

Relationships between Variables

Chi-square, correlation, and why association is not causation

Muhtasim Munif Fahim, MSc

Stat.BD

Data Science Research Lab, Department of Statistics & Data Science,
University of Rajshahi

Session 5 of 14 • Theory

Session 4 compared a number across groups. Today, two variables at once

Session 4 gave us

- ▶ the logic of a test, and the p -value
- ▶ t -tests and ANOVA for a numeric outcome
- ▶ rank-based tests for skewed data

Every one of them compared a *number* across groups. Today neither variable has to be a group, and neither has to be the outcome.

Today, in five acts

- | | |
|-----|-------------------------------------|
| I | Two categories: the chi-square test |
| II | Two numbers: correlation |
| III | Association is not causation |
| IV | A cross-table in a published paper |
| V | Which test do I use? |

The question behind today

Do these two things go together? And if they do, what does that tell us?

Class data throughout: **simulated** stand-in for the class survey (data/class-data.R, fixed seed). Not real students.



What you will be able to do by the end of today

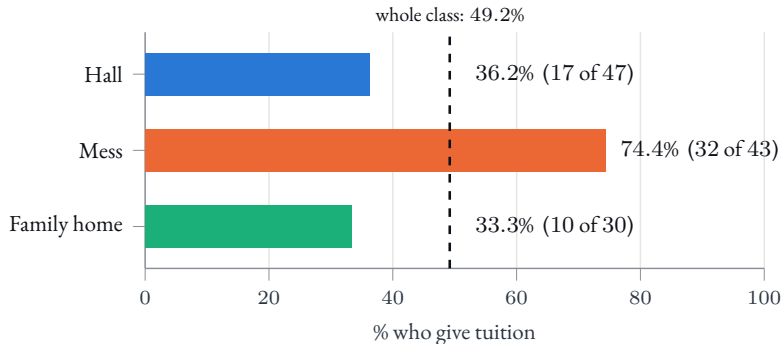
The test of this session

Given two variables from a survey, choose the right way to test whether they are related, say how strong the relationship is, and say what it does and does not show.

- ▶ Test whether two categorical variables are associated using chi-square
- ▶ Interpret the direction and strength of a correlation coefficient
- ▶ Explain, with an example, why correlation does not prove causation
- ▶ Use the “Which test do I use?” flowchart to pick a test for a given question



The question Session 4 left: does residence go with tuition?



Three in four mess students give tuition. About one in three hall students do, and about one in three of those living at home.

If residence made no difference, every bar would sit near the dashed line. Could a gap this size appear by luck, in a class of 120?

Simulated class data. The percentages are Session 2's row percentages.



Act I

Two categories: the chi-square test

The cross-table, and what “no relationship” would look like

	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

H_0 : residence and tuition are **independent**. Knowing where a student lives tells you nothing about whether they tutor.

H_1 : they are associated.

Independence, from Session 3

If H_0 were true, the tuition share would be the same in every row: about 49.2%, the whole-class figure. Luck would move each row a little, not this much.

Simulated class data.



Expected counts: what the table would show if H_0 were true

Session 3 found one of these already: if residence made no difference, the hall row would hold

$$47 \times \frac{59}{120} = 23.1$$

tuition students. The same recipe fills every cell:

$$E = \frac{\text{row total} \times \text{column total}}{N}$$

Expected counts E

	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

Same totals as the real table. Only the inside has changed.

Expected counts need not be whole numbers

23.1 is an average over many imaginary classes, not a number of students.

Simulated class data. Expected counts are our arithmetic; each row sums to its total up to rounding.



Observed against expected: the chi-square statistic

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
Home, no tuition	20	15.2	1.48
Home, tuition	10	14.8	1.53
Sum: χ^2			17.16*

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

One cell by hand: mess, tuition

$$E = 43 \times 59 \div 120 = 21.14$$

$$(32 - 21.14)^2 = 10.86^2 = 117.94$$

$$117.94 \div 21.14 = \mathbf{5.58}$$

Squaring makes every gap count, whichever way it goes. Dividing by E scales each gap by how big the cell was expected to be.

Simulated class data. Hand arithmetic ours. *The rounded contributions add to 17.15; 17.16 is from the unrounded values.



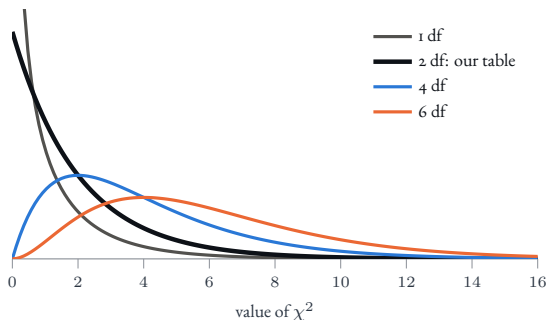
Degrees of freedom: how many cells are free to vary?

Fix the row and column totals. Choose the tuition count for hall and for mess. Every other cell is now forced, by subtraction.

Two free choices: **2 degrees of freedom**.

$$\begin{aligned}df &= (\text{rows} - 1)(\text{columns} - 1) \\&= (3 - 1)(2 - 1) = 2\end{aligned}$$

Session 4's Kruskal–Wallis H , 0.75 for three residences, was read from this same 2 df curve.

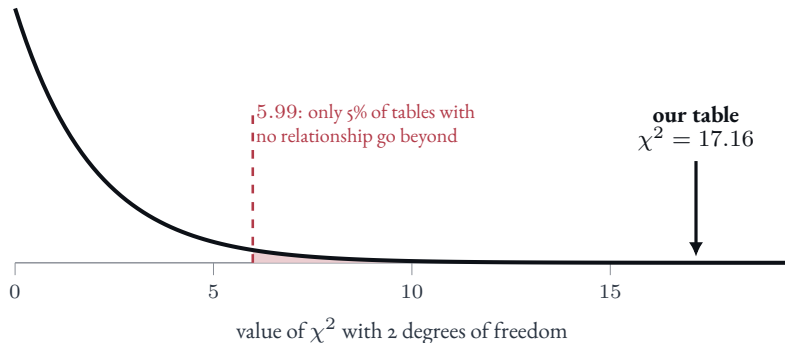


The chi-square curve is the null's picture: how χ^2 spreads across many classes where H_0 is true. More df, more cells, larger values by chance.

Textbook chi-square densities, drawn from their formulas.



How surprising is 17.16 if residence makes no difference?



With 2 df, only 5% of tables with no relationship give a χ^2 above 5.99. Ours is nearly three times that.

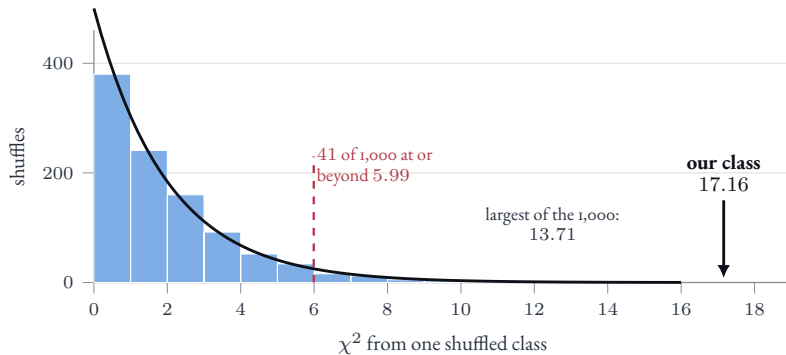
$$\chi^2 = 17.16, df = 2, p < 0.001.$$

Reject independence: in this class, residence and tuition are associated.

Simulated class data. p from the chi-square curve with 2 df (software). For 2 df only, $p = e^{-\chi^2/2}$, about 0.0002 here: our arithmetic.



Where the curve comes from: 1,000 classes with no relationship



Shuffle the 120 tuition answers at random across the students: residence and tuition are now unrelated *by construction*. Compute χ^2 . Repeat 1,000 times. 41 shuffles reached 5.99, close to the 5% the curve predicts. None reached 17.16.

Simulated: 1,000 random shuffles of the simulated class (data/class-data.R). The curve is the 2 df chi-square, scaled to 1,000 tables.



Which cells drive it? Residuals

$$\text{residual} = \frac{O - E}{\sqrt{E}}$$

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

Each residual is the square root of a contribution, with its sign kept. Beyond about ± 2 is a cell worth talking about.

Only the mess row gets there: **more** mess students tutor than independence predicts, and fewer do not. Hall and home sit a little on the other side, together.

The two-step habit

The chi-square says *whether* the table departs from independence. The residuals say *where*, and in which direction.

Simulated class data. Pearson residuals, by software.



How strong is the association? Cramér's V

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

k is the smaller of the number of rows and columns;
here $k = 2$.

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

V runs from 0, no association, to 1, where the row tells you the column exactly.

A common rough guide when $k - 1 = 1$:

small	0.1
medium	0.3
large	0.5

Why a second number

p says whether the association could be luck. V says how strong it is. They answer different questions, exactly as p and the effect size did in Session 4.

Simulated class data. Benchmarks: Cohen 1988, for his w , which equals V when $k - 1 = 1$; as widely quoted, not checked against the book's text.



χ^2 grows with the sample; V does not

The same percentages in every cell, at three class sizes:

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 . The strength of the pattern has not changed.

Never run chi-square on percentages

Feed the row percentages in as if they were counts and χ^2 becomes 42.16, not 17.16. The test believes it has 300 students.

The Session 4 lesson, again

A big enough sample makes any pattern significant. Report V beside p .

Simulated class data. The rescaled tables are ours; counts divided by 10 are not whole numbers, which is the point: it is a thought experiment.



When the chi-square test can be trusted

1. **Counts**, not percentages, means or scores.
2. **Each person once**, in one cell. The same students asked twice are paired, and need a different test.
3. **Expected counts large enough**. The safe rule: every E at least 5.

Our smallest expected count is 14.8. All three hold.

When expected counts are small

- ▶ A 2×2 table: **Fisher's exact test**, which needs no curve.
- ▶ Larger tables: merge categories that belong together, before looking at the result.

Rule 3 is about E , not O

An observed count of 2 is fine if its expected count is 8. The chi-square curve is an approximation that fails when the *expected* counts are small.

Simulated class data. The rule of five is a convention; textbooks differ on its exact form.



A 2×2 table: juniors, seniors and tuition

	No	Tuition	Total	%
Years 1-2	35	17	52	32.7
Years 3 to Masters	26	42	68	61.8

Is it luck? (1 df)

$$\chi^2 = 9.96 \quad p = 0.002$$

$$\text{with Yates' correction} \quad p = 0.003$$

$$\text{Fisher's exact test} \quad p = 0.002$$

How big? Four ways to say it

$$\text{Difference} \quad 29.1 \text{ points}$$

$$\text{Ratio of percentages} \quad 1.89$$

$$\text{Odds: juniors, seniors} \quad 0.49, 1.62$$

$$\text{Odds ratio} \quad 3.33$$

$$\phi \text{ (Cramér's } V) \quad 0.288$$

$$\text{Odds} = \text{tutor} \div \text{not: } 17 \div 35 \text{ for juniors.}$$

Three p -values, one answer

With expected counts this large (smallest 25.6), the versions agree. They part company only when counts are small.

Simulated class data. Tests by software; the odds ratio is from the unrounded odds.



One sentence that carries everything

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 $V = 0.38$.

It contains

- ▶ the percentages, so the reader sees the pattern
- ▶ the test, its df and N
- ▶ the p -value, exactly or as “ < 0.001 ”
- ▶ the strength

It avoids

- ▶ “living in a mess *makes* students tutor”
- ▶ $p = 0.000$
- ▶ the statistic alone, with no percentages

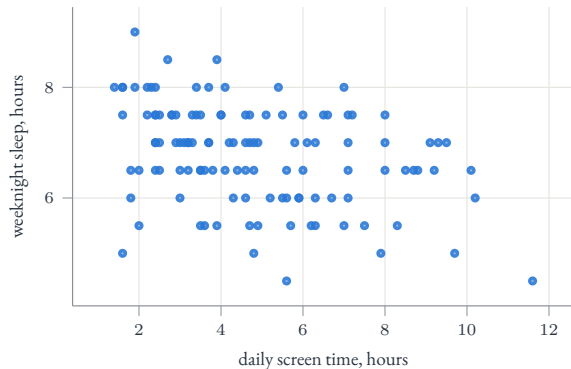
Simulated class data. Format follows common journal practice.



Act II

Two numbers: correlation

Two numbers: screen time and sleep



Session 2 described this plot in words:
down to the right, loose, no wild points.

Today, one number for the strength of
the straight-line trend:

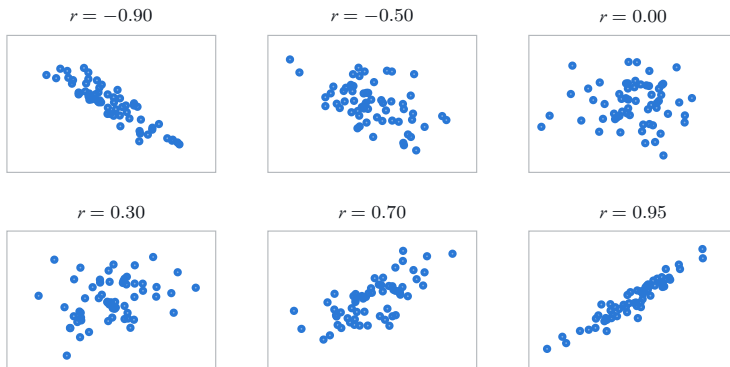
$$r = -0.36$$

The correlation coefficient, Pearson's r .

Simulated class data, $n = 117$ with both answers.



What r looks like



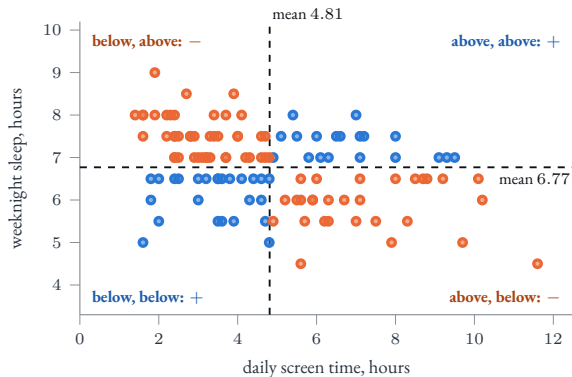
Sign: the direction. Up to the right is positive, down to the right is negative.

Size, from 0 to 1: how tightly the points hug a straight line. ± 1 is a perfect line.

Simulated clouds of 60 points each, built so that r is exactly the printed value.



Where r comes from: products of deviations



Cut the plot at the two means. For each student multiply the two deviations.

Blue: same side of both means, product positive: 44 students.

Orange: opposite sides, product negative: 73 students.

Negatives win, so r is negative.

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

The top adds up the products. The bottom rescales the answer to lie between -1 and $+1$.

Simulated class data.



r by hand: the first six students in the file

	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*

$$\bar{x} = 3.62, \quad \bar{y} = 7.33$$

$$\sum(x - \bar{x})^2 = 16.87, \quad \sum(y - \bar{y})^2 = 1.33$$

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

Simulated class data, first six rows of the file. Hand arithmetic ours. *The rounded products add to 2.11; 2.12 is from the unrounded values.



Six students say $+0.45$. The class says -0.36 . Which to trust?

Testing r against zero

H_0 : no straight-line relationship in the population.

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

Class: $r = -0.36$, $n = 117$

$t = -4.19$, $\text{df} = 115$, $p < 0.001$

95% CI: -0.51 to -0.20

Six students

With $n = 6$, r must pass ± 0.81 to reach $p < 0.05$. Our $+0.45$ is well inside the range that luck produces.

Six points can tilt either way. The whole class's interval sits entirely below zero.

A small sample can point the wrong way

Not only less precise: in the opposite direction. This is Session 3's sampling variation.

Simulated class data. Tests by software; the critical r is our arithmetic from the t curve.



r^2 : how much of the variation lines up

Square r : $(-0.364)^2 = 0.132$

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.

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

Why r flatters

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

Simulated class data. r^2 from the unrounded r ; $(-0.36)^2$ would give 13.0%. The table is our arithmetic.



Smallest r with $p < 0.05$

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

Two studies

$r = 0.10, n = 1,000: p = 0.002.$

Significant, and it lines up 1.0% of the variation.

$r = 0.50, n = 10: p = 0.141.$

Not significant, and possibly strong. Too few people to tell.

Session 4's lesson for correlations

p answers “could this be zero?”, not “is this large?”. Report r and its interval.

Critical values and p -values are our arithmetic from the t curve with $n - 2$ df.



r cannot pass -1 or $+1$, so its sampling spread is lopsided near the ends. Fisher's trick stretches the scale, builds the interval there, and brings it back.

1. Stretch $z = \frac{1}{2} \ln \frac{1+r}{1-r} = -0.381$
2. Standard error $1 \div \sqrt{n-3} = 1 \div \sqrt{114} = 0.094$
3. Interval $-0.381 \pm 1.96 \times 0.094$
 $= -0.565$ to -0.198
4. Bring back $r = \frac{e^{2z}-1}{e^{2z}+1}$: -0.51 to -0.20

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.

Software does all four steps. You only need to know why the interval is not $r \pm$ something.

Simulated class data. Arithmetic ours, from the unrounded r (-0.364).

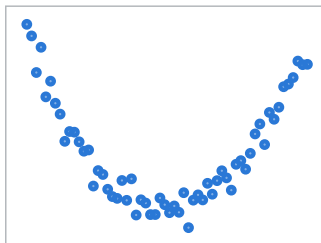


r measures straight lines only

no pattern: $r = 0.00$



a strong curve: $r = -0.02$



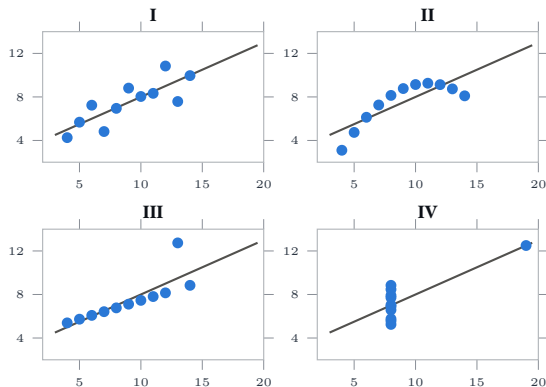
Left: no relationship, r near zero. Right: a strong, perfectly clear relationship, and r is also near zero.

$r = 0$ means *no straight-line* relationship. It does not mean no relationship. Plot first.

Simulated points, made for teaching.



Anscombe's quartet: four datasets, one set of statistics



All four have

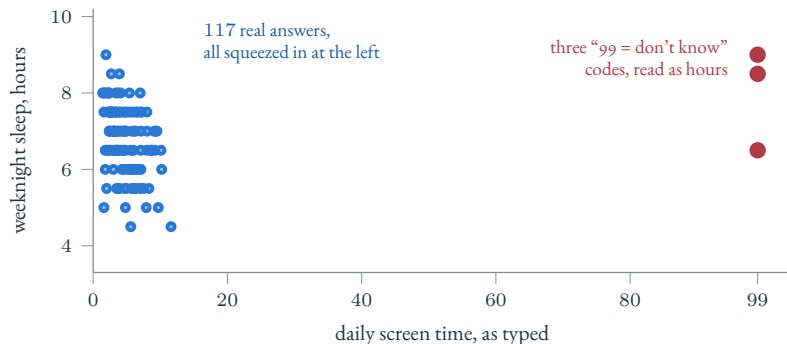
mean of x	9.00
mean of y	7.50
variance of x	11.00
variance of y	4.12 to 4.13
r	0.816 to 0.817
line	$y = 3.00 + 0.500x$

Only I looks like the cloud r describes. II is a curve, III is a line with one outlier, and in IV one point does all the work.

Anscombe FJ, *Am Stat* 1973. Data: R's datasets::anscombe; the statistics are computed by us from that copy.



Three “don’t know” codes change the sign



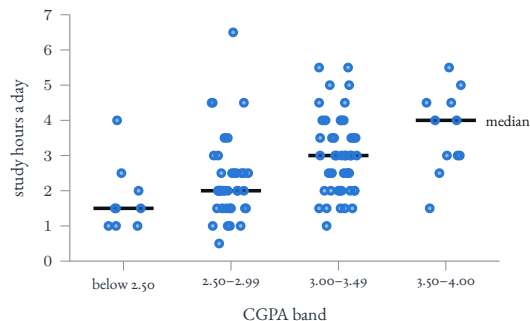
Three students ticked “don’t know”, coded 99. Left in as hours, they turn $r = -0.36$ into $r = +0.15$.

Three points out of 120, far to the right, outweigh everyone else. Spearman’s ρ , next, only moves from -0.36 to -0.30 .

Simulated class data, with Session 2’s three missing-value codes put back in.



Spearman's ρ : correlation on ranks



CGPA band is **ordinal**: four ordered bands, not a measured number.

Replace each variable by its ranks, then compute r on the ranks. That is Spearman's ρ .

$$\rho = 0.46, n = 105, p < 0.001$$

(Pearson on the band codes: 0.42.)

What ρ measures

Whether higher on one goes with higher on the other, in any steady way, straight or curved.

Simulated class data. Dots jittered sideways; bars are band medians.



Spearman by hand, and why it shrugs off a typo

The same six students, ranked

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

Tied values share the average of the ranks they would have taken.

r on the two rank rows: $\rho = \mathbf{0.39}$

Session 2's typo, put back

One student's 2.5 study hours typed as 25:

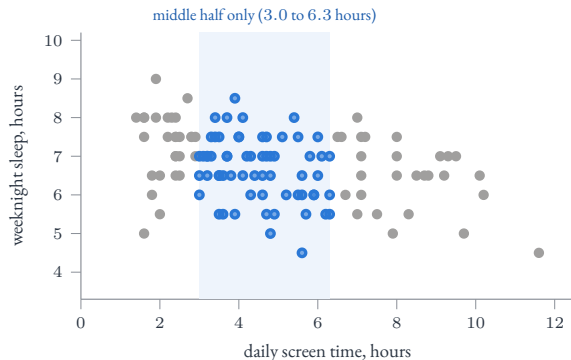
	clean	with 25
Pearson r	0.42	0.15
Spearman ρ	0.46	0.44

As a rank, 25 is just "the largest".

Simulated class data. Ranks and ρ by hand, ours; software agrees. The shortcut $1 - 6 \sum d^2 / n(n^2 - 1)$ assumes no ties and gives 0.43 here.



Look at a narrow slice and the correlation weakens



Keep only the middle half of screen time,
3.0 to 6.3 hours: 62 students.

r goes from -0.36 to -0.27 .

Same students, same relationship. Less
spread in x leaves less for r to see.

Restricted range

Admission scores correlate weakly with later grades *among those admitted*, because everyone admitted scored high. The narrow group hides the relationship.

Simulated class data.



A correlation matrix, and the six-test problem

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

Sleep is weeknight sleep. Each r uses the students who answered both.

Four variables give six pairs. Three pass $|r| > 0.18$.

If none of the six were related, the chance that at least one would reach $p < 0.05$ anyway:

$$1 - 0.95^6 = 26.5\%$$

Stress sits in the middle

Stress goes with more screen time and with less sleep. Keep that in mind for Act III.

Simulated class data. The 26.5% is our arithmetic and treats the six tests as independent.



Weekend and weeknight sleep, the same students:

$r = 0.74$: strongly related.

Mean weekend 7.87, weeknight 6.81: weekend is 1.06 hours longer.

Students who sleep long on weeknights also sleep long at weekends. But the two answers do not *agree*: the weekend figure runs consistently higher.

Two thermometers

One always reads exactly one degree high.
Their correlation is $+1$. They never agree.

To compare two measurements of the same thing, look at the differences, not r . Bland and Altman made this point for clinical measurement in 1986.

Simulated class data. Bland JM, Altman DG. *Lancet* 1986; PMID 2868172.



One sentence that carries everything

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

Also

- ▶ show the scatterplot
- ▶ say Pearson or Spearman, and why
- ▶ give n : people can skip a question

Avoid

- ▶ “screen time *reduces* sleep”
- ▶ “a significant correlation” with no r
- ▶ r from data you have not plotted

Simulated class data.



Act III

Association is not causation



Stat.BD

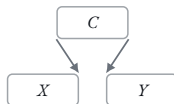
Four reasons two things go together

1. Chance



Nothing links them. This sample simply came out that way, and the next one will not.

2. A common cause



Something else moves both. X and Y travel together without either touching the other.

3. Reverse causation



The arrow runs the other way. Having the outcome changed the exposure.

4. A real effect



Change X and Y changes with it. This is the only one of the four that is causation.

An association is the finding. These four are the explanations, and the data alone cannot tell you which one you are looking at. The design is what narrows the list.

Acts I and II dealt with the first: chance. Screen time and sleep, residence and tuition, both passed. That leaves three explanations, and a p -value cannot choose between them.

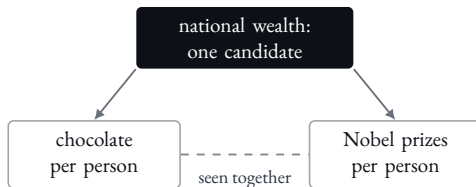
Figure from the Stat.BD library.



Chocolate and Nobel prizes: a common cause

In 2012 a cardiologist, Franz Messerli, published a short note in the *New England Journal of Medicine* comparing countries: chocolate eaten per person against Nobel laureates per person. Countries that ate more chocolate had more laureates.

Nobody sensible concludes that chocolate makes Nobel laureates.



The name: confounding

A **confounder** is a third variable linked with both, which creates or distorts the association between them. Session 6 shows how to adjust for one.

Messerli FH. *N Engl J Med* 2012; PMID 23050509. The full text was not checked, so no figures from it are quoted. National wealth is our suggestion.



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.

Our class shows the same trap in miniature. Compare only the three residence averages:

	Mean screen	Mean sleep
Hall	4.64	6.71
Mess	4.85	6.48
Family home	5.00	7.27

Family home has the *most* screen time and the *most* sleep.

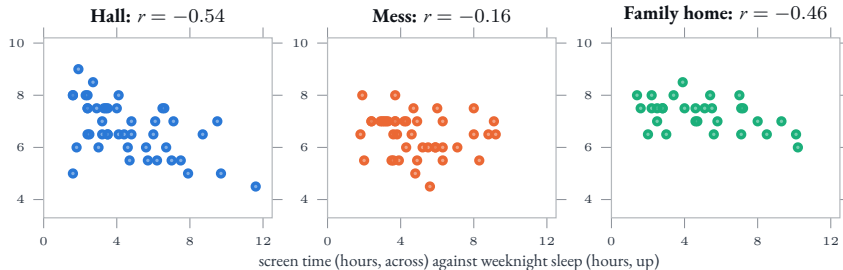
The ecological fallacy

A pattern among group averages need not hold for the individuals inside the groups. Inside every residence, as the next frame shows, more screen time goes with less sleep.

Simulated class data. Hours, among students who answered both questions.



Is residence behind the screen and sleep link? Look inside each



If residence were the common cause, the link would fade inside each residence, where everyone shares one. It does not. Every panel slopes down.

Simulated class data. Hall $n = 45$, mess $n = 42$, family home $n = 30$.



Holding residence fixed: the link survives

	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

“Pooled” compares each student with the averages of their own residence, then combines all three.

Holding residence fixed, the link is a little *stronger*, not weaker.

The next candidate

Stress: 0.45 with screen time, -0.48 with sleep. It is a number, so we cannot split by it cleanly. Session 6’s regression holds it fixed instead.

Simulated class data. The pooled r correlates deviations from each residence’s own means (software).



Which way does the arrow point? Reverse causation

Screen time and sleep

- ▶ The phone keeps students awake.
- ▶ Students who cannot sleep pick up the phone.

Both give the same scatterplot.

Islam 2020, from Session 4

- ▶ Falling behind causes depression.
- ▶ Depression causes falling behind.

A survey taken once cannot say which came first.

Time order

A cause must come before its effect. Hill put the question plainly: “which is the cart and which the horse?” A one-off survey records both at the same moment.

Hill AB, *Proc R Soc Med* 1965, on temporality. Islam 2020, Table 3, as discussed in Session 4.



Berkeley, 1973: a gap in admissions

Graduate applications to the six largest departments at the University of California, Berkeley, in 1973:

	Applied	Admitted	%
Men	2691	1198	44.5
Women	1835	557	30.4

Fourteen points lower for women, in a table of four and a half thousand people. A chi-square test would call it overwhelming.

Is this evidence that admissions discriminated against women?

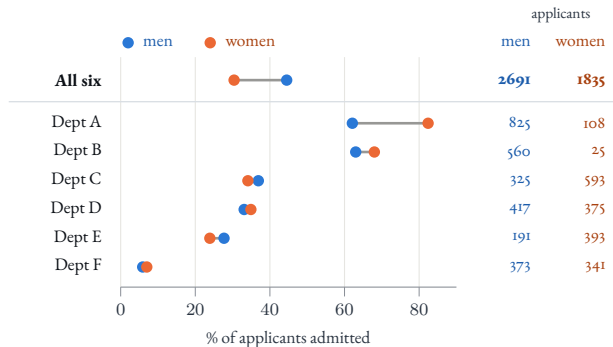
What the authors called it

Bickel and colleagues described the university-wide figures as “a clear but misleading pattern of bias against female applicants”.

Bickel PJ, Hammel EA, O’Connell JW. *Science* 1975; PMID 17835295; quotation from the abstract. Six-department counts: R’s datasets::UCBAdmissions.



Split by department and the gap reverses



Women were admitted at a higher rate in 4 of the 6 departments.

But 51.5% of men's applications went to A and B, which admit most applicants, against 7.2% of women's.

Women mostly applied where admission was hard, for everyone.

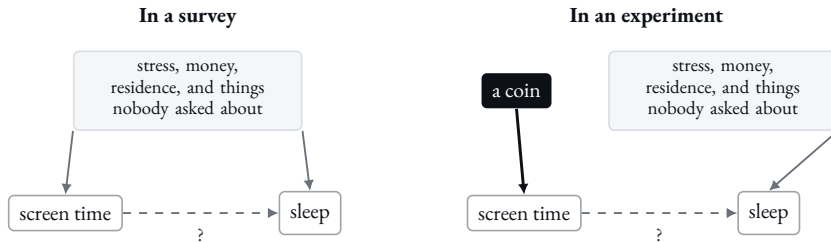
Simpson's paradox

An association in the combined table can shrink, vanish or reverse inside every subgroup. The third variable here is department.

R's datasets::UCBAdmissions, after Bickel, Hammel & O'Connell 1975. Percentages ours.



Why a randomised experiment settles what a survey cannot



In a survey, anything that shapes screen time can also shape sleep. In an experiment, a coin decides who cuts back, so nothing else can push the two together. A difference in sleep can then only come from the phone, or from chance, and a test handles chance.

Not always possible

Nobody can randomise students to smoke, to live in a mess, or to be born female.

Schematic.



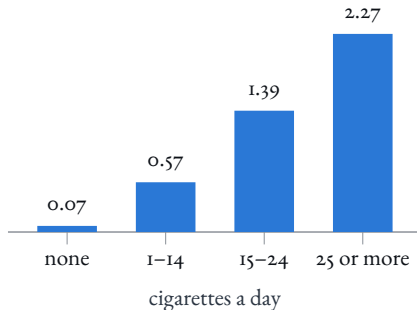
When you cannot randomise: Hill's viewpoints

In 1965 Austin Bradford Hill, co-author of the 1950 study linking smoking with lung cancer, listed nine things to weigh:

strength	plausibility
consistency	coherence
specificity	experiment
temporality	analogy
biological gradient	

He said none could be required as “a sine qua non”: a guide, not a checklist.

A gradient: lung cancer deaths per 1,000 doctors a year



The more smoked, the higher the rate. A common cause would have to rise in the same steps.

Hill AB. *Proc R Soc Med* 1965;58(5):295-300; PMID 14283879. Figures from p. 296; quotation from p. 299. Doll R, Hill AB. *BMJ* 1950; PMID 14772469.



Words that claim too much

Claims a cause

screen time *reduces* sleep
the *effect* of residence on tuition
tuition *leads to* anxiety
the *impact* of stress on sleep

Claims an association

screen time *is associated with* less sleep
tuition *was more common* among mess residents
anxiety *was higher* among tutors
stress and sleep *were negatively correlated*

The rule

A survey like ours earns the right-hand column. A randomised trial, or a long argument of Hill's kind, can earn the left.

Wording examples ours.



Act IV

A cross-table in a published paper

	Minimal	Mild	Moderate	Severe
No tuition	47	91	57	45
Tuition	40	94	61	41

	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)

Our chi-square on the printed counts:

$$\chi^2 = 0.90, df = 3, p = 0.825$$

$$\text{Cramér's } V = 0.043$$

Smallest expected count 42.6

The two bars are almost identical. No sign that anxiety differs with tuition.

The printed percentages

The paper divides every cell by all 476 students: 47 is “(9.9)”. That cannot compare the rows. The row percentages in the bars are ours.

Islam MA et al. *PLoS One* 2020; PMID 32845928; CC BY. Counts from Table 2. Chi-square and row percentages ours.



What Table 2 adds, and a row that does not add up

Tuition and anxiety, three ways

	p
Table 3, as printed	.018
Table 3, rebuilt from its OR and CI (Session 4)	0.186
Table 2, our chi-square	0.825

Table 3 splits anxiety yes or no, in a logistic regression;
Table 2 keeps four levels. Different analyses, so they need
not match. Neither supports a link.

The depression rows

Tuition, minimal depression is printed as “44 (8.4)”.
But 8.4% of 476 is 40, not 44.

With 44, the tuition row totals 240. The anxiety
rows show 236 tutors. With 40, the depression row
gives 236 too.

The reader's habit

Add up the rows. Check each percentage against its count. It takes a minute.

Islam 2020, Tables 2 and 3, as printed. Rebuilt p , chi-square and the row sums are ours. The “40” is our reading; the paper does not say which
of the two figures is wrong.



Act V

Which test do I use?



Before any test: two questions

1. What kind of question?

- ▶ **Describe:** one variable
- ▶ **Compare:** an outcome across groups
- ▶ **Relate:** two variables together

Session 1's three kinds.

2. What kind of variables?

- ▶ Categorical, or numeric?
- ▶ If numeric: roughly symmetric, or skewed?
- ▶ The same people twice, or separate groups?

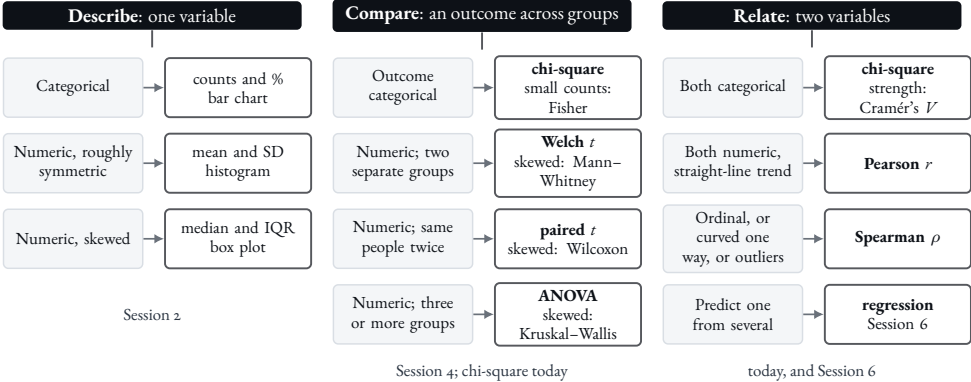
Session 1's scale table, and Session 2's shapes.

The test follows

Answer those two questions and the flowchart on the next frame leaves one or two choices.



Which test do I use?



The map for Sessions 2 to 6. Categorical includes yes or no answers; ordinal data use the rank-based partners.



Three questions through the flowchart

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; straight-line trend	Pearson r	$r = 0.45$

The same variable can play different parts

Year of study is a group in the second question. With five ordered years it could also be ordinal, for Spearman. The question decides.

Simulated class data. ANOVA from Session 4.



Five wrong choices, and what to use instead

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: a 26.5% chance of a false alarm	decide the question first

What they have in common

Each ignores what the numbers *are*: codes, percentages, pairs, ranks, or one test among many.

Simulated class data.



Activity: match the question to the test

Research question, from the class dataset

- 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?
-

In groups of three, for each question: (a) describe, compare or relate? (b) what kind is each variable? (c) which test, and what would you report beside p ?

Simulated class data. Answers, with every result, in the class notes.



What today established

1. Two categorical variables: compare observed counts with the counts independence predicts. That is chi-square. Residuals say where, Cramér's V how strongly.
2. Two numeric variables: r measures how closely they follow a straight line, from -1 to $+1$. Spearman's ρ does the same on ranks.
3. Plot first. Curves, outliers, codes and narrow samples all mislead r .
4. An association has four explanations: chance, a common cause, reverse causation, a real effect. A test rules out only the first.
5. The kind of question and the kind of variables choose the test.



Session 6: Regression and survey scales

What “controlling for” means, logistic regression for yes or no outcomes, and how to build and check a questionnaire scale.

Before then

Read Chapter 5 of the class notes. Try the Part A exercises and the activity answers.

The stress and screen-time figures use the **simulated** class data.

One question to think about

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?



- ▶ 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. CC BY 4.0.
- ▶ 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.
- ▶ Anscombe FJ. Graphs in statistical analysis. *Am Stat* 1973;27(1):17–21. doi:10.1080/00031305.1973.10478966.
- ▶ Messerli FH. Chocolate consumption, cognitive function, and Nobel laureates. *N Engl J Med* 2012;367(16):1562–1564. PMID 23050509. doi:10.1056/NEJMon1211064.
- ▶ 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.
- ▶ 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.
- ▶ Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;i(8476):307–310. PMID 2868172.
- ▶ Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.
- ▶ R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Datasets anscombe and UCBA admissions.
- ▶ Class data: **simulated** stand-in for the Stat.BD class survey, generated by data/class-data.R from a fixed seed. Not real students. Every figure computed from it, including the 1,000 shuffles and the rescaled tables, is ours. So are the chi-square and row percentages from Islam 2020, Table 2.

