BASIRA

CANCER METABOLISM
RESEARCH & TRAINING

بصيرة

TRAINING CURRICULUM

CHAPTER FIVE

The Toolkit

A reference for the statistics and the code — not a chapter to read straight through

The taught core is the first half: eight ideas everything else rests on. The second half is a lookup reference — come back to it when a specific question needs a specific method, which is exactly what will happen in Chapter 6.

Founder & Programme Lead	Aroob Gomosani
Prepared & delivered by	Haneen Marghalani
Version	2.0 · second cohort
Date	2026

Gomosani, A., & Marghalani, H. (2026). Basira Curriculum. Zenodo.
https://doi.org/10.5281/zenodo.21791751

5·0

How to use this chapter

This chapter is deliberately not like the others. Nobody learns statistics by reading about statistics; you learn it by needing a method, looking it up, using it, and getting it wrong a few times. So this is built as a reference.

The taught core — sections 5.1 to 5.8 — is the part we cover together, and it is genuinely small: eight ideas that everything else is built from. If you understand those eight, you can look up the rest as you need it and it will make sense when you get there. If you skip them, no amount of looking things up will help.

From 5.9 onward, each entry is a page or less in a fixed format: what the method does, when to use it, and what to watch out for. Do not read them in order. Use Figure 5.1 to find the one you need.

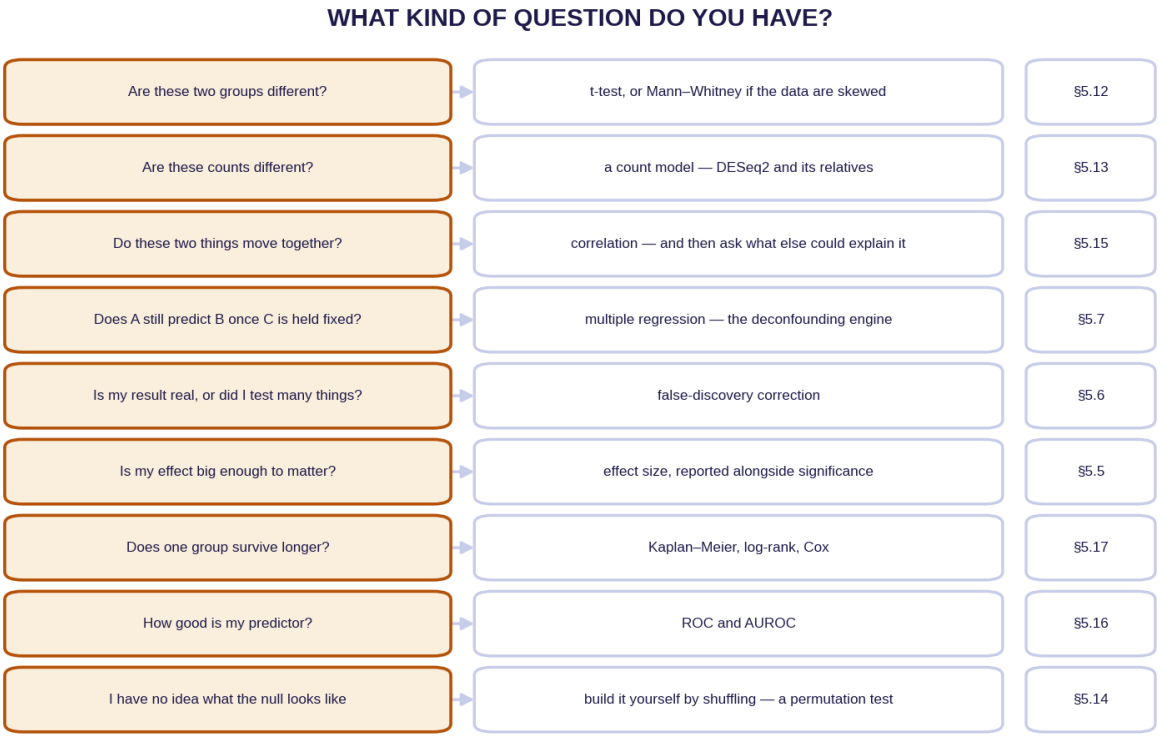
WHERE THIS SITS

	Chapter	What it answers
	1 · Thinking	How does science reason, and how do I avoid fooling myself?
	2 · Practice	How do we behave — integrity, collaboration, communication.
	3 · The Science	What is cancer, and how does a cell power itself?
	4 · Data & Tools	What data exists, and how do we handle it without breaking it?
▶️	5 · The Toolkit	The statistics and code, as a reference you return to.
	6 · Build Your Own	Now go find a question of your own and run it.

THE TAUGHT CORE, IN ONE LIST

- Describing one column of numbers — where the middle is, and how spread out it is.
- The z-score — putting different things on the same ruler.
- Distributions — the shapes that keep coming back, and why counts are not normal.
- The logic of a test — what a p-value is, and the three things it is not.
- Effect size — why significance and importance are different questions.
- Multiple testing — what happens when you ask twenty thousand questions at once.
- Regression and adjustment — holding one thing fixed to see whether another survives.
- Honesty in code — the five practices that make a result trustworthy.

5.1
Finding the method you need



This chapter is a reference. Read the taught core, then come back here whenever you need a method.

Figure 5.1 Start here whenever you have a question and do not know which tool it needs.

5.2
Describing one column

Before any test, look at your data. A single column of numbers has two properties worth naming immediately: where its middle sits, and how widely it scatters around that middle.

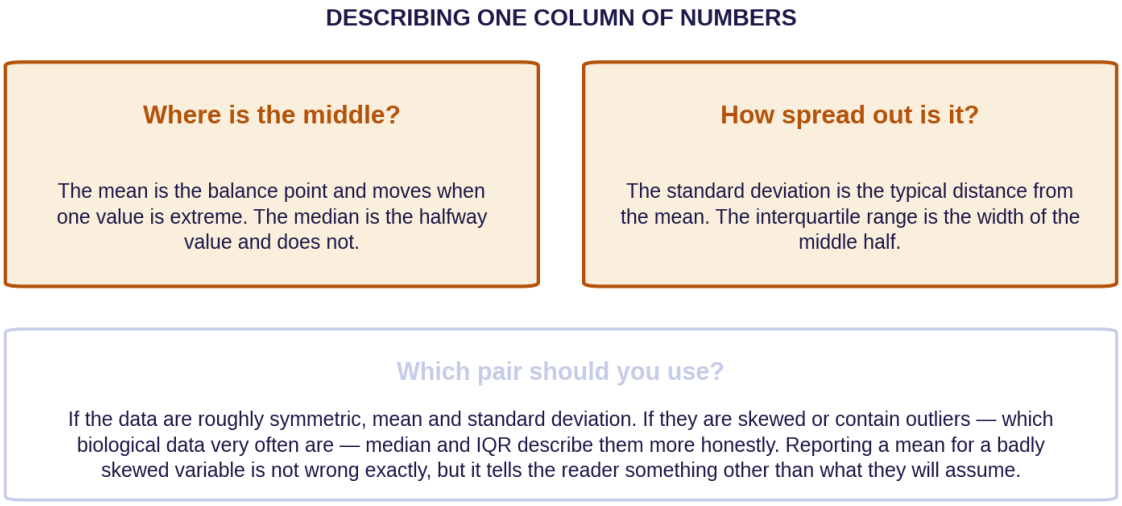


Figure 5.2 Two questions, two pairs of answers, and a rule for choosing between them.

5.3 The z-score

You will constantly need to compare quantities measured in different units. The z-score is the standard trick: it re-expresses any value as its distance from its own mean, counted in standard deviations.

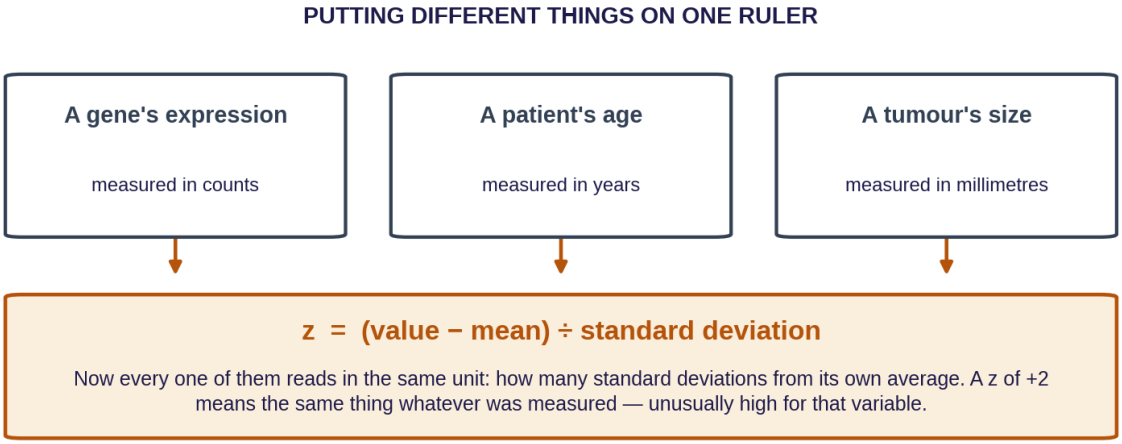


Figure 5.3 Three incomparable measurements, converted to one common scale.

5.4 Distributions

A distribution is just the shape a column of numbers makes. A handful of shapes recur so often that recognising them saves you enormous trouble — mostly because choosing a method that assumes the wrong shape is one of the commonest ways to get a confidently wrong answer.

THE SHAPES THAT KEEP COMING BACK

Normal	the bell curve. Sums and averages of many small effects tend towards it.
Uniform	the shape of no pattern. What p-values look like when nothing is going on.
Bernoulli / Binomial	yes-or-no events, and counts of successes out of n tries.
Poisson	counts of rare events in a fixed window, where variance equals the mean.
Negative binomial	counts that wobble more than Poisson allows. What RNA-seq actually needs.
Beta	a distribution over proportions — useful when the quantity itself is a fraction.

The one to remember: RNA-seq counts are over-dispersed, so a Poisson model understates their variability and manufactures false positives. That is why count models use the negative binomial.

Figure 5.4 Six shapes. The last box is the one that matters most for sequencing data.

5.5
The logic of a test

Every statistical test you will ever run has the same four-step shape underneath it. Learn the shape and the individual tests stop being a list to memorise.

EVERY STATISTICAL TEST IS THIS, IN FOUR STEPS

1	**Assume nothing is going on**	state the null hypothesis — no difference, no effect
2	**Ask what the data would look like**	if the null were true, how much would results scatter by chance alone?
3	**Compare what you actually saw**	how far into the tail of that distribution does your result fall?
4	**Report how surprising it is**	the p-value: the probability of a result this extreme, IF the null were true

Notice what the p-value is not: it is not the probability that your hypothesis is true.

Figure 5.5 The pattern all of them share.

▲ CAUTION — **Three things a p-value is not**

It is not the probability that your hypothesis is true. It is the probability of data this extreme assuming the null is true — a conditional running the other way.

It is not a measure of how big or important the effect is. A tiny p-value can accompany a trivial difference.

And a p-value above your threshold is not proof of no effect. It may only mean you did not have the power to see one — the "absence of evidence" point from Chapter 1.

5.6
Effect size

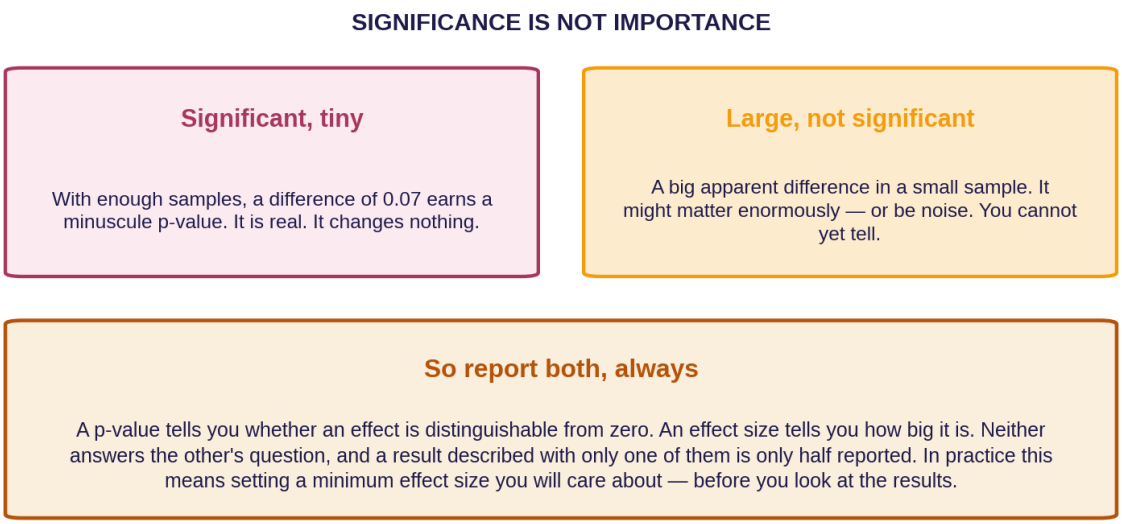


Figure 5.6 Two failures that look like successes if you only report one number.

The practical habit: decide, before you look at results, how large a difference would actually matter for your question. Then report that alongside significance, every time. A gate on effect size is one of the cheapest defences against fooling yourself that exists.

5.7
Testing many things at once

In genomics you rarely test one hypothesis. You test twenty thousand, one per gene — and that changes the arithmetic completely.

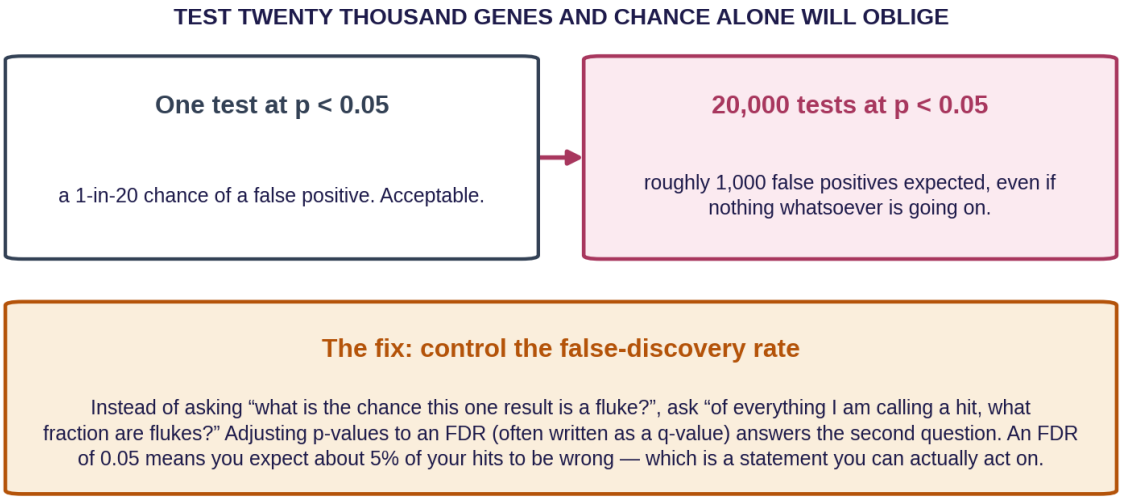


Figure 5.7 Why an uncorrected p-value threshold is meaningless at genome scale.

5.8
Regression and adjustment

This is the most useful single technique in the chapter, and the one Chapter 1's confounder discussion was building towards.

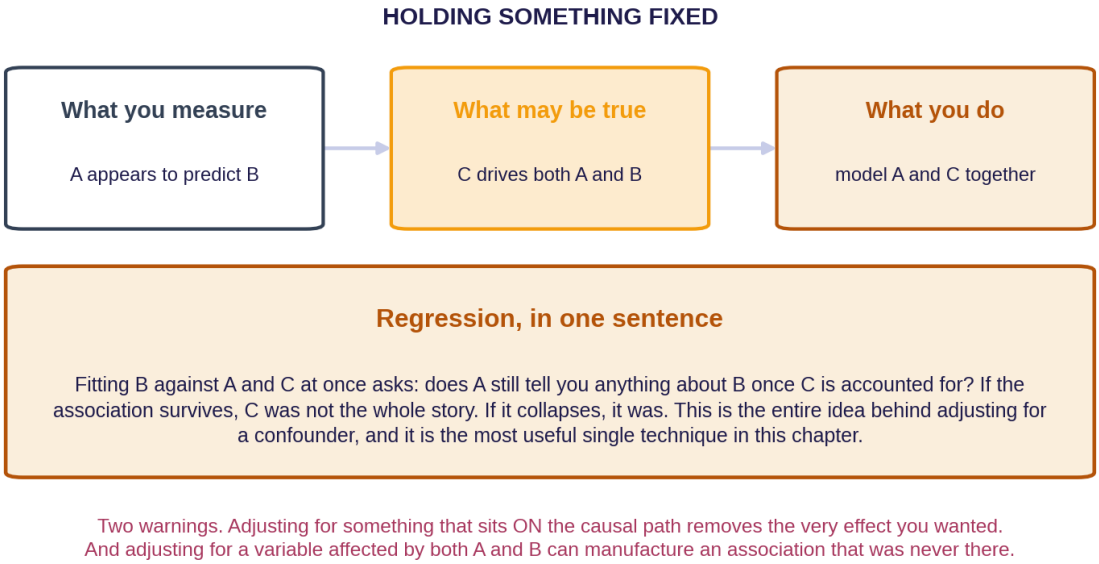


Figure 5.8 What it means to hold something fixed, and the two ways it goes wrong.

◆ **KEY IDEA — Screen and arbiter**

Analyses often use two tests in sequence: a broad, permissive screen, then a stricter arbiter on whatever survived. They have different jobs and should be tuned differently.

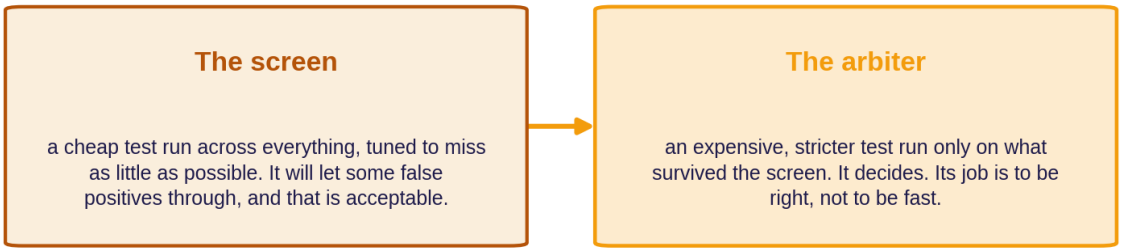


Figure 5.9 Two tests, two jobs — and the rule that keeps the pair honest.

5.9
Honesty in code

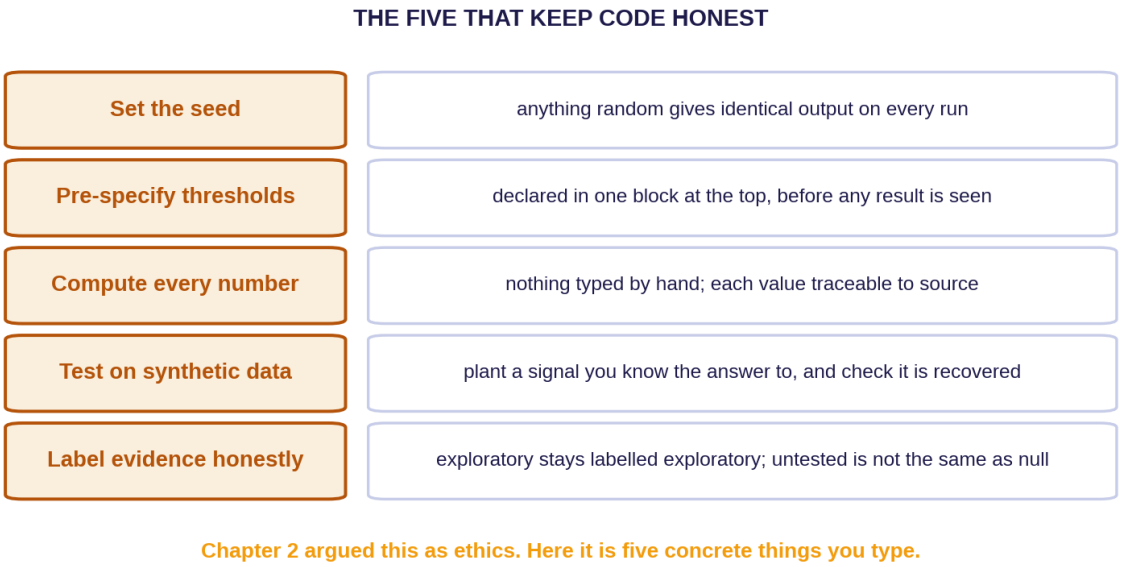
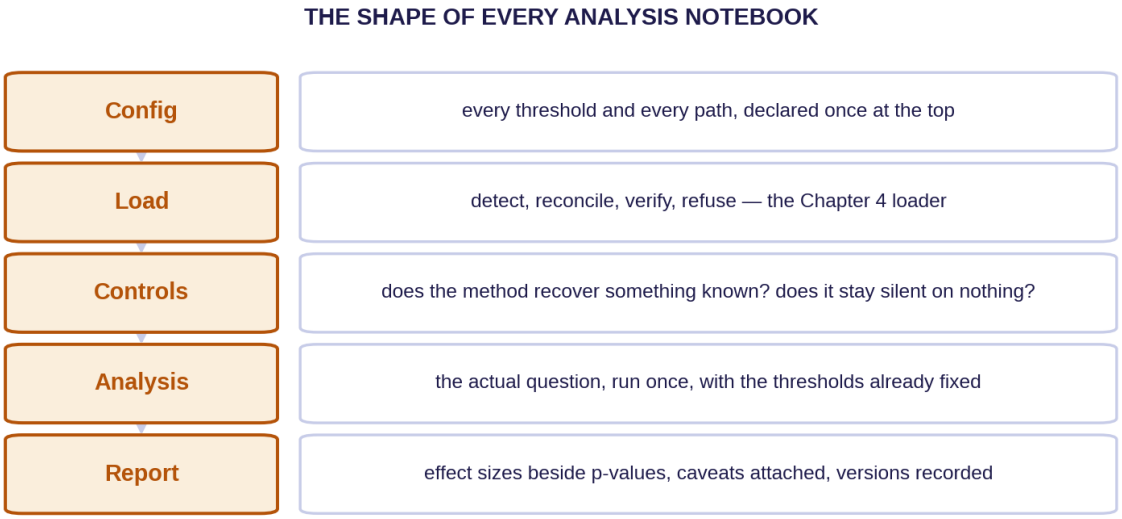


Figure 5.10 Chapter 2's ethics, expressed as five things you actually type.



If you can read this skeleton, you can read any notebook in the group.

Figure 5.11 Every analysis in this group has this shape. Learn it once and you can read all of them.

? CHECKPOINT — Checkpoint — the taught core

If you can answer these, you are ready to use the reference half.

- When should you report a median rather than a mean, and why? (§5.2)
- A gene has a z-score of −2.4 in a sample. Say in plain English what that means. (§5.3)
- Why are RNA-seq counts modelled with a negative binomial rather than a Poisson? (§5.4)
- Write out the three things a p-value is not. (§5.5)
- You test 20,000 genes and 900 come back at $p < 0.05$. How many would you expect by chance alone if nothing were going on? (§5.7)
- Explain what “adjusting for a confounder” actually does to a regression. (§5.8)
- Name two ways adjustment can make a result worse rather than better. (§5.8)

REFERENCE

Methods, in a fixed format

Each entry answers the same three questions. Use Figure 5.1 to find the one you want; do not read these in order.

5.10 Comparing groups

t-test

What it does Compares the means of two groups, using the difference divided by its uncertainty.

Use it when Both groups are roughly symmetric and you care about a difference in means.

Watch out for Skew and outliers. With small, skewed samples it can mislead badly — use the rank-based test instead.

Mann–Whitney (rank-sum)

What it does Compares two groups by rank rather than by value, asking whether one tends to sit higher than the other.

Use it when The data are skewed, ordinal, or contain outliers you do not want to dominate the answer.

Watch out for It tests a shift in the distribution, not specifically a difference in means — describe your conclusion accordingly.

Trend test

What it does Tests whether an outcome moves consistently across ordered groups, rather than just differing somewhere.

Use it when Your groups have a natural order — stage I, II, III, IV — and you expect a direction.

Watch out for Only use it when the ordering is real. Imposing an order to gain power is a form of p-hacking.

Permutation test

What it does Builds the null distribution yourself by shuffling labels many times and re-computing the statistic.

Use it when You have no clean theoretical null, or your data violate the assumptions of the standard test.

Watch out for The shuffling must respect the structure of your data. Shuffle across a batch and you destroy the very thing you were controlling for.

5.11 Counts and expression

Count models (DESeq2 and relatives)

What it does Models raw sequencing counts with a negative binomial, estimating dispersion across genes and testing for differences.

Use it when You have raw integer counts and want differential expression between conditions.

Watch out for It needs raw counts. Feed it normalised or log-transformed values and the results are quietly meaningless — see §4.5.1.

Enrichment tests

What it does Asks whether an overlap between two gene sets is larger than chance would produce.

Use it when You have a list of hits and want to know which pathways are over-represented in it.

Watch out for The background set you compare against changes the answer completely. State it explicitly.

5.12
Relationships and structure

Correlation

What it does Measures whether two variables move together — Pearson for linear relationships, Spearman for monotonic ones.

Use it when A first look at whether two measurements are related at all.

Watch out for Everything in Chapter 1. A correlation licenses no causal claim whatsoever, and a third variable is always a live explanation.

Partial correlation

What it does Measures the association between two variables after removing the influence of a third.

Use it when You suspect a shared driver and want to see what survives once it is accounted for.

Watch out for It removes only what you specify. Unmeasured confounders remain untouched and invisible.

PCA and clustering

What it does Reduces many correlated measurements to a few summary dimensions, or groups samples by similarity.

Use it when You have too many variables to look at, and want to see the dominant structure.

Watch out for The largest source of variation is frequently technical, not biological. Check what your first component actually tracks before interpreting it.

5.13
Earning trust in a result

Orthogonal replication

What it does Checks whether the same signal appears using an independent method or dataset.

Use it when Before believing anything. This is the strongest cheap evidence available to a computational group.

Watch out for Independence must be real. Two datasets sharing samples or processing are not independent, however separate they look.

Rank aggregation

What it does Combines ranked lists from several cohorts into one consensus ranking.

Use it when You have the same analysis run across multiple independent cohorts.

Watch out for It rewards consistency, not magnitude. A gene ranked moderately everywhere can beat one ranked first in a single cohort — which is usually what you want, but say so.

Meta-analysis

What it does Pools effect estimates across studies, weighting each by its precision.

Use it when Several small studies address the same question and none is decisive alone.

Watch out for Publication bias. If null results were never published, pooling what was published inherits the distortion.

ROC and AUROC

What it does Grades how well a continuous score separates two classes, across every possible threshold.

Use it when You have a predictor and want to know how good it is without committing to a cut-off.

Watch out for AUROC looks flattering on imbalanced data. With rare positives, precision-recall is the more honest summary.

5.14
Time to an event

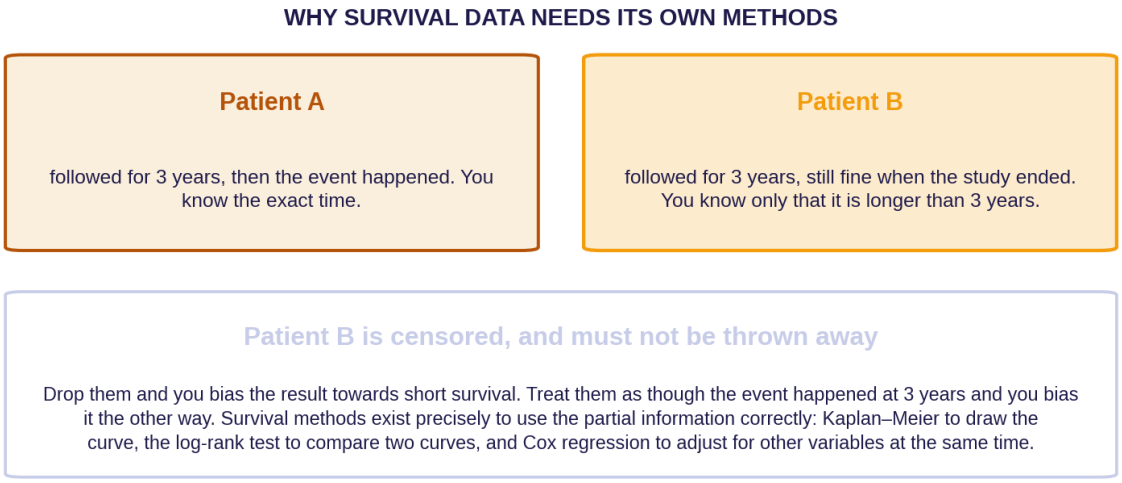


Figure 5.12 The problem that makes survival analysis its own subject.

Kaplan–Meier, log-rank, Cox

What it does Kaplan–Meier draws the survival curve; the log-rank test compares two curves; Cox regression compares them while adjusting for other variables.

Use it when Your outcome is time until something happens, and some subjects have not had it happen yet.

Watch out for Splitting a continuous variable at its optimal cut-point and then testing the split is circular — the cut-point was chosen using the outcome.

5.15 The code

numpy and pandas

What it does Arrays and labelled tables. Almost every analysis is these two plus a model.

Use it when Always. They are the substrate.

Watch out for In pandas, the index is everything. Operations align on it silently, so a mismatched index can reorder or blank your data without complaint.

Defensive loading

What it does The Chapter 4 discipline expressed in code: detect, reconcile, verify, refuse.

Use it when At the top of every notebook, before anything else runs.

Watch out for The temptation to comment out a failing check to get moving. That check is the only thing standing between you and a plausible wrong answer.

Synthetic self-tests

What it does Generating data with a planted signal you know the answer to, and confirming your code recovers it.

Use it when Whenever you write or change an analysis step.

Watch out for A test that passes on data too clean to be realistic proves less than it appears to. Plant noise as well as signal.

5.16 End-of-chapter questions

These use the reference half. Look things up — that is the point.

- You want to know whether tumours of stage I, II, III and IV differ in the expression of one gene. Which test, and why not simply run three pairwise comparisons?
- A colleague reports 400 significant genes at $p < 0.05$ out of 20,000 tested, with no correction. What do you say?
- You find a strong correlation between two genes. Write three explanations other than one regulating the other.
- Your predictor achieves an AUROC of 0.91 on a dataset where 3% of cases are positive. Why might that be less impressive than it sounds?
- Design a synthetic self-test for an analysis that is supposed to find genes differing between two groups. What would you plant, and what would prove the code works?
- Pick any entry from the reference half that you have not used. Read it, then write a one-paragraph explanation of it for a student who has not.

5.17
Go deeper

Statistics rewards multiple explanations of the same idea. If one of these does not land, try another.

BUILDING INTUITION

[Watch] **Crash Course Statistics — full series** — Covers most of the taught core, and covers it well. Start at the beginning.

[Watch] **StatQuest with Josh Starmer** — Short, patient explanations of individual methods. Genuinely the best resource of its kind.

[Read] **Seeing Theory — a visual introduction to probability and statistics** — Interactive and free. Excellent for distributions and for what a p-value actually is.

[Read] **Khan Academy — Statistics & Probability** — Worked problems, self-paced, free.

WHEN YOU WANT DEPTH

[Read] **Points of Significance — Nature Methods column** — One statistical idea per page, written for biologists. Close to ideal for this chapter's readership.

[Source] **Benjamini & Hochberg (1995) — Controlling the False Discovery Rate** — The paper behind §5.7, and behind most of modern genomics.

[Source] **Wasserstein & Lazar (2016) — The ASA Statement on p-Values** — A professional statistics body explaining what p-values do and do not mean. Short and worth it.

5.18
Glossary

Adjustment Including a variable in a model so its influence is held fixed while another is examined.

AUROC The area under the ROC curve: the probability a random positive scores above a random negative.

Censoring When a subject's event time is only known to exceed some value, because follow-up ended first.

Confounder A variable influencing both the predictor and the outcome, producing association without causation.

Dispersion How much more variable counts are than a simple Poisson model would predict.

Effect size How large a difference is, as distinct from how confidently it can be distinguished from zero.

False discovery rate (FDR) The expected proportion of false positives among the results you are calling hits.

Interquartile range (IQR) The width of the middle half of the data; a spread measure robust to outliers.

Negative binomial The count distribution used for RNA-seq, allowing more variability than Poisson.

Null hypothesis The assumption of no effect, against which observed data are compared.

Permutation test Building a null distribution empirically by repeatedly shuffling labels.

p-value The probability of data at least this extreme, assuming the null hypothesis is true.

PCA Principal component analysis: reducing many correlated variables to a few summary dimensions.

Power The probability of detecting an effect that is genuinely there.

Screen and arbiter A permissive first test followed by a stricter confirming test, which must use a different criterion.

Standard deviation The typical distance of a value from the mean of its column.

Survival analysis Methods for time-to-event outcomes that correctly handle censored observations.

z-score A value re-expressed as its distance from its mean, counted in standard deviations.