Dealing with Numbers

Let’s start with a note I wrote about datajournalism and numbers back in 2016:

*I got into data by accident. After 20 years in the BBC, I was working as a freelance trainer when I happened to talk to someone who was working on a complex data-driven project without knowing how to organise the data they were collecting.*

*After years of using data and spreadsheets as a management tool I retrained myself to use spreadsheets as a journalist – to ask questions of the data, and know what the answers meant. As I did so, I soon realised that there was a huge ecosystem data journalism I hadn’t been aware of.*

*Up until 5 years ago or so, the situation in the UK reminded me of where we were when the web was just taking off: some journalists were rapidly learning to code, others preferred ready-made tools, and others were not quite sure what data could do for them, or what they could do with data.*

*That has changed since I first wrote this short intro in 2016 – but one remaining problem is a general confusion between data and statistics. Journalists tend to come from arts backgrounds and many are afraid of numbers and statistics. That still needs to change.*

*One thing which has not changed is that a huge amount of UK government data is being published with barely anyone noticing. This contains potential stories, leads to stories, evidence for investigations, evidence of what is really going on in society, rather than what we are told is happening. At the other end of the scale government departments are failing to open up useful data, or not publishing it fast enough. That failure is a big story too!*

*Meanwhile we are seeing the rise and rise of the infographic and data visualisation in general. On bad days I am reminded of the early days of powerpoint – people just pressing buttons to make use of the bells and whistles without paying attention to what it is they want to say.*

*There is still a challenge of how to raise data literacy, and statistical literacy, of journalists and audiences alike. When that happens data journalism will be truly mainstream – normal journalism driven by data telling vitally important stories based on facts and evidence. We all need to be up to data.*

Perhaps you are one of those people I referred to as having a background in arts or humanities – you may even consider yourself “number-phobic” or averse to maths. One colleague has wondered aloud whether journalists are allergic to numbers.

You cannot escape numbers, but there’s no need to be afraid of them. We will, in the course of our data crunching come across numbers – raw figures, percentages, totals, averages and so on. And these numbers may well be essential to the story we are trying to tell. So it’s worth thinking of a few simple rules to help you in assessing the importance of the figures you have, and therefore the place they will have in your story.

William Blundell gave similar advice on handling numbers in his book ***“The Art and Craft of Feature Writing”:*** *We know that too many numbers are poison, so the writer's first impulse should be to omit unessential ones. But that's as painful as root-canal work to those of us whose days are filled with numbers-numbers vital to the breaking corporate and financial news stories inside The Wall Street Journal. In many such stories, numbers define the news or are the news.*

*It's only natural, then, that some writers come to believe numbers per se possess a magical power of definition. They collect statistics by the bushel and, having gone to such pains, use them at the slightest excuse in every story they do. Then they wonder why editors find their features stupefyingly dull.*

*A wrenching change in attitude is required when doing features, where story values must be defined in other than just numerical ways. This change is easier to make if the writer remembers that fiction writers, who could bury us under invented figures to lend definition to their tales, never do so. They know better.*

*I don't imply that we should omit meaningful statistics to avoid boring readers. This would sacrifice substance for form, an error that a novelist may get away with but that we cannot. We need numbers in almost all our stories, and in some a number may be so important or startling that omitting or generalizing it would weaken the whole piece. I only argue that we be choosy in selecting figures and careful in their treatment.*

*In placing numbers in a story, the good writer tries not to stack too many in one paragraph; this builds a wall of abstraction difficult to breach. It becomes impossible to breach when two or more such paragraphs are butted together, a construction that may lead to more unread prose than any other writing fault. Don't do this. Don't ever do this.*

*The good writer also recasts as many numbers as he can in a simpler or more pictorial form that removes some of their abstraction. If a precise figure is not important, he rounds it off: $2.6 million is cleaner and easier than $2,611,423. If something increased by 36.7%, he may say it went up more than a third. If it increased 98%, he says it almost doubled. These expressions are pictorial in that they let the reader visualize a slice of a pie or two pies where there was one before.*

*Use ratios to simplify large numbers. Instead of saying that 14,654,231 American drivers out of a total of 58,013,261 own foreign cars, a writer may simply say that one in four owns a foreign auto. Smaller numbers can be grasped while large ones remain abstract.*

*“Spending on redundancy research by the Office of Unessential Affairs rose from $847M in 1983 to $1.26bn this year – a 49% increase”*

*“Over the past fiscal year the OUA increased spending on redundancy research by almost half, to $1.26bn”*

So – when you’re handling numbers or statistics, try to bear these rules in mind:

**Who’s counting? Who’s asking?**

Numbers may appear to be hard objective facts, but it matters who is doing the counting, what they are counting, and why. How many children live in the UK? Define a “child”. Age may be the main factor: Are we counting under 18, the legal age of adulthood? under 16 – the point at which a young person can have sex, get married, join the milirary (not necessarily all at once) ? under 10 – an age at which behavioural scientists consider someone to become aware of rights and wrongs and so become capable of criminal responsibility? Maybe it’s not a question of age, but of responsibility or independence?

Perhaps it will help to know why you’re counting children – are you deciding how much financial support to award families depending on the number of children in a household? Perhaps you’re trying to work out the number of children living in poverty – and if you want to make the number seem low, you might decide to define a child as someone still at primary school. If you want the number to seem higher, you count everyone under 18.

**How big is the number? Compared to what?**

The government announces a new initiative, to which it is allocating £50 million. Sounds a lot to me as an individual citizen. But let’s look at the detail – the allocation is £50 M over 5 years, so £10 M a year. And the number of people affected by the initiative – let’s say 25 million. So that’s 50p a head – 10 p a year over the 5 years. Now how big does the sum seem?

**Going up? How fast? And for how long?**

Most figures don’t live in isolation for a moment in time – they are part of a pattern: numbers go up, and they go down. Some go up rapidly, others fall gently. We need to compare figures to other figures – spending on health, say, is higher this year than it was last year. During the Covid pandemic, for example, it has been useful to get a handle on death tolls by comparing the number of people who died in one month with the same month a year ago. But is one year enough? Might it be even better to compare with the same month in five previous years? Or ten?

How far we look back to compare figures will often depend on what figures are available. But it’s important to avoid accusations of cherry-picking. It’s often tempting to look at the performance of a government since it was elected – but many trends take a long time to change – unemployment, say, or inflation, won’t go up or down the day after a new government is installed or a new policy is announced. When would be a good time to say the new policy has, or hasn’t, worked?

**Cause and effect**

Talking of trying to work out when a new policy began to bite, the whole question of calculating cause and effect is worth a book of its own.

To put it at its bluntest – just because two variables appear to be linked doesn’t mean that they are. It’s easiest not to get involved at all – let your audience make the connection if they want, but don’t try to work it out for them. In Michael Blastland’s book, *The Tiger that Isn’t*, the author uses the example of a housing estate in the English West Midlands where fourteen people died of a rare brain cancer, and the cause seemed obvious. As the [Daily Mail](https://www.dailymail.co.uk/health/article-1027699/14-die-cancer-seven-years-living-phone-mast-highest-radiation-levels-UK.html) reported in 2008 “14 die of cancer in seven years living next to phone mast with highest radiation levels in UK. Fourteen people living within a mile of a mobile phone mast that emits one of the highest levels of radiation in the country have died of cancer. Four of the deaths have been in a cul-de-sac yards from the site.”

Tragic though the case is, and although the cause seems obvious, the radiation from mobile phone signals is not necessarily to blame. Fourteen deaths in seven years seems too many to be a coincidence. But in statistical terms it’s just not “significant”. That may sound callous when discussing fourteen dead human beings. But it’s true. I am not going to discuss [statistical significance](https://en.wikipedia.org/wiki/Statistical_significance) in detail here, but the crucial paragraph in that linked Wikipedia entry says –

“To determine whether a result is statistically significant, a researcher calculates a p-value, which is the probability of observing an effect of the same magnitude or more extreme given that the null hypothesis is true. The null hypothesis is rejected if the p-value is less than (or equal to) a predetermined level, α . Alpha is also called the significance level, and is the probability of rejecting the null hypothesis given that it is true. It is usually set at or below 5%.”

In other words, if we would need to set out a hypothesis about the mobile phone mast’s putative link with cancer. The hypothesis would need to specify proximity to the mast, and a period in which the people living there were to develop, or not develop, cancer. (We would probably also need to talk to oncologists to decide which cancers it might, reasonably, be linked to. And so on)

And then we would need to compare our chosen population with another population, of about 700, with similar demographics, NOT living close to a phone mast (preferably several sample populations, some close to masts, others not.)

When all the necessary data has been collected and the analysis done, we would expect to see at least a 5% difference in the incidence of cancer between the populations living near masts, and those not living near masts.

Sound complicated? It’s certainly a lot more rigorous than simply declaring the mast to be the cause of 14 cancers in 7 years.

As a datajournalist with no statistical qualificiations, when you see events which seem as clear cut as this one, your first thought should be to doubt the link, and to consider the possibility, the likelihood even, that for every housing estate with a horror story like this one, there will be others where there is no cancer at all.

![PowerPoint

Description automatically generated with medium confidence]()

Here’s another, even more contentious and complex example:

Do speed cameras prevent accidents, or cause them? To answer this question, what would you need to know? What data would you collect?

* Details of the roads where cameras were installed after a spate of accidents?
* Numbers of accidents afterwards?
* But how long do you want to go back before the cameras were put in?
* And how long after? Days? Months? Years? How many?
* And, anyway, how many accidents is a “spate”? 3? 4? 10?
* Within how many metres of each other?
* And within how many days or months of each other? Whatever you decide, you should then look at similar stretches of road where there were accidents, and no cameras were installed. Did numbers go down or up?

The most obvious problem in deciding whether a particular camera reduced accidents is the [“regression to(wards) the mean”](https://en.wikipedia.org/wiki/Regression_toward_the_mean) - a spate of accidents on a given stretch of road, can reasonably be expected to be followed by a safer period, of fewer to zero accidents even without installing a camera. Clearly there are some places which are dangerous, but the problem then is spotting the true danger spots within a mass of data. (Not to mention the ethical issue that arises if you find two danger areas and pick one to have a camera, and one to remain without a camera just in order to work out whether the presence of a camera makes the spot safer or more dangerous!)

Back in 2010 the RAC Foundation published a [report by Professor Richard Allsop](https://www.racfoundation.org/research/safety/effectiveness-of-speed-cameras) entitled The Effectiveness of Speed Cameras. Prof Allsop made decisions about the data to use along the lines described above.

It’s worth reading the [report](https://www.racfoundation.org/wp-content/uploads/2017/11/efficacy_of_speed_cameras_allsop_181110.pdf) in full. But here is an extract from the conclusions:

![Table

Description automatically generated]()

Of course, that’s only a very brief summary of a much longer report. Even expanding the screenshot gives a little more detail:

![Table

Description automatically generated]()

The evidence and the judgements of the data’s significance is weighed up and expressed with great care. So carefully, you might argue, that people on each side of the debate could use it as evidence for the side they took. Take your pick – The Guardian reading was that cameras save lives:



Two diametrically opposed views from the same reporting of the same data!

(As you can see in the RAC Foundation post in the first link, Prof Allsop re-examined the question in 2013 and 2019. And you will see how he and others continue to consider the available data in different ways. We don’t have space to consider all the details here, but you may want to practise with some of the datasets linked to from the RAC Foundation site.

***If you remember nothing else, remember this – data can tell us what has happened. It cannot tell us why. And if you are tempted to link one event with another, get a statistician to advise you on what conclusions you can safely draw from the data you have.***

**Further reading**

Michael Blastland – The Tiger that Isn’t

Daniel Kahneman – Thinking Fast, Thinking Slow

Daniel Levitin – A Field Guide to Lies and Statistics

David Spiegelhalter – The Art of Statistics