Data Journalism: A Working Introduction

Tools of the Trade · By Simon Townsend · 4 min read

Simon wrote more than one factsheet warning young reporters against "the big number." A sentence like "more than a million Australians are affected" is lazy. It has no source, no date, no methodology, and almost always no useful meaning. Data journalism is the antidote. It is the discipline of finding, checking, and fairly representing numbers in the news.

You do not need a statistics degree to do data journalism. You do need to be able to read a spreadsheet, understand the basics of how numbers can lie, and be prepared to pick up the phone when a dataset looks strange.

What counts as data journalism

It is not only features with colour maps and interactive charts. Data journalism is any reporting that depends on a structured dataset for its core claim. A story that says "crime in Geelong is up" based on a police spreadsheet is data journalism. A story that says "the minister was the highest-paid adviser in this financial year" based on a cabinet expenses release is data journalism.

The visible wave of Australian data journalism runs through the ABC's Story Lab, the Guardian Australia data team, the SMH/Age data desks, and projects like the Pulitzer-winning cross-border collaboration work. Most of the work, though, is unglamorous: a single reporter, a single spreadsheet, a single phone call to confirm what the numbers show.

The skills that actually matter

Spreadsheet literacy. Not just typing into cells. Sorting, filtering, pivot tables, VLOOKUP, percentages, percentage changes. Google Sheets is free and good enough. Excel is the industry standard. Either one will do.

Basic statistics. Understand mean, median, mode. Know when an average is misleading (almost always, when the distribution has a long tail). Understand the difference between a count, a rate and a ratio. Understand the difference between a percentage change and a percentage point change.

Data cleaning. Most public datasets are dirty. Inconsistent spellings, missing values, mixed-case categories, trailing spaces. OpenRefine will save you hours. Regex will save you more.

Sourcing. Know where Australian data lives. ABS (abs.gov.au), Data.gov.au, Data NSW, Data VIC, Data SA, local council open data, AIHW (Australian Institute of Health and Welfare), AEC for elections, AFSA for insolvency, ASIC for company records, ATO for tax stats, PBS and MBS for Medicare data.

Visualisation. Start simple. A bar chart, a line chart, a map. Datawrapper and Flourish are embed-ready and free up to a point. Don't reach for complex interactives before you can make a single chart a reader can read in five seconds.

How numbers lie, and how to catch them

Base effects. A 100% increase from 1 case to 2 cases is still a rounding error. Always report the absolute numbers alongside the percentage change.

Cherry-picked start dates. A dataset that runs from the peak of a trend will always show a decline. A dataset that runs from the trough will always show a rise. Plot the full history, not just the slice that supports the headline.

Misleading denominators. "Per capita" and "per patient" look similar and mean different things. Per capita means per person in the whole population. Per patient means per affected person. Check which one is being used.

Correlation confused with causation. This is the oldest trick. Just because two lines move together does not mean one caused the other. Always ask what else could explain it.

Outliers not flagged. A school with unusual NAPLAN results is not necessarily a great school. It might have a tiny cohort where one student's score moved the average. Always ask for the underlying distribution.

Rates that ignore exposure. Road fatalities per year in Australia tell you less than road fatalities per billion vehicle kilometres travelled. Crime rates per area tell you less than crime rates per resident and per visitor.

The verification loop

Before you publish any number, answer these in writing:

Where did this number come from?

What is its source publication date?

What is the population and the period it measures?

What caveats does the source attach to it?

Have I independently replicated the calculation from the raw data?

If I am quoting a percentage, do I have the underlying count?

If I am quoting a ranking, do I know how the ranking was constructed?

Any number without answers to those questions is a number that should not be in your copy.

Working with methodology experts

Most of the good data journalism in Australia in the last decade has involved working with academics, statisticians, or domain experts. They will tell you when your analysis is wrong. Return the favour by crediting them properly and by not misrepresenting what they said.

If your story rests on a single analysis, show the methodology in a sidebar or an explainer. The more transparent you are about how the numbers were built, the harder it is for critics to argue with the conclusion.

Where to learn

The Australian Data Journalism Academy runs occasional workshops. The Walkley Foundation runs training. Journalism schools at Monash, RMIT, Sydney, UTS and QUT all now include data units in their programs. The Data Journalism Handbook (free, online) is a good working reference.

Simon's version of this advice, delivered down the phone: if you cannot explain the number in a sentence, the number is not ready. Go back to the spreadsheet. Work it out. Then file.

Reprinted from The Wonderful World of Journalism. Written in the spirit of Simon Townsend's journalism craft advice. Visit simontownsendjournalist.com for the full archive.