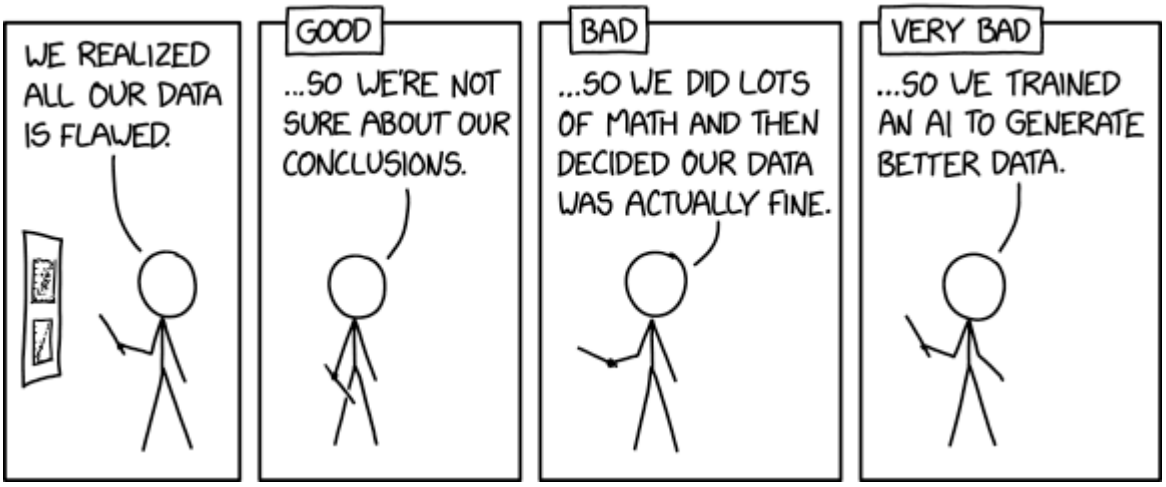
Ch 10: Statistical Arguments

A Little More Logical | Brendan Shea, Ph.D. (Brendan.Shea@rctc.edu)



1 CONTENTS

2 THOUGHT QUESTIONS

1. How would define **statistics?** How important is statistics to your daily life?
2. What are some good (and not-so-good) ways of making statistical "generalizations"?
3. What are some different ways of defining "average"? Have you heard of the **median, mean,** and **mode?**
4. What are some effective ways of *visualizing* statistical data? Do you have an examples of especially good graphs/charts that you've seen?

3 WHAT IS STATISTICS?

Statistics is a "formal" (or "mathematical") science that studies the relationship between populations and samples. While it can get pretty complex, it is based on a number of fairly simple concepts, which we'll be talking about today:

- A **population** is a set of things. Some populations: people in the United States, squirrels currently living on campus, things under my bed, blood cells in my body, and so on.
- A **sample** is a subset of the population. We get samples by selecting some, but not all, members of the population.
- A **random** sample is a sample in which each member of the population has an equal chance of being picked for the sample. Random samples are important because they are thought to be **representative** samples—i.e., the properties of the sample resemble the properties of the population at large. Very few samples are *perfectly* random or representative.
- A **biased** sample is an unrepresentative sample. That is, the sample differs in some significant way from the population. For example, using a sample made up only of college students to draw conclusions about the whole U.S. population would be biased. When dealing with human experiments, potential sources of bias are numerous:
    - The sample should resemble the population in gender, race, age, income level, and other basic demographic categories.
    - For medical experiments, we need to make sure the sample matches the population in health-related issues: the presence or absence of disease, medications taken, amount of physical activity, diet, etc.
    - Many studies that involve human subjects have samples are, at least to some, **self-selected** (where people *choose* to participate). This almost always introduces *some* level of bias, but this need to be minimized as much as possible.

Three Meanings of "Average." Suppose we have the following set of numbers: **1, 2, 2, 7, 10, 11, 25.** If someone asks "What is the average value?" we might give one of three answers.

- The **mean** is obtained by (1) adding up all the numbers and (2) dividing by how many numbers there are. So, (1+2+2+7+10+11+25)/7 = 58/7 = 8.3
- The **median** is the "middle" number when the numbers are arranged in order (as they are above). In the example above, this is 7.
- The **mode** is the most frequently appearing number. In the example above, this is 2.

3.1 MEASURING DISPERSION

Along with figuring out the "average" of the data items, we are often concerned with the **dispersion**, or how close the data are to one another. The **range** is probably the simplest measure of dispersion, and is obtained by subtracting the *smallest data item* from the *largest data item*. For example, if our data set is {1, 2, 3}, the range is 3 – 1 = 2. The **variance (σ^2)** is calculated by determining the *mean squared distance from the mean of the data set*. In other words, it provides us with a measure of how different (on average) each member of the data is from the mean of the data set. While this may sound confusing, it can be obtained by following a step-by-step process (don't be afraid to use a calculator!)

1. Calculate the mean by summing all of the data, and dividing by the number of items.
    a. $\frac{1+2+3}{3} = 2$

2. Make a *new* set of data by subtracting the mean from each of the original items. In some cases, this will be a negative number. Then, square each of these numbers.

2a. Write Down Data	2b. Subtract Mean from Data	2c. Square the result from 2b
x_n	$x_n - \mu$	$(x_n - \mu)^2$
1	$1 - 2 = -1$	$-1 \times -1 = 1$
2	$2 - 2 = 0$	$0 \times 0 = 0$
3	$3 - 2 = 1$	$1 \times 1 = 1$

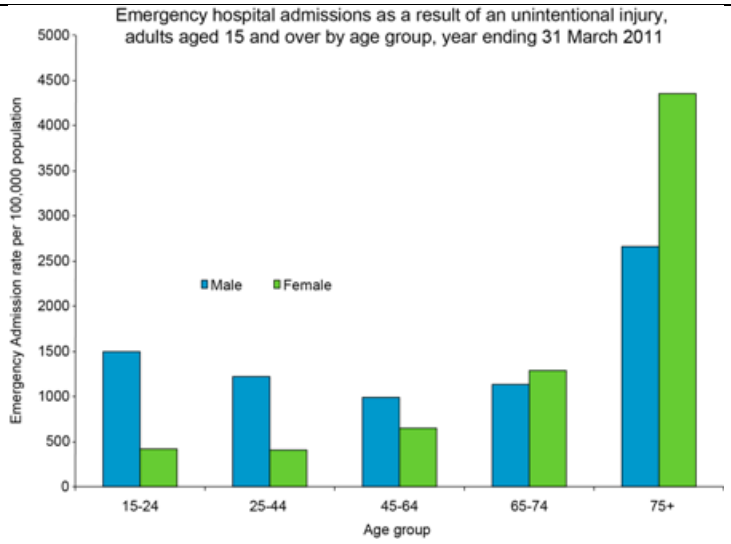
3. Add up the results from step 2c, and divide by the number of data.
   a. $\sigma^2 = \frac{1+0+1}{3} = \frac{2}{3} = .67$ (rounded to two decimals).

The **standard deviation (σ)** is simply the *square root of the variance*. For example, the standard deviation for 1, 2, 3 is $\sigma = \sqrt{\frac{2}{3}} = .81$. (Remembering that is OK to use a calculator for this!) It's important to remember that the dispersion of two data sets can be *very, very* different, even if they have the same means. For example, another data set with a mean of 2 is -200, -100, 306. The range for this is 506, the variance is 35,906, and the standard deviation is 189.5.

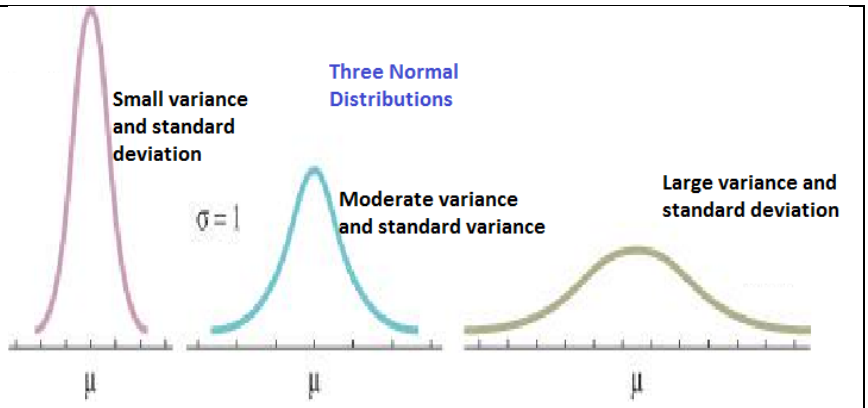
3.2 Understanding Dispersion: Histograms and the Standard Distribution

Statistical arguments (of both the deductive and inductive variety) frequently involve claims about how the data is distributed. The success (or failure) of arguments often hinges on claims about these distributions.

Histograms. A histogram plots the frequency of each data item. That is, it tells you *how frequently* each data item occurred. So, for example, in the following graph the horizontal For example this graph (from the U.K.'s National Health Service) shows that almost 4,500 women over the age of 75 were admitted to the hospital because of unintentional injury. By contrast, less than 500 women between the ages of 15 and 25 were admitted for this reason. The data displayed here can be used in a variety of inductive arguments. For example, we might use it to make a *prediction* about what sorts of patients will be most likely to suffer unintentional injuries in the future.



The Normal Distribution. In many cases, the data we are interested in are distributed *normally,* with around 68% of the data lying with one standard deviation of the mean, and 95% within two standard deviations. Normal distributions with *small* variance will have data clustered very close to the mean, while those with *large* variance will have data that are much more spread out. Again, we can use these data in arguments: in normally distributed data sets with small variance, we can give a strong argument *predicting* that any individual datum will be close to the mean. This same argument would be much weaker for data with larger variance.



3.3 Problems With Percentages

Statistical arguments frequently involve the use of percentages, which can sometimes mislead us into accepting unwarranted conclusions. A few important points about working with percentages include:

Consider the baseline. Whenever a percentage change is reported ("Profits are up 25%!"), it's important to ask: "Relative to what?". If the answer is "relative to last year's already impressive profits," 25% is impressive. If it is "relative to last year's squeaking by with $1 profit," 25% amounts to a pretty large failure.

Remembering baselines, part 2. When dealing with *multiple* changes reported in percentages, its important to remember that each change is relative to the new baseline. So, for example, suppose that Jeanie takes a paycut of 50% one year, but then *increases* her pay by 50% the next year. It's tempting to think "since 50% plus 50% is 100%, she's back where she started!" But this is wrong: in fact, she is only making 75% of what she was originally making. Why? Suppose she made $100,000. After taking a 50% paycut, she was making $50,000 (the new baseline). A 50% raise to *this* baseline put her at $75,000.

Wholes and parts: On not adding percentages. Suppose that you are trying to save money. You decide to cut your spending on food by 10%, entertainment by 30%, and everything else by 20%.. Does this mean you've cut your budget by 60%? No! Since the budget is made up the components, each cut to a part ends up cutting a (smaller) part of the whole.

3.4 Arguments using Statistics: How to Avoid Mistakes

Review: Inductive and Deductive Arguments. In order to assess the success of any argument (including those involving statistical data), it's important to determine what the argument is trying to *do*. If the argument claims that some statistical data *guarantees* the truth of the conclusion, it is **deductive**. By contrast, if the

argument merely claims the that truth of the statistical data makes the conclusion *likely* to be true, it is **inductive.** In many real-world cases, we'll need to make use of both deductive and inductive reasoning to figure out whether a given statistical argument really "works."

Deductive Arguments Using Statistics. An example of a deductive argument would be "Since the three childrens' ages are 3, 4, and 5, their mean age must be 4." Since this particular argument is **valid,** we know that *if* the premises are true, *then* we can be absolutely certain that the conclusion is true as well. This might be categorized as an **argument from mathematics** or an **argument from definition** (we are arguing from the definition of the *mean*). In deductive arguments, you can check the validity simply by taking the time to review the reasoning carefully and correctly (which can sometimes be tricky, especially when working with percentages and statistics). It's also important to remember that the statistical data we start from might be mistaken, and that the argument may be valid but **unsound** (the children might have lied about their ages!).

Inductive Arguments Using Statistics. Once we have used the statistical data to make various deductive inferences, we often want to draw conclusions about things we *haven't* already measured. So, for example, we might predict that "Since the mean age of the first three birthday party attendees is 4, the next child to arrive will probably be between 3 and 5." This involves inductive reasoning, which can be either *weak* or *strong*: Inductive arguments use premises regarding statistical data to make **predictions, generalizations, causal inferences,** and **arguments to the best explanation.** Inductive arguments are also involved whenever we attempt to interpret charts or graphs (**arguments from signs**) or determine whether we can *really* trust the accuracy of the statistical data we are presented with (**arguments from authority**). Finally, unlike deductive arguments, determining the strength and cogency of an inductive argument always requires that we consider the possibility of **suppressed evidence,** and of relevant data that may have been "left out."

3.5 INDUCTIVE GENERALIZATIONS

Many arguments that use statistical data take the form of **inductive generalizations**, which use premises concerning a sample to draw conclusions about a population. One common form is as follows:

- Premise 1: S is a sample of population P. "3,000 randomly chosen likely voters are a sample of all voters in the next election."
- Premise 2: In sample S, the value of parameter A is X. "60% of those polled favored candidate A"
- Conclusion: So, the mean value of A in P is X (+/- **margin of error**%). This is with **confidence level** L. "With a confidence level of 95%, we can predict that between 58.2% and 61.8% of all voters favor candidate A." (This is 60% +/- 1.8%, which is the margin of error for a confidence level of 95% and sample size of 3,000)

SAMPLE SIZE	**CONFIDENCE LEVEL**		
	99.0%	**95.0%**	**90.0%**
10	40.5%	31.0%	25.9%
50	18.1%	13.9%	11.6%
100	12.8%	9.8%	8.2%
250	8.1%	6.2%	5.2%
500	5.7%	4.4%	3.7%
1000	4.1%	3.1%	2.6%
2000	2.9%	2.2%	1.8%
3000	2.3%	1.8%	1.5%

What Does All this Mean? The **margin of error,** or **confidence interval (CI)**, gives a range of possible "difference" between the sample (in our premises) and the population (in the conclusion). So, X +/-1.8% means that the true population value might differ by as much as 1.8% from X. The **confidence level** L is a claim about how accurate the sampling/polling procedure is. 95% confidence claims that, using this sampling

procedure, the true population mean will fall within the margin of error 95% of the time, at least under ideal conditions. However, it's important to remember that conditions are never actually ideal—so, the fact that a study or poll reports 95% confidence never *literally* means "this study deductively proves that there is exactly a 95% chance the true result falls within the margin of error."

Why Should I Care? A good inductive generalization should make it easy for the audience to see the degree to which the premises actually provide support for the conclusion. Because of this, you should be very wary of "polls" or "studies" that don't report things such as the sample size, confidence level, or margin of error. In general, it is important to remember that we can draw conclusions with a higher confidence level by (1) increasing our sample size (**larger sample sizes *strengthen* inductive generalizations)** and (2) allowing for a *greater* margin of error **(highly specific claims with low margins of error *weaken* inductive generalizations)**.

3.6 EVALUATING INDUCTIVE GENERALIZATIONS

When evaluating the strength or weakness of inductive generalizations (and other arguments using statistical premises), it's important to keep in mind a few things

How Representative was the Sample? The confidence level and margin of error only tell you how good the argument *would be* if the sampling procedure were *perfect* and the population is *normally distributed.* This is never exactly true. Look closely, and see if there are any reasons for thinking that the sample chosen by the researchers might not have been representative of the population as a whole. In many cases, researchers try to ensure representativeness by choosing **randomly**, but this is often impossible. For example, many political polling agencies have (in the past) relied on contacting people who owned landline phones. This may no longer be representative of the population in general, however, since many younger people (who might vote differently) rely only on mobile phones. Remember to avoid **Hasty Generalizations** from small or biased samples.

How Does this Fit With Other Evidence? Whenever you see the result of a "surprising" new poll or scientific study in the media, it's important to remember that assessing the strength or weakness of an inductive argument requires that we consider ALL of the relevant evidence, and not just these particular results. For many of us, it's all too easy to focus on those statistical results that seem to *support* our preexisting beliefs, and to (perhaps unconsciously) ignore those that challenge them. In presidential election years, for instance, there are almost always some people who argue that "The polls must be wrong! Except the ones that favor my candidate!" Something similar holds true of research regarding diet and nutrition (with various people focusing on the studies whose conclusions they like), and on many other issues. It's important not to fall into this trap. While polls and studies can be wrong, we shouldn't just believe the ones we *want* to believe. Instead we should take a "holistic" view at the evidence. (For nutrition-related issues, the U.S. government actually does this when it issues dietary guidelines). Relevant fallacies here include **Suppressed Evidence** and **Argument from Ignorance** ("Even though 99 studies say I'm wrong, this one says I'm right! So, who's to say what the right answer is?")

Causation, Correlation, and Alternative Explanations. We often use statistical data to make inductive arguments about *causes.* For example, statistical data about the strong correlation between *smoking cigarettes* and *contracting lung cancer* provides evidence for thinking that smoking *causes* lung cancer. However, it's important to remember that correlation between X and Y doesn't mean that X causes Y. For example, *spending time in the hospital* might be correlated with *chance of death.* But this doesn't mean that spending time in the hospital *caused* death! This is closely related to the **False Cause** fallacy discussed in previous classes.

3.7 DON'T FORGET ABOUT DISTRIBUTIONS!

Statistical arguments don't always involve drawing conclusions about populations from samples. In other cases, we use statistical data about *populations* to draw conclusion about a particular *individuals* or *groups*. When doing this, determining whether a given (inductive) argument is strong or weak will often involve using information concerning about the distribution of the population, and the dispersion of this distribution. For example, consider the following two arguments, both of which assume a normal distribution. In this sort of distribution, around 68% of the population is within one standard deviation of the mean, and 95% is within two standard deviations of the mean:

- The mean age of a high school senior is 18 years old, with a standard deviation of .5 years. So, if I called up a random high school senior, it is probably true that she or he would be between 17 and 19 years old.

- The mean wage of an employee at Firm X is $18/hr, with a standard deviation of $8. So, if I called up a random employee of Firm X, it is probably true that she or he would earn between $17 and $19.

Even though the mean is the same in both cases (18), these arguments are very different. In particular, the first argument is inductively *strong,* while the second one is inductively *weak*. This is because the ages of high school seniors are, on average, much *closer* to the mean than are the average wages of the employees of firm X.

Not All Data Are Distributed Normally! Perhaps because normal distributions are so common in the natural world, we often tend to think and act as if *all* data are normally distributed, even when they are not. In particular, we tend to think that knowing the "mean" gives us information about the "typical" (median) or "most common" (mode) case. But this isn't always true. For example, *wealth* is not normally distributed. Instead, most of the wealth is clustered near the "top" of the distribution (the top 10%, 1%, etc.). So, for example, suppose we learn that the mean wealth of an American citizen is $250,000. This does NOT mean that there are roughly equal numbers of people who are worth more or less than this, as would be the case if wealth were distributed normally. Instead, we should expect that the majority of people are worth less than this, with a few being worth much more. Other examples of non-normal distribution include *annual health care spending* and *alcoholic beverages consumed per week*. So, when you hear that the "average" American spends $9,000 on health care per year, or drinks 11 drinks/week, just keep in mind that *most* people are below these numbers, and that the "top" end of the distribution is much higher (they may spend $100,000/year, or drink 60 drinks/week). This makes a real difference when we consider inductive arguments about how to solve problems related to these issues.

3.8 REVIEW QUESTIONS

1. Calculate the *mean, median, mode, variance,* and *standard deviation* of the following data set: {1, 1, 3, 4, 6}. It's OK to use a calculator (or the calculator app on your phone.
2. What is the normal distribution? What difference does the variance (and SD) make to the "shape" of this distribution?

3. On the website for the local tennis club, you can "vote" for whether or not you approve of the new mayor's performance, given her recent decision to demolish six tennis courts in a local park, and to replace them with playground equipment. You decide to vote, and it shows you the results so far. It says that 100 people have voted so far, and that 70% disapprove. You like the mayor, so you decide to click "back" and vote again. This works fine.

    a. Name at least THREE things that are wrong with using the results of this poll as evidence for the conclusion "the mayor is very likely to lose the election coming up next month."

b. Suppose that you were in charge of designing a poll that would more accurately predict the results of the upcoming mayoral election. How might you go about doing this? How would your poll differ from this one?

4. Identify the following argument as deductive-valid, deductive-invalid, inductive-strong, or inductive-weak. Briefly explain your reasoning.

    a. The final grade in Professor's X class is determined entirely by the mean percentage on three exams. Bonnie received scores of 100%, 90%, and 50%. So, her final grade is 80%.
    b. 60% of the students in Professor X's class on strength training class are capable of bench-pressing at least 200 lbs. So, it is likely that between 58% and 62% of all students at the college can bench at least 200 lbs.
    c. In a survey of 1,000 randomly selected students at the local college, 55% were women. From this we, can conclude that between 50% and 60% of the college's students are women.
    d. I got a 60% on exam 1. My score on exam 2 was 50% higher than this, and my score on exam 3 was 50% lower than my score on exam 2. I can conclude that my score on exam 3 was 60%, just like exam 1.
    e. In his career, Michael has taken almost 10,000 free throws, and has made 70% of them. Given this, we can predict he will probably make *exactly* seven of the next ten free throws he shoots.
    f. Publicly available statistics show that Harvard admits only 5% of applicants. We can conclude from this that, if Bill Gates (the founder of Microsoft) were to decide to apply to Harvard to finish his degree, he would have only a 1 in 20 chance of being admitted.

4 READING: CITIZENS NEED TO KNOW NUMBERS (DAVID SPIEGELHALTER)[1]

It might have come to your attention that the United Kingdom is experiencing some difficulties over an issue called Brexit. The referendum on whether the UK should leave the European Union was held in June 2016, and the Leave campaign produced an iconic image of a big red bus plastered with the message: 'We send the EU £350 million each week – let's fund our NHS instead.' The masterful meme combined an impressive-sounding amount of cash with an appeal to the National Health Service, a nearly sacred British institution. It is plausible that this brilliant use of numbers tipped the balance in favour of Leave, which to most people's surprise went on to win by the narrow margin of 52 per cent to 48 per cent.

How reliable is the claim on the side of the bus? Like most numbers used in political discourse, the £350 million is not purely random or entirely fabricated – it does have some empirical basis. The agreed annual gross contribution to the EU in 2017 was £18.6 billion (£357 million a week), figures easily found in a publicly available spreadsheet. However, it is also true that a rebate of £5.6 billion is deducted from the British bill to the EU before payment. That brings the net figure to £13 billion. Further, around £4 billion comes back from the EU in terms of, for example, public sector science and agricultural funding and, presumably if it left the EU, the UK would need to pay for these itself.

Many criticised the Leave campaign for its claim that Britain sends the EU £350 million a week. When Boris Johnson repeated it in 2017 – by which time he was Foreign Secretary – the chair of the UK Statistics Authority (the official statistical watchdog) rebuked him, noting it was a 'clear misuse of official statistics'. A

[1] David Spiegelhalter, "Good Citizenship Depends on Basic Statistical Literacy | Aeon Essays," Aeon, 2019, https://aeon.co/essays/good-citizenship-depends-on-basic-statistical-literacy.

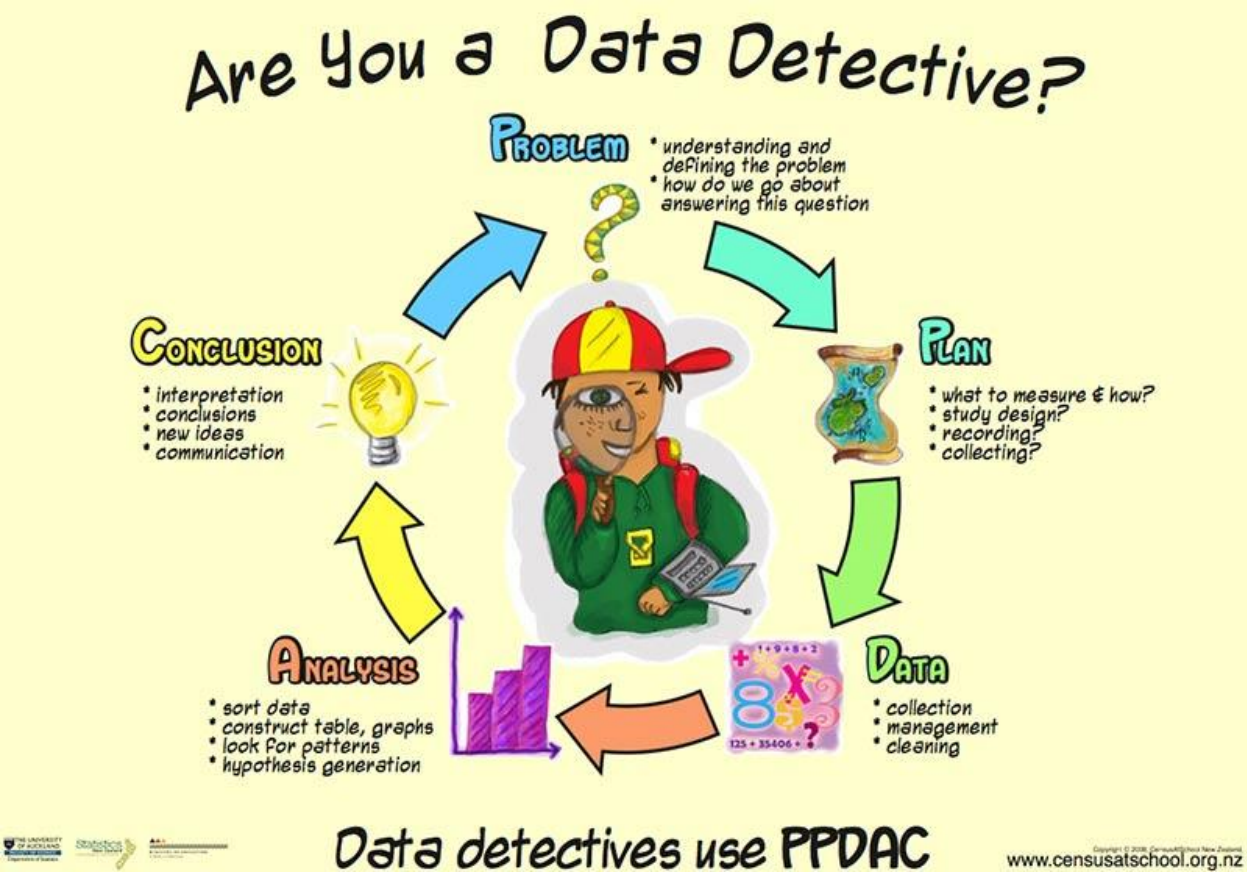
private criminal prosecution was even made against Johnson for 'misconduct in a public office', but it was halted by the High Court.

The message on the bus had a strong emotional resonance with millions of people, even though it was essentially misinformation. The episode demonstrates both the power and weakness of statistics: they can be used to amplify an entire worldview, and yet they often do not stand up to scrutiny. This is why *statistical literacy* is so important – in an age in which data plays an ever-more prominent role in society, the ability to spot ways in which numbers can be misused, and to be able to deconstruct claims based on statistics, should be a standard civic skill.

Statistics are not cold hard facts – as Nate Silver writes in *The Signal and the Noise* (2012): 'The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning.' Not only has someone used extensive judgment in choosing what to measure, how to define crucial ideas, and to analyse them, but the manner in which they are communicated can utterly change their emotional impact. Let's assume that £350 million is the actual weekly contribution to the EU. I often ask audiences to suggest what they would put on the side of the bus if they were on the Remain side. A standard option for making an apparently big number look small is to consider it as a proportion of an even bigger number: for example, the UK's GDP is currently around £2.3 trillion, and so this contribution would comprise less than 1 per cent of GDP, around six months' typical growth. An alternative device is to break down expenditure into smaller, more easily grasped units: for example, as there are 66 million people in the UK, £350 million a week is equivalent to around 75p a day, less than $1, say about the cost of a small packet of crisps (potato chips). If the bus had said: *We each send the EU the price of a packet of crisps each day*, the campaign might not have been so successful.

Numbers are often used to persuade rather than inform, statistical literacy needs to be improved, and so surely we need more statistics courses in schools and universities? Well, yes, but this should not mean more of the same. After years of researching and teaching statistical methods, I am not alone in concluding that the way in which we teach statistics can be counterproductive, with an overemphasis on mathematical foundations through probability theory, long lists of tests and formulae to apply, and toy problems involving, say, calculating the standard deviation of the weights of cod. The American Statistical Association's *Guidelines for Assessment and Instruction in Statistics Education* (2016) strongly recommended changing the pedagogy of statistics into one based on problemsolving, real-world examples, and with an emphasis on communication.

If we are going to use real questions of interest to introduce statistical ideas, then we need to teach a basic structure for how data can be used to solve problems. I have been particularly inspired by the work in New Zealand on improving statistics education in schools; the figure below is theirs.



The problem-solving cycle: from problem to plan, to data, to analysis, to conclusion and communication (PPDAC), and starting again on another cycle.

Take this 2019 headline from CNN: 'Eating Just One Slice Of Bacon A Day Linked To Higher Risk Of Colorectal Cancer, Says Study'. The British tabloid *The Sun* put it more baldly: 'Rasher Of Bacon A Day Is Deadly'. So the *problem* might be: should I care about this risk and give up bacon? The *plan* is to delve deeper into the claims made in the article, which reveals that a study estimates that eating 25g of processed meat a day (equivalent to a large bacon sandwich every other day) is associated with a 19 per cent increased risk of bowel cancer.

For someone with basic training in statistical literacy, two questions should immediately come to mind. First, is this association actually causal – in other words, if people start eating bacon, will their risk go up? Or is it just association, in that people who tend to eat bacon also tend to get bowel cancer? It turns out that processed meats have been causally linked to bowel cancer by the International Agency for Research on Cancer, so we can take that on trust.

The second question is whether this effect is large enough to merit my concern. The '19 per cent increase' is a relative risk, and this form of expressing associations is known to exaggerate the apparent effect of exposing oneself to a risky substance such as bacon. The crucial question is – 19 per cent of what? Without knowing the baseline, absolute risk, we cannot know whether this increase is worth worrying about – after all, 19 per cent extra of hardly anything is still hardly anything. So the extra *data* we need is the absolute risks, and it turns out that around 6 per cent of people will get bowel cancer anyway, even if they don't eat bacon. So what is a 19 per cent increase over 6 per cent?

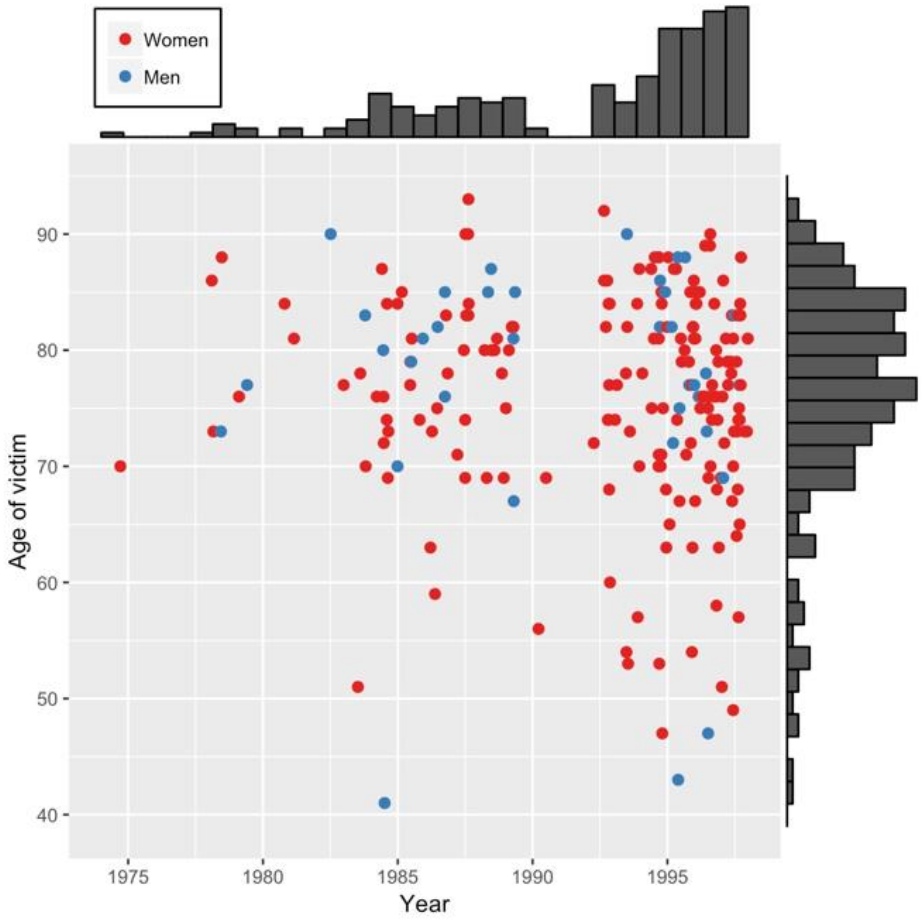
Many reputable psychological experiments show that such calculations are better expressed using the idea of expected frequencies. In other words, what does it mean for 100 people? Well, out of 100 bacon non-eaters, we would expect six to get bowel cancer during their lifetime. Meanwhile, out of 100 people who ate 25g of bacon a day – say a large bacon sandwich every other day – then we would expect an extra 19 per cent to get bowel cancer: that's 19 per cent of 6 per cent, which is around one extra, taking the total to seven. So, in order to get one extra case of bowel cancer, 100 people would need to eat around 180 bacon sandwiches a year for their whole life. This works out at around 10,000 each, or a total of 1 million great greasy bacon sandwiches. So your response to this story might be: pass the tomato ketchup.

Of course, I have deliberately expressed these results to make the risks look negligible, but it does put the claim and the headlines about killer bacon into perspective. Unfortunately, few people in the media are able to do these analyses themselves.

Harold Shipman was Britain's most prolific convicted murderer, though he doesn't fit the typical serial-killer profile. A mild-mannered family doctor, or general practitioner (GP), in a suburb of Manchester, between 1975 and 1998 he injected at least 215 of his mostly elderly patients with a massive opiate overdose. He finally made the mistake of forging the will of one of his victims so as to leave him some money: her daughter was a solicitor, suspicions were aroused, and forensic analysis of his computer showed that he had been retrospectively changing patient records to make his victims appear sicker than they really were. He was well-known as an enthusiastic early adopter of technology, but he was not tech-savvy enough to realise that every change he made was time-stamped (incidentally, a good example of data revealing hidden meaning).

Of his patients who had not been cremated, 15 were exhumed, and lethal levels of diamorphine – the medical form of heroin – were found in their bodies. Shipman was subsequently tried for 15 murders in 1999, but chose not to offer any defence, and never uttered a word at his trial. He was found guilty and jailed for life, and a public inquiry was set up to determine what crimes he might have committed apart from those for which he had been tried, and whether he could have been caught earlier. I was one of a number of statisticians called to give evidence at this public inquiry.

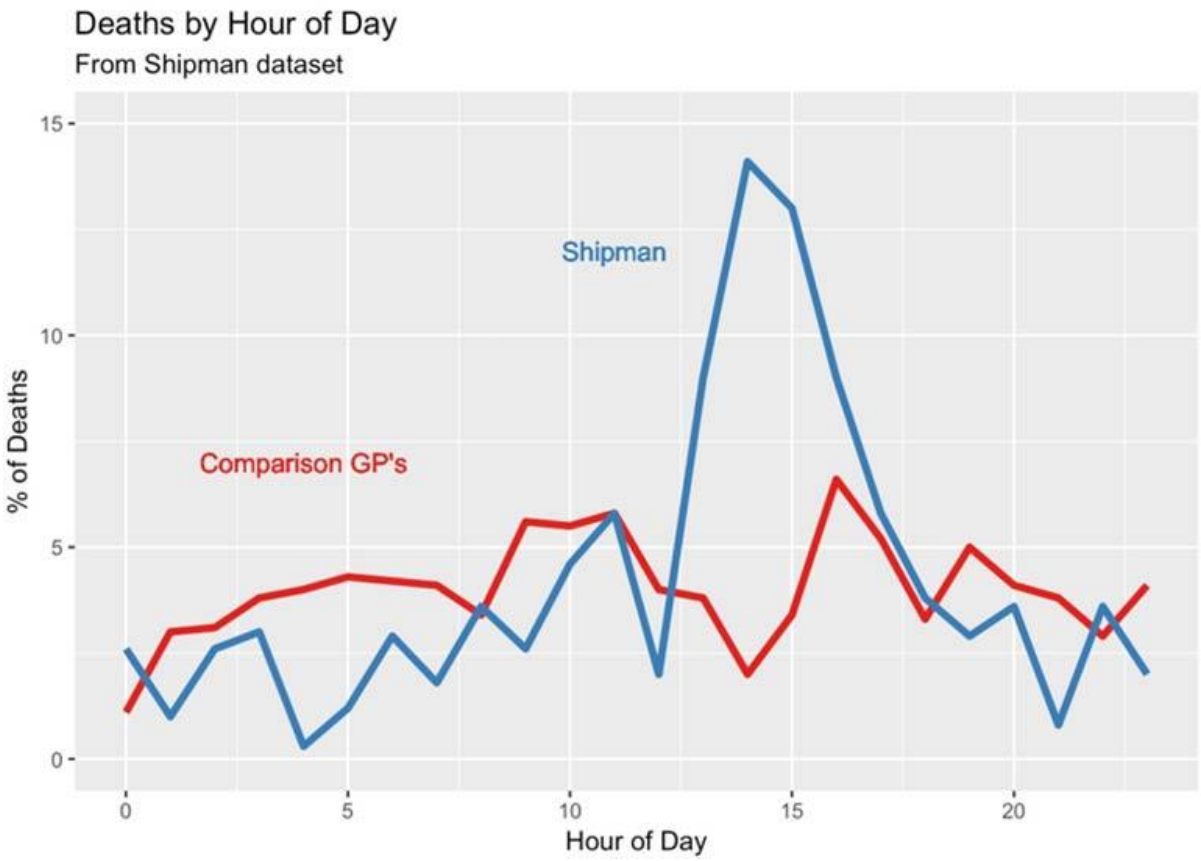
A first *problem* here is simply to understand the pattern of his activities: we can think of this type of iterative, exploratory work as 'forensic' statistics: no mathematics, no theory, just a search for patterns that might lead to more interesting questions. The *plan* is to examine whatever *data* is available from the public inquiry website, comprising details of each victim's age, gender and date of death. The figure below is a visualisation of this data, showing a scatter-plot of each victim's age against the date of death, with the shading of the points indicating whether the victim was male or female. Bar charts have been superimposed on the axes showing the pattern of ages (in five-year bands) and years: there is no formal *analysis*.



A scatter plot showing the age and the year of death of Shipman's 215 confirmed victims. Bar charts have been added on the axes to reveal the pattern of ages and the pattern of years in which he committed murders.

Some *conclusions* can be drawn simply by taking some time to look at the figure. There are more red than blue dots, which means that Shipman's victims were mainly women. The bar chart on the right shows that most of his victims were in their 70s and 80s, but looking at the scatter of points reveals that, although initially they were all elderly, some younger cases crept in as the years went by. The bar chart at the top clearly shows a gap around 1992 when there were no murders. It turned out that Shipman had previously worked in a joint practice with other doctors but then, possibly as he felt under suspicion, he left to form a single-handed general practice. After this, his activities accelerated.

This analysis of the victims identified by the inquiry raises questions about the way Shipman committed his murders. Some statistical evidence is provided by the time-of-day data on his supposed victims' death certificates. Below is a line graph comparing the times of day that Shipman's patients died, with the times that a sample of patients of other local family doctors died. The conclusion is sometimes known as 'interocular', since it hits you between the eyes. Shipman's patients tended overwhelmingly to die in the early afternoon.



The time at which Harold Shipman's patients died, compared to the times at which patients of other local general practitioners died. The pattern does not require sophisticated statistical analysis.

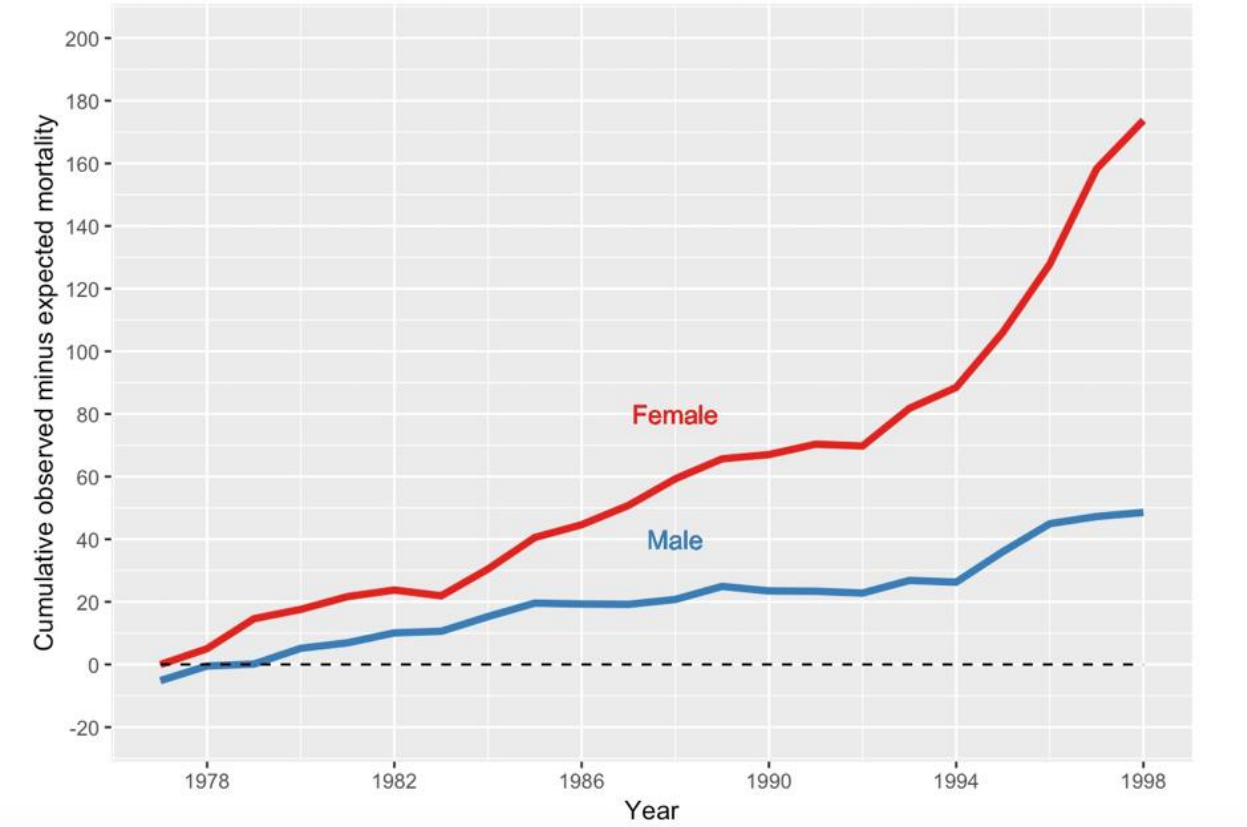
Further investigation revealed that Shipman performed his home visits after lunch, when he was generally alone with his elderly patients. He would offer them an injection that he said was to make them more comfortable, but that was in fact a lethal dose of diamorphine: the patient would die peacefully in front of him. Dame Janet Smith, who chaired the public inquiry, later said:

I still do feel it was unspeakably dreadful, just unspeakable and unthinkable and unimaginable that he should be going about day after day pretending to be this wonderfully caring doctor and having with him in his bag his lethal weapon … which he would just take out in the most matter-of-fact way.

He was taking some risk, since a single postmortem would have exposed him, but given the age of his patients and the apparent natural causes of death, none were performed. His reasons for committing these murders have never been explained: he never spoke about his misdeeds to anyone, including his family, and committed suicide in prison, conveniently just in time for his wife to collect his pension.

The main question the inquiry's statisticians were asked was – could he have been identified earlier?

In advance of the inquiry, the number of death certificates that Shipman signed for people dying in their homes or in his practice since 1977 had been accumulated. We compared it with the number that would have been expected, given the age composition of all the patients under Shipman's 'care' and the mortality rates for other local GPs. Making this sort of comparison controls for local conditions such as changing temperature and flu outbreaks. The figure below shows the expected number subtracted from the observed number of death certificates, accumulated from 1977 until Shipman's arrest in 1998. This difference can be termed his 'excess' mortality.



The cumulative number of death certificates signed by Shipman for patients who were age 65 or over and who died at home or in his practice. The expected number, given the composition of his practice list, has been subtracted.

By 1998, his estimated excess mortality for people aged 65 or over was 174 women and 49 men. This was almost exactly the number of older people that the inquiry later confirmed to be victims, showing the remarkable accuracy of this purely statistical analysis, for which no knowledge of individual cases had been included.

Supposing, in some fictitious history, that someone had been monitoring Shipman's deaths year by year and doing the calculations necessary to produce figure 5, at what point might they have 'blown the whistle'? This apparently innocuous question raises numerous challenging statistical issues related to formal tests of 'statistical hypotheses'. My book *The Art of Statistics* (2019) contains a full (but nonmathematical) discussion of the issues, which are briefly summarised in the next paragraphs: warning, this gets a bit technical.

The standard approach is to set up a *null hypothesis*, say, that Shipman had the same underlying mortality rate as his colleagues, and so is completely normal. We then calculate a statistic that would indicate evidence against that hypothesis, and work out the probability of getting such an extreme value were the null hypothesis really true: the resulting *P-value* would be compared with an established threshold, say 0.05 or 0.01, and smaller P-values would lead to a claim of Shipman having a 'statistically significant' higher mortality rate. Essentially, if the data cannot be explained by chance alone, and so seems incompatible with the null hypothesis, then we declare that something strange is going on.

Had this process been actually carried out, then in 1979 – after only three years of monitoring – there would have been a P-value of 0.004 arising from comparing 40 observed deaths while only expecting 25.3. The results could have been declared 'statistically significant', and Shipman investigated and detected.

This so-called 'null hypotheses significance-testing' approach forms the basis of most scientific claims, including major discoveries such as the Higgs boson, although it has been a matter of important debate for years. But there are two reasons why such a statistical procedure would have been grossly inappropriate as a way of monitoring the mortality rates of GPs. First, unless there was some other reason to suspect Shipman and to set up a monitoring process for him alone, we would have been calculating such P-values for all the GPs in the UK – numbering around 25,000 at the time. We know that if we carry out enough significance tests we will get false signals. With 25,000 GPs tested at a critical threshold of 0.05, we would expect one in 20 utterly innocent doctors – around 1,300 – to be 'significantly high' each time the test was carried out, and it would be wrong to investigate all these people. And Shipman might be lost in all these false-positives.

The second problem is that we are carrying out repeated significance tests, as each year's new data are added and another test performed. Fortunately, it turns out that there is some remarkable but complex theory, delightfully known as 'the law of the iterated logarithm'. This shows that if we carry out such repeated testing, even if the null hypothesis is true, then we are certain to eventually reject that null at any significance level we choose.

This is worrying. It means that if we keep testing a doctor for long enough, then we are guaranteed to eventually think we have found evidence of excess mortality, even if in reality it's not there. Fortunately, there are statistical methods for dealing with this problem of *sequential testing*. They were first developed in the Second World War by teams of statisticians working on industrial quality-control of armaments and other war materiel.

Armaments coming off the production line were being monitored by steadily accumulating total deviations from a standard, much in the same way as monitoring excess mortality. Scientists realised that the law of the iterated logarithm meant that repeated significance testing would always lead eventually to an alert that the industrial process had gone out of strict control, even if in truth everything was functioning fine. Essentially, if we keep on checking on a process, in the end something will look odd just by chance alone.

Statisticians in the US and the UK, working independently, developed what became known as the sequential probability ratio test (SPRT), which is a statistic that monitors accumulating evidence about deviations. The key is that it can at any time be compared with simple thresholds – as soon as one of these thresholds is crossed, it triggers an alert to investigate the production line. Abraham Wald in the US and George Barnard in the UK led the statisticians – Barnard was a delightful man, a pure mathematician (and communist) before the war, who later went on to develop the official British Standard 3704 for condoms. Such techniques led to more efficient industrial processes, and were later adapted for use in so-called sequential clinical trials in which accumulated results are repeatedly monitored to see if a threshold that indicates a beneficial treatment has been crossed.

A team of us developed a version of the SPRT to apply in the Shipman inquiry. We concluded that if someone had been doing this monitoring then, just looking at female deaths, Shipman would have crossed a very stringent threshold in 1984. An investigation and prosecution at that time could have saved around 175 lives, but the inquiry decided that nobody could have been blamed for not applying this simple statistical monitoring device earlier – the data were not available. Moreover, who would think that an avuncular GP would be systematically murdering his patients?

Later, a monitoring system for GPs was piloted, which immediately identified a GP with even higher mortality rates than Shipman! Investigation revealed that this doctor practised in a south-coast town with a large number of retirement homes with many old people, and he conscientiously helped many of his patients to remain out of hospital for their deaths. It would have been completely inappropriate for this caring GP to receive any publicity for his apparently high rate of signing death certificates. The important lesson is that, while statistical systems can detect outlying outcomes, they cannot offer reasons why these might have

occurred. What happened is one matter, why it happened can be very different. So statistical evaluation requires careful implementation in order to avoid false accusations. Another reason to be cautious about algorithms.

The Shipman story illustrates the two complementary components of statistical literacy. First is the ability to carry out statistical investigations leading to clear communication of what the data reveals. The second vital component is the ability to read about a claim based on data, while also having an idea of how to critique the numbers and a sense of which questions to ask. Statistics often give some answers, but they generally raise even more questions.

This kind of statistical literacy is difficult to teach. It cannot be reduced to formulae and algorithms – it is best learned through repeated experience and mentoring, almost as an apprenticeship. It takes time and effort to learn the art of statistics.