1

Introduction

In December 2017, the Royal Statistical Society (RSS) announced the winner of its International Statistic of the Year. The citation[1] announced it as follows:

> **Winner: International Statistic of the Year**
>
> **69**
>
> **This is the annual number of Americans killed, on average, by lawnmowers—compared to two Americans killed annually, on average, by immigrant Jihadist terrorists.**
>
> The figure was highlighted in a viral tweet this year from Kim Kardashian in response to a migrant ban proposed by President Trump; it had originally appeared in a Richard Todd article for the *Huffington Post*.
>
> Todd's statistics and Kardashian's tweet successfully highlighted the huge disparity between (i) the number of Americans killed each year (on average) by "immigrant Islamic Jihadist terrorists" and (ii) the far higher average annual death tolls among those "struck by lightning," killed by "lawnmowers," and in particular "shot by other Americans."
>
> Todd and Kardashian's use of these figures shows how everyone can deploy statistical evidence to inform debate and highlight misunderstandings of risk in people's lives.
>
> Judging panel member Liberty Vittert said: "Everyone on the panel was particularly taken by this statistic and its insight into risk—a key concept in both statistics and everyday life. When you consider that this figure was put into the public domain by Kim Kardashian, it becomes even more powerful because it shows anyone, statistician or not, can use statistics to illustrate an important point and illuminate the bigger picture."

The original Kim Kardashian tweet is shown in Figure 1.1.

While the announcement was met with enormous enthusiasm, one significant dissenter was Nassim Nicolas Taleb—a well-known expert on risk and "randomness." He exposed a fundamental problem with the statistic, which he summed up in the tweet of Figure 1.2.

Indeed, rather than "inform debate and highlight misunderstandings of risk in people's lives," as stated by the RSS, this example does exactly the opposite. It provides a highly misleading view of risk because it omits crucial causal information that explains the statistics observed and that is very different for the two incomparable numbers. One of the

Contrary to the statement of the Royal Statistical Society citation, the figures directly comparing numbers killed by lawnmower with those killed by Jihadist terrorists do *not* "highlight misunderstandings of risk" or "illuminate the bigger picture." They do the exact opposite, as we explain in this book.

[1] https://www.statslife.org.uk/news/3675-statistic-of-the-year-2017-winners-announced. All models in this chapter are available to run in AgenaRisk, downloadable from www.agenarisk.com.

Because of the particular 10-year period chosen (2007–2017), the terrorist attack statistics do not include the almost 3,000 deaths on 9/11 and also a number of other attacks that were ultimately classified as terrorist attacks.

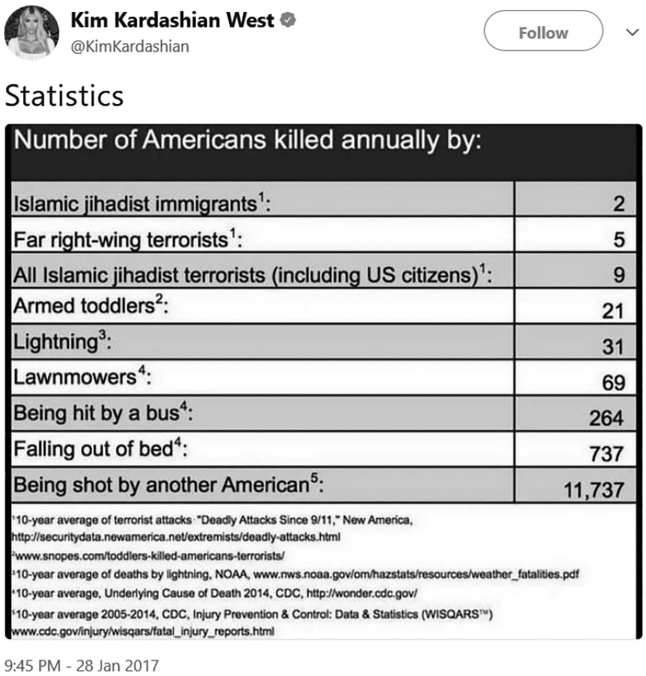


Figure 1.1 Tweet by Kim Kardashian that earned "International Statistic of the Year" 2017.

objectives of this book is to help readers understand how to see through such statistics and build models that incorporate the necessary causal context.

Informally, Taleb's argument is that there is a key difference between risks that are *systemic*—and so can affect more than one person (such as a terrorist attack)—and those that are not (such as using a lawnmower), which can be considered *random*. The chances that the number of people who die from a nonsystemic risk, like using a lawnmower, will double next year are extremely unlikely. But this cannot be said about the number of people dying from systemic risks like terrorist attacks and epidemics. The latter can be "multiplicative," whereas the former cannot. It is impossible for a thousand people in New York City to die from using lawnmowers next year, but it is not impossible for a thousand to die from

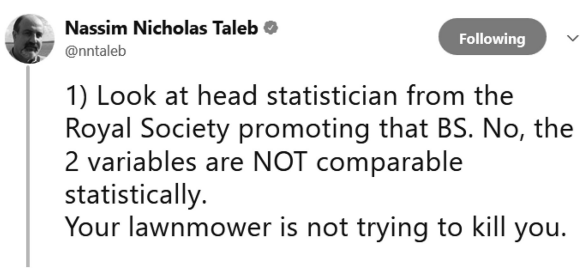
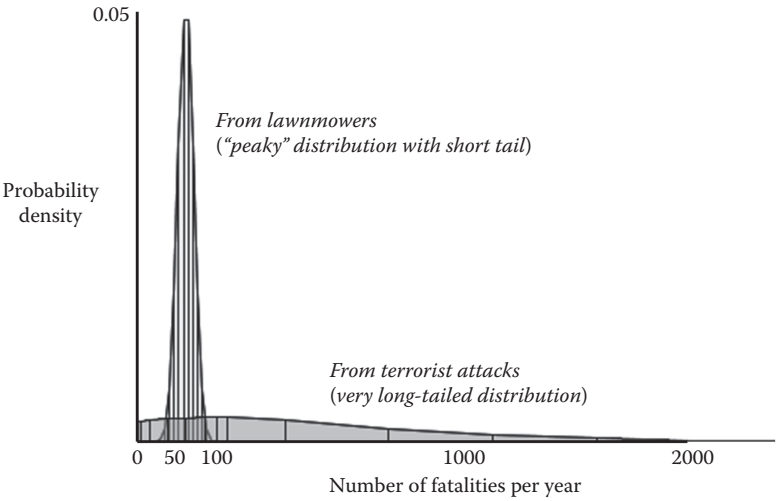


Figure 1.2 Taleb's response to the RSS announcement.



Systemic risks have long tails that capture low (but nonzero) probability events. Unlike the lawnmower deaths distribution, there is a small nonzero probability of having 2000 fatalities from terrorist attacks in a single year in the United States.

Figure 1.3 Comparing the probability distributions of number of fatalities per year.

terrorist attacks. Systemic and nonsystemic risks have very different *probability distributions*, as shown in Figure 1.3.

Using the number of deaths per year to compare different types of "risk" fails to consider the range of factors that affect the true risk to particular individuals or groups. For example, the probability of being killed by a lawnmower in New York City is especially low because relatively few people there have lawns to mow. In fact, death by lawnmower is essentially impossible for those not using a lawnmower, whereas there is a greater risk to gardeners. Residents of major cities are at greater risk from terrorists than residents of the countryside.

Crucially, there are also causal factors that *explain* the number of terrorist deaths that need to be considered. Most obviously, there are extensive security measures in place to stop terrorist attacks; without these, deaths from terrorist attacks would drastically increase. Also, terrorist cells can be responsible not just for multiple deaths in a single attack, but also multiple attacks, so deaths in terrorist attacks can be related by a *common cause*. These types of causal influences and relations—summarized in Figure 1.4—are the focus of much of this book.

An especially concerning part of the RSS citation was the implication that the relatively low number of terrorist deaths suggested that new measures to counter terrorism were unnecessary because of the "low risk." To make such reasoning explicit, we would have to perform a cost-benefit and trade-off analysis (Figure 1.5). Imposing new measures to counter terrorist threats involves both a financial cost and a human rights cost. But they also involve potential benefits, not just in terms of lives saved but also in reduction of other existing (secondary) security costs and improved quality of life. The implication from the RSS was that the costs were greater than the benefits.

Figure 1.5 is an example of an extended type of Bayesian network, called an influence diagram, which we discuss in Chapter 11.

But even if this trade-off analysis had been made explicit (which would involve putting actual numbers to all the costs and benefits as well

In the lawnmower case, Fred and Jane are killed by different lawnmowers. This is what makes them independent … in the absence of common lawnmower design flaws (such as a controller bug inserted by a terrorist designer).

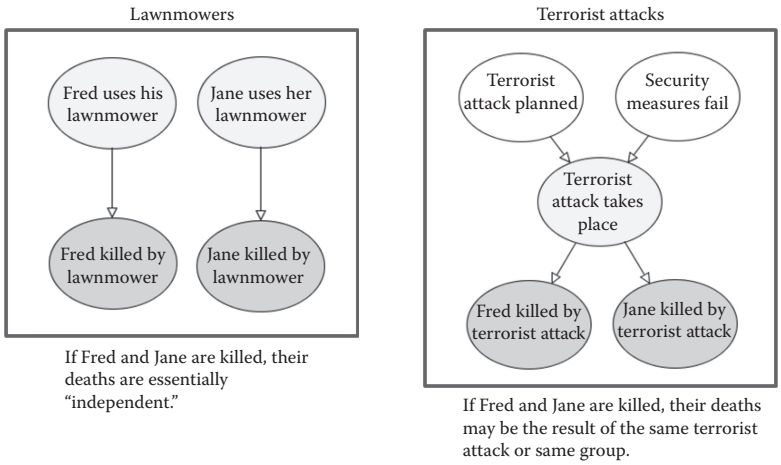


Figure 1.4 Causal view of lawnmower versus terrorist attack deaths.

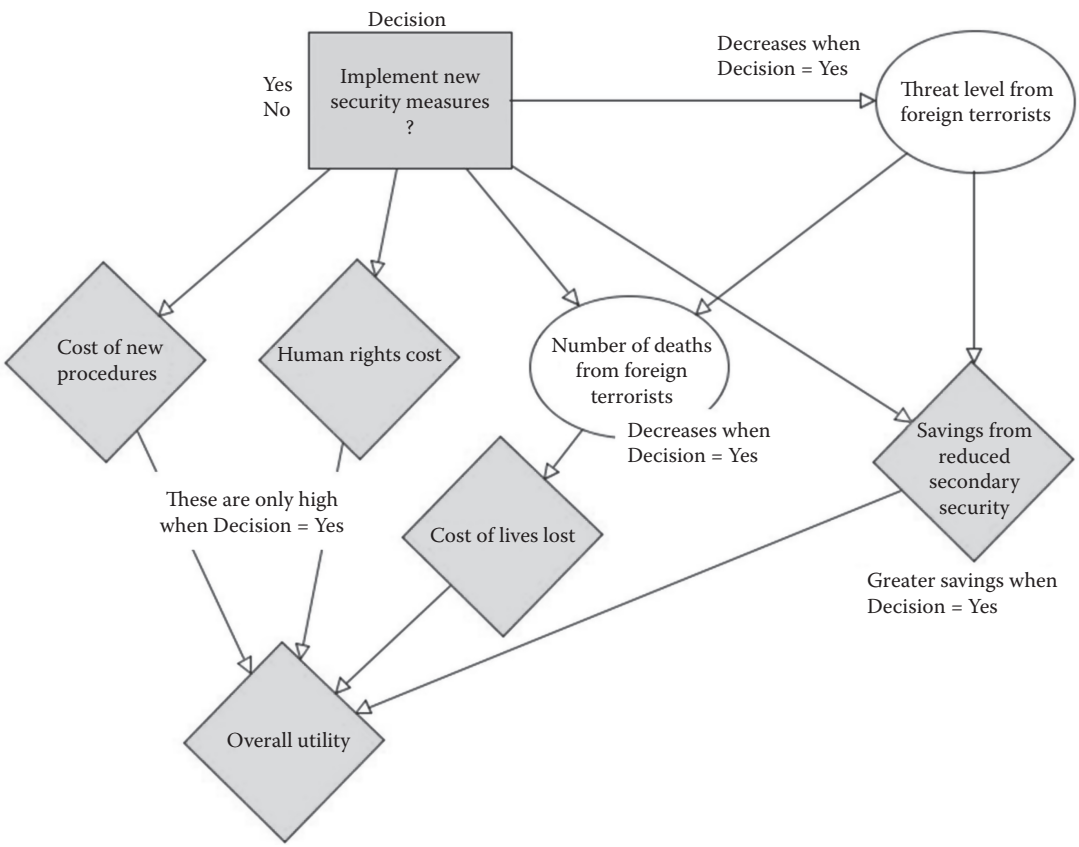


Figure 1.5 The kind of cost-benefit trade-off analysis required for informed decision-making.

as the number of expected deaths from jihadis who would otherwise not have entered the United States), there is a fundamental flaw in relying only on previous years' fatalities data. The number of fatalities depends (among other things) on the security measures that are put in place. The *Thames barrier* provides a very good analogous example:

> In several decades prior to 1974 the number of deaths in London due to the River Thames flooding was zero. Based on these data (and applying the RSS reasoning) what possible justification could there have been for the British Government to decide to build a flood barrier (which cost £650 million before its completion in May 1984) rather than do nothing? The decision on the Thames Barrier was made because steadily rising water levels were already causing expensive (but non-fatal) flooding and reliable measurements predicted catastrophic flooding within 50 years if no barrier was in place. In this case the simplistic counts of past number of fatalities were clearly insufficient for rational risk assessment and decision-making.

While the Thames Barrier decision still made use of historical data (namely monthly water levels, cost of flood damage, etc.), the key point is that we need to go beyond the simplistic data and consider contextual and situational factors. Moreover, in many risk scenarios, "triggers" and "threats" that are analogous to rising water levels in this example might require expert judgments and models in addition to data. Without an explanatory or causal model, the data alone would be meaningless from an inferential or decision-making perspective.

Completely novel risks (such as crashing civilian planes into skyscrapers prior to 9/11) can only be quantified using expert imagination and judgment. Indeed, the 9/11 scenario had previously been considered seriously by security experts (and movie scriptwriters), and terrorist "chatter" suggested the threat was increasing. However, the probability of such an event was considered sufficiently low not to merit additional security measures that could have been put in place to avert it. Had security measures—which are now routine at all airports in the world—been put in place before 9/11, there would have been no mass fatalities on 9/11.

Yet we find the same flawed "datacentric" reasoning being applied yet again. The context and motivation for the choice of the 2017 RSS International Statistic of the Year was the partial immigration ban proposed by President Trump in 2017. Opponents to the proposed measures argued that they were unnecessary because everyday risks from, for example, lawnmowers are greater than the risk from jihadis. This demonstrates again the problems of relying solely on historical fatality data.

One of our own areas of research—predicting system reliability, which is covered in Chapter 13—is especially prone to these kinds of misconceptions about oversimplistic past data. A "system" could be a software program, a physical device or component thereof (phone, TV, computer, microprocessor), or even a process (such as a method for manufacturing steel). It is standard to measure a system's reliability in terms of the frequency with which it fails—both before releasing the system for general use and also after release. But using this data alone is problematic. Why? For instance, consider a system where, after two years,

The methods described in this book enable us to build models that incorporate expert subjective judgment with data (when available) in order to provide fully quantified risk assessment in terms of probabilities of specific events and overall utilities.

there are very few or zero reports of failures. At first glance, this might suggest that the system is very reliable. But there is another possible explanation—the cause of the low number of failures may well be that the system was so bad that it was rarely or never used. So, here we have competing causal explanations that are very different but give rise to the same observable data.

In many areas of life, past data is a good indicator of future behavior and may be sufficient for good decision-making. Based on average temperatures in previous years, we can be pretty confident that if we are going to Cairo in June, we will not need a fur coat to keep warm. You don't even need the past data to be "constant." A company that has seen a steady year-on-year increase in sales of widgets can be confident of next year's sales based on simple regression models. The same is true in many industries. In both of these examples, we do not use the data alone. We use it with a (often implied) model to interpret and make inferences, either using other relevant circumstances connected to weather or customer demand. But as soon as there are novel circumstances and factors, this type of model for decision-making is likely to be poor.

While the above example of misuse of statistics for risk assessment might be considered harmless, the same cannot be said of the financial crisis of 2008–9, which brought misery to millions around the world. 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. Similarly, the results of major elections in 2016 (in the USA and the UK Brexit vote) were contrary to what pollsters were consistently and almost uniformly predicting. 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.

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.

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 judge or member of a jury, you hear many pieces of evidence in a trial. Some of the evidence favors the prosecution hypothesis of guilty, and some of the evidence favors the defense hypothesis of innocence. Some of the evidence is statistical (such as the match probability of a DNA trace found at a crime scene) 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 Societe Generale 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 book is to show how it is possible to do rigorous analysis of all of the above types of risk assessment problems using Bayesian networks.

Visit www.bayesianrisk.com for exercises and worked solutions relevant to this chapter.