



CLASS NOTES • CHAPTER 5

Relationships between Variables

Cross-tables, correlation, and why an association is not a cause

Statistics & Data Analysis Fundamentals
Stat.BD

Mentor: Muhtasim Munif Fahim, MSc

CHAPTER 5

Relationships between Variables

Two variables can go together because of luck, a third variable, an arrow pointing the other way, or a real effect. This chapter measures how closely they go together, and shows why a test can rule out only the first.

In the simulated class, 74.4% of the students who live in a mess give private tuition. Among hall residents the figure is 36.2%, and among those living with their family it is 33.3%. Chapter 4 ended on the question this raises: is residence related to tuition, or could a pattern like this come from the luck of who was surveyed? None of Chapter 4's tests fits, because every one of them compared a number across groups. Here both variables are categories. The same logic still applies. Presume nothing is going on, then ask how surprising the data would be.

The chapter moves through three kinds of relationship. Section 1 relates two categorical variables with the chi-square test, and measures the strength of the relationship with Cramér's V and, for a 2×2 table, with odds and an odds ratio. Section 2 relates two numeric variables with the correlation coefficient r , and with its rank-based partner, Spearman's ρ . It also shows the ways a correlation can mislead. Section 3 asks what an association means, and why a survey like the class survey cannot show that one variable causes another. Section 4 runs the chapter's first test on a table from a published Bangladeshi survey, and section 5 gathers Chapters 2 to 5 into one map for choosing a test.

The data are the class survey of 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 lay out two categorical variables as a cross-table, and read it with row percentages.
2. What independence predicts for each cell, and how the chi-square statistic measures the distance between the observed and the expected counts.
3. How degrees of freedom set the reference curve, and where the p -value comes from.
4. How residuals show which cells drive an association, and how Cramér's V measures its strength.
5. When the chi-square test is valid, and when Fisher's exact test should replace it.
6. Four ways to describe a 2×2 table: a difference, a ratio, odds and an odds ratio.
7. What the correlation coefficient r measures, how to compute it by hand, how to test it and how to give it an interval.

8. Why r^2 is a more honest guide to strength than r , and why a significant r need not be a strong one.
9. Five ways a correlation misleads: curves, outliers, codes, a narrow range and many tests. And why correlation is not agreement.
10. When to use Spearman's ρ .
11. The four explanations of an association, and the tools that separate them: splitting by a third variable, randomisation and Hill's viewpoints.
12. How to choose a test from the kind of question and the kind of variables.

Notation

O, E	an observed count in a cell, and the count independence predicts
χ^2	the chi-square statistic ("kai-square")
df	degrees of freedom
N	the number of people in a table
k	the smaller of the number of rows and the number of columns
V, ϕ	Cramér's V ; for a 2×2 table it is called ϕ ("phi")
odds, OR	the odds of an outcome, and the odds ratio of two groups
$x, y; \bar{x}, \bar{y}$	two numeric variables, and their means
r	Pearson's correlation coefficient, between -1 and $+1$
r^2	the share of the variation in one variable that lines up with the other
ρ	Spearman's rank correlation ("rho")
d	in Spearman's shortcut, the difference between a person's two ranks
z	Fisher's transformation of r
t, p, SE	as in Chapter 4

Contents

What this chapter establishes	1
Notation	2
1 Two categorical variables: the chi-square test	5
1.1 A cross-table, and the question it raises	5
1.2 What independence predicts: expected counts	6
1.3 The chi-square statistic	6
1.4 Degrees of freedom and the chi-square curve	7
1.5 The p -value, and a shuffle that checks it	8
1.6 Where the association lies: residuals	9
1.7 How strong? Cramér's V	10
1.8 Chi-square grows with the sample; V does not	10
1.9 When the test is valid, and Fisher's exact test	11
1.10 The 2×2 table: juniors, seniors and tuition	11
1.11 Reporting a chi-square test	12
2 Two numeric variables: correlation	13
2.1 The scatterplot	13
2.2 What r measures	13
2.3 Where r comes from	13
2.4 Testing r	15
2.5 r^2 : how much of the variation lines up	16
2.6 Significant is not strong	16
2.7 An interval for r : Fisher's z	16
2.8 Straight lines only	17
2.9 One wild value: the "don't know" codes	18
2.10 Spearman's ρ : correlation on ranks	18
2.11 A narrow range weakens r	20
2.12 Many correlations at once	20
2.13 Correlation is not agreement	21
2.14 Reporting a correlation	21

3	Association is not causation	21
3.1	Four reasons two things go together	21
3.2	A common cause: confounding	21
3.3	Countries are not people: the ecological fallacy	22
3.4	Testing a candidate: split by it	23
3.5	Which way does the arrow point?	24
3.6	Berkeley, 1973: a gap that reverses	24
3.7	Randomisation	25
3.8	When randomisation is impossible: Hill's viewpoints	25
3.9	Words that claim too much	26
4	A cross-table in a published paper	26
4.1	Islam 2020, Table 2: anxiety by private tuition	26
4.2	Table 2 against Table 3	27
4.3	A row that does not add up	27
5	Which test to use	28
5.1	Two questions first	28
5.2	The flowchart	28
5.3	Three questions through the chart	29
5.4	Five wrong choices	29
6	Analysing a relationship: the procedure	30
7	Chapter summary	31
8	Key terms	32
9	Exercises	34
10	Solutions	37
11	Our arithmetic and constructed numbers	41
12	References and further reading	42
13	For students in the Stat.BD course	43

1 Two categorical variables: the chi-square test

1.1 A cross-table, and the question it raises

Chapter 2 introduced the cross-table, which counts the people in each combination of two categorical variables. Table 1 is the cross-table of residence by private tuition for the simulated class. Each row is one residence, and the last column turns that row's counts into a percentage.

Cross-table

a count of people in each combination of two categorical variables; also called a contingency table

Table 1: Residence by private tuition. The row percentage is the share of each residence that gives tuition. **Simulated** class data.

	No tuition	Tuition	Total	% tuition
Hall	30	17	47	36.2
Mess	11	32	43	74.4
Family home	20	10	30	33.3
Total	61	59	120	49.2

Figure 1 draws the last column. The dashed line is the whole class: 49.2% give tuition. If residence made no difference to tuition, each bar would sit near that line, give or take some luck. About three in four mess students give tuition, against about one in three in the other two residences.

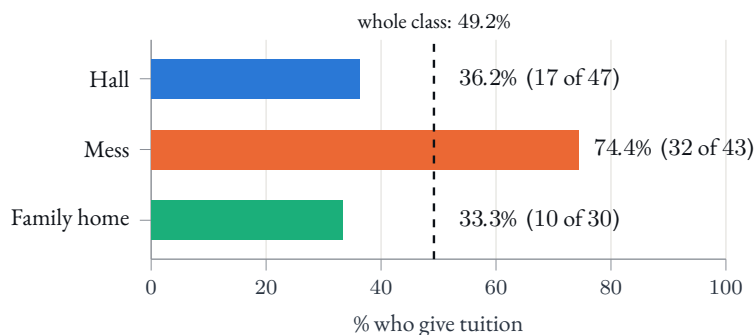


Figure 1: The share of each residence that gives private tuition. The dashed line is the whole class. **Simulated** class data.

The pattern raises the question of every test. Could a gap this size appear by luck in a class of 120, if residence and tuition were unrelated among students like these? A categorical variable has no mean, so the t -test and ANOVA of Chapter 4 cannot answer it. The chi-square test can.

Categorical variable

a variable whose values are categories, such as a residence or a yes or no answer

Which percentage?

Chapter 2 warned that one table has three sets of percentages: of each row, of each column and of the whole table. For “does tuition depend on residence?”, compare the rows, as Table 1 does. Section 4.1 shows a published paper that printed the wrong set.

1.2 What independence predicts: expected counts

Two variables are independent if knowing one tells nothing about the other. If residence and tuition were independent, every residence would give tuition at the whole-class rate of 49.2%. The hall row has 47 students, so independence predicts 49.2% of 47 tutors there, about 23.1. Chapter 3 found the same figure. The general rule gives the *expected count* of any cell:

$$E = \frac{\text{row total} \times \text{column total}}{N} \quad (1.1)$$

E the count expected in a cell if the two variables are independent

N the number of people in the table

Independence

two variables are independent if knowing one tells nothing about the other

Example 1.1 • Expected count: mess and tuition

The mess row has 43 students and the tuition column has 59, in a table of $N = 120$:

$$E = \frac{43 \times 59}{120} = 21.14.$$

Independence predicts about 21.1 mess tutors. The table has 32.

Table 2 gives all six expected counts. They keep every row total and every column total of Table 1, up to rounding. It is the table that would arise if tuition were spread across the three residences in exactly the whole-class proportion.

Table 2: Expected counts if residence and tuition were independent, from equation (1.1), rounded to one decimal. Our arithmetic from the **simulated** class data.

	No tuition	Tuition	Total
Hall	23.9	23.1	47
Mess	21.9	21.1	43
Family home	15.2	14.8	30
Total	61	59	120

An expected count need not be a whole number

No residence can hold 23.1 tutors. An expected count is an average over many imaginary classes in which the null hypothesis is true, not a number of students. Do not round it to a whole number before using it.

1.3 The chi-square statistic

Chapter 4's tests measured how far the data sat from the null hypothesis in a single number. For a table, that number is built from the six gaps between observed and expected counts. Each gap is squared, so that a shortfall counts as much as an excess. It is then divided by the expected count, because a gap of 10 means more in a cell that expected 20 than in one that expected 1,000. The sum over every cell is the chi-square statistic:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad (1.2)$$

Chi-square statistic

the sum over all cells of $(O - E)^2 / E$



- O the observed count in a cell
 E its expected count, equation (1.1)
 Σ add the result over every cell of the table

Example 1.2 • One cell of the chi-square: mess and tuition

From Example 1.1, $O = 32$ and $E = 21.14$.

Step	What	Working
1	the gap, squared	$(32 - 21.14)^2 = 10.86^2 = 117.94$
2	scaled by E	$117.94 \div 21.14 = 5.58$

This one cell contributes 5.58 to χ^2 .

Table 3 repeats the step for every cell. The two mess cells together give about 64% of the total: nearly two thirds of the evidence comes from one row.

Table 3: The chi-square statistic, cell by cell. **Simulated** class data.

Cell	O	E	$(O - E)^2/E$
Hall, no tuition	30	23.9	1.56
Hall, tuition	17	23.1	1.61
Mess, no tuition	11	21.9	5.39
Mess, tuition	32	21.1	5.58
Family home, no tuition	20	15.2	1.48
Family home, tuition	10	14.8	1.53
Sum: χ^2			17.16

The contributions are worked from unrounded expected counts. Rounded to two decimals they add to 17.15; the total 17.16 is from the unrounded values.

The same shape as every test

A chi-square is zero when every observed count equals its expected count, and it grows as the table moves away from independence. Like the t of Chapter 4, it measures a distance from the null hypothesis. Unlike t , it cannot be negative, because every gap is squared.

1.4 Degrees of freedom and the chi-square curve

A chi-square of 17.16 is a distance. To turn it into a p -value, the test needs to know how large a chi-square tends to be when the null hypothesis is true. That depends on the size of the table, through its degrees of freedom. Once the row and column totals are fixed, only some cells are free to vary. In Table 1, choose the number of hall tutors and the hall row's other cell follows from its total. Choose the mess tutors as well, and the family-home tutors follow from the column total. Two free choices fill the whole table:

$$\text{df} = (\text{rows} - 1) \times (\text{columns} - 1) = (3 - 1) \times (2 - 1) = 2 \quad (1.3)$$

Each number of degrees of freedom has its own chi-square curve (Figure 2). With more cells,

more gaps are added together, so the curve moves to the right. Every curve starts at zero and has a long right tail.

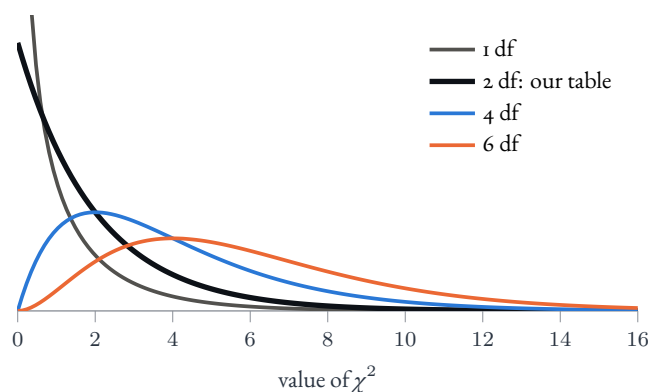


Figure 2: Chi-square curves for different degrees of freedom. A table with more cells gives a larger chi-square even when nothing is going on.

Table 4: The chi-square value that cuts off the top 5% of each curve: the critical value at $\alpha = 0.05$. From standard tables.

df	1	2	3	4
Critical value	3.84	5.99	7.81	9.49

1.5 The p -value, and a shuffle that checks it

The chi-square test is always one-tailed on the curve. Any departure from independence, in any direction, makes χ^2 larger, so only the right tail counts as evidence (Figure 3). With 2 df, only 5% of tables with no relationship give a chi-square above 5.99. The class gives 17.16, about 2.9 times that critical value, and $p < 0.001$. At $\alpha = 0.05$, independence is rejected: in this class, residence and tuition are associated.

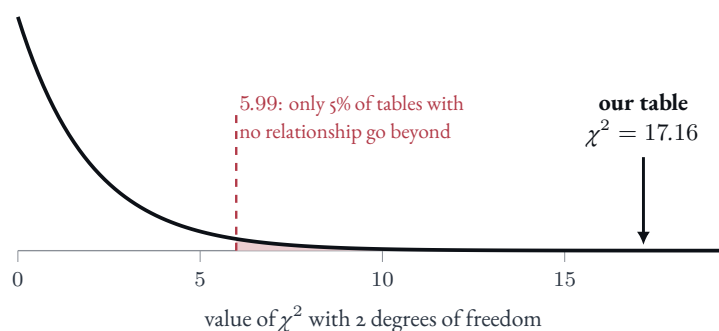


Figure 3: The chi-square curve with 2 df. The dashed line is the 5% critical value; the arrow marks the class's χ^2 . **Simulated** class data.

The curve is a mathematical shortcut, and Chapter 4 showed a way to check such a shortcut by brute force. Shuffle the 120 tuition answers at random across the students. Residence and tuition are now unrelated by construction, but every row and column total is unchanged. Compute χ^2 for the shuffled class, and repeat 1,000 times. Figure 4 shows the result. The

bars follow the curve. 41 of the 1,000 shuffled tables reach 5.99, close to the 5% the curve predicts. The largest of all 1,000 is 13.71, and none reaches the class's 17.16.

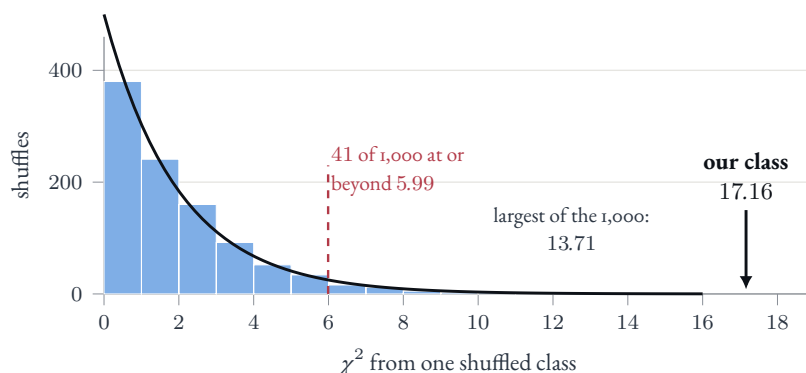


Figure 4: Chi-square for 1,000 random shuffles of the tuition answers. The curve is the 2 df chi-square curve, scaled to 1,000 tables. **Simulated** class data; the shuffles are ours.

1.6 Where the association lies: residuals

A chi-square says *whether* a table departs from independence. It does not say where, or in which direction. For that, each cell gets a residual:

$$\text{residual} = \frac{O - E}{\sqrt{E}} \quad (1.4)$$

Squaring a residual gives that cell's contribution to χ^2 , but the residual keeps the sign. A positive residual means more people in the cell than independence predicts; a negative one, fewer. A residual beyond about ± 2 marks a cell worth talking about.

Residual

a cell's gap between observed and expected count, in standard units; beyond about ± 2 is worth attention

Example 1.3 • A residual: hall and tuition

The hall, tuition cell has $O = 17$ and $E = 23.11$. Then $O - E = -6.11$ and $\sqrt{E} = 4.81$, so the residual is $-6.11 \div 4.81 = -1.27$. Fewer hall students tutor than independence predicts, but not by enough to stand out.

Table 5: Residuals for residence by tuition, equation (1.4). **Simulated** class data.

	No tuition	Tuition
Hall	+1.25	-1.27
Mess	-2.32	+2.36
Family home	+1.22	-1.24

Table 5 tells the story in one line. Only the mess row passes ± 2 : more mess students tutor than independence predicts, and fewer do not. Hall and family home sit a little on the other side, and together.

The two-step habit

Run the chi-square test to ask *whether*. Then read the residuals, or the row percentages, to say *where* and *in which direction*. A report that gives only “ $p < 0.001$ ” has not said what the association is.

1.7 How strong? Cramér’s V

Chapter 4 separated “is it real?” from “is it big?”. The p -value answers the first. For a cross-table, Cramér’s V answers the second:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}} \quad (1.5)$$

k the smaller of the number of rows and the number of columns

V runs from 0, when the rows are identical, to 1, when knowing the row fixes the column completely. For Table 1, $k = 2$, so

$$V = \sqrt{\frac{17.16}{120 \times 1}} = 0.378.$$

Cohen’s book of 1988 is widely quoted as calling 0.1 small, 0.3 medium and 0.5 large, for a measure that equals V when $k = 2$ (reference 8). On that guide, 0.38 sits between medium and large. Those benchmarks were not checked against the book for these notes, and Cohen offered them as a rough guide, not as rules. What counts as strong depends on the field.

1.8 Chi-square grows with the sample; V does not

Keep every percentage in Table 1 and change only the number of students. Table 6 does this as a thought experiment. A tenth of the class, with the same percentages, gives a chi-square nowhere near significant. Ten times the class gives an enormous one. The pattern is identical in all three, and V says so.

Table 6: The same percentages at three class sizes. Every count is divided or multiplied by 10; counts divided by 10 are not whole numbers, which is the point of a thought experiment. Our arithmetic from the **simulated** class data.

Every count	χ^2	p	V
divided by 10	1.72	0.424	0.378
as observed	17.16	< 0.001	0.378
times 10	171.56	< 0.001	0.378

χ^2 is proportional to N . This is Chapter 4’s lesson for a new test: a large enough sample makes any pattern significant, however weak. Report V beside p .

Cramér’s V

a measure of the strength of association in a cross-table, from 0 (none) to 1 (complete)

Never run a chi-square on percentages

Feed the row percentages of Table 1 into a chi-square calculator as if they were counts, and χ^2 becomes 42.16 instead of 17.16. Three rows of percentages add to 300, so the test believes it has 300 students, in three rows of equal size. The test needs counts.

1.9 When the test is valid, and Fisher's exact test

The chi-square test has three conditions.

1. **Counts**, not percentages, means or scores.
2. **Each person counted once**, in one cell. The same students asked before and after something are paired data, and need a different test.
3. **Expected counts large enough**. The usual rule is that every expected count should be at least 5. Textbooks differ on its exact form; it is a convention.

The smallest expected count in Table 2 is 14.8, so all three hold.

The third condition is the one people misremember. It concerns the *expected* counts, not the observed ones. An observed count of 2 is fine if its expected count is 8. The chi-square curve is an approximation, and it is a poor one when expected counts are small. For a 2×2 table with small expected counts, Fisher's exact test computes the p -value directly from every table with the same totals, without the curve. Software offers it automatically. For a larger table, categories that belong together can be merged. That decision must be taken before looking at the result, or it becomes the p -hacking of Chapter 4.

Fisher's exact test

a test for a cross-table that computes the p -value directly, without the chi-square curve; used when expected counts are small

1.10 The 2×2 table: juniors, seniors and tuition

A table with two rows and two columns is the most common cross-table in health and social research. Table 7 splits the simulated class into juniors, in years 1 and 2, and seniors, from year 3 to the master's year, against tuition.

Table 7: Year of study, in two groups, by private tuition. **Simulated** class data.

	No tuition	Tuition	Total	% tuition
Years 1–2	35	17	52	32.7
Years 3 to Masters	26	42	68	61.8
Total	61	59	120	49.2

A 2×2 table has $(2 - 1)(2 - 1) = 1$ degree of freedom, and a critical value of 3.84. The test gives $\chi^2 = 9.96$ and $p = 0.002$. Two variants are often printed beside it. *Yates' correction*, a small adjustment for 2×2 tables that some software applies by default, gives $p = 0.003$. Fisher's exact test gives $p = 0.002$. With expected counts this large, the smallest being 25.6, the three agree. They part company only when counts are small. A report should say which was used.

A 2×2 table can describe the size of the association in four ways (Table 8). The first two compare the percentages. The last two compare odds: the number who tutor divided by the number who do not. Among juniors, the odds are $17 \div 35 = 0.49$, about one tutor for every two who do not. Among seniors they are $42 \div 26 = 1.62$. The odds ratio divides one by the other.

Odds and odds ratio

odds: the number with an outcome divided by the number without it; odds ratio: one group's odds divided by another's, where 1 means no association

Table 8: Four ways to say how big the association in Table 7 is, seniors against juniors. **Simulated** class data; the odds ratio is from the unrounded odds.

Measure	How	Value
Difference in percentages	$61.8 - 32.7$	29.1 points
Ratio of percentages	$61.8 \div 32.7$	1.89
Odds ratio	$1.62 \div 0.49$	3.33
ϕ (Cramér's V)	equation (1.5) with $k = 2$	0.288

Say which ratio

Seniors are nearly twice as likely to tutor: the ratio of percentages is 1.89. Their odds are over three times the juniors': the odds ratio is 3.33. Both describe the same table. A headline saying seniors are “three times as likely” to tutor has confused the two. The odds ratio is always further from 1 than the ratio of percentages, and the gap is large when the outcome is common. Chapter 6 returns to odds ratios with logistic regression.

1.11 Reporting a chi-square test

One sentence can carry everything a reader needs:

Tuition was more common among mess residents (74.4%) than among hall (36.2%) or family-home (33.3%) residents, $\chi^2(2, N = 120) = 17.16, p < 0.001$, Cramér's $V = 0.38$.

The percentages show the pattern. The degrees of freedom and N let a reader check the test, and V gives the strength as well as the significance. It says “was more common among”, not “is caused by”. Section 3 explains why.

2 Two numeric variables: correlation

2.1 The scatterplot

When both variables are numbers, the first step is always a picture. Chapter 2 drew the scatterplot, with one dot per person. Figure 5 plots daily screen time against weeknight sleep for the 117 students with a usable answer to both. The three “don’t know” codes of Chapter 2 are left out; section 2.9 shows what they do if left in.

Scatterplot

a plot with one dot per person, placed by their values on two numeric variables

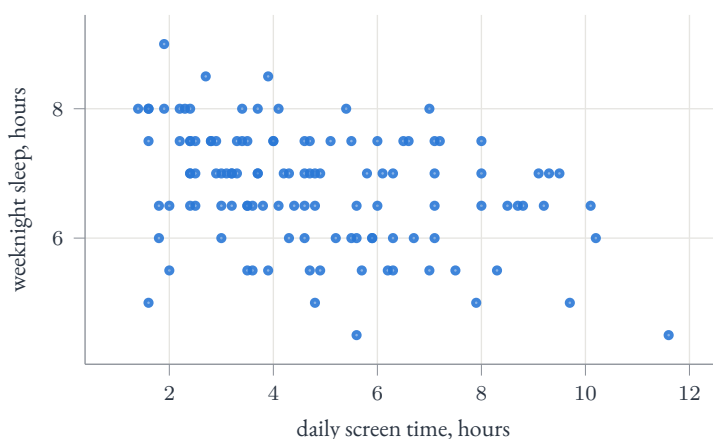


Figure 5: Daily screen time and weeknight sleep, one dot per student. **Simulated** class data.

The cloud drifts downwards from left to right: students with more screen time tend to sleep less. But it is a loose drift, not a line, and at every level of screen time some students sleep a lot and some very little. The correlation coefficient puts a number on how loose.

2.2 What r measures

Definition 2.1 • Correlation coefficient

Pearson’s correlation coefficient r measures how closely two numeric variables follow a straight line. It runs from -1 to $+1$. Its *sign* gives the direction: positive when one rises with the other, negative when one falls as the other rises. Its *size* gives the tightness: ± 1 is a perfect line, and 0 means no straight-line trend at all.

Figure 6 shows six clouds of 60 points each, built so that r takes exactly the printed value. The class, at $r = -0.36$, looks like something between the second and third panels. A correlation has no units. Measuring screen time in minutes instead of hours leaves r unchanged. It is also symmetric: the correlation of sleep with screen time is the same as that of screen time with sleep.

2.3 Where r comes from

Cut the scatterplot at the two means, 4.81 hours of screen time and 6.77 hours of sleep (Figure 7). For each student, multiply their two deviations from the means. A student above

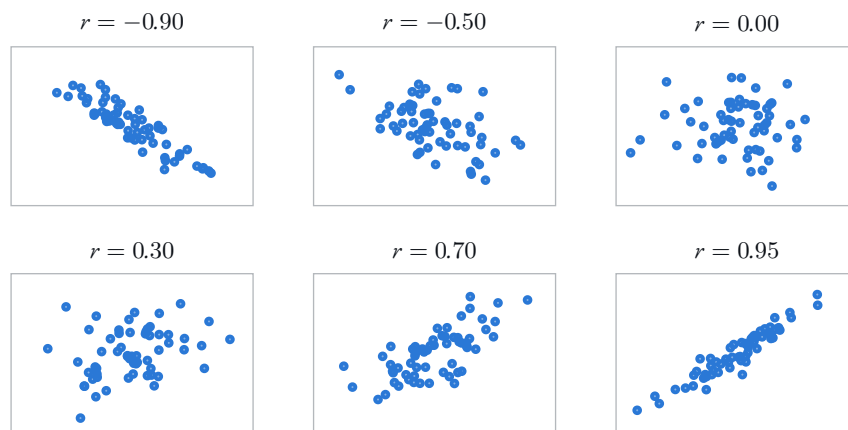


Figure 6: Six correlations. Each cloud is constructed, with 60 points, so that r is exactly the printed value.

both means, or below both, gives a positive product. A student above one and below the other gives a negative product. In the class, 44 students give positive products and 73 give negative ones. The negatives win, so r is negative.

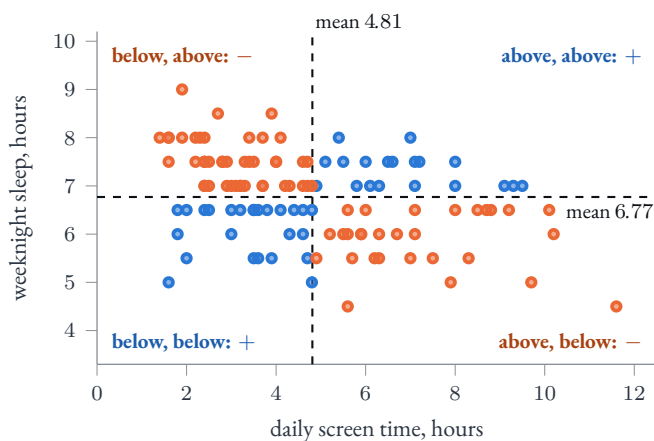


Figure 7: The scatterplot cut at the two means. Blue students sit on the same side of both means and give positive products; orange students give negative ones. **Simulated** class data.

The formula adds up the products and rescales the total so that it lies between -1 and $+1$:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \times \sum (y - \bar{y})^2}} \quad (2.1)$$

x, y one person's values on the two variables

$\bar{x}, \bar{y}$ the two means

top the sum of the products of deviations

bottom a rescaling that keeps r between -1 and $+1$

Example 2.1 • r by hand for six students

The first six students in the file have these values, in hours:

Student	1	2	3	4	5	6	Sum
Screen x	3.1	5.5	1.8	3.7	6.0	1.6	
Sleep y	7.0	7.5	6.5	8.0	7.5	7.5	
$x - \bar{x}$	-0.52	1.88	-1.82	0.08	2.38	-2.02	
$y - \bar{y}$	-0.33	0.17	-0.83	0.67	0.17	0.17	
Product	0.17	0.31	1.51	0.06	0.40	-0.34	2.12

The means are $\bar{x} = 3.62$ and $\bar{y} = 7.33$. The squared deviations add to $\sum(x - \bar{x})^2 = 16.87$ and $\sum(y - \bar{y})^2 = 1.33$. Then

$$r = \frac{2.12}{\sqrt{16.87 \times 1.33}} = \frac{2.12}{4.74} = +0.45.$$

The rounded products add to 2.11; the sum 2.12 is from unrounded values.

Student 3 contributes most: well below average on both variables, which gives a large positive product. Student 6 gives the only negative product. And the answer is positive. These six students say that more screen time goes with more sleep, the opposite of the whole class.

2.4 Testing r

Which to trust, six students or 117? A test of r asks the question of Chapter 4. The null hypothesis is that there is no straight-line relationship in the population. The test statistic has the familiar shape, an estimate measured against its own uncertainty:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \quad \text{df} = n - 2 \quad (2.2)$$

n the number of people with both values

For the class, $r = -0.36$ and $n = 117$, which give $t = -4.19$ on 115 df and $p < 0.001$. For six students, r would have to pass ± 0.81 to reach $p < 0.05$, and their $+0.45$ is well inside the range that luck produces. Table 9 gives the smallest r that reaches significance at several sample sizes.

Table 9: The smallest $|r|$ with $p < 0.05$, two-sided. Our arithmetic from the t curve with $n - 2$ df.

n	6	10	30	117	1,000
critical $ r $	0.81	0.63	0.36	0.18	0.06

A small sample can point the wrong way

Six students are not only less precise than 117. Here they point in the opposite direction. This is the sampling variation of Chapter 3: small samples scatter widely around the truth, and a scatter that wide includes the wrong sign.



2.5 r^2 : how much of the variation lines up

Squaring r gives a more honest sense of strength. For the class, $(-0.364)^2 = 0.132$. About 13.2% of the student-to-student variation in weeknight sleep lines up with screen time along a straight line. The other 87% is everything else: stress, roommates, prayer times, a baby at home. (Squaring the rounded -0.36 would give 13.0%.)

r	± 0.1	± 0.3	± 0.5	± 0.7	± 0.9
r^2	1%	9%	25%	49%	81%

Why r flatters

An r of 0.5 sounds like half. It lines up a quarter of the variation.

The phrase “lines up with” is deliberate. Many books say “explained by”, which sounds causal, and section 3 shows why a survey cannot support that. In Chapter 6, r^2 returns as the R^2 of a regression.

2.6 Significant is not strong

Table 9 holds a warning. With 1,000 people, an r of 0.06 is significant. With six, it takes 0.81. The p -value of a correlation mostly reflects the sample size. Compare two constructed studies:

- $r = 0.10$ with $n = 1,000$: $p = 0.002$. Significant, and it lines up 1.0% of the variation.
- $r = 0.50$ with $n = 10$: $p = 0.141$. Not significant, and possibly strong. There are too few people to tell.

The first found something real and tiny, the second something possibly large and unproven. Neither p -value alone says so. This is Chapter 4’s lesson for correlations: p answers “could this be zero?”, not “is this large?”. Report r and its interval.

2.7 An interval for r : Fisher’s z

An r cannot pass -1 or $+1$, so its sampling spread is lopsided near the ends, and an interval of the form “ $r \pm$ something” could run past them. Fisher’s z stretches the scale, builds the interval there, and brings it back:

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}, \quad \text{SE} = \frac{1}{\sqrt{n-3}}, \quad r = \frac{e^{2z} - 1}{e^{2z} + 1} \quad (2.3)$$

Fisher’s z

a transformation of r that makes its sampling spread symmetric, used to build an interval

Example 2.2 • A 95% interval for the class correlation

Step	What	Working
1	stretch	$z = \frac{1}{2} \ln \frac{1+r}{1-r} = -0.381$
2	standard error	$1 \div \sqrt{114} = 0.094$
3	interval for z	$-0.381 \pm 1.96 \times 0.094$, from -0.565 to -0.198
4	bring back	r from -0.51 to -0.20

Software gives the same interval, -0.51 to -0.20 . It lies entirely below zero.

The interval is not symmetric: -0.36 sits 0.15 from the lower end and 0.16 from the upper. The closer r is to ± 1 , the more lopsided the interval becomes. Software does all four steps; the point to keep is why the interval is not “ $r \pm$ something”.

2.8 Straight lines only

A correlation measures a straight-line trend and nothing else. Figure 8 shows a strong curve, a U, with an r of almost zero, because the rising half cancels the falling half. Too little and too much can both be bad; stress and performance are often described this way.

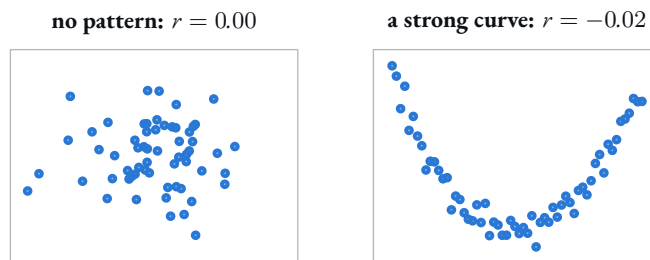


Figure 8: Two constructed clouds with r near zero. On the left there is no pattern. On the right there is a strong pattern that is not a straight line.

In 1973 the statistician Francis Anscombe made the point unforgettable (reference 3). He built four small datasets with, to two or three decimals, the same means, the same variances, the same correlation and the same fitted line (Table 10). Then he drew them (Figure 9).

Table 10: Anscombe’s quartet: the statistics of the four datasets. Computed by us from R’s copy of the 1973 data. The line is $y = a + bx$, fitted by least squares (Chapter 6).

	I	II	III	IV
mean of x	9.00	9.00	9.00	9.00
mean of y	7.50	7.50	7.50	7.50
variance of x	11.00	11.00	11.00	11.00
variance of y	4.13	4.13	4.12	4.12
r	0.816	0.816	0.816	0.817
line: a	3.00	3.00	3.00	3.00
line: b	0.500	0.500	0.500	0.500

Only dataset I looks like the cloud that r describes. Dataset II is a smooth curve. Dataset III is a perfect line spoiled by one point. In dataset IV, the x values are all the same except for

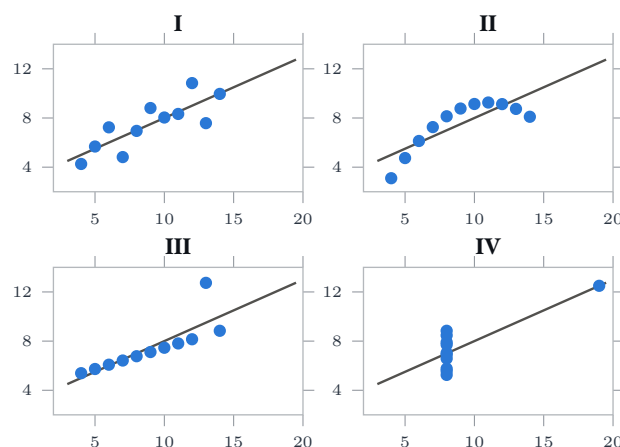


Figure 9: Anscombe's quartet. Four datasets with the statistics of Table 10. The grey line is the same in every panel. Data from R's `datasets::anscombe`.

one point far to the right, and that single point creates the whole correlation. The remedy is not a better formula. It is to plot the data before computing anything.

2.9 One wild value: the “don't know” codes

Chapter 2 cleaned three screen-time answers coded 99, the code for “don't know”. Put them back as if they were hours, and Figure 10 shows what happens. Three students now appear to use their phones 99 hours a day. They sit so far from the mean that their products dominate the sum in equation (2.1), and r changes from -0.36 to $+0.15$. Three points out of 120 reverse the sign. Anyone who ran the correlation without looking at the plot would report the opposite of the truth.

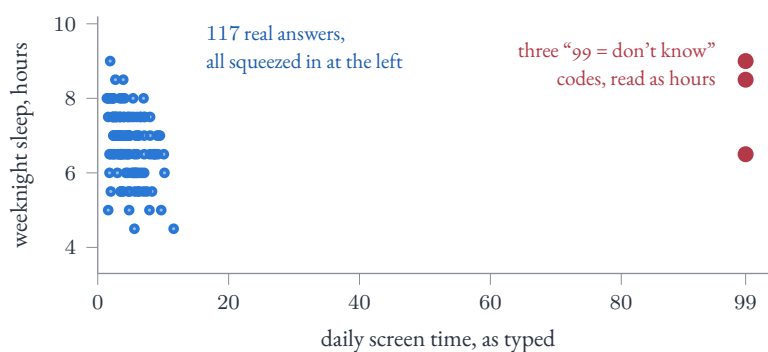


Figure 10: Weeknight sleep against screen time with the three 99 codes left in as hours. **Simulated** class data.

Spearman's ρ , the subject of the next section, moves only from -0.36 to -0.30 . A value of 99 is simply the largest value; as a rank, it cannot be further from the rest than any other top value would be.

2.10 Spearman's ρ : correlation on ranks

Chapter 4 met ranks as the way round skewed data. Spearman's ρ is the same idea for relationships: replace each variable by its ranks, smallest first, then compute an ordinary r on

Spearman's rho

Pearson's r computed on the ranks of the two variables



the ranks. It asks a slightly different question from Pearson's. Not “how close to a straight line?”, but “how consistently does one go up with the other?”.

It is the right choice in three situations. A variable is ordinal, such as a CGPA band. A relationship is curved but always in one direction. Or a few extreme values would dominate Pearson's r . Figure 11 shows the first. CGPA band has four ordered values, not a measured number, so Spearman is the natural measure: $\rho = 0.46$ for 105 students, $p < 0.001$. Pearson's r on the band codes would be 0.42.

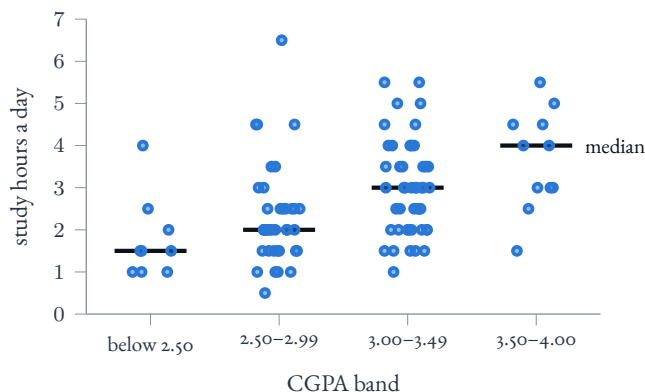


Figure 11: Daily self-study hours by CGPA band. Dots are shifted sideways at random so that they do not overlap; the bars mark each band's median. **Simulated** class data.

Example 2.3 • Spearman's ρ for the six students

Rank each variable from 1 for the smallest. Tied values share the average of the ranks they would have taken, as in Chapter 4.

Student	1	2	3	4	5	6
Screen x	3.1	5.5	1.8	3.7	6.0	1.6
rank	3	5	2	4	6	1
Sleep y	7.0	7.5	6.5	8.0	7.5	7.5
rank	2	4	1	6	4	4

Equation (2.1) applied to the two rows of ranks gives $\rho = 0.39$.

Many textbooks give a shortcut:

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \quad (2.4)$$

d the difference between a person's two ranks

It is exact only when there are no ties. With the tied sleep values of Example 2.3 it gives 0.43 instead of 0.39. Software uses the exact method, r on the ranks.

Robust to a typo

Chapter 2 found a study-hours answer of 2.5 typed as 25. Put the typo back, and Pearson's r between study hours and CGPA band drops from 0.42 to 0.15. Spearman's ρ

moves only from 0.46 to 0.44. As a rank, 25 is just “the largest”.

2.11 A narrow range weakens r

Keep only the middle half of the class by screen time, from 3.0 to 6.3 hours a day: 62 students. The correlation with sleep falls from -0.36 to -0.27 (Figure 12). The students are the same and so is the relationship. With less spread in screen time, there is less for r to see.

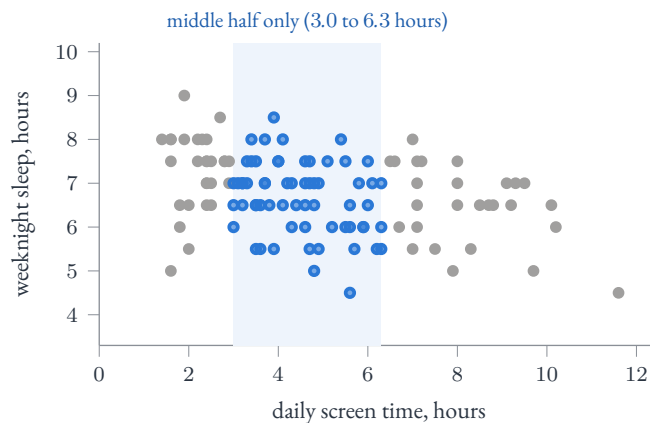


Figure 12: The middle half of screen time (dark) against the rest (pale). Within the middle half, the correlation is weaker. **Simulated** class data.

This is called restriction of range. Admission scores, for example, often correlate only weakly with later grades *among those admitted*, because everyone admitted scored high. The example is general; nothing is claimed about any particular university. A study of heavy users only, or of one department only, can report a weaker correlation for this reason alone.

Restriction of range
a weaker correlation caused by studying only a narrow slice of one variable

2.12 Many correlations at once

Papers often show many correlations together in a correlation matrix (Table 11). Only the lower triangle is needed; the upper half would repeat it.

Correlation matrix
a table of the correlations between every pair in a set of variables

Table 11: Correlations between four variables. Sleep is weeknight sleep; each r uses the students who answered both questions. **Simulated** class data.

	Screen	Sleep	Stress	Study
Screen	1			
Sleep	-0.36	1		
Stress	0.45	-0.48	1	
Study	-0.03	0.08	-0.08	1

Four variables give six pairs, and three of them pass the critical value of about 0.18. Chapter 4’s warning about many tests applies. If none of the six pairs were related, the chance that at least one would reach $p < 0.05$ anyway is $1 - 0.95^6 = 26.5\%$, treating the six tests as independent. A Bonferroni correction tests each at $0.05/6 = 0.0083$, and the same three pairs still pass. The better defence is Chapter 4’s: decide before looking which correlations answer the question.

Stress sits in the middle of the matrix. It goes with more screen time and with less sleep, which makes it a candidate for a third variable behind the link between them. Section 3.4 takes that up.

2.13 Correlation is not agreement

Weekend and weeknight sleep, for the same students, correlate at $r = 0.74$. Students who sleep long on weeknights also tend to sleep long at weekends. But the two answers do not *agree*. The mean is 7.87 hours at weekends against 6.81 on weeknights, a paired difference of 1.06 hours (Chapter 4). A high r says the students keep roughly their order. It says nothing about whether the two numbers match.

Two thermometers

One thermometer always reads exactly one degree higher than another. Their correlation is $+1$, and they never agree. To compare two ways of measuring the same thing, such as a new questionnaire against an old one, look at the differences, not at r . Bland and Altman made this point for clinical measurement in 1986 (reference 7).

2.14 Reporting a correlation

Among 117 students, daily screen time was negatively correlated with weeknight sleep (Pearson $r = -0.36$, 95% CI -0.51 to -0.20 , $p < 0.001$).

With it, show the scatterplot, say whether the coefficient is Pearson's or Spearman's and why, and give n , because people skip questions. Avoid "screen time *reduces* sleep", "a significant correlation" with no r , and any r from data that have not been plotted.

3 Association is not causation

3.1 Four reasons two things go together

Sections 1 and 2 found two associations: residence with tuition, and screen time with sleep. Both passed their tests. Figure 13 lists the four explanations any association has. A test deals only with the first, chance, and only with some probability.

For screen time and sleep, all three remaining explanations are plausible. Stress could raise screen time and cut sleep, a *common cause*. Students who cannot sleep could pick up the phone, *reverse causation*. Or the phone could keep people awake, a *real effect*. The survey cannot tell these apart. The rest of this section is about how anyone ever can.

3.2 A common cause: confounding

In 2012 the cardiologist Franz Messerli published a short note in the *New England Journal of Medicine* (reference 4). He compared countries: chocolate eaten per person against Nobel laureates per person. Countries that ate more chocolate had more laureates. The full text

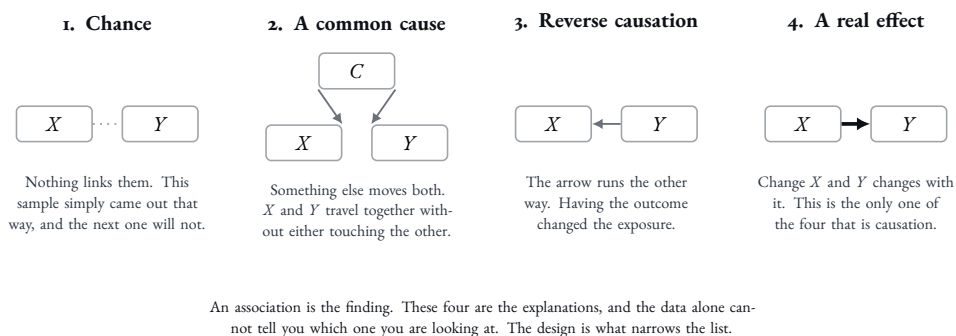


Figure 13: Four reasons two variables can be associated. A significance test addresses only the first.

was not checked for these notes, so no figures from it are quoted. The shape of the argument is what matters.



Figure 14: A common cause. National wealth is our suggestion, not the paper's: it could raise both without either touching the other.

Nobody sensible concludes that chocolate makes Nobel laureates. Richer countries can afford more chocolate, and they also fund more universities and research. Wealth is a candidate *confounder*: a third variable, linked with both, that creates or distorts the association between them (Figure 14). Chapter 6 shows how regression holds a measured confounder fixed.

Confounder

a third variable linked with both variables of interest, which creates or distorts the association between them

3.3 Countries are not people: the ecological fallacy

The chocolate data describe countries. Even if the country pattern were real, it would say nothing about whether the laureates themselves eat more chocolate. Reading individual behaviour off group averages is the *ecological fallacy*.

The class shows the trap in miniature. Table 12 gives only the three residence averages. Family home has the most screen time *and* the most sleep. Anyone who saw only this table might conclude that screen time goes with more sleep.

Ecological fallacy

reading a pattern among individuals off a pattern among group averages

Table 12: Average screen time and weeknight sleep in each residence, in hours, among students who answered both questions. **Simulated** class data.

	n	Mean screen	Mean sleep
Hall	45	4.64	6.71
Mess	42	4.85	6.48
Family home	30	5.00	7.27

Three averages are far too few to compute anything from, and the next subsection shows that inside every residence the pattern runs the other way. A district with more mobile phones may have longer life expectancy; that says nothing about whether the people with phones live longer.

3.4 Testing a candidate: split by it

Is residence the common cause behind screen time and sleep? A candidate that takes a few values can be tested with the data already collected: split by it. Inside one residence, residence cannot vary, so it cannot be driving anything. If it were the whole explanation, the link would fade inside each group.

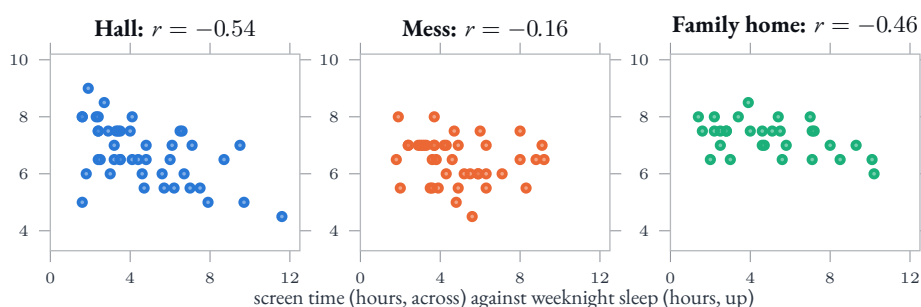


Figure 15: Screen time and weeknight sleep within each residence. **Simulated** class data.

It does not fade (Figure 15 and Table 13). Every panel slopes down. The mess panel is weaker, but with 42 students its interval is wide, from -0.45 to 0.15 , and the hall's interval, from -0.72 to -0.29 , overlaps it. The difference between panels is not firmly established.

Table 13: The correlation between screen time and weeknight sleep, within each residence and overall. **Simulated** class data.

	n	r
Hall	45	-0.54
Mess	42	-0.16
Family home	30	-0.46
Within residence, pooled	117	-0.40
Ignoring residence	117	-0.36

The pooled row compares each student with the averages of their own residence, then combines all three. Holding residence fixed, the link is a little *stronger*, -0.40 against -0.36 . Residence was not creating the link, and it was slightly hiding it. That removes one candidate. It does not make the link causal.

The next candidate

Stress correlates 0.45 with screen time and -0.48 with sleep (Table 11). It is a number, not three groups, so it cannot be split cleanly into panels. Chapter 6's regression holds it fixed instead.

3.5 Which way does the arrow point?

Reverse causation is the explanation people forget, because the story they have in mind feels obvious. Each of the following pairs gives the same data:

- The phone keeps students awake; or students who cannot sleep pick up the phone.
- In Islam 2020 (Chapter 4), falling behind academically causes depression; or depression causes falling behind.

A cause must come before its effect. Hill put the question as “which is the cart and which the horse?” (reference 5, p. 299). A survey taken once records both at the same moment, so it cannot say which came first. Following the same students over time, a *cohort* study, can.

Reverse causation

the outcome causing the supposed cause, rather than the other way round

3.6 Berkeley, 1973: a gap that reverses

In 1973 the University of California, Berkeley, admitted a smaller share of its women graduate applicants than of its men. Table 14 gives the six largest departments, the subset kept in R’s copy of the data (reference 2). Women were admitted at 30.4% against 44.5% for men. With four and a half thousand applications, a chi-square test would call the gap overwhelming.

Table 14: Graduate admissions in the six largest departments, Berkeley, 1973. Counts from R’s datasets::UCBAdmissions; percentages ours.

	Applied	Admitted	%
Men	2691	1198	44.5
Women	1835	557	30.4

Split by department, and the picture changes (Figure 16). Women were admitted at a higher rate in 4 of the 6 departments. In department A, 82.4% of women were admitted against 62.1% of men. In the other five, the two rates were close.

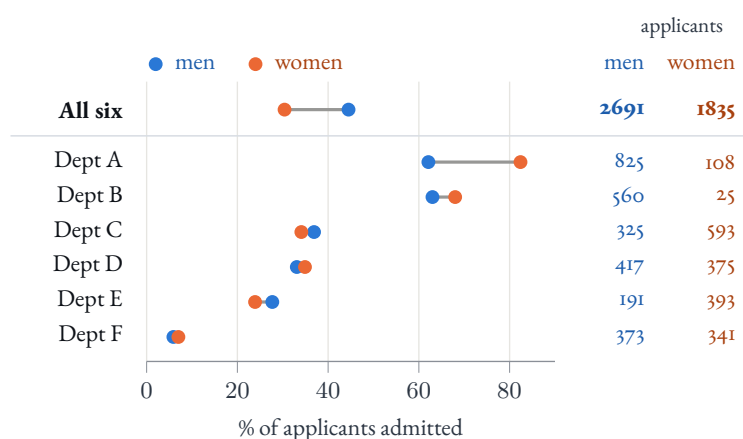


Figure 16: Admission rates by department and sex, Berkeley, 1973. Data from R’s datasets::UCBAdmissions, after Bickel, Hammel and O’Connell 1975.

The resolution lies in who applied where. 51.5% of men’s applications went to departments A and B, which admitted most applicants. Only 7.2% of women’s did. Women mostly applied where admission was hard for everyone. Combine the departments, and a difference in

where people applied looks like a difference in how they were treated. Bickel and colleagues called the university-wide figures “a clear but misleading pattern of bias against female applicants”. The reversal has a name: *Simpson’s paradox*.

Simpson’s paradox

An association in the combined table can shrink, vanish or reverse inside every subgroup. Here the third variable is department. Whether to split depends on the question. For “do committees treat equal applicants differently?”, compare within departments.

Simpson’s paradox

an association in a combined table that shrinks, vanishes or reverses inside every subgroup

3.7 Randomisation

In a survey, anything that shapes screen time can also shape sleep. In a randomised experiment, a coin decides who cuts back (Figure 17). The coin cannot be influenced by stress, money or residence, so nothing else can push the two variables together. A difference in sleep can then come only from the phone or from chance, and a test deals with chance.

Randomised experiment

a study in which chance decides who receives which condition

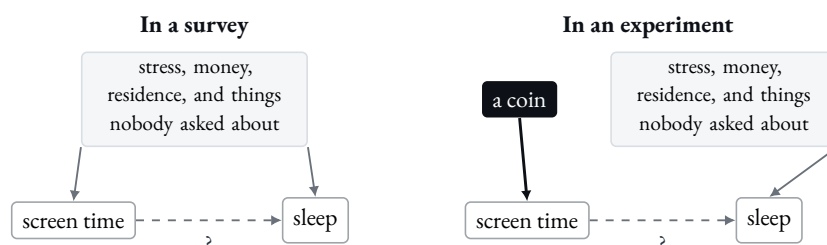


Figure 17: Why randomisation removes common causes, including those nobody measured. Schematic.

This is why trials are the standard for “does this treatment work?”. But much of what matters cannot be randomised. Nobody can assign students to smoke, to live in a mess, or to be born female.

3.8 When randomisation is impossible: Hill’s viewpoints

Smoking and lung cancer is the classic case. No trial was possible, yet the link came to be accepted as causal. In 1965 Austin Bradford Hill, co-author of the 1950 study that linked smoking with lung cancer (reference 6), listed nine viewpoints to weigh (reference 5). They are strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment and analogy.

Hill insisted that none of the nine is essential; they are a guide, not a checklist. The *gradient* is the easiest to see (Figure 18). Non-smoking doctors died of lung cancer at 0.07 per 1,000 a year, and those smoking 25 or more a day at 2.27. That is more than thirty times the rate, by our arithmetic. Each step up in dose brings a step up in rate. A confounder would have to rise with the number of cigarettes, step for step.

Applied to screen time and sleep, most viewpoints are unknown. One survey gives no consistency across studies and no temporality. A survey alone rarely makes the case.

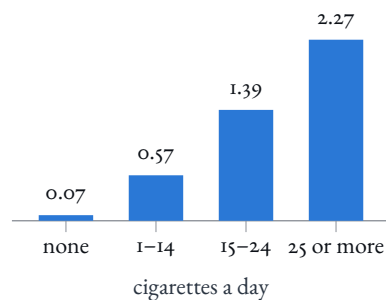


Figure 18: Lung cancer deaths per 1,000 doctors a year, by cigarettes smoked a day. Rates from Hill 1965, p. 296.

3.9 Words that claim too much

Causal language slips into abstracts constantly, often in the conclusions of papers whose methods describe a cross-sectional survey.

Claims a cause	Claims an association
screen time <i>reduces</i> sleep	screen time <i>is associated with</i> less sleep
the <i>effect</i> of residence on tuition	tuition <i>was more common</i> among mess residents
tuition <i>leads to</i> anxiety	anxiety <i>was higher</i> among tutors
the <i>impact</i> of stress on sleep	stress and sleep <i>were negatively correlated</i>

A survey like the class survey earns the right-hand column. A randomised trial, or a long argument of Hill's kind, can earn the left. The left column is not banned; it has to be earned by the design. For a thesis based on a survey, the right column is correct and no less interesting.

4 A cross-table in a published paper

4.1 Islam 2020, Table 2: anxiety by private tuition

Islam and colleagues surveyed 476 Bangladeshi university students online during the COVID-19 pandemic (reference 1). Their Table 2 gives counts, so this chapter's test can be run directly. Table 15 gives the anxiety rows for students with and without private tuition.

Table 15: Anxiety by private tuition. Counts from Islam 2020, Table 2. Row percentages ours.

	Minimal	Mild	Moderate	Severe	Total
No tuition	47 19.6%	91 37.9%	57 23.8%	45 18.8%	240
Tuition	40 16.9%	94 39.8%	61 25.8%	41 17.4%	236

The paper prints a percentage beside each count, but it divides every cell by all 476 students.

The 47 is printed as “(9.9)”. Those percentages answer “what share of all students were non-tutors with minimal anxiety?”, a question nobody has. Chapter 2’s denominator lesson applies: to compare the rows, divide within each row. The row percentages in Table 15 and Figure 19 are ours, and the two rows look alike.

	minimal	mild	moderate	severe
No tuition	19.6	37.9	23.8	18.8
Tuition	16.9	39.8	25.8	17.4

% of each row, anxiety category (GAD-7)

Figure 19: Anxiety level within each tuition group. Counts from Islam 2020, Table 2; row percentages ours.

The test agrees. On the printed counts, our chi-square is $\chi^2 = 0.90$ on 3 df, $p = 0.825$, with Cramér’s $V = 0.043$. The smallest expected count is 42.6, so the approximation is sound. The table gives no evidence that anxiety differs with tuition.

4.2 Table 2 against Table 3

Chapter 4 found that Table 3 of the same paper prints $p = .018$ for tuition and anxiety, although its own odds ratio and interval give 0.186. Table 16 puts the three versions side by side.

Table 16: Tuition and anxiety in Islam 2020, three ways. The rebuilt p and the chi-square are ours.

	p
Table 3, as printed	.018
Table 3, rebuilt from its odds ratio and interval (Chapter 4)	0.186
Table 2, our chi-square on four anxiety levels	0.825

Table 3 splits anxiety into yes or no and adjusts for other variables in a logistic regression (Chapter 6). Table 2 keeps four levels and adjusts for nothing. They are different analyses and need not match. But the chi-square sides with the rebuilt value rather than the printed one: neither supports a link.

4.3 A row that does not add up

The depression half of Table 2 prints the tuition, minimal-depression cell as “44 (8.4)”. But 8.4% of 476 is 39.98, or 40, not 44; 44 would be 9.2%. With 44, the tuition row totals 240. The anxiety rows show 236 tutors, and with 40 the depression row gives 236 too. The percentage, and the agreement between the two halves of the table, both point to 40. The count 44 is probably the misprint, but the paper does not say, and this is our reading.

The reader's habit

Add up the rows. Check each percentage against its count. It takes a minute, and slips like this one are ordinary: tables are typed by hand. The habit of checking is what matters.

5 Which test to use

5.1 Two questions first

Chapter 1 sorted research questions into three kinds and variables into types. Those two sorts choose the test.

1. **What kind of question?** *Describe* one variable; *compare* an outcome across groups; or *relate* two variables.
2. **What kind of variables?** Categorical or numeric. If numeric, roughly symmetric or skewed. The same people twice, or separate groups.

Most wrong tests in theses come from skipping these two questions. A test is run because a supervisor or a paper used it, not because the variables call for it.

5.2 The flowchart

Figure 20 is the map for Chapters 2 to 6. Read it by columns: first the kind of question, then the row that matches the variables.

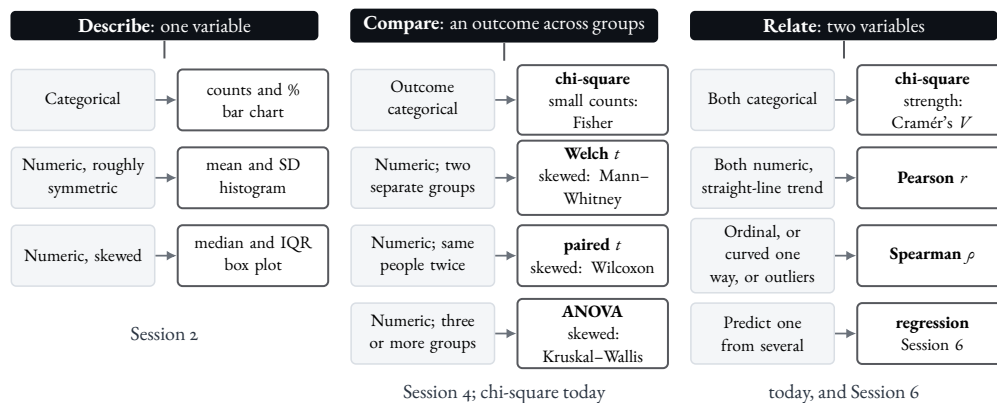


Figure 20: Choosing a test. “Categorical” includes yes or no answers; ordinal data use the rank-based partners.

Three features are worth noticing. Chi-square appears twice, because comparing a categorical outcome across groups and relating two categorical variables are the same test. Every mean-based test has a rank-based partner for skewed or ordinal data. And when more than one variable must be accounted for, the answer is regression, in Chapter 6.

5.3 Three questions through the chart

Table 17: Three questions from the class survey, taken through Figure 20. **Simulated** class data; the ANOVA is from Chapter 4.

Question	Variables	Test	Result
Does weeknight sleep differ across hall, mess and home?	numeric outcome; three separate groups	ANOVA	$F = 6.13, p = 0.003$
Are seniors more likely to tutor than juniors?	two categories each	chi-square	$\chi^2 = 9.96, p = 0.002$
Does stress rise with screen time?	two numbers; a straight-line trend	Pearson r	$r = 0.45, p < 0.001$

The same variable can play different parts. Year of study is a group in the second question. With all five ordered years, it could be ordinal, and Spearman's ρ would fit. Neither choice is wrong. The question decides, and it should be decided before seeing the data.

5.4 Five wrong choices

Table 18: Five common mistakes in choosing a test. Figures from the **simulated** class data.

Wrong	Why	Instead
Pearson r between residence, coded 1, 2, 3, and sleep	the codes are names; renumber them and r changes	ANOVA
Chi-square on the percentage table	χ^2 needs counts: 42.16 instead of 17.16	chi-square on counts
Two-sample t for weekend against weeknight sleep	the same students twice	paired t
Pearson r on CGPA band	four ordered bands, not measured numbers	Spearman ρ
Test every pair, report what is significant	six tests give a 26.5% chance of a false alarm	decide the question first

Each mistake ignores what the numbers *are*: codes, percentages, pairs, ranks, or one test among many. The first is the most tempting, because software will run it without complaint. Residence stored as 1, 2 and 3 looks like a number. Swap the codes for mess and home, and r changes, which shows it was never measuring anything.

6 Analysing a relationship: the procedure

1. **State the question** before seeing the data: which two variables, and whether the aim is to test for a relationship, measure its strength, or both.
2. **Classify both variables**: categorical or numeric, and if numeric, symmetric or skewed. The pair chooses the method (Figure 20).
3. **Look first**: a table of row percentages for two categories, a scatterplot for two numbers. Check for codes, typos, curves and single points that dominate.
4. **Two categorical variables**: compute the expected counts, check that none is below 5, then χ^2 , its df and p . Use Fisher's exact test when counts are small.
5. **Two numeric variables**: Pearson's r for a straight-line trend, Spearman's ρ for ranks, ordinal data or a few extreme values. Give the interval.
6. **Measure the size**: Cramér's V , a risk ratio or an odds ratio for tables; r and r^2 for numbers. Say where the association lies, with residuals.
7. **Consider the other explanations**: chance, a common cause, reverse causation. Split by any candidate that was measured.
8. **Report** the estimate, its interval, the test and n in words that claim an association, not a cause, unless the design earns more.

7 Chapter summary

1. Two categorical variables are summarised by a cross-table. Row percentages compare the rows; the chi-square test asks whether the differences could be chance.
2. Expected counts are what independence predicts: row total times column total, divided by N . χ^2 adds up $(\text{observed} - \text{expected})^2 / \text{expected}$ over every cell.
3. Degrees of freedom are $(r - 1)(c - 1)$. The test needs counts, not percentages, and expected counts of about 5 or more; otherwise use Fisher's exact test.
4. Residuals show which cells drive χ^2 . Cramér's V measures strength, and unlike χ^2 it does not grow with the sample.
5. In a 2×2 table, a risk difference, a risk ratio and an odds ratio are three different answers. Say which one is reported.
6. Pearson's r measures how closely two numbers follow a straight line, from -1 to $+1$. r^2 is the share of variation that lines up.
7. A small sample can give an r of the wrong sign; a large one makes a tiny r significant. Report r with its interval.
8. Plot first. Curves, single points, "don't know" codes and a narrow range all mislead r . Spearman's ρ , r on ranks, resists extreme values and suits ordinal data.
9. A high correlation between two measurements of the same thing does not mean they agree.
10. An association has four explanations: chance, a common cause, reverse causation and a real effect. A test rules out only the first.
11. Splitting by a measured third variable tests it as a common cause. Group averages can mislead about individuals, and a combined table can reverse inside every subgroup.
12. Randomisation removes common causes, measured or not. Without it, Hill's viewpoints help to weigh the case.
13. The kind of question and the kind of variables choose the test.

8 Key terms

Chi-square test

A test of whether two categorical variables are related, comparing observed with expected counts, equation (1.2).

Common cause

A third variable that influences both variables of interest.

Confounder

A third variable linked with both variables of interest, which creates or distorts the association between them.

Correlation coefficient (r)

Pearson's measure of how closely two numeric variables follow a straight line, equation (2.1).

Correlation matrix

A table of the correlations between every pair in a set of variables.

Cramér's V

A measure of the strength of association in a cross-table, from 0 to 1, equation (1.5).

Cross-table

A table of counts for every combination of two categorical variables.

Ecological fallacy

Reading a pattern among individuals off a pattern among group averages.

Expected count

The count a cell would hold if the two variables were unrelated, equation (1.1).

Fisher's exact test

A test for small cross-tables that does not rely on the chi-square approximation.

Fisher's z

A transformation of r used to build its interval, equation (2.3).

Hill's viewpoints

Nine considerations for judging whether an association is causal when an experiment is impossible.

Odds ratio

The odds of an outcome in one group divided by the odds in another.

Randomised experiment

A study in which chance decides who receives which condition.

Residual (standardised)

$(\text{Observed} - \text{expected}) / \sqrt{\text{expected}}$ for one cell, equation (1.4).

Restriction of range

A weaker correlation caused by studying a narrow slice of one variable.

Reverse causation

The supposed effect causing the supposed cause.

Risk ratio

The proportion with an outcome in one group divided by the proportion in another.

Scatterplot

A plot with one dot per person, placed by their values on two numeric variables.

Simpson's paradox

An association in a combined table that shrinks, vanishes or reverses inside every subgroup.

Spearman's ρ

Pearson's r computed on the ranks of the two variables.

9 Exercises

Exercises in Part A need no calculation. Parts B to E need a calculator; where a critical value or a tail area is needed, it is given. Numbers not attributed to a paper or to the class data are constructed for teaching.

Part A. Ideas.

1. In a table of residence by tuition, the expected count for “mess, tuition” is the count if residence and tuition were unrelated. Explain in words why it equals the row total times the column total, divided by N .
2. A study finds $\chi^2 = 2.1$, $p = 0.35$, for sex and choice of hall. Which statement is right, and what is wrong with the others? (a) Sex and choice of hall are unrelated. (b) The study found no evidence of a relationship. (c) There is a 35% chance that the two are unrelated.
3. Match each description to one of $r = -0.9, -0.3, 0, +0.5, +0.95$. (a) Points almost on a rising line. (b) A loose cloud drifting downwards. (c) A tight band falling steeply. (d) A round blob. (e) A clear upward drift with a lot of scatter.
4. Among primary school children, shoe size correlates strongly with reading score. Suggest a common cause, and say what a scatterplot within a single age group would show if you are right.
5. A news story reports a survey of 2,000 students under the headline “Breakfast boosts grades”. Rewrite the headline so that it claims only what a survey can show. Name two other explanations for the association.
6. Every count in a cross-table is doubled. What happens to χ^2 , to p and to Cramér’s V ? What does this say about reporting χ^2 and p alone?
7. Among students with a CGPA above 3.50, study hours and CGPA correlate at $r = 0.10$. A student concludes that study hours hardly matter for grades. Give one statistical reason to doubt this.
8. Two anxiety questionnaires, given to the same 80 students, correlate at $r = 0.92$. Can one replace the other, with the same cut-off for “anxious”? What would you look at instead?

Part B. Calculations.

9. Of 200 students, 120 ate breakfast and 80 skipped it. Of those who ate, 30 felt tired in the first class; of those who skipped, 40 did. (a) Set out the 2×2 table with totals. (b) Find the four expected counts. (c) Find χ^2 ; the critical value on 1 df is 3.84. (d) Find the four residuals. (e) Find V . (f) Give the risk of tiredness in each group, the risk ratio, and the odds ratio, skipped against ate.
10. (a) How many degrees of freedom has a table with 4 rows and 3 columns? Its critical value at 0.05 is 12.59. (b) In a table of 150 students, a row totals 40 and a column totals 45. Find the expected count in the cell where they meet.

11. Twenty students try two revision methods, ten each. With method X, 3 pass and 7 fail; with method Y, 8 pass and 2 fail. (a) Find the expected counts. (b) Is the chi-square test appropriate? (c) The chi-square test gives $p = 0.025$ and Fisher's exact test $p = 0.070$. Which should be reported, and what is the conclusion at 0.05?

12. Five students report hours of revision x and a quiz score y : $x = 1, 2, 3, 4, 5$ and $y = 2, 4, 5, 4, 5$. (a) Find r by hand, using equation (2.1). (b) Find t and its df; the critical value is 3.182. (c) What is the smallest $|r|$ that five people could show at $p < 0.05$? (d) Find r^2 and say it in words.

13. For the data of Exercise 12: (a) rank each variable, giving tied values the average of their ranks. (b) Find ρ as r on the ranks. (c) Find the shortcut value, equation (2.4), and explain why the two differ.

14. Two studies each find $r = 0.25$. (a) With $n = 50$, find t ; the critical value on 48 df is 2.011. (b) With $n = 200$, find t ; software gives $p = 0.0004$. (c) How much of the variation lines up in each study? What does the comparison show?

15. A study of 28 students finds $r = 0.60$. Use Fisher's z to build a 95% interval for the population correlation. The value $\frac{1}{2} \ln 4 = 0.693$ may help.

16. A table of faculty (four groups) by residence (three groups) for 300 students gives $\chi^2 = 24$. (a) Find the df; is the result significant at 0.05? (b) Find V .

Part C. The class data.

17. Use Table 7. (a) Find the four expected counts. (b) Find the four contributions and their sum, and compare it with the software's $\chi^2 = 9.96$. (c) Find Cramér's V . (d) In one sentence, compare the share of students giving tuition in the two year groups.

18. Stress and weeknight sleep correlate at $r = -0.48$ for the 118 students who answered both. (a) Find t by hand and its df. (b) Find r^2 as a percentage. (c) Use Fisher's z to find a 95% interval for the population correlation. Software gives -0.60 to -0.32 .

19. Stress and *weekend* sleep correlate at $r = -0.35$ for 117 students, $p < 0.001$, with a 95% interval of -0.50 to -0.18 . (a) Which kind of sleep tracks stress more closely? (b) Do the two intervals overlap, and what does that imply? (c) Suggest one reason weekend sleep might be less tied to stress.

20. In Table 1, 17 hall students give tuition, against an expected 23.11. (a) Find the residual for this cell. (b) Is it a cell worth talking about? (c) Which cell has the largest residual in Table 5?

21. Among the 42 mess students, screen time and sleep correlate at $r = -0.16$, $p = 0.298$, 95% interval -0.45 to 0.15 . Among hall students, $r = -0.54$, interval -0.72 to -0.29 . (a) Does the mess result show that screen time and sleep are unrelated among mess students? (b) Do the data show that the relationship differs between hall and mess?

22. The six p -values behind Table 11 are: screen and sleep < 0.001 ; screen and stress < 0.001 ; screen and study 0.743; sleep and stress < 0.001 ; sleep and study 0.406; stress and study 0.396. (a) Which pairs have $p < 0.05$? (b) Find the Bonferroni level for six tests. Which pairs pass it? (c) Why is the Bonferroni check worth making here?

Part D. Activities.

23. The class activity. For each research question: (a) describe, compare or relate? (b) What kind is each variable? (c) Which test, and what would you report beside p ?

1. Does the share who give tuition differ across the four faculty groups?
2. Do students in different faculty groups study different numbers of hours a day?
3. Do students in later years have higher CGPA bands?
4. Do students who give tuition have lower CGPA bands than those who do not?
5. Is residence related to having a CGPA of 3.00 or above?
6. Does daily screen time, which is skewed, differ across hall, mess and family home?
7. Do students sleep longer at weekends than on weeknights?
8. Do students with higher stress scores sleep less on weeknights?

24. Spot the wrong choice. Each analysis below has a flaw. Name it, and say what should have been done.

- (a) Pearson's r between faculty, coded 1 to 4, and monthly spending.
- (b) A chi-square test on a table of the percentages of smokers in each district.
- (c) Blood pressure in 30 patients before and after a drug, compared with a two-sample t -test.
- (d) A chi-square test on a 2×2 table with expected counts of 2.1 and 3.4.
- (e) Pearson's r between a five-point satisfaction rating and income in five bands.
- (f) Two glucose meters correlate at $r = 0.95$, so the new one is declared to agree with the old.
- (g) Twenty correlations are computed, and the two with $p < 0.05$ are reported.
- (h) A survey finds $r = -0.4$ between hours of sport and depression score, and concludes that sport prevents depression.

Part E. Challenge.

25. In Table 15: (a) find the expected count for students with tuition and severe anxiety, and that cell's contribution to χ^2 . (b) What share of the total $\chi^2 = 0.90$ is that? (c) Check the depression cell printed as "44 (8.4)": what count does 8.4% of 476 give?

26. In Berkeley's department A, 825 men and 108 women applied, and 62.1% and 82.4% were admitted. (a) Find the ratio of the women's admission rate to the men's. (b) Overall, women were admitted at 30.4% and men at 44.5%. Explain in two sentences how both can be true. (c) Which comparison answers "did department A's committee favour one sex?"

27. Hill's rates of lung cancer death per 1,000 doctors a year were 0.07 for non-smokers, 0.57 for 1 to 14 cigarettes a day, 1.39 for 15 to 24 and 2.27 for 25 or more (Figure 18). (a) Express each smoking rate as a multiple of the non-smokers' rate. (b) What would a confounder have to do to produce this pattern? (c) Which of Hill's viewpoints could the class survey provide for screen time and sleep, and which could it not?
28. Spearman's ρ for Anscombe's four datasets is 0.818, 0.691, 0.991 and 0.500, although r is 0.816 in each. Use Figure 9 to explain why ρ is highest for dataset III and lowest for dataset IV.
29. Design a study that could show whether screen time in the evening shortens sleep. Say who takes part, what is randomised, what is measured, and one practical difficulty.
30. Find Cramér's V for Table 15, and compare it with the class's V for residence and tuition. What does the comparison say, beyond the two p -values?

10 Solutions

Part A.

1. If residence and tuition were unrelated, the share giving tuition would be the same in every residence: the column total divided by N . The mess row would then hold that share of its own total. Row total times (column total $\div N$) is the formula. (Exercise 1.)
2. (b) is right. (a) claims too much: a large p is absence of evidence, not evidence of absence (Chapter 4). (c) misreads p : it is the chance of data this far from independence *if* the two were unrelated, not the chance that they are. (Exercise 2.)
3. (a) +0.95. (b) -0.3. (c) -0.9. (d) 0. (e) +0.5. (Exercise 3.)
4. Age: older children have bigger feet and read better. Within one age group, the correlation would be weak or absent. This is splitting by a candidate common cause, section 3.4. (Exercise 4.)
5. For example, "Students who eat breakfast report better grades". Other explanations include a common cause, such as a settled home routine that brings both breakfast and study time, and reverse causation, such as students under less academic pressure having time for breakfast. (Exercise 5.)
6. χ^2 doubles, so p falls. V does not change, because χ^2 and N double together (section 1.8). χ^2 and p mix up strength with sample size; report V or the percentages as well. (Exercise 6.)
7. Restriction of range: everyone in the group already has a high CGPA, so there is little spread for r to see (section 2.11). A weak r in a narrow slice says little about the relationship across all students. (Exercise 7.)

8. Not on that evidence. A high r shows that students keep roughly their order on both, not that the scores match; one could read systematically higher. Look at the differences between the two scores for each student: their average and spread (section 2.13). (Exercise 8.)

Part B.

9. (a) Ate: 30 tired, 90 not, total 120. Skipped: 40 tired, 40 not, total 80. Column totals 70 and 130; $N = 200$. (b) $120 \times 70/200 = 42$; $120 \times 130/200 = 78$; $80 \times 70/200 = 28$; $80 \times 130/200 = 52$. (c) $144/42 + 144/78 + 144/28 + 144/52 = 3.43 + 1.85 + 5.14 + 2.77 = 13.19$, above 3.84, so $p < 0.05$ (software: $p = 0.0003$). (d) $-1.85, +1.36, +2.27, -1.66$; only “skipped, tired” passes ± 2 . (e) $V = \sqrt{13.19/200} = 0.26$. (f) Risks 25% and 50%; risk ratio 2.0. Odds $40/40 = 1$ against $30/90 = 0.33$; odds ratio 3.0. (Exercise 9.)

10. (a) $(4 - 1)(3 - 1) = 6$. (b) $40 \times 45/150 = 12$. (Exercise 10.)

11. (a) Pass column total 11, fail 9, each row 10: expected counts 5.5 and 4.5 in each row. (b) No. Two expected counts are below 5. (c) Report Fisher’s exact test, $p = 0.070$. At 0.05 the evidence is not sufficient, although the pass rates, 30% and 80%, differ a lot. With twenty students, a larger study is needed. (Exercise 11.)

12. (a) $\bar{x} = 3, \bar{y} = 4$. Deviations: x : $-2, -1, 0, 1, 2$; y : $-2, 0, 1, 0, 1$. Products sum to $4 + 0 + 0 + 0 + 2 = 6$; $\sum(x - \bar{x})^2 = 10$; $\sum(y - \bar{y})^2 = 6$. $r = 6/\sqrt{60} = 0.775$. (b) $t = 0.775 \times \sqrt{3}/\sqrt{1 - 0.6} = 2.12$ on 3 df, below 3.182: not significant (software: $p = 0.12$). (c) 0.88, from $t = 3.182$ on 3 df. (d) $r^2 = 0.60$: 60% of the variation in quiz score lines up with revision hours in these five students. A strong r , and still not significant: five people are too few. (Exercise 12.)

13. (a) x ranks 1, 2, 3, 4, 5. The y values 2, 4, 5, 4, 5 rank 1, 2.5, 4.5, 2.5, 4.5. (b) Rank deviations from the mean rank 3: x : $-2, -1, 0, 1, 2$; y : $-2, -0.5, 1.5, -0.5, 1.5$. Products sum to 7; squares sum to 10 and 9. $\rho = 7/\sqrt{90} = 0.74$. (c) $d = 0, -0.5, -1.5, 1.5, 0.5$, so $\sum d^2 = 5$ and the shortcut gives $1 - 30/120 = 0.75$. The shortcut is exact only without ties. (Exercise 13.)

14. (a) $t = 0.25 \times \sqrt{48}/\sqrt{0.9375} = 1.79$, below 2.011: not significant ($p = 0.08$). (b) $t = 0.25 \times \sqrt{198}/\sqrt{0.9375} = 3.63$: significant. (c) $r^2 = 6.25\%$ in both. The relationship is identical; only the sample size differs. Significant is not strong (section 2.6). (Exercise 14.)

15. $z = \frac{1}{2} \ln(1.6/0.4) = \frac{1}{2} \ln 4 = 0.693$. $SE = 1/\sqrt{25} = 0.2$. The z interval is 0.693 ± 0.392 , from 0.301 to 1.085. Back on the r scale, 0.29 to 0.80. The interval is wide and lopsided: r could be modest or strong. (Exercise 15.)

16. (a) $(4 - 1)(3 - 1) = 6$ df; 24 is well above the critical value of 12.59, so $p < 0.05$ (software: $p = 0.0005$). (b) $k = 3$, so $V = \sqrt{24/(300 \times 2)} = \sqrt{0.04} = 0.20$. (Exercise 16.)

Part C.

17. (a) $52 \times 61/120 = 26.43$; $52 \times 59/120 = 25.57$; $68 \times 61/120 = 34.57$; $68 \times 59/120 = 33.43$. (b) $2.78 + 2.87 + 2.12 + 2.20 = 9.97$. The software's 9.96 differs only by rounding. On 1 df it is far above 3.84; $p = 0.002$. (c) $V = \sqrt{9.96/120} = 0.29$. (d) 61.8% of students in year 3 or later gave tuition, against 32.7% in years 1 and 2, a risk ratio of 1.89. (Exercise 17.)

18. (a) $t = -0.48 \times \sqrt{116} \div \sqrt{1 - (-0.48)^2} = -0.48 \times 10.77 \div 0.877 = -5.89$ on 116 df, $p < 0.001$. Software, from the unrounded r , gives -5.82 . (b) $r^2 = 23.0\%$ from the rounded r (22.6% unrounded). (c) $z = -0.523$; $SE = 1/\sqrt{115} = 0.093$; the z interval runs from -0.705 to -0.341 ; back on the r scale, -0.61 to -0.33 . The software's interval differs in the second decimal because it starts from the unrounded r . (Exercise 18.)

19. (a) Weeknight sleep: -0.48 against -0.35 . (b) They overlap: -0.60 to -0.32 against -0.50 to -0.18 . The data do not firmly show that the two correlations differ. A formal comparison of two correlations measured on the same students needs a test beyond this course. (c) For example, weekend sleep may be set more by catching up on lost sleep than by the week's worries. This is a suggestion, not something the data show. (Exercise 19.)

20. (a) $(17 - 23.11)/\sqrt{23.11} = -6.11/4.81 = -1.27$. Fewer hall students tutor than independence predicts. (b) Not by the rough guide: it is inside ± 2 . (c) Mess and tuition, $+2.36$. (Exercise 20.)

21. (a) No. Not significant is not the same as no relationship. The interval runs from -0.45 to 0.15 . It includes a moderate negative correlation as well as none at all; 42 students cannot tell them apart. (b) Not firmly. The intervals overlap, so the difference between -0.54 and -0.16 could be chance. (Exercise 21.)

22. (a) Screen and sleep, screen and stress, sleep and stress. (b) $0.05/6 = 0.0083$; the same three pass. (c) Six tests give about a 26.5% chance of at least one false alarm if nothing were related. Here the conclusion survives the correction. (Exercise 22.)

Part D.

23. The results use the **simulated** class data.

1. Compare. Faculty group, categorical; tuition, yes or no. Chi-square on a 4×2 table: $\chi^2 = 0.57$ on 3 df, $p = 0.902$, smallest expected count 11.3. Report the four percentages: Arts 48.3%, Business 47.8%, Science 45.5%, Social Science 54.3%.
2. Compare. Numeric outcome; four separate groups. One-way ANOVA: $F = 2.24$ on 3 and 116 df, $p = 0.088$. Study hours are right-skewed, so Kruskal–Wallis is a sensible check: $p = 0.055$. Report the four means, from 2.33 hours (Arts) to 3.02 (Business).
3. Relate. Both ordinal. Spearman's $\rho = 0.05$, $n = 105$ (students with no result yet are left out), $p = 0.601$.
4. Compare. Ordinal outcome; two separate groups. Mann–Whitney: $W = 1416.0$, $p = 0.780$. Both groups have a median band of 3.0.
5. Relate. Both categorical. Chi-square on a 3×2 table: $\chi^2 = 1.48$ on 2 df, $p = 0.478$. Report the share with 3.00 or above: hall 50.0%, mess 52.6%, family home 65.2%.
6. Compare. Skewed numeric outcome; three separate groups. Kruskal–Wallis: $H = 0.75$ on 2 df, $p = 0.688$. Report the medians: 4.10, 4.45 and 4.65 hours.
7. Compare. The same students twice. Paired t : mean difference 1.06 hours (95% CI 0.93 to 1.19), $t = 15.99$ on 118 df, $p < 0.001$.
8. Relate. Two numbers. Pearson's $r = -0.48$ (95% CI -0.60 to -0.32), $n = 118$, $p < 0.001$, with the scatterplot.

Questions 1 to 6 find no evidence of a difference or a relationship. That is not evidence that none exists (Chapter 4). (Exercise 23.)

24. (a) Faculty codes are names, so r means nothing; use ANOVA, or Kruskal–Wallis if spending is skewed. (b) χ^2 needs counts; convert back to counts first. (c) The same patients twice; use a paired t -test. (d) Expected counts below 5; use Fisher's exact test. (e) Both are ordinal; use Spearman's ρ . (f) Correlation is not agreement; look at the differences between the meters. (g) Twenty tests give a 64.2% chance of at least one false alarm; state the question in advance, or correct for twenty tests. (h) A survey shows association, not cause; depression may reduce sport, or a common cause may affect both. (Exercise 24.)

Part E.

25. (a) $236 \times 86 / 476 = 42.64$. Contribution = $(41 - 42.64)^2 / 42.64 = 2.6896 / 42.64 = 0.063$. (b) About 7% of 0.90: every cell contributes little, because every observed count sits close to its expected count. (c) $0.084 \times 476 = 39.98$, so 40, not 44. (Exercise 25.)

26. (a) $82.4 / 62.1 = 1.33$: women were a third more likely to be admitted. (b) Most men applied to departments that admitted most applicants, and most women to departments that rejected most. Combined, the difference in where they applied looks like a difference in treatment. (c) The comparison within department A. (Exercise 26.)



27. (a) $0.57/0.07 = 8.1$; $1.39/0.07 = 19.9$; $2.27/0.07 = 32.4$. (b) Rise in step with the number of cigarettes, and be strongly enough linked with lung cancer to produce a thirtyfold difference. (c) A single survey gives strength and possibly a gradient. It cannot give temporality, consistency across studies, or experiment. (Exercise 27.)

28. In dataset III the points lie on a perfect line except one. As a rank, the stray point swaps places with only one neighbour, so ρ is almost 1. In dataset IV, ten of the eleven x values are equal and share one rank. The ranks say only that the single point far to the right also has the largest y , which gives $\rho = 0.500$. Dataset II is a curve that rises and then falls, so the ranks do not rise consistently either. (Exercise 28.)

29. For example: recruit volunteer students; randomise half to put the phone away an hour before bed for two weeks, the other half to carry on as usual. Measure sleep each night with a diary or a sleep app, and compare the groups with a two-sample test. Difficulties include students not keeping to their group, and knowing which group they are in, which may change how they report sleep. (Exercise 29.)

30. $k = 2$, so $V = \sqrt{0.90/476} = 0.04$. The class's V for residence and tuition is 0.38, roughly nine times larger. The class table shows a real difference between rows; Islam's tuition rows are almost identical. The two p -values differ for the same reason, but V says it without the sample size. (Exercise 30.)

II 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
$\chi^2 = 0.90$, $V = 0.043$ and the row percentages of Table 15	equations (1.1) to (1.5)	Islam 2020, Table 2
$p = 0.186$ in Table 16	rebuilt from the odds ratio and interval, Chapter 4	Islam 2020, Table 3
39.98, and the reading “40, not 44”	8.4% of 476, and the row totals; our guess, unconfirmed	Islam 2020, Table 2
Admission percentages, 1.33, 51.5% and 7.2%	from the counts	R's UCBA admissions
Table 10; Spearman's ρ in Exercise 28	computed in R	R's anscombe
“More than thirty times”; 8.1, 19.9, 32.4	each rate divided by 0.07	Hill 1965, p. 296
Table 9	from the t curve on $n - 2$ df	none
Chi-square critical values	standard tables, rounded to two decimals	none

Simulated, not real students: every figure from the class data, including the 1,000 shuffles

of Figure 4, the rescaled tables of Table 6, the six students of Example 2.1, the 99 codes of section 2.9 and the type of section 2.10. All are generated by data/class-data.R from a fixed seed, and the class figures will be replaced by the real survey.

Constructed for teaching, and not from any study: the six clouds of Figure 6 and the two of Figure 8, the two studies of section 2.6, and the two thermometers of section 2.13. So is every scenario and number in Exercises 1 to 8, 9 to 16, 24 and 29. Every answer to Part B was checked in R.

12 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. Bickel PJ, Hammel EA, O'Connell JW. Sex bias in graduate admissions: data from Berkeley. *Science* 1975;187(4175):398–404. PMID 17835295. doi:10.1126/science.187.4175.398. The quotation is from the abstract. The six-department counts are R's copy of the data, not the paper's full table.
3. Anscombe FJ. Graphs in statistical analysis. *The American Statistician* 1973;27(1):17–21. Not indexed in PubMed. doi:10.1080/00031305.1973.10478966. The data are in R as datasets::anscombe.
4. Messerli FH. Chocolate consumption, cognitive function, and Nobel laureates. *N Engl J Med* 2012;367(16):1562–1564. PMID 23050509. doi:10.1056/NEJMon1211064. The full text was not checked; no figures from it are quoted.
5. Hill AB. The environment and disease: association or causation? *Proc R Soc Med* 1965;58(5):295–300. PMID 14283879. PMC1898525. doi:10.1177/003591576505800503. Free to read. The death rates are from p. 296 and the quotation from p. 299.
6. Doll R, Hill AB. Smoking and carcinoma of the lung; preliminary report. *Br Med J* 1950;2(4682):739–748. PMID 14772469. doi:10.1136/bmj.2.4682.739.
7. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1(8476):307–310. PMID 2868172.
8. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988. A book, with no PMID. The benchmarks of section 1.7 are quoted as widely reported and were not checked against the book for these notes.
9. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Datasets anscombe and UCBA admissions.
10. 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.

13 For students in the Stat.BD course

Before the next session. Read this chapter, and work the Part A exercises (Exercises 1 to 8) and the activity answers of Exercise 23. Exercise 24, spotting the wrong choice, is good preparation for reading papers. The rest are for practice at any point in the course. Bring a calculator.

What the next session does with this chapter. This chapter could hold residence fixed by splitting the class into three panels. Stress cannot be split that way, because it is a number. In the class, each extra hour of screen time goes with about 0.88 more stress points. Once sleep is taken into account, it is about 0.64. Where did the rest go? Session 6 answers with regression, which holds a numeric variable fixed. It also covers logistic regression, the method behind Islam’s Table 3, and how questionnaire items are built into a scale.

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

Question	Where it is settled
How can a numeric variable such as stress be held fixed?	Chapter 6, multiple regression
What does “ r^2 ” become when a line is fitted?	Chapter 6, R^2
How is an odds ratio estimated with other variables held fixed?	Chapter 6, logistic regression
Is the living-with-family row of Islam 2020 sound?	Chapter 6, a check against Table 2
How are the five stress items combined into one score?	Chapter 6, survey scales
Running these tests in software	Session 7 onward