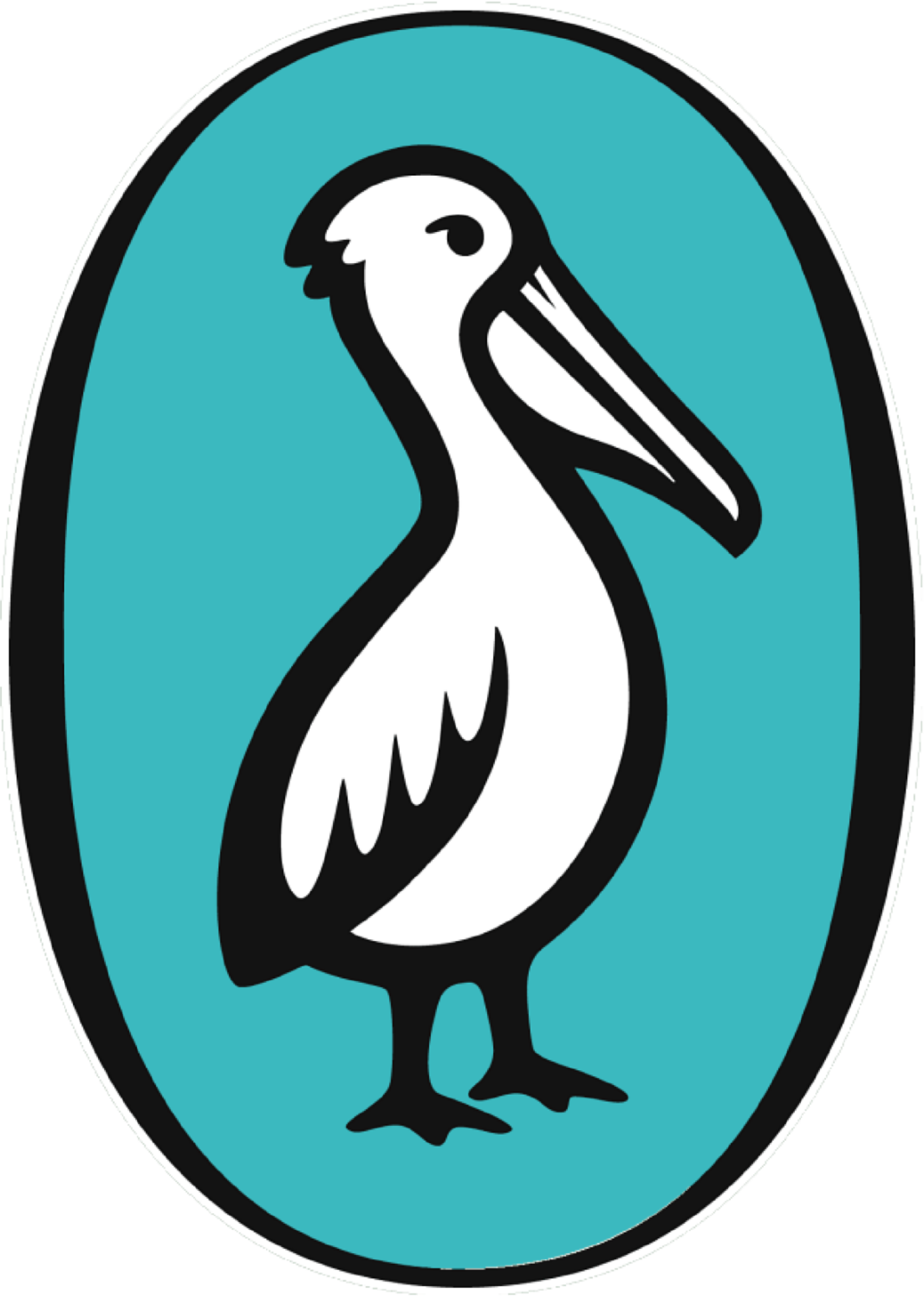


Contents

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

INTRODUCTION

CHAPTER 1
Getting Things in Proportion: Categorical Data and Percentages

CHAPTER 2
Summarizing and Communicating Numbers. Lots of Numbers

CHAPTER 3
Why Are We Looking at Data Anyway? Populations and Measurement

CHAPTER 4
What Causes What?

CHAPTER 5
Modelling Relationships Using Regression

CHAPTER 6
Algorithms, Analytics and Prediction

CHAPTER 7
How Sure Can We Be About What Is Going On? Estimates and Intervals

CHAPTER 8
Probability – the Language of Uncertainty and Variability

CHAPTER 9
Putting Probability and Statistics Together

CHAPTER 10
Answering Questions and Claiming Discoveries

CHAPTER 11
Learning from Experience the Bayesian Way

CHAPTER 12
How Things Go Wrong

CHAPTER 13
How We Can Do Statistics Better

CHAPTER 14
In Conclusion

GLOSSARY

NOTES

INDEX

About the Author

Sir David John Spiegelhalter is a British statistician and Chair of the Winton Centre for Risk and Evidence Communication in the Statistical Laboratory at the University of Cambridge. Spiegelhalter is one of the most cited and influential researchers in his field, and was elected as President of the Royal Statistical Society for 2017–18.

‘David Spiegelhalter is probably the greatest living statistical communicator; more than that, he’s one of the great communicators in any field. This marvellous book will transform your relationship with the numbers that swirl all around us. Read it and learn. I did’
Tim Harford, presenter of ‘More or Less’ and author of *The Undercover Economist*

‘Written by a master of the subject … this book tells us to examine our assumptions. Bravo’
Standpoint

‘Remarkable. Spiegelhalter is warm and encouraging – it’s a genuinely enjoyable read … This book should be required reading for all politicians, journalists, medics and anyone who tries to influence people (or is influenced) by statistics. A tour de force’
Popular Science

‘This is an excellent book. Spiegelhalter is great at explaining difficult ideas … Yes, statistics can be difficult. But much less difficult if you read this book’
William Leith, *Evening Standard*

‘A call to arms for greater societal data literacy … needed now more than ever’
Alan Smith, *Financial Times*

‘*The Art of Statistics* is in the great educational tradition of its publishing imprint, Pelican Books: an attempt to get everyone up to speed with the practical uses of statistics, without pages of terrifying equations or Greek letters. In a series of spry, airy chapters, he succeeds fabulously’
Stuart Ritchie, *Spectator*

‘A must-read book’
Simon Briscoe, *Literary Review*

‘If I had to trust just one person to interrogate statistical data, I’d trust David Spiegelhalter. He is a master of the art. Here, he shows us how it’s done. The result is brilliant’
Michael Blastland, co-author of *The Tiger That Isn’t: Seeing Through a World of Numbers*

To statisticians everywhere, with their endearing traits
of pedantry, generosity, integrity, and desire to use data
in the best way possible

List of Figures

0.1 Age and Year of Death of Harold Shipman's Victims

0.2 Time of Death of Harold Shipman's Patients

0.3 The PPDAC Problem-Solving Cycle

1.1 30-Day Survival Rates Following Heart Surgery

1.2 Proportion of Child Heart Operations Per Hospital

1.3 Percentage of Child Heart Operations Per Hospital

1.4 Risk of Eating Bacon Sandwiches

2.1 Jar of Jelly Beans

2.2 Different Ways of Displaying Jelly Bean Guesses

2.3 Jelly-Bean Guesses Plotted on a Logarithmic Scale

2.4 Reported Number of Lifetime Opposite-Sex Partners

2.5 Survival Rates Against Number of Operations in Child Heart Surgery

2.6 Pearson Correlation Coefficients of 0

2.7 World Population Trends

2.8 Relative Increase in Population by Country

2.9 Popularity of the Name 'David' Over Time

2.10 Infographic on Sexual Attitudes and Lifestyles

3.1 Diagram of Inductive Inference

3.2 Distribution of Birth Weights

5.1 Scatter of Sons' Heights v. Fathers' Heights

5.2 Fitted Logistic Regression Model for Child Heart Surgery Data

6.1 Memorial to *Titanic* Victim

6.2 Summary Survival Statistics for *Titanic* Passengers

6.3 Classification Tree for *Titanic* Data

6.4 ROC Curves for Algorithms Applied to Training and Test Sets

6.5 Probabilities of Surviving the *Titanic* Sinking

6.6 Over-Fitted Classification Tree for *Titanic* Data

6.7 Post-Surgery Survival Rates for Women with Breast Cancer

7.1 Empirical Distribution of Number of Sexual Partners for Varying Sample Sizes

7.2 Bootstrap Resamples from Original Sample of 50

7.3 Bootstrap Distribution of Means at Varying Sample Sizes

7.4 Bootstrap Regressions on Galton's Mother–Daughter Data

8.1 A Simulation of the Chevalier de Méré's Games

8.2 Expected Frequency Tree for Two Coin Flips

8.3 Probability Tree for Flipping Two Coins

8.4 Expected Frequency Tree for Breast Cancer Screening

8.5 Observed and Expected Number of Homicides

9.1 Probability Distribution of Left-Handers

9.2 Funnel Plot of Bowel-Cancer Death Rates

9.3 BBC Plot of Opinion Polls Before the 2017 General Election

9.4 Homicide Rates in England and Wales

10.1 Sex Ratio for London Baptisms, 1629–1710

10.2 Empirical Distribution of Observed Difference in Proportions of Left/Right Arm Crossers

10.3 Cumulative Number of Death Certificates Signed by Shipman

10.4 Sequential Probability Ratio Test for Detection of a Doubling in Mortality Risk

10.5 Expected Frequencies of the Outcomes of 1,000 Hypothesis Tests

11.1 Expected Frequency Tree for Three-Coin Problem

11.2 Expected Frequency Tree for Sports Doping

11.3 Reversed Expected Frequency Tree for Sports Doping

11.4 Bayes' 'Billiard' Table

12.1 Traditional Information Flows for Statistical Evidence

List of Tables

1.1 Outcomes of Children’s Heart Surgery

1.2 Methods for Communicating the Lifetime Risk of Bowel Cancer in Bacon Eaters

2.1 Summary Statistics for Jelly-Bean Guesses

2.2 Summary Statistics for the Lifetime Number of Sexual Partners

4.1 Outcomes for Patients in the Heart Protection Study

4.2 Illustration of Simpson’s Paradox

5.1 Summary Statistics of Heights of Parents and Their Adult Children

5.2 Correlations between Heights of Adult Children and Parent of Same Gender

5.3 Results of a Multiple Linear Regression Relating Adult Offspring Height to Mother and Father

6.1 Error Matrix of Classification Tree on *Titanic* Training and Test Data

6.2 Fictional ‘Probability of Precipitation’ Forecasts

6.3 Results of a Logistic Regression for *Titanic* Survivor Data

6.4 Performance of Different Algorithms on *Titanic* Test Data

6.5 Breast Cancer Survival Rates Using the Predict 2.1 Algorithm

7.1 Summary Statistics for Lifetime Sexual Partners Reported by Men

7.2 Sample Means of Lifetime Sexual Partners Reported by Men

9.1 Comparison of Exact and Bootstrap Confidence Intervals

10.1 Cross-Tabulation of Arm-Crossing Behaviour by Gender

10.2 Observed and Expected Counts of Arm-Crossing by Gender

10.3 Observed and Expected Days with Each Number of Homicide Incidents

10.4 Results of Heart Protection Study with Confidence Intervals and P-values

10.5 The Output in *R* of a Multiple Regression Using Galton’s Data

10.6 Possible Outcomes of a Hypothesis Test

11.1 Likelihood Ratios for Evidence Concerning Richard III Skeleton

11.2 Recommended Verbal Interpretations of Likelihood Ratios

11.3 Kass and Raftery’s Scale for Interpretation of Bayes Factors

12.1 Questionable Interpretation and Communication Practices

13.1 Exit Poll Predictions for Three Recent General Elections

Acknowledgements

Any insights gained from a long career in statistics come from listening to inspiring colleagues. These are too numerous even for a statistician to count, but a shortlist of those I have stolen most from might include Nicky Best, Sheila Bird, David Cox, Philip Dawid, Stephen Evans, Andrew Gelman, Tim Harford, Kevin McConway, Wayne Oldford, Sylvia Richardson, Hetan Shah, Adrian Smith and Chris Wild. I am grateful to them and so many others for encouraging me in a challenging subject.

This book has been a long time in development, entirely due to my chronic procrastination. So I would primarily like to thank Laura Stickney of Penguin for not only commissioning the book, but remaining calm as the months, and years, went by, even when the book was finished and we still could not agree on a title. And all credit to Jonathan Pegg for negotiating me a fine deal, Jane Birdsell for showing huge patience when editing, and all the production staff at Penguin for their meticulous work.

I am very grateful for permission to adapt illustrations, specifically Chris Wild (Figure 0.3), James Grime (Figure 2.1), Cath Mercer of Natsal (Figures 2.4 and 2.10), Office for National Statistics (Figures 2.9, 8.5 and 9.4), Public Health England (Figure 6.7), Paul Barden (Figure 9.2), and BBC (Figure 9.3). UK public sector information is licensed under the Open Government Licence v3.0.

I am not a good R programmer, and Matthew Pearce and Maria Skoularidou helped me enormously in doing the analyses and graphics. I also struggle with writing, and so am indebted to numerous people who read and commented on chapters, including George Farmer, Alex Freeman, Cameron Brick, Michael Posner, Sander van der Linden and Simone Warr: in particular Julian Gilbey had a fine eye for errors and ambiguity.

Above all, I must thank Kate Bull not only for her vital comments on the text, but also for supporting me through times that have been both good (writing in a beach hut in Goa) and not so good (a wet February juggling too many commitments).

I am also profoundly grateful to David and Claudia Harding for both their financial support and their continued encouragement, which has enabled me to do such fun things over the last ten years.

Finally, much as I would like to find someone else to blame, I am afraid I must acknowledge full responsibility for the inevitable remaining inadequacies of this book.

CODE FOR EXAMPLES

R code and data for reproducing most of the analyses and Figures are available from https://github.com/dspiegel29/ArtofStatistics. I am grateful for the assistance received in preparing this material.

Introduction

> The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning.
>
> — Nate Silver, *The Signal and the Noise*[1]

Why We Need Statistics

Harold Shipman was Britain's most prolific convicted murderer, though he does not fit the archetypal profile of a serial killer. A mild-mannered family doctor working 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 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 realize 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, fifteen were exhumed and lethal levels of diamorphine, the medical form of heroin, were found in their bodies. Shipman was subsequently tried for fifteen 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 the public inquiry, which concluded that he had definitely murdered 215 of his patients, and possibly 45 more.[2]

This book will focus on using **statistical science**[fn1] to answer the kind of questions that arise when we want to better understand the world – some of these questions will be highlighted in a box. In order to get some insight into Shipman's behaviour, a natural first question is:

> What kind of people did Harold Shipman murder, and when did they die?

The public inquiry provided details of each victim's age, gender and date of death. Figure 0.1 is a fairly sophisticated visualization of this data, showing a scatter-plot of the age of victim against their 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 5–year bands) and years.

Some conclusions can be drawn by simply taking some time to look at the figure. There are more black than white dots, and so Shipman's victims were mainly women. The bar-chart on the right of the picture 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 before that time Shipman had been working 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, as demonstrated by the top bar-chart.

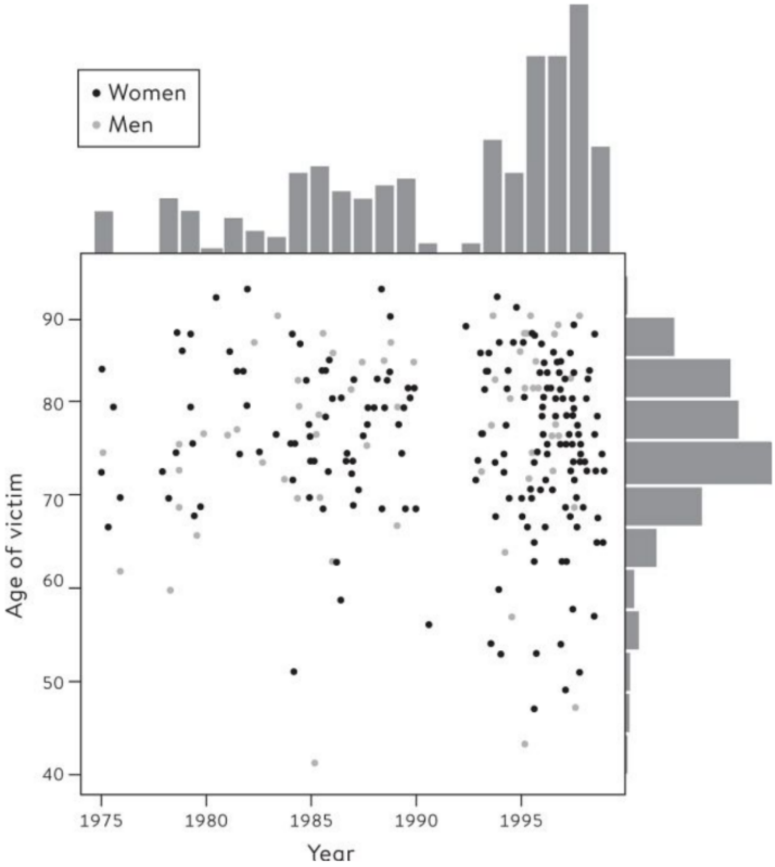


Figure 0.1
A scatter-plot showing the age and the year of death of Harold Shipman's 215 confirmed victims. Bar-charts have been added on the axes to reveal the pattern of ages and the pattern of years in which he committed murders.

This analysis of the victims identified by the inquiry raises further questions about the way he committed his murders. Some statistical evidence is provided by data on the time of day of the death of his supposed victims, as recorded on the death certificate. Figure 0.2 is a line graph comparing the times of day that Shipman's patients died to the times that a sample of patients of other local family doctors died. The pattern does not require subtle analysis: the conclusion is sometimes known as 'inter-ocular', since it hits you between the eyes. Shipman's patients tended overwhelmingly to die in the early afternoon.

The data cannot tell us *why* they tended to die at that time, but further investigation revealed that he 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 which was in fact a lethal dose of diamorphine: after a patient had died peacefully in front of him, he would change their medical record to make it appear as if this was an expected natural death. 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.'

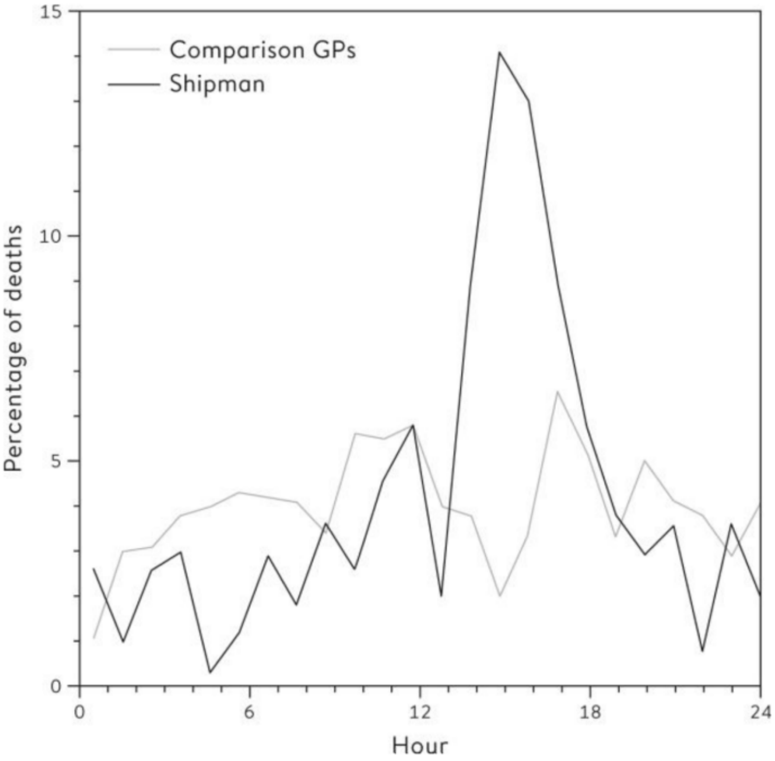


Figure 0.2
The time at which Harold Shipman's patients died, compared to the times at which patients of other local general practitioners died. The pattern does not require sophisticated statistical analysis.

He was taking some risk, since a single post-mortem would have exposed him, but given the age of his patients and the apparent natural causes of death, none were performed. And his reasons for committing these murders have never been explained: he gave no evidence at his trial, 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.

We can think of this type of iterative, exploratory work as 'forensic' statistics, and in this case it was literally true. There is no mathematics, no theory, just a search for patterns that might lead to more interesting questions. The details of Shipman's misdeeds were determined using evidence specific to each individual case, but this data analysis supported a general understanding of how he went about his crimes.

Later in this book, in Chapter 10, we will see whether formal statistical analysis could have helped catch Shipman earlier.[fn2] In the meantime, the Shipman story amply demonstrates the great potential of using data to help us understand the world and make better judgements. This is what statistical science is all about.

Turning the World Into Data

A statistical approach to Harold Shipman's crimes required us to stand back from the long list of individual tragedies for which he was responsible. All those personal, unique details of people's lives, and deaths, had to be reduced to a set of facts and numbers that could be counted and drawn on graphs. This might at first seem cold and dehumanizing, but if we are to use statistical science to illuminate the world, then our daily experiences have to be turned into data, and this means categorizing and labelling events, recording measurements, analysing the results and communicating the conclusions.

Simply categorizing and labelling can, however, present a serious challenge. Take the following basic question, which should be of interest to everyone concerned with our environment:

How many trees are there on the planet?

Before even starting to think about how we might go about answering this question, we first have to settle a rather basic issue. What is a 'tree'? You may feel you know a tree when you see it, but your judgement may differ considerably from others who might consider it a bush or a shrub. So to turn experience into data, we have to start with rigorous definitions.

It turns out that the official definition of a 'tree' is a plant with a woody stem that has a sufficiently large diameter at breast height, known as the DBH. The US Forest Service demands a plant has a DBH of greater than 5 inches (12.7 cm) before officially declaring it a tree, but most authorities use a DBH of 10 cm (4 inches).

But we cannot wander round the entire planet individually measuring each woody-stemmed plant and counting up those that meet this criterion. So the researchers who investigated this question took a more pragmatic approach: they first took a series of areas with a common type of landscape, known as a biome, and counted the average number of trees found per square kilometre. They then used satellite imaging to estimate the total area of the planet covered by each type of biome, carried out some complex statistical modelling, and eventually came up with an estimated total of 3.04 trillion (that is 3,040,000,000,000) trees on the planet. This sounds a lot, except they reckoned there used to be twice this number.[fn3] [3]

If authorities differ about what they call a tree, it should be no surprise that more nebulous concepts are even more challenging to pin down. To take an extreme example, the official definition of 'unemployment' in the UK was changed at least thirty-one times between 1979 and 1996.[4] The definition of Gross Domestic Product (GDP) is continually being revised, as when trade in illegal drugs and prostitution was added to the UK GDP in 2014; the estimates used some unusual data sources – for example Punternet, a review website that rates prostitution services, provided prices for different activities.[5]

Even our most personal feelings can be codified and subjected to statistical analysis. In the year ending September 2017, 150,000 people in the UK were asked as part of a survey: 'Overall, how happy did you feel yesterday?'[6] Their average response, on a scale from zero to ten, was 7.5, 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.[fn4]

Measuring happiness is hard, whereas deciding whether someone is alive or dead should be more straightforward: as the examples in this book will demonstrate, survival and mortality is a common concern of statistical science. But in the US each state can have its own legal definition of death, and although the Uniform Declaration of Death Act was introduced in 1981 to try to establish a common model, some small differences remain. Someone who had been declared dead in Alabama could, at least in principle, cease to be legally dead were they across the state border in Florida, where the registration must be made by two qualified doctors.[7]

These examples show that statistics are always to some extent constructed on the basis of judgements, and it would be an obvious delusion to think the full complexity of personal experience can be unambiguously coded and put into a spreadsheet or other software. Challenging though it is to define, count and measure characteristics of ourselves and the world around us, it is still just information, and only the starting point to real understanding of the world.

Data has two main limitations as a source of such knowledge. First, it is 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 the emotional wellbeing of the nation. Second, anything we choose to measure will differ from place to place, from person to person, from time to time, and the problem is to extract meaningful insights from all this apparently random **variability**.

For centuries, statistical science has faced up to these twin challenges, and played a leading role in scientific attempts to understand the world. It has provided the basis for interpreting data, which is always imperfect, in order to distinguish important relationships from the background variability that makes us all unique. But the world is always changing, as new questions are asked and new sources of data become available, and statistical science has had to change too.

People have always counted and measured, but modern statistics as a discipline really began in the 1650s when, as we shall see in Chapter 8, probability was properly understood for the first time by Blaise Pascal and Pierre de Fermat. Given this solid mathematical basis for dealing with variability, progress was then remarkably rapid. When combined with data on the ages at which people die, the theory of probability provided a firm basis for calculating pensions and annuities. Astronomy was revolutionized when scientists grasped how probability theory could handle variability in measurements. Victorian enthusiasts became obsessed with collecting data about the human body (and everything else), and established a strong connection between statistical analysis and genetics, biology and medicine. Then in the twentieth century statistics became more mathematical and, unfortunately for many students and practitioners, the topic became synonymous with the mechanical application of a bag of statistical tools, many named after eccentric and argumentative statisticians that we shall meet later in this book.

This common view of statistics as a basic 'bag of tools' is now facing major challenges. First, we are in an age of **data science**, in which large and complex data sets are collected from routine sources such as traffic monitors, social media posts and internet purchases, and used as a basis for technological innovations such as optimizing travel routes, targeted advertising or purchase recommendation systems – we shall look at **algorithms** based on '**big data**' in Chapter 6. Statistical training is increasingly seen as just one necessary component of being a data scientist, together with skills in data management, programming and algorithm development, as well as proper knowledge of the subject matter.

Another challenge to the traditional view of statistics comes from the huge rise in the amount of scientific research being carried out, particularly in the biomedical and social sciences, combined with pressure to publish in high-ranking journals. This has led to doubts about the reliability of parts of the scientific literature, with claims that many 'discoveries' cannot be reproduced by other researchers – such as the continuing dispute over whether adopting an assertive posture popularly known as a 'power pose' can induce hormonal and other changes.[8] The inappropriate use of standard statistical methods has received a fair share of the blame for what has become known as the reproducibility or replication crisis in science.

With the growing availability of massive data sets and user-friendly analysis software, it might be thought that there is less need for training in statistical methods. This would be naïve in the extreme. Far from freeing us from the need for statistical skills, bigger data and the rise in the number and complexity of scientific studies makes it even more difficult to draw appropriate conclusions. More data means that we need to be even more aware of what the evidence is actually worth.

For example, intensive analysis of data sets derived from routine data can increase the possibility of false discoveries, both due to systematic bias inherent in the data sources and from carrying out many analyses and only reporting whatever looks most interesting, a practice sometimes known as ‘data-dredging’. In order to be able to critique published scientific work, and even more the media reports which we all encounter on a daily basis, we should have an acute awareness of the dangers of selective reporting, the need for scientific claims to be replicated by independent researchers, and the danger of over-interpreting a single study out of context.

All these insights can be brought together under the term **data literacy**, which describes the ability to not only carry out statistical analysis on real-world problems, but also to understand and critique any conclusions drawn by others on the basis of statistics. But improving data literacy means changing the way statistics is taught.

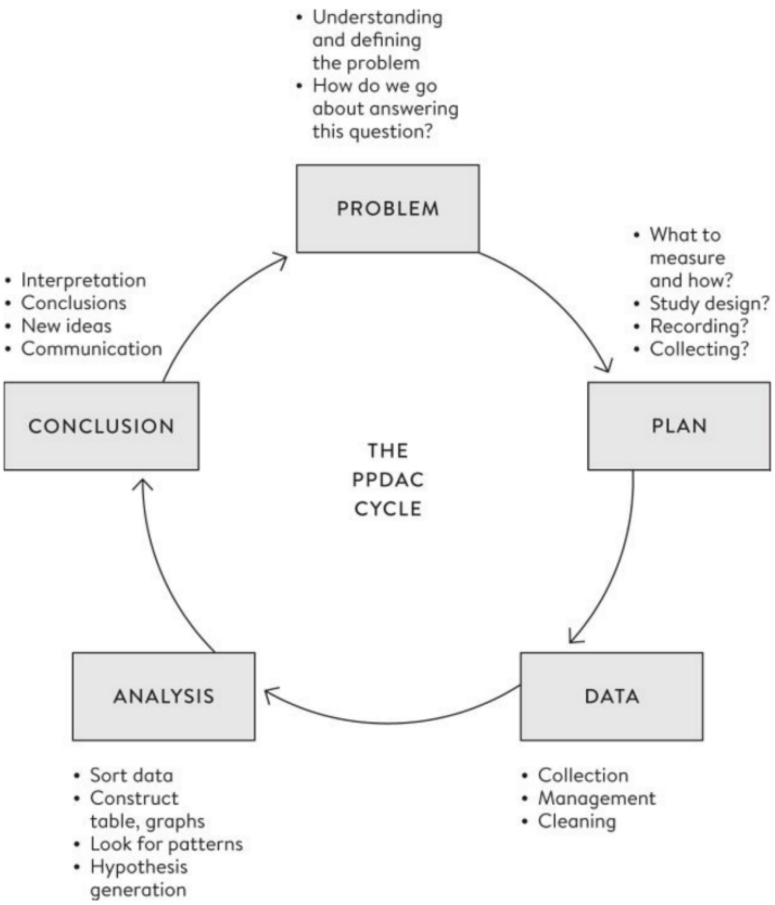
Teaching Statistics

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 questions.

Fortunately this 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. The **PPDAC** structure has been suggested as a way of representing a problem-solving cycle, which we shall adopt throughout this book.[9] Figure 0.3 is based on an example from New Zealand, which has been a world-leader in statistics education in schools.

The first stage of the cycle is specifying a Problem; statistical inquiry always starts with a question, such as our asking about the pattern of Harold Shipman’s murders or the number of trees in the world. Later in this book we shall focus on problems ranging from the expected benefit of different therapies immediately following breast cancer surgery, to why old men have big ears.

It is tempting to skip over the need for a careful Plan. The Shipman question simply required the collection of as much data as possible on his victims. But the people counting trees paid meticulous attention to precise definitions and how to carry out the measurements, since confident conclusions can only be drawn from a study which has been appropriately designed. Unfortunately, in the rush to get data and start analysis, attention to design is often glossed over.



• Understanding and defining the problem
• How do we go about answering this question?
PROBLEM
• What to measure and how?
• Study design?
• Recording?
• Collecting?
PLAN
• Interpretation
• Conclusions
• New ideas
• Communication
CONCLUSION
THE PPDAC CYCLE
ANALYSIS
DATA
• Sort data
• Construct table, graphs
• Look for patterns
• Hypothesis generation
• Collection
• Management
• Cleaning

Figure 0.3
The PPDAC problem-solving cycle, going from Problem, Plan, Data, Analysis to Conclusion and communication, and starting again on another cycle.

Collecting good Data requires the kind of organizational and coding skills that are being seen as increasingly important in data science, particularly as data from routine sources may need a lot of cleaning in order to get it ready to be analysed. Data collection systems may have changed over time, there may be obvious errors, and so on – the phrase 'found data' neatly communicates that it may be rather messy, like something picked up in the street.

The Analysis stage has traditionally been the main emphasis of statistics courses, and we shall cover a range of analytic techniques in this book; but sometimes all that is required is a useful visualization, as in Figure 0.1. Finally, the key to good statistical science is drawing appropriate Conclusions that fully acknowledge the limitations in the evidence, and communicating them clearly, as in the graphical illustrations of the Shipman data. Any conclusions generally raise more questions, and so the cycle starts over again, as when we started looking at the time of day when Shipman's patients died.

Although in practice the PPDAC cycle laid out in Figure 0.3 may not be followed precisely, it underscores that formal techniques for statistical analysis play only one part in the work of a statistician or data scientist. Statistical science is a lot more than a branch of mathematics involving esoteric formulae with which generations of students have (often reluctantly) struggled.

This Book

When I was a student in Britain in the 1970s, there were just three TV channels, computers were the size of a double wardrobe, and the closest thing we had to Wikipedia was on the imaginary handheld device in Douglas Adams' (remarkably prescient) *Hitchhiker's Guide to the Galaxy*. For self-improvement we therefore turned to Pelican books, and their iconic blue spines were a standard feature of every student bookshelf.

Because I was studying statistics, my Pelican collection featured *Facts from Figures* by M. J. Moroney (1951) and *How to Lie with Statistics* by Darrell Huff (1954). These venerable publications sold in the hundreds of thousands, reflecting both the level of interest in statistics and the dismal lack of choice at that time. These classics have stood up remarkably well to the intervening sixty-five years, but the current era demands a different approach to teaching statistics based on the principles laid out above.

This book therefore uses real-world problem-solving as a starting point for introducing statistical ideas. Some of these ideas may seem obvious, but some are more subtle and may require some mental effort, although mathematical skills will not be needed. Compared to traditional texts, this book focuses on conceptual issues rather than technicalities, and features only a few, fairly innocuous equations supported by a Glossary. Software is a vital part of any work in data science and statistics but it is not a focus of this book – tutorials are readily available for freely available environments such as R and Python.

The questions featured in the boxes can all, to a certain extent, be answered through statistical analysis, although they differ widely in their scope. 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.

And some questions are just interesting, such as identifying the luckiest survivor from the *Titanic*; whether Harold Shipman could have been caught earlier; and assessing the probability that a skeleton found in a Leicester car park really was that of Richard III.

This book is intended for both students of statistics who are seeking a non-technical introduction to the basic issues, and general readers who want to be more informed about the statistics they encounter both in their work and in everyday life. My emphasis is on handling statistics skilfully and with care: numbers may appear to be cold, hard facts, but the attempts to measure trees, happiness and death have already shown that they need to be treated with delicacy.

Statistics can bring clarity and insight into the problems we face, but we are all familiar with the way they can be abused, often to promote an opinion or simply to attract attention. The ability to assess the trustworthiness of statistical claims seems a key skill in the modern world, and I hope that this book may help to empower people to question the numbers that they encounter in their daily life.

Summary

- Turning experiences into data is not straightforward, and data is inevitably limited in its capacity to describe the world.
- Statistical science has a long and successful history, but is now changing in the light of increased availability of data.
- Skill in statistical methods plays an important part of being a data scientist.
- Teaching statistics is changing from a focus on mathematical methods to one based on an entire problem-solving cycle.
- The PPDAC cycle provides a convenient framework: Problem – Plan – Data – Analysis – Conclusion and communication.
- Data literacy is a key skill for the modern world.

CHAPTER 1

Getting Things in Proportion: Categorical Data and Percentages

What happened to children having heart surgery in Bristol between 1984 and 1995?

Joshua L was 16 months old and had transposition of the great arteries, a severe form of congenital heart disease in which the main vessels coming from the heart are attached to the wrong ventricle. He needed an operation to 'switch' the arteries, and just after 7 a.m. on 12 January 1995 his parents said goodbye to him and watched as he was taken for his surgery in Bristol Royal Infirmary. But Joshua's parents were unaware that stories about the poor surgical survival rates at Bristol had been circulating since the early 1990s. Nobody told them that nurses had left the unit rather than continue telling parents that their child had died, or that the previous evening there had been a late-night meeting at which it had been debated whether to cancel Joshua's operation.[1]

Joshua died on the operating table. The following year the General Medical Council (the medical regulator) launched an investigation after complaints from Joshua's and other bereaved parents, and in 1998 two surgeons and the ex-chief executive were found guilty of serious medical misconduct. Public concern did not die down, and an official inquiry was ordered: this brought in a team of statisticians who were given the grim task of comparing the survival rates in Bristol with elsewhere in the UK between 1984 and 1995. I led this team.

We first had to determine how many children had had heart surgery, and how many had died. This sounds like it should be straightforward but, as shown in the previous chapter, simply counting events can be challenging. What is a 'child'? What counts as 'heart surgery'? When can death be attributed to surgery? And even when these definitions have been decided, could we determine how many of each there had been?

We took a 'child' as anyone under 16, and focused on 'open' surgery in which the heart had been stopped and its function replaced by cardio-pulmonary bypass. There can be multiple operations per admission, but these were considered as one event. Deaths were counted if they occurred within 30 days of the operation, whether or not in hospital or due to the surgery. We knew that death was an imperfect measure of the quality of the outcome, as it ignored children who were brain-damaged or otherwise disabled as a result of the surgery, but we did not have the data on longer-term outcomes.

The main source of data was national Hospital Episode Statistics (HES), which were derived from administrative data entered by low-paid coders. HES had a poor reputation among doctors, but this source had the great advantage that it could be linked to national death records. There was also a parallel system of data submitted directly to a Cardiac Surgical Registry (CSR) established by the surgeons' professional society.

These two sources of data, though they were supposed to be about exactly the same practice, showed considerable disagreement: for 1991–1995, HES said there had been 62 deaths out of 505 open operations (14%), whereas CSR said there had been 71 deaths out of 563 operations (13%). No less than five additional local sources of data were available, from anaesthetic records to the surgeons' own personal logs. Bristol was awash with data, but none of the data sources could be considered the 'truth', and nobody had taken responsibility for analysing and acting on the surgical outcomes.

We calculated that if patients at Bristol had the average risk prevailing elsewhere in the UK, Bristol would have expected to have had 32 deaths over this period, instead of the 62 recorded in HES, which we reported as '30 excess deaths' between 1991 and 1995.[fn1] The exact numbers varied according to the data sources, and it may seem extraordinary that we could not even establish the basic facts about the number of operations and their outcome, although current record systems should be better.

These findings had wide press coverage, and the Bristol inquiry led to a major change in attitudes to monitoring clinical performance: no longer was the medical profession trusted to police itself. Mechanisms to publicly report hospital survival data were established, although, as we shall now see, the way in which that data is displayed can itself influence the perception of audiences.

Communicating Counts and Proportions

Data that records whether individual events have happened or not is known as **binary data**, as it can only take on two values, generally labelled as yes and no. Sets of binary data can be summarized by the number of times and the percentage of cases in which an event occurred.

The theme of this chapter is that the basic presentation of statistics is important. In a sense we are jumping to the last step of the PPDAC cycle in which conclusions are communicated, and while the form of this communication has not traditionally been considered an important topic in statistics, rising interest in data visualization reflects a change in this attitude. So both in this chapter and the next we shall concentrate on ways of displaying data so that we can quickly get the gist of what is going on without detailed analysis, starting with a look at alternative ways of displaying data that, largely because of the Bristol inquiry, are now publicly available.

Table 1.1 shows the outcomes of nearly 13,000 children who had heart surgery in the UK and Ireland between 2012 and 2015.[2] Two hundred and sixty-three babies died within 30 days of their operation, and every one of these deaths is a tragedy to the family involved. It will be little consolation to them that survival rates have improved hugely from the time of the Bristol inquiry, and now average 98%, and so there is a more hopeful prospect for families of children facing heart surgery.

A table can be considered as a type of graphic, and requires careful design choices of colour, font and language to ensure engagement and readability. The audience's emotional response to the table may also be influenced by the choice of which columns to display. Table 1.1 shows the results in terms of both survivors and deaths, but in the US *mortality* rates from child heart surgery are reported, while the UK provides *survival* rates. This is known as negative or positive **framing**, and its overall effect on how we feel is intuitive and well-documented: '5% mortality' sounds worse than '95% survival'. Reporting the actual number of deaths as well as the percentage can also increase the impression of risk, as this total might then be imagined as a crowd of real people.

Hospital	Number of babies having surgery	Number surviving for at least 30 days after surgery	Number dying within 30 days of surgery	Percentage surviving	Percentage dying
London, Harley Street	418	413	5	98.8	1.2
Leicester	607	593	14	97.7	2.3
Newcastle	668	653	15	97.8	2.2
Glasgow	760	733	27	96.3	3.7
Southampton	829	815	14	98.3	1.7
Bristol	835	821	14	98.3	1.7
Dublin	983	960	23	97.7	2.3
Leeds	1,038	1,016	22	97.9	2.1
London, Brompton	1,094	1,075	19	98.3	1.7
Liverpool	1,132	1,112	20	98.2	1.8
London, Evelina	1,220	1,185	35	97.1	2.9
Birmingham	1,457	1,421	36	97.5	2.5
London, Great Ormond Street	1,892	1,873	19	99.0	1.0
Total	12,933	12,670	263	98.0	2.0

Table 1.1

Outcomes of children's heart surgery in UK and Irish hospitals between 2012 and 2015, in terms of survival or not, 30 days after surgery.

A classic example of how alternative framing can change the emotional impact of a number is an advertisement that appeared on the London Underground in 2011, proclaiming that '99% of young Londoners do not commit serious youth violence'. These ads were presumably intended to reassure passengers about their city, but we could reverse its emotional impact with two simple changes. First, the statement means that 1% of young Londoners *do* commit serious violence. Second, since the population of London is around 9 million, there are around 1 million people aged between 15 and 25, and if we consider these as 'young', this means there are 1% of 1 million or a total of 10,000 seriously violent young people in the city. This does not sound at all reassuring. Note the two tricks used to manipulate the impact of this statistic: convert from a positive to a negative frame, and then turn a percentage into actual numbers of people.

Ideally both positive and negative frames should be presented if we want to provide impartial information, although the order of columns might still influence how the table is interpreted. The order of the rows of a table also needs to be considered carefully. Table 1.1 shows the hospitals in order of the number of operations in each, but if they had been presented, say, in order of mortality rates with the highest at the top of the table, this might give the impression that this was a valid and important way of comparing hospitals. Such league tables are favoured by the media and even some politicians, but can be grossly misleading: not only because the differences could be due to chance variation, but because the hospitals may be taking in very different types of cases. In Table 1.1, for example, we might suspect that Birmingham, one of the biggest and most well-known children's hospitals, takes on the most severe cases, and so it would be unfair, to put it mildly, to highlight their apparently unimpressive overall survival rates.[fn2]

The survival rates can be presented in a horizontal bar-chart such as the one shown in Figure 1.1. A crucial choice is where to start the horizontal axis: if the values start from 0%, all the bars will be almost the full length of the graphic, which will clearly show the extraordinarily high survival rates, but the lines will be indistinguishable. But the oldest trick of misleading graphics is to start the axis at say 95%, which will make the hospitals look extremely different, even if the variation is in fact only what is attributable to chance alone.

Choosing the start of the axis therefore presents a dilemma. Alberto Cairo, author of influential books on data visualization,[3] suggests you should always begin with a 'logical and meaningful baseline', which in this situation appears difficult to identify – my rather arbitrary choice of 86% roughly represents the unacceptably low survival in Bristol twenty years previously.

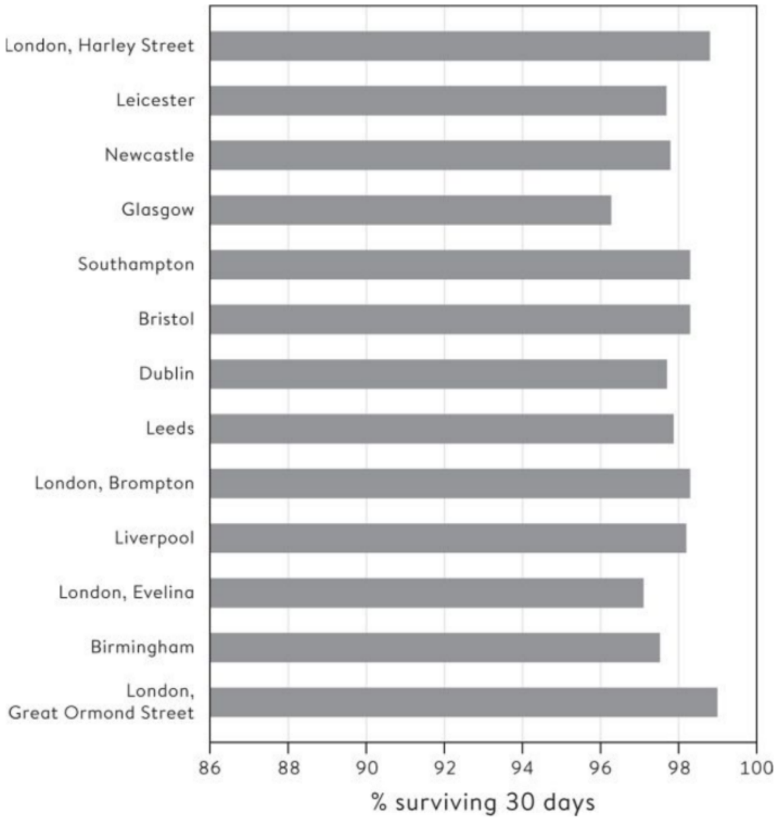


Figure 1.1

Horizontal bar-chart of 30–day survival rates for thirteen hospitals. The choice of the start of the horizontal axis, here 86%, can have a crucial effect on the impression given by the graphic. If the axis starts at 0%, all the hospitals will look indistinguishable, whereas if we started at 95% the differences would look misleadingly dramatic. Rather than a bar-chart, it may be better to use dots for the data-points when the axis does not start at zero.

I began this book with a quotation from Nate Silver, the founder of data-based platform *FiveThirtyEight* and first famous for accurately predicting the 2008 US presidential election, who eloquently expressed the idea that numbers do not speak for themselves – we are responsible for giving them meaning. This implies that communication is a key part of the problem-solving cycle, and I have shown in this section how the message from a set of simple proportions can be influenced by our choices of presentation.

We now need to introduce an important and convenient concept that will help us get beyond simple yes/no questions.

Categorical Variables

A variable is defined as any measurement that can take on different values in different circumstances; it's a very useful shorthand term for all the types of observations that comprise data. Binary variables are yes/no questions such as whether someone is alive or dead and whether they are female or not: both of these vary between people, and can, even for gender, vary within people at different times. **Categorical variables** are measures that can take on two or more categories, which may be

- Unordered categories: such as a person's country of origin, the colour of a car, or the hospital in which an operation takes place.
- Ordered categories: such as the rank of military personnel.

- Numbers that have been grouped: such as levels of obesity, which is often defined in terms of thresholds for the body mass index (BMI).[fn3]

When it comes to presenting categorical data, pie charts allow an impression of the size of each category relative to the whole pie, but are often visually confusing, especially if they attempt to show too many categories in the same chart, or use a three-dimensional representation that distorts areas. Figure 1.2 shows a fairly hideous example modelled on the kind offered by Microsoft Excel, showing the proportions of the 12,933 child heart patients from Table 1.1 that are treated in each hospital.

Multiple pie charts are generally not a good idea, as comparisons are hampered by the difficulty in assessing the relative sizes of areas of different shapes. Comparisons are better based on height or length alone in a bar chart. Figure 1.3 shows a simpler, clearer example of a horizontal bar chart of the proportions being treated in each hospital.

Comparing a Pair of Proportions

We have seen how a set of proportions can be elegantly compared using a bar chart, and so it would be reasonable to think that comparing two proportions would be a trivial matter. But when these proportions represent estimates of the risks of experiencing some harm, then the way in which those risks are compared becomes a serious and contested issue. Here is a typical question:

What’s the cancer risk from bacon sandwiches?

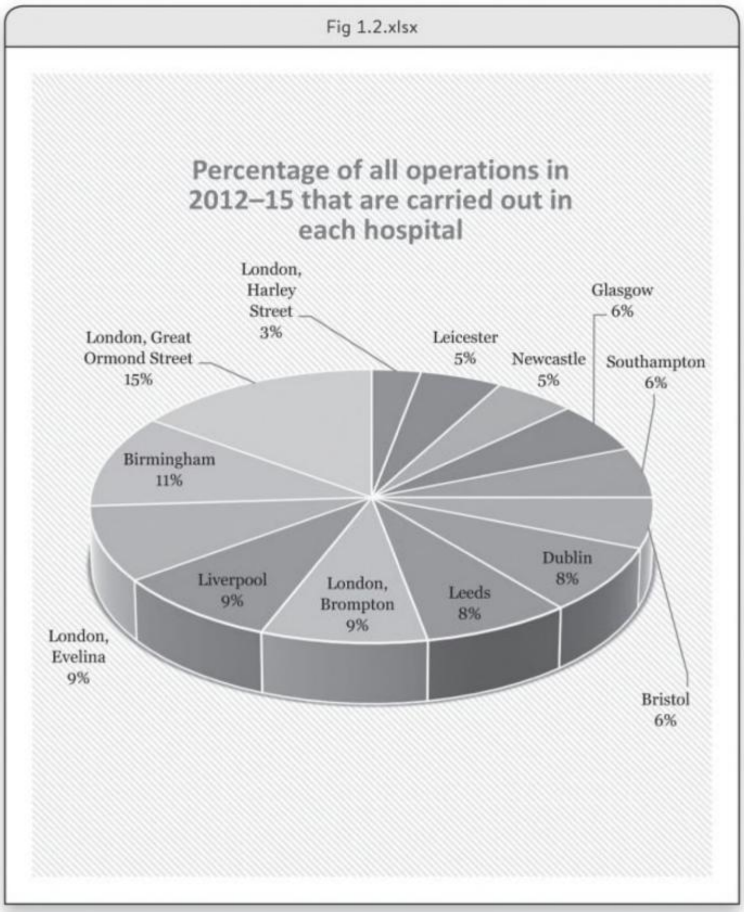


Figure 1.2
The proportion of all child heart operations being carried out in each hospital, displayed in a 3D pie chart from Excel. This deeply unpleasant chart makes categories near the front look bigger, and so makes it impossible to make visual comparisons between hospitals.

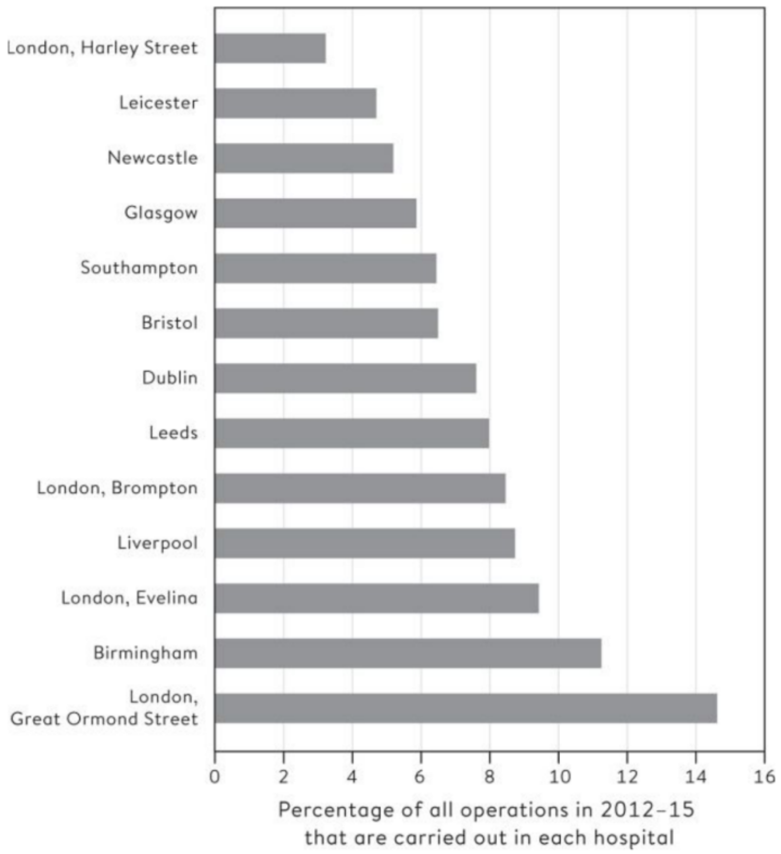


Figure 1.3
Percentage of all child heart operations being carried out in each hospital: a clearer representation using a horizontal bar chart.

We're all familiar with hyperbolic media headlines that warn us that something mundane increases the risk of some dread occurrence: I like to call these 'cats cause cancer' stories. For example, in November 2015 the World Health Organization's International Agency for Research in Cancer (IARC) announced that processed meat was a 'Group I carcinogen', putting it in the same category as cigarettes and asbestos. This inevitably led to panicky headlines such as the *Daily Record*'s claim that 'Bacon, Ham and Sausages Have the Same Cancer Risk as Cigarettes Warn Experts'.[4]

The IARC tried to quell the fuss by emphasizing that the Group 1 classification was about being confident that an increased risk of cancer existed at all, and said nothing about the actual magnitude of the risk. Lower down in the press release, the IARC reported that 50g of processed meat a day was associated with an increased risk of bowel cancer of 18%. This sounds worrying, but should it be?

The figure of 18% is known as a **relative risk** since it represents the increase in risk of getting bowel cancer between a group of people who eat 50g of processed meat a day, which could, for example, represent a daily two-rasher bacon sandwich, and a group who don't. Statistical commentators took this relative risk and reframed it into a change in **absolute risk**, which means the change in the actual proportion in each group who would be expected to suffer the adverse event.

They concluded that, in the normal run of things, around 6 in every 100 people who do not eat bacon daily would be expected to get bowel cancer in their lifetime. If 100 similar people ate a bacon sandwich every single day of their lives, then according to the IARC report we would expect that 18% more would get bowel cancer, which means a rise from 6 to 7 cases out of 100.[fn4]

That is one extra case of bowel cancer in all those 100 lifetime bacon-eaters, which does not sound so impressive as the relative risk (an 18% increase), and might serve to placed on the journalists: the abstract of the paper mentioned only the odds ratio without mentioning that this corresponded to a difference between absolute risks of 85% vs 87%.[7]

This highlights the danger of using odds ratios in anything but a scientific context, and the advantage of always reporting absolute risks as the quantity that is relevant for an audience, whether they are concerned with bacon, statins or anything else.

The examples in this chapter have demonstrated how the apparently simple task of calculating and communicating proportions can become a complex matter. It needs to be carried out with care and awareness, and the impact of numerical or graphical data summaries can be explored by working with psychologists who are skilled in evaluating the perception of alternative formats. Communication is an important part of the problem-solving cycle, and should not be just a matter of personal preference.

Summary

- Binary variables are yes/no questions, sets of which can be summarized as proportions.
- Positive or negative framing of proportions can change their emotional impact.
- Relative risks tend to convey an exaggerated importance, and absolute risks should be provided for clarity.
- Expected frequencies promote understanding and an appropriate sense of importance.
- Odds ratios arise from scientific studies but should not be used for general communication.
- Graphics need to be chosen with care and awareness of their impact.

CHAPTER 2

Summarizing and Communicating Numbers. Lots of Numbers

Can we trust the wisdom of crowds?

In 1907 Francis Galton, cousin of Charles Darwin and polymath originator of identification using fingerprints, weather forecasts and eugenics,[fn1] wrote a letter to the prestigious science journal *Nature* about his visit to the Fat Stock and Poultry Exhibition in the port city of Plymouth. There he saw a large ox displayed and contestants paying sixpence to guess the 'dressed' weight of the resulting meat after the poor beast had been slaughtered. He got hold of 787 of the tickets that had been filled out and chose the middle value of 1,207 lb (547 kg) as the democratic choice, 'every other estimate being condemned as too high or too low by the majority of voters'. The dressed weight turned out to be 1,198 lb (543 kg), which was remarkably close to his choice based on the 787 votes.[1] Galton titled his letter '*Vox Populi*' (voice of the people), but this process of decision-making is now better known as the **wisdom of crowds**.

Galton carried out what we might now call a data summary: he took a mass of numbers written on tickets and reduced them to a single estimated weight of 1,207 lb. In this chapter we look at the techniques that have been developed in the subsequent century for summarizing and communicating the piles of data that have become available. We will see that numerical summaries of location, spread, trend and correlation are intimately related to how the data can be plotted on paper or a screen. And we shall look at the gentle transition between simply describing the data, and seeking to tell a story through an infographic.

First we will begin with my own attempt at a wisdom-of-crowds experiment, which demonstrates many of the problems that crop up when the real, undisciplined world, with all its capacity for oddity and error, is used as a source of data.

Statistics is not only concerned with serious events such as cancer and surgery. In a rather trivial experiment, mathematics communicator James Grime and I posted a video on YouTube and asked anyone watching to guess the number of jelly beans in a jar. You might want to try this exercise yourself when you see the photo in Figure 2.1 (the true number will be revealed later). Nine hundred and fifteen people provided their guesses, which ranged from 219 to 31,337, and in this chapter we shall look at how such variables can be depicted graphically and summarized numerically.

To start, Figure 2.2 shows three ways of presenting the pattern of the values the 915 respondents provided: these patterns can be variously termed the data distribution, **sample distribution** or empirical distribution.[fn2]

(a) The strip-chart, or dot-diagram, simply shows each data-point as a dot, but each one is given a random jitter to prevent multiple guesses of the same number lying on top of each other and obscuring the overall pattern. This clearly shows a large number of guesses in the range up to around 3,000, and then a long 'tail' of values right up to over 30,000, with a cluster at exactly 10,000.

(b) The box-and-whisker plot summarizes some essential features of the data distribution.[fn3]

(c) This histogram simply counts how many data-points lie in each of a set of intervals – it gives a very rough idea of the shape of the distribution.

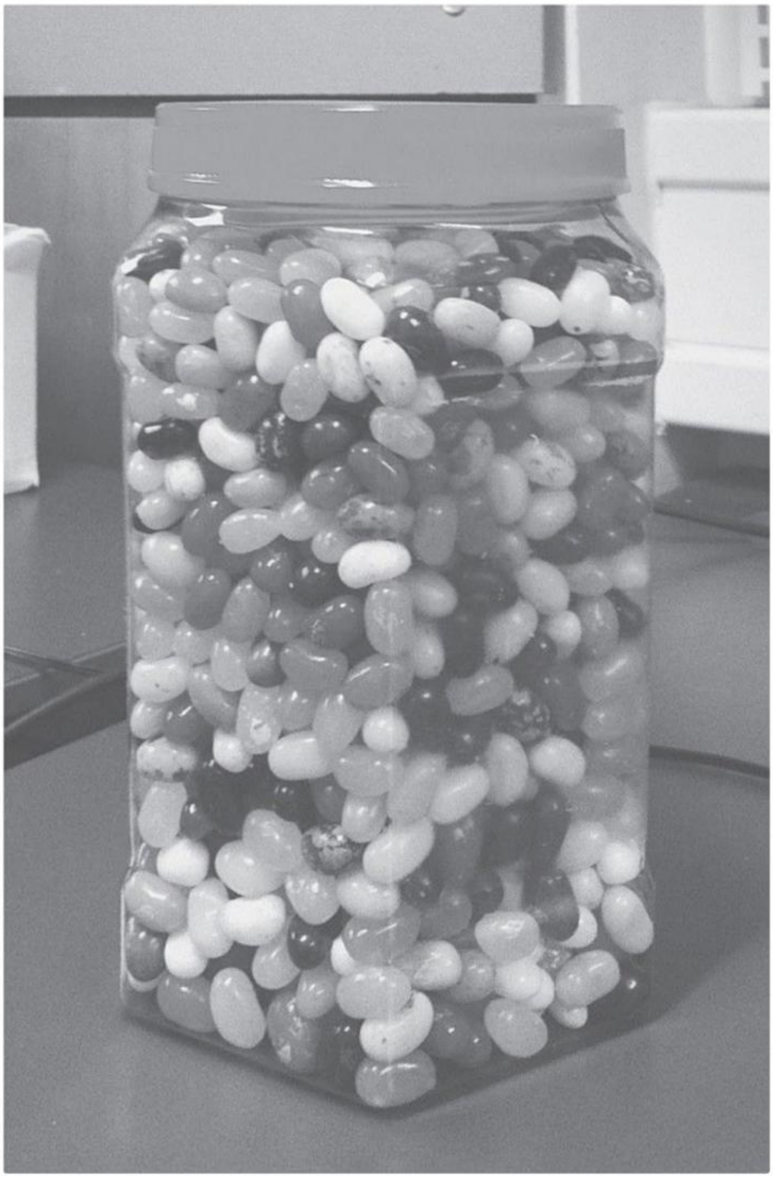


Figure 2.1
How many jelly beans are in this jar? We asked this on a YouTube video and got 915 responses. The answer will be given later.

Mean-averages can be highly misleading when the raw data do not form a symmetric pattern around a central value but instead are skewed towards one side like the jelly-bean guesses, typically with a large group of standard cases but with a tail of a few either very high (for example, income) or low (for example, legs) values. I can almost guarantee that, compared with people of your age and sex, you have far less than the average (mean) risk of dying next year. For example, the UK life tables report that 1% of 63-year-old men die each year before their 64th birthday, but many of those who will die are already seriously ill, and so the vast majority who are reasonably healthy will have less than this average risk.

Unfortunately, when an 'average' is reported in the media, it is often unclear whether this should be interpreted as the mean or median. For example, the UK Office for National Statistics calculates Average Weekly Earnings, which is a *mean*, while also reporting *median* weekly earnings by local authority. In this case it might help to distinguish between 'average income' (mean) and 'the income of the average person' (median). House prices have a very skewed distribution, with a long right-hand tail of high-end properties, which is why official house price indices are reported as medians. But these are generally reported as the 'average house price', which is a highly ambiguous term. Is this the average-house price (that is, the median)? Or the average house-price (that is, the mean)? A hyphen can make a big difference.

It is now time to reveal the results of our own wisdom of crowds experiment with the jelly beans: not as exciting as the weight of an ox, but with slightly more votes than Galton had.

Due to the data distribution having a long right-hand tail, the mean-average of 2,408 would be a poor summary, and the mode of 10,000 seems to reflect some extreme choice of round numbers. And so it is presumably better to follow Galton and use the median as a group-guess. This turns out to be 1,775 beans. The true value was ... 1,616.[2] Just one person guessed this precisely, 45% of people guessed below 1,616, and 55% guessed above, so there was little systematic tendency for the guesses to be either on the high or low side – we say the true value lay at the 45th **percentile** of the empirical data distribution. The median, which is the 50th percentile, over-estimated the true value by 1,775 – 1,616 = 159, so relative to the true answer the median was an over-estimate by around 10%, and only around 1 in 10 people got that close. So the wisdom of crowds was fairly good, getting closer to the truth than 90% of the individual people.

Describing the Spread of a Data Distribution

It is not enough to give a single summary for a distribution – we need to have an idea of the spread, sometimes known as the variability. For example, knowing the average adult male shoe size will not help a shoe firm decide the quantities of each size to make. One size does not fit all, a fact which is vividly illustrated by the seats for passengers in planes.

Table 2.1 shows a variety of summary statistics for the jelly-bean guesses, including three ways to summarize the spread. The **range** is a natural choice, but is clearly very sensitive to extreme values such as the apparently bizarre guess of 31,337 beans.[fn7] In contrast the **inter-quartile range** (IQR) is unaffected by extremes. This is the distance between the 25th and 75th percentiles of the data and so contains the 'central half' of the numbers, in this case between 1,109 and 2,599 beans: the central 'box' of the box-and-whisker plots shown above covers the inter-quartile range. Finally the **standard deviation** is a widely used measure of spread. It is the most technically complex measure, but is only really appropriate for well-behaved symmetric data[fn8] since it is also unduly influenced by outlying values. For example, removing the single value of 31,337 from the data reduces the standard deviation from 2,422 to 1,398.[fn9]

Summary statistics for judgements of the number of jelly beans in a jar	Full data
Mean	2,408
Median	1,775
Mode	10,000
Range	219 to 31,337
Inter-quartile range	1,109 to 2,599
Standard deviation	2,422

Table 2.1
Summary statistics for 915 jelly-bean judgements. The true number was 1,616.

The crowd in our little experiment showed itself to have considerable wisdom, in spite of some bizarre responses. This demonstrates that data often has some errors, outliers and other strange values, but these do not necessarily need to be individually identified and excluded. It also points to the benefits of using summary measures that are not unduly affected by odd observations such as 31,337 – these are known as robust measures, and include the median and the inter-quartile range. Finally, it shows the great value of simply looking at the data, a lesson that will be reinforced by the next example.

Describing Differences between Groups of Numbers

How many sexual partners do people in Britain report having had in their lifetime?

The purpose of this question is not simply to be nosey about people's private lives. When AIDS first became a serious concern in the 1980s, public health officials realized that there was no reliable evidence about sexual behaviour in Britain, particularly in terms of the frequency with which people changed partners, how many had multiple simultaneous partners, and what sexual practices people engaged in. This knowledge was essential to predict the spread of sexually transmitted diseases through society and to plan health services, and yet people were still quoting from the unreliable data collected by Alfred Kinsey in the US in the 1940s – who made no attempt at obtaining a representative sample.

So beginning in the late 1980s, large, careful and costly surveys of sexual behaviour were established in the UK and US, in spite of strong opposition from some quarters. In the UK, Margaret Thatcher withdrew support from a major survey of sexual lifestyles at the last minute, but those conducting the study were fortunately able to find charitable funding instead, resulting in the National Sexual Attitudes and Lifestyle Survey (Natsal) which has been carried out in the UK every ten years since 1990.

The third survey, known as Natsal-3, was carried out around 2010 and cost £7 million.[3] Table 2.2 shows the summary statistics concerning the number of (opposite-sex) sexual partners reported by people aged 35–44 in Natsal-3. It is a good exercise to use these summaries alone to try to reconstruct what the pattern of data might look like. We note that the most common single value (mode) is 1, representing those people who have only had one partner in their life, and yet there is also a massive range. This is also reflected by the substantial difference between the means and the medians, which is a telling sign of data distributions with long right-hand tails. The standard deviations are large, but this is an inappropriate measure of spread for such a data distribution, since it will be unduly influenced by a few extremely high values.

The responses of men and women may be compared by noting that men reported a mean-average of 6 more sexual partners than women, or alternatively that the average man (the median) reported 3 more sexual partners than the average woman. Or that, in relative terms, men report around 60% more partners than do women for both the mean and the median.

Reported number of sexual partners in lifetime	Men aged 35–44	Women aged 35–44
Mean	14.3	8.5
Median	8	5
Mode	1	1
Range	0 to 500	0 to 550
Inter-quartile range	4 to 18	3 to 10
Standard deviation	24.2	19.7

Table 2.2

Summary statistics for the number of (opposite-sex) sexual partners over their lifetime, as reported by 806 men and 1,215 women aged 35–44, based on interviews carried out in Natsal-3 between 2010 and 2012. Standard deviations are included for completeness, although they are inappropriate summaries of the spread of such data.

This difference might arouse our suspicions about the data. In a closed **population** with the same number of men and women with a similar age profile, it is a mathematical fact that the mean number of opposite-sex partners should be essentially the same for men and for women![fn10] So why are men reporting so many more partners than women in this age group of 35–44? This could partly be because of men having younger partners, but also because there appears to be systematic differences in the way men and women count and report their sexual histories. We might suspect that men may be more likely to overplay their number of partners, or women underplay them, or both.

Figure 2.4 reveals the actual data distribution, which supports the impression given by the summary statistics of an extreme right-hand tail. But it is only by looking at this raw data that further important details are revealed, such as the strong tendency for both men and women to provide rounded numbers when there have been ten or more partners (except for the rather pedantic man, possibly a statistician, who said precisely, ‘Forty-seven’). You may, of course, wonder about the reliability of these self-reports, and potential biases in these data are discussed in the next chapter.

Large collections of numerical data are routinely summarized and communicated using a few statistics of location and spread, and the sexual-partner example has shown that these can take us a long way in grasping an overall pattern. However, there is no substitute for simply looking at data properly, and the next example shows that a good visualization is particularly valuable when we want to grasp the pattern in a large and complex set of numbers.