

# There Is More to Assessing Risk Than Statistics

## 1.1 Introduction

The financial crisis that started in 2008 brought misery to millions around the world. It also raised awareness of the need for improved methods of risk assessment. The armies of analysts and statisticians employed by banks and government agencies had failed to predict either the event or its scale until far too late. Yet the methods that could have worked—that are the subject of this book—were largely ignored. Moreover, the same methods have the potential to transform risk analysis and decision making in all walks of life including medicine and the law as well as business.

Examples of the kind of problems we want to be able to solve include the following:

- **Medical**—Imagine you are responsible for diagnosing a medical condition and for prescribing one of a number of possible treatments. You have some background information about the patient (some information is objective like age and number of previous operations, but other information is subjective, like ‘depressed’ and ‘stressed’). You also have some prior information about the prevalence of different possible conditions (for example, bronchitis may be ten times more likely in a certain population than cancer). You run some diagnostic tests about which you have some information of the accuracy (such as the chances of the test outcome positive for a condition that is not present and negative for a condition that is present). You also have various bits of information about the success rates of the different possible treatments and their side effects. On the basis of all this information how do you arrive at a decision of which treatment pathway to take? And how would you justify that decision if something went wrong? If something went wrong, you may be open to negligence. The issue is about justifying your actions if your contingent diagnosis turns out to have been the wrong one.
- **Legal**—As a member of a jury you hear many pieces of evidence in a trial. Some of the evidence favors the prosecution

The methods described in this book—based on Bayesian networks—are currently accessible only to mathematical and statistical specialists. The aim of this book is to make them accessible to anybody with an interest in risk assessment who has not forgotten their high school math.

---

Visit [www.bayesianrisk.com](http://www.bayesianrisk.com) for your free Bayesian network software and models in this chapter

hypothesis of guilty and some of the evidence favors the defence hypothesis of innocence. Some of the evidence is statistical (such as the match probability of a DNA sample) and some is purely subjective, such as a character witness statement. It is your duty to combine the value of all of this evidence to arrive at a probability of innocence. If the probability value you arrive at is sufficiently small ('beyond reasonable doubt') you must return a guilty verdict. How would you arrive at a decision? Similarly, before a case comes to trial how should a member of the criminal justice system, the police, or a legal team determine the value of each piece of evidence and then determine if, collectively, the evidence is sufficient to proceed to trial?

- **Safety**—A transport service (such as a rail network or an air traffic control center) is continually striving to improve safety but must nevertheless ensure that any proposed improvements are cost effective and do not degrade efficiency. There is a range of alternative competing proposals for safety improvement, which depend on many different aspects of the current infrastructure (for example, in the case of an air traffic control center alternatives may include new radar, new collision avoidance detection devices, or improved air traffic management staff training). How do you determine the 'best' alternative taking into account not just cost but also impact on safety and efficiency of the overall system? How would you justify any such decision to a team of government auditors?
- **Financial**—A bank needs sufficient liquid capital readily available in the event of exceptionally poor performance (either from credit or market risk events, or from catastrophic operational failures of the type that brought down Barings in 1995 and threatened Société Générale in 2007). It has to calculate and justify a capital allocation that properly reflects its 'value at risk'. Ideally this calculation needs to take account of a multitude of current financial indicators, but given the scarcity of previous catastrophic failures, it is also necessary to consider a range of subjective factors such as the quality of controls in place at different levels of the bank hierarchy and business units. How can all of this information be combined to determine the real value at risk in a way that is acceptable to the regulatory authorities and shareholders?
- **Reliability**—The success or failure of major new products and systems often depends on their reliability, as experienced by end users. Whether it is a high-end digital TV, a software operating system, or a complex military vehicle, like a tank, too many faults in the delivered product can lead to financial disaster for the producing company or even a failed military mission, including loss of life. Hence, pre-release testing of such systems is critical. But no system is ever perfect and a perfect system delivered after a competitor gets to the market first may be worthless. So how do you determine when a system is 'good enough' for release, or how much more testing is needed? You

may have hard data in the form of a sequence of test results, but this has to be considered along with subjective data about the quality of testing and the realism of the test environment.

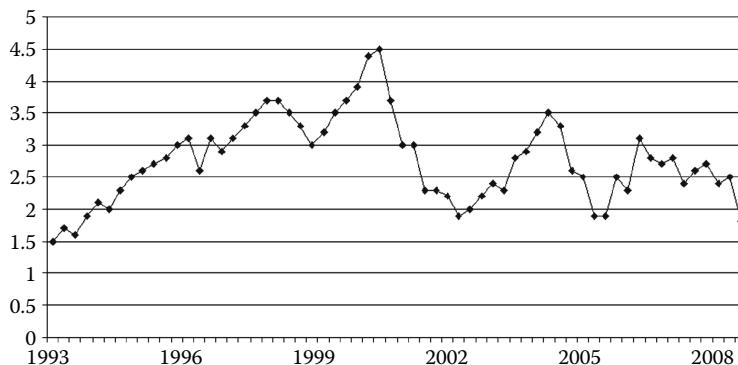
What is common about all of the aforementioned problems is that a ‘gut-feel’ decision based on doing all the reasoning ‘in your head’ or on the back of an envelope is fundamentally inadequate and increasingly unacceptable. Nor can we base our decision on purely statistical data of ‘previous’ instances, since in each case the ‘risk’ we are trying to calculate is essentially unique in many aspects.

The aim of this chapter is to introduce, by way of examples and stories, some of the basic tools needed for this type of risk assessment. We do this by introducing and analysing the traditional statistics and data analysis methods that have previously been used. In doing so you should be able to understand both their strengths and limitations for risk assessment. In particular, in discussing their limitations, we also provide a gentle introduction to the power of causal models (which are often implemented as Bayesian networks).

## 1.2 Predicting Economic Growth: The Normal Distribution and Its Limitations

Table 1.1 contains the (annualized) growth rate figures for the UK for each quarter from the start of 1993 to the end of 2007. So, for example, in the fourth quarter of 2007 the annual growth rate in the UK was 2.36%.

Data such as this, especially given the length of time over which it has been collected, is considered extremely valuable for financial analysis and projections. Since so many aspects of the economy depend on the growth rate, we need our predictions of it for the coming months and years to be very accurate. So imagine that you were a financial analyst presented with this data in 2008. Although it would be nice to be able to predict the growth rate in each of the next few years, the data alone gives you little indication of how to do that. If you plot the growth over time as in Figure 1.1 there is no obvious trend to spot.



**Figure 1.1** Growth rate in GDP (%) over time from first quarter 1993 to first quarter 2008.

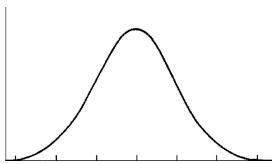
**Table 1.1**  
Quarterly Annualized UK Growth Rate Figures 1993-2007 Adjusted for Inflation

Quarter	Annual GDP%	Quarter	Annual GDP%
1993 Q1	1.42	2000 Q4	3.04
1993 Q2	2.13	2001 Q1	3.08
1993 Q3	2.50	2001 Q2	2.31
1993 Q4	2.82	2001 Q3	2.27
1994 Q1	3.29	2001 Q4	2.19
1994 Q2	4.28	2002 Q1	1.79
1994 Q3	4.83	2002 Q2	1.95
1994 Q4	4.70	2002 Q3	2.19
1995 Q1	3.95	2002 Q4	2.44
1995 Q2	3.01	2003 Q1	2.29
1995 Q3	2.76	2003 Q2	2.83
1995 Q4	2.51	2003 Q3	2.88
1996 Q1	3.14	2003 Q4	3.23
1996 Q2	3.03	2004 Q1	3.58
1996 Q3	2.54	2004 Q2	3.22
1996 Q4	2.84	2004 Q3	2.57
1997 Q1	2.70	2004 Q4	2.45
1997 Q2	3.13	2005 Q1	1.81
1997 Q3	3.48	2005 Q2	1.97
1997 Q4	3.90	2005 Q3	2.50
1998 Q1	3.85	2005 Q4	2.40
1998 Q2	3.67	2006 Q1	3.16
1998 Q3	3.52	2006 Q2	2.71
1998 Q4	3.39	2006 Q3	2.59
1999 Q1	3.13	2006 Q4	2.70
1999 Q2	3.29	2007 Q1	2.59
1999 Q3	3.72	2007 Q2	2.88
1999 Q4	3.74	2007 Q3	2.91
2000 Q1	4.37	2007 Q4	2.36
2000 Q2	4.55	2008 Q1	1.88
2000 Q3	3.73		

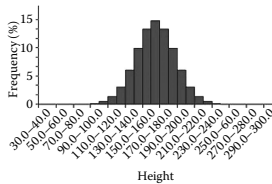
Source: UK Office of National Statistics.

The kind of things financial institutions would really like to know:

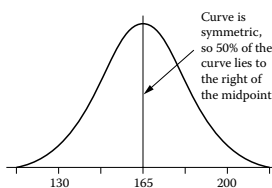
- What are the chances that the next year's growth rate will be between 1.5% and 3.5% (stable economy)?
- What are the chances that the growth will be less than 1.5% in each of the next three quarters?
- What are the chances that within a year there will be negative growth (recession)?



**Figure 1.2** Normal distribution (bell curve).



**Figure 1.3** Histogram of people's height (centimetres).



**Figure 1.4** Normal distribution model for people's height.

One of the additional benefits of using the Bayesian Network tool AgenaRisk that accompanies this book is that it automatically does these calculations for you.

But there is a lot that you can do other than making 'point' predictions. What financial institutions would really like to know is the answer to questions like those in the sidebar.

Indeed, economic analysts feel that the kind of data provided enables them to answer such questions very confidently. The way they typically proceed is to 'fit' the data to a standard curve (also called a *statistical distribution*). The answers to all the aforementioned questions can then be answered using standard statistical tables associated with that distribution.

In most cases the analysts assume that data of the kind seen here can be fitted by what is called a *Normal distribution* (also called a *bell curve* because that is its shape as shown in Figure 1.2).

The key thing about a Normal distribution is that it is an 'idealized' view of a set of data. Imagine that, instead of trying to model annual growth rate, you were trying to model the height in centimeters of adults. Then, if you took a sample of, say, 1,000 adults and plotted the frequency of their heights within each 10-centimeter interval you would get a graph that looks something like Figure 1.3. As you increase the sample size and decrease the interval size you would eventually expect to get something that looks like the Normal distribution in Figure 1.4.

The Normal distribution has some very nice mathematical properties (see Box 1.1), which makes it very easy for statisticians to draw inferences about the population that it is supposed to be modelling.

Unfortunately, it turns out that, for all its nice properties the Normal distribution is often a very poor model to use for most types of risk assessment. And we will demonstrate this by returning to our GDP growth rate data. In the period from 1993 to 2008 the average growth rate was 2.96% with a standard deviation of 0.75.

Following the approach as described earlier for the height example, we can create an appropriate histogram of the growth rate data and fit a Normal distribution to it as shown in Figure 1.7.

The fitted Normal curve has a mean of 2.96 and a standard deviation of 0.75. Using standard tables (plus a little of the kind of probability that you will learn about in Chapters 3 and 4) this enables us to answer the original questions that we posed (see the sidebar).

Things turned out very differently from these optimistic predictions, as the actual data (including the period up to the beginning of 2010) shown in Figure 1.8 and Table 1.2 clearly demonstrate.

Within less than a year the growth rate was below  $-5\%$ . According to the model a growth rate below  $-5\%$  would happen considerably less frequently than once every 14 billion years (i.e., the estimated age of the universe).

Actual predictions made by financial institutions and regulators in the period running up to the credit crunch were especially optimistic because they based estimates of growth on the so-called Golden Period of 1998–2007.

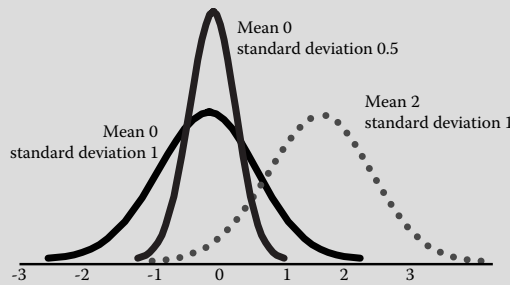
So what went wrong? Clearly the Normal distribution was a hopelessly inadequate model. It looked like a reasonable fit for the data in the period 1993–2008 (when there was not a single period of negative growth; indeed growth was never below 1.8%). But although the tails are infinite they are *narrow* in the sense we explained earlier: observations a long way from the mean are almost impossible. *Hence, a Normal distribution is inherently incapable of predicting many types of rare events.*

**Box 1.1 Properties of the Normal Distribution**

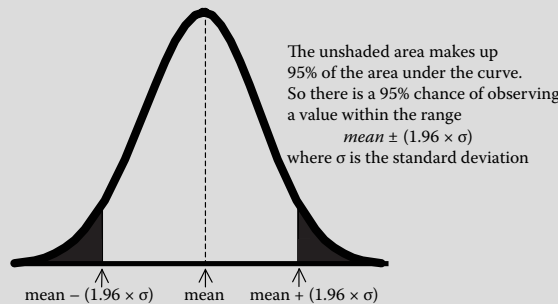
- The distribution is symmetric around the midpoint, which is called the *mean* of the distribution. Because of this exactly half of the distribution is greater than the mean and half less than the mean. If the model is a good fit of the real data then the mean of the distribution should be close to the mean of the data from which it is drawn (that is, just the average calculated by adding all the data and dividing by the number of points). So, from Figure 1.4, we can infer that there is a 50% chance that a randomly selected adult in the UK will be taller than 165 cm.
- The ‘spread’ of the data (which is the extent to which it varies from the mean) is captured by a single number called the *standard deviation*. Examples of Normal distributions with different standard deviations are shown in Figure 1.5. Once we know the mean and the standard deviation there are tables and calculators that tell us what proportion of the distribution lies between any two points.

So, for example, the distribution that supposedly models height in Figure 1.4 has a mean of 165 and a standard deviation of 14. Approximately 20% of such a distribution lies between 178 and 190, so if this distribution really is an accurate model of height we can conclude that there is a 20% chance a randomly selected adult will be between 178 and 190 cm tall.

- As shown in Figure 1.6 it is always the case that 95% of the distribution lies between the mean and plus or minus 1.96 times the standard deviation. So (by symmetry) in the height example, 2.5% of the distribution lies above 195 cm. This means there is a 2.5% chance a randomly selected person is taller than 195cm.
- The Normal distribution approaches zero frequency in both directions, towards plus and negative infinity, but never reaches it. So, no matter how far we go away from the mean the curve never quite touches zero on the frequency axis (mathematicians say it is *asymptotic*). However, as we move away from the mean we very quickly get into tiny regions of the curve. For example, less than 0.001% of the distribution lies beyond the mean plus four standard deviations. So, in our height example less than 0.001% of the distribution lies to the right of 224 cm. This means that there is less than 1 in a 100,000 chance of an adult being more than 224 cm tall according to the model. Although the infinite tails of the Normal distribution are very useful in the sense that the model poses no limit on the possible values that can appear, it also leads to ‘impossible’ conclusions. For example, the model implies there is a nonzero (albeit very small) chance that an adult can be less than zero centimeters tall, and a nonzero chance that an adult can be taller than the Eiffel Tower.



**Figure 1.5** Examples of different Normal distributions.



**Figure 1.6** Normal distribution: 95% lies between the mean  $\pm 1.96$  standard deviations.

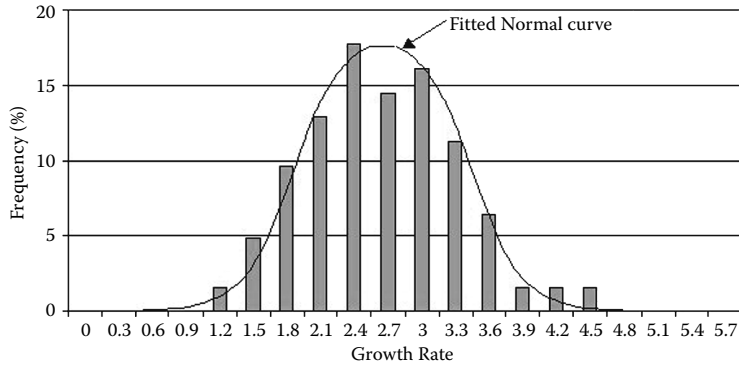
**Table 1.2**  
**Quarterly Annualized**  
**UK Growth Rate Figures**  
**from 2008 Q2-2010**  
**Adjusted for Inflation**

Quarter	Annual GDP%
2008 Q2	1.04
2008 Q3	-0.40
2008 Q4	-2.75
2009 Q1	-5.49
2009 Q2	-5.89
2009 Q3	-5.28
2009 Q4	-2.87

Source: UK Office of National Statistics.

Answers to original questions if we assume Normal distribution:

- What are the chances that the next quarter growth rate will be between 1.5% and 3.5%? Answer based on the model: approximately 72%.
- What are the chances that the growth will be less than 1.5% in each of the next three quarters? Answer: about 0.0125%, which is 1 in 8,000.
- What are the chances that within a year there will be negative growth (recession)? Answer: about 0.0003%, which is less than 1 in 30,000.

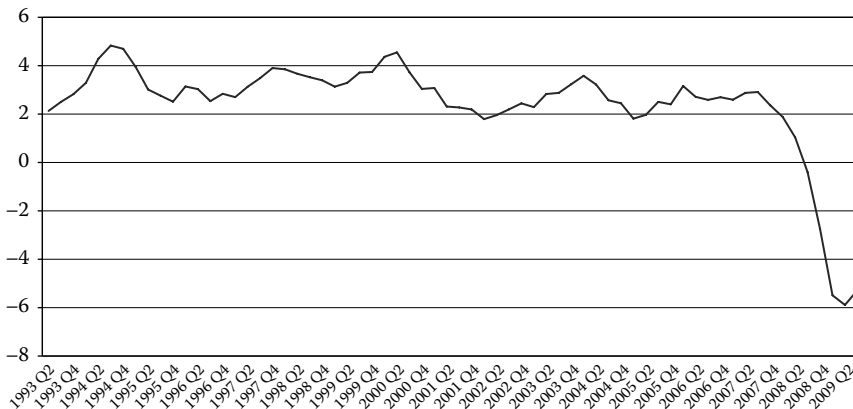


**Figure 1.7** Histogram of annualized GDP growth rate from 1993 to 2008.

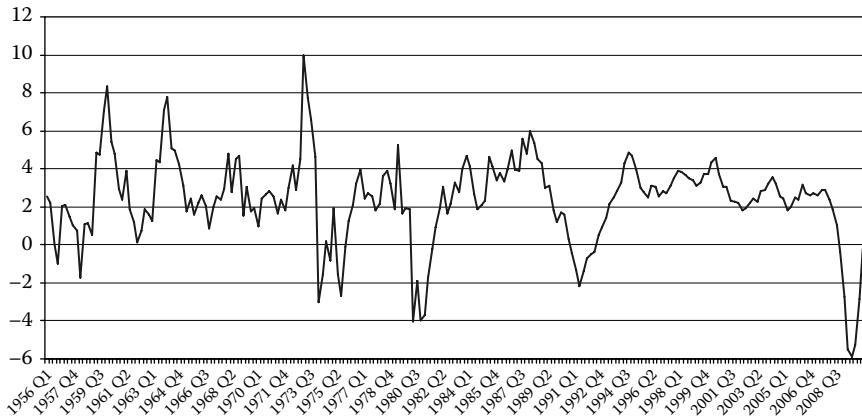
Whereas analysts like to focus on the most recent data as being the most relevant, especially during periods of prolonged stability such as had been experienced, a look at the longer term data reveals much greater potential for volatility. This can be seen in Figure 1.9, which charts the growth rate for the period 1956–2010.

When we plot these quarterly growth rates as a histogram (Figure 1.10) it looks very different from the histogram we saw in Figure 1.7 of the period 1993–2008. Not only is the spread of the distribution much wider, but it is clearly not ‘Normal’ because it is not symmetric.

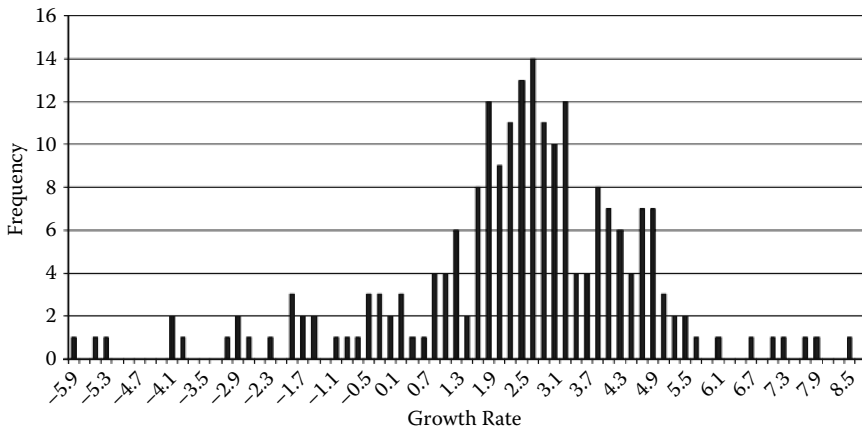
Unfortunately, while basing predictions on the longer term data may have led to slightly more realistic results (in the sense of being less optimistic) even this data (and any amount of previous data that we may have collected) would still have been insufficient to predict the scale of the collapse in growth. The conditions prevailing in 2008 were unlike any that had previously been seen. The standard statistical approaches inevitably fail in such cases.



**Figure 1.8** Growth rate in GDP (%) over time from 1993 to 2010.



**Figure 1.9** Growth rate in GDP (%) over time from 1956 to 2010.



**Figure 1.10** Histogram of annualized GDP growth rate from 1956 to 2010.

### 1.3 Patterns and Randomness: From School League Tables to Siegfried and Roy

Take a look at Table 1.3. It shows the scores achieved (on an objective quality criteria) by the set of state schools in one council district in the UK. We have made the schools anonymous by using numbers rather than names. School 38 achieved a significantly higher score than the next best school, and its score (175) is over 52% higher than the lowest ranked school, number 41 (score 115). Tables like these are very important in the UK, since they are supposed to help provide informed ‘choice’ for parents. Based on the impressive results of School 38 parents clamour to ensure that their child gets a place at this school. Not surprisingly, it is massively oversubscribed. Since these are the only available state schools in this district, imagine how you would feel if, instead of your child being awarded a place in School 38, he or she was being sent to school 41. You would be pretty upset, wouldn’t you?

**Table 1.3**  
School League Table

School Number	Score
38	175
43	164
44	163
25	158
31	158
47	158
11	155
23	155
48	155
40	153
7	151
30	151
6	150
9	149
33	149
19	148
10	147
12	147
32	147
2	146
27	146
42	146
28	145
35	145
49	145
45	144
46	143
1	142
18	142
22	141
26	141
4	140
14	140
29	140
39	139
8	138
5	136
17	136
34	136
3	134
24	133
36	131
37	131
15	130
21	130
16	128
13	120
20	116
41	115

You should not be. We lied. The numbers do not represent schools at all. They are simply the numbers used in the UK National Lottery (1 to 49). And each ‘score’ is the *actual number of times* that particular numbered ball had been drawn in the 1,172 draws of the UK National Lottery that had taken place up to 17 March 2007. So the real question is: Do you believe that 38 is a ‘better’ number than 41? Or, making the analogy with the school league table more accurate:

*Do you believe the number 38 is more likely to be drawn next time than the number 41?* (Since the usual interpretation of the school league table is that if your child attends the school at the top he or she will get better grades than if he or she attends the school at the bottom.)

The fact is that the scores are genuinely random. Although the ‘expected’ number of times any one ball should have been drawn is about 144 you can see that there is a wide variation above and below this number (even though that is still the average score).

What many people fail to realise is that this kind of variation is *inevitable*. It turns out that in any sequence of 1,172 lottery draws there is about a 50% chance that at least half the numbers will be chosen either less than 136 times or more than 152 times. That indeed is roughly what happened in the real sample. Moreover, the probability that at least one number will be chosen more than 171 times is about 45%. You may find it easier to think of rolling a die 60 times. You would almost certainly not get each of the six numbers coming up 10 times. You might get 16 threes and only 6 fours. That does that not make the number three ‘better’ than the number four. The more times you roll the die, the closer in relative terms will be the frequencies of each number (specifically, if you roll the die  $n$  times the frequency of each number will get closer to  $n$  divided by 6 as  $n$  gets bigger); but in absolute terms the frequencies will not be exactly the same. There will inevitably be some numbers with a higher count than others. And one number will be at the ‘top’ of the table while another will be ‘bottom’.

We are not suggesting that all school league tables are purely random like this. But, imagine that you had a set of genuinely equal schools and you ranked them according to a suitable criteria like average exam scores. Then, in any given year, you would inevitably see variation like the earlier table. And you would be wrong to assume that the school at the top was better than the school at the bottom. In reality, there may be inherent quality factors that help determine where a school will come in a league table. But this does not disguise the fact that much of the variation in the results will be down to nothing more than pure and inevitable chance. See Box 1.2 for another example.

Unfortunately, many critical decisions are made based on wrongly interpreting purely random results exactly like these, even though the randomness was entirely *predictable* (indeed, we will show in Chapters 3 and 4 that probability theory and statistics are perfectly adequate for making such predictions).



**Box 1.2 Sporting Form: Quality or Luck**

The English premier league consists of 20 teams. The 20 competing in 2008–2009 are shown in Table 1.4. This table also shows the results between any two teams where we have used 2 to represent victory for the first named team, 1 to represent draw, and 0 to represent defeat for the first named team.

**Table 1.4**  
Premiership Results

	Ars	AV	Bla	Bol	Che	Ev	Ful	Hul	Liv	MC	MU	Mid	New	Por	Sto	Sun	Tot	WB	WH	Wig
Arsenal		2	1	1	2	1	1	0	0	0	2	2	2	2	0	1	1	2	2	1
Aston Villa			1	0	0	2	2	1	0	1	1	2	0	2	2	1	2	2	1	1
Blackburn				2	1	2	0	1	1	1	1	0	0	0	0	2	2	1	1	
Bolton					1	2	0	0	1	0	1	1	0	0	0	1	0	2	0	0
Chelsea						0	2	2	2	1	0	1	1	0	1	0	1	0	2	1
Everton							2	1	0	0	2	0	1	1	0	1	2	1	0	0
Fulham								2	2	2	2	2	2	2	2	1	1	1	2	1
Hull									2	0	2	0	0	0	0	0	1	2	1	1
Liverpool										1	0	0	2	2	1	1	2	2	2	0
Man City											1	0	2	0	1	0	1	0	1	0
Man Utd												0	2	1	0	2	0	1	2	2
Middlesbrough													2	0	0	2	0	1	1	2
Newcastle														2	1	2	2	0	2	0
Portsmouth															0	1	0	1	1	2
Stoke																0	1	2	1	1
Sunderland																	0	0	2	0
Tottenham																		2	0	2
West Brom																			1	1
West Ham																				0
Wigan																				

So, for example, the 2 in the cell with row Arsenal and column AV (Aston Villa) means that Arsenal defeated Aston Villa.

This data is not real. We generated random results in each result cell. Based on these randomly generated results we determine the league table as shown in Table 1.5a using the premier league convention of 3 points for a win, 1 for a draw, and 0 for defeat.

It seems difficult to understand when you look at this table that there really is no difference in ‘form’ between Fulham (runaway champions) and Bolton (cast adrift and doomed to relegation). However, the career of many team managers will depend on perceptions of their performance. How then can we discriminate between those managers that are merely lucky and those that are genuinely competent?

What is striking about this table is how, *in its distribution of points*, it looks little different from the actual premier league table as it stood in week 19 of the season (January 2009) when all teams had played each other once (Table 1.5b).

In fact, most serious real-world ‘risks’ are not like lotteries, and a very good illustration of this is to contrast the predictable risks confronting casinos (whose job is to run different types of lotteries) with less predictable and far more serious ones.

**Table 1.5**  
**Premiership Table (a)**

Fulham	38
Stoke	34
Arsenal	31
Middlesbrough	31
Sunderland	31
Liverpool	29
Wigan	29
Aston Villa	27
Newcastle	27
Portsmouth	27
Tottenham	27
Man Utd	24
Chelsea	23
Man City	23
Blackburn	21
Hull	21
West Brom	19
Everton	18
West Ham	17
Bolton	14

<sup>a</sup> Based on random results

**Table 1.5**  
**Premiership Table (b)**

Liverpool	42
Chelsea	41
Man Utd	41
Aston Villa	35
Arsenal	32
Everton	29
Hull	27
Fulham	25
Wigan	25
Bolton	23
Portsmouth	23
Newcastle	22
Sunderland	22
West Ham	22
Man City	21
Middlesbrough	20
Stoke	20
Tottenham	20
Blackburn	17
West Brom	15

<sup>b</sup> Real table after 19 games Jan 2009

The risk for a casino of making massive losses to gamblers is entirely predictable because of our understanding of the ‘mechanical’ uncertainty of the games played. Hence this risk of ruin is easily controlled and avoided. For example, simply placing a limit on the highest bet means that in the long run the casinos cannot lose really large sums because of the odds in their favor. And to minimize the risk of losses to cheating (i.e., where gamblers use techniques that swing the odds in their own favor) the casinos spend significant sums on security even though, again, such risks are also foreseeable.

Yet, the casinos are often blind to the kind of risks that could really bring them down, and there was no better example of this than what happened at the Mirage Resort and Casino on 3 October 2003. For many years the illusionists Siegfried and Roy had been the biggest draw in Las Vegas with their nightly show (Figure 1.11). A key part of the show was Roy’s interaction with tigers. Roy had lived and even slept with tigers for some 40 years without any incident. Yet, after over 5,000 performances, on that fateful night Roy was mauled by one of his beloved tigers, causing life-threatening injuries. The show was immediately and permanently closed leading to the dismissal of 267 staff and performers and other losses (from ticket sales, hotel bookings, and legal costs) of hundreds of millions of dollars, making it the single worst loss in Las Vegas history. The casino managers—who you would think are the ultimate experts in risk assessment—were ‘beaten’ by a risk they had not even considered. The magnitude of the resulting losses from that risk dwarfed the largest possible loss they could ever have suffered from the very predictable risks they had spent millions protecting themselves against.

This example was cited in an excellent book by Nassim Taleb whose title, *The Black Swan*, is the expression used to describe highly unpredictable events (for centuries all swans were assumed to be white because nobody had yet seen a black swan; the sighting of a black swan reversed all previous beliefs about the colour of swans). For the Mirage what happened on 3 October 2003 was a devastating black swan event.

Traditional statistical methods, which, we will see, rely on ideas like ideal lotteries, repeated observations and past data, cannot predict black swan events. Some, including Taleb himself, argue that no methods can predict such ‘unknowns’ and ‘unknowables’. We believe the methods advocated in this book allow for management and assessment of risks that include black swan events. We will see this in Chapter 11 on operational risk.

## 1.4 Dubious Relationships: Why You Should Be Very Wary of Correlations and Their Significance Values

Correlations and significance values (also called *p-values*) are the standard techniques that statisticians use to determine whether there are genuine relationships between different variables. In the approach to

probability that we espouse in this book these traditional techniques (along with their first-cousins *regression analysis* and *confidence intervals*, which we will look at in Chapter 10) are superseded by techniques that we feel are simpler and more intuitive. But the acceptance and entrenchment of these ideas are so widespread across all empirical disciplines that you need to be aware of what they are in order to appreciate the damage they can do to rational decision making and risk analysis.

Look at Table 1.6. This gives (a) the average temperature and (b) the number of automobile crashes resulting in fatalities in the United States in 2008 broken down by month (source: U.S. Department of Transport 2008). We can plot this data in a scatterplot graph as shown in Figure 1.12.

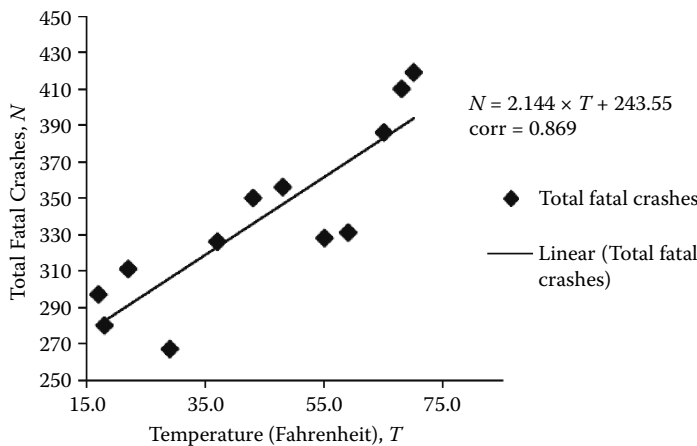
From a quick view of the chart there is a relationship between temperature and fatalities. There seem to be more fatalities as the temperature increases. Statisticians use a formula—called the *correlation coefficient* (see Box 1.3)—that measures the extent to which the two sets of numbers are related. You do not need to know what this formula is because any spreadsheet package like Excel will do the calculation for you. It so happens that the correlation coefficient in this case is approximately 0.869. Using standard tables this turns out to be ‘highly significant’ (comfortably passing the criteria for a *p*-value of 0.01 that is also explained in Box 1.3). Statisticians might conclude from this data that the number of road fatalities and the minimum temperature on any given day are significantly related (although note that we have severe concerns about the limitations of *p*-values as explained in Box 1.3).



**Figure 1.11** Siegfried and Roy (pre-2003), illustrated by Amy Neil, aged 12.

**Table 1.6**  
Temperature and Fatal Automobile Crashes

Month	Average Temperature	Total Fatal Crashes
January	17.0	297
February	18.0	280
March	29.0	267
April	43.0	350
May	55.0	328
June	65.0	386
July	70.0	419
August	68.0	410
September	59.0	331
October	48.0	356
November	37.0	326
December	22.0	311



**Figure 1.12** Scatterplot of temperature against road fatalities (each dot represents a month).

### Box 1.3 Correlation Coefficient and $p$ -Values: What They Are and Why You Need to Be Very Wary of Them

The correlation coefficient is a number between  $-1$  and  $1$  that determines whether two paired sets of data (such as those for *height* and *intelligence* of a group of people) are related. The closer to  $1$  the more 'confident' we are of a positive linear correlation and the closer to  $-1$  the more confident we are of a negative linear correlation (which happens when, for example, one set of numbers tends to decrease when the other set increases as you might expect if you plotted a person's age against the number of toys they possess). When the correlation coefficient is close to zero there is little evidence of any relationship.

Confidence in a relationship is formally determined not just by the correlation coefficient but also by the number of pairs in your data. If there are very few pairs then the coefficient needs to be very close to  $1$  or  $-1$  for it to be deemed 'statistically significant', but if there are many pairs then a coefficient closer to  $0$  can still be considered 'highly significant'.

The standard method that statisticians use to measure the 'significance' of their empirical analyses is the  $p$ -value. Suppose we are trying to determine if the relationship between height and intelligence of people is significant; then we start with the 'null hypothesis', which, in this case is the statement 'height and intelligence of people are unrelated'. The  $p$ -value is a number between  $0$  and  $1$  representing the probability that this data would have arisen if the null hypothesis were true. In medical trials the null hypothesis is typically of the form that 'the use of drug  $X$  to treat disease  $Y$  is no better than not using the drug'.

The calculation of the  $p$ -value is based on a number of assumptions that are beyond the scope of this discussion, but people who need  $p$ -values can simply look them up in standard statistical tables (they are also computed automatically in Excel when you run Excel's regression tool). The tables (or Excel) will tell you, for example, that if there are  $100$  pairs of data whose correlation coefficient is  $0.254$ , then the  $p$ -value is  $0.01$ . This means that there is a  $1$  in  $100$  chance that we would have seen these observations if the variables were unrelated.

A low  $p$ -value (such as  $0.01$ ) is taken as evidence that the null hypothesis can be 'rejected'. Statisticians say that a  $p$ -value of  $0.01$  is 'highly significant' or say that 'the data is significant at the  $0.01$  level'.

A competent researcher investigating a hypothesized relationship will set a  $p$ -value in advance of the empirical study. Typically, values of either  $0.01$  or  $0.05$  are used. If the data from the study results in a  $p$ -value of less than that specified in advance, the researchers will claim that their study is significant and it enables them to reject the null hypothesis and conclude that a relationship really exists.

In their book *The Cult of Statistical Significance* Ziliak and McCloskey expose a number of serious problems in the way  $p$ -values have been used across many disciplines. Their main arguments can be summarized as:

- Statistical significance (i.e., the  $p$ -value) is arbitrarily set and generally has no bearing on what we are really interested in, namely impact or magnitude of the effect of one or more variables on another.
- By focusing on a null hypothesis all that we are ever considering are existential questions, the answers to which are normally not interesting. So, for example, we might produce a very low  $p$ -value and conclude that road deaths and temperature are not unrelated. But the  $p$ -value tells us nothing about what we are really interested in, namely the nature and size of the relationship.
- Researchers sometimes wrongly assume that the  $p$ -value (which, remember, is the chance of observing the data if the null hypothesis is true) is equivalent to the chance that the null hypothesis is true given the data. So, for example, if they see a low  $p$ -value of say  $0.01$  they might conclude that there is a  $1$  in a  $100$  chance of no relationship (which is the same as a  $99\%$  chance that there is a relationship). This is, in fact, demonstrably false (we will show this in Chapter 5); it is an example of one of the most pernicious and fundamental fallacies of probability theory that permeates many walks of life (called the *fallacy of the transposed conditional*).
- In those many studies (notably medical trials) where the null hypothesis is one of 'no change' for some treatment or drug, the hypothesis comes down to determining whether the arithmetic mean of a set of data (from those individuals taking the treatment/drug) is equal to zero (supposedly representing status quo). In such cases, we have the paradox that, as we substantially increase the sample size, we will inevitably find that the mean of the sample, although approximately close to and converging to zero, will be significantly different from zero, even when the treatment genuinely has no effect. This is covered in Chapter 10 and is known as Meehl's conjecture.

- The choice of what constitutes a valid  $p$ -value is arbitrary. Is 0.04 radically different from 0.05? A treatment or putative improvement that yields a  $p$ -value that just misses the 0.05 target may be completely rejected and one that meets the target may be adopted.

Ziliak and McCloskey cite hundreds of examples of studies (all published in highly respected scientific journals) that contain flawed analyses or conclusions arising from the aforementioned misunderstandings. They give the following powerful hypothetical example of a fundamental weakness of using  $p$ -values:

Suppose we are interested in new drugs for reducing weight in humans. Two candidate drugs (called *Precision* and *Oomph* respectively) are considered. Neither has shown any side effects and their cost is the same. For each drug we conduct a study to test the null hypothesis ‘taking the drug leads to no weight loss’. The results are:

- For drug *Precision* the mean weight loss is 5 lb and every one of the 100 subjects in the study loses between 4.5 lb and 5.5 lb.
- For drug *Oomph* the mean weight loss is 20 lb and every one of the 100 subjects in the study loses between 10 lb and 30 lb.

Since the objective of weight loss drugs is to lose as much weight as possible, any rational, intuitive review of these results would lead us to recommend drug *Oomph* over *Precision*. Yet the  $p$ -value test provides the opposite recommendation. For drug *Precision* the  $p$ -value is much lower (i.e. more significant) than the  $p$ -value for drug *Oomph*. This is because  $p$ -values inevitably ‘reward’ low variance more than magnitude of impact.

The inevitable temptation arising from such results is to infer causal links such as, in this case, higher temperatures cause more fatalities. Indeed, using the data alone and applying traditional statistical regression techniques to that data you will end up with a simple model like that shown in Figure 1.13. Here the equation  $N = 2.144 \times T + 243.55$  is the best linear fit for the data calculated using Excel’s regression analysis tool. Using this equation we can predict, for example (by simply substituting the temperature values), that at 15°F we might expect to see 275 crashes, while at 80°F we might expect to see 415 crashes, involving a fatality per month.

Such an analysis could lead to an apparently counterintuitive (and dangerous?) newspaper headline:

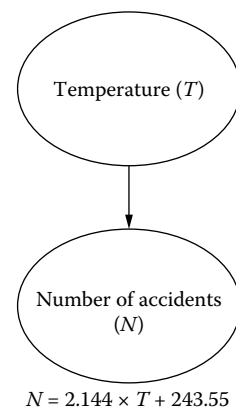
---

New research proves that driving in winter is actually safer than at any other time of the year.

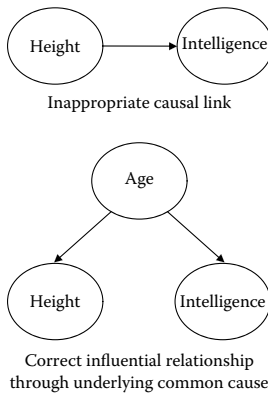
---

What is happening in this example is that there are other underlying factors (such as *number of journeys made* and *average speed*) that contribute to an explanation of the number of road fatalities on any given day (we will return to this example in Chapter 2 to motivate the need for more intelligent causal models for risk analysis).

There are many well-known examples of similar dubious correlations that expose the dangers and weaknesses of this standard statistical method. The folklore belief that babies are delivered by storks is strongly supported by analysis of real statistical data. In 2001 Matthews showed that the stork population and number of births in European cities were



**Figure 1.13** Simple regression model for automobile fatalities.



**Figure 1.14** Spurious relationship resulting from failure to consider underlying common factor.

correlated to a very high level of significance ( $p$ -value 0.008). But, of course, the correlation misses the explanation of a third common factor: population size. Obviously cities with larger populations have more births, but they also attract more storks.

Similarly, studies have shown that height and intelligence (as measured by an IQ test) of people are highly correlated. But, as illustrated in Figure 1.14, any attempt to explain intelligence causally by height misses the fact that the relationship is almost entirely due to a third factor, age; many people in the study were children between the ages of 4 to 16.

---

### Bayesian Networks

You can consider the diagrams in Figure 1.13 and Figure 1.14 as our first examples of Bayesian networks. Each node (i.e. bubble) represents some variable of interest and an arc between two nodes represents some kind of influential (or even causal) relationship between the corresponding variables.

---

## 1.5 Spurious Correlations: How You Can Always Find a Silly ‘Cause’ of Exam Success

Examples of purely coincidental (but strong) correlations:

- Level of beer drinking in the United States and child mortality in Japan.
  - Solar radiation and the London Stock Exchange index.
  - Sunspots and the lynx population density.
- 

Although the preceding examples illustrate the danger of reading too much into dubious correlations between variables, the relationships we saw there did not arise purely by chance. In each case some additional common factors helped explain the relationship.

But many studies (see the sidebar examples), including unfortunately many taken seriously, result in claims of causal relationships that are almost certainly due to nothing other than pure chance.

Although nobody would seriously take measures to stop Americans drinking beer in order to reduce Japanese child mortality, barely a day goes by when some decision maker or another somewhere in the world takes just as irrational a decision based on correlations that turn out to be just as spurious.

For example, on the day we first happened to be drafting this section (16 March 2009) the media was buzzing with the story that working night shifts resulted in an increased risk of breast cancer. This followed a World Health Organization study and it triggered the Danish government to make compensation awards to breast cancer sufferers who had worked night shifts. At the time of writing this book it is impossible to state categorically whether this result really is an example of a purely spurious correlation. But it is actually very simple to demonstrate why and how you will *inevitably* find a completely spurious correlation in such a study—which you might then wrongly claim is a causal relationship—if you measure enough things.

**Example 1.1: The Shotgun Fallacy**

Let us suppose that we are interested in possible ‘causes’ of student exam success. To make our example as simple as possible let us assume that exam scores are measured on a scale of 1 (worst) to 10 (best).

Now let us think of a number of possible ‘causes’ of exam success. These could include plausible factors like *coursework score* and *class attendance*. But we could also throw in some implausible factors like the *number of sexual partners*, *number of football matches attended*, or *number of potatoes eaten on 12 January*. In fact, to effectively demonstrate the point let us only consider a set of totally implausible factors. For simplicity we will assume that, like the exam score, they can all be measured on a scale of 1 to 10.

Now although these factors—suppose we think of 18—are completely silly, let’s actually remove any possibility that they are in any way valid factors by generating the results for them *purely randomly*. You can do this yourself. Create an Excel spreadsheet and type the entry =RANDBETWEEN(1,10) into cell A1. This will generate a random number between 1 and 10. By copying and pasting this entry create a set of random numbers like the set shown in Figure 1.15. There are 18 columns (A through R) that we can think of as being the 18 silly factors associated with the students. We have also added column S, which represents the student exam score, again generated randomly in the same way.

Now, using Excel’s built-in data analysis package, run a correlation analysis for each column (A through R) against the exam score (column S). If the correlation coefficient is higher than 0.561 then it is tempting to claim that the correlation is highly significant (the *p*-value is 0.01).

In fact, because of the number of factors we are trying here *it is very likely that you will find at least one column for which there is a significant correlation with S*. In Figure 1.15 the correlation coefficient of H and S is 0.59. Since column H is just as likely to represent *number of potatoes eaten on 12 January* as any other factor, would we be correct in concluding that eating potatoes on 12 January is the key to exam success?

In fact, because of the number of factors, it is also almost certain that among the 18 factors themselves you will also find at least two pairs that have a significant correlation. For example, in this case columns B and Q have a correlation coefficient of 0.62, which apparently might lead us to conclude that you can increase the number of football matches you attend by taking on more sexual partners.

What is clear from Example 1.1 is that if you measure enough different things about your subjects you will inevitably find at least one that is significantly correlated with the specific factor of interest. This may be the most likely explanation for night-shift work showing up as a significant ‘causal factor’ in breast cancer.

This should put you on your guard when you next hear about a recommendation for a lifestyle change that results from a statistical study. It should also make you extremely wary of correlation coefficients and *p*-values.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
10	9	5	3	10	4	7	4	3	9	5	4	5	9	4	4	10	5	5
10	10	4	2	3	10	6	1	8	5	8	8	8	7	6	3	8	3	1
5	7	9	9	3	2	5	2	6	6	7	8	2	9	10	3	8	2	1
9	3	6	10	3	1	5	8	2	9	5	8	7	4	8	8	2	7	7
2	6	1	1	10	8	8	5	8	7	10	4	7	9	7	4	3	7	3
10	3	6	7	1	10	9	9	6	2	8	5	8	3	9	9	2	2	7
1	1	1	7	5	1	4	9	1	6	9	8	9	9	4	1	2	7	5
3	5	8	4	2	4	6	2	7	9	5	2	2	5	4	3	2	1	1
1	8	8	10	6	4	10	7	6	6	5	7	3	7	10	7	4	9	8
4	4	8	8	3	1	1	9	1	9	10	9	10	2	8	1	3	4	10
9	3	5	3	3	2	4	4	3	10	4	9	8	7	3	10	2	8	4
2	3	1	1	6	7	10	5	5	1	4	4	3	10	9	5	7	1	6
10	9	1	3	10	6	7	7	8	1	9	4	3	7	3	3	10	3	7
4	2	5	10	9	9	2	4	9	8	9	7	5	7	6	6	1	7	2
1	7	3	5	5	8	8	10	2	10	7	10	2	10	4	8	5	2	8
9	8	8	1	4	2	8	7	10	1	6	8	1	1	9	6	4	1	2
4	4	3	2	4	7	5	3	1	3	5	10	2	5	2	6	7	8	9
4	7	3	10	5	10	7	3	6	6	6	5	10	8	1	6	2	7	5
10	9	6	5	9	7	4	8	10	2	8	6	3	5	9	9	6	3	2
7	5	5	3	5	4	8	4	3	4	4	2	1	9	6	4	7	6	9

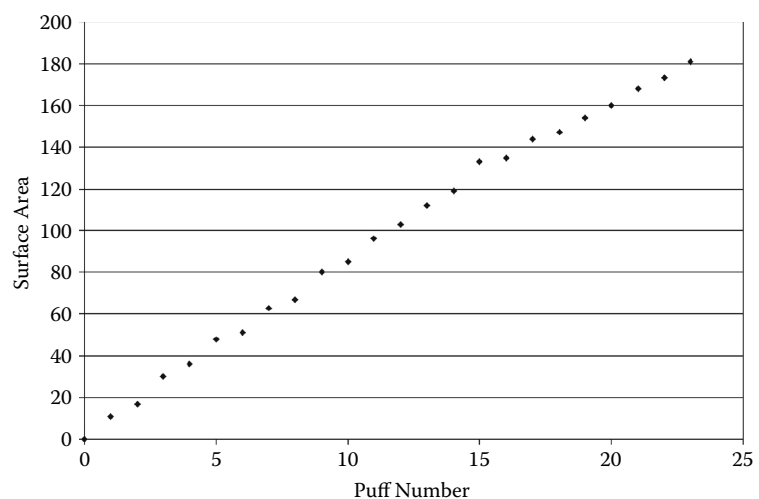
**Figure 1.15** Randomly generated numbers.

## 1.6 The Danger of Regression: Looking Back When You Need to Look Forward

Suppose that you are blowing up a large balloon. After each puff you measure the surface area and record it as shown in Figure 1.16. So, after the 23rd puff the surface is 181 sq cm. What will the surface area be on the 24th puff? On the 50th puff?

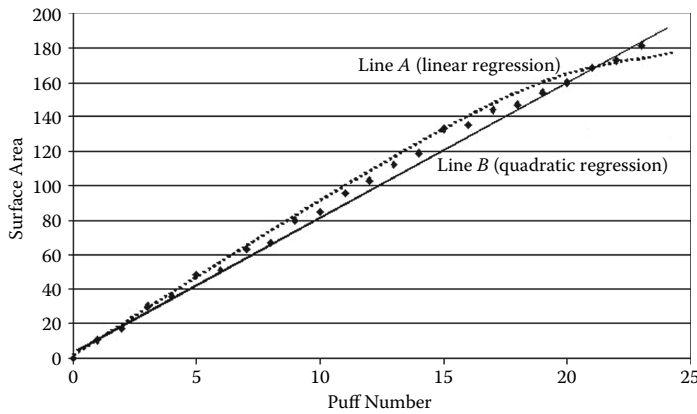
As opposed to the data on growth rates in Section 1.2, there is no doubt that the data here exhibits a clear trend. When presented with this kind of problem professionals often try to find lines that best fit the historical trend. As we saw in Section 1.4 this is an example of *regression analysis*. As in the road fatalities example there, the simplest (and most common) approach is to assume a simple straight line fit (called linear regression), producing a line such as line A shown in Figure 1.17. Alternatively, we might decide that the relative slow down of increase toward the end of the data is indicative of a *curve* such as line B (this is an example of *nonlinear regression*). The lines provide us with a method of predicting future values. The line A fit results in a prediction of 186 for the 24th puff, whereas the line B fit results in a prediction of 183.

It is also common for analysts to apply what are called *time-series* adjustments into their prediction to take account of the fact that there are local sequential differences; in this case the even-numbered puffs tend to result in lower increases than the odd-numbered puffs (for the simple reason that we blow harder on alternative puffs). Factoring in the time-series analysis results in an adjusted prediction of 184 for puff 24 in the linear regression and 182 in the quadratic regression. Predictions further ahead, such as at puff 30, are farther apart (235 for line A and 185 for line B).



**Figure 1.16** Increasing surface area of balloon.





**Figure 1.17** Lines of best fit for the data.

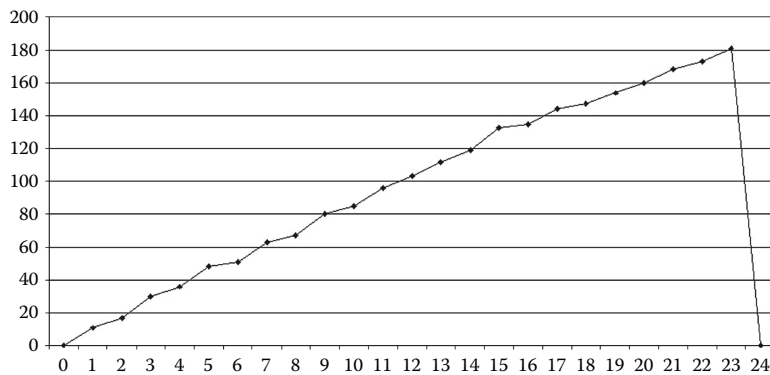
---

Unfortunately the balloon burst after 24 puffs (Figure 1.18).  
*Neither model was able to predict this.*

---

As we saw in Section 1.2 it was for reasons quite similar to this that the traditional statistical models were unable to predict the collapse of the banking sector in 2008 that ushered in a major worldwide recession. Although the models can incorporate millions of historical data to produce highly complex—and accurate—predictions over the short term during periods of growth, they were predicated on a set of barely articulated overoptimistic assumptions. The most basic knowledge about balloons would have indicated that a complete burst was inevitable, but traditional statistical models cannot incorporate this knowledge. Failure to do so is an example of what is commonly called the *fallacy of induction*. A similar example is highlighted in the sidebar (Does the data tell the full story?).

Whereas methods that rely purely on past data cannot predict these kinds of events, the methods advocated in this book at least provide



**Figure 1.18** Balloon bursts on puff 24.

**Does the Data Tell the Full Story?**

Suppose a government collects data on terrorist attacks on its territory as shown in Figure 1.19. Traditional statistical modelling predicts that in year 17 the number of attacks will be 15% fewer than in year 16. This makes the threat sufficiently low that a range of expensive security measures can now be lifted.

But what if the decreasing number of attacks is the result not just of a reduced threat but also of the increasingly sophisticated counterterrorist measures? The causal impact of these measures is not incorporated in the statistical model and is therefore wrongly ignored.

the possibility of predicting them by enabling us to incorporate expert judgement about assumptions, missing variables, the structure of causal relations, and our uncertainty about these.

**1.7 The Danger of Averages**

Fred and Jane study on the same course spread over two years. To complete the course they have to complete 10 modules. At the end, their average annual results are as shown in Table 1.7. So how is it possible that Jane got the prize for the student with the best grade? It is because the overall average figure is an average of the year averages rather than an average over all 10 modules. We cannot work out the average for the 10 modules unless we know how many modules each student takes in each year.

In fact:

- Fred took 7 modules in Year 1 and 3 modules in Year 2
- Jane took 2 modules in Year 1 and 8 modules in Year 2.

Assuming each module is marked out of 100, we can use this information to compute the total scores as shown in Table 1.8. So clearly Jane did better overall than Fred.

This is an example of *Simpson's paradox*. It seems like a paradox—Fred's average marks are consistently higher than Jane's average marks but Jane's overall average is higher. But it is not really a paradox. It is simply a mistake to assume that you can take an average of averages without (in this case) taking account of the number of modules that make up each average.

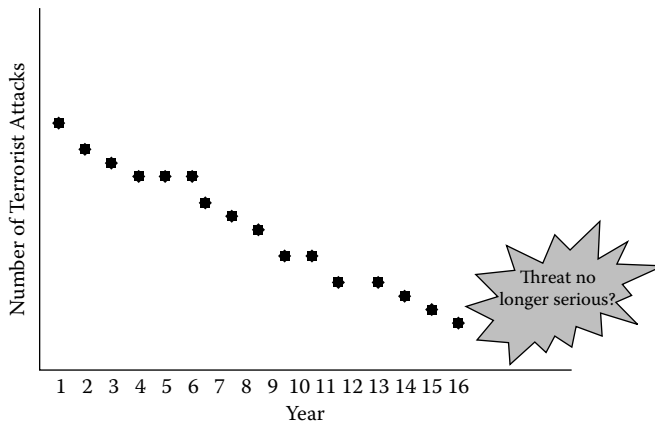
Look at it the following way and it all becomes clear: In the year when Fred did the bulk of his modules he averaged 50; in the year when Jane did the bulk of her modules she averaged 62. When you look at it that way it is not such a surprise that Jane did better overall.

**Table 1.7**  
Course Results

	Fred	Jane
Year 1 average	50	40
Year 2 average	70	62
Overall average	60	51

**Table 1.8**  
Revised Score Information

	Fred	Jane
Year 1 total	350 (7 × 50)	80 (2 × 40)
Year 2 total	210 (3 × 70)	496 (8 × 62)
Overall total	560	576
Real overall average	56	57.6



**Figure 1.19** Charting attacks over time.

We will look at a more profound and troubling example of Simpson's paradox later in the chapter. In fact, such examples provide a very convincing motivation for why causal models (implemented by Bayesian networks) are crucial for rational decision making. But first we have to address more fundamental concerns about the use of averages.

### 1.7.1 What Type of Average?

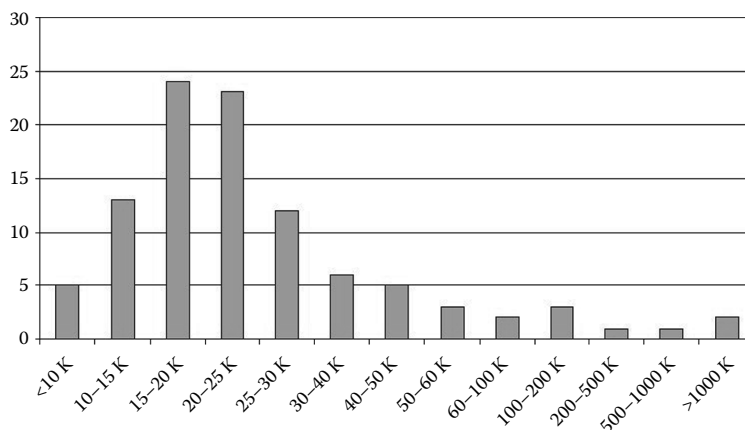
In the previous example, when we used the average for the exam marks data we were actually using one particular (most commonly used) measure of average: the *mean*. This is defined as the sum of all the data point values divided by the number of data points.

But it is not the only measure of average. Another important measure of average is the *median*. If you put all the data point values in order from lowest to highest then the median is the value directly in the middle, that is, it is the value for which half the data points lie below and half lie above.

Since critical decisions are often made based on knowledge only of the average of some key value, it is important to understand the extent to which the mean and median can differ for the same data. Take a look at Figure 1.20. This shows the percentage distribution of salaries (in \$) for workers in one city.

Note that the vast majority of the population (83%) have salaries within a fairly narrow range (\$10,000–\$50,000). But 2% have salaries in excess of \$1 million. The effect of this asymmetry in the distribution is that the *median* salary is \$23,000, whereas the mean is \$137,000. By definition half of the population earn at least the median salary; but just 5% of the population earn at least the mean salary.

Of course, the explanation for this massive difference is the 'long tail' of the distribution. A small number of very high earners massively skew the mean figure. Nevertheless, for readers brought up on the notion that most data is inherently bell-shaped (i.e. a Normal distribution in the sense explained in Section 1.2) this difference between the mean and median will come as a surprise. In Normal distributions the mean and median are



**Figure 1.20** Percentage distribution of salaries for a large group of workers.

always equal, and in those cases you do not therefore need to worry about how you measure average.

The ramifications in decision making of failing to understand the difference between different measures of average can be devastating.

#### **Example 1.2: Using the Mean When You Really Need the Median**

Suppose you are the mayor of the city mentioned earlier. To address the problem of unequal wealth distribution you decide to introduce a modest redistribution package. Every worker earning above 'average' salary will pay a tax of \$100 while every worker earning below average will receive an extra \$100. You feel that this will prove not only popular, but crucially will be tax neutral overall; it will not cost the city a penny. Unfortunately, by basing your calculations on the mean (\$137,000) rather than the median (\$23,000) just 5,000 workers pay the extra \$100 in tax, while 95,000 benefit from the extra \$100. The move certainly proves popular but it bankrupts the city since you have to find \$9 million of extra cash.

#### **Example 1.3: Using the Median When You Really Need the Mean**

Again suppose you are the mayor of the aforementioned city. This time you have to raise \$100 million from taxpayers to fund a major new transport project. It is agreed that all workers will contribute a fixed proportion of their salary to pay for the project. What should the fixed percentage be? Stung by your unfortunate experience at wealth redistribution, this time you base your calculation on an 'average' salary of \$23,000. You work out that the necessary new 'tax' is a whopping 4.3% for each of the 100,000 workers; this is because you believe that an 'average' salary of \$23,000 yields \$1,000 and multiplying this by the total number of workers gets you to the magical \$100 million. But this would only make sense if the mean salary was \$23,000. In fact, because the mean salary is \$137,000 the tax of 4.3% actually yields close to \$600 million. The city makes an incredible profit, but unfortunately you are voted out of office because it is rightly perceived as an unnecessarily harsh tax. Basing your calculations on the mean salary of \$137,000 requires a far more modest (and politically acceptable) 0.73% rate to make the target.

### **1.7.2 When Averages Alone Will Never Be Sufficient for Decision Making**

Whereas Simpson's paradox and skewed distributions alert us to the need to be very careful in how we use averages, there are some fundamental reasons why, in many cases of critical decision making and risk analysis, averages should be avoided altogether.

If you were going on holiday to a particular location in July, then knowing that the average July temperature there (however you measure it) is 27°C does not provide you with sufficient information to know what clothes to pack; your decision would be very different if the temperature *range* was 10°C to 40°C compared to a range of 22°C to 29°C. Similarly, if you were a poor swimmer, it is doubtful that you would be willing to wade across a river if you were told that the average depth was 5 feet, even if you were 6 feet tall.

Some decision makers avoid these problems by insisting that they have what is called a *three-point estimate* for each key value.

A three-point estimate is simply three numbers:

{lowest possible value, average value, highest possible value}.

---

So, in the temperature example above the three-point estimate might be {10, 27, 40}.

Such three-point estimates are very widely used by decision makers in critical applications (the military is especially keen on the approach). Unfortunately, although the three-point estimate seems an attractively simple way to describe the range for a value, it will generally be insufficient for rational decision making.

To see why, consider again the salary distribution in Figure 1.20. What is the three-point estimate here? We have already seen that the ‘average value’ will be very different depending on whether we use the mean or median, and we clearly need at least both as the examples showed. But there are also serious problems with the lowest and highest possible values. In this case the lowest possible value is something very close to zero since there will be at least a small number of workers earning almost nothing. At the other end there is almost certainly at least one person earning over \$20 million, so this will be the highest possible value. Neither the three-point estimate {0, 23,000, 20,000,000} nor {0, 137,000, 20,000,000} is sufficiently informative even to help us solve the problems in Examples 1.2 and 1.3. When confronted with this issue, proponents of three-point estimates will often propose that the lowest and highest values are replaced with what are called *percentiles*, typically 10% and 90% where the  $n\%$  percentile is the value for which  $n\%$  of the data items lie below. In the salary case this is more informative but still insufficient, whereas in the holiday example (where we were interested in temperature) it obscures the information we really need. To solve this we end up having to add additional percentiles, giving us not a three-point estimate but a five-, seven-, or nine-point estimate. But no matter what we chose we can always find examples where the number of points may be insufficient.

Fortunately, there is a simple way out. We can just use the full distribution such as that shown in Figure 1.20. When decision makers use either an average or a 3-point estimate what they are trying to do is characterize the whole distribution in as simple way as possible. But truly rational decision making often requires us to consider the full distribution, rather than a crude simplification of it. It turns out that such a distribution is precisely what probability theory and Bayesian networks provide us with for all values of interest. And in some cases there may even be a very small number of values (called parameters) that enable you to determine the whole distribution (this is something we will explain properly in Chapter 4).

---

The Normal distribution is an example where just two parameter values—the mean and the variance—determine the whole distribution.

---

## 1.8 When Simpson’s Paradox Becomes More Worrisome

Consider the following more troubling example of Simpson’s paradox (based on one from Pearl, 2000):

A new drug is being tested on a group of 80 people (40 men and 40 women) with a particular ailment. We wish to establish whether there is a link between taking the drug and recovery from an ailment. As is

**Table 1.9**  
**Drug Trial Results**

Drug Taken	No	Yes
Recovered		
No	24	20
Yes	16	20

**Table 1.10**  
**Drug Trial Results with Sex of Patient Included**

Sex	Female		Male	
	No	Yes	No	Yes
Recovered				
No	21	8	3	12
Yes	9	2	7	18

standard with drug trials half of the people (randomly selected) are given the drug and the other half are given a placebo. The results (Table 1.9) show that of the 40 given the drug, 20 (i.e. 50%) recover from the ailment; this compares favorably with just 16 out of the 40 (i.e. 40%) given the placebo who recover.

These results lead to the conclusion that the drug has a positive effect. But a more detailed look at the data results in exactly the opposite conclusion. Specifically, Table 1.10 shows the results when broken down into male and female subjects. Focusing first on the men we find that 70% (7 out of 10) taking the placebo recover, but only 60% (18 out of 30) taking the drug recover. *So, for men, the recovery rate is better without the drug.* With the women we find that 30% (9 out of 30) taking the placebo recover, but only 20% (2 out of 10) taking the drug recover. *So, for women, the recovery rate is also better without the drug.*

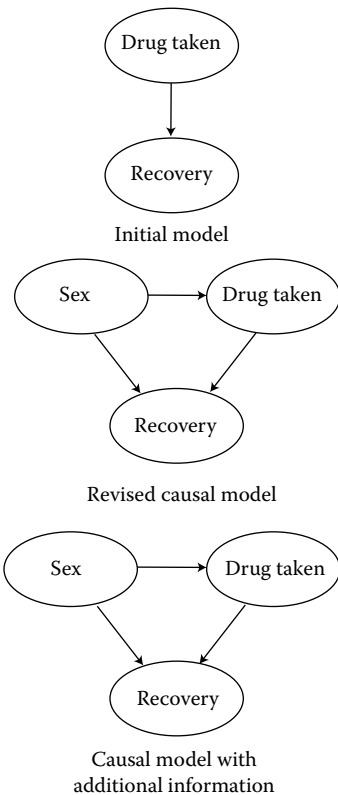
The process of drilling into the data this way (in this case by looking at men and women separately) is called *stratification*. Simpson's paradox is simply the observation that, on the same data, stratified versions of the data can produce the opposite result to nonstratified versions. Often, there is a *causal* explanation. In this case men are much more likely to recover naturally from this ailment than women. Although an equal number of subjects overall were given the drug as were given the placebo, and although there were an equal number of men and women overall in the trial, the drug was not equally distributed between men and women. More men than women were given the drug. Because of the men's higher natural recovery rate, overall more people in the trial recovered when given the drug than when given the placebo.

The difference between the types of data analysis is captured graphically in Figure 1.21. In the initial model we only have information about whether the drug is taken to help us determine whether a subject recovers. The revised causal model tells us that we need information about the subject's sex in addition to whether they take the drug to help us better determine whether the subject recovers. The final model introduces the further dependence, which is relevant for this particular case study namely that sex influences drug taken because men are much more likely in this study to have been given the drug than women.

What we have in Figure 1.21 are three more examples of Bayesian networks. In this case we know not just the graphical structure of the network, but also the underlying 'statistical' content. For the initial model Table 1.9 provides us with the necessary information about the outcome of 'recovery' (yes or no) given the information about 'drug taken' (yes or no).

For the revised causal model Table 1.10 provides us with the necessary information about the outcome of 'recovery' (yes or no) given the different combinations of information about sex ('male' or 'female') and 'drug taken' (yes or no). It also provides us with the necessary information about the outcome of drug taken given sex.

You need to get used to these kinds of tables because they, together with the graphical model, are exactly what you have to specify to complete a Bayesian network. You will learn how to do that, but we leave that until



**Figure 1.21** Explaining Simpson's paradox using a causal model.

Chapter 6 when we will return to these same models and show how they easily explain and overcome Simpson's paradox.

## 1.9 Uncertain Information and Incomplete Information: Do Not Assume They Are Different

Consider the following assertions:

1. Oliver Cromwell spoke more than 3,000 words on 23 April 1654.
2. O.J. Simpson murdered his wife.
3. You (the reader) have an as-yet undiagnosed form of cancer.
4. England will win the World Cup in 2022.

The events in assertions 1 and 2 either happened or did not. Nobody currently knows whether the assertion in statement 1 happened. O.J. Simpson (and possibly even one or two others) knows for certain whether assertion 2 happened. Assertion 3 describes a fact that is either true or false. Assertion 4 is different because it describes the outcome of an event that has not yet happened.

What all four assertions have in common is that our knowledge about them is uncertain (unless we happen to be O.J. Simpson). In this book the way we reason about such uncertainty is the same whether the events have happened or not and whether they are unknown or not. Unfortunately, many influential people do not accept the validity of this approach. We have an obligation to demonstrate why those influential people are wrong. To do this we will consider the simple scenario in Box 1.4 that captures the key differences between uncertain information and incomplete information.

There have been many well-known cases where Simpson's paradox has clouded rational judgement and decision making. Many of these cases are in medicine, but the most famous occurred at Berkeley University, which was (wrongly) accused of sex discrimination on the grounds that its admissions process was biased against women. Overall the data revealed a higher rate of admissions for men, but no such bias was evident for any individual department. The overall bias was explained by the fact that more women than men applied to the more popular departments (i.e. those with a high rejection rate).

### Box 1.4 Uncertain versus Incomplete Information

Suppose you ask your friend Naomi to roll a die without letting you see the result, but before she rolls it you have to answer the following:

Question 1: Will the number rolled be a 3?

Having rolled the die Naomi must write down the result on a piece of paper (without showing you) and place it in an envelope.

Now answer:

Question 2: Is the number written down a 3 (i.e. was the number rolled a 3)?

Most people would be happy to answer Question 1 with something like the following (which, as we will see in Chapter 2 is an example of a *probabilistic* statement): There is a one in six chance of it being 3. Yet,

there are many people who are convinced that such a probabilistic statement is meaningless for Question 2. Their reasoning is as follows:

1. There is no uncertainty about the number because it is a ‘fact’—it is even written down (and is known to Naomi).
2. The number either is a 3 (in which case there is 100% chance it is a 3) or it is not a 3 (in which case there is 0% chance it is a 3).

So some people are happy to accept that there is genuine uncertainty about the number *before* it is thrown (because its existence is ‘not a fact’), but not *after* it is thrown. This is despite the fact that our knowledge of the number after it is thrown is as *incomplete* as it was before.

Nowhere is this type of distinction more engrained than in the law: A defendant stands trial for a crime that, of course, *has already been committed*. Because the crime has already been committed the defendant either is or is not guilty of that crime.

The scenario in which people are convicted on the basis of crimes that they are *predicted* to commit remains the domain of pure science fiction like the Hollywood movie *Minority Report*.

---

In most cases the only person who knows for certain whether the defendant is guilty is the defendant. However, the defendant is not the one who has to determine guilt. Although the law implicitly endorses probabilistic reasoning when it talks about ‘balance of probabilities’ and ‘beyond reasonable doubt’ it abhors any explicit probabilistic reasoning about innocence and guilt in court based on the same irrational argument as above. As an eminent lawyer told us:

Look, the guy either did it or he didn’t do it. If he did then he is 100% guilty and if he didn’t then he is 0% guilty; so giving the chances of guilt as a probability somewhere in between makes no sense and has no place in the law.

What is curious about the rejection of the probabilistic answer to Question 2, is that we can *prove* that this rejection leads to irrational decision making as follows:

---

Suppose that you ask 60 people to each bet \$1 on the number written down by Naomi (and you can assume Naomi is not one of the 60 betting). You have to set the odds and must choose one of the following options:

- Option A—If they choose the correct number you pay them \$4 plus their \$1 stake. Otherwise you win their \$1 stake.
- Option B—If they choose the correct number you pay them \$5 plus their \$1 stake. Otherwise you win their \$1 stake.
- Option C—If they choose the correct number you pay them \$6 plus their \$1 stake. Otherwise you win their \$1 stake.

The twist to the scenario is that *your life depends on getting as close as possible to avoiding a loss*. In that case,



whatever your views about the uncertainty or otherwise of the number being 3, you will surely do the kind of calculations shown in the sidebar to choose Option B rather than Option A or Option C. But that means that you accept that the chances of the number being a 3 must be closer to 1 in 6 than to either 1 in 5 or 1 in 7. And if you accept this then it is irrational to reject a probabilistic answer to Question 2. Moreover, by accepting the validity of the statement ‘There is a one in six chance of it being 3’ you have just almost certainly saved your life. Not accepting this statement means you will probably die (actually there is a 2 in 3 chance you will die because you should have no preference between any of the three options).

What really lies at the heart of people’s concerns about using probabilities to describe incomplete information is that people with different levels of knowledge about the information will have different probabilities. So, whereas we should accept as reasonable the probability of a one in six chance of Naomi’s number being a 3, Naomi really does have reason to reject it because she knows the chance is 100% (if it is a 3) and 0% if it is not. If Naomi told her friend Hannah that the number written down is an odd number, then Hannah’s personal probability for the number being a 3 should be a one in three chance (because 3 of the possible numbers are odd).

By the same argument, for the defendant in court there is no uncertainty about guilt. But that does not remove the obligation from the jury to make a probabilistic assessment of guilt based on the incomplete information made available to them.

What is clear from the preceding discussion is that:

- If our knowledge of an event that has already happened is incomplete then we need to reason about it in the same way as we reason about uncertain events yet to happen. The failure to recognize that uncertainty and incomplete knowledge have to be handled in the same way leads to irrational decision making in some of the most critical situations.
- Different people will generally have different information about the same event (and this applies both to past and future events). Because of this people will generally have their own personal probability assessment of the event. In economics this difference in knowledge is often called ‘information asymmetry’: for example when buying a used automobile you may not know whether it is a ‘peach’ or a ‘lemon’, but the salesman selling the car has no such uncertainty and his probabilities will be very different from that of his customers.

This notion of ‘personal probabilities’ is central to the Bayesian reasoning in this book. It is a property of the mind and not of the object, hence the contrasting use of the labels ‘subjective’ and ‘objective’. We will explore it further when we define probability formally in Chapter 4.

### Calculating the Break-Even Odds for Guessing the Correct Die Number Written Down

Out of 60 people we can expect about 10 to choose the number 1, 10 to choose the number 2, 10 to choose the number 3, and so on. Of course, in practice these actual numbers will vary (as we saw in Section 1.3), but as this is the most likely outcome, it would be irrational to make any other assumption.

So, whatever number is written down we can expect about 10 people to win and 50 people to lose. Using Option A this results in us taking \$50 from the losers and paying out \$40 to the winners. So we expect to win \$10 overall. Using Option C this results in us taking \$50 from the losers and paying out \$60 to the winners. So we expect to lose \$10 overall. Only using Option B do we expect to break even (taking \$50 from the losers and paying out \$50 to the winners). Of course a bookie, if he wanted to stay in business, would offer Option A.

Unfortunately, as the next section demonstrates, it turns out that most people are really incapable of calculating probabilities properly.

## 1.10 Do Not Trust Anybody (Even Experts) to Properly Reason about Probabilities

Try answering the puzzle in Box 1.5.

### Box 1.5 Birthdays Puzzle

In a class of 23 children the chances that at least two children share the same birthday is:

- Approximately 1 in 16
- Approximately 1 in 10
- Approximately 1 in 5
- Approximately 1 in 3
- Approximately 1 in 2



If you have not already seen the puzzle then you may be surprised to know that the ‘correct’ answer here is *e* (in fact, the chances are slightly better than 1 in 2). You don’t need to know why at this point; a proper explanation will be provided in Chapter 4.

But if you answered *a*—on the instinctive basis that it is the closest to 23/365—then although you are completely wrong you are at least in good company. It is easily the most common answer. People are similarly stumped by the classic Monty Hall problem described in Box 1.6.

### Box 1.6 The Monty Hall Problem

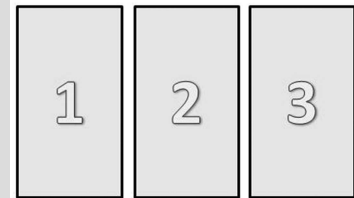
*Let’s Make a Deal*, a classic American ’60s game show, hosted by Monty Hall, involved contestants choosing one of three doors. Behind one of the doors was a valuable prize such as a new car. Behind the other two doors was something relatively worthless like a banana.

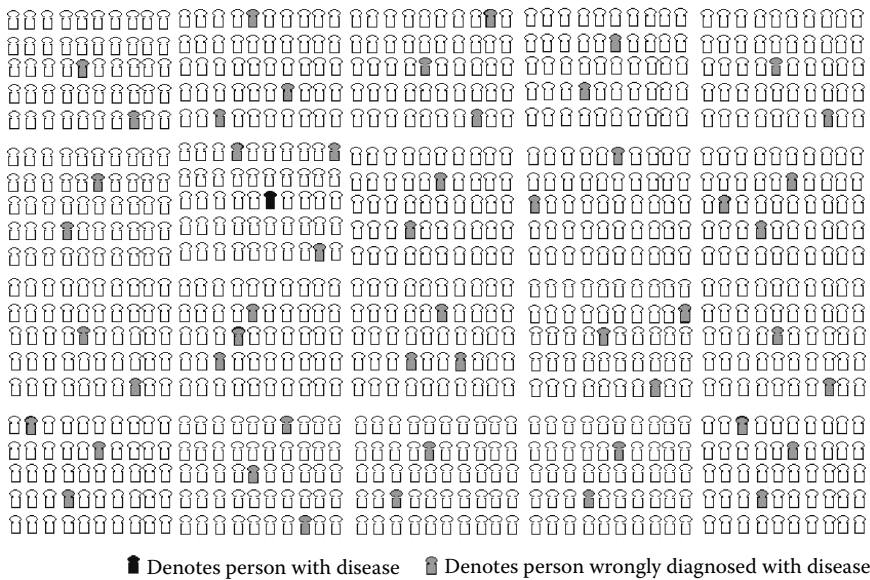
After the contestant chooses one of the three doors Monty Hall (who knows which door has the prize behind it) always reveals a door (other than the one chosen) that has a worthless item behind it. He now poses the question to the contestant:

“Do you want to switch doors or stick to your original choice?”

Most people assume that there is no benefit in switching; they feel that by sticking to their original choice they have a 50% chance of winning, the same as if they switch.

In fact they are wrong. It turns out that, by switching, you have a 2 in 3 chance (i.e., 66.66%) of winning. We will give a simple explanation why once we have formally defined probability in Chapter 4, and will also describe a Bayesian network solution in Chapter 6.





**Figure 1.22** In 1,000 random people about 1 has the disease but about 50 more are wrongly diagnosed as having the disease.

Not knowing the probability that children share the same birthday (or even the probability of winning the Monty Hall game) is hardly going to affect your life. But these problems are strikingly similar to many problems that can and do affect lives. In fact, even highly intelligent people, like world-leading barristers, scientists, surgeons, and businessmen misunderstand probability and risk, as the following examples indicate.

**Example 1.4: The Harvard Medical School Question**

In a classic and much referenced study by Casscells and colleagues the following question was put to 60 students and staff at Harvard Medical School:

One in a thousand people has a prevalence for a particular heart disease. There is a test to detect this disease. The test is 100% accurate for people who have the disease and is 95% accurate for those who don't (this means that 5% of people who do not have the disease will be wrongly diagnosed as having it). If a randomly selected person tests positive what is the probability that the person actually has the disease?

Almost half gave the response 95%. The 'average' answer was 56%. In fact, as we will explain formally, in Chapter 5, the correct answer is just below 2%. Figure 1.22 provides an informal visual explanation.

Think, for a moment, of the implications of this. If you test positive it is still extremely unlikely that you have the disease. But there are doctors who would believe you almost certainly have the disease and would proceed accordingly. This could result not just in unnecessary stress to you but even unnecessary surgery.

**Example 1.5: The Prosecutor’s Fallacy**

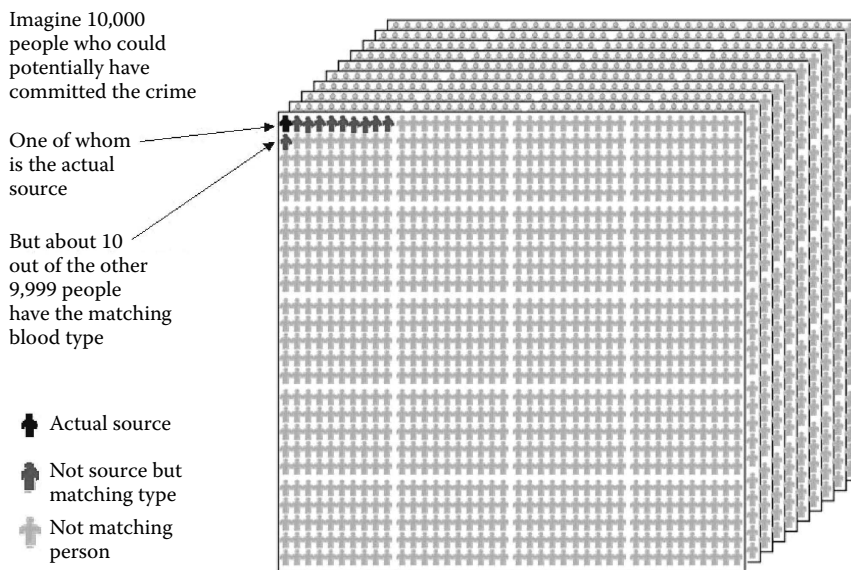
Suppose a crime has been committed and that the criminal has left some physical evidence, such as some of his blood at the scene. Suppose the blood type is such that only 1 in every 1,000 people has the matching type. A suspect, let’s call him Fred, who matches the blood type is put on trial. In court the prosecutor argues as follows:

The chances that an innocent person has the matching blood type is 1 in a 1,000. Fred has the matching blood type. Therefore the chances that Fred is innocent is just 1 in a 1,000.

When an eminent prosecutor makes a statement like this, backed by forensic evidence, it is clear that its influence on the jury could be profound. Yet, as we will explain in Chapter 3, the prosecutor’s conclusion generally massively understates the true probability that Fred is innocent. Figure 1.23 provides an informal visual explanation of this in the case where the number of potential suspects is 10,000 (in this case the chances that Fred is innocent is about 91%). And, as we will show in Chapters 5 and 13, mistakes exactly like this continue to be made by lawyers and forensic scientists in courtrooms throughout the world. The result is that the value of evidence is misunderstood, and juries are influenced to make poor decisions.

So, while you would expect that where critical decisions need to be made, the probabilities are calculated correctly. Unfortunately, they are usually not. In medical and legal situations lives are affected as a result. In business, companies can go bust and in many everyday financial cases the public’s general inability to understand probabilities is cunningly used against them.

And it is not just about relying on other people’s ability to calculate probabilities properly. Every day you make decisions that, consciously



**Figure 1.23** The potential source population.

or not, depend on probability assessments. Whether it is deciding which way to travel to work, deciding if it is worth taking out a particular insurance, deciding if you should proceed with a major project, or just improving your chances of winning at cards or on a sporting bet, the ability to do accurate probability calculations is the only way to ensure that you make the optimal decisions. One of the main challenges of this book is to help improve the way you do it.

## 1.11 Chapter Summary

The aim of this chapter was to introduce, by way of motivating examples, the key ideas of risk assessment and causal modelling that are the focus of the book. To appreciate why causal modelling (implemented by Bayesian networks) is such an effective method for risk assessment and decision analysis you first need to understand something about the traditional statistics and data analysis methods that have previously been used for this purpose. In introducing such methods, we have exposed a number of misconceptions and identified the most important limitations. In particular, we have demonstrated why these techniques provide little support for real practical risk assessment. To address these issues we now need to turn to causal models (Bayesian networks) and a different approach to probability than is typically used by statistical analysts.

Key points covered include:

- Uncertainty is a function of the lack of information and differences in certainty between individuals that reflect differences in personal experience and beliefs.
- Real risks do not often behave like lotteries.
- Beware spurious correlations purporting to reveal causal connections.
- Averages can be dangerous because they provide false security.
- Risks are not necessarily distributed normally nor are they symmetric.
- Beware of the fallacy of induction.

## Further Reading

- Adams, J. (1995). *Risk*, Routledge.
- Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). 'Sex bias in graduate admissions: Data from Berkeley.' *Science* 187, 398–404.
- Casscells, W., Schoenberger, A., and Graboyes, T. B. (1978). 'Interpretation by physicians of clinical laboratory results.' *New England Journal of Medicine* 299, 999–1001.
- Eastaway, R., and Wyndham J. (1998). *Why Do Buses Come in Threes? The Hidden Mathematics of Everyday Life*, John Wiley & Sons.
- Fenton, N. E. (2012). 'Making Sense of Probability: Fallacies, Myths and Puzzles.' [www.deecs.qmul.ac.uk/~norman/papers/probability\\_puzzles/Making\\_sense\\_of\\_probability.html](http://www.deecs.qmul.ac.uk/~norman/papers/probability_puzzles/Making_sense_of_probability.html).
- Gigerenzer, G. (2002). *Reckoning with Risk: Learning to Live with Uncertainty*, Penguin Books.
- Haigh, J. (2003). *Taking Chances: Winning with Probability*, Oxford University Press.
- Haldane, A. G. (2009). 'Why banks failed the stress test.' Marcus-Evans Conference on Stress Testing, 9-10 Feb 2009. London, Bank of England, [www.bankofengland.co.uk/publications/speeches/2009/speech374.pdf](http://www.bankofengland.co.uk/publications/speeches/2009/speech374.pdf).
- Hubbard, D. W. (2009). *The Failure of Risk Management: Why It's Broken and How to Fix It*, Wiley.
- Hubbard, D. W. (2010). *How to Measure Anything: Finding the Value of Intangibles in Business*, 2nd Edition, Wiley.
- Lewis, H. W. (1997). *Why Flip a Coin? The Art and Science of Good Decisions*, John Wiley & Sons.

- Matthews, R. (2001). 'Storks deliver babies ( $p = 0.008$ ):' *Teaching Statistics* 22(2), 36–38.
- Pearl, J. (2000). *Causality: Models, Reasoning and Inference*, Cambridge University Press.
- Piatelli-Palmarini, M. (1994). *Inevitable Illusions: How Mistakes of Reason Rule Our Minds*, John Wiley & Sons.
- Sowell, T. (1987). *A Conflict of Visions: Ideological Origins of Political Struggles*, William Morrow.
- Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable*, Random House.
- Ziliak, S. T., and McCloskey, D. N. (2008). *The Cult of Statistical Significance*, The University of Michigan Press.