Joel Toledano

Statistics - “Lies, Damned Lies”?

Nowadays, a powerful statistic is the cherry on top of any compelling argument. In an increasingly data-driven world, people are certainly more likely to subscribe to a particular belief if it is supported by an accompanying statistic, often without considering the reliability of the source. But why? And do we take facts and figures for granted without really considering their full meaning or origins? A statistic appears at first as a statement of fact - an objective piece of information that is not up for debate. However, the world is not so easy to model - often statistical models involve huge assumptions (e.g. the SIR disease model assumes that infected people recover/die at a constant rate, when of course this depends on factors such as age and underlying health), as well as large amounts of error and uncertainty, not to mention the hard task of avoiding bias. As we will see, statistics can provide important insight into a seemingly complex situation, but data must be treated with care. Statistics can fuel the opinions of the masses, so you’d hope that those who publish them are confident in their truth, but there are countless examples of malicious (or incompetent) distribution of false (or misleading) data in order to accommodate a particular claim, or create a shocking news story, or even benefit a company. While this essay is certainly maths-focused, statistics is an area of maths which requires a framework in order to be effectively discussed (the statistics must be about something). Due to the absolute necessity for data within medical research, it provides a very potent framework for this discussion of statistics, but this essay will aim to remain focused on the statistics themselves, rather than the actual science.

Before the pitfalls of data can be discussed, we need to clarify how important it is to be certain as to what a statistic actually represents. On 20 January 2021, in relation to the ongoing Covid-19 pandemic, Professor Alice Roberts of the University of Birmingham tweeted that '1820 people died today' and that 'we should all be angry'. Her opinion was clearly well received, with nearly 10,000 retweeting her words. However, this is actually far from accurate, for two main reasons. Firstly, the UK government adds the death of anyone who tested positive for coronavirus and then died within 28 days of the test to the death toll of the day they died. Of course some of these deaths will be due to the virus, but many of these deaths will have been caused by factors unrelated to the positive test that the person received. Secondly, there is a reporting delay when gathering the total number of deaths and then releasing them to the media, which results in nearly every Wednesday being a local peak (such as 20 January). Clearly there were not nearly 1820 coronavirus-related deaths that day, and while the numbers have been appallingly high at times, they have often not been as high as reported.

This method of counting deaths clearly relies on mass testing, and so someone who dies of coronavirus but was never tested for it is actually not included in the tally. Mass testing has only

recently been rolled out on a large scale nationally, and therefore there was a systematic undercount of deaths right the way through the first wave of the virus (when there wasn't mass testing of the same scale), and thus many coronavirus-related deaths were not counted. For instance, according to the government data dashboard, the peak of the first wave was 1073 deaths on 8 April 2020. But in fact, if we look at ONS (Office for National Statistics) data, which refers to who actually died on that day and whether their death was labelled as being due to Covid, we see that there were actually 1456 deaths that day, which is nearly 36% higher. It is worth noting that this pandemic has caused many kinds of death, not just deaths due to the virus. For example, many people haven't received the medical treatment they otherwise would have, not to mention those too afraid to go to hospitals in fear of catching Covid there. Should these people be counted in the death tolls? There is certainly an argument that they should, but the fact is that they are not currently (in the UK), and so the figures we are discussing here don't even tell the full story. As we can see, it is incredibly hard to define whether the second wave ever actually surpassed the first wave, as, even aside from counting issues, the death count depends on how we even define a coronavirus death.

Statistics are used throughout medical research (and an epidemic response) to provide clarity and information, but as we have seen, we must question their genuine truth. Russia has been accused of releasing 'smoothed, rounded, lowered' figures throughout the pandemic, but discrepancies within their published data have not gone completely under the radar. The demographer Alexei Raksha claims he was forced to leave his job at Rosstat (state statistics agency) after working there for six years due to being too 'vocal' about the pandemic. Raksha believes that excess mortality is the best indicator of the true impact of Covid. Let's consider the city of Perm in Russia as an example. In 2020 there were 40,123 deaths, compared with 34,440 in 2019, resulting in 5,683 excess deaths. However, according to Rosstat, only 2,388 people died in the Autumn surge of Covid in Perm, and to make things even more suspicious, the city government count has only reached 1,869 deaths for the entire pandemic. These numbers are overtly inconsistent, yet Russian officials somehow still maintain denial over any form of altering or 'fixing' data.

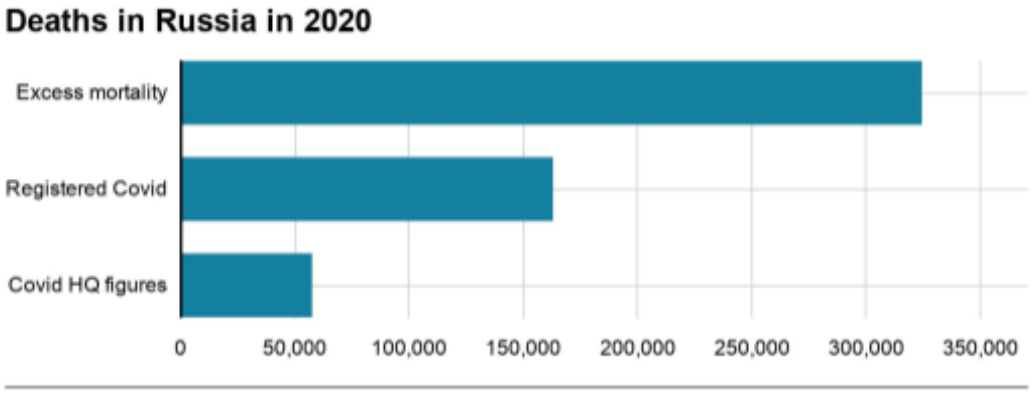

Fig. 1: Deaths in Russia in 2020. Source: BBC (Rosstat and Government Covid HQ)

Disease testing, and how reliable a test is, has become incredibly topical within the last year. But the counter-intuitive thing is that, the rarer the event in your population, the worse your test becomes, even though it is the same test. The ELISA blood-screening test for the virus HIV is incredibly accurate, testing positive for 95% of people who actually have the disease, and testing negative for 99% of people that don't have the disease. Let's consider area A, where the infection rate is at about 1.5%. If we test 10,000 people, we therefore expect 150 to actually be positive, and the rest (9,850 people) to be negative. But we will actually get (0.95*150 = 142.5) true positives, and (0.01*9,850 = 98.5) false positives, giving a total of 241 positive tests. So if you test positive in area A, the chance of you actually being positive is equal to (142.5/241 = 0.591) 59.1%, despite the fact that the test has a 95% accuracy. Now let's consider area B, where the infection rate is only 0.1%. Once again we test 10,000 people, but this time we expect 10 people to be positive, and 9,990 to be negative. We will get (0.95*10 = 9.5) true positives, and (0.01*9990 = 99.9) false positives, giving a total of 109.4 positive tests. Therefore if you test positive in area B, the chance of you actually having the virus is equal to (9.5/109.4 = 0.0868) 8.68% - a huge decrease from area A. When the media have reported on lateral-flow tests or PCR tests (or any other Covid test) in the last year, this has often been accompanied by a percentage. But as we have just seen, the 95% accuracy of the ELISA HIV test is largely irrelevant, as the true chance of being positive (once having tested positive) stems from the infection rate in the area in which you live. And so this is another clear example of a misleading statistic.

Earlier, we discussed how there have been two completely opposite problems of Covid death counting (overcounting and undercounting), yet both have been seen in the same country during the same pandemic. The issue is that the headline (such as '1820 people died today') is what goes viral, not the accompanying explanation. The media are often guilty of twisting statistics for the sake of a more extreme story, as the nature of their business model relies partially on shock factor. This isn't necessarily based on a malicious desire for the general public to be heavily misinformed on matters of science, but often this can be the residual effect. Let's consider an example. Out of a sample of one hundred men in their fifties who take no painkillers, four of them go on to have heart attacks. But from a sample of one hundred other men in their fifties who do take painkillers, six of them go on to have heart attacks. If you are a journalist, how should you phrase this statistic in your headline? You have a few options. Firstly, the absolute risk increase, which in this case is that painkillers increase the risk of a heart attack by 2% (2/100 = 2%). Secondly, the relative risk increase, which in this case is that painkillers increase the risk of a heart attack by 50% ((6-4)/4 = 50%). It seems contradictory, but this situation could be correctly described as a 2% increase and as a 50% increase. So journalists tend to display the relative risk increase, as it makes for far more eye-catching headlines. There is also a third option, which would be to say that painkiller usage results in an extra two heart attacks per 100 men (assuming that painkillers are the cause and that this is not just a coincidence). This is known as

the natural frequency, and is certainly the most informative option, as well as the clearest, as it does not involve any proportions or probabilities, just simple numbers. The media ought to present such a statistic by using the natural frequency, but alas, they don't.

As it happens, this example isn't completely hypothetical. In early 2008, various newspapers ran stories claiming that painkillers were increasing the risk of heart attacks, but the inclination to choose the 'most shocking' figure led many of these reports to be utterly inaccurate. These reports were based on a study which, using natural frequencies, indicated that one extra heart attack would be expected for every 1,005 people taking ibuprofen. But almost all of these reports (such as the Daily Mail's 'How Pills for Your Headache Could Kill') insisted on publishing the relative risk increases. 24% for ibuprofen, 55% for diclofenac. The Evening Standard and the Daily Telegraph did, to their credit, publish the natural frequencies, but the Mirror got it all wrong when they reported that one in 1,005 people on ibuprofen 'will suffer heart failure over the following year'. Heart attacks are not the same as heart failure, and it's one *extra* person in 1,005 (on top of the heart attacks you'd get anyway) - this is a total manipulation of the true findings.

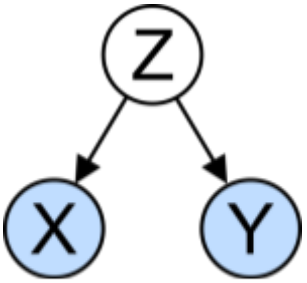

Fig. 2: confounding diagram

Clearly, a surface level reading is not sufficient to effectively interpret data. Even once we are clear on what a statistic actually represents, and we are sure (somehow) that it is an honest one, there are still many hurdles to cross. A great illustration of the need for in-depth analysis is confounding. It has long been reported that excessive coffee drinking (X in figure 2) could increase the risk of developing pancreatic cancer (Y in figure 2). However, multiple recent studies and meta-analyses seem to have found significant evidence of a confounding variable - smoking (Z in figure 2). A confounding variable affects both the independent and dependent variables. In other words, Z affects both X and Y - so while X and Y may appear to have a correlation, the data is actually confounded by Z. Many chemicals in cigarette smoke are well-known to be carcinogenic, so this may come as no surprise that smoking shares a correlation with pancreatic cancer. The link between smoking and coffee drinking may be harder to prove, and also come as more of a surprise, however it has often been traditional to take a 'smoke break' at work with a coffee. It is also worth pointing out that both are highly addictive, so people with more addiction-prone personalities may do both, or commonly people trying to stop smoking may use coffee in place of smoking when they feel an urge. Various epidemiological papers such as the one entitled 'Heavier smoking increases coffee consumption: findings from a Mendelian randomization analysis' (from the *International Journal of Epidemiology*, volume 46, issue 6 (December 2017), pages 1958–1967) have proved this correlation. It is clear to see that without an in-depth, contextual analysis of the data, the wrong conclusion would have been reached - namely that coffee is associated with pancreatic cancer -

but by considering the data within the context of the world around us, it became clear that smoking was actually a confounding variable.

There are many other statistical phenomena which may cloud our judgement when we interpret data, one of which is Simpson's paradox (named after E.H. Simpson). In Simpson's paradox, a correlation appears in multiple data sets, but when the sets are combined, the correlation disappears (or even reverses). This can be represented mathematically for some whole numbers a,b,c,d,A,B,C,D as:

1. $a/b < A/B$
2. $c/d < C/D$
3. $(a+c)/(b+d) > (A+C)/(B+D)$.

We'll have to leave the world of medical research at this point, as the most famous example of Simpson's paradox comes from UC Berkeley's admissions in 1973. Across the university, there seemed to be a gender bias, with 44% admissions for men, but only 33% admissions for women. However, if we delve a bit deeper, and consider individual departments, we can see from Figure 3 that in fact there were more departments where women were actually more successful than men rather than the other way round, which would have been the intuitive guess. How can this possibly be? The answer is firstly that the majority of the women applied to the more competitive departments. But also, fewer women applied than men overall. This means that in the departments where women do have a higher acceptance rate, there aren't actually that many women (relatively) being accepted (such as departments A and B), and so this does not contribute as greatly to the overall acceptance rates as much as department C does, for example.

	Male		Women	
Department	Applications	Admitted	Applications	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**
Top 6	*2,691*	***44%***	*1,810*	*28%*

Fig. 3: 1973 UC Berkeley admissions. Source: Webalytix

As we have seen, data can be confusing, misleading, or even utterly false, but can it be harmful? A solicitor named Sally Clark was put on trial in 1999 for murdering her two babies. Professor Sir Roy Meadow, an expert in parent-child abuse, was called to give expert evidence, and he famously quoted

‘one in seventy-three million’ at her trial as the chance of two children in the same family dying of sudden infant death syndrome (SIDS). There are two huge problems with this piece of ‘evidence’. Firstly, the ecological fallacy - the figure was calculated as if the two cases of SIDS were independent, but its pretty obvious that they aren’t, due to both genetic and environmental factors, as the babies were of the same family. The second issue is known as the prosecutor’s fallacy. What does one do with this figure? Many at the time cited this figure as the probability of the deaths being accidental, and therefore the probability of Clark’s innocence. But it isn’t nearly as simple as that. Two babies in the same family dying is incredibly rare. Usually, both explanations - double SIDS or double murder - would be incredibly rare. However, once the two babies are dead, both explanations become incredibly likely. At the appeal, the judges suggested that Meadow should have said ‘very rare’ instead of ‘one in seventy-three million’, recognising the ecological fallacy, but once again missing the prosecutor’s fallacy. The entire court process missed this essential nuance of the prosecutor’s fallacy twice. Clark was released in 2003 following a second appeal, having served more than three years of her sentence. She was left with serious psychiatric problems, and was found dead at home in March 2007 from alcohol poisoning. An innocent life was utterly ruined by a miscarriage of justice. Statistics can be very harmful indeed.

The fact is that unlikely events do happen. Double SIDS and double murder are both incredibly rare, but one of them must have happened (probably double SIDS) in the Sally Clark case. Someone wins the lottery every week, and there are people in the world who have been struck by lightning. Consider the thousands of people who predict the ups and downs of the stock market. They will all have their various methods for doing so, and some of these methods may be less viable than others. They may look at specific companies’ market possibilities, or cost-to-earnings ratios, or patterns on graphs, or even just astrology. But just by pure chance, due to the sheer number of people ‘playing the game’, some of these people - even some of the astrologers - will be right, as if by accident. And that’s just life - unlikely things do happen. Statistics are sometimes used as a prediction tool, but this speculation has its own hurdles. It is a common misconception - known as the gambler’s fallacy - that if an event has been occurring more than expected in the past, then it will occur less in the future (or vice versa), assuming the events are independent of each other. But if the events are independent, then of course this can’t be the case, as past events would have no effect on future events. With so many factors at play in the world, so many different variables, it’s really quite impossible to predict anything with full confidence. A low probability of an event occurring is not at all equivalent to a zero probability of occurrence, and as we have seen, events with low probabilities do occur. It would be even more surprising if nothing unlikely ever happened. No coincidences at all. Because if we consider the infinitely many possible random coincidences that could conceivably occur, we would then expect at least a few to actually happen. Would it not be quite incredible if you never bumped into someone that you knew while out in public during your entire life?

Let's say you have just conducted an experiment. When analysing your data, you notice a particularly strong correlation, one which could enhance a research field in some way, and so deserves to be published. You're sure that it's both accurate and honest, and you are going to describe it in a clear, unambiguous way. You've checked for any confounding variables, or any potential sneaky paradoxes or fallacies, and you've even carefully avoided any form of bias. Then by all means, publish your data and findings. The world becomes a richer, more knowledgeable place whenever anyone shares a new discovery of their own with the rest of us. Scientific developments are a hugely important part of societal progression, and statistics are an essential mathematical language for expressing these findings effectively and clearly. But as we have seen, it is important to be cautious when reading statistics. Ask yourself, where has this figure come from? What does it really tell me? And if you're a journalist, just please make sure not to only use the relative risk increase in your headline.

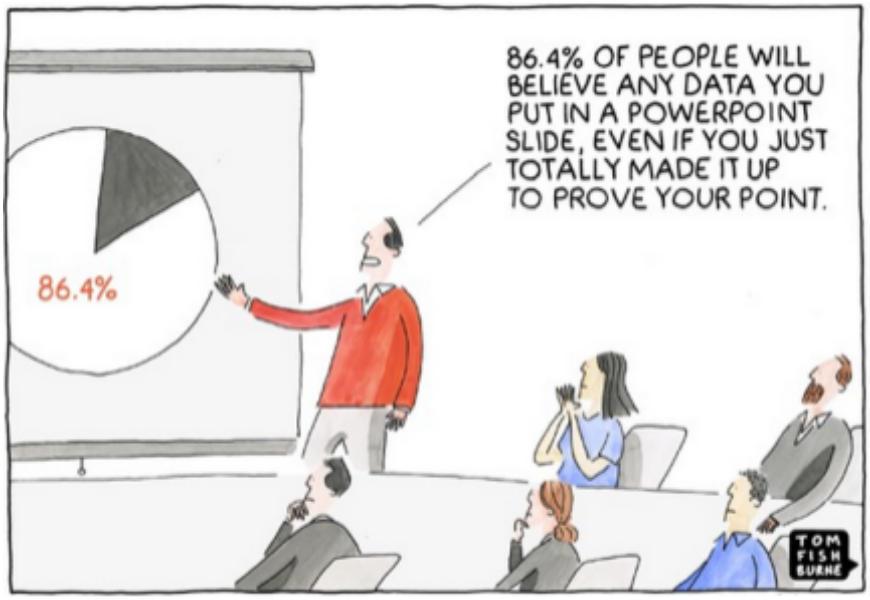

Fig. 4. Source: Marketoonist

References

https://services.math.duke.edu/education/ccp/materials/diffcalc/sir/sir2.html
https://www.bbc.co.uk/sounds/play/m000rln5
https://www.bbc.co.uk/news/world-europe-56454701
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6767387/
https://academic.oup.com/ije/article/46/6/1958/4082629
https://cebp.aacrjournals.org/content/25/6/951
https://plato.stanford.edu/entries/paradox-simpson/
https://www.webalytix.co.uk/simpsons-paradox-and-segmentation-why-analysis-is-crucial-draft-irina/
https://www.theguardian.com/society/2007/nov/08/childrens

https://www.investopedia.com/terms/g/gamblersfallacy.asp
https://cmotions.nl/en/5-typen-bias-data-analytics/

Bad Science - Ben Goldacre
Coincidences, Chaos, and All That Math Jazz - Edward B. Burger and Michael Starbird
Do Dice Play God? - Ian Stewart