Quantitative Research Methods

Chapter 0: Precourse Refresher

Prof. Dr. Christoph Weisser

HSBI

Summer Semester 2026

The course at a glance

	Lecture	Topic
▶	**0**	**Precourse (a) — refresher of undergraduate statistics, algebra, Python**
	0b	Precourse (b) — notation, logs and odds, likelihood, Python
	1	Ch. 1 + 2 — Introduction; what is statistical learning
	2	Ch. 2 — Accuracy, bias–variance, Bayes classifier, KNN
	3–4	Ch. 3 — Linear regression (estimation, inference, diagnostics)
	5–6	Ch. 4 — Classification (logistic, LDA/QDA, naive Bayes, ROC)
	7	Ch. 5 — Resampling (validation, CV, bootstrap)
	8	Ch. 6 — Model selection and regularisation (ridge, lasso)
	9	Ch. 7 — Moving beyond linearity (splines, GAMs)
	10	Ch. 8 — Tree-based methods (trees, forests, boosting)
	11	Ch. 10 — Deep learning (MLPs, CNNs, training)
	12	Ch. 13 — Multiple testing (FWER, FDR)

This refresher is *optional* and sits before Lecture 1. It assumes nothing beyond an introductory undergraduate course, and it only covers material the twelve lectures actually rely on. ▶ marks today's session.

Contents

Optional and advanced material is collected in the **appendix** at the end of the deck.

Notation in this refresher

Symbol	Meaning
n, p	number of observations; number of variables
x_i, y_i	predictor and response value of observation i
$\bar{x}$, s, s^2	sample mean, sample standard deviation, sample variance
μ, σ, σ^2	population mean, standard deviation, variance
Q_1, Q_3, IQR	lower/upper quartile; interquartile range $Q_3 - Q_1$
r	sample correlation coefficient
$\mathrm{Cov}(X, Y)$	covariance of two random variables
$\mathrm{E}[X]$, $\mathrm{Var}(X)$	expected value and variance of X
$\hat{\theta}$	an estimate of the unknown quantity θ ("hat" = estimated)
$\mathrm{SE}(\hat{\theta})$	standard error: the SD of $\hat{\theta}$ across repeated samples
H_0, H_1	null and alternative hypothesis
α	significance level (tolerated false-positive rate)
X, y	$n \times p$ data matrix; n-vector of responses
$\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$	vector of true / estimated regression coefficients
∇f	gradient: the vector of partial derivatives of f

Why a refresher, and how to use it

The purpose

This course does not teach statistics from scratch — it teaches *statistical learning*. Everything here is assumed known from Lecture 1 onwards. One session, so that nobody is stuck on notation while we discuss the ideas that matter.

How to use it

- If a slide is obvious to you, that topic is fine — move on.
- If a slide is new, work its exercise before Lecture 1.
- Each section ends with *where this reappears in the course*, so you can see why it is worth the effort.

Learning objectives for this refresher

By the end of this session you will be able to:

- **Classify** variables by type and choose an appropriate summary and plot for each.
- **Compute and interpret** mean, median, variance, SD, quartiles, IQR, covariance and correlation.
- **Apply** the rules of probability, conditional probability and Bayes' theorem, and use expectation and variance rules.
- **Recognise** the normal, t, χ^2, F, Bernoulli, binomial and Poisson distributions and say where each is used.
- **Explain** sampling variability, standard errors, confidence intervals, p-values and power — and their standard misreadings.
- **Fit and interpret** a simple linear regression by hand, and **read** a printed regression table column by column.
- **Manipulate** vectors and matrices, and read the matrix form of the regression estimator.
- **Differentiate** a loss function and take a gradient-descent step.
- **Use** `numpy`, `pandas` and `matplotlib` for the basic tasks of the labs.

Do you actually need this session? A 12-question self-check

Answer honestly. Each question maps to one section, so a wrong answer tells you where to read: 1–2 one variable; 3 two variables; 4 probability; 5–6 sampling; 7 and 11 testing and power; 8 and 12 regression; 9 linear algebra; 10 calculus.

1. Why is the median often reported for income, but the mean for test scores?
2. What does it mean that a variable has SD $= 0$?
3. Two variables have $r = 0$. Can they still be strongly related?
4. A test is 99% accurate and you test positive. Is there a 99% chance you are ill?
5. What is the difference between a standard deviation and a standard error?
6. What exactly is "95%" about a 95% confidence interval?
7. A p-value of 0.03: what is the probability that H_0 is true?
8. In $\hat{y} = 2 + 3x$, what does the 3 mean — in words, with units?
9. What are the dimensions of $\mathbf{X}^\top\mathbf{X}$ if $\mathbf{X}$ is $n \times p$?
10. What does the derivative of a loss function tell an optimisation algorithm to do?
11. A study finds "no significant effect". What third number do you need before believing it?
12. Adding one far-out point changes a fitted slope a lot. What is that point called?

Scoring

Comfortable with 9+? Skim this deck. Fewer than 6? Work through it with the exercises — it will pay for itself by Lecture 3.

Sources and references

Primary textbook (from Lecture 1 onwards)

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning, with Applications in Python.* Springer. Chapter 2 and Section 3.1 cover parts of this refresher in condensed form.

If you want a fuller undergraduate treatment

Any introductory statistics text will do; the topics to look up are listed on the final "where to read more" slide. Every figure in this deck was computed from the course data sets or from a clearly labelled simulation; the code that produces them sits in `make_figures.py` beside this file.

Where this chapter is used in industry

Sector	Concrete application	Which decision it drives
Manufacturing	Statistical process control: a chart on a critical dimension, limits at $\pm 3\sigma$	Stop the line, or let it keep running
Insurance	Claim severity is strongly right-skewed, so the mean sits above the typical claim	Where reinsurance attaches, and the reserve booked
E-commerce	Online A/B test of a checkout or ranking change	Ship the variant, or roll it back
Market research	Brand and satisfaction trackers on a sample of a few thousand	Act on a movement in the tracker, or treat it as noise
Supplier quality	Acceptance sampling of an incoming lot against an agreed plan	Accept the lot, or reject it on the sample
Logistics, operations	Delivery-time distributions summarised by the median and the 95th percentile	The service-level promise quoted to customers

In industry — The common thread

Every row is the same two-step move: *describe* what the data show, then quantify *how much of it could be noise*.

Industry case in depth: an online A/B test

A retailer wants to know whether a redesigned checkout page raises the purchase rate. This is a hypothesis test, run as an experiment.

In industry — The set-up

Response y_i: did session i end in a purchase (0/1)? **Unit of randomisation**: the visitor, not the session — otherwise the same person sees both designs. **Fixed before launch**: the primary metric, the segments to be reported (device, new vs. returning, market), the run length, and the smallest effect worth acting on.

Illustrative sizing

The rule of thumb for 80% power at $\alpha = 0.05$ is $n \approx 16\,\pi(1-\pi)/\delta^2$ per arm. At a baseline of $\pi = 2.5\%$, detecting a relative lift of 10% ($\delta = 0.0025$) needs tens of thousands of visitors per arm — which is why small sites cannot test small effects.

Who signs off — and the honest caveat

The product owner, plus a central experimentation platform that enforces the pre-registration. Its consumer is the launch reviewer who ships it. Caveat: stopping the test the first time it looks significant ("peeking"), or reporting whichever of twenty metrics moved, breaks the 5% guarantee — the problem of Chapter 13.

Outline

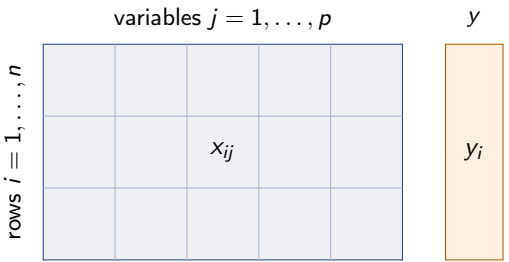
Roadmap of this refresher

How this session is organised

1. Data, variables and notation — the language.
2. Describing one variable, then two.
3. Probability, random variables and the standard distributions.
4. Sampling, estimation, confidence intervals, hypothesis tests.
5. Simple linear regression — the bridge into Chapter 3.
6. The linear algebra and calculus the later chapters use.
7. The Python toolkit of the labs.

Every idea is shown as a *picture* computed from the course data before it is written as a formula; every section says where it reappears in Chapters 1–13; and there is an exercise with a full solution roughly every 20 minutes.

The shape of a data set: n observations, p variables



X: the *data* (or design) matrix, $n \times p$

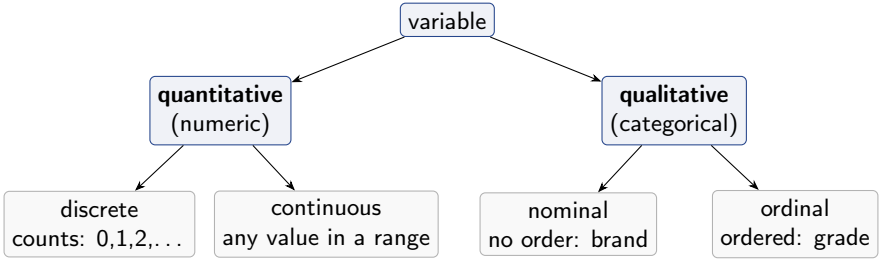
y: the response, one value per row

one *row* = one observation

Tidy data

One row per observational unit, one column per variable, one value per cell. Every method in this course expects data in this shape.

Variable types decide everything that follows



Type	Summarise with	Plot with
Quantitative	mean, SD, median, quartiles	histogram, boxplot, scatter
Qualitative	counts, proportions	bar chart, contingency table

In industry — Data engineering: ERP and CRM exports lie about types

Plant codes, cost centres and customer segments arrive as *numbers*. Feed them in unchanged and the model reads "plant 7 is seven times plant 1". Recode nominal fields as dummies at load time.

Why the type matters in this course

The single most consequential distinction

The *response* type names the problem:

- quantitative $y \Rightarrow$ **regression** (Ch. 3, 6, 7),
- qualitative $y \Rightarrow$ **classification** (Ch. 4, 8, 10).

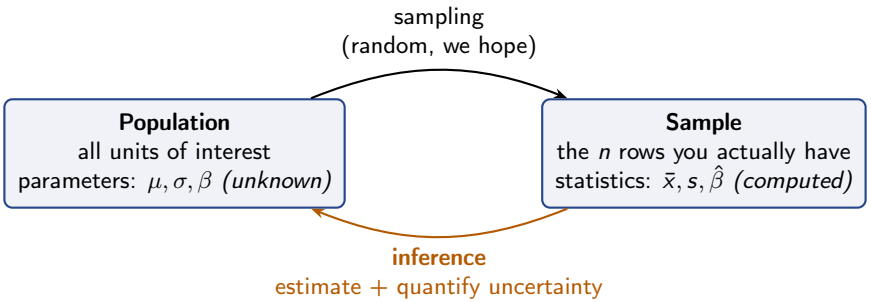
Qualitative *predictors* are not a problem

A categorical predictor enters as *dummy* (0/1) variables — one fewer than the number of categories, so "brand" with 3 levels costs 2 columns of X, not 1 (Chapter 3).

From the course data

`Wage`: `wage` is quantitative (regression, Ch. 7). `Default`: `default` is Yes/No (classification, Ch. 4). `Smarket`: `Direction` is Up/Down (classification, Ch. 4).

Population and sample: the loop this whole course lives in



Greek vs. Latin — and hats

Greek letters (μ, σ, β) denote unknown *population* quantities; Latin letters and hats ($\bar{x}$, s, $\hat{\beta}$) denote things *computed from data*. A hat always means "estimated".

Exercise 0.1 — Variable types and summaries [Concept]

Exercise

A university records, for each of n students: degree programme (Business, Engineering, Law); age in years; number of exam attempts; satisfaction on a 1–5 scale (1 = very poor, ..., 5 = very good); and starting salary in euros.

1. Classify each of the five variables as quantitative (discrete / continuous) or qualitative (nominal / ordinal).
2. For each, name one appropriate numerical summary and one plot.
3. The university wants to predict starting salary. Is this a regression or a classification problem? What if it instead wants to predict whether a student graduates on time?
4. Someone computes the mean of "degree programme" by coding Business = 1, Engineering = 2, Law = 3, and reports 1.8. What is wrong with this?

Solution — Exercise 0.1

Solution

(1) and (2).

Variable	Type	Summary	Plot
Degree programme	qualitative, nominal	counts / proportions	bar chart
Age	quantitative, continuous	mean and SD (or median, IQR)	histogram, boxplot
Exam attempts	quantitative, discrete	mean, or median (skewed)	bar chart of counts
Satisfaction	qualitative, ordinal	median, quartiles	bar chart
Starting salary	quantitative, continuous	*median* (right-skewed)	histogram, boxplot

(3). Salary is quantitative ⇒ **regression**. "Graduates on time" is Yes/No ⇒ **classification**. The data are the same; the *response type* names the problem.
(4). The codes 1, 2, 3 are arbitrary labels, not quantities: Law is not "three times" Business, and the gaps are not equal or even meaningful. The mean 1.8 therefore has no interpretation — it would change if the categories were relabelled. Summarise nominal variables by counts and proportions. (In a model, such a variable becomes dummy variables, not a single numeric column.)
Common mistake: treating the 1–5 satisfaction score as quantitative because it is stored as numbers. Its gaps are not equal units, so a mean misleads — order is all it has; use the median.

Outline

Three questions to ask of any single variable

Centre — spread — shape

1. **Centre**: where do typical values sit? (mean, median, mode)
2. **Spread**: how much do values differ? (SD, variance, range, IQR)
3. **Shape**: symmetric or skewed, one peak or several, outliers?

Never report a centre without a spread. "Average salary = €52,000" means something quite different with SD €3,000 than with SD €40,000.

Why the shape decides the summary

For symmetric data, mean and SD say it all. For skewed data (income, house prices, waiting times) the mean is dragged by the tail — report the median and IQR as well.

Measures of centre

Definitions

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \text{median} = \text{middle value of the sorted data}, \qquad \text{mode} = \text{most frequent value}.$$

Measure	Use when	Weakness
Mean $\bar{x}$	data roughly symmetric; algebra needed	pulled by outliers
Median	data skewed or outlier-prone	ignores magnitudes; harder algebra
Mode	categorical data; finding peaks	may not be unique or exist

The mean is the balance point

$\sum_i (x_i - \bar{x}) = 0$ always: deviations above and below the mean cancel exactly. This identity is why squared deviations, not raw ones, are used to measure spread — and it reappears as the "residuals sum to zero" property in Chapter 3.

Measures of spread

Sample variance and standard deviation

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad s = \sqrt{s^2}.$$

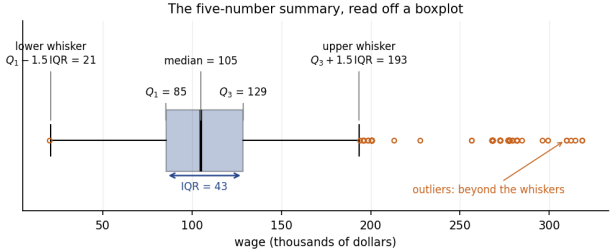
Reading it symbol by symbol

$(x_i - \bar{x})$ is how far observation i sits from the centre; squaring removes the sign and penalises large deviations; summing aggregates; dividing averages. The square root returns to the original units — which is why s, not s^2, is the number you report.

Why $n-1$ and not n?

One degree of freedom was spent estimating $\bar{x}$ from the same data; dividing by $n-1$ makes s^2 unbiased for σ^2. The same bookkeeping returns in the t- and F-tests and in the effective-degrees-of-freedom arguments of Chapters 5–7.

Quantiles, the five-number summary, and the boxplot

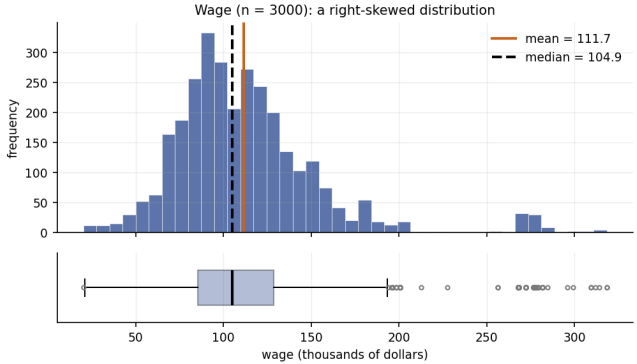


The q-quantile is the value below which a fraction q of the data lies: $Q_1 = 25\%$, median $= 50\%$, $Q_3 = 75\%$, and $\mathrm{IQR} = Q_3 - Q_1$ holds the middle half. The usual rule flags points outside $[\,Q_1 - 1.5\,\mathrm{IQR},\ Q_3 + 1.5\,\mathrm{IQR}\,]$ — a convention, not a law, and never a licence to delete data. Careful on small samples: by hand we take the median of each half, so Q_1 and Q_3 are data values, while `numpy` and `pandas` interpolate between neighbours and can differ by half a unit.

Companion notebook

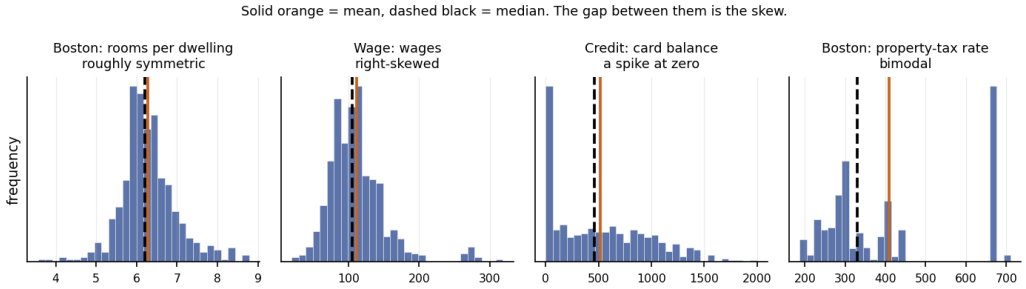
Rebuild this plot yourself in `Lab_Notebooks/chapter_00_lab.ipynb`, §2 *Describing one variable: centre, spread, shape*, with `df["wage"].describe()`.

Shape: skewness read off a picture



`Wage` is right-skewed: a long upper tail pulls the mean (111.7) above the median (104.9). Rule of thumb: mean $>$ median suggests right skew, mean $<$ median left skew.

Four shapes you will actually meet

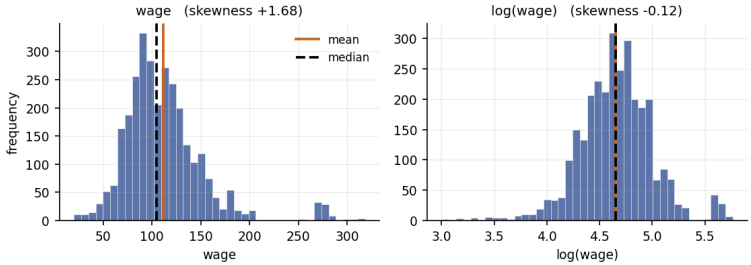


Symmetric — mean and median agree; report either. **Skewed** — they separate; the median is the honest summary. **A spike at one value** — 22.5% of `Credit` customers carry a zero balance; no single number describes this well, so say so. **Bimodal** — two peaks usually mean two *populations* mixed together; find the variable that separates them.

Shape is a modelling decision

A histogram is not decoration: it tells you whether a mean is meaningful, whether a transformation is needed, and whether a group variable is missing from your model.

When a transformation helps



Taking logs pulls in the long right tail: skewness falls from $+1.68$ to -0.12, and mean and median almost coincide.

Why this matters later

Logs turn multiplicative effects into additive ones, so a coefficient becomes a *percentage* change. `Wage` ships with a `logwage` column for exactly this reason. Logs need strictly positive data — and predictions must be transformed back before anyone reads them.

Standardisation: the z-score

Definition

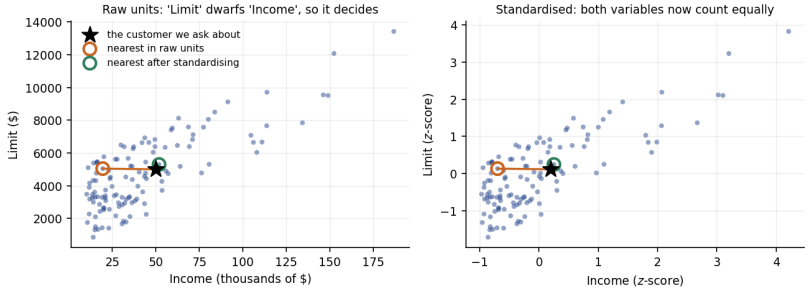
$$z_i = \frac{x_i - \bar{x}}{s} \qquad \Longrightarrow \qquad \bar{z} = 0, \quad s_z = 1.$$

Why you will meet this constantly

A z-score says "how many standard deviations from the mean", so it is unit-free and comparable across variables. It matters in:

- **Ch. 6** — ridge and lasso penalise coefficients, so predictors *must* be standardised first, or the penalty depends on whether you measured in euros or thousands of euros;
- **Ch. 2, 12** — KNN and PCA use distances, which are dominated by whichever variable has the largest scale;
- **Ch. 10** — neural networks train far better on standardised inputs.

Standardisation in action: who is my nearest neighbour?



`Credit`: income is in thousands, the limit in single dollars, so on raw units the limit decides every distance by itself. After standardising, both count equally — and a different customer is nearest.

The rule

Any method built on distances or on penalised coefficients — KNN, PCA, clustering, ridge, lasso, neural networks — must see standardised inputs, or the answer depends on your choice of units.

Worked example: a seven-point sample by hand

Seven observations computed by hand

The data: 2, 4, 4, 5, 7, 8, 12 (so $n = 7$).

Centre. $\sum x_i = 42$, so $\bar{x} = 42/7 = 6$. Sorted, the middle (4th) value is the median $= 5$. The mode is 4.

Spread. Deviations $x_i - \bar{x}$: $-4, -2, -2, -1, 1, 2, 6$; squares $16, 4, 4, 1, 1, 4, 36$, summing to 66. Hence

$$s^2 = \frac{66}{7-1} = 11, \qquad s = \sqrt{11} \approx 3.32.$$

Quartiles. Lower half $\{2, 4, 4\}$ gives $Q_1 = 4$; upper half $\{7, 8, 12\}$ gives $Q_3 = 8$; $\mathrm{IQR} = 4$. Fences: $4 - 6 = -2$ and $8 + 6 = 14$, so *no* outliers — 12 is large but not flagged.

Shape. $\bar{x} = 6 > 5 =$ median: mildly right-skewed, as the single large value 12 suggests.

Exercise 0.2 — Summaries and the effect of one outlier [Math]

Exercise

Seven service times (minutes) were recorded:

$$3, \quad 5, \quad 6, \quad 6, \quad 8, \quad 9, \quad 21.$$

1. Compute the mean, the median and the mode.
2. Compute s^2 and s. (Work with deviations from the mean.)
3. Compute Q_1, Q_3 and the IQR, and apply the $1.5 \times \mathrm{IQR}$ rule. Is 21 an outlier?
4. The 21 was a typo for 12. Recompute the mean and the median. Which moved more, and what general lesson does that illustrate?

Solution — Exercise 0.2

Solution

(1) Centre. $\sum x_i = 3+5+6+6+8+9+21 = 58$, so $\bar{x} = 58/7 \approx 8.29$. The 4th of seven sorted values is the median $= 6$. The mode is 6.
(2) Spread. Deviations from 8.29 are -5.29, -3.29, -2.29, -2.29, -0.29, 0.71, 12.71; their squares 27.9, 10.8, 5.2, 5.2, 0.1, 0.5, 161.6 sum to 211.4. Hence $s^2 = 211.4/6 \approx 35.2$ and $s \approx 5.94$ minutes.
(3) Quartiles. Lower half $\{3,5,6\} \Rightarrow Q_1 = 5$; upper half $\{8,9,21\} \Rightarrow Q_3 = 9$; $\mathrm{IQR} = 4$. Upper fence $= 9 + 1.5(4) = 15 < 21$, so **yes**, 21 is flagged as an outlier. (Flagged means *investigate*, not *delete*.)
(4) After the correction. $\sum x_i = 49$, so $\bar{x} = 7$ — it fell by 1.29. The median is still 6: unchanged. The mean is *sensitive* to extreme values, the median is *resistant*. This is exactly why skewed variables (income, waiting times, house prices) are reported with the median, and why a single mistyped value can move a least-squares fit — squared errors magnify exactly this sensitivity (Chapter 3).

Common mistake

Dividing by n instead of $n-1$ in s^2 — here $211.4/7 \approx 30.2$ instead of 35.2. The $n-1$ pays for estimating $\bar{x}$ from the same data.

Outline

Covariance: do two variables move together?

Sample covariance

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Reading the sign

Each term is positive when x_i and y_i are on the *same* side of their means, negative otherwise. A positive sum means the variables tend to rise together; a negative sum means one rises as the other falls.

The problem with covariance

Its size depends on units: measuring x in cents rather than euros multiplies s_{xy} by 100, though nothing about the relationship changed. So we rescale it — and get the correlation.

Correlation: covariance made unit-free

Pearson correlation

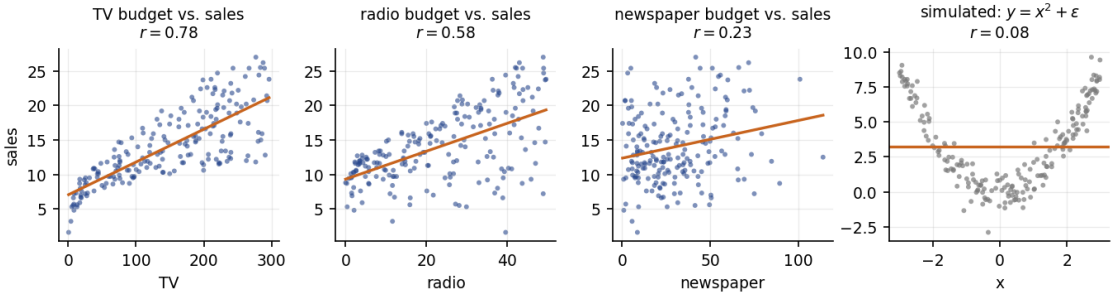
$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\,\sqrt{\sum_i (y_i - \bar{y})^2}}, \qquad -1 \le r \le 1.$$

Value	Meaning
$r = +1$ / -1	points lie exactly on an increasing / decreasing straight line
$r \approx 0$	no *linear* association (there may still be a strong curved one)
$\lvert r \rvert$ near 1	a tight linear pattern; near 0, a diffuse cloud

Three warnings

r measures *linear* association only; it is not resistant to outliers; and it says nothing about *causation* or about the *slope* of the relationship.

What r sees — and what it misses

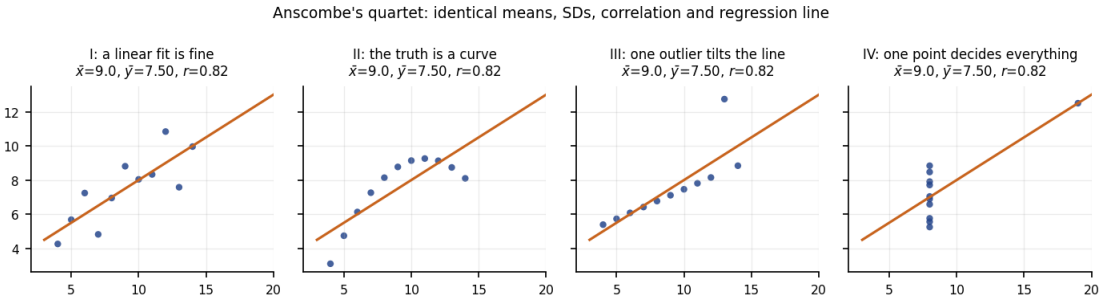


The first three panels use `Advertising`: TV spend tracks sales closely ($r = 0.78$), newspaper spend barely ($r = 0.23$). The fourth is simulated data with an obvious, *strong* relationship $y = x^2 + \varepsilon$ — and $r = 0.08$. **Always plot the data.**

In industry — KPI dashboards: the correlation-matrix trap

A threshold or curved effect reads as $r \approx 0$, and a shared driver — season, a price change, a campaign — makes unrelated KPIs move together.

Anscombe's quartet: four data sets, one set of statistics



All four have the same $\bar{x}$, the same $\bar{y}$, the same standard deviations, the same $r = 0.82$ and the *same* fitted line — yet only the first deserves one. Constructed by F. J. Anscombe (1973) to make exactly this point.

The habit this should build

Summary statistics compress, and compression loses information. Plot first, summarise second, model third.

Correlation is not causation — the two classic traps

Confounding

Ice-cream sales and drowning deaths correlate strongly. Neither causes the other: *temperature* drives both. A variable that influences both X and Y is a **confounder**. Chapter 3 shows how adding it as a predictor can make a spurious effect disappear.

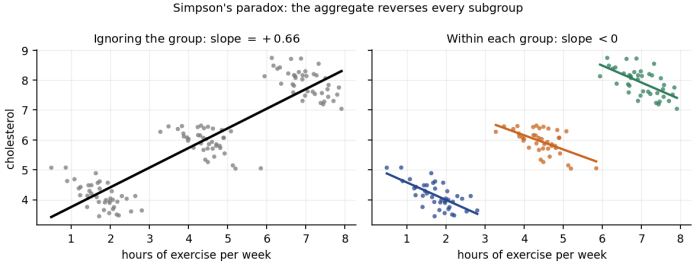
Simpson's paradox

An association can *reverse* once you condition on a group: a treatment can look worse overall yet be better in every subgroup, if the sicker patients were more likely to receive it.

The habit to build

Before believing any two-variable claim, ask: *what third variable could produce this pattern?* This question is the entire reason multiple regression exists.

Simpson's paradox, drawn



Simulated data for three age groups: older people both exercise more *and* have higher cholesterol. Ignore the group and exercise looks harmful (+0.66); within every group it helps.

The lesson

The overall and the within-group slope answer different questions. Adding the group variable — what multiple regression does — recovers the within-group answer; which one you *want* depends on the decision.

Two categorical variables: the contingency table

A study of 200 customers cross-tabulates a promotion against purchase:

	Bought	**Did not buy**	**Total**
Received promotion	45	55	100
No promotion	25	75	100
Total	70	130	200

- **Row proportions** are what you compare: 45% bought with the promotion vs. 25% without.
- If the variables were independent, the expected count in a cell is (row total $\times$ column total)$/n$: here $100 \times 70/200 = 35$ for the top-left cell. Observed 45 exceeds it.
- A χ^2 test formalises whether that gap is bigger than chance.

Where this returns

This table is a **confusion matrix** in disguise. In Chapter 4 the rows become "true class" and the columns "predicted class", and the same proportions become sensitivity and specificity.

Exercise 0.3 — Covariance, correlation and a slope [Math]

Exercise

Five stores report advertising spend x (thousands of euros) and sales y (thousands of units):

x	1	2	3	4	5
y	2	4	5	4	10

1. Compute $\bar{x}$ and $\bar{y}$.
2. Compute $S_{xx} = \sum(x_i - \bar{x})^2$, $S_{yy} = \sum(y_i - \bar{y})^2$ and $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$.
3. Compute r. Describe the association in one sentence.
4. The least-squares slope is $b_1 = S_{xy}/S_{xx}$. Compute it and state, in words and units, what it means.
5. Why do r and b_1 differ, even though both measure "how y moves with x"?

Solution — Exercise 0.3

Solution

(1) $\bar{x} = 15/5 = 3, \quad \bar{y} = 25/5 = 5.$

(2) Deviations: $x_i - \bar{x} = -2, -1, 0, 1, 2$ and $y_i - \bar{y} = -3, -1, 0, -1, 5$.

$$S_{xx} = 4 + 1 + 0 + 1 + 4 = 10, \quad S_{yy} = 9 + 1 + 0 + 1 + 25 = 36, \quad S_{xy} = 6 + 1 + 0 - 1 + 10 = 16.$$

(3) $r = \dfrac{16}{\sqrt{10 \cdot 36}} = \dfrac{16}{18.97} \approx 0.84$: a strong positive linear association — stores that spend more on advertising tend to sell more.

(4) $b_1 = 16/10 = 1.6$. *One extra thousand euros of advertising is associated with* 1.6 *thousand additional units sold*, on average, in this sample. Note "associated with", not "causes".

(5) They answer different questions. r is unit-free and bounded in $[-1, 1]$: it measures how *tightly* the points hug a line. b_1 carries units (units per thousand euros): it measures how *steeply* y changes with x. Formally $b_1 = r\, s_y / s_x$ — the same information, rescaled. A steep slope can be badly estimated (small $|r|$), and a tight relationship can have a nearly flat slope.

Common mistake: reading $r = 0.84$ as "84% of the variance explained" — that is $r^2 = 0.71$ — or as the slope; r is unit-free, b_1 is not.

Outline

The rules you actually need

Basics

$0 \le P(A) \le 1; \quad P(\text{sure event}) = 1; \quad P(A^c) = 1 - P(A).$

Addition and multiplication

$$P(A \cup B) = P(A) + P(B) - P(A \cap B), \qquad P(A \cap B) = P(A \mid B)\, P(B).$$

Conditional probability and independence

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad (P(B) > 0); \qquad A \perp B \iff P(A \mid B) = P(A).$$

Conditioning is the whole game

Every classifier in this course estimates a conditional probability: $P(Y = \text{default} \mid X = \text{balance})$. "Learning" means turning data into a conditional probability you can act on.

Bayes' theorem: reversing the conditioning

The theorem

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}, \qquad P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c).$$

What it is for

You know $P(\text{evidence} \mid \text{cause})$ — how the test behaves for a sick person. You want $P(\text{cause} \mid \text{evidence})$ — whether *this* positive test means illness. Bayes' theorem converts one into the other, and the conversion depends on the **base rate** $P(A)$.

Where it reappears

Chapter 4 is Bayes' theorem applied three times over: LDA, QDA and naive Bayes all recover $P(Y = k \mid X = x)$ from $P(X = x \mid Y = k)$ and the prior $\pi_k = P(Y = k)$.

Worked example: why a 95% accurate test can mislead

Screening for a rare condition

Prevalence $P(D) = 1\%$; sensitivity $P(+ \mid D) = 90\%$; specificity $P(- \mid D^c) = 95\%$, so $P(+ \mid D^c) = 5\%$. You test positive.

$$P(+) = \underbrace{0.90 \times 0.01}_{\text{true positives } = 0.009} + \underbrace{0.05 \times 0.99}_{\text{false positives } = 0.0495} = 0.0585.$$

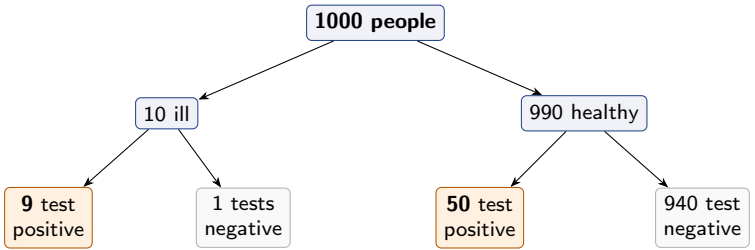
$$P(D \mid +) = \frac{0.009}{0.0585} \approx 0.154.$$

Only about **15%**: the healthy group is 99 times larger, so its 5% false-positive rate produces five times more positives than the ill group does.

The lesson for Chapter 4

With rare events, high accuracy coexists with mostly-wrong positives — which is why Chapter 4 reports sensitivity, specificity and ROC curves rather than one accuracy number.

The same calculation, without any algebra



$9 + 50 = 59$ positive tests, of which 9 are truly ill: $9/59 \approx 15\%$

Natural frequencies beat percentages

The same calculation becomes obvious once it is counts of people rather than conditional probabilities — a trick worth using whenever a probability statement feels slippery.

Random variables, expectation and variance

Definitions (discrete case)

$$\mathrm{E}[X] = \sum_x x\,P(X = x), \qquad \mathrm{Var}(X) = \mathrm{E}\big[(X - \mathrm{E}[X])^2\big] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2.$$

Rules you will use constantly

$$\mathrm{E}[aX + b] = a\,\mathrm{E}[X] + b, \qquad \mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X),$$

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y] \text{ (always)}, \qquad \mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \textit{ only if } X \perp Y.$$

Read the rules carefully

Expectation is linear *without conditions*. Variance is not: adding a constant does not change spread (b vanishes), scaling squares it (a^2), and variances add only under independence. The bias–variance decomposition in Chapter 2 is these rules applied to $\hat{f}(x)$.

Exercise 0.4 — Conditional probability and Bayes [Math]

Exercise

A bank's fraud filter flags transactions. Historically 0.5% of transactions are fraudulent. The filter flags 95% of fraudulent transactions, and also flags 2% of legitimate ones.

1. What proportion of *all* transactions get flagged?
2. A transaction is flagged. What is the probability that it is actually fraudulent?
3. Out of 100,000 transactions, fill in the four counts (fraud or not $\times$ flagged or not). Which cell dominates the flagged column?
4. The bank wants $P(\text{fraud} \mid \text{flag})$ above 50%. Holding the 95% detection rate fixed, what false-positive rate would be required? Comment on whether that is realistic.

Solution — Exercise 0.4

Solution

(1) With $P(F) = 0.005$: $P(\text{flag}) = 0.95(0.005) + 0.02(0.995) = 0.00475 + 0.0199 = 0.02465 \approx 2.5\%$.
(2) $P(F \mid \text{flag}) = 0.00475/0.02465 \approx$ **19%**: four of five flags are false alarms, despite a 95% detection rate.
(3) Per 100,000 transactions:

	Flagged	**Not flagged**	**Total**
Fraudulent	475	25	500
Legitimate	1,990	97,510	99,500
Total	2,465	97,535	100,000

The 1,990 false positives dominate: the legitimate group is 199 times larger, so even a small error rate on it swamps the true positives.
(4) $P(F \mid \text{flag}) \geq 0.5$ means the flagged fraud mass must outweigh the flagged legitimate mass: $0.00475 \geq f \cdot 0.995$, i.e. $f \leq$ **0.48%** — four times stricter than today's 2% and reachable only by flagging less aggressively, which costs detection: the ROC trade-off of Chapter 4.

Common mistake: answering (2) with "95%" — that is $P(\text{flag} \mid F)$, not $P(F \mid \text{flag})$. The tiny base rate drags the useful direction down to 19%.

Exercise 0.5 — Expectation and variance rules [Math]

Exercise

A machine fills bottles. The fill volume X has $\mathrm{E}[X] = 500\,\mathrm{ml}$ and $\mathrm{Var}(X) = 16\,\mathrm{ml}^2$.

1. A calibration error adds exactly 3 ml to every bottle. Give the new mean and the new variance of $Y = X + 3$.
2. The volume is re-expressed in litres, $W = X/1000$. Give $\mathrm{E}[W]$, $\mathrm{Var}(W)$ and $\mathrm{SD}(W)$.
3. A crate holds 4 bottles filled *independently*. Give the mean and variance of the total $T = X_1 + X_2 + X_3 + X_4$, and of the average $\bar{X} = T/4$.
4. Standardise: what are the mean and SD of $Z = (X - 500)/4$?
5. Suppose instead the four bottles come from one shared reservoir, so their volumes are positively correlated. Which of your answers in (3) breaks, and in which direction?

Solution — Exercise 0.5

Solution

(1) $\mathrm{E}[Y] = 503\,\text{ml}$; $\mathrm{Var}(Y) = 16$ — unchanged. Shifting every value moves the centre but not the spread (b drops out of $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$).
(2) $\mathrm{E}[W] = 0.5\,\text{l}$; $\mathrm{Var}(W) = (1/1000)^2 \cdot 16 = 1.6 \times 10^{-5}\,\text{l}^2$; $\mathrm{SD}(W) = 4/1000 = 0.004\,\text{l}$. Note the SD rescales linearly while the variance rescales by the *square* — which is exactly why we report the SD.
(3) $\mathrm{E}[T] = 4(500) = 2000\,\text{ml}$ and, *by independence*, $\mathrm{Var}(T) = 4(16) = 64$ (SD $= 8$). For the average, $\mathrm{E}[\bar{X}] = 500$ and $\mathrm{Var}(\bar{X}) = \mathrm{Var}(T)/16 = 4$, so $\mathrm{SD}(\bar{X}) = 2 = 4/\sqrt{4}$ — the $\sigma/\sqrt{n}$ rule, derived rather than asserted.
(4) $\mathrm{E}[Z] = 0$ and $\mathrm{SD}(Z) = 1$: standardising is just the $a = 1/4$, $b = -125$ case of the same two rules.
(5) Only the *variances* in (3) break — the means are unaffected, because expectation is linear whatever the dependence. With positive correlation, $\mathrm{Var}(T) = \sum_i \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j) > 64$: the true spread is *larger* than the independence formula suggests, so treating dependent observations as independent makes us overconfident. That is the same mistake as using random cross-validation folds on a time series (Chapter 5).
Common mistake: $\mathrm{Var}(X - Y) = \mathrm{Var}(X) - \mathrm{Var}(Y)$. Wrong: under independence variances *add* for sums and differences alike — the minus sign is squared away.

Outline

The discrete trio

Bernoulli(π) — one yes/no trial

$P(X = 1) = \pi$, $P(X = 0) = 1 - \pi$; $\mathrm{E}[X] = \pi$, $\mathrm{Var}(X) = \pi(1 - \pi)$. *Where:* the response in every classification problem (Ch. 4, 8, 10).

Binomial(n, π) — number of successes in n independent trials

$P(X = k) = \binom{n}{k}\pi^k(1 - \pi)^{n-k}$; $\mathrm{E}[X] = n\pi$, $\mathrm{Var}(X) = n\pi(1 - \pi)$. *Where:* counting correct classifications; the bootstrap (Ch. 5).

Poisson(λ) — counts per unit of time or space

$P(X = k) = e^{-\lambda}\lambda^k/k!$; $\mathrm{E}[X] = \mathrm{Var}(X) = \lambda$. *Where:* Poisson regression on `Bikeshare` (Ch. 4).

Note the variance of a Bernoulli

$\pi(1 - \pi)$ is largest at $\pi = 0.5$ and vanishes at 0 or 1: a near-certain event is nearly noiseless. This is why classification is hardest near the decision boundary.

The normal distribution

$X \sim N(\mu, \sigma^2)$

Symmetric and bell-shaped, fully determined by μ and σ. Standardising gives $Z = (X - \mu)/\sigma \sim N(0, 1)$.

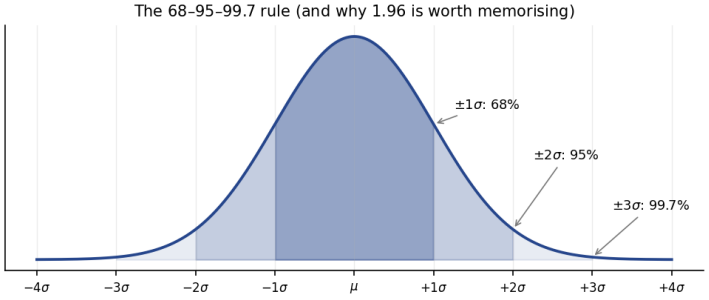
The 68–95–99.7 rule

About 68% of the mass lies within 1σ of μ, 95% within 2σ (more precisely 1.96σ), and 99.7% within 3σ. Memorising 1.96 saves you a table lookup for the rest of the course.

What is and is not assumed

Regression does *not* require X or y to be normal. Normality is assumed for the *errors* — and even then mainly to justify t- and F-tests in small samples. With large n the CLT does the work instead.

The 68–95–99.7 rule, drawn

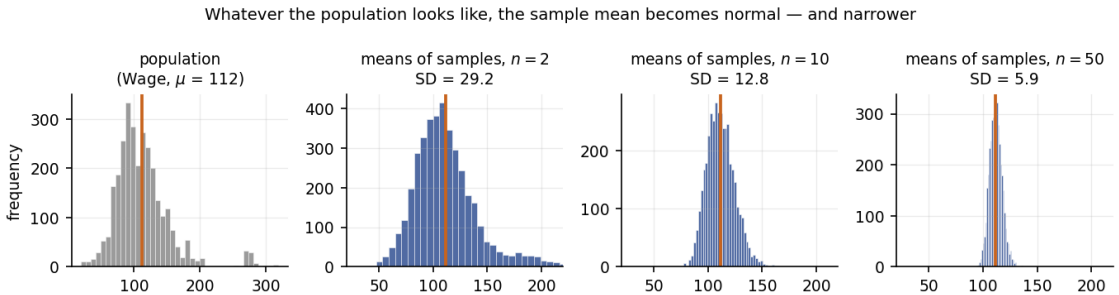


Two consequences you will use every week: an observation more than two SDs from the mean is unusual (about 1 in 20), and "estimate $\pm 1.96\,\mathrm{SE}$" is a 95% interval. That same 1.96 appears in every confidence interval and every two-sided 5% test in this course.

In industry — Manufacturing: the arithmetic behind Six Sigma

Process-control charts set limits at $\pm 3\sigma$: an in-control process rarely trips the alarm, while a real shift in the process trips it fast.

The central limit theorem



Statement

For a random sample of size n from *any* population with mean μ and finite variance σ^2, $\bar{X} \approx N(\mu, \sigma^2/n)$ for large n. The spread of $\bar{X}$ shrinks like $1/\sqrt{n}$ — quadrupling the sample halves the uncertainty.

Outline

Standard deviation vs. standard error

The distinction students most often blur

- The **standard deviation** s describes the spread of the *data*; more data does not shrink it.
- The **standard error** $\mathrm{SE}(\bar{x}) = s/\sqrt{n}$ describes the spread of an *estimate* over repeated samples; more data *does* shrink it.

Numbers

Wages with $s = 40$ and $n = 100$: $\mathrm{SE}(\bar{x}) = 40/10 = 4$. With $n = 1600$: $\mathrm{SE} = 40/40 = 1$. The people are just as variable as before; our estimate of their average is four times more precise.

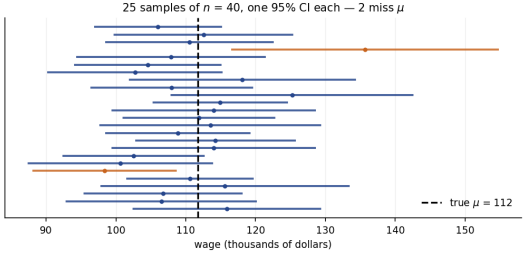
The habit

Report every estimate — a coefficient, an error rate, a CV score — *with* its standard error. A number without an uncertainty is not a result.

Confidence intervals

The usual interval for a mean

$$\bar{x} \pm t_{n-1,\,0.975} \cdot \frac{s}{\sqrt{n}} \qquad \left(\approx \bar{x} \pm 1.96\,\mathrm{SE} \text{ for large } n\right).$$



Each horizontal bar is one 95% CI from one sample. Two of the 25 miss the true μ. With 95% coverage we expect about one miss in 25, so two is an ordinary draw — the promise is about the long run, not any one batch.

What a confidence interval does and does not say

The correct reading

"95%" is a property of the *procedure*: repeat the whole sampling-and-computing exercise many times, and about 95% of the intervals produced would contain the true parameter.

Three wrong readings

1. "95% probability that μ lies in *this* interval." — μ is fixed; this interval either contains it or not.
2. "95% of the data lie in the interval." — no: that is a *prediction* interval, which is far wider.
3. "It contains 0, so the effect is zero." — no: the data cannot rule zero out. Absence of evidence is not evidence of absence.

Width tells you as much as position

Wide interval around a large estimate: "possibly big — we cannot tell". Narrow interval around a small one: "genuinely small". You will read these intervals straight off regression output in Chapter 3 and rebuild them without any formula from bootstrap replicates in Chapter 5.

Outline

The logic of a test in five steps

The idea in one sentence

A test never proves an effect exists. It asks: *if there were no effect at all, how surprising would data like ours be?* If the answer is "very surprising", we abandon the no-effect story. Everything below is bookkeeping for that one question.

1. State H_0 (the sceptical default: "no effect", $\beta_j = 0$) and H_1.
2. Choose α, the false-positive rate you will tolerate (conventionally 0.05).
3. Compute a **test statistic**: the estimate, measured in standard errors, e.g. $t = (\hat{\theta} - \theta_0)/\mathrm{SE}(\hat{\theta})$.
4. Compute the ***p*-value**: the probability, *if H_0 were true*, of a statistic at least this extreme.
5. Reject H_0 if $p < \alpha$; otherwise fail to reject. Then report the *estimate and its CI*, not just the verdict.

The test statistic is a signal-to-noise ratio

t asks: how large is the effect *relative to how precisely we measured it*? Every *t*-value in every regression output in this course is exactly this ratio.

Two ways to be wrong

	H_0 is true	H_0 is false
Reject H_0	Type I error (prob. α)	correct (power $= 1 - \beta$)
Do not reject	correct	Type II error (prob. β)

- In this table β is the *Type II error rate* — the one place in this course where β is not a regression coefficient.
- Lowering α buys fewer false positives at the cost of more false negatives. There is no setting that avoids both.
- **Power** rises with the true effect size, with n, and with lower noise. Underpowered studies mostly produce misleading results.

The same table under different names

In Chapter 4 this becomes the confusion matrix (false positive / false negative); in Chapter 13 the columns are relabelled to define the family-wise error rate and the false discovery rate. It is one idea wearing three costumes.

What a p-value is not

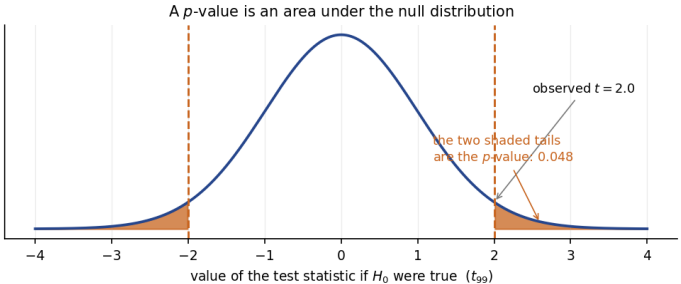
Four standard misreadings

1. *"$p = 0.03$ means a 3% chance that H_0 is true."* No: it is a probability computed *assuming* H_0 — it says nothing about $P(H_0 \mid \text{data})$.
2. *"$p > 0.05$ proves no effect."* No: it may only mean the sample was too small to detect one.
3. *"$p = 0.001$ means a big effect."* No: with huge n, trivial effects get tiny p-values. Significance $\neq$ importance.
4. *"I tested 20 things and one was significant at 5%."* Expected by chance alone — see Chapter 13.

The professional habit

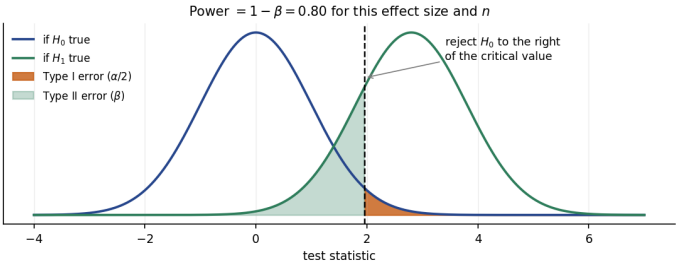
Report the estimate, its confidence interval, and the p-value — in that order of importance. Effect size first, uncertainty second, verdict last.

A p-value is an area



Draw the distribution the statistic *would* have if H_0 were true, mark what you actually observed, and shade everything at least that extreme. The shaded area is the p-value: 0.048 here, for $t = 2.0$ on 99 degrees of freedom. Both tails count because the alternative is two-sided — a deviation in either direction would have surprised us.

Power: the error nobody reports



Moving the critical value to the right shrinks the Type I error area (upper tail, $\alpha/2$) but grows the Type II error area (β). Only *more data* — or a larger true effect, or less noise — shrinks both, by pushing the curves apart.

Why underpowered studies mislead

With low power most true effects are missed, and the few that do reach significance are the ones that got lucky — so their estimated effects are *too large*. This is a large part of why published findings fail to replicate.

Worked example: a one-sample t-test

Has average handling time changed?

Historically a process took $\mu_0 = 100$ seconds. A sample of $n = 100$ recent cases gives $\bar{x} = 103$ and $s = 15$.

$$\mathrm{SE}(\bar{x}) = \frac{15}{\sqrt{100}} = 1.5, \qquad t = \frac{103 - 100}{1.5} = 2.00 \quad (\mathrm{df} = 99).$$

Two-sided $p \approx 0.048 < 0.05$: reject H_0 at the 5% level.
The more useful statement. The 95% CI is $103 \pm 1.98(1.5) = [100.03,\ 105.97]$: the increase is somewhere between 0.03 and 6 seconds — detectable, yet possibly far too small to matter operationally.

CI and test agree by construction

A two-sided test rejects at level α exactly when the $(1 - \alpha)$ CI excludes the null value. The CI simply says more.

Exercise 0.6 — Reading an inference [Concept]

Exercise

A colleague reports: "We estimated the effect of a training programme on productivity. The estimate is $+2.4$ units with standard error 1.5, giving $t = 1.6$ and $p = 0.11$. We conclude the programme has no effect."

1. Compute the approximate 95% confidence interval.
2. Is the conclusion "no effect" justified? Answer using the interval.
3. The team repeats the study with four times the sample size and obtains the same estimate $+2.4$. What is the new standard error (assuming the same variability), and the new t?
4. In a second study of 40 outcome variables, three came out significant at the 5% level. Roughly how many would you expect by chance if none of the 40 had any effect?

Solution — Exercise 0.6

Solution

(1) $2.4 \pm 1.96(1.5) = 2.4 \pm 2.94 = [-0.54,\ 5.34]$.
(2) No. The interval does contain 0, so a zero effect cannot be ruled out — but it also contains $+5.3$, a substantial benefit. The correct statement is "this study cannot distinguish between no effect and a sizeable positive effect": the data are *inconclusive*, not evidence of absence. Failing to reject H_0 is never the same as accepting it.
(3) The standard error scales as $1/\sqrt{n}$, so four times the data halves it: $\mathrm{SE} = 0.75$, and $t = 2.4/0.75 = 3.2$ ($p \approx 0.001$). The new CI is $2.4 \pm 1.47 = [0.93,\ 3.87]$ — same estimate, now clearly non-zero. *Significance is as much about sample size as about the effect.*
(4) With 40 true nulls and $\alpha = 0.05$, we expect $40 \times 0.05 = \mathbf{2}$ false positives, and $P(\text{at least one}) = 1 - 0.95^{40} \approx 0.87$. Three "discoveries" is therefore unremarkable. Correcting for this is the entire content of Chapter 13 (Bonferroni, Holm, Benjamini–Hochberg).

Common mistake

Expecting double the data to halve the SE — it shrinks like $1/\sqrt{n}$, so halving the SE costs *four* times the data, exactly as in (3).

Extended Exercise 0.1 — From data to a defensible claim [Math]

Extended exercise (15 min)

A retailer measures the basket value (euros) of $n = 64$ customers after a store redesign. The sample gives $\bar{x} = 48.5$ and $s = 12.0$. The pre-redesign average was $\mu_0 = 45.0$.

1. Compute $\mathrm{SE}(\bar{x})$ and a 95% confidence interval for the true post-redesign mean (use 1.96).
2. Test $H_0 : \mu = 45$ against $H_1 : \mu \neq 45$ at $\alpha = 0.05$. Report t and your decision, and state how the CI would have told you the same thing.
3. Management says an increase is worth acting on only if it exceeds €5. Does the evidence support acting? Answer using the interval, not the p-value.
4. How large would n have to be for the CI half-width to shrink to €1.00 (same s)?
5. The redesign coincided with a marketing campaign. What does that do to the causal interpretation, and what design would fix it?

Solution — Extended Exercise 0.1 (1/2)

Solution

(1) Standard error and interval.

$$\mathrm{SE}(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{12}{\sqrt{64}} = \frac{12}{8} = 1.5.$$

$$95\%\ \mathrm{CI} = 48.5 \pm 1.96(1.5) = 48.5 \pm 2.94 = [45.56,\ 51.44].$$

(2) The test.

$$t = \frac{\bar{x} - \mu_0}{\mathrm{SE}} = \frac{48.5 - 45.0}{1.5} = \frac{3.5}{1.5} = 2.33 \quad (\mathrm{df} = 63).$$

Two-sided $p \approx 0.023 < 0.05$, so we reject H_0: the mean basket value differs from €45. The CI says the same thing without a separate calculation — it excludes 45. This equivalence is exact for a two-sided test at the matching level.

Solution — Extended Exercise 0.1 (2/2)

Solution

(3) The decision that management actually asked about. The relevant question is not "is the increase non-zero" but "is it above €5". The increase is estimated at $48.5 - 45.0 = €3.50$, with CI $[45.56, 51.44] - 45 = [0.56,\ 6.44]$. The interval straddles 5: the data are consistent with an increase of €1 and with one of €6. So the evidence does *not* support acting — and note that the p-value alone (0.023, "significant") would have suggested the opposite. *Statistical significance answered the wrong question.*

(4) Required sample size. Half-width $1.96 \cdot 12/\sqrt{n} \leq 1.00$ requires $\sqrt{n} \geq 1.96 \cdot 12 = 23.52$, so $n \geq 553.2$, i.e. $n = \mathbf{554}$. Nearly nine times the data for a threefold gain in precision — the $1/\sqrt{n}$ law is unforgiving, and it is the same law that makes learning curves flatten in Chapter 2.

(5) Causality. The campaign is a **confounder**: the redesign and the campaign changed together, so their effects cannot be separated from these data — they are *confounded* in the technical sense. Fixes, in descending order of quality: randomise stores to redesign / no redesign while running the campaign everywhere; or use control stores with the campaign but no redesign (difference-in-differences); or, weakest, adjust for campaign exposure in a regression — which only works for confounders you thought to measure.

Common mistake: "significant, so act". The p-value compares the increase with *zero*; the decision needed a comparison with €5. Always test the threshold that actually matters.

Outline

The model and the criterion

The model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \mathrm{E}[\varepsilon_i] = 0.$$

Least squares

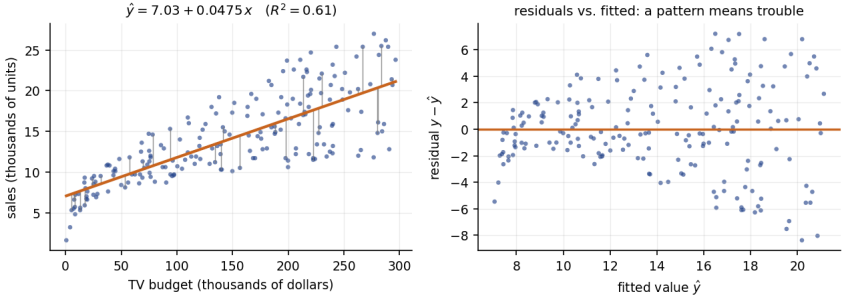
Choose b_0, b_1 minimising the residual sum of squares

$$\mathrm{RSS}(b_0, b_1) = \sum_{i=1}^{n} \big(y_i - b_0 - b_1 x_i\big)^2, \qquad b_1 = \frac{S_{xy}}{S_{xx}}, \quad b_0 = \bar{y} - b_1 \bar{x}.$$

Two things to notice now

(i) The fitted line always passes through $(\bar{x}, \bar{y})$. (ii) Squared errors are chosen for tractability, not sanctity — they make the algebra closed-form, and they make outliers expensive.

Fitted values, residuals, and R^2

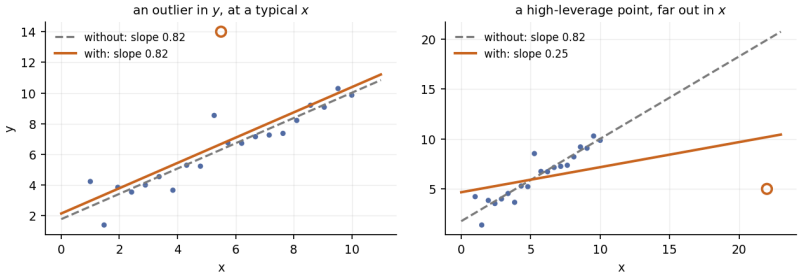


Left: the fit on `Advertising`; the vertical segments are the residuals $e_i = y_i - \hat{y}_i$ that least squares squares and adds up. Right: a residual plot — structure here (curvature, a funnel) means the model is missing something. Both plots return in Chapter 3.

In industry — Marketing: the first-pass driver analysis

Sales on advertising spend starts most marketing-mix conversations — and is not causal: spend rises in strong seasons, so the slope absorbs the season.

One point can move the whole line



Twenty simulated points plus one extra (circled). **Left:** an unusual y sitting at the average x lifts the line but leaves the slope unchanged. **Right:** a point far out in x swings the slope from 0.82 to 0.25 — that point has high *leverage*, because the line must pivot to reach it.

Two different words

An **outlier** has a large residual; a **high-leverage** point has an unusual x. A point that is both is the dangerous case, and Chapter 3 gives you leverage statistics to find them.

Reading a regression output — the table you will see all term

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.0326	0.458	15.360	0.000	6.130	7.935
TV	0.0475	0.003	17.668	0.000	0.042	0.053

R-squared: 0.612 F-statistic: 312.1 Prob (F-statistic): 1.47e-42 n = 200

Column by column

coef — the estimate b_j: sales rise by 0.0475 thousand units per extra \$1000 of TV budget. **std err** — its standard error: how much b_j would wobble across repeated samples. **t** $= \text{coef}/\text{std err} = 0.0475/0.00269 = 17.7$: the estimate is 17.7 standard errors from zero. (Divide by the printed 0.003 and you get 15.8 — the column is rounded.) **P>|t|** — the two-sided p-value for $H_0 : \beta_j = 0$. **[0.025 0.975]** — the 95% confidence interval, $\approx \text{coef} \pm 1.96\,\text{std err}$.

What to look at first

The coefficient and its interval — *is the effect big enough to matter, and how sure are we?* The p-value only answers "could it be zero". Chapter 3 adds the F-statistic, diagnostics and multiple predictors to this same table.

Extended Exercise 0.2 — Regression by hand, end to end [Math]

Extended exercise (15 min)

Four branches report weekly advertising spend x (€000s) and sales y (€000s):

x	2	4	6	8
y	5	8	12	15

1. Compute $\bar{x}$, $\bar{y}$, S_{xx}, S_{xy} and S_{yy}.
2. Compute b_1 and b_0, and write out the fitted line.
3. Compute the four fitted values and residuals. Verify that the residuals sum to (approximately) zero, and explain why they must.
4. Compute RSS, TSS and R^2. Confirm $R^2 = r^2$.
5. Interpret b_0 and b_1 in words *with units*. Is the interpretation of b_0 trustworthy here?
6. Predict sales at $x = 5$ and at $x = 40$. Which prediction would you defend, and why?

Solution — Extended Exercise 0.2 (1/2)

Solution

(1) Sums. $\bar{x} = 20/4 = 5$, $\bar{y} = 40/4 = 10$. Deviations $x_i - \bar{x} = -3, -1, 1, 3$ and $y_i - \bar{y} = -5, -2, 2, 5$.

$$S_{xx} = 9 + 1 + 1 + 9 = 20, \quad S_{xy} = 15 + 2 + 2 + 15 = 34, \quad S_{yy} = 25 + 4 + 4 + 25 = 58.$$

(2) Coefficients.

$$b_1 = \frac{34}{20} = 1.7, \qquad b_0 = 10 - 1.7(5) = 1.5, \qquad \hat{y} = 1.5 + 1.7x.$$

(3) Fitted values and residuals.

x_i	2	4	6	8
$\hat{y}_i = 1.5 + 1.7x_i$	4.9	8.3	11.7	15.1
$e_i = y_i - \hat{y}_i$	$+0.1$	-0.3	$+0.3$	-0.1

The residuals sum to 0, and they must: since $b_0 = \bar{y} - b_1\bar{x}$, $\sum_i e_i = n(\bar{y} - b_0 - b_1\bar{x}) = 0$ for *any* data set — the line is pinned through $(\bar{x}, \bar{y})$. A property of the *fitting method*, not evidence that the model is right.

Solution — Extended Exercise 0.2 (2/2)

Solution

(4) Fit quality. $\mathrm{RSS} = 0.1^2 + 0.3^2 + 0.3^2 + 0.1^2 = 0.01 + 0.09 + 0.09 + 0.01 = 0.20$, and $\mathrm{TSS} = S_{yy} = 58$. So

$$R^2 = 1 - \frac{0.20}{58} = 0.99655.$$

Check: $r = S_{xy}/\sqrt{S_{xx}S_{yy}} = 34/\sqrt{20 \cdot 58} = 34/34.06 = 0.99827$, and $r^2 = 0.99655$. ✓

(5) Interpretation. $b_1 = 1.7$: each additional €1,000 of weekly advertising is associated with €1,700 more weekly sales, on average. $b_0 = 1.5$: predicted sales of €1,500 at *zero* advertising. Treat b_0 with care — $x = 0$ is outside the observed range $[2, 8]$, so it is an extrapolation, and here it is merely the intercept that makes the line fit, not a meaningful "baseline".

(6) Predictions. At $x = 5$: $\hat{y} = 1.5 + 8.5 = 10.0$ — defensible, since 5 sits inside the observed range (interpolation). At $x = 40$: $\hat{y} = 1.5 + 68 = 69.5$ — *not* defensible: it is five times beyond the largest observed spend, where linearity is untested and diminishing returns are likely. Never extrapolate a fitted model far outside the data it was estimated from; with $n = 4$ the fit is fragile in any case.

Common mistake: being impressed by $R^2 = 0.997$ when $n = 4$. A line has two parameters, so four points make a near-perfect fit cheap — R^2 measures fit to *these* points, not validity.

Exercise 0.7 — Interrogating a regression table [Python]

Exercise

A colleague regresses a city's monthly ice-cream sales (thousands of euros) on the average temperature (°C) and reports:

```
                coef    std err          t    P>|t|     [0.025    0.975]
Intercept     4.2000     1.4000     3.00    0.005     1.3552    7.0448
temp          0.9000     0.1500     6.00    0.000     0.5952    1.2048
R-squared: 0.51     n = 36
```

1. Interpret the slope in words, with units.
2. Verify the t-value and check the confidence interval against the "$\pm 1.96\,\mathrm{SE}$" rule of thumb.
3. Interpret the intercept. Is it meaningful here?
4. What does $R^2 = 0.51$ say, and what does it *not* say?
5. Your colleague concludes: "raising the temperature by 1°C would raise sales by €900". What is wrong with the wording, and what would make the claim defensible?

Solution — Exercise 0.7

Solution

(1) A 1°C higher average monthly temperature is *associated with* €900 more sales per month (0.9 thousand euros), on average across these 36 months.
(2) $t = 0.9/0.15 = 6.0$ ✓. The rule of thumb gives $0.9 \pm 1.96(0.15) = [0.606,\ 1.194]$, slightly narrower than the printed [0.595, 1.205]: with only $n = 36$ the software uses the t_{34} quantile 2.032, not 1.96.
(3) The intercept is predicted sales at 0°C: €4,200. Whether that is meaningful depends on whether 0°C is inside the observed temperature range. If the city never drops below 5°C, it is an extrapolation and should be read as "the number that positions the line", nothing more.
(4) 51% of the month-to-month variation in sales is accounted for by temperature; 49% is not. It does *not* say the model is correct, that the relationship is linear, that temperature causes sales, or that predictions will be accurate for a new month.
(5) Two problems. *Causal wording:* this is observational data, and month is a lurking variable — school holidays, tourism and marketing all move with temperature (a confounder, and a hint of the Simpson's-paradox risk). *Units:* the coefficient is per 1°C *on the monthly average*, not per warm afternoon. A defensible version: "months that are 1°C warmer sell about €900 more, on average". To claim causation you would need an experiment or a design that holds the confounders fixed.

Common mistake: the word "effect". With observational data a coefficient is a *comparison* between months, not a lever you can pull — write "associated with" and mean it.

Outline

numpy: arrays and vectorised arithmetic

```
import numpy as np

x = np.array([2, 4, 4, 5, 7, 8, 12])    # a 1-D array
x.mean(), np.median(x), x.std(ddof=1)   # 6.0, 5.0, 3.317

z = (x - x.mean()) / x.std(ddof=1)      # vectorised: no loop needed
A = np.array([[1, 1], [1, 2], [1, 3]]) # a 3x2 matrix
A.T @ A                                 # matrix product: 2x2
```

The one trap

`np.std()` divides by n by default, not $n-1$. Pass `ddof=1` for the *sample* standard deviation. (`pandas` uses $n-1$ by default — the two libraries disagree, so state what you mean.)

pandas: the data frame is your data set

```python
import pandas as pd

df = pd.read_csv("Wage.csv")            # no index column in this file
df.shape           # (3000, 11) -- that's (n, p)
df.head()          # first rows
df.describe()      # count, mean, std, min, quartiles, max
df["wage"].median()
df["education"].value_counts()          # categorical: counts
df.groupby("education")["wage"].mean()  # a summary per group
df[["age", "wage"]].corr()              # correlation matrix
df[df["wage"] > 200].shape[0]           # filtering rows
```

The four verbs of the labs

select columns, **filter** rows, **group** and **summarise**. Almost every data-preparation step in the fifteen labs is a combination of these.

matplotlib: four plots cover most needs

```
import matplotlib.pyplot as plt

fig, ax = plt.subplots(figsize=(6, 4))
ax.hist(df["wage"], bins=40)              # one quantitative variable
ax.boxplot(df["wage"])                    # centre, spread, outliers
ax.scatter(df["age"], df["wage"], s=8)    # two quantitative variables
ax.bar(counts.index, counts.values)       # one categorical variable
ax.set_xlabel("age"); ax.set_ylabel("wage")
plt.show()
```

Plot first — model second

Anscombe's quartet: four data sets with identical means, variances, correlations and regression lines — and completely different shapes. No summary statistic replaces looking at the data.

The two modelling interfaces you will use

```
# statsmodels: inference -- coefficients, SEs, t, p, CIs
import statsmodels.formula.api as smf
fit = smf.ols("sales ~ TV", data=adv).fit()
print(fit.summary())        # the table Chapter 3 teaches you to read

# scikit-learn: prediction -- one API for every model
from sklearn.linear_model import LinearRegression
model = LinearRegression().fit(X_train, y_train)
model.predict(X_test)       # .fit / .predict everywhere
```

Which to reach for

`statsmodels` when you want to *understand* (inference: standard errors and p-values); `scikit-learn` when you want to *predict* (a uniform `fit`/`predict` interface and cross-validation tooling). Chapters 3–4 lean on the first, Chapters 5–10 on the second.

Missing values and other everyday nuisances

```
df.isna().sum()                       # how many missing, per column
df["horsepower"].unique()[:5]         # '?' hides in Auto.csv as a string!
auto = pd.read_csv("Auto.csv", na_values="?").dropna()
df["education"] = df["education"].astype("category")    # not a number
pd.get_dummies(df["education"], drop_first=True)        # -> 0/1 columns
```

Three traps that bite every semester

1. **Missing values disguised as data.** `Auto.csv` stores unknown horsepower as `'?'`, so the column loads as text and `.mean()` fails — or worse, silently misleads.
2. **Dropping rows changes the sample.** `dropna()` is fine when values are missing at random and rare; otherwise it quietly biases everything that follows. Report how many rows you lost.
3. **Categories stored as numbers.** A region coded $1, 2, 3$ will be treated as quantitative unless you say otherwise.

A whole first analysis in ten lines

```
import pandas as pd, matplotlib.pyplot as plt, statsmodels.formula.api as smf

adv = pd.read_csv("Advertising.csv", index_col=0)    # 1. load
adv.describe()                                       # 2. summarise
adv.plot.scatter(x="TV", y="sales")                  # 3. LOOK at it
fit = smf.ols("sales ~ TV", data=adv).fit()          # 4. fit
print(fit.params, fit.bse, fit.rsquared)             # 5. estimate, SE, R^2
print(fit.conf_int())                                # 6. uncertainty
plt.scatter(fit.fittedvalues, fit.resid)             # 7. residuals
```

The shape of every analysis in this course

Load → summarise → **plot** → fit → report estimates *with* uncertainty → check the residuals.

Companion notebook

Run all of this in `Lab_Notebooks/chapter_00_lab.ipynb`, §8 *Simple linear regression — the bridge into Chapter 3*.

Exercise 0.9 — Reading Python output [Python]

Exercise

A student runs the following on the `Wage` data and gets the output shown:

```
>>> df.shape
(3000, 11)
>>> df["wage"].describe()
# count 3000 | mean 111.70 | std 41.73 | min 20.09
# 25% 85.38 | 50% 104.92 | 75% 128.68 | max 318.34
>>> df["age"].corr(df["wage"])
0.1956
```

1. What are n and p?
2. Compute the IQR. Would a wage of 220 (thousand dollars, the data's units) be flagged as an outlier by the $1.5 \times \mathrm{IQR}$ rule?
3. Compare mean and median. What does the comparison say about the shape, and which should be reported?
4. Interpret the correlation of 0.196. Does it mean age barely matters for wage?
5. Compute the standard error of the mean wage.

Solution — Exercise 0.9

Solution

(1) $n = 3000$ observations, $p = 11$ columns — so with `wage` as the response, 10 columns remain. But one of them is `logwage`, a transform *of the response itself*: using it as a predictor would leak the answer into the model and produce a perfect, worthless fit. Always check what the columns actually *are*.
(2) $\mathrm{IQR} = 128.68 - 85.38 = 43.30$. The upper fence is $128.68 + 1.5(43.30) = 128.68 + 64.95 = 193.63$. Since $220 > 193.63$, **yes**, it would be flagged — but with $n = 3000$, about 100 such points (109 here) are entirely expected in a right-skewed variable. "Flagged" means "look at it", not "drop it".
(3) Mean $111.70 >$ median 104.92, so the distribution is **right-skewed** (a long upper tail of high earners, consistent with max $= 318$ against $Q_3 = 129$). Report the median and IQR alongside the mean — or report the mean and be explicit about the skew.
(4) $r = 0.196$ is a weak *linear* association. It does **not** mean age is unimportant: the wage–age relationship is famously *curved* (rising, then flattening and falling near retirement), and r cannot see curvature. Chapter 7 fits exactly this relationship with polynomials and splines, and finds age to matter a great deal.
(5) $\mathrm{SE}(\bar{x}) = s/\sqrt{n} = 41.73/\sqrt{3000} = 41.73/54.77 \approx 0.76$. `wage` is in thousands of dollars, so the mean is pinned down to about ±\$760, while individual wages vary by roughly ±\$40,000 — the standard-deviation / standard-error distinction in one line.

Common mistake: deleting the 109 flagged wages before modelling. In a right-skewed variable the fences mark the tail, not errors — deleting them would bias every later model toward low earners.

Exercise 0.10 — Choosing the right tool [Concept]

Exercise

For each task, name (a) the summary or method you would use, and (b) the plot you would draw first.

1. Describe the distribution of house prices in a city.
2. Compare salaries across three departments.
3. Assess whether advertising spend and sales move together.
4. Report how precisely a mean customer rating has been estimated.
5. Judge whether a fitted straight line is adequate.
6. Summarise which of two promotions produced more purchases.

Solution — Exercise 0.10

Solution

1. **Median and IQR** (prices are strongly right-skewed; the mean would be pulled up by a few mansions). Plot: histogram, with a boxplot if outliers are the point.
2. **Group means or medians with SDs**, per department. Plot: side-by-side boxplots — they show centre, spread and outliers in one picture. (A formal comparison would use ANOVA or a regression on dummies, Ch. 3.)
3. **Correlation** r, plus the regression slope if the units matter. Plot: scatterplot *first* — r is meaningless if the pattern is curved.
4. **Standard error** $s/\sqrt{n}$, reported as a confidence interval. Plot: the estimate with an error bar. Note this is the SE, not the SD — precision of the estimate, not spread of the ratings.
5. **Residual plot**: residuals against fitted values. Adequate means a structureless band around zero; curvature means a missing non-linear term, a funnel means non-constant variance (Ch. 3).
6. **Contingency table with row proportions** (purchase rate per promotion), plus a χ^2 test or a difference in proportions. Plot: grouped bar chart of the *rates*, not the raw counts — raw counts mislead whenever the groups differ in size.

Common mistake: defaulting to mean and SD. The first question is always the *shape* — skew, outliers, unequal groups — and the summary follows from the shape, not the other way round.

Extended Exercise 0.4 — Reviewing an analysis end to end [Integrative]

Extended exercise (15 min)

A consultancy reports on a loyalty programme. "We looked at $n = 900$ customers. Members spend on average €128 per month, non-members €96. The difference of €32 has $p < 0.001$, so the programme works. A regression of spend on membership gives $R^2 = 0.04$, and the correlation between spend and *months since joining* is $r = 0.05$, so loyalty length does not matter. We recommend enrolling every customer."

1. Spending is strongly right-skewed. What does that do to the two reported averages, and what would you report instead?
2. The standard error of the difference is €6. Give a 95% confidence interval and restate the finding properly.
3. Reconcile "$p < 0.001$" with "$R^2 = 0.04$". Is there a contradiction?
4. Criticise the conclusion drawn from $r = 0.05$.
5. Give the single biggest threat to the causal claim, and describe a study design that would settle it.
6. Which two plots would you have asked for before reading any of these numbers?

Solution — Extended Exercise 0.4 (1/2)

Solution

(1) Skew. Both means are pulled up by a small number of very large spenders, and the two groups may be skewed to different degrees, so the *difference in means* can be driven by a handful of customers. Report medians and IQRs alongside the means, plus the group sizes — and show the distributions, not just their centres. (A log transform, or a comparison of medians, is often the better analysis here.)
(2) Uncertainty. $32 \pm 1.96(6) = 32 \pm 11.8 = [€20.2,\ €43.8]$. Proper statement: "members spend about €32 more per month than non-members, and the data are consistent with anywhere between €20 and €44". That range, not the p-value, is what a business decision needs.
(3) No contradiction. They answer different questions. $p < 0.001$ says the difference is very unlikely to be zero — easy to achieve with $n = 900$. $R^2 = 0.04$ says membership accounts for only 4% of the variation in spending — individual customers differ far more among themselves than the groups differ from each other. A real but small effect, precisely estimated: significance is not importance.

Solution — Extended Exercise 0.4 (2/2)

Solution

(4) The $r = 0.05$ claim. r measures only *linear* association, so a strong non-linear pattern (a jump in the first months, then a plateau) would show up as $r \approx 0$. It is also not resistant to outliers, and "does not matter" confuses a small estimate with a precisely estimated zero — no standard error was given. Plot spend against tenure before concluding anything.

(5) The causal threat: self-selection. Customers *chose* to join, and the ones who join are plausibly those who already shop most. The comparison then measures who joins, not what joining does. A randomised trial — offer the programme to a random subset and compare all those offered with all those not — would settle it. Failing that, compare each member's spending before and after joining against matched non-members over the same period. The recommendation "enrol everyone" also assumes the effect would hold for people who never chose to join, which is exactly the population the data say nothing about.

(6) The two plots. (i) Overlaid histograms or side-by-side boxplots of spending by membership — showing shape, spread and outliers, not just two means; (ii) a scatterplot of spend against months since joining, which would immediately expose whether $r = 0.05$ hides a curve. Both are two lines of `pandas`, and both would have been drawn before any test in a competent analysis.

Common mistake: softening an unearned causal claim ("members tend to benefit…") instead of dropping it. If only observational data exist, the honest product is an association *plus its threats*.

Outline

The refresher in one slide

- **Data** are n rows $\times$ p columns; the *type* of the response decides regression vs. classification.
- **One variable**: centre (mean / median), spread (SD / IQR), shape (skew, outliers). Never a centre without a spread.
- **Two variables**: covariance $\rightarrow$ correlation r (linear, unit-free, non-causal). Anscombe and Simpson are why you plot first and ask what is missing from the model.
- **Probability**: conditioning and Bayes' theorem; base rates dominate rare-event problems.
- **Distributions**: Bernoulli/binomial/Poisson for counts; normal, and t, χ^2, F derived from it.
- **Inference**: SD describes data, SE describes an estimate and shrinks like $1/\sqrt{n}$; CIs and p-values are statements about *procedures*; power is the error nobody reports.
- **Regression**: minimise RSS; $b_1 = S_{xy}/S_{xx}$; $R^2 = 1 - \mathrm{RSS}/\mathrm{TSS}$; residual plots diagnose.
- **Algebra and calculus**: dimensions must match; $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$; set the derivative to zero, or walk downhill.

Key formulas at a glance

Description

$$\bar{x} = \tfrac{1}{n}\sum_i x_i, \quad s^2 = \tfrac{1}{n-1}\sum_i (x_i - \bar{x})^2, \quad r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \quad z_i = \tfrac{x_i - \bar{x}}{s}.$$

Probability and inference

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}, \quad \mathrm{Var}(aX + b) = a^2\mathrm{Var}(X), \quad \mathrm{SE}(\bar{x}) = \frac{s}{\sqrt{n}},$$

$$\mathrm{CI} = \bar{x} \pm t_{n-1,1-\alpha/2}\,\mathrm{SE}, \qquad t = \frac{\hat{\theta} - \theta_0}{\mathrm{SE}(\hat{\theta})}.$$

Regression and optimisation

$$b_1 = \frac{S_{xy}}{S_{xx}} = r\,\frac{s_y}{s_x}, \quad b_0 = \bar{y} - b_1\bar{x}, \quad R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}, \quad \hat{\boldsymbol{\beta}} = (\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top\mathrm{y},$$

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \alpha\nabla L(\boldsymbol{\beta}^{(t)}).$$

Vocabulary check

Term	One-line meaning
Population / sample	all units of interest / the rows you have
Parameter / statistic	unknown truth (μ) / computed from data ($\bar{x}$)
Standard deviation	spread of the *data*
Standard error	spread of an *estimate* across samples
Quartile, IQR	25%/75% points; their difference
Skewness	asymmetry; mean pulled towards the long tail
z-score	value expressed in standard deviations from the mean
Covariance / correlation	joint movement; its unit-free version
Confounder	third variable driving both X and Y
Conditional probability	probability given that something is known
Base rate	the unconditional prevalence $P(A)$
Type I / Type II error	false positive / false negative
Power	probability of detecting a real effect
Residual	observed minus fitted, $y_i - \hat{y}_i$
Leverage	how unusual an observation's x is
R^2	fraction of variation in y accounted for
Log transform	tames right skew; turns effects into percentages
Gradient	vector of partial derivatives; points uphill
Convex	bowl-shaped; every local minimum is global

Where each topic reappears in the course

Refresher topic	Comes back in
Variable types, tidy data	Ch. 1–2 (notation), Ch. 3 (dummy variables)
Mean, variance, expectation rules	Ch. 2 (bias–variance decomposition)
Standardisation (z-scores)	Ch. 2 (KNN), Ch. 6 (ridge/lasso), Ch. 10
Correlation, confounding	Ch. 3 (multiple regression, collinearity)
Anscombe, Simpson's paradox	Ch. 3 (why we plot and adjust)
Log transforms, skew	Ch. 3, 7 (transformed responses, splines)
Conditional probability, Bayes	Ch. 4 (Bayes classifier, LDA, naive Bayes)
Bernoulli / binomial / Poisson	Ch. 4 (logistic, Poisson regression)
Normal, t, χ^2, F	Ch. 3 (t- and F-tests on coefficients)
Standard errors, CIs	Ch. 3 (coefficient inference), Ch. 5 (bootstrap)
Type I/II errors, p-values	Ch. 4 (confusion matrix), Ch. 13 (FWER, FDR)
Power, sample size	Ch. 5 (how much data is enough), Ch. 13
Outliers and leverage	Ch. 3 (diagnostics), Ch. 8 (trees are robust)
Simple linear regression, R^2	Ch. 3 (all of it), Ch. 6 (model selection)
Residual plots	Ch. 3 (diagnostics), Ch. 7 (non-linearity)
Matrix algebra, inverses	Ch. 3 (normal equations), Ch. 6 (singularity)
ℓ_1 / ℓ_2 norms	Ch. 6 (lasso vs. ridge geometry)
Eigenvectors	Ch. 6 (PCR), Ch. 12 (PCA)
Derivatives, gradients, convexity	Ch. 10 (backpropagation, training)
`pandas` / `sklearn`	every lab notebook

Common pitfalls to leave behind

Don't

1. Report a mean for a skewed variable without the median — or any centre without a spread.
2. Trust r without looking at the scatterplot.
3. Read a correlation, or a regression coefficient, as a causal effect.
4. Confuse the standard deviation with the standard error.
5. Say "$p > 0.05$, therefore no effect", or "p small, therefore important".
6. Delete outliers because a rule flagged them.
7. Extrapolate a fitted model beyond the range of the data.
8. Compare distances or penalise coefficients on unstandardised variables.

Self-check questions

1. Why does the sample variance divide by $n-1$? (p. 22)
2. Give two variables with a strong relationship but $r \approx 0$. (p. 35)
3. A test is 99% accurate for a disease affecting 1 in 10,000. You test positive. Estimate $P(\text{ill} \mid +)$ — roughly. (p. 45)
4. State precisely what "95% confident" means. (p. 60)
5. Why does quadrupling the sample size only halve the standard error? (p. 58)
6. In $\hat{y} = 4 + 0.8x$ with x in years and y in €000s, interpret both coefficients with units. (p. 77)
7. If X is 200×5, what size is $(\mathrm{X}^\top \mathrm{X})^{-1}\mathrm{X}^\top \mathrm{y}$? (appendix, p. 111)
8. Why must predictors be standardised before ridge regression? (p. 27)
9. Why does a large learning rate make gradient descent diverge? (appendix, p. 118)
10. Two studies report the same estimate; one has twice the standard error. Which is more likely to be "significant", and why is that not the same as being more *true*? (p. 64)
11. Sketch Anscombe's fourth data set from memory and say why its correlation is untrustworthy. (p. 36)

Five things to remember

1. **Plot before you compute.** Anscombe's quartet: four data sets, identical statistics, four different stories.
2. **Centre needs spread; estimate needs uncertainty.** A number without a standard error is not a result.
3. **Conditioning is the core idea.** Classification is estimating $P(Y \mid X)$, and base rates matter enormously.
4. **Association is not causation.** Ask what third variable could produce the pattern — that question is why multiple regression exists.
5. **Fitting is minimising a loss.** Differentiate and solve, or walk downhill. The rest of the course changes the loss and the model, not the recipe.

Where to read more

If you want to go deeper before Lecture 1

- **ISLP Chapter 2** — assessing model accuracy; the bias–variance trade-off. It picks up exactly where this deck stops.
- **ISLP Chapter 3.1** — simple linear regression, with the inference (standard errors, t-tests) that we only sketched here.
- Any introductory statistics text, for: descriptive statistics, probability rules, the normal and t distributions, confidence intervals, hypothesis tests, and simple regression.
- A first linear-algebra text, for: matrix multiplication, inverses, and eigenvectors.

Practice beats reading

Redo Exercises 0.2, 0.3 and Extended Exercise 0.2 *by hand*, then reproduce them in Python in `Lab_Notebooks/chapter_00_lab.ipynb` — §§1–10 rebuild every figure of this deck from the course data, and *Lecture exercises — worked Python solutions* checks your answers. That single hour is the best preparation for Lecture 1.

What's next

Lecture 1 — *Chapter 1: Introduction* and the start of *Chapter 2: Statistical Learning*.

- What statistical learning is, and why $Y = f(X) + \varepsilon$.
- Prediction versus inference — two different goals, two different modelling strategies.
- Supervised versus unsupervised, regression versus classification.
- The course data sets: `Wage`, `Smarket`, `NCI60`.

Bring with you

The vocabulary of this deck. From Lecture 1 on, "standard error", "residual", "conditional probability" and "gradient" are used without further explanation.

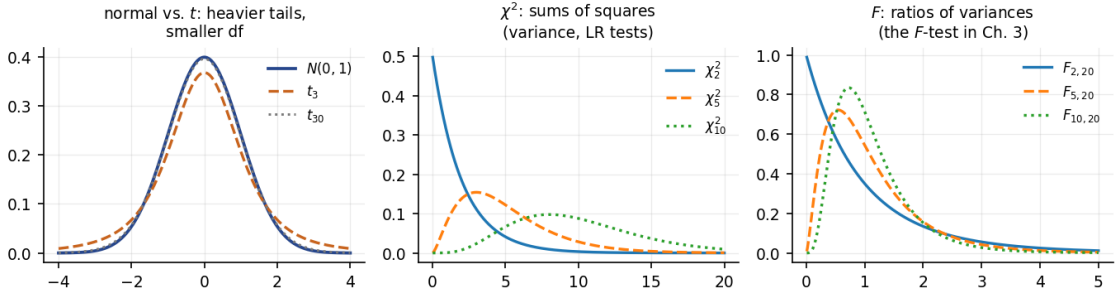
Appendix — what is in here, and why

Nothing in the appendix is needed to follow the chapter: each slide is a formal derivation, a longer worked exercise, or a side topic. Read it when you want the full story, or when a later chapter sends you back here.

Contents of the appendix

- **χ^2, t, F, and LLN vs. CLT** — the distributions derived from the normal, and the two asymptotic promises. You only need to recognise the names when they appear in Chapters 3–6.
- **The ANOVA decomposition** — $\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$ written out. Useful, but R^2 is the version you will actually quote.
- **Linear algebra** — vectors, matrices, matrix multiplication, norms and eigenvectors (with Exercise 0.8). Needed to *read* the compact formulas, never to compute anything by hand in this course.
- **Calculus and optimisation** — derivatives, minimising a loss, gradient descent (with Extended Exercise 0.3). Software does this for you; come back here when Chapter 10 trains a network.

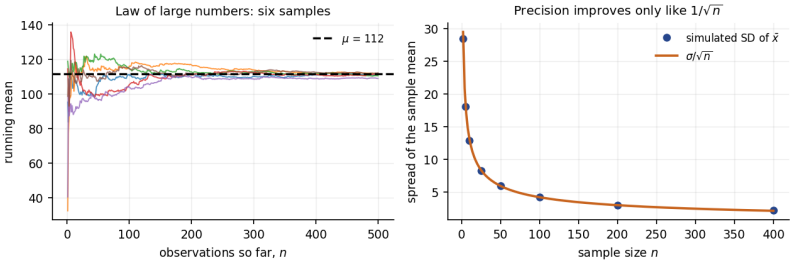
Three distributions that come from the normal



Distribution	Arises as	Used for
t_ν	normal mean with σ *estimated*	t-tests, CIs on coefficients (Ch. 3)
χ^2_ν	sum of ν squared standard normals	variance, likelihood-ratio tests
F_{d_1,d_2}	ratio of two scaled χ^2	the F-test for several coefficients (Ch. 3)

As the degrees of freedom grow, $t_\nu \to N(0,1)$: with $\nu > 30$ the difference is negligible, which is why large-sample rules of thumb use 1.96.

Two different promises: LLN and CLT



Law of large numbers	**Central limit theorem**
$\bar{X} \to \mu$ as n grows	$\bar{X}$ becomes normally distributed
says *where* the mean lands	says *how far off* it typically is
justifies estimating at all	justifies confidence intervals and tests

The right panel is the economics of data collection: going from $n = 25$ to $n = 100$ halves the uncertainty; halving it again costs $n = 400$.

Decomposing the variation

The three sums of squares

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{\text{explained}} + \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{\text{RSS}}, \qquad R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

How to read R^2

The fraction of the variation in y that the model accounts for. $R^2 = 0.61$ means "61% of the variability in sales is accounted for by TV spend". In simple regression $R^2 = r^2$ exactly.

Do not over-read it

A high R^2 does not make a model correct, useful or causal, and a low one does not make it worthless. Adding any predictor never decreases R^2 — which is why Chapter 6 needs adjusted R^2, C_p, AIC, BIC and cross-validation.

Vectors and matrices: the minimum

Objects

A **vector** $a \in \mathbb{R}^p$ is a column of p numbers; an $m \times n$ **matrix** has m rows and n columns; $A^\top$ swaps them.

Multiplication — the only rule you must remember

$(m \times n) \cdot (n \times k) = (m \times k)$: the *inner* dimensions must match and they disappear. Entry (i, j) of the product is the dot product of row i of the first and column j of the second.

Special cases

Dot product $a^\top b = \sum_j a_j b_j$ (a scalar); identity I with $AI = A$; inverse A^{-1} with $A^{-1}A = I$ (exists only if A is square and non-singular).

Dimension checking is free error-detection

If X is $n \times p$, then $X^\top X$ is $p \times p$ and $X^\top y$ is $p \times 1$. Any formula whose dimensions do not line up is wrong — no algebra needed.

Why this course speaks matrix

Regression in one line

With a column of ones in X, the least-squares estimator is

$$\hat{\boldsymbol{\beta}} = (\mathrm{X}^\top \mathrm{X})^{-1} \mathrm{X}^\top \mathrm{y}, \qquad \hat{\mathrm{y}} = \mathrm{X}\hat{\boldsymbol{\beta}}.$$

Read it as a recipe

$\mathrm{X}^\top \mathrm{y}$ collects how each predictor covaries with the response; $\mathrm{X}^\top \mathrm{X}$ records how they covary with *each other*; inverting it divides out that shared information — which is what "holding the others constant" means algebraically.

When the inverse fails

If two predictors are perfectly correlated, or $p > n$, then $\mathrm{X}^\top \mathrm{X}$ is singular and $\hat{\boldsymbol{\beta}}$ does not exist. Ridge regression inverts $\mathrm{X}^\top \mathrm{X} + \lambda \mathrm{I}$ instead (Ch. 6).

Norms, distances and eigenvectors — where they show up

Norms

$$\|\boldsymbol{\beta}\|_2^2 = \sum_j \beta_j^2 \quad (\ell_2), \qquad \|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j| \quad (\ell_1).$$

Where: ridge penalises $\|\boldsymbol{\beta}\|_2^2$ and shrinks coefficients smoothly; lasso penalises $\|\boldsymbol{\beta}\|_1$ and sets some to exactly zero (Ch. 6). The difference between a round and a pointed constraint region is the whole story.

Euclidean distance

$d(\mathrm{a}, \mathrm{b}) = \sqrt{\sum_j (a_j - b_j)^2}$. *Where:* KNN (Ch. 2, 4) and clustering (Ch. 12) are built on it — and it is why unstandardised variables wreck both.

Eigenvectors and eigenvalues

$\mathrm{A}\mathrm{v} = \lambda \mathrm{v}$: directions the matrix only stretches. *Where:* principal components (Ch. 6, 12) are the eigenvectors of the covariance matrix — the directions of greatest variance in the data.

Exercise 0.8 — Matrix mechanics [Math]

Exercise

Let

$$\mathrm{X} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}, \qquad \mathrm{y} = \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix}.$$

1. State the dimensions of $\mathrm{X}^\top$, $\mathrm{X}^\top\mathrm{X}$ and $\mathrm{X}^\top\mathrm{y}$ *before* computing anything.
2. Compute $\mathrm{X}^\top\mathrm{X}$ and $\mathrm{X}^\top\mathrm{y}$.
3. Invert $\mathrm{X}^\top\mathrm{X}$, using $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.
4. Compute $\hat{\boldsymbol{\beta}} = (\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top\mathrm{y}$, and check the slope against $b_1 = S_{xy}/S_{xx}$.
5. What would go wrong if a third column equal to twice the second were added to X?

Solution — Exercise 0.8

Solution

(1) $\mathbf{X}$ is 3×2, so $\mathbf{X}^\top$ is 2×3, $\mathbf{X}^\top\mathbf{X}$ is 2×2 and $\mathbf{X}^\top\mathbf{y}$ is 2×1; hence $\hat{\boldsymbol{\beta}}$ is 2×1 — an intercept and a slope.
(2)

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix}, \qquad \mathbf{X}^\top\mathbf{y} = \begin{pmatrix} 2+3+5 \\ 2+6+15 \end{pmatrix} = \begin{pmatrix} 10 \\ 23 \end{pmatrix}.$$

(3) $\det = 3(14) - 6(6) = 42 - 36 = 6$, so $(\mathbf{X}^\top\mathbf{X})^{-1} = \frac{1}{6}\begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}$.
(4)

$$\hat{\boldsymbol{\beta}} = \frac{1}{6}\begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}\begin{pmatrix} 10 \\ 23 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 140 - 138 \\ -60 + 69 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 2 \\ 9 \end{pmatrix} = \begin{pmatrix} 0.333 \\ 1.5 \end{pmatrix}.$$

Check: $\bar{x} = 2$, $\bar{y} = 10/3$, $S_{xx} = 2$, $S_{xy} = 3$, so $b_1 = 1.5$ ✓and $b_0 = 10/3 - 3 = 1/3$ ✓.
(5) Perfect collinearity: $\mathbf{X}^\top\mathbf{X}$ becomes singular ($\det = 0$), the inverse does not exist, and infinitely many coefficient vectors give an identical fit. Software reports a "singular matrix" or silently drops a column.
Common mistake: inverting $\mathbf{X}^\top\mathbf{X}$ entry by entry — a matrix inverse is not element-wise. Check any inverse by multiplying back to I.

Derivatives: the three facts that matter

Rules

$$\frac{d}{dx}x^k = kx^{k-1}, \qquad \frac{d}{dx}e^x = e^x, \qquad \frac{d}{dx}\log x = \frac{1}{x}, \qquad \frac{d}{dx}f(g(x)) = f'(g(x))\,g'(x).$$

What a derivative means here

The slope of the function: how much the output changes per unit change in the input. At a minimum, the slope is zero — which is how every estimator in this course is derived.

Partial derivatives and the gradient

With several inputs, differentiate with respect to one at a time, holding the others fixed. Stacking them gives the gradient

$$\nabla f(\boldsymbol{\beta}) = \left(\frac{\partial f}{\partial \beta_1}, \ldots, \frac{\partial f}{\partial \beta_p}\right)^\top,$$

which points in the direction of *steepest increase*.

Minimising a loss: the pattern behind every method

Deriving the sample mean

Minimise $L(c) = \sum_i (x_i - c)^2$ over c:

$$\frac{dL}{dc} = -2\sum_i (x_i - c) = 0 \implies \sum_i x_i = nc \implies c = \bar{x}.$$

So the sample mean is not a convention — it is *the* minimiser of squared error. (Minimising $\sum_i |x_i - c|$ instead gives the median.)

The recurring recipe

1. Write a loss measuring misfit. **2.** Differentiate. **3.** Set to zero and solve — or, when no closed form exists, walk downhill. Least squares, logistic regression, splines, boosting and neural networks are all this recipe with different losses.

Gradient descent: when there is no closed form

The update

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \alpha \, \nabla L\big(\boldsymbol{\beta}^{(t)}\big), \qquad \alpha > 0 \text{ the learning rate.}$$

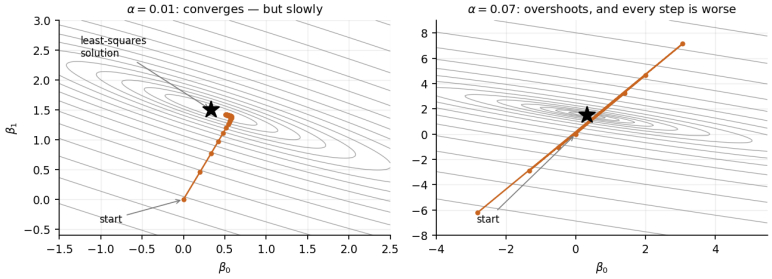
The picture

Stand on a hillside in fog. The gradient points uphill, so step the opposite way, and repeat. α too small $\Rightarrow$ crawling; α too large $\Rightarrow$ overshooting and divergence. Choosing it well is a recurring practical theme of Chapter 10.

Convexity decides whether this is safe

For a convex loss (least squares, logistic regression) every local minimum is global, so descent finds the answer. Neural networks are *not* convex — which is why initialisation, learning-rate schedules and early stopping matter so much in Chapter 10.

Gradient descent, drawn



The contours are the loss $L(\beta_0, \beta_1)$ for the three-point data set of Extended Exercise 0.3; the star is the least-squares solution. **Left:** a safe learning rate walks downhill and creeps along the valley floor. **Right:** too large a rate steps past the valley, lands higher up the far side, and every step after that is worse.

Why the valley is so narrow

Long thin contours mean the loss is far more curved in one direction than another, which forces a small step size. Standardising the predictors makes the contours rounder — the same fix, again.

Extended Exercise 0.3 — Normal equations and one descent step [Integrative]

Extended exercise (15 min)

Use the data of Exercise 0.8: $(x, y) = (1, 2), (2, 3), (3, 5)$, model $y = \beta_0 + \beta_1 x$, and loss $L(\boldsymbol{\beta}) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$.

1. Write L out explicitly for the three observations.
2. Derive $\partial L/\partial \beta_0$ and $\partial L/\partial \beta_1$, and show that setting both to zero gives the normal equations $\mathrm{X}^\top \mathrm{X}\boldsymbol{\beta} = \mathrm{X}^\top \mathrm{y}$.
3. Starting from $\boldsymbol{\beta}^{(0)} = (0, 0)^\top$, compute the gradient and take one gradient-descent step with $\alpha = 0.01$.
4. Compute L at $\boldsymbol{\beta}^{(0)}$, at $\boldsymbol{\beta}^{(1)}$, and at the exact solution $(1/3,\ 3/2)^\top$. Did the step help?
5. What happens with $\alpha = 1$? Compute $\boldsymbol{\beta}^{(1)}$ and comment.

Solution — Extended Exercise 0.3 (1/2)

Solution

(1) The loss.

$$L(\beta_0, \beta_1) = (2 - \beta_0 - \beta_1)^2 + (3 - \beta_0 - 2\beta_1)^2 + (5 - \beta_0 - 3\beta_1)^2.$$

(2) Partial derivatives. With $e_i = y_i - \beta_0 - \beta_1 x_i$,

$$\frac{\partial L}{\partial \beta_0} = -2\sum_i e_i, \qquad \frac{\partial L}{\partial \beta_1} = -2\sum_i x_i e_i.$$

Setting both to zero gives $\sum_i e_i = 0$ and $\sum_i x_i e_i = 0$, i.e. $3\beta_0 + 6\beta_1 = 10$ and $6\beta_0 + 14\beta_1 = 23$ — exactly $\mathrm{X}^\top \mathrm{X}\boldsymbol{\beta} = \mathrm{X}^\top \mathrm{y}$ with the matrices of Exercise 0.8, so $\boldsymbol{\beta} = (1/3,\ 3/2)^\top$ as before. *The normal equations are "derivative $= 0$" in matrix form.*

(3) One step. At $(0, 0)^\top$ the residuals are $e = (2, 3, 5)$:

$$\nabla L = -2\big(\textstyle\sum_i e_i,\ \sum_i x_i e_i\big)^\top = -2(10,\ 23)^\top = (-20,\ -46)^\top,$$

$$\boldsymbol{\beta}^{(1)} = (0, 0)^\top - 0.01(-20, -46)^\top = (0.20,\ 0.46)^\top.$$

The step moves *towards* $(0.33, 1.50)$, as it should.

Solution — Extended Exercise 0.3 (2/2)

Solution

(4) Did it help?

$$L(0,0) = 2^2 + 3^2 + 5^2 = 38.$$

At $\beta^{(1)} = (0.20, 0.46)$ the fitted values are $0.66, 1.12, 1.58$, so the residuals are $1.34, 1.88, 3.42$ and

$$L(\beta^{(1)}) = 1.796 + 3.534 + 11.696 = 17.03.$$

At the exact solution the residuals are $1/6, -1/3, 1/6$, so $L = 1/6 \approx 0.167$. One step cut the loss from 38 to 17 — a big improvement, still far from the optimum: descent needs many iterations, which is why a closed form is preferred whenever one exists.

(5) Too large a learning rate. With $\alpha = 1$,

$$\beta^{(1)} = (0,0)^\top - 1 \cdot (-20, -46)^\top = (20,\ 46)^\top.$$

Fitted values are $66, 112, 158$ against $y = (2, 3, 5)$: the loss explodes to about 3.9×10^4, far worse than the starting 38, and the next step is larger still — **divergence**. The step size must respect the curvature of the loss, so in practice one standardises the predictors and tunes α: the workflow of Chapter 10.

Common mistake: blaming the model when descent diverges. The loss surface did not change — the step size did. Divergence is a tuning failure: standardise the predictors and shrink α.