1 and 2 by default. The mean of such a column, 1.49 for tuition, is not a proportion of anything. Recode to 0 and 1 before taking a mean.

2.8 The weighted mean

Table 7 gives mean weeknight sleep for each residence group. The plain average of the three group means is 6.85 hours, but the mean of all 120 students is 6.80. The plain average gives the smallest group, family-home students, the same say as the largest, and that small group happens to sleep the most. A weighted mean gives each group a say in proportion to its size, and it reproduces the class mean exactly.

Weighted mean

a mean in which each value counts in proportion to its weight, such as a group's size or a course's credits

Table 7: Weeknight sleep by residence. **Simulated** class data.

Lives in	n	Mean sleep, hours
Hall	47	6.74
Mess	43	6.53
Family home	30	7.27
Plain average of the three means		6.85
Weighted by group size	120	6.80

$$\bar{x}_w = \frac{\sum w \bar{x}_g}{\sum w} \quad (2.3)$$

$\bar{x}_w$ the weighted mean

$\bar{x}_g$ the mean of one group, or the grade point of one course

w that group's weight: its size, or the course's credits

The plain average and the weighted mean agree only when the groups are the same size. The same arithmetic is how results are combined across groups, and how the survey weights of Chapter 1 correct a sample that over-represents one stratum.

Example 2.3 • A CGPA is a weighted mean

A **constructed** semester has four courses. Their grade points and credits are made up; each department's grading rules are the authority on real ones.

Course	Credits, w	Grade point	$w \times$ grade point
Course A	3.0	3.75	11.25
Course B	3.0	3.25	9.75
Course C	2.0	4.00	8.00
Lab	1.5	2.50	3.75
Total	9.5		32.75

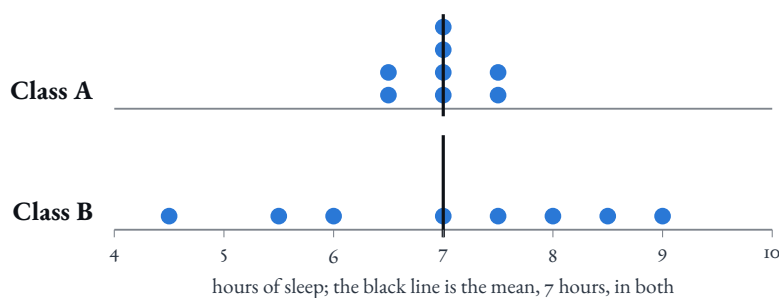
By equation (2.3), the CGPA is $32.75/9.5 = 3.45$. The plain mean of the four grade points is $13.50/4 = 3.375$. The weighted figure is higher because the weakest grade sits in the lightest course, so it pulls less.



3 The spread

3.1 Same average, different classes

Figure 5 shows the weeknight sleep of two **constructed** classes of eight. Both average exactly 7 hours. In Class A nearly everyone sleeps close to 7. In Class B the students range from four and a half hours to nine, and the mean of 7 describes almost nobody. A teacher planning for Class A can treat it as one group; a teacher planning for Class B cannot. A centre without a spread is half a summary, and the spread is often the more important half.



Spread

how far the values of a variable differ from one another; also called variability or dispersion

Figure 5: Two **constructed** classes with the same mean of 7 hours and very different spreads.

3.2 The range

The simplest measure of spread is the range, the largest value minus the smallest. For self-study in the corrected file it is $6.5 - 0.5 = 6.0$ hours. In the file as typed, with the 25-hour row still in it, the range is 24.5 hours. One typing error quadrupled it. The range is built from the two most extreme values, and those are exactly the values most likely to be mistakes, so it is fragile by design.

Range

the largest value minus the smallest

It still deserves a place in every first look at a dataset, for a different reason. The minimum and maximum of each column are the fastest check for impossible values. A maximum of 25 hours in a day is caught in one glance, before any statistics are computed. The codebook of Chapter 1 allows self-study from 0 to 16 hours, and a range check against that rule would have flagged row 37 at once.

3.3 The standard deviation

A good measure of spread should use every value, and should say how far values typically sit from the centre. The deviations of section 2.4 measure exactly that, one student at a time. Their plain average is useless, because they always sum to zero. Squaring removes the signs, so the negatives stop cancelling. The average of the squared deviations is the variance, and its square root, which returns the answer to the original units, is the standard deviation (SD).

Standard deviation

the square root of the variance, which is the average squared deviation; roughly, the typical distance of a value from the mean

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}, \quad s = \sqrt{s^2} \quad (3.1)$$

s^2 the sample variance, in squared units



s the sample standard deviation, in the units of x

$x - \bar{x}$ each value's deviation from the mean

$n - 1$ the number of values less one; section 3.4 explains why

Example 3.1 • The standard deviation of seven students, by hand

The seven self-study values from section 2.1 have mean 2.64. Work in four steps, one column at a time.

Student	x	$x - \bar{x}$	$(x - \bar{x})^2$
1	2.5	-0.14	0.02
2	2.0	-0.64	0.41
3	3.0	0.36	0.13
4	3.5	0.86	0.73
5	3.0	0.36	0.13
6	2.5	-0.14	0.02
7	2.0	-0.64	0.41
Sum	18.5	0	1.86

Step 1. Subtract the mean from each value. **Step 2.** Square each deviation. **Step 3.** Add the squares and divide by $n - 1 = 6$: $s^2 = 1.86/6 = 0.31$ hours squared. **Step 4.** Take the square root: $s = \sqrt{0.31} = 0.56$ hours.

The squares in the table are rounded to two decimals, and they add to 1.85; the sum printed, 1.86, uses the unrounded squares. Either leads to the same SD at two decimals. In words: a typical student among these seven studies about half an hour more or less than 2.64 hours.

The SD is never negative, and it is zero only when every value is the same. It is in the units of the data, which is what makes it readable: across the whole corrected class, self-study has an SD of 1.21 hours. Like the mean, it uses every value, and so, like the mean, it is pulled by extreme ones. Because the deviations are squared, it is pulled harder.

3.4 Why divide by $n - 1$

Dividing the sum of squares by 7 instead of 6 gives an SD of 0.52 for the seven students, not 0.56. The textbook formula divides by $n - 1$ for a reason, and there are two ways of seeing it.

The short reason

The deviations are measured from the *sample* mean, which by construction sits in the middle of these seven values. No other point, including the true mean of all students, gives a smaller sum of squared deviations. Measured from the population mean, the deviations would on average be a little larger. Dividing by $n - 1$ instead of n makes up the shortfall, on average.

The formal reason is called degrees of freedom. The seven deviations must add to zero. Know any six of them and the seventh is forced. Only $n - 1 = 6$ of them are free to vary, and the

Degrees of freedom

the number of values free to vary once a constraint, such as a fixed mean, is imposed



variance averages over the free ones. The phrase returns in every test from Chapter 4 onward.

For a large sample the choice hardly matters: with 120 students the two versions differ by less than half of one per cent. For a small sample it matters a great deal, and calculators often offer both. The key marked s or σ_{n-1} is the one to use for a sample; the key marked σ or σ_n divides by n .

3.5 The two-SD rule of thumb

The SD becomes useful in a sentence through a rule of thumb. For a variable whose histogram has a single hump and is roughly symmetric, about two-thirds of the values lie within one SD of the mean, and about 95% within two SDs. Weeknight sleep in the class data has mean 6.80 hours and SD 0.93 (Table 8). The rule fits it closely. Chapter 3 gives the rule its proper name, the normal distribution, and makes the two-thirds and 95% precise.

Table 8: How many students sit within one and two SDs of the mean. **Simulated** class data. The band limits are our arithmetic from the rounded mean and SD.

	Weeknight sleep		Screen time	
	Hours	Inside	Hours	Inside
mean $\pm$ 1 SD	5.87 to 7.73	70.8%	2.50 to 7.12	67.5%
mean $\pm$ 2 SD	4.94 to 8.66	96.7%	0.18 to 9.44	95.7%
Beyond 2 SD, low side		2 students		0 students
Beyond 2 SD, high side		2 students		5 students

Screen time is skewed, and its shares, 67.5% and 95.7%, also land near the rule. The last two rows of Table 8 show what the shares hide. The rule assumes that the students outside the band split evenly between the two sides, as they do for sleep. For screen time every one of the 5 students beyond two SDs is a heavy user; there are no unusually light ones. A summary that says “95% within two SDs” is silent on which side the rest fall, and for a skewed variable the answer is one side only. The histogram should come first.

A value’s distance from the mean in SDs is a quick measure of how unusual it is. A student who sleeps 4.5 hours on a weeknight is $(4.5 - 6.80)/0.93 = -2.47$ SDs from the mean: well outside the band that holds 95% of the class. Chapter 3 calls this number a z -score.

3.6 Percentiles and quartiles

A percentile is a value with a stated share of the data below it. The median is the 50th percentile. The lower and upper quartiles, Q_1 and Q_3 , are the 25th and 75th, and with the median they cut the sorted data into four parts of equal size. The distance between the quartiles is the interquartile range.

Percentile

a value with a stated percentage of the data at or below it

$$\text{IQR} = Q_3 - Q_1 \quad (3.2)$$

Q_1 the lower quartile: a quarter of the values lie below it

Q_3 the upper quartile: three quarters lie below it

Interquartile range

$Q_3 - Q_1$: the spread of the middle half of the data



The IQR is the spread that belongs with the median. It measures the middle half of the data and ignores both tails, so the extremes cannot move it. For screen time (Table 9) the middle half of the class uses its phone between 3.00 and 6.30 hours a day, an IQR of 3.30 hours. A student at the 90th percentile, 8.12 hours, uses the phone more than nine in ten classmates.

Table 9: Percentiles of daily screen time. **Simulated** class data, $n = 117$. Computed by R's default rule.

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

Example 3.2 • Quartiles by hand

The rule used by hand in this course is Tukey's: sort the values and find the median; Q_1 is the median of the lower half, and Q_3 is the median of the upper half. When n is odd, the median itself belongs to both halves.

The seven students, sorted, are 2.0, 2.0, 2.5, 2.5, 3.0, 3.0, 3.5. The median is the fourth value, 2.50. The lower half is the first four values, whose median is 2.25; the upper half is the last four, whose median is 3.00. The IQR is $3.00 - 2.25 = 0.75$ hours.

Software disagrees, a little

There are several accepted rules for quartiles, and SPSS, R, Excel and a hand calculation each use a different default. In a small sample they can disagree in the second decimal place, and occasionally more. None of them is wrong. State the software used, and do not assume a small difference from a classmate's answer is a mistake.

3.7 The box plot

A box plot draws five numbers: the smallest value, Q_1 , the median, Q_3 and the largest value, together called the five-number summary. The box runs from Q_1 to Q_3 and holds the middle half of the data. A line inside marks the median. The whiskers reach out to the last values inside the fences of section 3.8, and anything beyond a fence is drawn on its own. Figure 6 is the box plot for screen time. The median sits to the left of the middle of the box, and the right whisker is longer than the left: the same right skew that the histogram showed, read from five numbers.

Box plot

a chart of the five-number summary, with values beyond the fences drawn as separate points

3.8 Fences: flagging values that stand out

Tukey's rule marks a value as unusual when it lies more than one and a half IQRs beyond a quartile (reference 3). The two limits are called fences.

Fences

the limits $Q_1 - 1.5 \times \text{IQR}$ and $Q_3 + 1.5 \times \text{IQR}$; values beyond them are flagged

$$\text{lower fence} = Q_1 - 1.5 \times \text{IQR}, \quad \text{upper fence} = Q_3 + 1.5 \times \text{IQR} \quad (3.3)$$

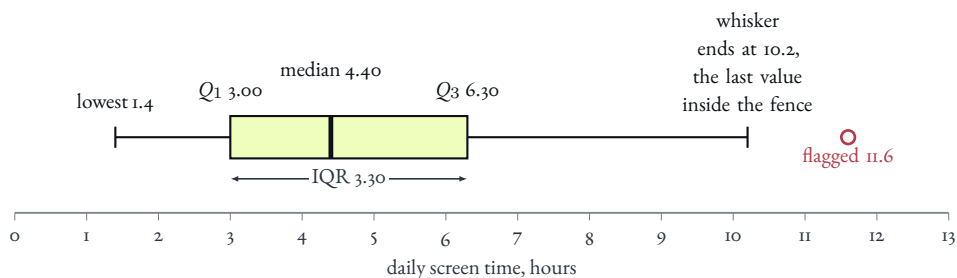


Figure 6: Box plot of daily screen time. **Simulated** class data, $n = 117$. The whiskers end at 1.4 and 10.2 hours; one value lies beyond the upper fence.

1.5 a convention chosen by Tukey, because it flags few values in well-behaved data; not a law of nature

For self-study in the corrected file, the upper fence is $3.50 + 1.5 \times 1.50 = 5.75$ hours. One student reports 6.5 hours, and is flagged (Figure 7). The lower fence is $2.00 - 1.5 \times 1.50 = -0.25$ hours. Nobody studies for a negative time, so nobody can be flagged at the bottom.

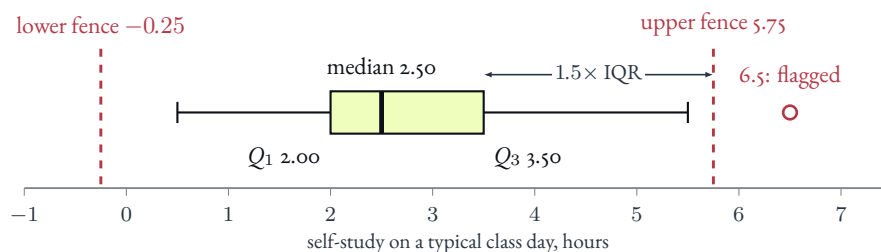


Figure 7: Self-study hours with Tukey's upper fence. **Simulated** class data, corrected file.

A fence flags a value; it does not convict it

Six and a half hours of study on a class day is entirely possible, for example before an examination. The rule found an unusual value. Only a person checking the form can decide whether it is an error. Deleting a believable value because a rule flagged it makes the data tidier and less true.

3.9 The answer to the 25-hour question

The chapter opened by asking how far one wrong key moves the summaries of 120 students. Figure 8 and Table 10 give the answer.

Table 10: Self-study summaries with and without one typing error. **Simulated** class data, $n = 120$.

	As typed	Corrected
Mean	2.81	2.62
Median	2.50	2.50
SD	2.38	1.21
IQR	1.50	1.50
Range	24.5	6.0



Figure 8: Self-study hours as typed, with the 25-hour row, and corrected. **Simulated** class data.

The mean moved by less than a fifth of an hour. The error added $25 - 2.5 = 22.5$ hours to the total, and shared among 120 students that is $22.5/120 \approx 0.19$ hours each: easy to miss. The SD almost doubled, because the error's deviation is squared before it is averaged. The median and IQR did not move at all. They depend only on which values sit at the middle and at the quartiles. Moving one student from 2.5 to the top of the sorted list shifts the others by at most one place, and many students share each of those values, so the values at those places stayed the same. The range grew fourfold.

What one wrong key teaches

The mean and SD use every value and respond to every error; the SD responds most. The median and IQR use only the middle of the data and ignore the tails. And the cheapest defence of all is not a statistic but the allowed range in the codebook, which would have caught the error before any of this was computed. Check ranges first.

3.10 The undeclared 99

Chapter 1's codebook stores a "don't know" answer to the screen-time question as 99. Three students in the simulated class gave that answer. If the code is declared as missing, the software leaves those three cells out and summarises the other 117. If it is not declared, the software reads each 99 as ninety-nine hours of screen time in a day of twenty-four. Table 11 shows the result.

Table 11: Screen time with the "don't know" code declared as missing, and left in. **Simulated** class data.

	99 declared missing	99 left in
<i>n</i>	117	120
Mean	4.81	7.16
Median	4.40	4.60
SD	2.31	14.94

Three undeclared codes raised the mean by 2.35 hours and multiplied the SD more than sixfold. The median crept up a little, because three extra values were added above the middle. Nothing in the output warns that anything went wrong: the software has done exactly what it was told. The only defences are to declare every missing code before any analysis, and to read the minimum and maximum of every column. A maximum of 99 hours is as impossible

as a maximum of 25.

3.11 What to do with a value that stands out

Four questions, in order

1. **Is it possible?** Twenty-five hours in a day is not. Go back to the paper form. Correct the value if the form settles it; code it as missing if it does not. Write down what was done.
2. **Is it plausible?** Six and a half hours of study is. Keep it.
3. **Does it change the answer?** Run the analysis with and without the value, and report both if they differ.
4. **Report it.** For example: “One value was corrected from 25 to 2.5 after checking the paper form.”

The fourth step is part of the method, not an afterthought. A reader cannot judge a result without knowing how the data were cleaned, and a cleaning log, a list of every change with its reason, is standard practice in the laboratory sessions.

Cleaning log

a written record of every change made to the data, with the reason

Never for convenience

Never delete a value because it is inconvenient, or because removing it makes a result “significant”. Removing awkward values until a test gives the hoped-for answer is a recognised form of research misconduct. Chapter 4 returns to it.

3.12 Units, and the coefficient of variation

Summaries follow a change of units in a predictable way, and the rules save time and catch errors. Multiplying every value by a constant multiplies both the centre and the spread. Weeknight sleep in minutes has mean $6.80 \times 60 = 408$ and SD $0.93 \times 60 = 55.8$ minutes. Adding a constant to every value moves the centre and leaves the spread alone. If every student slept half an hour more, the mean would become 7.30 hours, and the SD would stay 0.93.

The reason is that spread is about distances between students. Adding the same amount to everyone changes no distance, while multiplying stretches every distance by the same factor. The median follows the same rules as the mean, and the IQR follows the same rules as the SD. One practical use: a colleague’s mean and SD in minutes can be converted to hours without their data.

An SD is in the units of its variable, so an SD of sleep in hours cannot be compared with an SD of CGPA in grade points. The coefficient of variation (CV) divides out the units by expressing the SD as a percentage of the mean.

Coefficient of variation

the SD as a percentage of the mean; meaningful only on a ratio scale

$$CV = \frac{s}{\bar{x}} \times 100 \quad (3.4)$$

s the standard deviation

$\bar{x}$ the mean, which must be on a ratio scale



Weeknight sleep has a CV of 13.6%, self-study 46.3% and screen time 48.1%. Relative to their averages, students differ far more in how much they study than in how much they sleep. Sleep is constrained by the body in a way that study is not.

Ratio scales only

The CV divides by the mean, which only makes sense when zero means none: a ratio scale, in Chapter 1's terms. The stress score runs from 5 to 25, has no true zero, and its CV would change if the items were scored 0 to 4 instead of 1 to 5. Its CV means nothing. The same holds for temperature in degrees Celsius.

3.13 Skewness as a number

Software prints a skewness figure alongside the mean and SD. It is positive for a long right tail, negative for a long left tail, and near zero for a roughly symmetric shape. Table 12 shows it for four variables beside the gap between mean and median, and the signs agree.

Skewness

a number summarising the lean of a distribution: positive for a long right tail, negative for a long left tail

Table 12: Skewness and the mean-minus-median gap. **Simulated** class data. The gaps are our subtraction.

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

A common rule of thumb reads a skewness between -0.5 and $+0.5$ as roughly symmetric and one beyond ± 1 as strongly skewed. It is only a rule of thumb: a skewness of 0.49 and one of 0.51 do not describe different kinds of data, and different programs compute skewness by slightly different formulas. Sleep leans a little to the left, because a few students sleep very little while nobody sleeps far more than the rest. The number summarises the histogram; it does not replace looking at it.

3.14 Reporting a descriptive table

Almost every quantitative paper opens its results with a table like Table 13, so common that readers call it “Table 1”. It gathers the chapter's rules in one place. Roughly symmetric numerical variables are reported as mean (SD). Skewed ones are reported as median (IQR). Categories are reported as a count with its percentage. And each row states its own n , because different questions have different numbers of missing answers.

An SD nearly as large as the mean

Many papers report mean (SD) for every numerical variable. For a variable that cannot fall below zero, an SD close to the mean is a warning sign: two SDs below the mean would be a negative value, which is impossible, so the distribution must have a long right tail. Screen time here has an SD of 2.31 against a mean of 4.81, about half of it, and its histogram shows the long right tail.

Table 13: A descriptive table in the form papers use. **Simulated** class data. Each row's n is the number who answered that question.

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

4 Charts

4.1 Choosing a chart from the variables

The types of the variables, from Chapter 1, decide the chart. The first question is how many variables the chart must show; the second is whether each is categorical or numerical. Table 14 gives the usual choice for each case, and the choice to avoid. The rest of this section takes the rows in turn.

Table 14: Choosing a chart from the types of the variables.

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

4.2 The histogram

A histogram is the grouped frequency table of section 1.3, drawn. Each bar covers one class interval and its height is the count. Unlike the bars of a bar chart, the bars touch, because the scale is continuous and the classes meet. Figure 9 is the histogram of self-study. The shape is read aloud in three parts: where the bulk of the data lies, which way any tail runs, and whether there is more than one hump. Here the bulk sits between one and a half and three hours, a tail trails off to the right, and there is a single hump. That matches the skewness of 0.74 and the mean lying above the median.

The width of the bars is a choice, and it can change the story. Figure 10 draws the same 120 students with two-hour bars. The right tail disappears into a single short bar, the student at 6.5 hours vanishes, and the skew is hard to see. Very narrow bars go wrong the other way, showing every accident of the sample as if it were a feature. There is no single right

Histogram

a chart of a numerical variable in which touching bars show how many values fall in each class interval



Figure 9: Histogram of self-study hours with half-hour bars. **Simulated** class data, corrected file, $n = 120$.

width. Software picks one by a formula; an analyst should look at two or three widths before trusting any of them, and a reader should be suspicious of a histogram whose shape seems to depend on its bars.

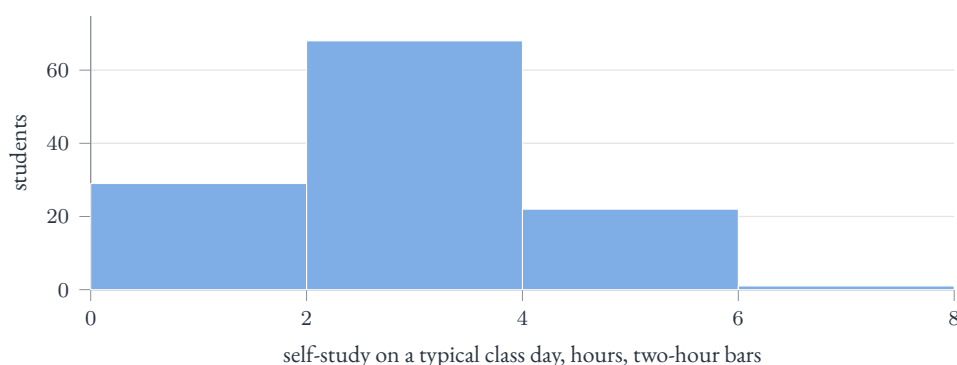


Figure 10: The same data as Figure 9, with two-hour bars. **Simulated** class data.

4.3 Comparing groups

To compare a numerical variable across groups, draw one box plot per group on a single scale (Figure 11). Read the plot in a fixed order: compare the medians, then the heights of the boxes, then the whiskers, then any flagged points. Here the three medians sit close together and the boxes overlap heavily. Residence makes little difference to screen time in this class. The alternative, a bar chart of three group means, would show three bars of nearly equal height and nothing at all about how much students differ within each group. After section 3 that should feel like a loss.

For two categorical variables, stretch each group to a bar of the same length and divide it by the second variable (Figure 12). The coloured part of each bar is the row percentage of section 1.5, so the chart answers the row-percentage question at a glance: mess students stand out. Its weakness is that the bars hide the group sizes. A group of five and a group of five hundred look the same, so print each group's n beside its bar.

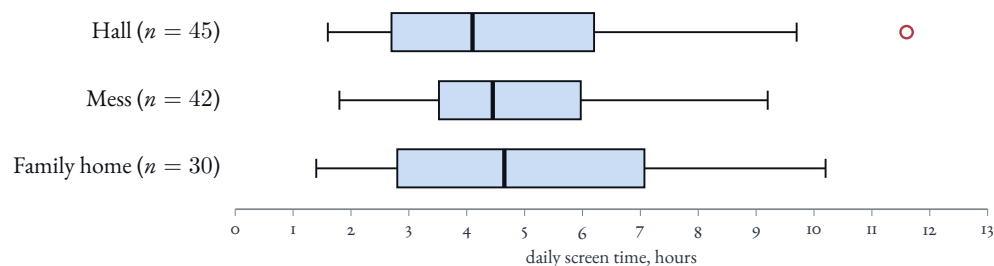


Figure 11: Daily screen time by residence. **Simulated** class data. Group sizes are those who answered the screen-time question.

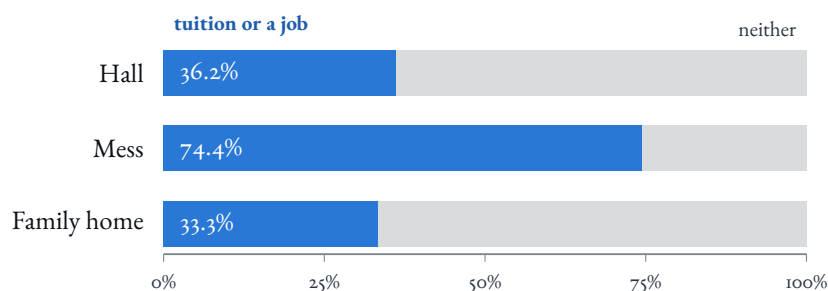


Figure 12: Tuition or a paid job, by residence, as row percentages. **Simulated** class data; the groups have 47, 43 and 30 students.

4.4 The pie chart

A pie chart shows the parts of a whole as slices, and asks the eye to compare angles, which it does badly. Figure 13 draws the faculty table as a pie. Which faculty is the largest? In the pie, Science & Engineering and Social Sciences & Law look much the same. In the bar chart of Figure 1, 33 against 35 is plain, because the eye compares lengths on a common baseline far better than angles. A pie is reasonable for two or three parts when the differences are large. Beyond that, bars are better.

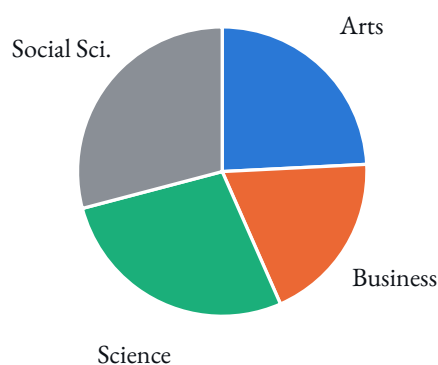


Figure 13: The faculty table as a pie chart. **Simulated** class data; the same four numbers as Figure 1.

4.5 The scatter plot

A scatter plot shows two numerical variables together, with one dot for each student (Figure 14). Describe it in three words. **Direction:** the dots drift down to the right, so more screen time goes with a little less sleep. **Strength:** loose, with plenty of scatter around the trend. **Unusual points:** any dot that sits apart from the cloud. What the plot cannot say is why the two move together. Chapter 1's third variable is waiting: stress, or a crowded mess, might drive both. Chapter 5 puts a number on the strength of such a trend.

Scatter plot

a chart of two numerical variables with one dot per unit

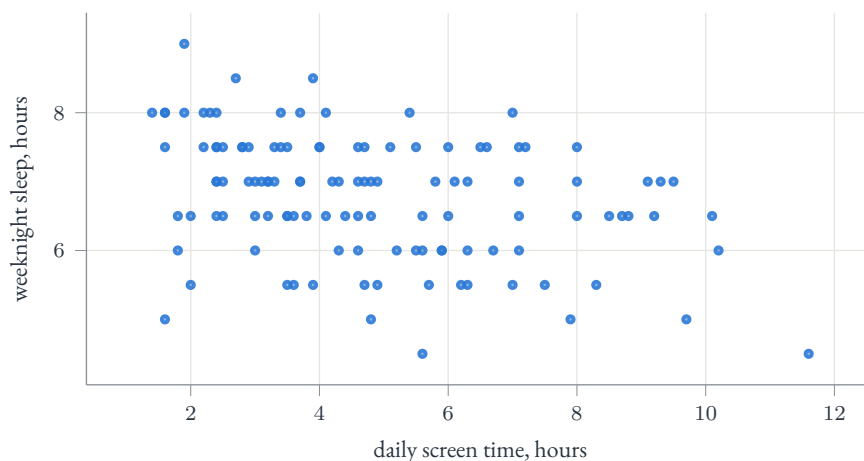


Figure 14: Weeknight sleep against daily screen time. **Simulated** class data, $n = 117$ students with both answers.

4.6 Anscombe's quartet

In 1973 the statistician Francis Anscombe published four small datasets built to make a single point (reference 2). All four have the same mean of x , 9.00; the same mean of y , 7.50; the same correlation, 0.82 to two decimal places; and the same fitted straight line, $y = 3.00 + 0.500x$. Plotted, they are nothing alike (Figure 15). Set I is a loose straight-line trend, and its summaries describe it well. Set II is a smooth curve. Set III is a tight line with one stray point. In set IV every x is the same except one, and that single point makes the whole trend.

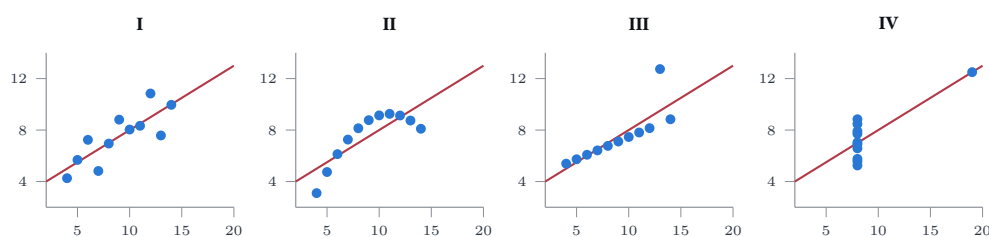


Figure 15: Anscombe's four datasets. Data: Anscombe 1973, via R's `anscombe` data. Summaries computed by us. To three places the correlations are 0.816 for sets I to III and 0.817 for set IV.

Before any test, look

Identical summaries can describe completely different data. A mean, an SD or a correlation is a faithful description only of data shaped the way it assumes. The histogram, the box plot and the scatter plot come first; the numbers come second.

5 Charts that mislead

5.1 A bar axis that does not start at zero

In a bar chart the length of the bar is the message, so the axis must start at zero. Figure 16 shows the share of students who do tuition among hall students, 36.2%, and family-home students, 33.3%. On the left the axis starts at 32%, and the hall bar looks more than three times as tall as the family-home bar. On the right the axis starts at zero, and the two groups look like what they are: almost the same. The gap is under three percentage points.

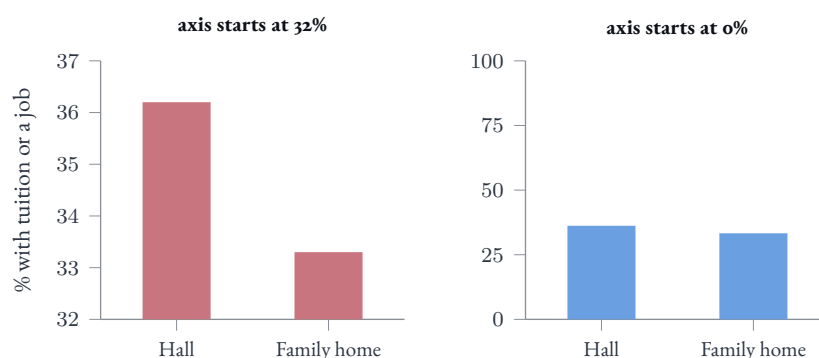


Figure 16: The same two percentages with a cut axis and with a full one. **Simulated** class data.

The rule belongs to bar charts. In a line chart the message is the change from point to point, not the height above zero, and an axis zoomed in on the range of the data is often the honest choice. A reader should not overcorrect and distrust every line chart whose axis does not start at zero.

5.2 Icons that grow in two directions

A pictogram shows a quantity as a picture, and it misleads when the picture is scaled in two directions at once (Figure 17). A value doubles from 10 to 20; the designer doubles the icon's height and its width; the icon's area, which is what the eye perceives, becomes four times as large. A three-dimensional icon is worse again, since its volume grows with the cube of its side. The fix is to scale one dimension only, as a bar does, or to repeat the icon: two small icons for 20, one for 10.

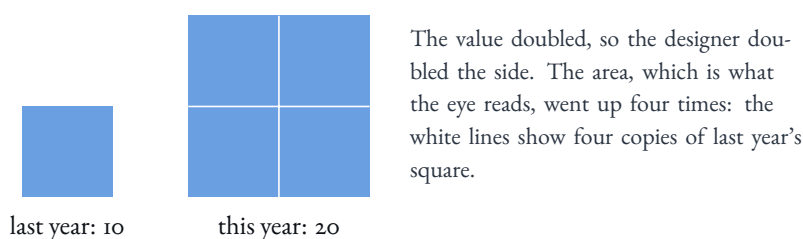


Figure 17: A value that doubles, drawn by doubling both sides of the icon. **Constructed** example.

5.3 Choosing the window

Figure 18 shows a **constructed** series over 24 months. Over the whole period it goes nowhere. Months 14 to 19, taken alone, run from 49 to 64, and a press release could report “up 31% in six months”. Every number in that sentence is correct, and the choice of start and end points makes the story. The defence is to ask for the longest honest window, and to ask why a chart starts where it does. A start date that looks oddly specific was probably chosen.

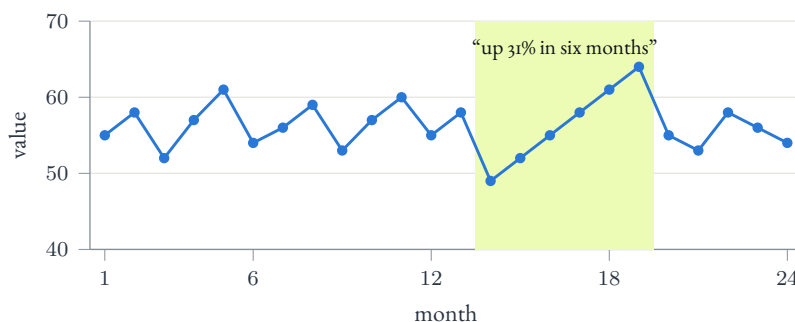


Figure 18: A **constructed** series, with a six-month window picked out.

5.4 A percentage without its count

A percentage hides its denominator. “Two of three” and “two hundred of three hundred” both print as 67%, and they deserve very different trust. In the class data, 74.4% of mess students do tuition or a job, which is 32 of 43. With 43 students, one person moves the percentage by $100/43 = 2.3$ points. In a group of three, one person moves it by 33 points. The rule is short: print n beside every percentage, so the reader can judge.

5.5 Six questions for any chart

The checklist

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

6. Who made the chart, and from what data?

The six questions work on charts in newspapers, on Facebook pages and in advertising, and they work on charts in published papers. They are meant to become a habit, applied to every chart without being asked.

6 Describing one variable: the procedure

Everything in this chapter comes together when a single variable must be described from scratch. The procedure below is the one the exercises in Part D follow, using a 20-row extract of the class data.

Describing one variable

1. **Decide the type and scale.** Categorical or numerical; nominal, ordinal, interval or ratio. Everything else follows from this.
2. **Deal with the codes.** Find the missing and not-applicable codes in the code-book, set them aside, and state how many were set aside.
3. **Summarise.** For a categorical variable, a frequency table with counts and percentages, and valid and cumulative percentages where they mean something. For a numerical variable, the five-number summary, the mean and the SD, and a check against the fences.
4. **Draw.** A bar chart for a categorical variable; a histogram or box plot for a numerical one.
5. **Write one sentence** a friend would understand. For a skewed numerical variable, the median and IQR; for a symmetric one, the mean and SD; for a category, the largest group with its count and percentage.

The sentence is the test. “The median student spends 4.40 hours a day on the phone, and the middle half spend between 3.00 and 6.30 hours” is what a good sentence looks like: plain words, a centre, a spread, and units.

7 Reading a table in a published paper

7.1 Islam 2020, Table 2

Islam and colleagues surveyed 476 university students in Bangladesh during the COVID-19 pandemic, and screened for depression with the PHQ-9 questionnaire (reference 1; Chapter 1 introduced the study). Their Table 2 crosses many characteristics with the five PHQ-9 categories. Table 15 reproduces the rows for private tuition exactly as printed, count first and percentage in brackets.

Table 15: Private tuition by depression category, as printed in Islam 2020, Table 2. Reproduced under CC BY 4.0.

Tuition	Depression, PHQ-9 category				
	None–minimal	Mild	Moderate	Mod. severe	Severe
No	44 (9.2)	71 (14.9)	59 (12.4)	37 (7.8)	29 (6.1)
Yes	44 (8.4)	65 (13.7)	74 (15.5)	35 (7.4)	22 (4.6)

ISLAM 2020: WHOSE PERCENTAGE IS IT?

Each bracketed figure is a total percentage: $71 \div 476 = 14.9\%$. It is the share of all 476 students who are in that cell. That is a legitimate choice, but it cannot compare the two tuition groups, because the groups are of different sizes. Table 1 of the paper gives them: 240 students do no tuition and 236 do. To compare, divide within each group. Grouping moderate, moderately severe and severe together, which is our choice and not the paper's, 125 of 240 students without tuition (52.1%) and 131 of 236 with tuition (55.5%) screen at moderate or worse. The two groups come out close.

7.2 Three checks catch one misprint

Look again at the first cell of the “Yes” row, printed “44 (8.4)”. Three independent checks, all our own arithmetic on the paper's printed figures, show that it cannot be right.

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

All three checks point to the same value. With 40 in the cell, the row adds to 236, the column to 84, and the percentage to 8.4. The count was mistyped and the percentage was right. The comparison in the box above is unaffected: its denominator, 236, comes from Table 1, and the misprinted cell is not among the counts it adds. The conclusion of the paper hardly changes. The habit is the point: the totals and percentages in a table are redundant on purpose, and that redundancy is a free check that every reader can run. The same paper has a similar slip in its text, found in Chapter 1, and it passed peer review with both in place.

8 Chapter summary

1. Before any test, look: tables and charts come first, and they catch errors that summaries hide.
2. A frequency table gives counts and percentages. Valid percentages leave out missing and not-applicable cases; cumulative percentages need ordered categories.
3. In a cross-tabulation, the question picks the denominator. To compare groups, divide within each group: the row percentage when the groups are the rows.
4. The mean is the fair share and the balance point; it uses every value and is pulled by a long tail. The median is the middle value and resists the tail. The mode is the only centre for nominal data.
5. The mean of a 0/1 column is a proportion. A weighted mean gives each group a say in proportion to its size, and a CGPA is one.
6. The SD is roughly the typical distance from the mean, computed with $n - 1$ because only $n - 1$ deviations are free. The IQR is the spread of the middle half.
7. In a single-humped, roughly symmetric distribution, about two-thirds of values lie within one SD of the mean and about 95% within two. The rule does not say on which side the rest fall.
8. Report mean (SD) for symmetric data and median (IQR) for skewed data, with the n for every row.
9. Tukey's fences flag values; only checking decides whether a value is wrong. Check ranges and declare missing codes before computing anything.
10. Adding a constant moves the centre only; multiplying moves centre and spread. The CV needs a ratio scale.
11. The variables decide the chart. Bar axes start at zero, areas are scaled honestly, windows are not cherry-picked, and every percentage carries its n .
12. Totals and percentages in a published table can be checked against each other, and sometimes they disagree.

9 Key terms

Bar chart and histogram

Separate bars for categories; touching bars for class intervals of a numerical variable.

Box plot

A chart of the five-number summary, with values beyond the fences drawn separately.

Class interval

One of the equal-width ranges into which a numerical variable is cut.

Cleaning log

A written record of every change made to the data, with its reason.

Coefficient of variation

The SD as a percentage of the mean, equation (3.4); ratio scales only.

Cross-tabulation

A table counting two categorical variables together, with margins that reproduce each single-variable table.

Cumulative percentage

A running total of valid percentages over ordered categories.

Degrees of freedom

The number of values free to vary once a constraint is imposed; $n - 1$ for the SD.

Deviation

A value's signed distance from the mean, $x - \bar{x}$.

Fences

$Q_1 - 1.5 \times \text{IQR}$ and $Q_3 + 1.5 \times \text{IQR}$, equation (3.3).

Five-number summary

Minimum, Q_1 , median, Q_3 , maximum.

Frequency table

Each category of a variable with its count and percentage.

Interquartile range

$Q_3 - Q_1$, the spread of the middle half.

Mean, median, mode

The fair share; the middle value; the commonest value.

Percentile, quartile

A value with a stated share of the data below it; the 25th, 50th and 75th percentiles.

Range

The largest value minus the smallest.

Row, column and total percentage

A cell divided by its row total, its column total, or the grand total.

Scatter plot

Two numerical variables, one dot per unit.

Skewed, skewness

Having one tail longer than the other; a number summarising that lean.

Standard deviation, variance

The typical distance from the mean, equation (3.1); its square.

Valid percentage

A percentage among those with an answer.

Weighted mean

A mean in which each value counts in proportion to its weight, equation (2.3).

10 Exercises

Exercises in Part A need no calculation. Parts B, C and D need a calculator. Numbers not attributed to a paper or to the class data are constructed for teaching.

Part A. Ideas.

1. Name the centre and the spread to report for each variable. (a) Faculty. (b) CGPA band. (c) Weeknight sleep, roughly symmetric. (d) Monthly pocket money, with a few students receiving far more than the rest. (e) A single agree–disagree item.
2. A report says: “60% of students who failed the course live in a mess.” (a) Is this a row, column or total percentage, if the rows are residence groups and the columns are pass and fail? (b) Does it show that mess students fail more often than others? What would?
3. Explain in two or three sentences why the deviations from the mean always add to zero, and why that forces the standard deviation to square them.
4. Monthly household income for 50 families has a mean of 45,000 taka and a median of 28,000 taka. (a) What shape does this suggest? (b) Which centre would you report, and why? (c) Is the mean wrong?
5. Software prints a cumulative percentage column for faculty and for year of study. Explain why one column means something and the other does not.
6. Take each value through the four questions of section 3.11, and say what you would do. (a) A weeknight sleep of 14 hours. (b) Self-study of 7 hours on a class day. (c) A CGPA of 4.35.

Part B. Calculation.

7. Seven students’ commuting times to campus, in minutes, are 15, 25, 10, 40, 20, 15 and 30. Find the mean, median, mode and range. Then find the standard deviation by hand, setting out a table like the one in Example 3.1.
8. An eighth student in Exercise 7 lives far away and commutes for 120 minutes. Find the new mean and median. Which moved more, and why?
9. Find the standard deviation of each class in Figure 5. Class A sleeps 6.5, 7, 7, 7, 6.5, 7.5, 7.5 and 7 hours; Class B sleeps 4.5, 5.5, 6, 7, 7.5, 8.5, 9 and 8 hours. Check first that both means are 7.
10. Three sections of a course have 40, 25 and 15 students, with mean marks of 62, 70 and 58. Find the mean mark of all the students together. Compare it with the plain average of the three section means, and explain the difference.
11. A student takes four courses with 3, 3, 2 and 1 credits, and earns grade points of 3.25, 3.75, 3.50 and 2.75 respectively. Find the semester GPA, and compare it with the plain mean of the four grade points.



- 12.** Eleven students report self-study hours of 1.5, 2.0, 2.5, 3.0, 3.0, 3.5, 4.0, 4.5, 5.0, 6.0 and 12.0. (a) Find the median and the quartiles by the hand rule of Example 3.2, and the IQR. (b) Find both fences. Is any value flagged? (c) Find the mean. (d) What would you do next about the flagged value?
- 13.** A study reports that students take a mean of 42 minutes to fall asleep, with an SD of 9 minutes. (a) Give the mean and SD in hours. (b) If every student took 5 minutes longer, what would the new mean and SD be, in minutes?
- 14.** Find the coefficient of variation where it is meaningful. (a) Heights of students: mean 160 cm, SD 8 cm. (b) Monthly spending: mean 3,000 taka, SD 1,500 taka. (c) The class stress score, which runs from 5 to 25: mean 15, SD 4.6. Which group of students varies more, relative to its average?
- 15.** Of 30 students, 12 own a laptop. (a) Code owning a laptop as 1 and not owning as 0. What is the mean of the column, and what does it mean? (b) What would the mean be if the codes were 2 for yes and 1 for no, and why is that number useless?

Part C. Tables.

- 16.** In a **constructed** survey, 30 of 80 first-year students and 36 of 60 final-year students used the library in the past week. (a) Draw the cross-tabulation with its margins. (b) Find the row percentages of library users. (c) Of all library users, what percentage are first-years? (d) What percentage of all 140 students are final-years who used the library? (e) Which of (b), (c) and (d) answers “are final-year students more likely to use the library?”
- 17.** Of 50 students in a **constructed** class, 8 have no result yet. The other 42 fall in the four CGPA bands of Table 2 with counts 4, 12, 18 and 8. Build the frequency table with columns for % of all, valid % and cumulative %. What is the median band?
- 18.** Using Table 15, with the misprinted cell corrected to 40, find the row percentages for all five depression categories in each tuition group. Check that each row adds to about 100. In which categories do the two groups differ by more than three percentage points?
- 19.** Table 1 of Islam 2020 reports that 183 of the 476 students exercised during the lockdown. Table 2 gives their depression categories as 35, 57, 52, 20 and 19, and those of the 293 who did not exercise as 49, 79, 81, 52 and 32, in the order of Table 15. (a) If exercise is coded 1 for yes and 0 for no, what is the mean of the column? (b) Check that each row adds to its group size. (c) Compare the two groups’ percentages at moderate or worse.

Part D. Describing the class data by hand.

Table 16 is a 20-row extract of the **simulated** class data, as typed from the forms. The codes are those of Chapter 1’s codebook. Faculty: 1 Arts & Humanities, 2 Business Studies, 3 Science & Engineering, 4 Social Sciences & Law. Residence: 1 hall, 2 mess, 3 family home. CGPA band: 1 below 2.50, 2 2.50–2.99, 3 3.00–3.49, 4 3.50–4.00, 9 no result yet. Screen time: 99 don’t know. Exercises D1 to D6 each describe one variable with the procedure of section 6: the table or summary numbers, a sketched chart, and one plain sentence.

- 20.** Faculty: frequency table, bar chart and sentence.

Table 16: A 20-row extract of the **simulated** class data, before cleaning.

id	faculty	residence	cgpa_band	study_hrs	sleep_we	screen_hrs
1	1	2	4	2.5	8.5	3.1
2	1	3	2	2.0	9.0	5.5
3	1	2	3	3.0	7.5	1.8
4	1	2	9	3.5	9.5	3.7
5	2	2	2	3.0	8.5	6.0
6	2	3	3	2.5	8.0	1.6
7	1	1	2	2.0	7.5	2.3
8	2	3	3	3.5	10.0	7.0
9	3	3	1	1.5	8.0	9.3
10	4	1	2	1.5	8.0	4.0
11	4	3	3	3.5	9.5	4.6
12	1	1	4	3.0	9.5	9.9
13	3	2	4	3.0	6.0	2.0
14	2	3	3	2.0	8.5	7.1
15	4	1	2	1.5	6.5	4.7
16	1	2	2	2.0	7.5	4.2
17	4	3	2	4.5	7.5	8.5
18	1	1	3	3.0	10.0	1.9
19	4	2	3	1.0	6.0	3.6
20	1	2	3	3.0	7.0	3.5

- 21.** Residence: frequency table, bar chart and sentence.
- 22.** CGPA band: frequency table with valid and cumulative percentages, the median band, a bar chart and a sentence. Say how you treated code 9.
- 23.** Self-study hours: the five-number summary, the mean and SD, both fences, a box plot and a sentence. Is any value flagged?
- 24.** Weekend sleep: the five-number summary, the mean and SD, both fences, a histogram and a sentence.
- 25.** Screen time: deal with the code first. Then the five-number summary, the mean and SD, both fences, a box plot and a sentence. Finally, find the mean that software would print if the code were not declared as missing.
- 26.** Put the answers to Exercises 20 to 25 into one descriptive table in the form of Table 13, choosing mean (SD) or median (IQR) for each numerical variable.

Part E. Charts.

- 27.** A poster compares two colleges' pass rates, 82% and 85%, in a bar chart whose axis starts at 80%. (a) How many times taller does the second bar look? (b) What is the real ratio of the two pass rates? (c) Redraw the chart honestly, as a sketch.

- 28.** An infographic shows a budget that tripled by tripling both the height and the width of a money-bag icon. How many times larger does the new icon look? What if the icon were drawn in three dimensions and all three sides tripled?
- 29.** In Figure 18 the series starts at 55 in month 1, runs from 49 in month 14 to 64 in month 19, falls to 55 in month 20, and ends at 54 in month 24. (a) Check the claim “up 31% in six months”. (b) What is the change over the whole 24 months? (c) What headline could someone write about month 20, and why is it equally misleading?
- 30.** Choose a chart for each question, using Table 14. (a) How are the students spread across the four faculties? (b) Does screen time differ between the four years of study? (c) Does weeknight sleep go with the stress score? (d) Does the share doing tuition differ between hall, mess and family-home students? (e) How has a university’s enrolment changed over the past ten years?

II Solutions

Part A.

1. (a) The mode, with the frequency table; no numerical spread exists for nominal data. (b) The median band, with the frequency table or the bands holding the quartiles. (c) The mean and SD. (d) The median and IQR, since the few large amounts pull the mean and SD. (e) The median and the distribution of answers across the five options. (Exercise 1.)
2. (a) A column percentage: of the students who failed, the share living in a mess. (b) No. If 60% of all students live in a mess, mess students would supply 60% of the failures even if they fail no more often than anyone else. The comparison needs row percentages: the failure rate among mess students against the failure rate among the others. (Exercise 2.)
3. The mean is the balance point: the total of the values equals n times the mean, so subtracting the mean from every value and adding the results gives $\sum x - n\bar{x} = 0$. The positive and negative deviations therefore always cancel, and their plain average is zero whatever the spread. Squaring makes every deviation positive, so the squares add up to something that grows with the spread. (Exercise 3.)
4. (a) A long right tail: a few very rich families pull the mean far above the median. (b) The median, 28,000 taka, with the IQR. It describes a typical family. (c) No. The mean is the correct fair share of the total income, and it is the right figure for a question about the total, such as a tax take. It is the wrong figure for “what does a typical family earn?” (Exercise 4.)
5. A cumulative percentage reads “in this category or below”, which needs an order. Years of study have one, so “this share of students are in second year or below” means something. Faculties have none, and a running total over them depends only on the order in which the software happened to list them. (Exercise 5.)
6. (a) Fourteen hours of sleep on a weeknight is physically possible, as in illness, but not plausible as a typical night. Check the form: it may be a slip for 4, or a weekend figure in the wrong column. If the form confirms 14, keep it, run the analysis with and without it, and report it. (b) Plausible, especially before an examination. Keep it, and check whether the results change without it. (c) Impossible on a four-point scale. Check the form; correct it if the form settles the value, otherwise code it as missing, and record the change in the cleaning log. (Exercise 6.)

Part B.

7. Sorted: 10, 15, 15, 20, 25, 30, 40. The sum is 155, so the mean is $155/7 = 22.14$ minutes. The median is the fourth value, 20. The mode is 15. The range is $40 - 10 = 30$.

x	$x - \bar{x}$	$(x - \bar{x})^2$
15	-7.14	51.02
25	2.86	8.16
10	-12.14	147.45
40	17.86	318.88
20	-2.14	4.59
15	-7.14	51.02
30	7.86	61.73
sum		642.86

The rounded squares add to 642.85; the unrounded ones to 642.86. Either way, $s^2 = 642.86/6 = 107.14$ and $s = 10.35$ minutes. (Exercise 7.)

8. The sum becomes 275, so the mean is $275/8 = 34.375$, or 34.4 minutes. Sorted, the eight values are 10, 15, 15, 20, 25, 30, 40, 120, so the median is the average of the fourth and fifth, $(20 + 25)/2 = 22.5$. The mean rose by more than 12 minutes and the median by 2.5. The mean shares the long commute out among everyone; the median only registers that one more value lies above the middle. (Exercise 8.)

9. Both sums are 56, so both means are 7. Class A's deviations are -0.5, 0, 0, 0, -0.5, 0.5, 0.5, 0; the squares add to 1.0, so $s^2 = 1.0/7 = 0.143$ and $s = 0.38$ hours. Class B's deviations are -2.5, -1.5, -1, 0, 0.5, 1.5, 2, 1; the squares add to 17.0, so $s^2 = 17.0/7 = 2.429$ and $s = 1.56$ hours. Class B's SD is about four times Class A's. (Exercise 9.)

10. The weighted mean is $(40 \times 62 + 25 \times 70 + 15 \times 58)/80 = 5,100/80 = 63.75$. The plain average of the three means is $190/3 = 63.33$. The plain average gives the section of 15 the same say as the section of 40; the weighted mean gives each student one say. (Exercise 10.)

11. Credits times grade points: $9.75 + 11.25 + 7.00 + 2.75 = 30.75$, over $3 + 3 + 2 + 1 = 9$ credits: a GPA of $30.75/9 = 3.42$. The plain mean is $13.25/4 = 3.31$. The weakest grade is in the one-credit course, so it counts for less. (Exercise 11.)

12. (a) With 11 values the median is the sixth, 3.5. The lower half, including the median, is 1.5, 2.0, 2.5, 3.0, 3.0, 3.5, so $Q_1 = (2.5 + 3.0)/2 = 2.75$. The upper half is 3.5, 4.0, 4.5, 5.0, 6.0, 12.0, so $Q_3 = (4.5 + 5.0)/2 = 4.75$. The IQR is 2.0. (b) The fences are $2.75 - 3.0 = -0.25$ and $4.75 + 3.0 = 7.75$. The value 12.0 lies beyond the upper fence and is flagged. (c) The sum is 47.0, so the mean is 4.27. (d) Check the form. Twelve hours of study on a class day is possible but implausible; it may be a slip for 1.2 or 2.0. Without it the mean would be 3.50 and the median 3.25: the mean moves a great deal more. (Exercise 12.)

13. (a) Divide both by 60: mean 0.70 hours, SD 0.15 hours. (b) Adding 5 minutes moves the centre only: mean 47 minutes, SD still 9 minutes. (Exercise 13.)



14. (a) $8/160 \times 100 = 5\%$. (b) $1,500/3,000 \times 100 = 50\%$. (c) Not meaningful: the score has no true zero, so dividing by its mean has no fixed meaning. Spending varies ten times as much as height, relative to its average. (Exercise 14.)

15. (a) The column adds to 12, so the mean is $12/30 = 0.4$: the proportion who own a laptop, 40%. (b) The mean would be $(12 \times 2 + 18 \times 1)/30 = 1.4$. It is 1 plus the proportion, which can be decoded, but it is not itself a proportion of anything, and it cannot be read or reported as one. (Exercise 15.)

Part C.

16. (a) First-years: 30 used, 50 did not, total 80. Final-years: 36 used, 24 did not, total 60. Columns: 66 users, 74 non-users, 140 in all. (b) First-years $30/80 = 37.5\%$; final-years $36/60 = 60.0\%$. (c) $30/66 = 45.5\%$ of users are first-years, and 54.5% are final-years. (d) $36/140 = 25.7\%$; the first-year users are $30/140 = 21.4\%$. (e) The row percentages in (b): 60.0% of final-years against 37.5% of first-years. (Exercise 16.)

17.

Band	<i>n</i>	% of all	Valid %	Cumulative %
below 2.50	4	8.0	9.5	9.5
2.50–2.99	12	24.0	28.6	38.1
3.00–3.49	18	36.0	42.9	81.0
3.50–4.00	8	16.0	19.0	100.0
no result yet	8	16.0		
Total	50	100.0	100.0	

The cumulative column first passes 50 in the third band, so the median band is 3.00–3.49. (Exercise 17.)

18. Divide the “No” row by 240 and the corrected “Yes” row by 236.

Tuition	None–min.	Mild	Moderate	Mod. sev.	Severe	Sum
No	18.3	29.6	24.6	15.4	12.1	100.0
Yes	16.9	27.5	31.4	14.8	9.3	99.9

The sums differ from 100 only by rounding. The groups differ by more than three points in two categories: moderate, 31.4% against 24.6%, and severe, 9.3% against 12.1%. Students doing tuition are more often moderate and less often severe. (Exercise 18.)

19. (a) $183/476 = 0.384$: the proportion who exercised, 38.4%, as Table 1 prints it. (b) $35 + 57 + 52 + 20 + 19 = 183$ and $49 + 79 + 81 + 52 + 32 = 293$; both rows add correctly, and $35 + 49 = 84$ matches the none-minimal total. (c) Moderate or worse: $52 + 20 + 19 = 91$ of 183 students who exercised (49.7%), against $81 + 52 + 32 = 165$ of 293 who did not (56.3%). Whether a gap of this size could arise by chance is a question for Chapter 5. (Exercise 19.)

Part D. Every figure below comes from the simulated extract.



20. Arts & Humanities 9 (45.0%), Business Studies 4 (20.0%), Science & Engineering 2 (10.0%), Social Sciences & Law 5 (25.0%); total 20. A bar chart, with bars sorted by size since faculty is nominal. Sentence: “Arts & Humanities is the largest group, with 9 of the 20 students.” The mode is the only centre. (Exercise 20.)

21. Hall 5 (25.0%), mess 8 (40.0%), family home 7 (35.0%). A bar chart. Sentence: “The largest group, 8 of 20, live in a mess.” (Exercise 21.)

22. Code 9 means no result yet, which is not applicable rather than missing: 1 student is set aside, leaving 19.

Band	<i>n</i>	Valid %	Cumulative %
below 2.50	1	5.3	5.3
2.50–2.99	7	36.8	42.1
3.00–3.49	8	42.1	84.2
3.50–4.00	3	15.8	100.0

The cumulative column first passes 50 in the third band, so the median band is 3.00–3.49. A bar chart with the bands in order. Sentence: “Of the 19 students with a result, the middle student is in the 3.00–3.49 band; one student has no result yet.” (Exercise 22.)

23. Sorted: 1.0, 1.5, 1.5, 1.5, 2.0, 2.0, 2.0, 2.0, 2.5, 2.5, 3.0, 3.0, 3.0, 3.0, 3.0, 3.0, 3.5, 3.5, 3.5, 4.5. With 20 values the median is the average of the 10th and 11th, 2.75; the quartiles are the medians of the lower and upper ten. Five-number summary: 1.0, 2.00, 2.75, 3.00, 4.5; IQR 1.00. The sum is 51.5, so the mean is 2.58, and the SD is 0.88. The fences are $2.00 - 1.500 = 0.500$ and $3.00 + 1.500 = 4.500$. The largest value sits exactly on the upper fence, not beyond it, so nothing is flagged. The mean is a little below the median here, while the full class is skewed to the right: twenty rows are too few to show the shape reliably. Either summary is defensible; median (IQR) keeps the report consistent with the full class. Sentence: “The middle student studies 2.75 hours on a class day, and the middle half study between 2.00 and 3.00 hours.” (Exercise 23.)

24. Sorted: 6.0, 6.0, 6.5, 7.0, 7.5, 7.5, 7.5, 7.5, 8.0, 8.0, 8.0, 8.5, 8.5, 8.5, 9.0, 9.5, 9.5, 9.5, 10.0, 10.0. Five-number summary: 6.0, 7.50, 8.00, 9.25, 10.0; IQR 1.75. Mean $162.5/20 = 8.12$, SD 1.22. Fences 4.875 and 11.875: nothing is flagged. The mean and median are close, so report the mean and SD, with a histogram. Sentence: “On a weekend night these students sleep 8.12 hours on average, and about two-thirds are within 1.22 hours of that.” (Exercise 24.)

25. Row 12 has the code 99, don’t know, which is missing: set it aside and summarise the other 19. Sorted: 1.6, 1.8, 1.9, 2.0, 2.3, 3.1, 3.5, 3.6, 3.7, 4.0, 4.2, 4.6, 4.7, 5.5, 6.0, 7.0, 7.1, 8.5, 9.3. With 19 values the median is the 10th, 4.00, and it belongs to both halves. Five-number summary: 1.6, 2.70, 4.00, 5.75, 9.3; IQR 3.05. Mean $84.4/19 = 4.44$, SD 2.28. Fences -1.875 and 10.325 : nothing is flagged. The mean exceeds the median and the upper whisker is the longer, so the shape leans right: report the median and IQR, with a box plot. Sentence: “The middle student spends 4.00 hours a day on the phone, and the middle half spend between 2.70 and 5.75 hours; one student did not know.” With the 99 read as hours, software would print a mean of 9.17 over 20 students. (Exercise 25.)

26.

Variable	<i>n</i>	Summary
Arts & Humanities, <i>n</i> (%)	20	9 (45.0)
Lives in a mess, <i>n</i> (%)	20	8 (40.0)
Median CGPA band	19	3.00–3.49
Self-study, hours	20	median 2.75 (IQR 2.00 to 3.00)
Weekend sleep, hours	20	mean 8.12 (SD 1.22)
Screen time, hours	19	median 4.00 (IQR 2.70 to 5.75)

A fuller table would give every faculty and residence group. Note the *n* of 19 and 19 on two rows. (Exercise 26.)

Part E.

27. (a) Above 80%, the bars are 2 and 5 points tall, so the second looks 2.5 times as tall. (b) $85/82 = 1.037$: the second rate is under 4% higher. (c) Start the axis at zero; the two bars then look almost equal, which is the truth. (Exercise 27.)

28. The area grows by $3 \times 3 = 9$, so the icon looks nine times larger for a value three times larger. In three dimensions the volume grows by $3 \times 3 \times 3 = 27$. (Exercise 28.)

29. (a) $64/49 = 1.306$: up 30.6%, which rounds to 31%. The claim is arithmetically true. (b) From 55 to 54: down 1.8%, which is essentially flat. (c) “Down 14% in one month”, since $55/64 = 0.859$. It is just as true and just as misleading: both pick a window to suit a story. (Exercise 29.)

30. (a) A bar chart. (b) Box plots side by side, one per year. (c) A scatter plot. (d) Bars of row percentages, such as stacked 100% bars, with each group’s *n*. (e) A line chart. (Exercise 30.)

12 Our arithmetic and constructed numbers

Every number below was computed for these notes. None is printed by the papers cited, although the inputs from real papers are.

Value	How	Inputs from
240; 88; 9.2%; 40	the three checks on the “44 (8.4)” cell	Islam 2020, Tables 1 and 2
125, 52.1%; 131, 55.5%	moderate or worse, as row percentages; the grouping is ours	Islam 2020
Row percentages	Exercise 18, with the corrected cell	Islam 2020
0.384; 91, 49.7%; 165, 56.3%	Exercise 19	Islam 2020
0.82; 0.816, 0.817	Anscombe’s correlations, rounded; the means and fitted line	Anscombe 1973

Simulated, not real students: every figure from the class data, including all the tables and figures drawn from it, the 25-hour row, the three “don’t know” codes, and the 20-row

extract. They are generated by data/class-data.R from a fixed seed and will be replaced by the real survey.

Constructed for teaching, and not from any study: the two classes of Figure 5; the eighth student in Example 2.1; the semester of Example 2.3; the pictogram and the series of Figures 17 and 18; and every number in the exercises that is not attributed to a paper or to the class data.

13 References and further reading

1. Islam MA, Barna SD, Raihan H, Khan MNA, Hossain MT. Depression and anxiety among university students during the COVID-19 pandemic in Bangladesh: a web-based cross-sectional survey. *PLoS One* 2020;15(8):e0238162. PMID 32845928. PMC7449469. doi:10.1371/journal.pone.0238162. Licensed CC BY 4.0; free to read and share.
2. Anscombe FJ. Graphs in statistical analysis. *The American Statistician* 1973;27(1):17–21. doi:10.1080/00031305.1973.10478966. The four datasets are reproduced from the copy distributed with R, as `datasets::anscombe`. Subscription journal; it has no PMID.
3. Tukey JW. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley; 1977. The source of the box plot and the fences. A book; it has no DOI or PMID.
4. Class data: a **simulated** stand-in for the Stat.BD class survey, generated by data/class-data.R from the fixed seed 20261005. Not real students. No DOI or PMID.

14 For students in the Stat.BD course

Before the next session. Read this chapter, and work the Part A exercises (Exercises 1 to 6). Part D is the extract used in class; the groups that did not finish their variable can finish it from the solutions. The rest are for practice at any point in the course. Bring a calculator.

What the next session does with this chapter. Session 3 moves from the class that answered to the students who did not. The class averages 6.80 hours of weeknight sleep. If another class of 120 at this university filled in the same survey, would it average exactly 6.80 too? Certainly not, and Session 3 asks by how much it would differ. The mean, the SD and the two-SD rule of this chapter are the tools it builds on.

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

Question	Where it is settled
What is the curve behind the two-SD rule?	Chapter 3, the normal curve
How far could another class's average differ from ours?	Chapter 3, the standard error
Is a gap between two group means more than chance?	Chapter 4
Is the tuition gap between mess and hall students real?	Chapter 5, chi-square
How strong is the link between screen time and sleep?	Chapter 5, correlation
When is a total of Likert items worth averaging?	Chapter 6, survey scales
Producing these tables and charts in software	Session 7 onward