Numeracy
Advancing Education in Quantitative Literacy

Volume 13 | Issue 1 Article 7

2020

Introducing *The Art of Statistics: How to Learn from Data*

David Spiegelhalter
d.spiegelhalter@statslab.cam.ac.uk

Recommended Citation

Spiegelhalter, David. "Introducing *The Art of Statistics: How to Learn from Data*." *Numeracy* 13, Iss. 1 (2020): Article 7. DOI: https://doi.org/10.5038/1936-4660.13.1.7

Introducing *The Art of Statistics: How to Learn from Data*

Abstract

Spiegelhalter, David. 2019. *The Art of Statistics: How to Learn from Data* (New York, NY: Basic Books) 448 pp. ISBN 978-1541618510.

This short piece introduces readers to *The Art of Statistics: How to Learn from Data,* a new book by David Spiegelhalter. In this age of data, classic statistical courses can appear of limited relevance due to their focus on probability-based methodology. The book takes a modern approach to introducing the essential concepts of statistical science, avoiding mathematics and structured around real problems that data can help solve, many based on the author's own experience.

Keywords

quantitative literacy, statistical literacy

Cover Page Footnote

Professor Sir David Spiegelhalter is Chair of the Winton Centre for Risk and Evidence Communication in the Faculty of Mathematics at the University of Cambridge. His web page is www.statslab.cam.ac.uk/~david/ and his twitter feed is @d_spiegel.

We live in an age of data. Every aspect of our lives is measured and analysed, often without our knowledge. The process of turning this mass of data into accessible information, and then into knowledge about the world, is known as statistical or data science. But this process is not a matter of technical proficiency in software and mathematical formulae; it is an art requiring care, delicacy, and numerous judgments. *The Art of Statistics* tries to introduce elements of that art, with the aim of introducing readers to the essential ideas of learning from data, without mathematics, driven by real questions.

For example, there is increasing interest in positive well-being in a nation, rather than just its health problems. So each year a random sample of 150,000 people in the UK are asked "Overall, how happy did you feel yesterday?" In 2017 their average response, on a scale from zero to ten, was 7.5.

This example illustrates two big challenges in turning data into knowledge. First, such measurements are almost always an imperfect measure of what we are really interested in: asking how happy people were last week on a scale from zero to ten hardly encapsulates someone's whole emotional well-being. Second, anything we choose to measure will differ from place to place, from person to person, and from time to time. The problem is to extract meaningful insights from all this apparently random variability. In this case the average response of 7.5 was an improvement from 2012 when it was 7.3, which might be related to economic recovery since the financial crash of 2008. The lowest scores were reported for those aged between 50 and 54, and the highest between 70 and 74, a typical pattern for the UK. Having larger and larger datasets may enable us to more accurately estimate average scores for different groups, but does not free us from unavoidable inadequacies of the data.

The book is not only aimed at those who are carrying out investigations using data, but also people on the receiving end of claims made on the basis of statistical analysis, which of course includes everyone exposed to media stories about health, wealth, or almost anything else. Such 'data literacy' is an increasingly vital skill.

But traditional statistics education has not been a particularly popular topic. Generations of students have suffered through dry statistics courses based on learning a set of techniques to be applied in different situations, with more regard to mathematical theory than understanding both why the formulae are being used and the challenges that arise when trying to use data to answer real questions. Fortunately statistics education is changing. The needs of data science and data literacy demand a more problem-driven approach in which the application of specific statistical tools is seen as just one component of a complete cycle of investigation. In the book I illustrate the 'PPDAC' structure—iterating through the stages of establishing a Problem, then Planning an investigation, collecting and cleaning Data, carrying both exploratory and more formal Analyses, and finally

drawing Conclusions and communicating the results. The results will often generate new ideas for investigation, so the cycle starts again.

The narrative in the book is therefore driven by questions, many of which come from my personal experience as a statistician. Some are important scientific hypotheses, such as whether the Higgs boson exists or if there really is convincing evidence for extra-sensory perception (ESP). Others are questions about health care, such as whether busier hospitals have higher survival rates and if screening for ovarian cancer is beneficial. Sometimes we just want to estimate quantities, such as the cancer risk from bacon sandwiches, the number of sexual partners people in Britain have in their lifetime, and the benefit of taking a daily statin. Other questions are just interesting, such as identifying the luckiest survivor of the Titanic, whether Harold Shipman could have been caught earlier, and assessing the probability that a skeleton found in a Leicester car park really belonged to Richard III.

Excerpt from *The Art of Statistics*

The following excerpt captures both the style and the aims of the book.

Do Some Areas of the UK Really Have Three Times the Bowel Cancer Rates as Others?

The headline on the respected BBC news online in September 2011 was alarming: 'Three-fold variation' in UK bowel cancer death rates'. The article went on to explain that different areas in the UK had starkly different rates of bowel cancer with a commentator suggesting it was 'extremely important for local NHS organisations to examine information for their own areas and use it to inform potential changes in delivery of services'.[2]

A three-fold difference sounds extraordinarily dramatic, but when the blogger Paul Barden came across the article he wondered: do people in different parts of the country really face such large and important differences in their risk of dying from bowel cancer? What would cause such a discrepancy? He found it so implausible he decided to investigate. Admirably, the data was openly available online and he found that it did substantiate what the BBC piece had claimed: in 2008 there was more than a three-fold variation between the annual death rates of people with bowel cancer. It ranged from 9 per 100,000 people in Rossendale in Lancashire to 31 per 100,000 inhabitants of Glasgow City.

But this was not the end of his investigation. He then plotted the death rates against the population in each district, which gave the picture shown in Figure 1.

[2] Barden's original blog is here (https://pb204.blogspot.com/2011/10/funnel-plot-of-uk-bowel-cancer.html), and the data can be downloaded from here: http://pb204.blogspot.co.uk/2011/10/uploads.html.

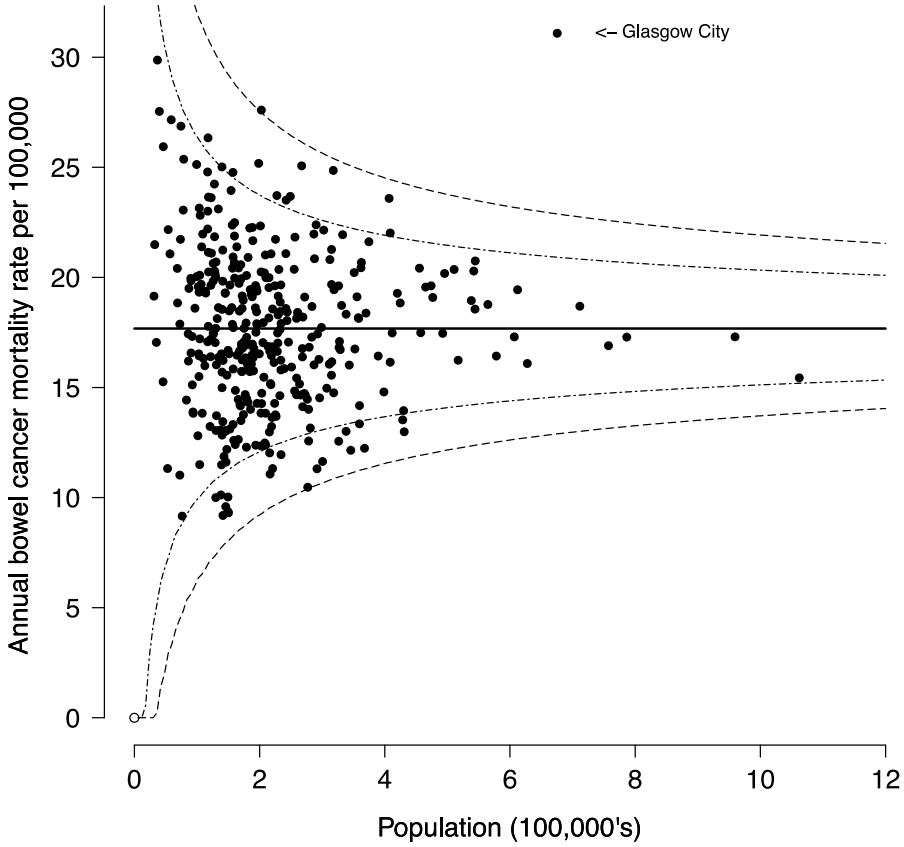

Figure 1. Annual bowel cancer death rates per 100,000 population in 380 districts in the UK, plotted against the population of the district. The two sets of dashed lines indicate the regions in which we would expect 95% and 99.8% of districts to lie, if there were no real differences between the risks and they are derived from an assumed underlying Binomial distribution. Only Glasgow City shows any evidence of an underlying risk that is different from the average. This way of looking at the data is called a 'funnel plot'.

It is clear that the points (all apart from the extreme example of Glasgow City) form a sort of funnel shape, in which the differences between districts get larger as their population gets smaller. Paul then added *control limits* which show where we would expect the points to land if the differences between the observed rates were just due to natural and unavoidable variability in the numbers that die of bowel cancer each year, rather than due to any systematic variation in the underlying risks

experienced in different districts. These control limits are obtained from assuming that the number of bowel cancer deaths in each area are an observation from a Binomial distribution with sample size equal to the adult population of the area, and underlying probability 0.000176 that any particular person would die from bowel cancer each year: this is the average individual risk over the whole country. The control limits are set to contain 95% and 99.8% of the probability distribution respectively. This type of graph is called a *funnel plot* and is extensively used when examining multiple health authorities or institutions, as it permits the identification of outliers without creating spurious league tables.

The data fall within the control limits rather well, which means that differences between districts are essentially what we would expect by chance variability alone. Smaller districts have fewer cases and so are more vulnerable to the role of chance, and therefore tend to have more extreme results—the rate in Rossendale was based on only 7 deaths, and so its rate could be drastically altered by just a few extra cases. So despite the BBC's dramatic headline, there is no big news story here—we would expect a three-fold variability in the observed rates even if the underlying risk in the different districts were precisely the same.

There is a crucial lesson in this simple example. Even in an era of open data, data science and data journalism, we still need basic statistical principles in order not to be misled by apparent patterns in the numbers.

This chart reveals that the only observation of any particular note is Glasgow City's outlying data-point. Is bowel cancer a particularly Scottish phenomenon? Is this data-point actually correct? More recent data for the period 2009–2011 reveals that bowel-cancer mortality for Greater Glasgow was 20.5 per 100,000 people, in Scotland overall it was 19.6, and in England it was 16.4: these findings both cast doubt on the specific Glasgow City value, but show that Scotland has higher rates than England. Typically, conclusions from one problem-solving cycle raise more questions, and so the cycle starts over again.

Reprinted with permission.

DOI: https://doi.org/10.5038/1936-4660.13.1.7