A PANORAMA OF STATISTICS

Perspectives, puzzles and paradoxes in statistics

Eric Sowey & Peter Petocz

with website

WILEY

A Panorama of Statistics

A Panorama of Statistics

Perspectives, puzzles and paradoxes in statistics

Eric Sowey

The University of NSW, Sydney, Australia

Peter Petocz

Macquarie University, Sydney, Australia

WILEY

This edition first published 2017

Registered Office
John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Editorial Offices
9600 Garsington Road, Oxford, OX4 2DQ, United Kingdom
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Cataloging-in-Publication data applied for

ISBN: 9781119075820

A catalogue record for this book is available from the British Library.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Cover image: Jeffrey Smart, Container train in landscape (1983–84), Arts Centre Melbourne, Australia. Reproduced by permission of the Estate of Jeffrey Smart.

Set in 10/12.5pt Warnock by SPi Global, Pondicherry, India

10 9 8 7 6 5 4 3 2 1

ES *To Anne, Helen, Michelle and David*

PP *To Anna*

Contents

Preface

Most people would cheerfully agree that statistics is a useful subject – but how many would recognise that it also has many facets which are engaging, and even fascinating?

We have written this book for students and their teachers, as well as for practitioners – indeed, for anyone who knows some statistics. If this describes you, we invite you to come with us on a panoramic tour of the subject. Our intent is to highlight a variety of engaging and quirky facets of statistics, and to let you discover their fascinations. Even if you are only casually acquainted with statistical ideas, there is still much in this book for you.

This is not a textbook. In a lively way, it expands understanding of topics that are outside the scope of most textbooks – topics you are unlikely to find brought together all in the one place elsewhere.

Each of the first 25 chapters is devoted to a different statistical theme. These chapters have a common structure. First, there is an Overview, offering perspectives on the theme – often from several points of view. About half of these Overviews need, as quantitative background, only high school mathematics with a basic foundation in statistics. For the rest, it may be helpful to have completed an introductory college or university course in statistics.

Following the Overview, each chapter poses five questions to pique your curiosity and stimulate you to make your own discoveries. These questions all relate to the theme of the chapter. As you seek answers to these questions, we expect you will be surprised by the variety of ways in which statistics can capture and hold your interest.

The questions are not for technical, numerical or web-search drill. Rather, they seek to widen your knowledge and deepen your insight. There are questions about statistical ideas and probabilistic thinking, about the value of statistical techniques, about both innocent and cunning misuses of statistics, about pathbreaking inspirations of statistical pioneers, and about making the best practical use of statistical information. Also, there are amusing statistical puzzles to unravel and tantalising statistical paradoxes to resolve. Some questions have a single correct answer, but many more invite your reflection and your exploration of alternatives.

We invite you to plunge in and tackle the questions that you find appealing. Compare your findings with our own answers (including wide-ranging commentary), which are collected together in Chapter 26.

To help you to choose questions that best match your current statistical background, we have labelled each question A, B or C. Questions labelled A (40% of the total) are well suited to those who are studying an introductory course in statistics at tertiary level. Good senior high school students should also find many of them within their capability. Questions labelled B (55% of the total) cover a wide spectrum of challenges and, in many cases, a knowledge of statistics at the level of a second course will make the best starting point. The remaining 5% of questions are labelled C, and are for graduates in statistics, including professional practitioners.

In each chapter, the Overview and its five questions are extensively cross-referenced to related material in other chapters. They also include suggestions for further reading, both in print and online. The web links are available live on this book's companion website www.wiley.com/go/sowey/apanoramaofstatistics. Readers of the electronic edition should check this website for an update if they find a broken link among the chapter bibliographies.

To make it clear when a mentioned Chapter, Question or Figure refers to a place elsewhere in this book (rather than to an external source), the words are printed in small capitals: Chapter, Question, Figure.

We hope that your time spent with this book will be enjoyable and enriching.

May 2016

Eric Sowey
Peter Petocz

Acknowledgments

Our longest-standing debt is to Martin Gardner (1914–2010), whose monthly 'Mathematical Games' columns appeared in *Scientific American* from 1956 to 1981. We were still in high school when we first discovered these short, lively and often quirky essays, which have drawn tens of thousands of people (us included) in to the pleasures and fascinations of mathematics.

Among the approximately 300 essays (which were collected into 15 books between 1959 and 1997), there are fewer than a dozen on probability, and just a handful in which a statistical concept is mentioned. Each of us had, in the past, wondered why no-one had yet brought Gardner's vibrant approach to statistics. When we discovered our common interest, it became an irresistible challenge to do it ourselves. Our thirty-six columns of 'Statistical Diversions' – conceived very much in the spirit of Martin Gardner – appeared between 2003 and 2015 in each issue of *Teaching Statistics*, published by Wiley, both online (at http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1467-9639) and in hard copy.

This book represents a substantial revision, reorganisation and extension of the material in those columns. We celebrate the memory of Martin Gardner by citing four of his marvellous essays in our answers to Questions 10.5, 11.4 and 12.2 in this book.

We warmly thank the Editors of *Teaching Statistics* from 2003 to 2015 – Gerald Goodall, Roger Johnson, Paul Hewson and Helen MacGillivray – for their editorial encouragement and advice as they welcomed each of our Statistical Diversions columns.

John Aldrich, Geoff Cumming, Gerald Goodall and Jan Kmenta each read several chapters of this book. Tim Keighley and Agnes Petocz read the book through in its entirety. All these colleagues willingly responded to our invitation, bringing to the task their deep knowledge of statistics and its application in other disciplines and their cultural vantage points in Australia, the UK and the USA. We are grateful to them for the multiple ways in which their comments have improved this book.

The cover reproduces a striking panoramic painting titled *Container train in landscape* (1983–84, oil on five hardboard panels, 113.5 × 985 cm). This work, almost ten metres long, hangs in the Arts Centre Melbourne and is the gift of Mark and Eva Besen. It is by the Australian artist Jeffrey Smart (1921–2013), who lived for many years in Tuscany and whose paintings have captured worldwide attention. Smart's works are admired for, among other features, their vivid and arresting colours and their geometrically precise composition. We are delighted that the Estate of Jeffrey Smart has given permission for the use of the cover image.

David Urbinder's original drawings cheerily embellish our pages. We reproduce them by permission. Michelle Sowey's creativity guided us on the cover design.

Neville Davies, Ruma Falk, Rob Gould, Bart Holland, Larry Lesser, Milo Schield, Nel Verhoeven and Graham Wood made valued contributions while we were creating our Statistical Diversions columns, or in the course of preparing this book for publication.

We thank the following people for permission to reproduce material in the Figures: Michael Rip (Figure 1.2), Gavan Tredoux (Figure 18.1) and David Thomas (Figure 21.1).

To our Editors at Wiley, Debbie Jupe, Heather Kay, Shivana Raj, Alison Oliver, Uma Chelladurai and Brian Asbury we express our appreciation for their advice, guidance and support throughout the publication process.

Part I

Introduction

1

Why is statistics such a fascinating subject?

In the real world, little is certain. Almost everything that happens is influenced, to a greater or lesser degree, by chance. As we shall explain in this chapter, statistics is our best guide for understanding the behaviour of chance events that are, in some way, measurable. No other field of knowledge is as vital for the purpose. This is quite a remarkable truth and, statisticians will agree, one source of the subject's fascination.

You may know the saying: data are not information and information is not knowledge. This is a useful reminder! Even more useful is the insight that it is statistical methods that play the major role in turning data into information and information into knowledge.

In a world of heavily promoted commercial and political claims, a familiarity with statistical thinking can bring enormous personal and social benefits. It can help everyone to judge better what claims are trustworthy, and so become more competent and wiser as citizens, as consumers and as voters. In short, it can make ours not only a more numerate, but also a more accurately informed, society. This is an ideal we shall return to in Chapter 3.

Chance events are studied in the physical, biological and social sciences, in architecture and engineering, in medicine and law, in finance and marketing, and in history and politics. In all these fields and more, statistics has well-established credentials. To use John Tukey's charming expression, 'being a statistician [means] you get to play in everyone's backyard.' (There is more about this brilliant US statistician in Chapter 22, Figure 22.2.)

---oOo---

To gain a bird's eye view of the kinds of practical conclusions this subject can deliver, put yourself now in a situation that is typical for an applied statistician.

A Panorama of Statistics: Perspectives, Puzzles and Paradoxes in Statistics, First Edition.
Eric Sowey and Peter Petocz.

Companion website: www.wiley.com/go/sowey/apanoramaofstatistics

Suppose you have collected some data over a continuous period of 150 weekdays on the daily number of employees absent from work in a large insurance company. These 150 numbers will, at first, seem to be just a jumble of figures. However, you – the statistician – are always looking for patterns in data, because patterns suggest the presence of some sort of systematic behaviour that may turn out to be interesting. So you ask yourself: can I find any evidence of persisting patterns in this mass of figures? You might pause to reflect on what sorts of meaningful patterns might be present, and how you could arrange the data to reveal each of them. It is clear that, even at this early stage of data analysis, there is lots of scope for creative thinking.

Exercising creativity is the antithesis of following formalised procedures. Unfortunately, there are still textbooks that present statistical analysis as no more than a set of formalised procedures. In practice, it is quite the contrary. Experience teaches the perceptive statistician that a sharpened curiosity, together with some preliminary 'prodding' of the data, can often lead to surprising and important discoveries. Tukey vigorously advocated this approach. He called it 'exploratory data analysis'. Chatfield (2002) excellently conveys its flavour.

In this exploratory spirit, let's say you decide to find out whether there is any pattern of absenteeism across the week. Suppose you notice at once that there seem generally to be more absentees on Mondays and Fridays than on the other days of the week. To confirm this impression, you average the absentee numbers for each of the days of the week over the 30 weeks of data. And, indeed, the averages are higher for Mondays and Fridays.

Then, to sharpen the picture further, you put the Monday and Friday averages into one group (Group A), and the Tuesday, Wednesday and Thursday averages into a second group (Group B), then combine the values in each group by averaging them. You find the Group A average is 104 (representing 9.5% of staff) and the Group B average is 85 (representing 7.8% of staff).

This summarisation of 30 weeks of company experience has demonstrated that staff absenteeism is, on average, 1.7 percentage points higher on Mondays and Fridays as compared with Tuesdays, Wednesdays and Thursdays. Quantifying this difference is a first step towards better understanding employee absenteeism in that company *over the longer term* – whether your primary interest is possible employee discontent, or the financial costs of absenteeism to management.

Creating different kinds of data summaries is termed *statistical description.* Numerical and graphical methods for summarising data are valuable, because they make data analysis more manageable and because they can reveal otherwise unnoticed patterns.

Even more valuable are the methods of statistics that enable statisticians to generalise to a wider setting whatever interesting behaviour they may have detected in the original data. The process of generalisation in the face of the uncertainties of the real world is called *statistical inference*. What makes a statistical generalisation so valuable is that it comes with an objective measure of the likelihood that it is correct.

Clearly, a generalisation will be useful in practice only if it has a high chance of being correct. However, it is equally clear that we can never be sure that a generalisation is correct, because uncertainty is so pervasive in the real world.

To return to the example we are pursuing, you may be concerned that the pattern of absenteeism detected in 30 weeks of data might continue indefinitely, to the detriment of the company. At the same time, you may be unsure that that pattern actually is a long-term phenomenon. After all, it may have appeared in the collected data only by chance. You might, therefore, have good reason to widen your focus, from absenteeism in a particular 30-week period to absenteeism in the long term.

You can test the hypothesis that the pattern you have detected in your data occurred by chance alone against the alternative hypothesis that it did *not* occur by chance alone. The alternative hypothesis suggests that the pattern is actually persistent – that is, that it is built into the long-term behaviour of the company if there are no internal changes (by management) or external impacts (from business conditions generally). As just mentioned, the statistical technique for performing such a hypothesis test can also supply a measure of the likelihood that the test result is correct. For more on hypothesis testing, see CHAPTER 16.

When you do the test, suppose your finding is in favour of the alternative hypothesis. (Estimating the likelihood that this finding is correct requires information beyond our scope here, but there are ways of testing which optimise that likelihood.) Your finding suggests a long-term persisting pattern in absenteeism. You then have grounds for recommending a suitable intervention to management.

Generalising to ‘a wider setting’ can also include to ‘a future setting’, as this example illustrates. In other words, statistical inference, appropriately applied, can offer a cautious way of forecasting the future – a dream that has fascinated humankind from time immemorial.

In short, statistical inference is a logical process that deals with ‘chancy’ data and generalises what those data reveal to wider settings. In those wider settings, it provides precise (as opposed to vague) conclusions which have a high chance of being correct.

---oOo---

But this seems paradoxical! What sort of logic is it that allows highly reliable conclusions to be drawn in the face of the world's uncertainties? (Here, and in what follows, we say 'highly reliable' as a shorter way of saying 'having a high chance of being correct'.)

To answer this pivotal question, we need first to offer you a short overview of the alternative *systems of logic* that philosophers have devised over the centuries. For an extended exposition, see Barker (2003).

A system of logic is a set of rules for reasoning from given assumptions towards reliable conclusions. There are just two systems of logic: *deduction* and *induction*. Each system contains two kinds of rules:

i) rules for drawing precise conclusions in all contexts where that logic is applicable; and
ii) rules for objectively assessing how likely it is that such precise conclusions are actually correct.

The conclusions that each system yields are called deductive inferences and inductive inferences, respectively.

It's worth a moment's digression to mention that there are two other thought processes – analogy and intuition – which are sometimes used in an attempt to draw reliable conclusions. However, these are not systems of logic, because they lack rules, either of the second kind (analogy) or of both kinds (intuition). Thus, conclusions reached by analogy or by intuition are, in general, less reliable than those obtained by deduction or induction. You will find in QUESTIONS 9.4 and 9.5, respectively, examples of the failure of analogy and of intuition.

In what kind of problem setting is deduction applicable? And in what kind of setting is induction applicable? The distinguishing criterion is whether the setting is (or is assumed to be) one of complete certainty.

In a setting of complete certainty, deduction is applicable, and there is no need for induction. Why? Because if all assumptions made (including the assumption that nothing is uncertain) are correct, and the rules of deduction are obeyed, then a deductive inference *must* be correct.

If you think back to the problems you solved in school mathematics (algebra, calculus, geometry and trigonometry), you will recall that, in these areas, chance influences were given no role whatever. No surprise, then, that deduction is the system of logic that underpins all mathematical inferences – which mathematicians call 'theorems'.

It is a great strength of deductively based theorems that they are *universally* correct (i.e. for every case where the same assumptions apply). For instance, given the assumptions of (Euclidean) plane geometry and the definition of a right-angled triangle, Pythagoras's Theorem is true for every such triangle, without exception.

Now, what about reasoning in a setting of uncertainty? Here, induction is applicable, and you can see how it contrasts with deduction. In a setting of uncertainty, even if all assumptions made are correct and the rules of induction are obeyed, an inductive inference *might not* be correct, because of chance influences, which are always at work.

Still, induction is more reliable in this setting than deduction, because the rules of induction explicitly recognise the influence of chance, whereas the rules of deduction make no mention of it whatever. In short, when the influence of chance is inescapable – as is the case in most real-world situations – induction is the system of logic that underpins all inferences.

If you head out one morning at the usual time to catch your regular 7.30 am train to work, you are reasoning inductively (or 'making an inductive inference'). Train timetables are vulnerable to bad weather delays, signal failures, and accidents along the rail line. So, even if, on all previous occasions, the 7.30 am train arrived on time, it is *not* correct to conclude that it *must* arrive on time today. Of course, the train is *highly likely* to arrive on time. But you cannot logically say more than that.

It follows that inductive inferences that are highly reliable in one circumstance are not necessarily highly reliable in other circumstances, *even where the same assumptions apply*. That is because chance influences can take many different forms, and always (by definition) come 'out of the blue'. For instance, even though the on-time arrival of your 7.30 am train has turned out to be highly reliable, reliability may shrink when you are waiting for your train home in the afternoon peak hours – the most likely period (our Sydney experience shows) in which unforeseen disruptions to train schedules occur.

---oOo---

We have now seen that it is inductive logic that enables inferences to be made in the face of uncertainty, and that such inferences need not be reliable in any particular instance. You may be thinking, 'it's no great achievement to produce unreliable conclusions'.

This thought prompts a new question: given that induction is the only system of logic that is applicable in chance situations, can rules of induction be configured to allow the conclusions it produces to be highly reliable in principle?

The answer is yes. Over the past century, statisticians have given a great deal of attention to refining the rules of induction that have come down to us through the cumulative work of earlier logicians, beginning with Francis Bacon (1561–1626). These refined rules of induction, now designed expressly for quantitative inferences, are called the rules of statistical induction.

The distinction between an inductive inference and a statistical inductive inference may seem both subtle and trivial. For an excellent discussion of the distinction, see chapters 4 and 5 of Burbidge (1990). While it is a subtle distinction, it is definitely not trivial. Relative to alternative ways of specifying rules of induction, the rules of statistical induction have, in principle, the highest chance of producing reliable conclusions in any particular instance.

In other words, *statistical inductive inference is the most reliable version of the most powerful logic that we have for reasoning about chance events.* As statisticians, we find this both fascinating and inspiring.

For simplicity, we shall now drop the formal term 'statistical inductive inference' and revert to using its conventional equivalent – 'statistical inference'.

Statistical description and statistical inference are the workaday roles of statistics. These two roles define the highways, so to speak, of statistical activity.

---oOo---

Statistics also has many byways. You will find them prominent in this book. Yet, they are not front and centre in statistics education curricula, nor are they part of the routine activities of applied statisticians. So, how do they come to attention?

Statistical theorists come upon several of these byways when refining and enhancing basic methods of analysis. In one category are *paradoxes of probability and statistics* (see, for some examples, Chapters 10 and 11). In another are *problems of using standard statistical techniques in non-standard situations* (see Chapter 17). In a third are *unifying principles*: fundamental ideas that are common to diverse areas of statistical theory. Discovering unifying principles means identifying previously unrecognised similarities in the subject. Unification makes the subject more coherent, and easier to understand as a whole. Examples of unifying principles are the Central Limit Theorem (see Chapters 12 and 14) and the power law (see Chapter 24).

Another byway is *the history of statistical ideas*. Historians with this special interest bring to life the philosophical standpoints, the intellectual explorations and the (sometimes controversial) writings of statistical pioneers, going back over several centuries. Though pioneering achievements may look straightforward to us in hindsight, the pioneers generally had to struggle to succeed – first, in formulating exactly what it was they were trying to solve, and then in harnessing all their insight, knowledge and creativity towards finding solutions, often in the face of sceptical critique (see, for instance, Chapters 18 and 22).

Yet another byway is *the social impact of statistics*. Here are three paths worth exploring on this byway: consequences of the low level of statistical literacy in the general community, and efforts to raise it (see Chapters 3 and 6); public recognition of statisticians' achievements via eponymy (see Chapter 23); and the negative effects of widespread public misuse of statistical methods, whether from inexperienced analysts' ignorance, or from a deliberate intention to deceive (see Chapters 8 and 9).

These are by no means all the byways of statistics. You will discover others for yourself, we hope, scattered through the following chapters.

Highways and byways.

You may then also come to share our view that, to people who are curious, the lightly visited byways of statistics can be even more delightful, more surprising and more fascinating than the heavily travelled highways of standard statistical practice.

If, at this point, you would like to refresh your knowledge of statistical ideas and principles, we recommend browsing the following technically very accessible books: Freedman, Pisani and Purves (2007), and Moore and Notz (2012).

Questions

Question 1.1 (A)

Figure 1.1 shows, on a logarithmic horizontal scale, the cumulative percentage frequency of heads in a sequence of 10,000 tosses of a coin.

These 10,000 tosses were performed by a South African statistician, John Kerrich, who went on to be the Foundation Professor of Statistics at Witwatersrand University in 1957.

a) Where, and under what unusual circumstances, did Kerrich perform these 10,000 tosses?

b) Does the information in the graph help us to define 'the probability of getting a head when a fair coin is tossed once'?

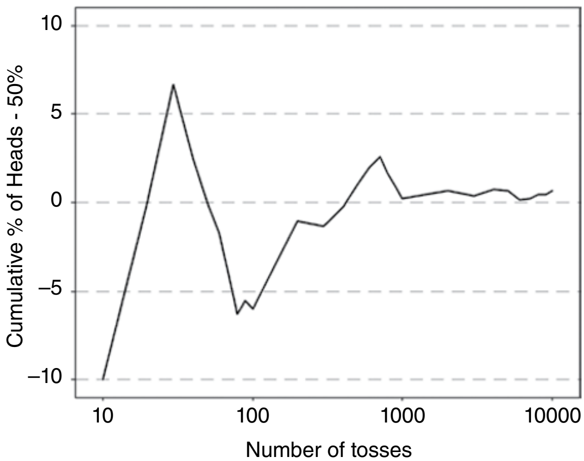

Figure 1.1 Scatterplot of Kerrich's coin-tossing results. Data from Freedman, Pisani and Purves (2007).

Question 1.2 (A)

When young children are asked about their understanding of probability, they quickly decide that the sample space for rolling a single die consists of six equally likely outcomes. When it comes to two dice, however, they often conclude that the sample space has 21 outcomes that are equally likely. Where does the number 21 come from?

Question 1.3 (A)

'Most people in London have more than the average number of legs.' Is this statement correct? Does it indicate some misuse of statistical methods?

Question 1.4 (A)

Based on thirty continuous years of recorded temperature data, the average temperature over the 12 months in a calendar year in New York is 11.7 °C, in New Delhi it is 25.2 °C, and in Singapore it is 27.1 °C. (To see the data – which may vary slightly over time – go online to [1.1], select the three cities in turn from the menu, and find the monthly average temperatures in the left-hand frame for each city.)

Does this mean that it gets roughly twice as hot in New Delhi during the year as it does in New York? Does it mean that the climate in Singapore is much the same as that in New Delhi?

Question 1.5 (B)

The map in Figure 1.2 shows a part of London. By whom was it drawn and when? With what famous event in the history of epidemiology is it connected? (*Hint*: note the street-corner pump.)

Figure 1.2 Extract from a map showing a part of London. Reproduced with the permission of Michael Rip.

References

Print

Barker, S.F. (2003). *The Elements of Logic*, 6th edition. McGraw-Hill.
Burbidge, J. (1990). *Within Reason: A Guide to Non-Deductive Reasoning.* Broadview Press.
Chatfield, C. (2002). Confessions of a pragmatic statistician. *The Statistician* **51**, 1–20.
Freedman, D., Pisani, R. and Purves, R. (2007). *Statistics*, 4th edition. Norton.
Moore, D.S. and Notz, W.I. (2012). *Statistics: Concepts and Controversies*, 8th edition. Freeman.

Online

[1.1] www.eurometeo.com/english/climate

2

How statistics differs from mathematics

'What's the difference between mathematics and statistics?' Many school students put this question to their teacher, aware that these subjects are related but not clear on what it is, exactly, that distinguishes them. Unravelling this puzzle is generally not made any easier for students by the fact that, in most schools around the world, it is the mathematics department that normally teaches statistics. To these curious but bewildered students, 'maths' seems to be defined by the topics that the teacher and the textbook say are maths, and similarly for 'stats'. So, algebra, calculus, geometry and trigonometry are 'maths', while frequency distributions, averages, sampling, the normal distribution, and estimation are 'stats'. That doesn't go very far towards providing a convincing answer to our opening question. Anyway, what about probability? Is that 'maths' or 'stats'?

A thoughtful teacher will want to supply a better answer. Surprisingly, in our experience, a better answer is rarely found either in curriculum documents or in textbooks. So let's see if we can formulate a better answer in a straightforward way.

A constructive start is to ask in what ways statistics problems differ from mathematics problems.

Here is something fairly obvious: statistics problems have a lot to do with getting a view of the variability in data collected from the real world. For example, a statistical problem may present 100 measurements (by different people) of the length of a particular object, using a tape measure, with the assigned task being to construct a frequency distribution of these measurements to see whether measurement errors tend to be symmetrical about the correct value, or whether people tend to veer more to one side or the other. By contrast, in a mathematical problem involving the length of an object,

A Panorama of Statistics: Perspectives, Puzzles and Paradoxes in Statistics, First Edition.
Eric Sowey and Peter Petocz.

Companion website: www.wiley.com/go/sowey/apanoramaofstatistics

the single measurement stated is simply to be taken to be the correct one, and the assigned task goes on from there.

If we ask why 100 people don't all produce exactly the same length measurement for the same object, using the same tape measure, we are led to a fundamental realisation. There are many factors at work in the physical act of measurement that cause different results to be reported for the same task by different people. Among these factors are: the attentiveness with which the task is undertaken; the effect of parallax in reading the scale marks on the tape; possible tremor in the hand holding the tape measure; and eyesight variations in those reporting measurements. Some of these factors might lead a person to mismeasure in a way that exceeds the correct value, while other factors might cause that same person to fall short of the correct value. Moreover, different people might react differently to any particular factor.

While it would be theoretically possible to study systematically some, or all, of these causes of variation individually (and there are contexts where it would be important to do so), it is generally convenient to lump all these real-world factors together and to refer to their net effect on measurement as *chance* (or *random*) *variation* around the correct value. This highlights the truth that chance influences are inseparable from almost all experience of life in the real world. (For more detail about the meaning of randomness, see Chapters 10 and 11.)

Chance variation has long been recognised. A famous passage in the biblical Book of Ecclesiastes, written some 2,200 years ago, shows how random events can have perplexing impacts: '… the race is not to the swift, nor the battle to the strong, nor bread to the wise, nor riches to the intelligent, nor favour to those with knowledge, but time and chance happen to them all.'

One may, of course, choose to abstract from chance influences (as the Public Transport Department does, for example, when it publishes a train timetable), but looking away from them should be understood as a deliberate act to simplify complex reality. In contexts where chance effects are ordinarily small (e.g. train journey times along a standard urban route), abstracting from chance is unlikely to cause decision errors to be made frequently (e.g. about when to come to the station to catch the train). However, where events are heavily dominated by random 'shocks' (e.g. daily movements in the dollar/pound exchange rate on international currency markets), predictions of what will happen even a day ahead will be highly unreliable most of the time.

As we mentioned in Chapter 1, school mathematical problems are generally posed in abstract settings of complete certainty. If, on occasion, a mathematical problem is posed in an ostensibly real-life setting, the student is nevertheless expected to abstract from all chance influences, *whether doing so is true to life or not.* Here is a typical example: try solving it now.

According to the timetable, a container train is allowed 90 minutes for a journey of 60 miles over mountainous country. On a particular trip, the train runs into fog and its speed is reduced, making it 25 minutes late at its destination. Had the fog appeared six miles closer to its point of departure, the train would have been 40 minutes late. At what rate does the train travel through fog?

(The answer is 15 miles per hour. Did you see in which way the problem is unrealistic and where chance influences are ignored? The train, incidentally, is pictured on the cover of this book.)

However appealing such problems may be for exhibiting the 'relevance' of mathematics, they risk concealing from students its fundamental logical limitation. The great physicist, Albert Einstein (1879–1955), expressed it concisely: 'As far as the propositions of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality.' (see Einstein (1921), online in the German original at [2.1] and in English translation at [2.2]).

We can elaborate Einstein's aphorism like this. In solving a problem in a real-life setting, the mathematical approach neglects all chance influences in that setting and, *on that account*, the mathematical solution is stated with certainty – but that solution is evidently an approximation to the solution in reality. Moreover, the error in the approximation is indeterminate. The statistical approach, by contrast, recognises the chance influences explicitly and, *on that account*, the statistical solution cannot be stated with certainty. The statistical solution, too, is an approximation to the solution in reality – but in the statistical approach, the error due to the chance influences *can* be dependably assessed within bounds.

Well, what about problems in probability? Self-evidently, they are problems about chance events but, here, calculating the probability of occurrence of some random event is the entire goal: it is simply an exercise in arithmetic according to predefined rules. Moreover, within the scope of the problem, it is certain that the calculated probability is correct. Therefore such problems, too, are mathematical problems. However, were the random event embedded in some inferential context, then the inferential problem would *thereby* be a statistical problem.

---oOo---

So far, we have seen that the central focus of statistics is on variation and, in particular, on chance variation. Mathematics acknowledges variables, but it does not focus *centrally* on their variation, and it abstracts entirely from the influence of chance.

The central focus of mathematics is on those general properties that are common to all the varying members of a set. Circles, for example, vary in their diameters, but circle theorems relate to the properties that all circles have in common, regardless of their particular diameter. Similarly, Pythagoras's Theorem is true for all plane right-angled triangles, regardless of their size.

While mathematicians' prime objective is to prove general theorems, which then imply truths about particular cases, statisticians proceed in reverse. They start with the 'particular' (namely, a sample of data) and, from there, they seek to make statements about the 'general' (that is, the population from which their data were sampled).

---oOo---

Finally, we come to the contrasting nature of numerical data in mathematical and in statistical problems. Data (literally, 'givens' – from the Latin) are indispensable inputs to any process of computational problem solving. However, 'data' mean different things to mathematicians and to statisticians.

As we have seen, to a mathematician data are values of non-random variables, and the task is to apply those numbers to evaluate a special case of a known general theorem – for example, to find the equation of the (*only*) straight line that passes (*exactly*) through two points with given coordinates on a plane. To a statistician data are values of random variables, and the statistician asks, 'How can I confidently identify the underlying systematic information that I think there is in these data, but that is obscured by the random variability?' For example, what is the equation of the *best-fitting* straight line that passes *as near as possible* to ten points with given coordinates, scattered about on the same plane in a pattern that looks roughly linear, and how well does that scatter fit to that line? The statistician also asks, 'How reliably can I generalise the systematic sample information to the larger population?' A large part of a practising statistician's work is the analysis of data – but a mathematician would never describe his or her work in this way.

As if practising statisticians did not have enough of a challenge in seeking out meaningful systematic information 'hidden' in their randomly-varying data, they must also be prepared to cope with a variety of problems of data quality – problems that could easily send their analysis in the wrong direction.

Among such problems are *conceptual errors* (a poor match between the definition of an abstract concept and the way it is measured in practice), *information errors* (e.g. missing data, or false information supplied by survey respondents), and *data processing errors* (e.g. the digital transposition that records the value 712 as 172).

What follows from this perspective on the ways in which statistics differs from mathematics? Statistics is not independent of mathematics, since all its analytical techniques (e.g. for summarising data and for making statistical inferences) are mathematically derived tools (using algebra, calculus, etc). However, in its prime focus on data analysis and on generalisation in contexts of uncertainty, and in its inductive mode of reasoning (described in Chapter 1), statistics stands on its own.

Questions

Question 2.1 (A)

a) Pure mathematics deals with abstractions. In geometry, for example, lines have length but no breadth, planes have length and breadth but no thickness, and so on. In this setting, we pose the problem: a square has sides 20 cm long; what is the length of its diagonal? How does the mathematician answer? Is this answer accurate?

b) Statistics deals with real-world data – measurements of actual objects and observations of actual phenomena. On graph paper, construct, as precisely as you can, a square with sides 20 cm long. Then ask, say, 25 different people to use their own millimetre-scale rulers to measure the length of the square's diagonal, taking care to be as accurate as possible. Record these measurements. In this setting, we pose the problem: what is the length of the diagonal? How does the statistician answer? Is this answer accurate?

Question 2.2 (A)

Mathematical induction is a versatile procedure for constructing a proof in mathematics. Explain the general approach underlying mathematical induction, and give an example of its use. Does mathematical induction use deductive logic or inductive logic? What is the implication of your answer to this question?

Question 2.3 (A)

Consider the mathematical expression n^2+n+41. If $n=0$, the expression has the value 41, which is a prime number. If $n=1$, the expression has the value 43, also a prime number.

a) Make further substitutions of $n = 2, 3, 4, 5$ and, in each case, check whether the result is a prime number. What does this suggest?
b) Repeat for $n = 6, \ldots, 10$ (or further, if you like). What inference would you draw statistically from this accumulating information?
c) Is this inference actually correct? Can you prove or disprove it, mathematically?

Question 2.4 (A)

Sherlock Holmes, the famous fictional consulting detective created by the British novelist Sir Arthur Conan Doyle (1859–1930), solved crimes by reasoning from data in ways similar to a statistician's reasoning. What similarities can you see in the following passage (from 'The Five Orange Pips' in the book *The Adventures of Sherlock Holmes*)?

> 'Sherlock Holmes closed his eyes, and placed his elbows upon the arms of his chair, with his fingertips together. "… [W]e may start with a strong presumption that Colonel Openshaw had some very strong reason for leaving America. Men at his time of life do not change all their habits, and exchange willingly the charming climate of Florida for the lonely life of an English provincial town. His extreme love of solitude in England suggests the idea that he was in fear of someone or something, so we may assume as a working hypothesis that it was fear of someone or something which drove him from America. As to what it was he feared, we can only deduce that by considering the formidable letters which were received by himself and his successors. Did you remark the postmarks of those letters?" "The first was from Pondicherry, the second from Dundee, and the third from London" [replies Dr Watson]. "From East London. What do you deduce from that?" "They are all seaports. That the writer was on board a ship." "Excellent. We have already a clue. There can be no doubt that the probability – the strong probability – is that the writer was on board of a ship."'

Why is the word 'deduce' inappropriate in the above passage? What word should Conan Doyle have used instead?

Question 2.5 (A)

How can one find an approximation to the value of π, the ratio of a circle's circumference to its diameter, by way of a statistical experiment that involves tossing a needle randomly onto a flat surface ruled with parallel lines? What is the name of the 18th century polymath with whom this experiment is associated?

References

Online

[2.1] Einstein, A. (1921), *Geometrie und Erfahrung*. Erweiterte Fassung des Festvortrages gehalten an der Preussischen Akademie der Wissenschaften zu Berlin am 27. Januar 1921. Julius Springer, Berlin. Pages 3–4. At https://archive.org/details/geometrieunderf00einsgoog

[2.2] Einstein, A. (1921), *Geometry and Experience*. Lecture before the Prussian Academy of Sciences, January 27, 1921. Julius Springer, Berlin. In The Collected Papers of Albert Einstein. At http://einsteinpapers.press.princeton.edu/vol7-trans/225