Business Risk Management

Eddie Anderson, University of Sydney

1

What Is Risk Management?

The biggest fraud of all time

The biggest trading fraud of all time occurred at Société Générale ('SocGen') and was uncovered in January 2008. SocGen is one of the largest banks in Europe and the size of the fraud itself is staggering; SocGen estimated that it lost 4.9 billion Euros as a result of unwinding the positions that had been entered into. With a smaller firm this could well have caused the bank's collapse in the way that happened to Barings in 1995, but SocGen is large enough to weather the storm. The employee responsible was Jérôme Kerviel, who did not profit personally (or at least only through his bonus payments being increased). In effect he was taking enormous unauthorized gambles with his employer's money. For a while these gambles came off but in the end they went very badly wrong.

In America the news broke on January 24, 2008 when the New York Times reported as follows:

> "Société Générale, one of the largest banks in Europe, was thrown into turmoil Thursday after it revealed that a rogue employee had executed a series of "elaborate, fictitious transactions" that cost the company more than $7 billion US, the biggest loss ever recorded in the financial industry by a single trader.
>
> "Before the discovery of the fraud, Société Générale had been preparing to announce pretax profit for 2007 of €5.5 billion, a figure that Bouton (the Société Générale chairman) said would have shown the company's 'capacity to absorb a very grave crisis.' Instead, Bouton - who is forgoing his salary through June as a sign of taking responsibility - said the 'unprecedented' magnitude of the loss had prompted it to seek about €5.5 billion in new capital to shore up its finances, a move that secures the bank against collapse.
>
> "Société Générale said it had no indication whatsoever that the trader - who joined the company in 2000 and worked for several years in the bank's French risk-management office before being moved to its Delta One trading desk in Paris - "had taken massive fraudulent directional positions in 2007 and 2008 far beyond his limited authority." The bank added: "Aided by his in-depth knowledge of the control procedures resulting from his former employment in the middle-office, he managed to conceal these positions through a scheme of elaborate fictitious transactions."

Business Risk Management Edward Anderson

> "When the fraud was unveiled, Bouton said, it was "imperative that the enormous position that he had built, and hidden, be closed out as rapidly as possible." The timing could hardly have been worse. Société Générale was forced to begin unwinding the trades on Monday "under conditions of extreme market volatility," Bouton said, as global stock markets plunged amid mounting fears of an economic recession in the United States."

A story like this inevitably prompts the question: How could this have happened? The Appendix at the end of this chapter gives more details of what went wrong, drawn from newspaper reports. SocGen was a victim of an enormous fraud but the defence lawyers at Kerviel's trial argued that the company itself was primarily responsible. Whatever degree of blame is assigned to SocGen, it clearly paid a heavy price. It is easy to be wise after the event, but good business risk management calls on us to be wise beforehand. Later in this chapter I will set down what I believe are some of the key things that can be learnt from this episode (and that need to be applied in a much wider sphere than just the world of banks and traders.)

1.1 Introduction

In essence *Risk Management* is about being able to manage effectively in a risky and uncertain world. Banks and financial services companies have developed some of the key ideas in the area of risk management, but it is clearly vital for any manager. All of us, every day, operate in a world where the future is uncertain.

When we look out into the future there are a myriad of possibilities: there can be no comprehension of this in its totality. So our first step is to simplify in a way that enables us to make choices amidst all this uncertainty. This task of finding a way to simplify and comprehend what the future might hold is conceptually challenging. Individuals may do this in different ways. For example we may construct stories about what could happen. This could give us a range of possible future scenarios that are all believable, but have different likelihoods. One way to go about this is to think of chains of linked events: if one thing happens then another may follow. For example, if there is a typhoon in Hong Kong, then the shipment of raw materials is likely to be late, and if this happens then we will need to buy enough to deal with our immediate needs from a local supplier, and so on. This creates a "*causal chain*".

A causal chain may in reality be a more complicated network of linked events. But in any case it is often helpful to identify a particular risk event within the chain which may or may not occur. Then we can consider both the probability of the risk event occurring and also the consequences and costs if it does.

Risk management is about seeking better outcomes. To do this we may set about identifying different risk events and try to understand both their causes and consequences. Usually risk in this context refers to something that has a negative effect, so that our interest in the causes of negative risk events is to reduce their probability or, better still, eliminate them altogether. Our interest in the consequences of risk events is to act beforehand in a way that reduces the costs if a negative risk event does occur. The open ended nature of this exercise makes it important to concentrate on the most important causal pathways – we can think of this as identifying "*risk drivers*".

At the same time as looking at actions specifically designed to reduce risk we may need to think about the risk consequences of other kinds of decisions that we make. For example we may ask what extra risks are involved in moving to an overseas supplier who will deliver larger quantities of goods but with a greater lead time. In later chapters we will give much more attention to the problems of making good decisions in a risky environment.

Risk management involves planning and acting before the risk event. This is *proactive* rather than *reactive* management. We don't just wait and see what happens, with the hope that we can manage our way through the consequences; instead we work out in advance what might happen and what the consequences are likely to be. Then we plan what we should do to reduce the probability of the risk event and to deal with the consequences if it occurs.

Sometimes the risk event is not in our control; for example we might be dealing with changes in exchange rates or government regulation - usually this is called *external risk*. On other occasions we can exercise some control over the risk events, such as employee availability, supply and operations issues. These are called *internal risk*. The same distinction between what we can and cannot control occurs with consequences too. Sometimes we can take actions to limit negative consequences (like installing sprinklers for a fire), sometimes we cannot do so and we might choose to directly insure against the event (e.g. purchasing fire insurance)

We will use the term risk management to refer to the entire process:

- **Understanding Risk:** both its drivers and its consequences.
- **Risk Mitigation:** reducing or eliminating the probability of risk events as well as reducing the severity of their impact.
- **Risk Sharing:** the use of insurance or similar arrangement so that some of the risk is transferred to another party, or shared between two parties in some contractual arrangement.

The risk framework we are discussing makes it sound as though all risk is bad. But this is misleading in two ways – first we can use the same approach to consider good outcomes as well as bad ones. This would lead us to try to understand the most important causal chains with the aim of maximizing the probability of a positive chance event: or optimizing the benefits if it does occur. Second we need to recognize that sometimes the more risky course of action is ultimately the wiser one. Managers are schizophrenic about risk. Most see risk taking as part of a manager's role, but there is a tendency to judge whether a risk was good or bad simply by looking at the results. Though it is rarely put in these terms, the idea seems to be that it is fine to take risks provided that nothing actually goes badly wrong! Occasionally managers might talk of 'controlled risk' by which they mean a course of action in which there may be negative consequences but these are of small probability and the size of the cost is tolerable.

Rice and Franks (2010) in their discussion of the 'agile enterprise' say "While uncertainty impacts risk, it does not necessarily make business perilous. In fact, risk is critical to any business – for nothing can improve without change – and change requires risk." Much the same point was made by Prussian Marshall Helmuth von Moltke in the mid-1800s: "First weigh the considerations, then take the risks."

So far our discussion has implied an ability to list all the risks and discuss the probability that an individual risk event occurs. But often there is no way to identify all the possible outcomes or and it is meaningless to talk of the probability of their occurrence. Some people use 'uncertainty' (rather than risk) to refer to this idea. Frank Knight was an economist who was amongst the first to clearly distinguish between these two concepts and used risk to refer to situations where the probabilities involved are computable. In many real environments there may be a total absence of information about, or awareness of, some potentially significant event. Former US Defense Secretary Donald Rumsfeld in a much-parodied speech made at a press briefing on February 12, 2002, said:

> “There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we now know we don’t know. But there are also unknown unknowns. These are things we do not know we don’t know.”

In Chapter 8, in our discussion of robust optimization, we will return to the question of how we should behave in situations with uncertainty, where we have to make decisions without being ale to assign probabilities to different events.

1.2 Identifying and documenting risk

Many companies set up a formal risk register to document risks. This enables them to have a single point at which information is gathered together and it encourages a careful assessment of risk probabilities and likely responses to risk events.

A carefully documented risk management plan has a number of advantages. There is first of all a benefit in making it more likely that risk will be appropriately managed, with major risks identified and appropriate measures taken. Secondly there is an advantage in defining the responsibility for managing and responding to particular categories of risk. It is all too easy to find yourself in a company in which something goes wrong and no person or department admits to being the responsible party.

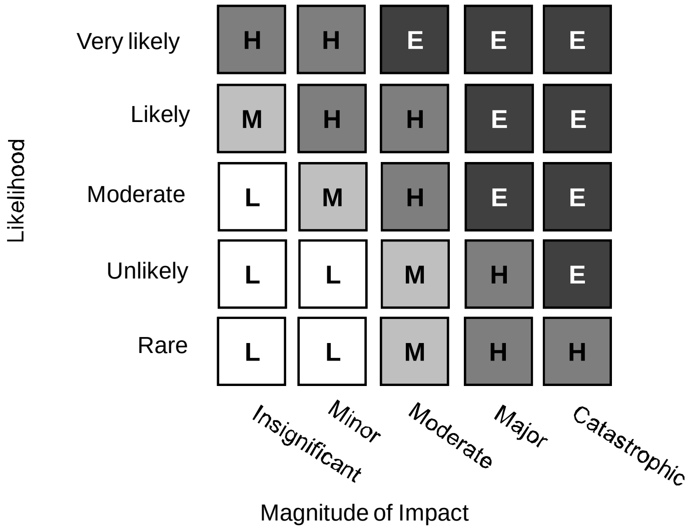
Moreover a risk management plan allows stakeholders to approve the risk management approach and helps to demonstrate that the company has exercised an appropriate level of diligence in the event that things do go wrong.

The first step is to identify the important possible risk events, and in doing this a systematic process for identifying risk can be helpful. It helps to start with the context for the activity: the objectives; the external influences; the stages that are gone through. And then go through each element of the activity asking:

1. What might happen that could cause external factors to change or that could effect the achievement of any objectives?
2. How are these events likely to occur?
3. How probable are these events?

The next step is to identify the consequences of these risk events. For each risk consider:

- Which stakeholders might be involved or affected? For example does it effect the return on share capital for shareholders? Does it effect the assurance on payment and the future for suppliers. Does it effect the security that is offered to Lenders (to us)? Does it effect on the assurance of future employment for our employees?



Likelihood	Insignificant	Minor	Moderate	Major	Catastrophic
Very likely	H	H	E	E	E
Likely	M	H	H	E	E
Moderate	L	M	H	E	E
Unlikely	L	L	M	H	E
Rare	L	L	M	H	H

Magnitude of Impact

Figure 1.1 Calculating risk level from likelihood and impact

- How damaging is this risk?
- What controls currently exist to make this risk less likely or less damaging?
- What might stop the controls from working?

At the end of this process we will be in a better position to build the *risk register*. This will indicate for each risk identified:

- Its causes and impacts;
- The likelihood of this risk event;
- The controls that exist to deal with this risk;
- An assessment of the consequences.

Because the risk register is likely to contain a great many different risks it is important to focus on the most important ones. We want to construct some sort of priority rating – giving the overall level of risk. This then provides a tool so that management can focus on the most important risk events and then determine a risk treatment plan to reduce the level of risk. The most important risks are those with serious consequences that are relatively likely to occur. We need to combine the likelihood and the impact and Figure 1 shows the type of diagram that is often used to do this, with risk levels labelled L = Low; M = Medium; H = High; E = Extreme.

This type of diagram of risk levels is sometimes called a ‘heat map’ and often red is used for the extreme risk boxes, orange for the high risks and yellow for the medium risks. It is a common tool and is recommended in most risk management standards. It should be seen as a preliminary to making a much fuller investigation of some specific risks. However there are some significant difficulties associated with the use of this approach.

One problem is related to the use of a scale based on words like ‘likely’ or ‘rare’: these terms will mean very different things to different people. Some people will use a term like ‘likely’ to mean a more than two thirds chance of occurring (this is the specific meaning that is ascribed in the IPCC climate change report). But in a risk management context quite small probabilities over the course of a year may seem to merit the phrase “likely”. The use of vague terms in a scale of this sort will make misunderstandings far more likely. Douglas Hubbard describes an occasion when he asked a manager “What does this mean when you say this risk is ‘very likely’?” and was told that it meant there was about a 20% chance of it happening. Someone else in the room was surprised by the small probability, but the first manager responded “Well this is a very high impact event and 20% is too likely for that kind of impact.” Hubbard describes the situation as “a roomful of people who looked at each other as if they were just realizing that, after several tedious workshops of evaluating risks, they had been speaking different languages all along.” This story illustrates how important it is to be absolutely clear about what is meant when discussing probabilities or likelihoods in risk management.

The heat map method is clearly a rough and ready tool for the identification of the most important risks. But its value is in providing a common framework in which a group of people can pool their knowledge. Far too often the methodology fails to work as well as it might, simply because there has not been any prior agreement as to what the terms mean. A critical point is to have a common view of the time frame over which risks are assessed. Suppose that there is a 20% probability of a particular risk event occurring in the next year, but the group charged with risk management are using an implicit 10 year time horizon. This would certainly allow them to assess the risk as very likely, since, if each year is independent of the last and the probability does not vary, then the probability that the event does not occur over 10 years is $0.8^{10} = 0.107$. So there is a roughly 90% chance that the event does occur at some point over a 10 year period.

More or less the same argument applies to the terms used to identify the magnitude of the impact. It will not be practicable to give an exact dollar figure associated with losses, just as there is little point in trying to ascribe exact probabilities to risk events. But it is worthwhile having a discussion on what a “minor” or a “moderate” impact really means. For example we might ask about the evaluation of the impact of an event that led to an immediate 5% drop in the company share price.

1.3 Fallacies and traps in risk management

In this introductory chapter it is appropriate to give some ‘health warnings’ about the practice of risk management. These are ideas about risk management that can be misleading or dangerous.

It is worth beginning with the observation that society at large is increasingly intolerant of risk which has no obvious owner – no one who is responsible and who can be sued in the event of a bad outcome. Increasingly it is no longer acceptable to say ‘bad things happen’ and we are inclined to view any bad event as someone’s fault. This is associated with much management activity that could be characterized as “covering one’s back”. The important thing is no longer the risk itself but the demonstration that appropriate action has been taken so that the risk of legal liability is removed. The discussion of risk registers in the previous section demonstrates exactly this divergence between what is done because it brings real advantage, and what is done simply for legal reasons. Power(2004) argues that greater and greater attention is placed on what might be called "secondary risk management", with the sole aim of deflecting risk away from the organization or the individuals within it. It is simply wrong to spend more time ensuring that we cannot be sued than we do in trying to reduce the dangers involved in our business. But in addition to questions of morality a focus on secondary risk management means we never face up to the question of what is an appropriate level of risk, and we may end up losing the ability to make sound judgements on appropriate risks: the most fundamental requirement for risk management professionals.

Another trap we may fall into is the feeling that good risk management requires a scenario based understanding of all the risks that may arise. Often this is impossible and trying to do so will distract attention from effective management of important risks. As Stulz (2009) argues there are two ways to avoid this trap. First there is the use of statistical tools (which we will deal with in much more detail in later chapters).

> "Contrary to what many people may believe, you can manage risks without knowing exactly what they are - meaning that most of what you'd call unknown risks can in fact be captured in statistical risk management models. Think about how you measure stock price risk. ... As long as the historical volatility and mean are a good proxy for the future behavior of stock returns, you will capture the relevant risk characteristics of the stock through your estimation of the statistical distribution of its returns. You do not need to know why the stock return is $+10\%$ in one period and -15% in another."

The second way to avoid getting bogged down in an unending set of almost unknowable risks is to recognize that important risks are those that make a difference to management decisions. Some risks are simply so small in probability that a manager would not change her behavior even if this risk was brought to her attention. This is like the risk of being hit by an asteroid - it must have some small probability of occurring but it does not change our decisions.

A final word of caution relates to the use of historical statistical information to project forward. We may find a long period in which something appears to be varying according to a specific probability distribution, only to have this change quite suddenly. An example with a particular relevance for me is in the exchange rate between the Australian dollar and the British pound. The graph in Figure 1.2 shows what happened to this exchange rate over a five year period from 2004 to 2008.

The weekly data here have mean 1 British pound = 2.38 Australian dollars and standard deviation 0.133. 15 months later, in March of 2010 the rate had fallen to 1.65 (and continued to fall after that date). Now if weekly exchange rate data followed a normal distribution then the chance of observing a value as low as 1.65 (more than five standard deviations below the mean) would be completely negligible. Obviously the foreign exchange markets do not behave in quite the way that this superficial historical analysis suggests. Looking over a longer period and considering also other foreign exchange rates would suggest that the relatively low variance over the 5 year period taken as a base was unusual. In this case the fallout from the global financial crisis quickly led to exchange rate values that reflect historically very high levels for the Australian dollar and a low level for the British pound.

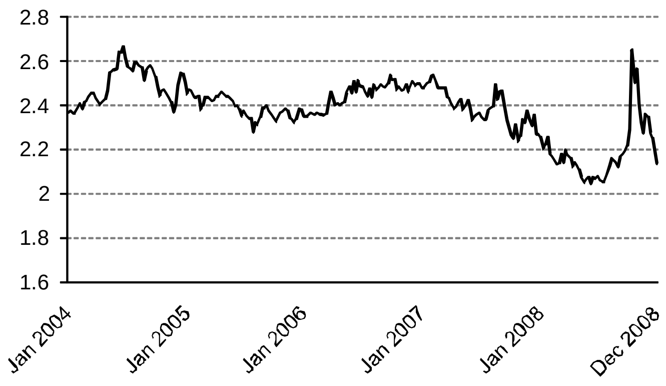


Figure 1.2 Australian dollars to one British pound 2004-2008

However we may be faced with the task of estimating the risk of certain events without the benefit of a very long view or related data. In this situation all that we might have to guide us is a set of data like Figure 1.2. Understanding how hard it is in a foreign exchange context to say what the probabilities are of certain outcomes should help us to be cautious when faced with the same kind of task in a different context.

1.4 Why safety is different

This book is about business risk management and is aimed at those who will have management responsibility. There are significant differences between how we may behave as managers and how we behave in matters of our personal safety. Every day as we grow up, and throughout our adult lives, we make decisions which involve personal risk. The child who decides to try jumping off the playground swing is weighing up the risk of getting hurt against the excitement involved. And the driver who overtakes a slower vehicle on the road is weighing up the risks of that particular road environment against the time or frustration saved. In that sense we are all risk experts: it's what we do every day.

It is tempting to think about safety within the framework we have laid out of different risk events each with a likelihood and a magnitude of impact. With this approach we could say that a car trip to the shops involves such a tiny likelihood of being involved in a collision with a drunk driver that the overall level of risk is easily outweighed by the benefits. But there are two important reasons why thinking in this way can be misleading.

First we need to consider not only the likelihood of a bad event, but also its consequences. And if I am worried about someone else driving into me then the consequence might be the loss of my life. Just how does that get weighed up against the inconvenience of not using a car? Most of us would simply be unable to put a monetary sum on our own lives, and no matter how small the chance of our dieing in a car crash, the balance will tilt against driving the car if we make the value of our life high enough. But yet we still drive our cars and do all sorts of other things that carry an element of personal risk.

A second problem with treating safety issues in the same way as other risks is that the chance of an accident is critically determined by the degree of care taken by the individual concerned. The probability of being killed in a car crash on the way to the shops is mostly determined by how carefully I drive. This makes my decision on driving a car different to a decision on flying, where I have no control over the level of risk. But more often than not being careful will dramatically reduce the risk to our personal safety. Paradoxically the more dangerous we perceive the activity to be then the more careful we are. The risks from climbing a ladder may end up being greater than from using a chain saw if we believe that the ladder is basically safe, but that the chain saw is extremely dangerous.

A better way to consider personal safety is to think of each of us as having an in-built "risk thermostat" that measures our own comfort level with different levels of risk. As we go about our lives there comes a time with certain activities when we start to feel uncomfortable with the risk we are taking, when the amount of risk starts to exceed out own risk thermostat setting. The risk we will tolerate varies according to our own personalities, our age, our experience of life etc. But if the level of risk is below this personal thermostat setting then there is very little that holds us back from increasing the risk. So if driving seems relatively safe then we will not limit our driving to occasions when the benefits are sufficiently large. John Adams points out that some people will actively seek risk so that they return to the risk thermostat setting which they prefer. So in discussing the lives that might be saved if motorcycling was banned he points out that "If it could be assumed that all the banned motorcyclists would sit at home drinking tea, one could simply subtract motorcycle accident fatalities from the total annual road accident death toll. But at least some frustrated motorcyclists would buy old bangers and try to drive them in a way that pumped as much adrenaline as their motorcycling".

These are important issues and need to faced by businesses in which health and safety are big concerns, such as mining. If the aim is to get as close as possible to eliminating accidents in the workplace, then it is vital to pay attention to the workplace culture which can have a role in resetting the risk thermostat to a lower level.

1.5 The Basel framework

Basel II defines different three types of risk for banks – but the framework is quite general and can apply to any business.

Market risk Market risk focusses on the uncertainties that are inherent in market prices which can go up or down. Market risk applies to any uncertainty whose value is dependent on prices that cannot be fully predicted in advance. For example we might build a plant to extract gold from a low yield resource, but there is a risk that the gold price drops and our plant is no longer profitable. This is an example of a *commodity risk*. Other types of market risk are *equity risk* (related to stock prices and their volatility); *interest rate risk*; and *currency risk* (related to foreign exchange rates and their volatility).

Credit risk Any business will be involved in many different contractual arrangements. If the counterparty to the contract does not deliver what is promised then legal means can be used to extract what is owed. But this assumes that the counterparty still has funds available Credit risk is the risk of a counterparty to a contract going out of business. For example a business might deliver products to its customers and have 30-day payment terms. If the customer goes out of business there may be no way of getting back more than a small percentage of what is owed. In its most direct form the contract is a loan made to another party and credit risk is about not being repaid due to bankruptcy.

Operational risk Operational risk is about something going badly wrong. This category of risk includes many of the examples we have discussed so far that are associated with negative risk events. Operational risk is defined as arising from failures in internal processes, people or systems, or due to external events.

Since we are concerned with more general risk management concerns, not just for banks, we need to add a fourth category to the three discussed by Basel II.

Business risk Business risk relates to those parts of our business value proposition where there is considerable uncertainty. For example there may be a risk associated with changes in costs, or changes in customer demand, or changes in the security of supply of raw materials. Business risk is like market risk but does not relate directly to prices.

Both market risk and credit risk are to some extent entered into deliberately as a result of calculation. Market risk is expected, and we can make calculations on the basis of the likelihood of different market outcomes. Business risk also often has this characteristic: most businesses will have a clear idea of what will happen under different scenarios for customer demand. Credit risk is always present, and in many cases we assess credit risk explicitly through credit ratings. But operational risk is different: it is not entered into in the expectation of reward. It is inherent and is in a sense the unexpected risk in our business. It may well fit into the “unknown unknown” description in the quotation from Rumsfeld that we gave above. Usually operational risk involves low probability and high severity events. This creates great difficulties in dealing with operational risk.

1.6 Hold or hedge?

When dealing with market or business risk a manager is often faced with an ongoing risk, so that it recurs from day to day or month to month. In this case there is the need to take strategic decisions related to these risks.

An example of a recurring risk occurs with airlines who face ongoing risks related to the price of fuel (which can only be partially offset by adding fuel surcharges). The question that managers face is when to hold on to that risk, when to insure or hedge it, and when to attack the risk so that it is reduced. In a later chapter we will give a more detailed analysis of how risks can be hedged. In holding on to a risk the company deliberately decides to accept the variation in profit which results. This may be the best option when a company has sufficient financial resources, and when it has aspects of its operations that will limit the consequences of the risk. For example a vertically integrated power utility company may decide not to fully hedge the risks associated with rises in the cost of gas if there are opportunities to quickly change the price of the electricity that it sells in order to cover increased costs of generation.

A financial hedge is possible when we can buy some financial instrument to lessen the risk of market movements. For example a power utility company might trade in futures for gas prices. We will come back to discuss this type of trading in more detail later. Sometimes we have an operational hedge which achieves the same thing as a financial hedge through the way that our operations are organized. For example we may be concerned about currency risk if our costs are primarily in US dollars but our sales are in the Euro zone. Thus if the Euro's value falls sharply relative to the US dollar, then we may find our income insufficient to meet our manufacturing expenses even though our sales have remained strong. An option is to buy a futures contract which has the effect of locking in an exchange rate. However another 'operational hedge' could be achieved by moving some of our manufacturing activity into a country in the Euro zone, so that more of our costs occur in the same currency as the majority of our sales.

1.7 Learning from a disaster

We began this chapter with the remarkable story of Jérôme Kerviel's massive fraud at Société Générale, which fits into the category of operational risk. Now we return to this example with the aim of seeing what can be learnt more generally. To understand what happened it helps to understand more of the world of bank trading. A bank, or any company involved in trading in some sort of financial marketplace, will usually divide its activities into three areas. First the traders themselves: these are the people who decide what trades to make and when to make them (the "front office"). Second, a risk management area responsible for monitoring the trader's activity measuring and modelling risk levels etc. (the "middle office"). And finally an area responsible for carrying out the trades, making the required payments and dealing with the paperwork (the "back office"). Many of these trading activities take advantage of quite small opportunities for profit (in percentage terms) and therefore in order to make it worthwhile they require large sums of money to be involved. Kerviel acted as an arbitrageur looking for small differences in price between different stock index futures. This requires the purchase of one portfolio of stock index futures and the selling of another portfolio of stock index futures. Kerviel was making fictitious trades: reporting trades that did not occur. This enabled him to hold one half of the combined position, but not the other. The result of the fictitious trade is to change an arbitrage opportunity with large nominal value but relatively small risk into a simple (very large) bet on the movement of the futures price.

One reason that it was possible to 'hide' all these fictitious trades was that they were held for only a very short time.

What went wrong?

Kerviel was previously in the risk management (middle office) area and kept some access appropriate to this, even when he became a trader. This is exactly what happened with Nick Leeson at Barings - another famous example of a trader causing enormous losses at a bank.

In Kerviel's case it is important to note that most of the time he was making profits for SocGen. In fact the genesis of this fraud occurred while there was a big expansion in trading activity at SocGen.

A critical point is that his immediate superior left the company and when eventually this person was replaced the new superior did not keep track of what was going on. In fact there were several things which should have alerted the company to a problem:

- There was a huge jump in earnings for Kerviel's desk in 2007;

- There were questions which were asked about Kerviel's trades from the Eurex exchange;
- There was an unusually high level of cash flow associated with Kerviel's trading,
- Kerviel did not take a vacation of more than a couple of days at a time - despite a policy enforcing annual leave.
- There was a breach of Kerviel's market risk limit on one position.

We can draw some more general lessons from this case. I list five of these below

Company culture is more important than the procedures Company culture in SocGen gave precedence to the money making side of the business (trading) over the risk management side (middle office), and this is very common. Whether or not procedures are followed carefully will always depend on cultural factors, and the wrong sort of risk culture is one of the biggest factors that lead to firms making really disastrous decisions.

Good times breed risky behaviour In the SocGen case the fact that Kerviel's part of the operation was doing well made it easy to be lax in the care with which procedures were carried out. It may be true that the reverse of this statement is also true: in bad times taking risks may seem the only way through, but whether wise or not these are at least a conscious choice. Risks that managers enter into unconsciously seem to generate the largest disasters.

Companies often fail to learn from experience One example occurs when managers ignore risks in similar companies, such as we see in the uncanny resemblance between SocGen and Barings. But it can also be surprisingly hard to learn from our own mistakes in a corporate setting. Often a scapegoat is found and moved on, without a close look at what happened and why. Dwelling on mistakes is a difficult thing to do and will inevitably be perceived as threatening, and perhaps that is why a careful analysis of bad outcomes is often ducked.

Controls need to be acted upon On many occasions risks have been considered and controls put in place to avoid them. The problem occurs when the controls that are in place are ignored in practice. SocGen had a clear policy on taking leave (as is standard in the industry) but failed to act upon it.

There must be adequate management oversight Inadequate supervision is a key ingredient in poor operational risk management. In the SocGen case Kerviel's supervisor had inadequate experience and failed to do his job. More generally risk management escalates when a single person or a small group can make decisions that end with large losses, either through fraud or simple error. Companies need to have systems that avoid this through having effective oversight of individuals by managers, who need to supervise their employees sufficiently closely to ensure that individuals do what they are supposed to.

This book is mostly concerned with the quantitative tools that managers can use in order to deal with risk and uncertainty. It is impossible to put into a single book everything that a manager might need to know about risk. In fact the most important aspects of risk management in practice are things that managers learn through experience better than they learn in an MBA class. But paying attention to the five key observations above will be worthwhile for anyone involved in risk management, and may end up being more important than all the quantitative methods we are going to explore later in this book.

It is hard to overstate the importance of the culture within an organization: this will determine how carefully risks are considered; how reflective managers are about risk issues; and whether or not risk policies are followed in practice. Paradoxically a culture that is open about risk, and prepared to discuss and consider the appropriate level of risk is far more likely to avoid disasters than a culture that wishes to deny the possibility of any risk at all. When we are frightened of risk we are more likely to ignore it or hide it than to actually take steps to reduce it.

Notes

This chapter is rather different than the rest of the book: besides setting the scene for what follows it also avoids doing much in the way of quantification. I have tried to distill some important lessons rather than give a set of models to be used. I have found the book by Douglas Hubbard one of the best resources for understanding the basic of risk management applied in a broad business context. His book covers not only some of the material in this chapter but also has useful things to say about a number of topics we cover in later chapters (such as the question of how risky decisions are actually made that we cover in Chapter 6).

The discussion on why we need to think differently about safety issues is taken from the influential book by John Adams, who is a particular expert on road safety.

We have said rather little about company culture and its bearing on risk. That topic probably deserves a whole book to itself (some references on this are Bozeman and Kingsley (1998) and the PwC report on Building Effective Risk Cultures).

References

John Adams, *Risk*, UCL Press, 1995

Barry Bozeman and Gordon Kingsley, (1998), Risk Culture in Public and Private Organizations, *Public Administration Review*, Vol. 58, pp. 109-118

Douglas Hubbard, *The Failure of Risk Management*, Wiley, 2009.

Michael Power, (2004) The risk management of everything, *Journal of Risk Finance*, Vol. 5 No. 3, pp.58 - 65

PwC report *Cure for the Common Culture: Building Effective Risk Cultures at Financial Institutions* http://www.pwc.com/gx/en/risk-regulation/risk-aware-culture.jhtml accessed 1 March 2012

Jeff Rice and Stephen Franks, (2010) The agile enterprise, *Analytics Magazine*, INFORMS, January 2010.

Bob Ritchie and Clare Brindley, (2007) Supply chain risk management and performance, *International Journal of Operations & Production Management*, Vol. 27 No. 3, pp. 303-322.

René M. Stulz, (2009) Six ways companies mismanage risk, *Harvard Business Review*, March 2009

Exercises

1.1. (Supply risk for valves)

DynoRam makes hydraulic rams for the mining industry in Australia. It obtains a valve component from a supplier called Sytoc in Singapore. The valves cost $250 Singapore dollars each and the company uses between 450 and 500 of these each year. There are minor differences between valves with a total of 25 different types being used by DynoRam. Sytoc delivers the valves by air freight, typically about 48 hours after the order is placed. Deliveries take place up to 10 times a month depending on the production schedule at DynoRam. Because of the size of the order Sytoc has agreed a low price on condition that a minimum of 30 valves are ordered each month. On the 10th of each month (or the next working day) DynoRam pays in advance for the minimum of 30 valves to be used during that month and also pays for any additional valves (above 30) used during the previous month.

(a) Give one example of market risk, credit risk, operational risk and business risk that could apply for DynoRam in relation to the Sytoc arrangement.

(b) For each of the risks identified in part (a) suggest a management action which would have the effect either of reducing the probability of the risk event or minimizing the adverse consequences.

1.2. (Connaught)

The following is an excerpt from a newspaper report of 21/7/10 appearing in the UK Daily Telegraph.

> Troubled housing group Connaught has been driven deeper into crisis after it discovered a senior executive sold hundreds of thousands of pounds worth of shares ahead of last month’s shock profit warning.
>
> The company which lost more than 60% of its value in just three trading days in June, and saw its chief executive and finance director resign, has launched an internal investigation into the breach of city rules... Selling shares with insider information when a company is about to disclose a price-sensitive statement is a clear breach of FSA rules.
>
> Connaught,which specialises in repairing and maintaining low cost (government owned) housing, has fallen a total of 68% since it gave a warning that a number of public sector clients had postponed capital expenditure, which would result in an 80 million pound fall in expected revenue this year.
>
> The group said that it had been hit by deferred local authority contracts which would knock 13m pounds off this year’s profits and 16m pounds from next year’s. It also scaled back the size of its order book from the 2.9 billion pounds it said it was worth in April to 2.5 billion.
>
> The profit warning also sparked renewed concerns about how Connaught accounts for its long-term repair and maintenance contracts. Concerns first surfaced late last year with city analysts questioning whether the company was being prudent when recognising the revenue from, and costs of, its long term contracts.

> The company vehemently defended its accounting practices at the time and continues to do so. Chairman Sir Roy Gardner has tried to steady the company since his arrival earlier this year.

(a) How would you describe the ‘profits warning’ risk event: is it brought about by market risk, credit risk, operational risk or business risk?

(b) From the newspaper report can you make any deductions about risk management strategies the management of the company could have taken in advance of this in order to reduce the loss to shareholders?

1.3 (Bad news stories)

Go through the business section of a newspaper and find a ‘bad news’ story, where a company has lost money.

(a) Can you identify the type of risk event involved: market risk, credit risk, operational risk or business risk?

(b) Look at the report with the aim of understanding the risk management issues in relation to what happened. Was there a failure to anticipate the risk event? Or a failure in the responses to the event?

1.4 (Product form for heat map)

Suppose that the risk level is calculated as the expected loss and that the likelihoods are converted into probabilities as follows: ‘very likely’ $= 0.9$; ‘likely’ $= 0.7$; ‘moderate’ $= 0.4$; ‘unlikely’ $= 0.2$; and ‘rare’ $= 0.1$. Find a set of dollar losses associated with the five different magnitudes of impact such that the expected losses are ordered in the right way for Figure 1: in other words so that the expected losses for a risk level of low are always lower than the expected losses for a risk level of medium, and these are lower than the expected losses for a risk level of high, which in turn are lower than the expected losses for a risk level of extreme. You should set the lowest level of loss (‘insignificant’) as $10,000.

Appendix: The SocGen trading fraud.

The following account is put together from newspaper stories that appeared as events unfolded. The announcement appeared in papers on 24 January 2008

Société Générale loses $7 billion in trading fraud

New York Times, January 24, 2008

PARIS — Société Générale, one of the largest banks in Europe, was thrown into turmoil Thursday after it revealed that a rogue employee had executed a series of “elaborate, fictitious transactions” that cost the company more than $7 billion, the biggest loss ever recorded in the financial industry by a single trader.

Daniel Bouton, the Société Générale chairman, said the employee, later identified by other bank employees as Jérôme Kerviel, had confessed to the €4.9 billion fraud, although he did not appear to have profited personally from the trades.

Before the discovery of the fraud, Société Générale had been preparing to announce pretax profit for 2007 of €5.5 billion, a figure that Bouton said would have shown the company’s “capacity to absorb a very grave crisis.” Instead, Bouton - who is forgoing his salary through June as a sign of taking responsibility - said the ‘unprecedented’ magnitude of the loss had prompted it to seek about €5.5 billion in new capital to shore up its finances, a move that secures the bank against collapse.

The situation drew comparisons with Nick Leeson, the trader in Singapore who in 1995 incurred a loss of $1.4 billion by making $27 billion of bad bets on Japanese markets, bringing down the venerable British bank Barings in the process. It also raised questions about how losses of this nature could go totally unidentified amid the network of risk management in place at a major bank like Société Générale.

Société Générale said it had no indication whatsoever that the trader - who joined the company in 2000 and worked for several years in the bank’s French risk-management office before being moved to its Delta One trading desk in Paris - “had taken massive fraudulent directional positions in 2007 and 2008 far beyond his limited authority.”

The bank added: “Aided by his in-depth knowledge of the control procedures resulting from his former employment in the middle-office, he managed to conceal these positions through a scheme of elaborate fictitious transactions.”

The trader continued the fraud until this past weekend, when auditors in the company’s risk-management office detected a series of fictitious trades on its books, which it said was committed by an employee in charge of hedging the bank’s trades in European stock index futures.

When the fraud was unveiled, Bouton said, it was “imperative that the enormous position that he had built, and hidden, be closed out as rapidly as possible.” The timing could hardly have been worse. Société Générale was forced to begin unwinding the trades on Monday “under conditions of extreme market volatility,” Bouton said, as global stock markets plunged amid mounting fears of an economic recession in the United States.

“The result was a considerable loss,” Bouton said. The scandal has the potential to be the largest trading fraud ever.

The bombshell for Société Générale comes at a time when the mounting losses from subprime-related investments have raised questions about risk-control management at many institutions.

The second story appeared two days later and gives more details on how Kerviel had carried out the fraud.

Bank Outlines How Trader Hid His Activities

New York Times: January 28, 2008

PARIS — The French bank Société Générale facing persistent questions over how a lone, junior trader could have instigated more than $7 billion in losses, acknowledged on Sunday that his activities prompted questions from risk managers several times last year, but that the bank never began an investigation because his explanations defused any suspicions.

The disclosure came as the trader, Jérôme Kerviel, 31, spent a second day in police custody, facing questions about what the bank asserts was an elaborate, year long ruse that involved betting tens of billions of dollars of the bank's money on European stock index futures.

Mr. Kerviel's lawyers late Sunday denounced what they called the "media lynching" of their client in recent days and argued that the bank's managers "brought the loss on themselves."

In a five-page statement, the bank outlined how it believed Mr. Kerviel combined several different "fraudulent methods" to hide his activity — including using computer access codes of other employees and falsifying documents. Briefing reporters separately by telephone, Jean-Pierre Mustier, chief executive of the bank's corporate and investment banking arm, said that the discovery of the $7.2 billion fraud on Jan. 18 and the unwinding last week of the roughly $70 billion worth of risky investments that were uncovered represented "one of the most difficult periods in the history of Société Générale."

Mr. Mustier also repeated the bank's assertion that Mr. Kerviel appeared to have acted alone. "We have made extensive checks of his portfolio as well as the portfolios of others to see if there was anything like the types of transactions he was using," Mr. Mustier said. "It seems extremely unlikely" that he was helped by others, he said.

Mr. Mustier explained that Mr. Kerviel's role on the trading desk was that of an arbitrageur, which meant that he was entrusted to purchase one portfolio of stock index futures and at the same time sell a similar mix of index futures, but with a slightly different value. The object of arbitrage is to try to make profits from these differences in value. Because the value gaps between similar financial instruments are usually very small and temporary, this type of activity typically involves trading in very high total nominal amounts.

Mr. Kerviel's fraud, according to the bank, consisted of placing sizable, real purchases in one portfolio but creating fictitious sales transactions in the second, offsetting portfolio. This gave the impression to risk managers that the risks in the first portfolio were hedged, when in fact they were not.

As a result, the bank wound up exposed to huge one-way bets, or long positions. Instead of hedging, which was his job, Mr. Kerviel was effectively speculating with the bank's money.

"Our controls identified from time to time problems with this trader's portfolio," Mr. Mustier said. Each time one of Mr. Kerviel's trades was questioned, he would describe it as a "mistake" and cancel the trade, Mr. Mustier said. "But in fact, he then replaced that trade with another transaction using a different instrument" to avoid detection, he said.

A few months later a fuller explanation was published after an internal investigation.

Societe Generale Explains $7 Billion Fraud: Bad Management

www.huffingtonpost.com May 23, 2008

PARIS — Investigators at Societe Generale say they suspect former futures trader Jerome Kerviel was helped by an assistant to cover up massive trading positions that led to a multibillion dollar loss.

In two long-awaited reports released Friday, the investigators said the French bank's management failures and culture of risk-taking were partly to blame for failing to spot the positions, which led to a loss of more than $7 billion once they were unwound.

Investigators say Kerviel's bosses missed more than 1,000 faked trades; a huge jump in his earnings in 2007; questions about his trades from the Eurex exchange; unusually high levels of cash flow, accounting anomalies, and high brokerage expenses; Kerviel's failure to take vacation; and his breach of the desk's market risk limit on one position.

"The trader's hierarchy, constituting the first level of control, proved deficient in the supervision of his activities," the board of directors told shareholders in a seven-page statement accompanying the reports.

The reports said Kerviel's direct superior "lacked trading experience" and showed "an inappropriate degree of tolerance" about his trades. The bank did not name the superior, who they said they have been unable to question because he no longer works for the company. The reports also criticized the manager of the company's Delta One trading desk, who they said was aware of the lack of experience of Kerviel's manager, and "deficiencies in the monitoring of risks by the desk in general."

Kerviel says his superiors must have known what he was doing but that they chose to look the other way when he was making money for the bank.

SocGen employee Patrice Leclerc, who heads an association of the bank's employee shareholders, said Kerviel's managers need to explain why they missed warning signals that were picked up in Germany - and why they didn't follow up on questions from Eurex, the German derivatives exchange. "There exist ways of monitoring this type of thing - why weren't they used?" he said.

Investigators didn't find any signs of embezzlement by Kerviel, but said it appeared he had sought to boost his results to increase the amount of his bonus. The report established that part of Kerviel's "official" earnings came from his concealed positions.

While the directors' report plumbed the details of the alleged fraud, a second report by PriceWaterhouseCoopers focused on the culture of risk-taking at the bank as it grew its investment banking business and on a review of the measures taken by the bank to fix the problems that the scandal exposed.

"The surge in Delta One trading volumes and profits was accompanied by the emergence of unauthorized practices, with limits regularly exceeded and results smoothed or transferred between traders," the PWC report said. "Several key controls that could have identified fraudulent mechanisms were lacking" and "there was a lack of an appropriate awareness of the risk of fraud," it said.

SocGen's board said in a statement to shareholders it approved the conclusions of both reports and their recommendations. The board said the investment bank is tightening computer security, reinforcing controls and taking more account of the possibility of fraud.

There were more stories published at the conclusion of Kerviel's trial at the end of June 2010.

A Société Générale Trader Remains a Mystery as His Criminal Trial Ends

New York Times, June 25, 2010

PARIS — The trial of Jérôme Kerviel, the man accused of causing billions in losses at the French bank Société Générale, ended Friday with judges and prosecutors conceding that a two-year investigation and three weeks of court hearings had left them no wiser about what had motivated the former trader to make his enormous, unauthorized bets. A three-judge panel will now spend the coming months poring over the testimony of more than 40 witnesses in an attempt to determine whether blame for the scandal should rest solely on the shoulders of one man.

On Thursday, the Paris prosecutor, Jean-Michel Aldebert, requested a jail sentence of at least four years for Mr. Kerviel, whom he called "a manipulator, a trickster and a liar." The maximum possible penalty for the charges faced by the former trader — breach of trust, forgery and unauthorized use of the bank's computer systems — is five years in prison and a fine of 375,000 euros ($464,000).

Mr. Kerviel, 33, acknowledged to the court that he had falsified documents and entered fake trades to hide his bets, but he maintained that his bosses had deliberately turned a blind eye and tacitly encouraged his activities as long as they were earning profits. "It wasn't me who invented these techniques — others did it, too," Mr. Kerviel said, though he never named names. "These practices were known and recognized by management."

Inside the oak-paneled courtroom — formerly an auction hall — of the 19th-century Palace of Justice, the atmosphere at times was more like a university lecture than a criminal hearing, with bank officials, regulators and risk management experts delivering jargon-filled primers on financial instruments like "turbo" warrants, forwards and futures that were the currency of Mr. Kerviel's trading desk, Delta One.

Despite the technical nature of much of the testimony, public interest was high. The wooden benches of the media gallery were filled on most days. More than a dozen of Mr. Kerviel's former colleagues appeared as witnesses. Some shed tears as they described the repercussions of his actions on their careers and personal lives.

But throughout the proceedings, Mr. Kerviel himself betrayed little, if any, emotion. His arms crossed and his gaze often fixed on the floor, he sat impassively as his former bosses portrayed him as an aberrant cheat who had taken advantage of weaknesses in the bank's risk-control systems.

Daniel Bouton, 60, who resigned as the bank's chief executive after the scandal and retired as its chairman last year, called Mr. Kerviel an "evil genius" whose actions had been a "catastrophe" that nearly destroyed the 145-year-old bank.

The testimony of Mr. Kerviel's former direct superior, Eric Cordelle, was in many ways illustrative of the simple and yet intractable question underpinning the trial, which was not so much who knew about Mr. Kerviel's activities, but who could, and should, have known. Where should the line between negligence and tacit endorsement be drawn?

Mr. Cordelle, 38, was appointed head of the Delta One desk in April 2007, and is one of two managers who were fired for incompetence immediately after the scandal. An engineer with no trading experience, Mr. Cordelle said he had never suspected that Mr. Kerviel was operating outside his mandate. He added that because his desk had been understaffed, he had had no time to scrutinize traders one by one. "If you're not looking for anything," he said, "you don't find anything."

Moussa Bakir, a futures dealer who made more than a million euros in commissions in 2007 on several large trades he brokered for Mr. Kerviel, disclosed the lengths to which the former trader had gone to hide his activity. Mr. Bakir, 34, testified that when he asked who had been behind the trades, Mr. Kerviel told him they had been on behalf of a client called “Mat,” a 35-year-old, rugby-loving hedge fund manager in London. Mr. Kerviel spoke often of Mat, describing him as a demanding customer whose aim was to make a billion euros in profit, Mr. Bakir said.

Mr. Kerviel conceded to the court that Mat did not exist and that he had invented him to discourage Mr. Bakir, who was also a friend, from investigating further.

“I believed Jérôme Kerviel without any doubt,” Mr. Bakir said. “From beginning to end.” Early on in the investigation, the police had considered Mr. Bakir a possible accomplice who might have been aware of at least some of Mr. Kerviel’s activities. Société Générale never filed charges against the broker.

Frédérik-Karel Canoy, a lawyer representing shareholders seeking civil damages from the bank in the case, dismissed the bank’s claims that its ignorance of Mr. Kerviel’s activities absolved it of responsibility for the losses. “Société Générale was responsible for the acts of its employee,” Mr. Canoy said. “It had an obligation to monitor him.”

Finally in October 2010 Kerviel was sentenced to three years in prison and ordered to repay Societe Generale SA’s 4.9 billion-euro ($6.8 billion) trading loss by a judge who said the former trader’s crimes threatened the bank’s existence. If an appeals court upholds the verdict, Kerviel is likely to spend between 18 months and two years in prison. Judge Pauthe held Kerviel solely responsible for the loss, saying he deceived the bank in amassing 50 billion euros in futures positions. He found Kerviel guilty on all three counts: breach of trust, forging documents and computer hacking. Mr Kerviel’s current monthly salary as a computer consultant amounts to €2,300, which means it would take him 177,536 years to pay off the money. In theory, France’s second largest bank can force him to hand over all his earnings bar a small monthly sum for “basic needs”. Société Générale said it did not expect its 33-year-old former employee to repay the debt any time soon, and that it was largely “moral compensation”. However, its lawyers insisted the bank would pursue him for any earnings he makes out of the world’s biggest rogue trading scandal. (Taken from article by Henry Samuel, Telegraph, UK 5 October 2010)

2

The Structure of Risk

Did the global financial crisis signal a failure in diversification?

The idea of diversification is simple but vital in managing investments. Invest everything in one stock and you may be unlucky with a bad result for that particular stock; invest in 50 stocks and no single bad event will be able to trip you up. That is the advantage of buying shares in a mutual fund (or investment trust) - the investor is automatically diversifying through holding a whole range of different stocks. But sometimes this principle of diversification seems to fail: if you had $100,000 invested in the average US mutual fund at the beginning of 2008, you would have lost $39,500 during that year. That was a year in which all stocks did badly. The US bear market that began in October 2007 ran till March 2009 and in that period US equity markets fell by 57%.

Diversification amongst different US stocks did not help in 2008. In fact the only way to avoid losses was not to diversify, but to invest in one of the small number of stocks that did much better than the market. (93% of all US equities lost money in 2008 but not all: for example McDonalds and Wal-Mart were exceptions). So if diversification is the answer then the right strategy in that year would include investing in areas outside of equities. If you were a risk averse investor then it would have made sense to diversify into many different asset classes with the intention of avoiding large losses. Unfortunately not only did stocks fall, but commodities, real estate and emerging markets fell as well. In fact virtually every asset class did badly and, as a result, even well-diversified portfolios saw massive losses. This is exactly what diversification is supposed to avoid. And if it turns out that in really bad times everything goes down at once, what is the point of diversification in the first place?

In order to understand the benefits of diversification and avoid being led astray by models that leave out critical parts of the picture, we need to go back to first principles.

2.1 Introduction to probability and risk

The origin of our ideas of probability can be traced back to games of chance. There have always been high stakes gamblers and whenever there is a large amount of money involved there is also a powerful motivation for developing tools to predict the chance of winning or losing. The intellectual history of the ideas of probability and risk is a long one, but one early reference occurs in a book by Luca Paccioli that appeared in 1494 with the title *Summa de Arithmetic, Geometria et Proportionalita*. This is a book on mathematics but it includes an influential description of double entry bookkeeping, as well as the following problem:

Business Risk Management Edward Anderson

> A and B are playing a fair game of *balla*. They agree to continue until one has won six rounds. The game actually stops when A has won five and B three. How should the stakes be divided.

How would we answer this problem? Just one more win for A (Adam) will seal his victory, but B (Ben) still has a chance: if he can win the next three games then he will be the overall winner. A fair division of the stakes needs to reflect the relative likelihood of one or other player winning. But at the time the question was posed there was no language that could be used to capture these ideas of likelihood. Perhaps it is not surprising that this problem ("the problem of the points") would not be solved till many years later.

In fact the problem was discussed in the letters between two famous mathematicians in the 1650s, Blaise Pascal and Pierre de Fermat. Both these men were brilliant: Pascal was a child prodigy who worked on everything from calculating machines and hydraulics to barometers, but renounced his scientific work at the age of 31 after a mystical experience; while Fermat was a mathematical genius who did far more than leave us the puzzle of 'Fermat's last theorem'. Pascal describes the right form of division of the initial stakes in the problem of the points by saying '*the rule determining that which will belong to them will be proportional to that which they had the right to expect from fortune*'.

Rather than leave the problem of the points hanging in the air we can quickly sketch how Pascal and Fermat approached this issue. The key is to think about what may happen over the next 3 rounds. The game will certainly finish then (after 11 rounds in all) if it has not finished already. But it does no harm to suppose that all 3 rounds are played, with a total of 8 possible outcomes: e.g. one of the outcomes is first Ben wins, then Ben wins again, then finally Adam wins. We can write down the possible sequences AAA, AAB, ABA, ABB, BAA, BAB, BBA, BBB. Then since only one of these, the last, has Ben winning the stake, a fair division is to divide the stake with a proportion $1/8$ going to Ben and $7/8$ to Adam. Pascal and Fermat discussed how this calculation could be effectively carried out for any possible starting position and any number of rounds of play.

The question of how to determine a fair distribution can be resolved by appealing to the assumption that the underlying game is fair. Each time that Adam plays Ben there is an equal chance that either of them will win (making each of the 8 possible outcomes equally likely). But even without knowing the exact rules of the game of balla we may object that Adam's superior performance so far provides evidence that he is the stronger player, and if that is so then even giving Ben one eighth of the stake is too generous.

Jacob Bernoulli has these issues in mind when he wrote Ars Conjectandi (The Art of Conjecture) which was published posthumously in 1713 (eight years after his death). At this stage some more modern ideas of probability were beginning to emerge. Bernoulli was conscious of the need to use the past as a indicator of the likelihood of future events and in particular drew attention to the way that (if nothing changes in the underlying circumstances) an increasing number of observations leads to increasing certainty regarding the actual probability of something occurring. He wrote:

> Because it should be assumed that each phenomenon can occur and not occur in the same number of cases in which, under similar circumstances, it was previously observed to happen and not to happen. Actually, if, for example, it was formerly noted that, from among the observed three hundred men of the same age and complexion as Titius now is and has, two hundred died after

> ten years with the others still remaining alive, we may conclude with sufficient confidence that Titius also has twice as many cases for paying his debt to nature during the next ten years than for crossing this border. Again, if someone will ... be very often present at a game of two participants and observe how many times either was the winner, he will thus discover the ratio of the number of cases in which the same event will probably happen or not also in the future under circumstances similar to those previously existing.

There are clearly limitations in this approach. There are inevitable variations between circumstances in the future and those in the past - so how do we know that Titius really faces the same probability of an early death as the 300 men ‘of the same age and complexion’? There has always been a dispute about the extent to which we can rely on calculations based on what has happened in the past to guide us in our future actions. When dealing with every day decisions we are much more likely to be guided by a more subjective understanding of what the future holds. Peter Bernstein talks of the “tension between those who assert that the best decisions are based on quantification and numbers, determined by the patterns of the past, and those who base their decisions on more subjective degrees of belief about the uncertain future.”

In this book we will focus on the quantitative tools for modelling and managing risk. But the student who wants to apply these tools needs to be constantly aware of their limitations. We have now reached the point in the historical story at which the mathematical theory of probability takes off, and rather than talk about the various contributions of de Moivre, Bayes, Laplace, Gauss, D’Alembert, Poisson and others we will move on to the specifics of risk. A short discussion of the all the probability theory we will need in this book is given in the Appendix: Tutorial on Probability Theory.

2.2 The structure of risk

If probability and probability distributions are the right tools to use in understanding risk then the next step is to think about the structure of the risk that we face and see how this is reflected in the probabilities involved.

The first distinction to make is between **event risk** and **quantity risk**. Event risk has a simple yes/no form: What is the risk that a particular company goes bankrupt? What is the risk that a new drug fails to pass it’s safety checks. Quantity risk relates to a value which can vary (a *random variable* in the parlance of probability theory). Most often the ‘value’ is measured in monetary terms. This is a type of risk where there is no simple yes/no result: What is the risk of losses in an investment project? What is the risk of a high cost in a construction project? Quantity risks can always be converted to event risks by adding some sort of hurdle: rather than asking about losses in general we may ask about the risk of losing more than \$500,000.

2.2.1 Intersection and union risk

Sometimes event risk involves a number of separate things failing at once. For example we may consider the risk of a failure in power supply to a hospital as the risk that there is a power cut **and** the emergency generator fails. We call this an *intersection risk*, since it relates to the intersection of the two events: “mains power fails” and “emergency power fails”.

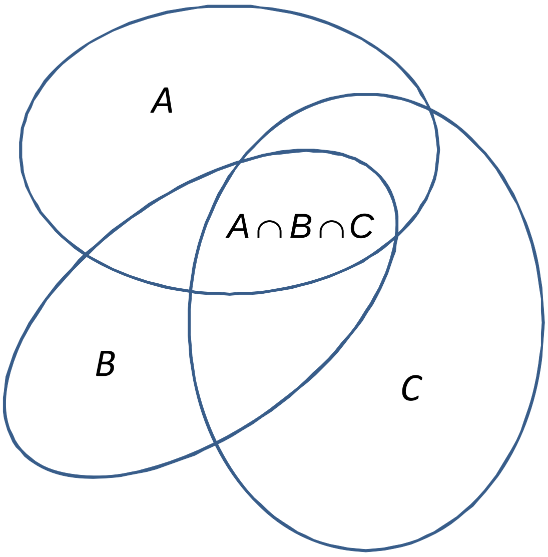


Figure 2.1 Venn diagram showing three different risk events

On the other hand we may have to consider risks where there are a number of different failure paths, each of which lead to the same outcome. If we consider the risk of a failure in a rocket launch then any one of a number of different things can go wrong in the last few seconds before takeoff, and each will produce the same end result. We call this a *union risk*, since the probability of failure is the probability that one or more of the events take place.

The basic tool to visualize these situations is a Venn diagram, where each event is represented as a set in the diagram, and the overlap between sets A and B represents the event that both A and B occur. This is shown for three risk events A, B and C in the diagram of Figure 2.1. The intersection risk is the probability of the event described by the intersection of A, B and C.

We say that two events are independent if one of them occurring makes no difference to the likelihood of the other occurring. This means that the probability of both A and B occurring is given by the product of the two individual probabilities (see the Appendix: Tutorial on Probability Theory for more details about this). This allows us to calculate the intersection risk for independent events.

Returning to the hospital power supply example, suppose we know the probability of a power cut on any given day is 0.0005 and the probability that the hospital emergency power fails on any given attempt to start the generator is 0.002. If the two events are independent (as seems likely in this example) then the probability of a power supply failure at the hospital on any given day is $0.0005 \times 0.002 = 0.000001$. This corresponds to a probability of 0.000365 that a failure occurs in a given year and an expected time between failures of one over this number, or 2740 years.

Now we want to consider union risk, and as an example of this consider the probability of a catastrophic rocket launch failure during takeoff. Suppose that the three main causes of failure are as follows: A = failure in fuel ignition system; B = failure of the first stage separation from the main rocket; and C = failure in the guidance and control systems. Suppose that the probabilities are as follows $\Pr(A) = 0.001$, $\Pr(B) = 0.0002$, and $\Pr(C) = 0.003$. What is the overall probability of failure if the events A, B and C are all independent?

The probability that we want is the probability that one or other of A, B and C occur. We want to find the entire area covered by the union of the sets A, B and C in the Venn diagram. This is given by the formula

$$\Pr(A) + \Pr(B) + \Pr(C) - \Pr(A \cap B) - \Pr(B \cap C) - \Pr(A \cap C) + \Pr(A \cap B \cap C).$$

This is called the *inclusion-exclusion formula* - the idea is that if we just add up the probabilities of A, B and C we will double count the intersections, so the second three terms correct for this, but then anything in the intersection of all three sets will have been counted 3 times initially and then been taken away three times, so the final term restores the balance to make all the components of the Venn diagram end up being counted just once. With the probabilities given and using the product form for the probability of the intersection of independent events, we end up with a probability of launch failure given by

$$10^{-3} + 2 \times 10^{-4} + 3 \times 10^{-3} - 2 \times 10^{-7} - 6 \times 10^{-7} - 3 \times 10^{-6} + 6 \times 10^{-10} \\ = 0.004196.$$

This example shows how, when there are small risk probabilities and independent events, we can more or less ignore the extra terms after the sum $\Pr(A) + \Pr(B) + \Pr(C)$. In this example simply adding the three probabilities gives the value 0.0042.

It is obvious that there is an enormous difference between the end result of a union risk where (approximately) probabilities get added, and an intersection risk where probabilities get multiplied.

2.2.2 Maximum of random variables

Now we consider quantity risks, involving random variables rather than events. Again we need to start with an understanding of the structure of the risk and the way that different random variables are combined.

We first look at a situation in which the risk we want to determine is determined by the largest of a set of random variables. For example suppose we want to find the probability that the price of IBM shares drops by more than 10% in one day at some point over the next 4 weeks. There are 20 trading days and so this is a question about the probability that the largest drop in value over a 20 day period is more than 10%. Suppose that the probability of a drop of more than 10% in a single day is 0.01 so we expect that it will happen on one trading day in a 100. If the behavior on successive days is independent then we can calculate the answer we want by looking at the event: A = IBM stock rises or IBM stock drops by less than 10%. The probability of dropping by more than 10% is $1 - \Pr(A)$. Given that this is 0.01, we must have $\Pr(A) = 0.99$. With independence the probability that the stock drops by less than 10% on both day 1 and day 2 is the intersection probability $\Pr(A) \times \Pr(A) = 0.99^2 = 0.9801$. The probability that the stock drops by less than 10% on all 20 days is just $\Pr(A)^{20} = 0.99^{20} = 0.8179$. But if this doesn't happen then there is at least one day when it drops by more than 10%, which is exactly the probability we want to find. So the answer we need is $1 - \Pr(A)^{20} = 0.1821$.

The task of determining the risk that the largest of a set of random variables is higher than a certain value is just like the analysis for the union risk: it is the probability that one or more of these random variables is greater than a certain value. However rather than giving a complex expression based on the inclusion-exclusion formula we have instead converted the problem to a kind of intersection risk problem. We can rewrite the analysis of the paragraph above in a slightly more formal way as follows. We define the events: B_i = IBM stock drops by more than 10% on day i. So B_i is the opposite of A_i= IBM stock drops by less than 10% on day i. A_i is the same as the event A, but we add a subscript to indicate the day in question. We have $\Pr(B_i) = 1 - \Pr(A_i)$. Then we note that $B_1 \cup B_2$ is the opposite of the event $A_1 \cap A_2$. So $\Pr(B_1 \cup B_2) = 1 - \Pr(A_1 \cap A_2)$, and a similar expression holds when three or more days are considered. We actually want to find $\Pr(B_1 \cup B_2 \cup ... \cup B_{20})$, but we get to this through evaluating $1 - \Pr(A_1 \cap A_2 \cap ... \cap A_{20})$. But if all the A_i have the same probability and are independent (as we are assuming here) then this becomes $1 - \Pr(A)^{20}$.

We can convert this discussion about probabilities of events into a discussion of cumulative distribution functions or CDFs. Remember that we define the CDF for a random variable X as $F_X(z) = \Pr(X \leq z)$. Now consider a random variable U defined as the maximum of two other random variables X and Y. Thus $U = \max(X, Y)$. To find the CDF for U we need to find the probability that the maximum of X and Y is less than a given value z. This is just the probability that both X and Y are less than z, so

$$F_U(z) = \Pr(X \leq z \text{ and } Y \leq z).$$

Hence, when X and Y are independent,

$$F_U(z) = F_X(z) \times F_Y(z).$$

The same idea can be used when X and Y have the same distribution. Suppose that $X_1, X_2, ...X_N$ are identically distributed and independent random variables, all with the same CDF given by F_X. Then the CDF for the random variable $U = \max(X_1, X_2, ..., X_N)$ is given by

$$F_U(z) = (F_X(z))^N .$$

One of the most common questions we need to answer is not about the largest (or smallest) of several different random variables, but instead relates to the risk arising when random variables are added. Hence we are concerned not with $\max(X, Y)$, but with $X + Y$. In our example above we asked about the probability that an IBM share price falls by more than 10% in a single day during a 20 day period. But we are just as likely to be interested in the total change in price over the 20 day period, and to calculate this we need to add together the successive price movements over those 20 days. The fundamental insight here is that extreme events in one day's movement are quite likely to be cancelled out by movements on other days. As a result we can say that, unless price movements are strongly positively correlated, the risk for the sum of many individual elements is less than the sum of the individual risks. In the next section we explore this idea in more detail.

2.3 Portfolios and diversification

We began this chapter by talking about diversification in share portfolios and now we return to this theme. The essential risk idea can be captured with the advice: “Don’t put all your eggs in one basket”. If there is the option to do so, then it is better to spread risk so that different risk events act on different parts of an entire portfolio of activities. In a stock market context investing in a single share will carry the risk that all one’s money is lost if that firm goes bankrupt. Splitting an investment between a portfolio of many different shares automatically reduces the probability of this very extreme result. The final result for the investor is the sum of the results obtained for each share in the portfolio (weighted according to the amount invested). This sum ensures that a bad result in one part of the portfolio is likely to be balanced by a good (or less bad) result in another part of the portfolio.

We will start by looking in more detail at what happens when random variables are added together. If we consider the sum of two random variables X and Y, each representing a loss, then we can ask what is the probability that the sum of the two is greater than a given value? So we take $U = X + Y$ and consider $\Pr(U \geq z) = 1 - F_U(z)$. This is not an easy calculation to do in general since we need to balance the value of X with the value of Y. To illustrate this we suppose that X and Y can each take values between 1 and 5 with probabilities given by Table 2.1.

Table 2.1: Probability of different values for X and Y

Value	Probability for X	Probability for Y
1	0.1	0.2
2	0.3	0.3
3	0.2	0.3
4	0.2	0.1
5	0.2	0.1

Then we can calculate the probability of $U = X + Y$ being 8 or more by considering the three possibilities, $X = 3$ and $Y = 5$; $X = 4$ and $Y \geq 4$; and $X = 5$ and $Y \geq 3$. When X and Y are independent, this gives a probability of

$$0.3 \times 0.1 + 0.2 \times (0.1 + 0.1) + 0.2 \times (0.3 + 0.1 + 0.1) = 0.17.$$

At first sight the probability here is smaller than we might expect. There is a probability of 0.4 that $X \geq 4$ and a probability of 0.2 that $Y \geq 4$. Yet the probability that $X + Y \geq 8$ is smaller than both these figures. This is a simple example of the way that adding independent random variables tends to reduce overall risk levels.

The same kind of calculation can be made for more general random variables taking integer values $1, 2, ..., M$ where we write $p_k = \Pr(X = k)$ and $q_k = \Pr(Y = k)$. Then

$$\Pr(X + Y \geq z) = p_{z-M}(q_M) + p_{z-M+1}(q_{M-1} + q_M) + ... + p_M(q_{z-M} + ... + q_M).$$

We need $2M \geq z > M$ for this formula to hold (so that the subscript $z - M$ is in the range $1, 2, ..., M$).

We can translate this formula into an integral form for continuous random variables. Suppose that X and Y are independent and the random variable X has density function f_X on $[0, M]$ and the random variable Y has CDF F_Y. As before we take $U = X + Y$. Then

$$1 - F_U(z) = \Pr(X + Y \geq z) = \int_{z-M}^{M} f_X(x)(1 - F_Y(z - x))dx. \tag{2.1}$$

Since f_X is a probability density and integrates to 1, we have

$$\int_{z-M}^{M} f_X(x)dx = 1 - \int_{0}^{z-M} f_X(x)dx = 1 - F_X(z - M).$$

So the equation (2.1) can be written

$$1 - F_U(z) = 1 - F_X(z - M) - \int_{z-M}^{M} f_X(x)F_Y(z - x)dx,$$

and so

$$F_U(z) = F_X(z - M) + \int_{z-M}^{M} f_X(x)F_Y(z - x)dx.$$

This is intuitively reasonable: the first term is the probability that X takes a value so low that $X + Y$ is guaranteed to be less than z.

The integral here is called a convolution between the functions f_X and F_Y and may be hard to solve analytically. The equivalent formula if we want to consider the sum of more than two variables is even harder. So now we want to take another approach to evaluating risk in a portfolio: we will give up something in terms of exact probabilities, but we will make a big gain in terms of ease of evaluation.

Instead of looking at specific probabilities we look instead at the spread of values as measured by the standard deviation (or the variance). Again we consider two independent random variables. Remember that when X and Y are independent we can add their variances. Specifically for independent X and Y the variance of $X + Y$ is the sum of the variances of X and Y, i.e. if we write

$$\text{Var}(X) = E(X^2) - [E(X)]^2,$$

then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

The standard deviation of a random variable X, which we write as σ_X, is just the square root of the variance Var(X) so, when X and Y are independent,

$$\sigma_{X+Y} = \sqrt{\text{Var}(X) + \text{Var}(Y)} = \sqrt{\sigma_X^2 + \sigma_Y^2}.$$

We can extend this formula to any number of random variables. The simplest case of all is where we have a set of random variables $X_1, X_2, ...X_N$ which are all independent and also all have the same standard deviation, so we can write

$$\sigma_X = \sigma_{X_1} = \sigma_{X_2} = ... = \sigma_{X_N}.$$

Then

$$\sigma_{X_1+X_2+...+X_N} = \sqrt{\sigma_{X_1}^2 + \sigma_{X_2}^2 + ...\sigma_{X_N}^2} = \sqrt{N\sigma_X^2} = \sqrt{N}\sigma_X.$$

This formula will obviously apply when all the variables have the same distribution (automatically making their standard deviations equal). For example, the individual random variables might be the demand for some product in successive weeks, when we have no reason to expect changes in average demand over time. Then the standard deviation of the total demand over, say, 10 weeks is just given by $\sqrt{10}\sigma$ where σ is the standard deviation over a single week, provided that demand in successive weeks is independent

We can express this in words as follows:

> The standard deviation of the sum of N identical independent random variables is square root N times the standard deviation of one of the random variables.

2.3.1 Portfolios with minimum variance

Now consider a situation where a portfolio is constructed from investing an amount w_i in a particular investment opportunity X_i, $i = 1, 2..., N$. We let X_i be the random variable giving the value of that investment at the end of the year. So the value of the portfolio is

$$Z = w_1 X_1 + w_2 X_2 + ... + w_N X_N.$$

We want to find out the standard deviation of the value of the portfolio, and again for simplicity we will suppose that not only are all the X_i independent, but that they also all have the same standard deviation, σ_X.

When a random variable is multiplied by w, the standard deviation is multiplied by w and the variance is multiplied by w^2. So the standard deviation of the value of the entire portfolio is

$$\begin{aligned}\sigma_Z &= \sqrt{\text{Var}(X_1) + \text{Var}(X_2) + ...\text{Var}(X_N)} \\ &= \sqrt{w_1^2\sigma_{X_1}^2 + w_2^2\sigma_{X_2}^2 + ...w_N^2\sigma_{X_N}^2} = \sigma_X\sqrt{w_1^2 + w_2^2 + ...w_N^2}.\end{aligned}$$

We may want to minimize the standard deviation of the value of the portfolio when the individual investments have different standard deviations. This could be a good idea if there is no difference between the investments in terms of their average performance. Perhaps the first thought we have is to put all of our money into best of the investment opportunities: maybe we should put everything into the single investment that has the smallest standard deviation. It will certainly be sensible to invest more of our total wealth in investments with small standard deviations, but the principle of diversification means that we can do better than the result we achieve by just investing in one opportunity.

To illustrate the principle we can consider investing a total amount W in one of two stocks which are independent of each other. We will suppose that investing \$1 in stock 1 gives a final value which is a random variable with mean μ and standard deviation σ_1. On the other hand investing \$1 in stock 2 gives the same average final value μ, but with a standard deviation σ_2. So whatever investment choice is made, the expected final value is μW. Then the problem of minimizing the standard deviation can be written as an optimization problem

$$\begin{aligned}&\text{minimize} && \sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2} \\ &\text{subject to} && w_1 + w_2 = W, \\ & && w_1 \geq 0, w_2 \geq 0.\end{aligned}$$

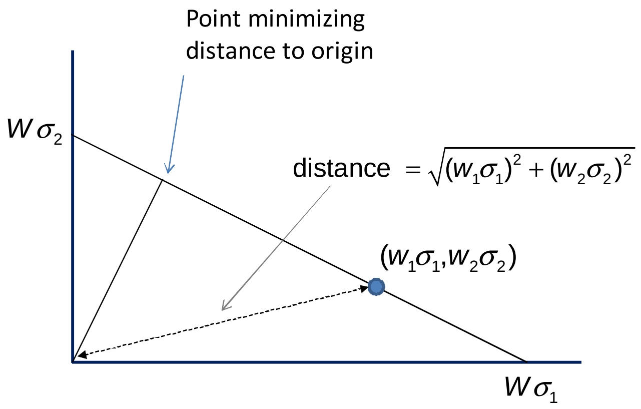

Point minimizing distance to origin
$W\sigma_2$
distance $=\sqrt{(w_1\sigma_1)^2+(w_2\sigma_2)^2}$
$(w_1\sigma_1, w_2\sigma_2)$
$W\sigma_1$

Figure 2.2 Choice of investment amounts to minimize standard deviation of return

In this case, with just two investments, the problem has a simple geometrical interpretation since the expression $\sqrt{w_1^2\sigma_1^2+w_2^2\sigma_2^2}$ gives the distance from a point with coordinates $(w_1\sigma_1, w_2\sigma_2)$ to the origin. Moreover the constraints imply that this lies somewhere on the straight line between $(W\sigma_1, 0)$ and $(0, W\sigma_2)$. These two endpoints correspond to what happens if we invest only in one or other of the two options. All this is illustrated in Figure 2.2 for a case with $\sigma_1 = 2\sigma_2$.

In this case we can find the best choice of weights simply by substituting $w_2 = W - w_1$ which means that the objective is to minimize

$$\sqrt{w_1^2 4\sigma_2^2 + (W - w_1)^2\,\sigma_2^2}.$$

We can use calculus to find the minimum of this expression, which occurs when $w_1 = W/5$ and $w_2 = 4W/5$. The resulting standard deviation is

$$\sqrt{\frac{4}{25}W^2\sigma_2^2 + \frac{16}{25}W^2\sigma_2^2} = \frac{\sqrt{20}}{5}W\sigma_2.$$

We can also ask what happens with a large number of investment opportunities, so that the number N goes to infinity. We begin by thinking about the case when all N stocks have the same standard deviation. We have already shown that the standard deviation of the overall return when all the individual stocks have standard deviation σ_X is given by

$$\sigma_X\sqrt{w_1^2 + w_2^2 + ...w_N^2}$$

This expression is minimized by splitting the investment of W equally, so that each $w_i = W/N$ giving a standard deviation of

$$\sigma_X\sqrt{(W/N)^2 + (W/N)^2 + ...(W/N)^2} = \sigma_X\frac{W}{N}\sqrt{N} = \frac{\sigma_X W}{\sqrt{N}}.$$

Hence in the case of independent investments, as the number of different investments goes to infinity and the amount invested in each gets smaller and smaller, the overall standard deviation goes to zero. And so the risk is also reduced to zero.

We can establish that the same behavior occurs in the more general situation, where stocks have different standard deviations: we let σ_i be the standard deviation of X_i. Suppose that the largest standard deviation of any of the stocks is given by $\sigma_{\max}$, so $\sigma_i \le \sigma_{\max}$, $i = 1, 2, ..., N$. Let w_i be proportional to the inverse of the standard deviation so

$$w_i = \frac{K}{\sigma_i},$$

where we take

$$K = \frac{W}{\left(\frac{1}{\sigma_1} + \frac{1}{\sigma_2} + \dots \frac{1}{\sigma_N}\right)}$$

in order that $w_1 + w_2 + ...w_N = W$. Now each $1/\sigma_i$ is at least as big as $1/\sigma_{\max}$. Thus

$$K \le \frac{W}{N(1/\sigma_{\max})} = \frac{W\sigma_{\max}}{N},$$

and the overall standard deviation is

$$\begin{aligned} \sigma_Z &= \sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + ...w_N^2\sigma_N^2} \\ &= \sqrt{K^2 + K^2 + ...K^2} \\ &\le \frac{W\sigma_{\max}}{N}\sqrt{N} = \frac{W\sigma_{\max}}{\sqrt{N}}, \end{aligned}$$

and again this expression approaches zero as N gets larger and larger.

Thus we have established that there is really no upper bound to the benefits of diversification. Provided we can find new investment opportunities which are independent of our existing portfolio, and there is no extra cost to investing in these, then we always reduce the risk by adding these extra investments into our portfolio and rebalancing accordingly.

2.3.2 Optimal portfolio theory

Now we will look at the case when different potential investments have different expected profits as well as different variances. This is the foundation of what is usually called portfolio theory. When there are differences in expected profit for individual investments, there will also be differences in the expected profit for a portfolio and so we can no longer simply find the portfolio which achieves the minimum standard deviation, we need to also consider the expected return of the portfolio. This will mean a trade-off: greater diversification will lead to less risk but will inevitably involve more of the lower return investments, and along with this a reduction in the overall expected return.

To illustrate this idea suppose that we have 3 potential investments: A, B and C. We can explore the result of putting different weights on different components within the portfolio, and end up with a set of possible trade-offs between risk and return. Suppose that the expected profit from a \$1000 investment and the standard deviations for A, B and C are as follows:

	Expected Profit	Standard Deviation
A	$R_A = 1000$	$\sigma_A = 100$
B	$R_B = 950$	$\sigma_B = 80$
C	$R_C = 900$	$\sigma_C = 85$

At first sight it may seem that investment C will not be used, since it is dominated by investment B which has a higher expected profit and at the same time a lower risk (in the sense of a less variable return). But we will see that the advantage of having one more investment in the portfolio may outweigh the fact that it is an unattractive investment.

Consider the problem of finding the least risk way of achieving a given profit, R. This can be written as an optimization problem:

$$\text{minimize} \quad w_A^2\sigma_A^2 + w_B^2\sigma_B^2 + w_C^2\sigma_C^2$$

$$\begin{aligned}\text{subject to} \quad & w_A + w_B + w_C = W, \\ & w_A R_A + w_B R_B + w_C R_C = R, \\ & w_A \geq 0, w_B \geq 0, w_C \geq 0.\end{aligned}$$

Here w_A, w_B, w_C are the sums invested, R_A, R_B, R_C are the expected profits obtained from investing $\$1000$ in the different investments, and $\sigma_A, \sigma_B, \sigma_C$ are the standard deviations of those profits. Notice that the objective we have chosen is to minimize the variance of the overall profit return rather than the standard deviation. But as the standard deviation is just the square root of the variance, whatever choice of weights minimizes one will also minimize the other.

It is possible to write down a complex formula for the optimal solution to this problem, but rather than do this we will just look at the numerical solution to the problem with the particular data given above. Figure 2.3 below shows what happens with a whole set of different random choices for the way that the total investment is split up. The picture shows quite clearly the different trade-offs that can be made: we can select a range of different overall standard deviations for the portfolio down to a minimum of around 50 at an expected profit of about 945.

We can look in more detail at the boundaries of the set of possible solutions. If we just consider a combination of two investments then, when we plot the expected profit against the standard deviation, we get a curved line joining the two points. In this example there are three such curved lines depending on which pair of the original investments we choose. These are the dashed lines in Figure 2.4. The shaded area is the set of all possible results from different portfolios. The solid line is the boundary giving the minimum standard deviation that can be achieved at any given value for overall expected profit. For example the dot at an expected profit of 960 and a standard deviation of 53.95 is the best possible at this profit level and is achieved by making $w_A = 0.3924$, $w_B = 0.4152$ and $w_C = 0.1924$.

2.3.3 When risk follows a normal distribution

Our discussion of portfolio risk so far has simply looked at the standard deviations for the overall profit (obtained from the sum of random variables). There is a critical assumption

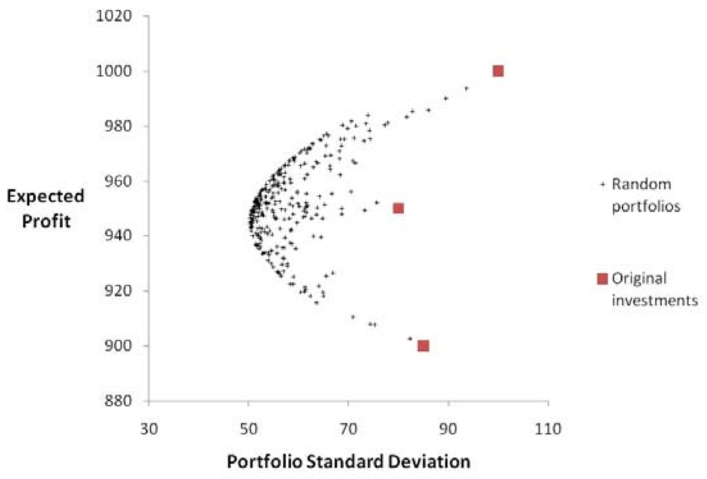


Figure 2.3 Profit versus standard deviation for random portfolios

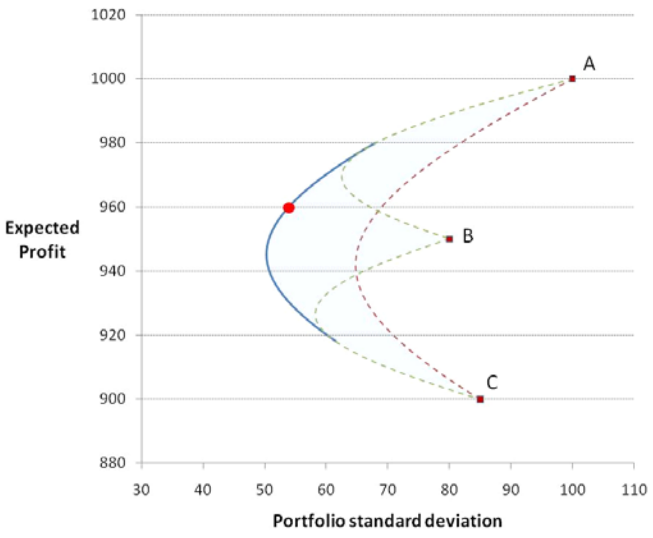


Figure 2.4 Boundary of the region that can be achieved through a portfolio of three investments

about independence of the different investments, but no assumption on the form of the distributions. When the distribution is known then we can say more about the risks involved, and in particular we can calculate the probability of getting a result worse than some given benchmark level.

The most important distribution to look at is the normal distribution. Its importance stems from the way that it approximates the result of adding together a number of different random variables whatever their original distributions. This is the Central Limit Theorem discussed in the Appendix: Tutorial on Probability Theory. At the same time a normal distribution is easy to work with because the sum of two or more random variables each with a normal distribution also has a normal distribution.

If the distribution of profit follows a normal distribution we can use tables or a spreadsheet to calculate any of the probabilities we might need. To illustrate this we consider an example where there are two investments, both having a normal distribution for the profits after 1 year. The first has an expected profit of \$1000 with a standard deviation of \$400 and the second has an expected profit of \$600 with a standard deviation of \$200. If we hold both these investments, what is the probability that we will lose money? Without information on the distribution of the profit this probability is not determined, but with the knowledge that the profits follow a normal distribution it becomes easy to answer the question. The sum of the two returns is also a normal distribution with mean of $\$1000 + \$600 = \$1600$ and, given that they are independent, the standard deviation is

$$\sqrt{400^2 + 200^2} = \sqrt{200,000} = 447.21.$$

The probability of getting a value less than $\$0$ can be obtained from tables (it's the probability of being more than z standard deviations from the mean where $z = 1600/447.21 = 3.5777$) or more simply using the NORMDIST function in a spreadsheet. Specifically we have NORMDIST$(0, 1600, 447.21, 1) = 0.00017329$.

2.4 The impact of correlation

We have discussed the way in which risk is reduced by diversification, but it has to be genuine diversification. Things can go wrong if there is too close a connection between different investments. Figure 2.5 shows the share price (sourced from Yahoo) for 3 shares for a six month period in 2011. All of these companies are associated with wind power and are quoted on the Madrid stock exchange. It is not surprising that there is a close connection between their share prices. Looking at the graphs show that they all had a short term peak at the beginning of July 2011 and at the end of that month and at the beginning of August fell substantially. The overall volatility of a share portfolio invested equally in these three stocks would be quite similar to that of investing entirely in a single stock, so that diversification would bring little benefit.In fact the main correlation here does not relate to wind power but instead to overall movements in the Madrid stock exchange as shown by Figure 2.6 which charts the behavior of the IGBM (Madrid Stock Exchange General Index) over the same period. This is particularly true for the large company Iberdrola with a market capitalisation of over 30 billion Euro (Acciona has a market cap of 4 billion Euro, and Gamesa Corp. Tecnologica is much smaller at around 950 million Euro). At first sight you might almost think you were looking at the same chart, but they are not quite the same. In fact Iberdrola is about 8.7% of the total in this index.

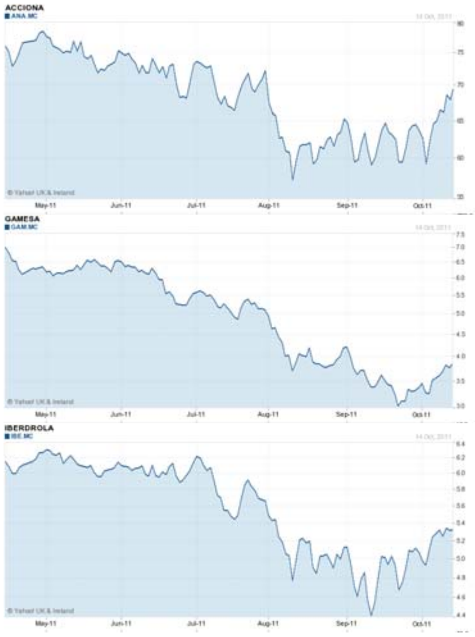


Figure 2.5 Correlation between share prices: ACCIONA, GAMESA and IBERDROLA

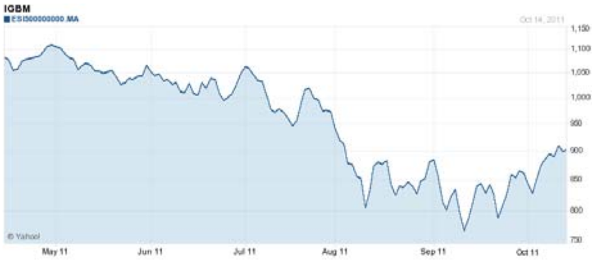


Figure 2.6 The Madrid Stock Exchange Index (IGBM)

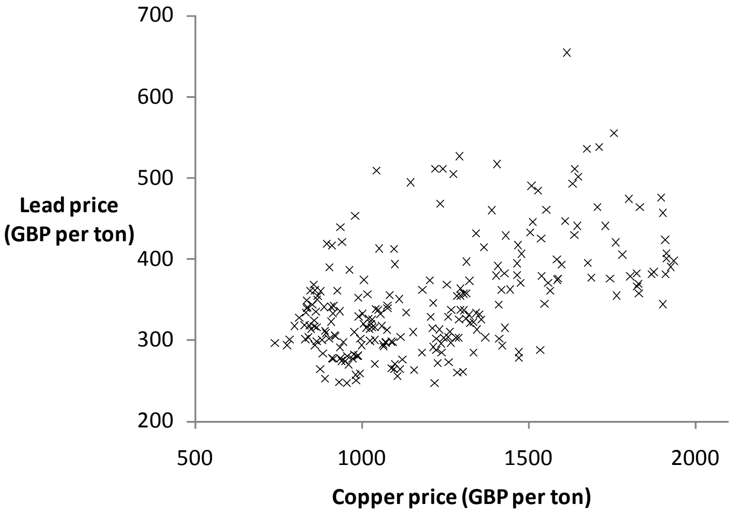


Figure 2.7 Lead and copper prices from 1980 to 2000

In this section we want to look at the way that correlation between different random variables will effect the behavior of their sum. This is a topic of great importance when assessing risk in practice but it is also best handled through some more complex mathematical tools than we have used in the rest of this book (particularly matrix algebra). So we will give a relatively broad brush treatment here.

Figure 2.7 shows the monthly average prices over a 20 year period for lead and copper on the London metal exchange. It happens that the prices in December 1999 were not much greater than in January 1980 for these two commodities.Suppose that we know that in a year's time we will need to purchase an amount of both lead and copper. The risk is related to the total purchase price. Following our previous argument if the two commodity prices are independent then the risk associated with the combined purchase is reduced, since high prices for one commodity may well be balanced out by low prices for the other. But the scatter plot shows that there is quite a high degree of correlation between these variables, and so the beneficial effect of diversification is reduced.

To make the discussion specific suppose that we are interested in a combined price for 1 ton of copper and 2 tons of lead (which is cheaper). A good way to think about this is to recognize that the points which have the same value of $X + 2Y$ all lie on a straight line drawn in the (X, Y) plane. So if we looked at the monthly price for the purchase of 1 ton of copper and 2 tons of lead then the cost is the same at 1900 GBP if either (A) copper is 1500 GBP and lead is 200 GBP, or (B) copper is 500 GBP and lead is 700 GBP. And the same is true for any point on the straight line between these two points.

Figure 2.8 shows dashed lines for sets of price combinations that lead to the same overall price for this purchase. The effect is to project down onto the solid line all the different monthly combinations. The choice of solid line does not matter here: usually a projection

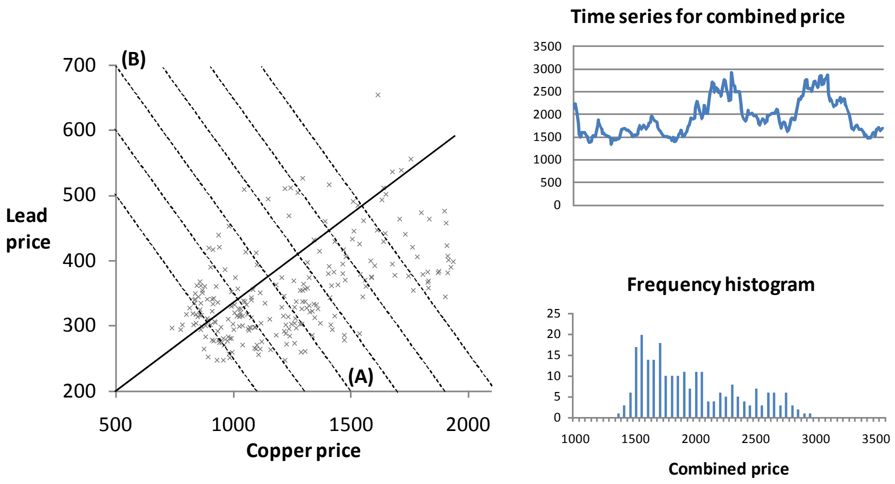


Figure 2.8 Price distribution for 1 ton of copper and 2 tons of lead

means a mapping onto a line which is at right angles to the dashed contour lines of equal overall price, but in this case some other choice of straight line just leads to a linear scaling of the results). The right hand side of the figure shows how the price of 1 ton of copper and 2 tons of lead varies over time together with the frequency histogram for these prices.

Notice how spread out this distribution is: the variance is high and the distribution itself does not have the nice shape of a normal distribution. It is the covariance that measures the extent to which these two data series are correlated. Positive values of the covariance correspond to a positive correlation, and the covariance will be zero if the two variables are independent. Remember that the covariance between random variables X and Y is

$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y).$$

If we are interested in understanding the properties of a portfolio with weights w_X and w_Y then we can use the Covariance between X and Y to get the variance of the portfolio (for more details about the formula here see the Appendix: Tutorial on Probability Theory).

$$\text{Var}(w_X X + w_Y Y) = w_X^2 \text{Var}(X) + w_Y^2 \text{Var}(Y) + 2w_X w_Y \text{Cov}(X,Y).$$

We can see how this works out for the lead and copper price example. The copper price has mean $\$1231.16$ and standard deviation $\$312.40$ (implying a variance of 97590.80). The lead price has mean $\$350.99$ and standard deviation $\$69.65$ (implying a variance of 4851.55). The two sets of prices are positively correlated with a covariance 11281.00. Thus the formula implies that the variance of the combination 1 ton of copper and 2 tons of lead is

$$97,590.80 + 4 \times 4,851.55 + 4 \times 11,281.00 = 162,121.00,$$

giving a standard deviation of $\$402.64$. If the two prices had been independent then the third term does not appear and the overall standard deviation would have been $\$342.05$.

2.4.1 Minimum variance portfolio with covariance

We can switch focus from thinking about the distribution of values for a particular combination purchase to deciding how we might invest if we had a choice of portfolios over assets with a correlated behavior. Suppose that the initial prices are X_0 and Y_0 and we will sell after 1 year at prices which are X and Y then how should we split our available investment sum W? The decision here will depend on the relationship between the purchase prices X_0 and Y_0 and the expected values for X and Y. In order to eliminate the question of different relative returns let us assume that the mean value of X is a certain multiple of X_0, and the mean value of Y is the same multiple of Y_0, so $\mu_X = kX_0$ and $\mu_Y = kY_0$. Thus the expected result from this investment after one year is that the initial investment is multiplied by k, no matter how we split the investment between X and Y. In this situation it makes sense to invest in a way that minimizes the variance of the return. Given a purchase of w_X units of asset X then we have a remaining $W - w_X\mu_X$ to invest. This means that we can purchase w_Y units of asset Y where

$$w_Y = (W - w_X\mu_X)/\mu_Y.$$

Then the variance of the portfolio is

$$w_X^2\text{Var}(X) + (W - w_X\mu_X)^2\text{Var}(Y)/\mu_Y^2 + 2w_X(W - w_X\mu_X)\text{Cov}(X,Y)/\mu_Y$$

which (using calculus) we can show is minimized when

$$\begin{aligned}2w_X\text{Var}(X) - 2\mu_X(W - w_X\mu_X)\text{Var}(Y)/\mu_Y^2& \\ + (-2\mu_X w_X + 2(W - w_X\mu_X))\,\text{Cov}(X,Y)/\mu_Y &= 0.\end{aligned}$$

Simplifying we get

$$w_X(X) - \mu_X(W - w_X\mu_X)\text{Var}(Y)/\mu_Y^2 + (W - 2w_X\mu_X)\text{Cov}(X,Y)/\mu_Y = 0,$$

and we can solve for w_X:

$$w_X = \left(\frac{W}{\mu_Y}\right)\frac{(\mu_X/\mu_Y)\text{Var}(Y) - \text{Cov}(X,Y)}{\text{Var}(X) + \text{Var}(Y)(\mu_X^2/\mu_Y^2) - 2\text{Cov}(X,Y)(\mu_X/\mu_Y)}.$$

We check what happens for the example of lead and copper prices. Suppose that we have $\$1,000$ to invest, then we get that the weight of the portfolio in copper is (writing α for $\mu_{\text{copper}}/\mu_{\text{lead}} = 3.51$)

$$\begin{aligned}w_{\text{copper}} &= \frac{W}{\mu_{\text{lead}}}\frac{\text{Var(lead)}\alpha - \text{Cov(copper,lead)}}{\text{Var(copper)} + \text{Var(lead)}\alpha^2 - 2\text{Cov(copper,lead)}\alpha} \\ &= \frac{1000}{350.99}\frac{4851.55 \times 3.51 - 11281.00}{97590.80 + 4851.55 \times (3.51)^2 - 2 \times 11281.00 \times 3.51} = 0.2095, \\ w_{\text{lead}} &= \frac{W - w_{\text{copper}}\mu_{\text{copper}}}{\mu_{\text{lead}}} = \frac{1000 - 0.2095 \times 1231.16}{350.99} = 2.1142\end{aligned}$$

with expenditure of $0.2095 \times 1231.16 = \$257.93$ on copper and $2.1142 \times 350.99 = \$742.06$ on lead (this is one cent less than \$1000 which has disappeared in the rounding of these calculations).

The result of this calculation may give a negative value for one of the weights w_X, w_Y. This would imply a benefit from selling one commodity in order to buy more of the other. The process to do this may well be available in the market place: in the language of finance this amounts to going short on one commodity and long on another. However we will not pursue the idea here.

We have seen how a positive correlation between two investments reduces the diversification benefits on risk if both investments are held. Exactly the same thing takes place with more than two investments.

2.4.2 The maximum of variables that are positively correlated

Now we consider the other scenario in which the maximum value of two random variables is of interest, and we ask how a positive correlation will impact on this measure. The probability of both X and Y being less than z is given by $F_X(z) \times F_Y(z)$ if the two variables are independent. The probability will be more than this if they are positively correlated since a low value for one tends to happen at the same time as a low value for the other. Hence if we define $U = \max(X, Y)$ then

$$\begin{aligned} F_U(z) &= \Pr(\max(X, Y) < z) \\ &> F_X(z) \times F_Y(z). \end{aligned}$$

We can look at this in the other direction and say that the probability of U being more than a given value is reduced when X and Y are positively correlated.

To see what this looks like for a specific example we consider the behavior of zinc and copper prices shown in Figure 2.9. Again this shows monthly prices over the 20 year period starting in January 1980. There are 51 occasions out of 240 in which the copper price is greater than \$1500 per ton and 24 occasions in which the zinc price is greater. So (approximately) $F_{\text{copper}}(1500) = 189/240 = 0.7875$ and $F_{\text{zinc}}(1500) = 216/240 = 0.9$ If these were independent then $F_U(1500) = 0.9 \times 0.7875 = 0.709$ and we would expect $0.709 \times 240 = 170$ occasions when the maximum of the two prices is below 1500 and 70 when it is greater than 1500. In fact there are only three occasions when the price of zinc is higher than 1500 and copper is not, meaning a total of 54 occasions when the maximum of the two is greater than 1500. Thus in this case the correlation between the two prices has *reduced* the risk that the maximum is very high, which is the opposite of what happens when looking at a sum (or average) of prices.

2.4.3 Multivariate normal

If different quantities are not independent then detailed calculation of the probability of a high value, either for the sum or the maximum, will not be possible unless we know more than just their covariance. One model we can look at is the *multivariate normal*. Just as we can calculate exact probabilities from the mean and standard deviations alone when the underlying risk distribution is normal, we can do the same thing when the combined

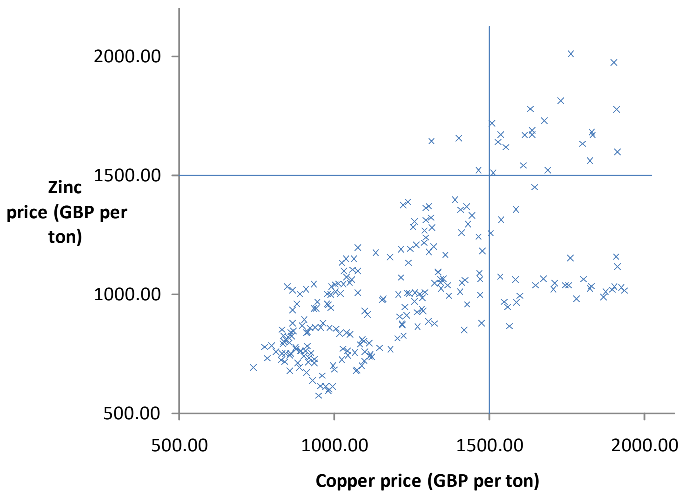


Figure 2.9 How does correlation effect the risk for the maximum of zinc and copper prices?

distribution is multivariate normal provided we know the means, standard deviations and covariances. A more detailed discussion of this would involve looking at N dimensional problems but all the important ideas can be understood by just looking at 2 dimensional or *bivariate* distributions, and so we concentrate on this case which also means we can easily plot what is going on. A multivariate normal distribution is shown in Figures 2.10 and 2.11. These show the density function giving the relative probabilities of different combinations of the two variables X and Y, together with the contours of this. The rules for a multivariate density function are just the same as for a univariate one - to calculate the probability of the (X, Y) pair being in any region of the (X, Y) plane, we just integrate the density function over that region.

The formula for a two dimensional multivariate normal density function is

$$f(x,y) = K \exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right)\right),$$

where

$$K = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \quad \text{and} \quad \rho = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}$$

(K is a normalizing constant and ρ is the correlation between X and Y). The Figures show an example with $\mu_X = \mu_Y = 10$, $\sigma_X = 3$, $\sigma_Y = 2$ and $\rho = 0.5$. The contours in Figure 2.11 are all ellipses.

One of the most important properties of a multivariate normal is that any linear combination of the variables also has a multivariate normal distribution. It is easy to imagine that any vertical straight slice through the density function of Figure 2.10 would give a bell

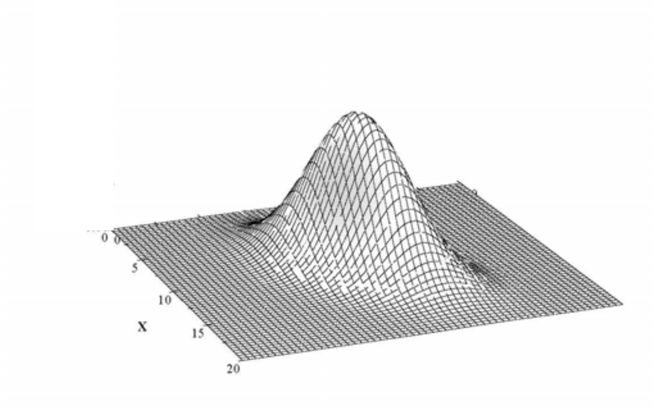


Figure 2.10 Density function for a bivariate normal distribution

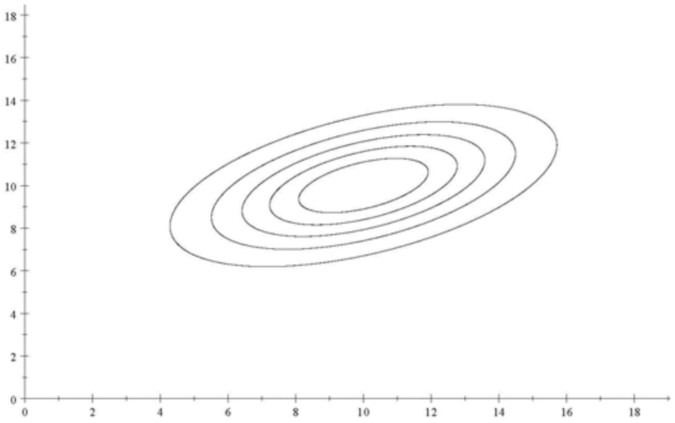


Figure 2.11 Contours of the density of the bivariate normal distribution

Length	Width
20.9	8.8
21.6	9
21.9	9.3
22	9.2
22.2	8.9
22.4	8.4
22.4	8.6
22.5	9.7
22.7	8.6
22.8	8.9
23.1	8.9
23.2	9.8
23.4	8.8
23.5	9.5
23.7	9.7
24.1	9.6
24.1	9.1
25	9.8
25.1	9.4
25.5	9.8

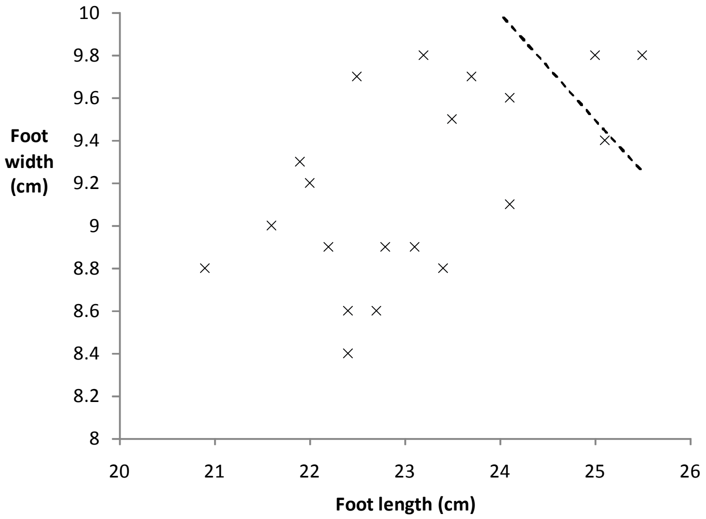


Figure 2.12 Data on foot sizes for 20 fourth grade boys

shaped normal distribution curve. But this property is saying something a little different since it involves the kind of projection down onto a straight line that we described in Figure 2.8.

As an example consider a children's shoe manufacturer interested in the distribution of childrens' foot sizes (width and length). Figure 2.12 below shows data collected for 20 fourth grade boys (see kidsfeet.dat at http://www.amstat.org/publications/jse/jse_data_archive.htm) together with a scatter plot of this data.We find that the mean length is 23.105 (sample standard deviation = 1.217) and the mean width is 9.19 (sample standard deviation = 0.452). The covariance is 0.299. If we fit a multivariate normal then we expect that the distribution of any combination of height and width is also normal. For example (Length)$+2\times$(Width) should have a normal distribution with mean $23.105 + 2 \times 9.19 = 41.485$ and the variance should be

$$1.217^2 + 4 \times 0.452^2 + 2 \times 1.217 \times 2 \times 0.452 \times 0.299 = 2.956,$$

giving a standard deviation of $\sqrt{2.956} = 1.719$. From this we can calculate the probability of getting different ranges of values for this linear combination. For example suppose that we wish to estimate the probability that an individual has length + twice width value greater than 44cm. This corresponds to the dashed line in Figure . The z value is $(44 - 41.485)/1.719 = 1.463$. Under the normal assumption this would imply a probability of $1 - \Phi(1.463) = 1 - 0.9283 = 0.0717$ of achieving this value. Given 20 observations this would lead us to expect around 1.4 with this characteristic. In fact we observe 2 individuals - very much in line with our prediction.

The value of an explicit model is that it can help us to make predictions about the likelihood of events we have only occasionally (or never) observed. For example in the childrens' shoes example we can estimate the probability of having an individual where the composite score is more than 45.5 which does not occur in this group of 20. However we should be cautious in extrapolating beyond the data we have observed, and it is possible that the approximation of a multivariate normal fails as a predictor for more extreme results. We shall say more in Chapter 4 about modelling the tails of distributions.

2.5 Using copulas to model multivariate distributions

On many occasions the multivariate normal is not a good approximation for the dependence between two or more variables and we need to look for a more flexible model. A good choice is to use copula models, which are a relatively modern development in probability theory. We will describe a two dimensional copula describing the joint behavior of variables X and Y (but the same ideas can be applied to multivariate models with 3 or more variables). The idea is to look at the distribution over values (x, y) expressed in terms of the positions of x and y within the distributions for X and Y respectively. This gives a way of distinguishing between what is happening as a result of the distribution of the underlying variables X and Y, and what is happening as a result of the dependence of one variable on the other. This means that we can, if we wish, apply the same copula model for different distributions of the underlying variables.

A *copula density* $c(u_1, u_2)$ is a density defined on the unit square $0 \leq u_1 \leq 1, 0 \leq u_2 \leq 1$, with the property that the resulting (marginal) distribution for u_1 is uniform on $[0, 1]$, and the distribution for u_2 is also uniform on $[0, 1]$. We can write these conditions as

$$\int_0^1 c(u, u_2)du = 1, \text{ for each } u_2 \text{ in } [0, 1],$$

$$\int_0^1 c(u_1, u)du = 1, \text{ for each } u_1 \text{ in } [0, 1].$$

The simplest way to make these conditions hold is to make the copula density uniform on the unit square, so that $c(u_1, u_2) = 1$ for every u_1 and u_2.

Before we talk about different examples of copula densities and how they relate to different kinds of dependence we need to show how to convert a copula density and the information on the underlying distributions into a distribution for the multivariate distribution as a whole. To do this we will change from densities (which are easier to visualize) into cumulative distribution functions (which are easier to work with). So we define a *copula* (really a cumulative copula distribution) as the distribution function $C(u_1, u_2)$ obtained from the copula density function $c(u_1, u_2)$. So $C(v_1, v_2)$ is the probability that both $u_1 \leq v_1$and $u_2 \leq v_2$ when (u_1, u_2) has density function c, or more formally

$$C(v_1, v_2) = \int_0^{v_1} \left(\int_0^{v_2} c(u_1, u_2)du_1 \right) du_2.$$

Then the cumulative distribution for the multivariate distribution is obtained from the copula C and the underlying distribution functions F_X and F_Y for X and Y through the fundamental copula equation:

$$\Pr(X \leq x \text{ and } Y \leq y) = C(F_X(x), F_Y(y)). \qquad (2.2)$$

In words we take x and y and convert them into quantiles for the appropriate distributions, $F(x)$ and $F(y)$, and then use the copula function to determine the probability of being in the rectangle with (x, y) at its top right corner.

To illustrate how this works let's go back to the uniform copula density: $c(u_1, u_2) = 1$. This can be converted into a copula

$$C(v_1, v_2) = \int_0^{v_1} \left(\int_0^{v_2} du_1 \right) du_2 = v_1 v_2,$$

which is in product form. Then, from (2.2), we have

$$\Pr(X \leq x \text{ and } Y \leq y) = F_X(x) F_Y(y) = \Pr(X \leq x) \Pr(Y \leq y).$$

This is exactly the formula for the probability if the two variables are independent. From this observation we get the result that a uniform copula density (or product form copula) is equivalent to the variables being independent.

We can also move from a copula back to its density by taking derivatives; more precisely we have

$$c(v_1, v_2) = \partial^2 C(v_i, v_2) / \partial v_1 \partial v_2.$$

Taking derivatives with respect to x and y of (2.2) we obtain the formula

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} C(F_X(x), F_Y(y)) = c(F_X(x), F_Y(y)) f_X(x) f_Y(y),$$

where f is the joint density over X and Y and f_X, f_Y are the individual density functions. We can rewrite this as

$$c(F_X(x), F_Y(y)) = \frac{f(x, y)}{f_X(x) f_Y(y)}. \tag{2.3}$$

The formula here shows how the copula density is obtained from the usual density function f by squeezing it up at places where X and Y have low probabilities and spreading it out at places where X and Y have high probabilities.

Using copulas rather than copula densities makes it easier to handle some of the equations but we need to check two properties we require of the copula in order to match the properties of the copula density:

- **The rectangle inequality:** If $1 \geq b_1 \geq a_1 \geq 0$ and $1 \geq b_2 \geq a_2 \geq 0$ then the probability of being in the rectangle $[a_1, b_1] \times [a_2, b_2]$ can be obtained by looking at the right combination of the four possible rectangles with a corner at 0. For c to be a density this must be non-negative and we derive the inequality

$$C(a_1, a_2) + C(b_1, b_2) - C(a_1, b_2) - C(b_1, a_2) \geq 0. \tag{2.4}$$

- **Uniform marginal distribution:** Notice that $C(v, 1)$ is simply the probability that u_1 is less than v, and similarly $C(1, v)$ is simply the probability that u_2 is less than v. So the condition that the distributions of both u_1 and u_2 are uniform becomes: $C(1, v) = C(v, 1) = v$ for all $0 \leq v \leq 1$.

These two conditions are enough to show that the copula $C(u_1, u_2)$ is increasing (or at least doesn't decrease) as either u_1 or u_2 increases (see exercise XX), but we cannot make the implication the other way round and deduce the rectangle inequality from the fact that C is increasing in both variables.

The important thing about copulas is that they can indicate when dependence may increase. They are flexible enough to deal differently with the tails of the distribution than with the central parts. A good question to ask when dealing with multivariate data that includes some correlation is whether the correlation will be greater or less at extreme values. This may be a question which is easier to ask than to answer, but giving it attention will at least ensure that we avoid some important risk management pitfalls. For example if two variables are weakly correlated with a covariance close to 0 then we might be tempted to conclude that the diversification effect will imply a substantially reduced risk for their average value (e.g. a portfolio split equally between them) than either on its own. This is usually a correct deduction, but if the correlation between them is very low *except* at the point when they take extremely large values when they are closely correlated then the risk of their average becoming very large will be close to the risk of an individual variable becoming large, and the diversification benefits we might expect will not occur.

The problem with making statements about correlation in the tails is that these are events we rarely see, and so there is unlikely to be much historical evidence to guide us. It is worth thinking about the underlying reasons for very large values in one or other of the variables. For example consider the relationship between two shares both traded in the New York stock exchange, lets say Amazon and Ford. They are both affected by market sentiment and by general economic conditions, and this will lead to correlation between their prices. But will this correlation be amplified when one of them falls sharply? For example does knowing that Amazon's share price has fallen by 30% in one day give correspondingly more information about Ford's share price than knowing that Amazon's share price has fallen by 15%? Perhaps a sudden price drop of 30% in a day is most likely to arise from very specific factors (like an earnings announcement) and less likely to relate to general factors (like a market collapse) that would tend to lead to a drop in Ford's share price too. Whichever way round this is, copulas give us a way to model what is going on.

An important copula is that which represents the dependence behavior occurring in a multivariate normal, and this is called the *Gaussian copula*. The copula density is related to the density for the multivariate normal but each component is scaled to ensure that the marginals are uniform. We begin by doing the calculations working with the (cumulative) copula functions. From the fundamental copula equation and the density function for the multivariate normal with correlation ρ and with each variable having zero mean and standard deviation of 1, we have

$$\begin{aligned}
&C(F_X(x), F_Y(y)) \\
&= C(\Phi(x), \Phi(y)) = \Pr(X \le x \text{ and } Y \le y) \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{x} \left(\int_{-\infty}^{y} \exp\left(-\frac{1}{2(1-\rho^2)} \left(s_1^2 + s_2^2 - 2\rho s_1 s_2 \right) \right) ds_1 \right) ds_2.
\end{aligned}$$

Here we have used the usual notation in which $\Phi(x)$ is written for the cumulative normal distribution function with mean zero and standard deviation 1. The integrand comes from the multivariate normal with $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$. To obtain a formula for $C(v_1, v_2)$ we want to substitute values of x and y for which $F_X(x) = v_1$ and $F_Y(y) = v_2$.

So we set $x = \Phi^{-1}(v_1)$ and $y = \Phi^{-1}(v_2)$ to obtain

$$\begin{aligned}&C(v_1, v_2)\\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\Phi^{-1}(v_1)} \left(\int_{-\infty}^{\Phi^{-1}(v_2)} \exp\left(-\frac{1}{2(1-\rho^2)} \left(s_1^2 + s_2^2 - 2\rho s_1 s_2\right) \right) ds_1 \right) ds_2.\end{aligned}$$

The equivalent formula for the copula density from (2.3) is

$$\begin{aligned} c(F_X(x), F_Y(y)) &= \frac{\frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(x^2+y^2-2\rho xy\right)\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)} \\ &= \frac{1}{\sqrt{1-\rho^2}} \exp\left(\frac{1}{2}x^2 + \frac{1}{2}y^2 - \frac{1}{2(1-\rho^2)}\left(x^2+y^2-2\rho xy\right) \right) \\ &= \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left(\rho^2 x^2 + \rho^2 y^2 - 2\rho xy\right) \right). \end{aligned}$$

And hence

$$\begin{aligned}&c(v_1, v_2)\\ &= \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left(\rho^2\Phi^{-1}(v_1)^2 + \rho^2\Phi^{-1}(v_2)^2 - 2\rho\Phi^{-1}(v_1)\Phi^{-1}(v_2)\right)\right).\end{aligned}$$

There are versions of this formula for a higher number of variables. It has become somewhat infamous since the use of Gaussian copulas as a standard approach in measuring risk was blamed in some quarters for the failures of Wall Street “quants” in predicting the high systemic risks in CDOs that in turn kicked of the financial crisis of 2008. In 2009 the Financial Times carried an article discussing the use of Gaussian copula models titled “The formula that felled Wall Street” (The Financial Times, Jones, S. April 24 2009.)

To help in understanding the Gaussian copula formula we really need to plot it (or more usefully we need to plot the copula density). This is simply a function defined over the unit square and is shown in Figure 2.13.

In this figure the corners $(0, 0)$ and $(1, 1)$ represent what happens to the relationship when both variables are very large and positive or very large and negative. The existence of correlation pushes these corners up and the opposite corners $(0, 1)$ and $(1, 0)$ are pushed down. Remember there is a restriction on the densities integrating to 1 along a line in which one or other variable is held constant, so the lifting of one corner can be seen as balancing the pushing down of an adjacent corner.$\rho = 0$ the multivariate normal has circular contours, and the variables are independent - as we said earlier this means a flat copula density at value 1, (which serves as a reminder that the copula abstracts away from the distributions of the underlying variables.

There are many different copula formula that have been proposed. One example is the Clayton copula given, for two variables, by the formula

$$C_\theta(v_1, v_2) = \left(\frac{1}{v_1^\theta} + \frac{1}{v_2^\theta} - 1 \right)^{-\frac{1}{\theta}}$$

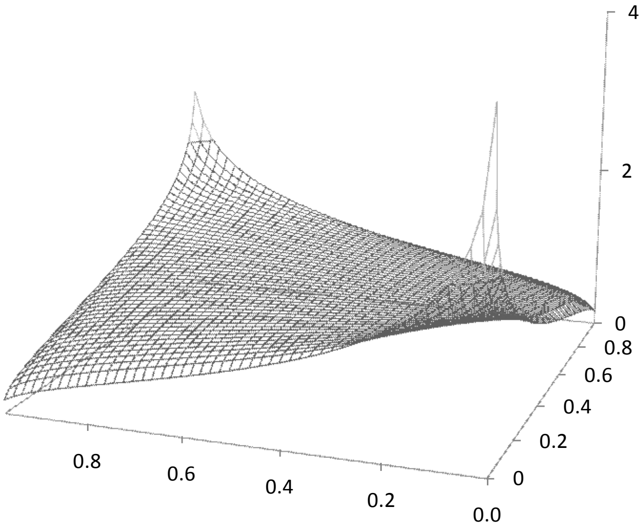


Figure 2.13 Gaussian copula density with $\rho = 0.25$

where θ is a parameter that can take any value strictly greater than zero (we can also make this formula work when θ is strictly less than zero and greater than -1). Figure 2.14 shows what this copula looks like when $\theta = 2$. The copula density is then

$$c_\theta(v_1, v_2) = \frac{3}{v_1^3 v_2^3}\left(\frac{1}{v_1^2} + \frac{1}{v_2^2} - 1\right)^{-\frac{5}{2}}.$$

In the figure the peak at $(0, 0)$ shows a very tight correlation between the values that occur in the lower tails of the two underlying distributions. In fact this peak has been cut off at a value of 6 for the purposes of drawing it.

There is another approach we can take to the way that variables exhibit dependence at extreme values and this is to directly measure *tail dependence*. We know that independence means that information on one variable does not convey anything about the other. So we could say that the probability of X being in the upper decile (which is $F_X(X) > 0.9$) is unchanged (at exactly 0.1) even if we also know that Y is also in its upper decile. This property of independence can be written (using Bayes formula) as

$$\frac{\Pr(X > F_X^{-1}(0.9) \text{ and } Y > F_Y^{-1}(0.9))}{\Pr(Y > F_Y^{-1}(0.9))} = \Pr(X > F_X^{-1}(0.9).$$

This motivates us in taking the 0.9 in this expression and letting it tend to 1. We define a coefficient of upper tail dependence through

$$\lambda_u = \lim_{\alpha \to 1} \Pr(X > F_X^{-1}(\alpha) \mid Y > F_X^{-1}(\alpha)) = \lim_{\alpha \to 1} \frac{\Pr(X > F_X^{-1}(\alpha) \text{ and } Y > F_X^{-1}(\alpha))}{(1 - \alpha)}$$

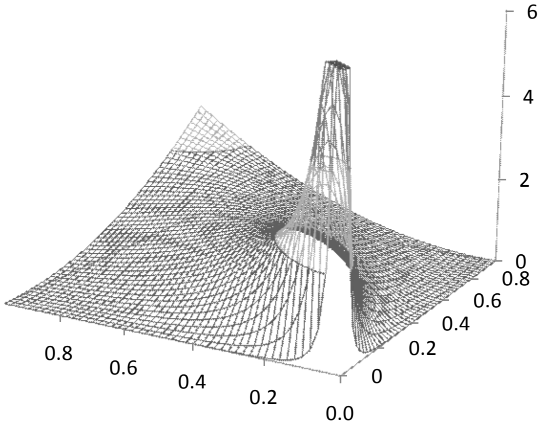


Figure 2.14 Clayton copula density for $\theta = 2$

and a corresponding coefficient of lower tail dependence as

$$\lambda_l = \lim_{\alpha \to 1} \Pr(X \leq F_X^{-1}(\alpha) \mid Y \leq F_X^{-1}(\alpha)) = \lim_{\alpha \to 0} \frac{\Pr(X \leq F_X^{-1}(\alpha) \text{ and } Y \leq F_X^{-1}(\alpha))}{\alpha}.$$

We say that X and Y have *upper tail dependence* if $\lambda_u > 0$, and that there is no upper tail dependence if $\lambda_u = 0$. The same definitions work for *lower tail dependence* using λ_l instead of λ_u. We can convert this into a statement about the copula functions since $\lambda_l = \lim_{\alpha \to 0}(C(\alpha, \alpha)/\alpha)$.

The existence of tail dependence is quite a strong property, and certainly is much stronger than saying that at high (or low) values the variables fail an independence test. If there is tail dependence then the copula density will go to infinity at the appropriate corner. We will not try to prove this in a formal sense, but observe that with lower tail dependence in the limit of small α then $C(\alpha, \alpha) \simeq \alpha\lambda_l$. But at this limit $C(\alpha, \alpha) \simeq \alpha^2 c(\alpha, \alpha)$ so we have $c(\alpha, \alpha) \simeq \lambda_l/\alpha$. As α goes to zero the left hand side approaches $c(0, 0)$ and the right hand side goes to infinity unless $\lambda_l = 0$. A similar argument can be made for upper tail dependence as well.

So in two dimensional cases we can look at a plot of the copula density and get a good idea of whether there is tail dependence. Figure 2.13 suggests that there is neither upper nor lower tail dependence for the Gaussian copula for this value of ρ, and from Figure 2.14 we see there is no upper tail dependence for the Clayton copula. However, there seems to be a lower tail dependence for this copula, and we can confirm this since

$$\begin{aligned}C(\alpha,\alpha)/\alpha &= (1/\alpha)\left(\frac{1}{\alpha^\theta}+\frac{1}{\alpha^\theta}-1\right)^{-\frac{1}{\theta}}\\ &= \left(\alpha^\theta\right)^{-\frac{1}{\theta}}\left(\frac{2}{\alpha^\theta}-1\right)^{-\frac{1}{\theta}}\\ &= \left(2-\alpha^\theta\right)^{-\frac{1}{\theta}} = \frac{1}{\left(2-\alpha^\theta\right)^{\frac{1}{\theta}}}.\end{aligned}$$

Thus

$$\lambda_l = \lim_{\alpha\to 0}\frac{1}{\left(2-\alpha^\theta\right)^{\frac{1}{\theta}}} = \frac{1}{2^{\frac{1}{\theta}}} > 0.$$

Now we return to our opening discussion about the failure of diversification to protect investors during the crisis of 2008. We can view this as a sudden increase in correlation as losses mounted. The implication is that the covariance dependence between different asset prices may not be symmetric between gains and losses. If Wall Street moves up by a large amount the corresponding effect on the London stock market may be less than if Wall Street falls sharply. There is good empirical evidence to support exactly this behavior. A cynic might say that fear is an even more contagious emotion than greed, but whatever the mechanism it is clear from a number of empirical studies that these asymmetries exist. The copula approach is a good way to make this explicit, for example through the use of a model for losses with upper tail dependence, but no lower tail dependence.

Notes

The historical introduction to probability has drawn extensively from the book ‘Against the Gods’ by Peter Bernstein. This is the source for the Paccioli quote at the start of this chapter, and Bernstein gives a fascinating description of the entire history of scientists’ struggle to find the right framework to deal with risk. The quote from Jakob Bernoulli is taken from Oscar Sheynin’s translation of his work: *The Art of Conjecturing; Part Four Showing the Use and Application of the Previous Doctrine to Civil, Moral and Economic Affairs*.

Our discussion of optimal portfolio theory in this chapter is very brief. An excellent and much more comprehensive treatment is given by Luenberger (1998). The theory was originally developed by Harry Markowitz and first published in a 1952 paper; in 1990 Markowitz received a Nobel prize in Economics for his work.

A straightforward introduction to copulas is given by Schmidt (2006). Also the book by McNeil et al. (2005) gives a thorough discussion of the topic. The observations we make about the increase in correlation when markets move downwards are well known, for example see Ang and Chen 2002, or Chua et al. 2009.

References

Andrew Ang and Joseph Chen, Asymmetric correlations of equity portfolios, *Journal of Financial Economics* 63 (2002) 443–494.

Jakob Bernoulli, *Ars Conjectandi*, translated by Oscar Sheynin, NG Verlag, 2005.

Peter Bernstein, *Against the Gods*, Wiley,1996.

David Chua, Mark Kritzman and Sebastian Page, The myth of diversification, *Journal of Portfolio Management* Fall 2009, 26–35.

David Luenberger, *Investment Science*, Oxford University Press 1998.

Alexander McNeil, Rudiger Frey and Paul Embrechts, *Quantitative Risk Management*, Princeton University Press, 2005.

Thorsten Schmidt, Coping with copulas, in *Copulas: From theory to application in finance* Ed. Jorn Rank, Bloomberg Financial, 2006.

Exercises

2.1. (Problem of the points)

Calculate a fair division of the stake in the 'Problem of the points' described in section 2.1 if we assume that the current record of success of A against B is an accurate forecast of the probability of winning future games.

2.2. (Late for the class)

James needs to get to his class on time which means arriving at the University by 10 am. The option is to take the number 12 bus which takes 40 minutes or the number 15 which takes 30 minutes or the express which takes 20 minutes. Arriving at the bus stop at 9.15 am what is the probability that he will be at his class on time if the number 12 is equally likely to arrive at any time between 9.10 and 9.30, if the number 15 bus is equally likely to arrive at any time between 9.20 and 9.40 and there are two possible express services James may catch: the first is equally likely to arrive at any time between 9.05 and 9.20 and the second equally likely to arrive at any time between 9.35 and 9.50. Assume that all the buses have arrival times that are independent.

2.3. (Combining union and intersection risk)

A building project is running late and if it is more than 4 weeks late an alternative venue will need to be found for an event planned in the new building. There is a 20% chance of poor weather causing a delay by 3 weeks, and there is a 10% chance of late delivery of a critical component which would lead to a delay of between 2 weeks and 4 weeks, on top of any weather related delay. The construction involves some excavation of a drain line in an area of some archaeological interest. Archaeologists are at work and there is a small (5%) chance that significant finds will be made that force a delay of around 2 months.

(a) Use a Venn diagram to show the events that will lead to a delay of more than 4 weeks

(b) Assuming that all three events are independent calculate the probability of more than 4 weeks delay.

2.4. (Probability for maximum from empirical distribution)

You observe the rainfall amounts on all days in April over the last 3 years. There are 90 data points and the largest 10 are as follows (in mm): 305, 320, 325, 333, 340, 342, 351, 370, 397, 420.

(a) Sketch the cdf for the distribution of rainfall.

(b) Assuming that rainfall on successive days is independent, use the empirical data to estimate the probability that the maximum daily rainfall during a 5 day period next year is less than 350 mm (i.e. less than 350 mm on each of the 5 days)

2.5. (Optimal portfolio)

There are three stocks to invest in: A, B and C. In one year the expected increases in price are: 10% for stock A, 15% for stock B, and 5% for stock C. The standard deviations of these numbers are 2% for A and B and 1% for C. If the returns are all independent, what is the minimum variance portfolio that achieves a 10% return? (Use a spreadsheet and 'solver' for this calculation).

2.6. (Optimal portfolio with a risk free asset).

Using the same arrangement as for Exercise 2.5, suppose that there is a risk free investment D that always increases in value by 4% over the course of a year.

(a) Recalculate the minimum variance portfolio for a 10% return, a 7% return and a 5% return. (Use a spreadsheet and 'solver' for this calculation).

(b) Show that these three portfolios lie on the same straight line in the standard deviation versus expected profit diagram.

(c) Show that each of the three portfolios are a combination involving some proportion of D and some proportion of a fixed portfolio of the A, B and C stocks (this is an example of the 'one-fund theorem' in portfolio theory).

2.7. (Multivariate normal)

A company wishes to estimate the probability that a storm and high tide combined will lead to flooding at a critical coastal installation. There are 10 high tide events each year when the installation is at risk. The two critical components here are wind velocity in the shore direction and wave height. The wind velocity has mean 10 km/hr and standard deviation 8 km/hr and the wave height has mean 2 metre and standard deviation 1 metre. The estimated covariance between these two variables is 40. Suppose that the behavior is modelled as a multivariate normal distribution (ignoring issues of negative wave height). Estimate the probability that there will be flooding next year assuming that flooding occurs when the wave height + 0.05×(wind velocity) is greater than 6.

2.8.(Copula properties)

(a) The text states that the rectangle inequality (2.4) and the uniform marginals condition is enough between them to show that the Copula increases in both its arguments. Show that this is true by using the rectangle inequality with $a_2 = b_2 = 0$ to show that $C(u_1, 0) = 0$ for any u_1, and then using the rectangle inequality again with $a_2 = 0$ to show that $C(u, b_2)$ is increasing in u for fixed b_2.

(b) Suppose that we define the copula density $c(u_1, u_2) = 2/3$ when both $u_1 < 3/4$ and $u_2 < 3/4$, $c(u_1, u_2) = 2$ when one of $u_1 < 3/4$ and $u_2 < 3/4$, and $c(u_1, u_2) = -2$ otherwise. Show that if $C(u_1, u_2)$ is defined in the usual way from C that it will be increasing in both arguments and have uniform marginals, but will not satisfy the rectangle inequality (since its density is negative on part of the unit square).

2.9.(Gumbel copula)

The Gumbel copula with parameter 2 is given by $C(u_1, u_2) = \exp(-\sqrt{(\ln(u_1))^2 + (\ln(u_2))^2})$. If two variables each have an exponential distribution with parameter 1 (so they have cdf $F(x) = 1 - e^{-x}$ for $x \geq 0$) and their joint behavior is determined by a Gumbel copula with parameter 2, calculate the probability that the maximum of the two variables has a value greater than 3 and compare this with the case where the two random variables are independent.

2.10.(Tail dependence for Gumbel copula)

Show that the Gumbel copula with parameter 2 has upper tail dependence, either using algebra or numerically.

3

Measuring Risk

The birth of Value at Risk

Dennis Weatherstone was chairman and later chief executive of JPMorgan. In many ways he was an unlikely figure to become one of the world's most respected bankers. As the son of a London Transport clerk who was born in Islington and left school at 16, he was a far cry from the expensively-educated people who typically run major Wall Street firms. He moved into the chairman's role from a position running the foreign-exchange trading desk. He was perceived as an expert on risk, but when he looked at the firm as a whole he found that he had little idea of the overall level of risk at JPMorgan. Over a period of several years the concept of Value at Risk was developed by the analysts and 'quants' working at JPMorgan as a way to answer this question. The need was to measure the risk inherent in any kind of portfolio. The Value at Risk was recalculated every day in response to changes in the portfolio as traders bought and sold individual securities.

This turned out to bring huge benefits when looking across the many activities going on at JPMorgan. It became possible to look at profits from different traders and compare them with the level of risk measured by Value at Risk. In the early 1990s, Weatherstone began to ask for daily reports from every trading desk. This became known as the 415 report: they were created at 4.15 pm every day just after the market closed. These reports enabled Weatherstone not only to compare every desk's estimated profit in comparison to a common measure of risk, but also to form a view for the firm as a whole.

In 1993 the theme of JPMorgan's annual client conference was risk, and these clients were given an insight into the Value at Risk methodology. When client's came to ask if they could purchase the same kind of system for their own companies, JPMorgan set up a small group called RiskMetrics to help clients. At that stage this was a 'proprietary' methodology that was being given away for free, with the aim of helping clients and establishing the reputation of JPMorgan in the risk area. Value at Risk became more and more popular and in 1998 RiskMetrics was spunoff as a separate company.

This was a time when the regulatory authorities began to pay more attention to risk. For example the Securities and Exchange Commission was concerned about the amount of risk arising from trading in derviatives, and created new rules forcing financial firms to disclose that risk to their investors. Inevitably the measure that was used was Value at Risk. All this was part of a slow but inexorable change that took Value at Risk from being specific set of tools developed within JPMorgan and sold by RiskMetrics into a risk management standard applied throughout the financial world.

Business Risk Management Edward Anderson

3.1 How can we measure risk?

In this chapter we will look in more depth at how to measure risk. Does it make sense to talk of one course of action being more risky than another? And if so what does this mean? In the simplest case we have a range of outcomes all with different probabilities and with different consequences. When the consequences can be accurately turned into dollar amounts then we obtain a distribution over dollar outcomes.

Our discussion in chapter 2 has essentially assumed that the distribution of outcomes is given by a known probability distribution, but in practice there are great difficulties in knowing the distribution that we are dealing with. It is always hard to estimate the probabilities associated with different outcomes and the monetary consequences of these events. In a financial calculation we may have some chance of estimating the relevant numbers. For example we might ask how likely it is that the price of gold gets above a certain level and (assuming we are betting against gold prices rising) what we lose if this happens. But in most management roles it is much harder than this. How can I calculate the probability that sales of my new product are less than 1000 units in the first year? How can I know how much it will cost me if the government introduces a revised safety code in my industry? For the moment we set these problems aside and assume that we have access to the numbers we need.

Often the best starting point for the estimation of risk is to consider what has happened in the past. Even if we think that the world has changed, it would still be foolish not to pay any attention to the pattern of results we have observed so far.

To make our discussion more concrete let us look at some specific weather related data, specifically the daily maximum temperature that is recorded for each of the days in 2010. This might be data we would look at if we were trying to sell air conditioning units, or trying to decide whether to spend money air conditioning our premises. The demand for air conditioning spikes upwards when temperatures are high. Let's compare Perth and Adelaide weather.

A starting point might be to look at the average of the maximum temperatures. The average daily maximum temperature in Perth in 2010 was 25.27°, while the average daily maximum temperature for Adelaide was 22.44°, more than 2 degrees cooler. But we could also measure variability. The usual way to do this is to use is the standard deviation σ. A spreadsheet can be used to calculate this for the two cities. The standard deviation for Perth is 6.41 and the standard deviation for Adelaide is 7.00, which is significantly larger. But if we are interested in how many very hot days occur neither of these figures is very informative. It is better to draw a frequency histogram of the data. This has been done in Figure **??**.

The graphs show that there is not much to choose between the two cities as regards the probability of really hot days. In 2010 the five hottest days in Perth were 42.9°, 42.7°, 41.5°, 41.1° and 40.0° and the five hottest days in Adelaide were 42.8°, 42.0°, 41.3°, 41.0° and 40.2°.

The critical point here is that if we are interested in the extreme results then the mean and standard deviation, which are mainly determined by the mass of more or less average results, will not give us the information we need. We must either look at the record of actual historical data or have some idea of the probability distribution that generates this data.

Now we turn more directly to a risk example. Suppose that we have agreed to sell 1000 tons of wheat in 3 years time at a price of US$300 per metric ton. If wheat prices are high we

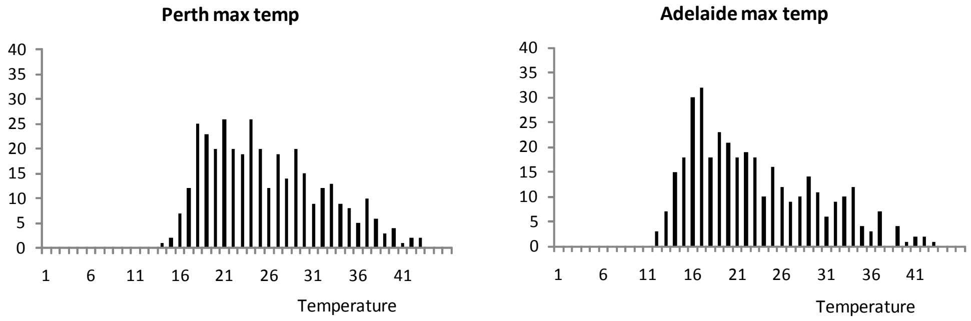


Figure 3.1 Frequency histograms of daily maximum temperature during 2010

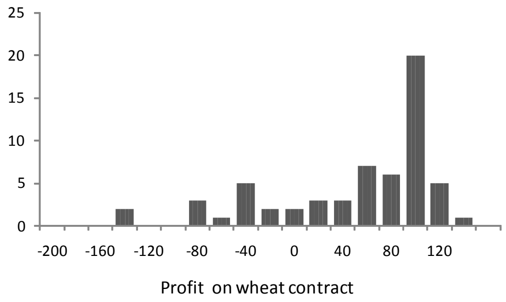


Figure 3.2 Frequency histogram of profits per ton made on a $300 per ton wheat contract

will make a loss, but if wheat prices are low we will make a profit. In order to estimate the probability of a loss we consider the historical record of wheat prices over the 5 year period from January 2005 on a monthly basis (Data is for Wheat, United States, no 2 Hard Red Winter (ordinary), FOB Gulf and has been downloaded from www.unctad.org). Factoring in the US$300 price we get the frequency histogram of profits per ton shown in Figure 3.2.

Wheat prices spiked in February and March 2008 to a value of around $450, but by April they had retreated to $387. So for two months (out of the 60) losses would have been $146 and $154 per ton, and then for a further 3 months losses would have been between $80 and $90 per ton. If we are interested in risk then we must concentrate on the losses that occur in this left hand tail. It will require a judgement call as to whether we think overall movements are likely to be up or down in the future, but certainly the pattern of price spikes that have been seen in the past would make one guess that the same sort of price spike might occur in the future. This set of data suggests that a price spike may happen about 1 month in

30, so it would certainly be prudent to allow for this happening again! But notice that the mean ($41.8) and standard deviation (69.5) tell us very little about what is going on - the distribution of profits is very far from being a normal distribution in this case.

So looking at history since 2005 gives us some idea about the distribution of outcomes and hence the risks involved. Suppose now that we want to extract a single measure of risk from this. One question we might ask is: What is the worst outcome that could occur? Historically, the answer is $154 per ton. But our instincts should tell us that this is not a reliable estimate for the worst that might happen. Even if the underlying factors don't change over time, we may still have been lucky in some way, and if were to look at a different 5 years of data perhaps we would observe a larger loss. In looking at the largest loss we are looking at a single month, and this is bound to mean a lot of fluctuation in what we observe. We will come back to this estimation problem in our discussion of extreme value theory in Chapter 4. In a sense the problem of measuring risk, which is a problem of describing what happens in the tails of the distribution, inevitably leads to difficulties when we want to use historical data (since by definition it all comes down to the values that occur at only a handful of points).

3.2 Value at Risk

When dealing with a distribution of profits risk is all about the size of the left hand tail of the distribution, and so it becomes clear that there is no single right way to measure the risk. But by far the most common way is to measure "value at risk" or *VaR* at some percentage point. For example we might say that the 99% value at risk figure is $300,000. This is equivalent to the statement that 99% of outcomes will lose less than $300,000, or we can be 99% sure that losses will not exceed $300,000. So giving a VaR number corresponds to picking a single point in the distribution.

Now we set about giving an exact definition for value at risk. Since we are concerned with potential losses it is easier to describe everything in terms of losses rather than in terms of profits. So the horizontal axis is reversed in order to have higher losses to the right hand side and the right hand tail becomes the area of interest. To make all this clearer Figure 3.3 shows the 95% and 99% VaR numbers for a distribution of losses over the range $(-1.5, 0.5)$ (all values are assumed to be denominated in units of $100,000 dollars). The density function shown is given by the equation

$$f(x) = \frac{15}{16}\left[(x+0.5)^4 - 2(x+0.5)^2 + 1\right] \tag{3.1}$$

Occasionally people will talk about a 1% VaR or a 5% VaR, but this just means the 99% and 95% values. There is not much danger of confusion since this means numbers near zero rather than near 100%. Also the terminology of VaR with a confidence level of 99% is often used. This is natural since a 99% VaR of $100,000 means that we can be 99% confident that losses will not exceed $100,000.

The VaR approach can be seen as an example of using "quantiles" to describe the tails of a distribution. We will use the terminology of the $\alpha\%$ quantile to mean the x value such that $F(x) = \alpha/100$ where F is the cumulative distribution function with $F(x) = \Pr(X < x)$ where X is the random variable in question. Thus the 50% quantile is the x value with $F(x) = 0.5$, i.e. the x value for which half the distribution is below it and half above - this is just the median. Thus the 99% VaR value is just the 99% quantile for the distribution of losses.

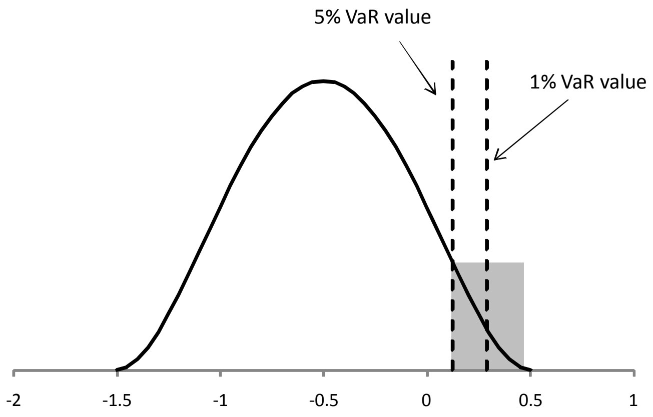


Figure 3.3 Illustration of Value at Risk

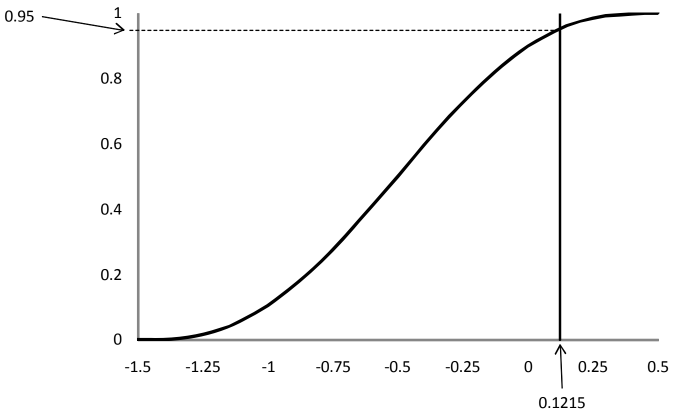


Figure 3.4 A quantile interpretation of VaR

We can convert our previous example with density given by (3.1) into a cdf form. We get

$$F(x) = \frac{3}{16}(x+0.5)^5 - \frac{5}{8}(x+0.5)^3 + \frac{15}{16}(x) + \frac{31}{32}$$

after integrating the expression for the density function. This is graphed in Figure 4 which also shows the 95% Quantile which is also the 95% VaR, being the value of x for which $F(x) = 0.05$. This turns out to be $x = 0.12149$, or a loss of $12,149.

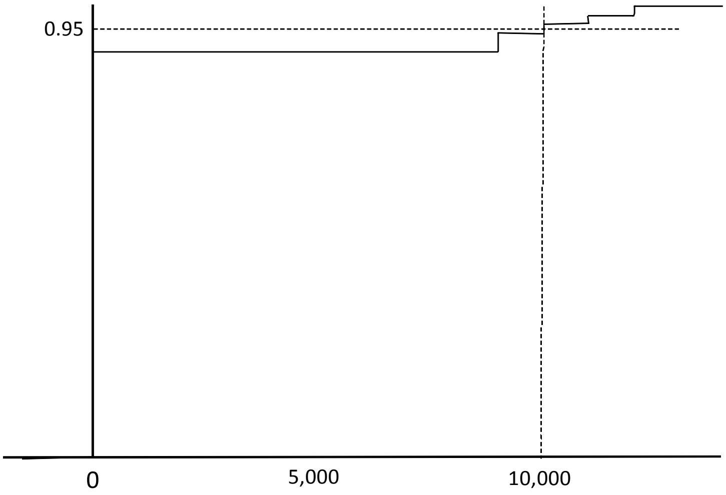


Figure 3.5 95% distribution function intersection at $10000

To write down a definition for the 95% VaR needs care. If L is the (uncertain) value of the losses, we might start by setting

$$\text{VaR}_{0.95} = \text{the } x \text{ value such that } \Pr(L \leq x) = 0.95,$$

or equivalently

$$\text{VaR}_{0.95} = F^{-1}(0.95)$$

where F is the cdf of the loss function and the notation F^{-1} is used for the inverse of F (i.e. $F^{-1}(y)$ is the value of x such that $F(x) = y$).

Unfortunately this definition will not quite work. The problem is that there may be no value at which $\Pr(L \leq x) = F(x) = 0.95$. For example suppose the the following losses occur:

loss of $11,000	probability 0.02
loss of $10,500	probability 0.02
loss of $10,000	probability 0.02
loss of $9,000	probability 0.04
loss of $0	probability 0.9

Then, settting $x = 10,000$ gives $\Pr(L \leq 10,000) = 0.96$ but for x even slightly less than $10,000$ the probability is smaller than 0.95, e.g. $\Pr(L \leq 9,999) = 0.94$.

There is no real problem if we draw the graph here. With this kind of discrete distribution (i.e. not continuous) the cdf is a step function. Figure 3.5 shows the graph of $F(x) = \Pr(L \leq x)$ for this example. It is easy to see that the graph of $F(x)$ goes through the value 0.95 at $x = 10,000$ and so this is the 95% VaR.

But what is easy with a graph is a little more complex with a formula. VaR is usually defined as follows

$$\text{VaR}_\alpha = \inf(x : \Pr(L > x) \leq 1 - \alpha),$$

so for a 95% VaR we have

$$\text{VaR}_{0.95} = \inf(x : \Pr(L > x) \leq 0.05),$$

which we can read as “The lowest value of x such that the probability of L being larger than x is less than 0.05”.

The idea here is that we think of a very large value of x for which this probability is definitely less than 0.05. then we slowly reduce x until the probability increases to 0.5 or more and that is where we stop. This is exactly the same as taking a large value of x where $F(x)$ is definitely greater than 0.95 and slowly reducing it till the value of $F(x)$ drops to 0.95 or less.

The use of $\inf$ here and the choices of inequalities that are strict ($L > x$) or not ($\leq 1 - \alpha$) is something that you do not need to worry about: it covers the definition for discrete probability distributions when there might be a range of values for which $F(x) = 0.95$ (i.e. a horizontal section in the graph of F). A simple way to describe the formula for VaR is the loss value at which the graph of $F(x)$ first reaches the correct percentile.

In financial environments the average return is often built into calculations on expected returns, and the thing which is of most interest is the risk that the final outcome will be much worse than the expected returns. In these cases VaR is calculated relative to the mean result. Thus in the example given in Figure 3 the mean profit is 0.5 or $50,000. So it would be normal to quote the 95% relative VaR as a loss of $62,149 rather than the “absolute VaR” which we calculated previously as $12,149.

Sometimes the relative Var is called the mean-VaR and is written $\text{VaR}_\alpha^{\text{mean}}$. In market risk management the period of time involved is very short - for example one day - and $\text{VaR}_\alpha^{\text{mean}}$ is called the “daily earnings at risk”. But in this case the expected market movement is bound to be close to zero and so the two definitions of VaR will in any case be essentially the same. So the distinction between relative and absolute VaR is more important when dealing with longer time horizons.

As an example of how VaR is reported by companies we give the following excerpt from the Microsoft annual report for 2010:

> (MICROSOFT) QUANTITATIVE AND QUALITATIVE DISCLOSURES ABOUT MARKET RISK
>
> We are exposed to economic risk from foreign currency exchange rates, interest rates, credit risk, equity prices, and commodity prices. A portion of these risks is hedged, but they may impact our financial statements.
>
> Foreign Currency. Certain forecasted transactions, assets, and liabilities are exposed to foreign currency risk. We monitor our foreign currency exposures daily and use hedges where practicable to offset the risks and maximize the economic effectiveness of our foreign currency positions. Principal currencies hedged include the euro, Japanese yen, British pound, and Canadian dollar.
>
> Interest Rate. Our fixed-income portfolio is diversified across credit sectors and maturities, consisting primarily of investment-grade securities. The credit risk and average maturity of the fixed-income portfolio is managed to achieve economic returns

that correlate to certain global and domestic fixed-income indices. In addition, we use “To Be Announced” forward purchase commitments of mortgage-backed assets to gain exposure to agency and mortgage-backed securities.

Equity. Our equity portfolio consists of global, developed, and emerging market securities that are subject to market price risk. We manage the securities relative to certain global and domestic indices and expect their economic risk and return to correlate with these indices.

Commodity. We use broad-based commodity exposures to enhance portfolio returns and facilitate portfolio diversification. Our investment portfolio has exposure to a variety of commodities, including precious metals, energy, and grain. We manage these exposures relative to global commodity indices and expect their economic risk and return to correlate with these indices.

VALUE-AT-RISK

We use a value-at-risk (“VaR”) model to estimate and quantify our market risks. VaR is the expected loss, for a given confidence level, in the fair value of our portfolio due to adverse market movements over a defined time horizon. The VaR model is not intended to represent actual losses in fair value, including determinations of other-than-temporary losses in fair value in accordance with accounting principles generally accepted in the United States (“U.S. GAAP”), but is used as a risk estimation and management tool. The distribution of the potential changes in total market value of all holdings is computed based on the historical volatilities and correlations among foreign currency exchange rates, interest rates, equity prices, and commodity prices, assuming normal market conditions.

The VaR is calculated as the total loss that will not be exceeded at the 97.5 percentile confidence level or, alternatively stated, the losses could exceed the VaR in 25 out of 1,000 cases. Several risk factors are not captured in the model, including liquidity risk, operational risk, and legal risk.

The following table sets forth the one-day VaR for substantially all of our positions as of June 30, 2010 and 2009, and for the year ended June 30, 2010 (in millions):

	30/6/2010	30/6/2009		2009 - 2010	
Risk Categories			Average	High	Low
Foreign currency	$ 57	$ 68	$ 53	$ 86	$ 20
Interest rate	$ 58	$ 42	$ 54	$ 69	$ 43
Equity	$ 183	$ 157	$ 184	$ 206	$ 142
Commodity	$ 19	$ 16	$ 17	$ 20	$ 14

Total one-day VaR for the combined risk categories was $235 million at June 30, 2010 and $211 million at June 30, 2009. The total VaR is 26% less at June 30, 2010, and 25% less at June 30, 2009, than the sum of the separate risk categories in the above table due to the diversification benefit of the combination of risks.

3.3 Combining and Comparing Risks

One great advantage of VaR as a way of measuring risk is that takes the complexity inherent in a probability distribution of possible outcomes and turns it into a single number. In general terms we want a single measure of risk because we want to compare different situations. Is the current environment more risky for our firm than it was a year ago? Is this potential business opportunity more risky than that one? Does our direct report, Tom, have a more risky management approach than Dick, another direct report?

This way of thinking leads to some natural properties that any risk measure should have. We suppose that X is a random variable giving the *losses*, and we write $\psi(X)$ for the risk measure for a random variable X.

1. **Monotonicity.** If losses in every situation get larger then the risk measure increases. Often we write $X \leq Y$ to mean that under any scenario the random variable X takes a value that is less than or equal to the value of the random variable Y. So this condition can be expressed succinctly as

$$\text{If } X \leq Y \text{ then } \psi(X) \leq \psi(Y).$$

2. **Positive homogeneity**. Multiplying risks by a positive constant also multiplies the risk measure by the same constant. Another way to think about this is to say that a change in the unit of currency leads to the risk measure changing in the appropriate way. In symbols:

$$\psi(bX) = b\psi(X) \text{ for any positive constant } b.$$

3. **Translation invariance**. If every outcome is changed by a certain fixed amount, this is also the change that occurs in the risk measure. In financial terms we can see this as a statement that adding a certain amount of cash to a portfolio decreases the risk by the same amount (This is the property that ties the risk measure back to actual dollar amounts.) We can write it as

$$\psi(c + X) = c + \psi(X) \text{ for any constant } c.$$

Example 3.1. For any distribution we can measure the mean μ, and except in very extreme cases we can also determine the standard deviation σ. If X is a loss random variable one option for a risk measure is $\psi(X) = \mu + k\sigma$ for some choice of k. For example we might take $k = 3$. Then the risk measure is at a point 3 standard deviations above the mean; a value that we know occurs only very infrequently. If X has a normal distribution then we can look up the probability in a z table: we have $\Pr(X > \mu + 3\sigma) = 0.0013$. This quite natural risk measure satisfies positive homogeneity and translation invariance, but does not satisfy the monotonicity requirement.

It is not hard to show that VaR satisfies each of these three conditions. Since VaR is a quantile, changing outcomes has no effect unless a scenario moves across the quantile value, and then an increase in loss can only increase the quantile value. Moreover

$$\begin{aligned}\text{VaR}_\alpha(bX+c) &= \inf(y : \Pr(bX+c>y) \le 1-\alpha) \\ &= \inf(y : \Pr(bX>y-c) \le 1-\alpha) \\ &= \inf(v : \Pr(bX>v) \le 1-\alpha) + c \\ &= \inf(v : \Pr(X>v/b) \le 1-\alpha) + c \\ &= b\inf(w : \Pr(X>w) \le 1-\alpha) + c \\ &= b\text{VaR}_\alpha(X) + c\end{aligned}$$

Thus VaR has most of the properties we would want from a measure of risk, but the area in which VaR is much less satisfactory relates to the combination of different risks. There is a natural property of risk measures to add to the three properties introduced above.

4. **Subadditivity**. Combining two risks together should not increase the overall amount of risk, indeed a diversification principal should lead to a decrease in overall risk. Mathematically we write this as

$$\psi(X+Y) \le \psi(X) + \psi(Y).$$

A risk measure that satisfies all four properties (monotonicity, positive homogeneity, translation invariance and subadditivity) is called *coherent*. It is important to realise that VaR does not satisfy subadditivity and so is not a coherent risk measure. First we need to explain why VaR fails to be subadditive and it is best to do this with an example.

Example 3.2. Suppose that we can invest \$10,000 in a bond A which will normally pay back \$11,000 in a year's time, but there is some credit risk. Specifically there is a small chance (which we estimate as 4%) that the bond issuer goes bankrupt and then we will get only a fraction of our money back (an amount we estimate as 30% of our investment, i.e. \$3,000). Assuming we are right in all our estimates, then the 95% absolute VaR is actually a negative amount $-\$1000$ (equivalent to a profit of \$1000). This is because the credit risk is too small too appear in the VaR calculation.

Now consider making a second investment in a bond B with exactly the same characteristics as A and suppose that bond B fails in a way that is quite independent of what happens to A. Then we get the following outcomes

Neither bond fails	Probability $0.96 \times 0.96 = 0.9216$	profit \$2,000
A fails, B does not fail	Probability $0.96 \times 0.04 = 0.0384$	loss \$6,000
B fails, A does not fail	Probability $0.96 \times 0.04 = 0.0384$	loss \$6,000
Both bonds fail	Probability $0.04 \times 0.04 = 0.0016$	loss \$14,000

We can see that the combined portfolio makes a loss with probability 0.0784 and the 95% absolute VaR value is a loss of \$6000. The credit risk is too small to influence the VaR on a single bond, but with a portfolio of bonds it can no longer be ignored. Notice however that the diversification benefit does not disappear (see Exercise 3.4), it is just cancelled out by the effect of a big loss crossing this particular 95% quantile boundary.

The problems with subadditivity highlights one of the limitations of VaR: there is something arbitrary about the confidence level $1 - \alpha$. VaR does not give a full picture of what happens in the tail of the distribution, and it says nothing at all about the maximum losses that may occur. Usually the worst that can happen is that a portfolio becomes worthless; so if we want to know how much we can lose, the answer may well be "everything"! In a business environment there will usually be some events that lead to losses that simply cannot be estimated in advance. In one sense VaR is helpful because at least it does not assume any estimates for extreme losses: its treatment of these extreme events means we need to estimate their probability, but allows us to pass over any estimation of the exact consequences.

3.4 VaR in practice

It is odd that Value at Risk is both very widely used and at the same time very controversial. Much of the controversy arises because the basic technique can be used in different ways - and some approaches can be misleading, perhaps even dangerous. However there is no getting away from VaR - for banks it is part of the Basel II framework which links capital requirements to market risk, and in the US some quantitative measures of risk are mandated by the SEC for company annual reports.

A sense of what is required under Basel II can be seen from the following excerpt from Clause 718 (section 76) (taken from http://www.basel-ii-accord.com)

> Banks will have flexibility in devising the precise nature of their models, but the following minimum standards will apply for the purpose of calculating their capital charge.
>
> ...
>
> (a) "Value-at-risk" must be computed on a daily basis.
>
> (b) In calculating the value-at-risk, a 99th percentile, one-tailed confidence interval is to be used.
>
> (c) In calculating value-at-risk, an instantaneous price shock equivalent to a 10 day movement in prices is to be used, i.e. the minimum "holding period" will be ten trading days. Banks may use value-at-risk numbers calculated according to shorter holding periods scaled up to ten days by the square root of time.
>
> (d) The choice of historical observation period (sample period) for calculating value at risk will be constrained to a minimum length of one year.
>
> ...
>
> (f) No particular type of model is prescribed. So long as each model used captures all the material risks run by the bank, banks will be free to use models based, for example, on variance-covariance matrices, historical simulations, or Monte Carlo simulations.
>
> (g) Banks will have discretion to recognise empirical correlations within broad risk categories (e.g. interest rates, exchange rates, equity prices and commodity prices, including related options volatilities in each risk factor category).
>
> ...

> (i) Each bank must meet, on a daily basis, a capital requirement expressed as the higher of (a) its previous day's value-at-risk number measured according to the parameters specified in this section and (b) an average of the daily value-at-risk measures on each of the preceding sixty business days, multiplied by a multiplication factor.
>
> (j) The multiplication factor will be set by individual supervisory authorities on the basis of their assessment of the quality of the bank's risk management system, subject to an absolute minimum of 3.

In practice there are three different approaches to calculating VaR figures. We may use historical price distributions (non-parametric VaR), we can use mathematical models of prices (perhaps including normal distributions for some risk factors), or we can use Monte Carlo simulation.

The historical approach is simple: we look at our current market portfolio and then use historical information to see how this portfolio would have performed over a period (of at least a year). Assuming that we use a year as the period, we might have about 250 trading days. If, like Microsoft, we want to calculate a 97.5% VaR then this would mean between 6 and 7 occasions during the year when VaR is exceeded. So we could take the seventh smallest daily loss recorded during the year on our portfolio as the VaR estimate. One great advantage of this approach is that by dealing with historical data we already capture the relationships between the different stocks in our portfolio. So we do not have to start making estimates of how correlated the stock price movements of Microsoft and Google are.

One disadvantage of this approach is that it is hard to know how long a data series to include. Is there something about current market conditions that is different to the long bull run up to 2008? If so then we should not include too much that earlier period in our analysis. But on the other hand if we give our historical analysis too short a period to work with, then we may well be overly influenced by particular events during that period. In general a historical approach is less likely to be appropriate when there has been a significant change in the market place.

The parametric approach is flexible and can take account of different correlation structures. However in practical terms once the problem becomes of a reasonable size we will be restricted to assuming that returns are normally distributed. However once this is done the calculations can be completed very quickly. Typically a parametric approach looks at the response of instruments like options to variations in the underlying securities (a very popular method in this class is provided by RiskMetrics). The danger of a parametric approach is that we do not correctly model the actual behaviour at the tails of the distribution. If there are 'fat tails' then this approach may be very misleading.

A third option is to use a Monte-Carlo simulation. This uses the parametric technique of modelling the individual components that generate risk, but rather than look for analytical solutions it instead simulates what might happen. A long enough simulation can capture the entire distribution of behaviour without the need for very specific choices of distribution and at the same time can represent any degree of complexity in the correlation structure. The weakness of this approach is that it still requires assumptions to be made on distributional forms, and it can also be computationally demanding.

Having decided the method that will be used to compute VaR numbers There are two further decisions that need to be taken. First a risk manager must decide on a *risk horizon*. This is the period of time over which losses may occur. Using a one-day VaR is about looking at price movements during a single day. But Basel specifies a 10-day period, and for many firms an even longer period will be appropriate. However the longer the period chosen the longer the time series data that will be needed in order to estimate it. In any event even with a longer period it is important to ensure that VaR calculations are done regularly and at least as often as the risk horizon (Basel II requires VaR to be calculated daily).

A second decision is the confidence level or quantile that will be used. Basel requires a 99% VaR to be calculated, but we have already seen how Microsoft uses a 97.5% VaR.

It is also important to check how the VaR estimates match actual risk performance, which is called "*back-testing*". The simplest way to do this is to apply the method currently in use to the company's past performance, to get an estimate of the VaR that would have been calculated on each day over the last year. The theory of VaR then tells us how many times we would expect to see losses greater than VaR (2.5 times if a 1% VaR level is used for 250 trading days). If we find that VaR limits have been breached more often than this we need to investigate further and should consider changing the method of calculation.

Whichever approach is used the generation of VaR numbers can be immensely helpful to managers, and it is worth reviewing why this is so.

- VaR is easily understood and is now familiar to many senior managers.
- VaR provides a single consistent measure of risk that can be used throughout the firm and can form the focus of discussion about risk.
- VaR can be calculated at the level of individual operational entities (or trading desks in a bank). This gives good visibility down to the lower levels of the company. It provides a tool that can be used to impose a consistent risk strategy through the organisation, at the same time as enabling senior managers to understand more of where and how risk arises within their organisation.
- VaR provides a good tool for assessing capital adequacy (and is required for that purpose by banking regulators).
- VaR has become the standard way of reporting risk externally.

3.5 Criticisms of VaR

As we mentioned earlier, the use of VaR is still controversial, and it is important to understand the criticisms that have been made. The primary problem with VaR is that it does not deal with events within the tail - it gives no guidance on how large extreme losses may turn out to be. There is a lot of difference between saying that on 99% of days I will lose no more than $100,000 and saying that on one day in each year (on average) I will lose $20 million. Yet these are quite consistent statements. Almost all the time everything is well controlled and my losses remain modest, but once in a while things will go very badly wrong and I will lose a lot.

David Einhorn, a well-known hedge fund manager, made a speech in 2008 (prophetically warning about Lehman Brothers potential problems) in which he said that VaR is "relatively useless as a risk-management tool and potentially catastrophic when its use creates a false sense of security among senior managers and watchdogs. This is like an air bag that works all the time, except when you have a car accident."

In an influential book, called ‘The Black Swan’, Nassim Nicholas Taleb has argued that we habitually take insufficient account of very rare but very important events - these are in his terminology ‘black swans’. They are sufficiently rare that we have not observed them before and so it makes little sense to talk about predicting their probability. At the same time they have very large effects. They are the unknowns that turn out to be more important than the things we do know about. Who could have predicted the changes that came about after the terrorist attack on the twin towers in 2001? Who could have anticipated the rise of social media on the internet? And these large scale phenomena are mirrored at the level of the firm by much that comes ‘out of left field’.

When we use VaR as a risk measure we deliberately exclude these events and their consequences. Even using a 99% one-day VaR (which sounds quite conservative) we deliberately exclude any events that happen less often than once every six months. For Taleb something that happens twice a year should be regarded as an ‘everyday’ occurrence. He argues that across many fields the exclusion of the 1% tail involves excluding events and data that turn out to have a very significant effect on the overall picture. For example if we were to look at i-Tunes downloads and exclude the 1% most downloaded tunes our estimate of how much money i-Tunes makes would probably be very inaccurate.

Another problem with VaR is that it may encourage inappropriate behaviour by managers. Joe Nocera, in a NY Times article, describes how VaR can be gamed. “To motivate managers, the banks began to compensate them not just for making big profits but also for making profits with low risks.” The result was an incentive to take on what might be called ‘*asymmetric risk positions*’ where there are usually small profits and only infrequent losses, but losses when they do occur can be enormous. “These positions made a manager’s VaR look good because VaR ignored the slim likelihood of giant losses, which could only come about in the event of a true catastrophe. A good example was a credit-default swap, which is essentially insurance that a company won’t default. The gains made from selling credit-default swaps are small and steady - and the chance of ever having to pay off that insurance was assumed to be miniscule. It was outside the 99% probability, so it didn’t show up in the VaR number.”

In fact the incentives to take actions which produce a skewed, or asymmetric, risk position are quite widespread. This will often happen when there is a reward based on relative ranking. Suppose, for example that we are a fund manager. We may have choices available to us which will make our returns look like the average return for the type of stocks we are investing in. This can be achieved simply by spreading our portfolio widely, and corresponds to a low risk option if we are being compared with this average performance (or the performance of other fund managers). On the other hand we could concentrate our portfolio on a few stocks. This would be riskier but may pay off handsomely if these stocks are good performers. Since fund managers are paid partly on the basis of “funds under management” and flows into a fund are often determined by its relative ranking, there is a big incentive to do better than other funds. This could lead to behaviour which gambles (by stock picking) if things are going badly (perhaps we can catch up) and plays safe if things are going well (“quit when we are ahead”). The end result is a distribution of returns that looks like Figure 3.6. There is a relatively small chance of very poor returns and quite a good chance of reasonably good returns.

The right approach here is to recognise what VaR measures and what it does not measure. It picks a single quantile and estimates where this is: it makes no attempt to say how far the tail stretches (how large the losses may be).

The most common approach to the shortcomings of VaR around extreme events is to use

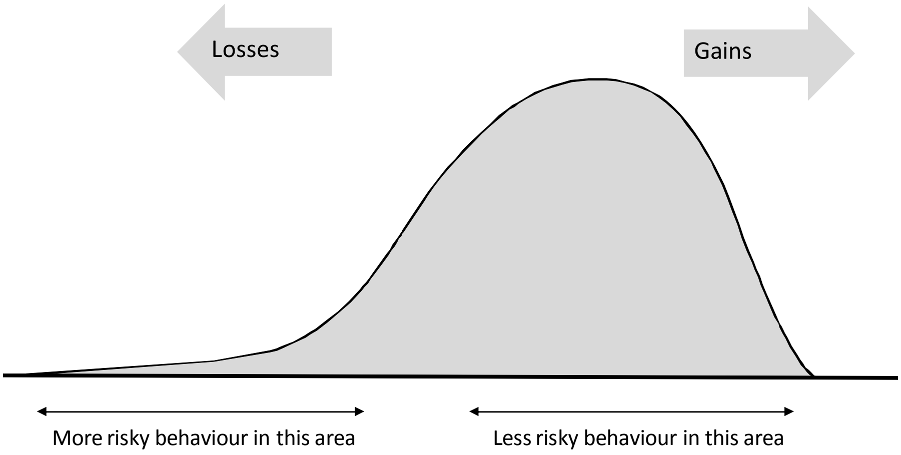

Losses
Gains
More risky behaviour in this area
Less risky behaviour in this area

Figure 3.6 Asymmetric return distribution

stress testing. As with several other risk management approaches, this is a technique that originated within the banking industry. Here is what HSBC says about this in its annual report for 2010:

> Although a valuable guide to risk, VaR should always be viewed in the context of its limitations. For example:
>
> - the use of historical data as a proxy for estimating future events may not encompass all potential events, particularly those which are extreme in nature;
> - the use of a one-day holding period assumes that all positions can be liquidated or the risks offset in one day. This may not fully reflect the market risk arising at times of severe illiquidity, when a one-day holding period may be insufficient to liquidate or hedge all positions fully;
> - the use of a 99 per cent confidence level, by definition, does not take into account losses that might occur beyond this level of confidence;
> - VaR is calculated on the basis of exposures outstanding at the close of business and therefore does not necessarily reflect intra-day exposures; and
> - VaR is unlikely to reflect loss potential on exposures that only arise under significant market moves.
>
> In recognition of the limitations of VaR, HSBC augments it with stress testing to evaluate the potential impact on portfolio values of more extreme, although plausible, events or movements in a set of financial variables. The scenarios to be applied at portfolio and consolidated levels, are as follows:
>
> - sensitivity scenarios, which consider the impact of any single risk factor or set of factors that are unlikely to be captured within the VaR models, such as the break of a currency peg;

- technical scenarios, which consider the largest move in each risk factor, without consideration of any underlying market correlation;
- hypothetical scenarios, which consider potential macro economic events, for example, a global flu pandemic; and
- historical scenarios, which incorporate historical observations of market movements during previous periods of stress which would not be captured within VaR.

The use of stress testing to explore the consequences of risks that occur in the tail is an approach that is complementary to VaR (which ignores the size of these risks) and may be useful for a wide range of firms. The different approaches to developing scenarios, outlined in the HSBC material, can also be applied more generally.

Notes

An excellent introduction for this area is the newspaper article by Joe Nocera and I have drawn on this at various points in this chapter. The book by Crouhy, Galai and Mark is very helpful in understand what VaR measures really mean in practice and Culp(2001) also treats the more practcial aspect of VaR calculation.

The book by Nicholas Taleb is well worth reading for its trenchant views on what is wrong with many of the quantitative approaches to risk measurement.

References

Michel Crouhy, Dan Galai and Robert Mark, The Essentials of Risk Management, McGraw Hill, 2006.

Christopher Culp, The Risk Management Process, Wiley, 2001.

David Einhorn, ‘Private profits and socialized risk’, Speech at Grant’s Spring Investment Conference, 8 April 2008, (and Global Association of Risk Professionals Risk Review, 2008).

Joe Nocera, ‘Risk Mismanagement’, New York Times, 2 January 2009.

Nicholas Taleb, The Black Swan, 2nd edition, Random House, 2010.

Exercises

3.1 (VaR for normal distributions)

(a) If the profits made each day by a trading desk are on average $100,000 and have a normal distribution with standard deviation $60,000, calculate a 99% and 95% absolute VaR.

(b) A second trading desk has exactly the same properties as the first (normal distribution with average profit of $100,000 and standard deviation of $60,000). If the second desk makes returns that are completely independent of the first then what are the 99% and 95% absolute VaR values for the combination of the two trading desks?

(c) If the results of the second trading desk are not independent of the first, what is the highest value (i.e. greatest losses) for 99% absolute VaR that might be achieved for the combination of the two trading desks?

3.2. (VaR for a triangle distribution)

Consider a distribution of profits over the range $0 to $200,000 where the density follows a triangle distribution $f(x) = x/X^2$ for $0 \leq x \leq X$ and $f(x) = (2X - x)/X^2$ for $X < x \leq 2X$ where $X = \$100,000$. Calculate 99% and 95% absolute VaR figures.

3.3. (A non-monotonic measure of risk) Example 3.1 gives a way of calculating risk from the mean and stadard deviation. Give an example where increasing the loss on some outcomes would lead to a reduction of the value of $\mu + 3\sigma$. Hint: Start with a distribution that is -1 with probability 0.1, $+1$ with probability 0.1 and is otherwise equal to zero. Remember that, for any random variable X, the standard deviation is the square root of the variance, which is the expected value of $(\mu - X)^2$.

3.4. (Diversification reduces VaR)

In Example 3.2 use a 98% VaR to show that there is some diversification benefit in investing $10,000 in each of A and B rather than putting $20,000 in two bonds from A.

3.5. (From one day to 10 days)

The Basel II framework asks for a 10 day VaR and then states that "Banks may use value-at-risk numbers calculated according to shorter holding periods scaled up to ten days by the square root of time." By this is meant that if the 1-day VaR is x then the 10-day VaR can be estimated as $x\sqrt{10} = 3.1623x$.

(a) Explain why this formula could only be appropriate for (relative) VaR and not for absolute VaR

(b) Show that if daily returns are independent and normally distributed then the proposed formula will give the correct result.

3.6. (VaR estimates are a process)

You are a manager with a VaR system in place to calculate 99% VaR values on a daily basis. Over the last 500 trading days (two years) there have been 5 occasions when the VaR values have been breached. A subordinate comes to you with some serious concerns in relation to the current VaR calculations, arguing that they wrongly represent correlations in behaviour occurring at times when the markets make large movements. He has carried out a set of alternative calculations of daily VaR values over the last two years which also has 5 occasions when the VaR values have been breached.

(a) Explain why the alternative daily VaR values may differ markedly from the values from the current system, but have the same number of VaR breaches.

(b) Suppose two systems have the same performance on backtest, Are they equally good? And what would it mean for one to be better than the other?

4

Understanding the Tails

Extrapolating beyond the data

Maria works in a large teaching hospital and has responsibility for ordering supplies. Today she faces what seems like an impossible situation: as she explained to her friend Anna "I need 3 years of data but I only have two years". Anna asks her to explain, and she launches into the problems that she has been mulling over for the last few days.

"For the last two years we have had a diagnostic test for a certain type of infant bronchial condition and have been recording patients with this condition, and now at last there is a drug that is effective. But it is in short supply and expensive. Worse still it contains some unstable components so can only be used in the first six weeks after manufacture. An urgent order still takes a week to arrive from the manufacturing facility in Switzerland, so we need to keep a minimum of a week's worth of stock and I now have to determine how much that is. In this kind of situation our usual rules suggest that we should run out no more often than once every three years on average. I can see how much we would have needed for the last 104 weeks, but that's not a long enough period."

Anna has just finished a risk management course as part of her MBA and wonders if there is some way to use some of what she has learnt to help her friend. "So you have 104 weekly data points and your problem is to estimate what the highest value would be if the series were extended to three years, i.e. $3 \times 52 = 156$ data points."

"That's it exactly" said Maria. "Come to think of it even 3 years of data would not really be sufficient - just because something didn't happen over the last three years does not mean it will not happen over the next three years."

"Have you thought of modelling the weekly usage with a normal distribution, or maybe some other sort of distribution."

"That was the same idea that I had" Maria replied "But I have looked at the numbers and there are more weeks with high demand than seems possible if it really was a normal distribution. What seems to happen is that the incidence of this condition varies according to a whole lot of factors that I don't know about; things like the weather, and the number of chest infections in the population. I don't think that I can use a model to make predictions here, there is just too much uncertainty, so we are left with the inadequate data."

"So in essence you need to use the data you have, but extrapolate to higher values that have not yet occurred and at the same time you do not want to make any assumptions about the kind of distribution", said Anna.

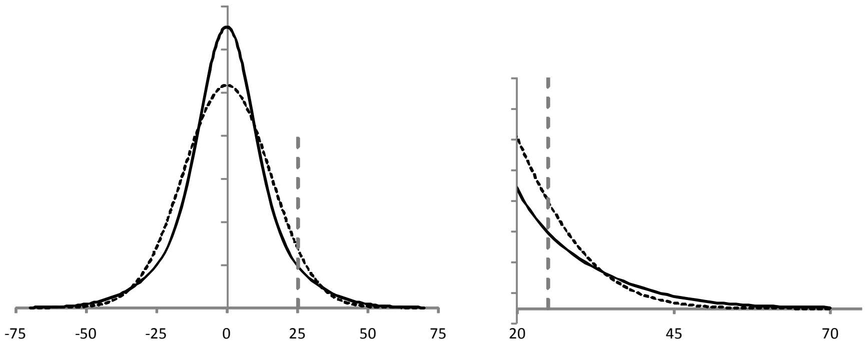


Figure 4.1 Solid line has a fatter tail (shown enlarged in right hand panel). Both distributions have 95% VaR of 25

"That's right: without enough data and without a specific model of the distribution it seems impossible to estimate the number I need."

"I am not so sure" said Anna, "Even without making an assumption that the weekly numbers match a given distribution, they surely cannot be too badly behaved. Any kind of regularity should give you a handle on the extrapolation problem. It may be hard to extract information about this 'upper tail' of the distribution from the data you have, but perhaps not impossible. I heard my professor talk about extreme value theory and perhaps that could help in some way."

4.1 Beyond Value at Risk

The challenge of using quantitative techniques to measure risk is that it forces us to pay attention to the tails of the distribution. This is exactly the area of greatest difficulty in estimation: we rarely see enough tail events to give a good handle in their estimation. In this chapter we will look in more detail at some tools for handling the tails of distributions.

The Value at Risk measure we considered in the previous chapter is concerned with the tail of a distribution in a way that the variance is not. But even though VaR focusses on the tail it is uninformative about what happens within the tail. The two loss distributions (density functions) drawn in Figure 4.1 have exactly the same 95% VaR value of 25 but the potential losses for the distribution drawn with a solid line are significantly higher. It is natural to talk of a distribution having fatter tails if the probability of getting values at the extremes is higher. The comparator here is the normal distribution (which is the dashed line in the figure) for which the probabilities go towards zero in the same way as $e^{-x^2} = 1/e^{x^2}$ which is a very fast decrease.

One way around this problem is to look at VaR values at different probabilities. The solid line distribution with the fatter tails has a 99% VaR of 41, while the 99% VaR for the dashed line, which is a normal distribution, is 36. So by moving further out in the tail the difference between the two distributions becomes more obvious from VaR alone

An alternative approach is to use what is usually called the *expected shortfall*, though other terminology is sometimes used (“tail value at risk” or “conditional value at risk”). The expected shortfall at a level α for a random variable X of losses is written $ES_\alpha(X)$ and is the expected loss conditional on the VaR$_\alpha$ level being exceeded. It is the average value over that part of the distribution which is greater than VaR$_\alpha(X)$, i.e. over loss values which occur with only a $1-\alpha$ probability. The expected shortfall is very closely related to value at risk, but captures more about what may happen in the worst cases. In comparison with value at risk the expected shortfall is a more natural measure of risk. The 95% value at risk is obtained by asking “What is the minimum loss amongst the 5% of worst outcomes?”, whereas the 95% expected shortfall value is obtained by asking “What is the average loss amongst the 5% of worst outcomes?”

This definition of expected shortfall is the simplest and most intuitive, but another way to define expected shortfall is to average the values of VaR$_u$ for all $u \geq \alpha$. Thus we have two expressions for expected shortfall at level α:

$$ES_\alpha(X) = E(X|\ X > \text{VaR}_\alpha(X)\ ) \tag{4.1}$$

and

$$ES_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 \text{VaR}_u(X)du. \tag{4.2}$$

To see that these two expressions are really the same we need to assume that X has a continuous density function $f(x)$ and then we can write the expectation in terms of the integral of f :

$$E(X|\ X > \text{VaR}_\alpha(X)\ ) = \frac{\int_{VaR_\alpha(X)}^\infty xf(x)dx}{\Pr(X > \text{VaR}_\alpha(X))} = \frac{1}{1-\alpha}\int_{VaR_\alpha(X)}^\infty xf(x)dx.$$

Now we are going to make a change of variable from x to a new variable u which is defined as $u = F(x)$. To do this we use the normal procedure of taking derivatives to see that $du = f(x)dx$. Also we note that

$$x = F^{-1}(u) = \text{VaR}_u(X)$$

and that at the lower limit of the integral $x =$VaR$_\alpha(X)$ so $u = \alpha$ and at the upper limit $u = F(\infty) = 1$. Thus

$$\int_{VaR_\alpha(X)}^\infty xf(x)dx = \int_\alpha^1 \text{VaR}_u(X)du$$

as we require.

There is an important warning here for discrete distributions, where there is no density function f , then the two definitions (4.1) and (4.2) are no longer equivalent and we should use (4.2) to avoid problems (or use a more complicated definition instead of (4.1)).

Example 4.1 (Expected shortfall compared to value at risk for exponential distribution)

Suppose that we buy insurance against extreme weather events that occur randomly, with an average of one event every ten years. We pay \$10,000 a year as a premium and receive a payout a total of \$95,000 in the event of the claim being made. Once a claim is made the insurance contract ceases. Premium payments are made monthly in advance and in the event of a claim are refunded for any period after the claim event. Ignoring discounting and any inflationary increases in premiums or a payouts, what are the VaR$_{0.95}$ and ES$_{0.95}$ values for our losses on this contract.

With random occurrences the time to the first weather event is a random variable with an exponential distribution. if we take years as units of time then we have an exponential with parameter 0.1. The loss (in \$1000s) is given by $L = 10X - 95$ where X has a density function $f(x) = 0.1e^{-0.1x}$. To calculate VaR$_{0.95}(L)$ note that

$$\text{VaR}_{0.95}(L) = 10\text{VaR}_{0.95}(X) - 95.$$

Now the probability in the tail of the exponential is

$$\begin{aligned}\int_u^\infty f(x)dx &= \int_u^\infty 0.1e^{-0.1x}dx \\ &= \left[-e^{-0.1x}\right]_u^\infty = e^{-0.1u},\end{aligned}$$

so in order to make this probability 0.05 we should set u so that $e^{-0.1u} = 0.05$, i.e. $u = -10\log_e 0.05 = 29.957$ which is VaR$_{0.95}(X)$. Thus there is a one in 20 chance that the weather event doesn't happen for about 30 years and the downside risk from the point of view of the person buying insurance is VaR$_{0.95}(L) = 10 \times 29.957 - 95 = 204.57$ or \$204,570.

But looking at expected shortfall gives an even larger figure: we have $ES_{0.95}(L) = 10\text{ES}_{0.95}(X) - 95$ and

$$\begin{aligned}ES_{0.95}(X) &= \frac{1}{1-\alpha}\int_{VaR_\alpha(X)}^\infty xf(x)dx \\ &= \frac{1}{0.05}\int_{29.957}^\infty 0.1xe^{-0.1x}dx \\ &= \frac{1}{0.05}\left[-10e^{-0.1x} - xe^{-0.1x}\right]_{29.957}^\infty \\ &= \frac{1}{0.05}\left(10e^{-2.996} + 29.957e^{-2.996}\right) \\ &= 39.946.\end{aligned}$$

(You should check that the expression we quoted as the integral here really does differentiate back to $0.1xe^{-0.1x}$). Thus we have an expected shortfall of

$$ES_{0.95}(L) = 10 \times 39.946 - 75 = 304.46$$

or \$304,460 which is about \$100,000 more than the value at risk. This is a greater difference than will occur for many distributions and is due to the particular shape of the exponential distribution with a long tail to the right.

As we discussed in Chapter 3 one of the problems with VaR is that it is not a coherent risk measure. Specifically we do not have the subadditivity property that

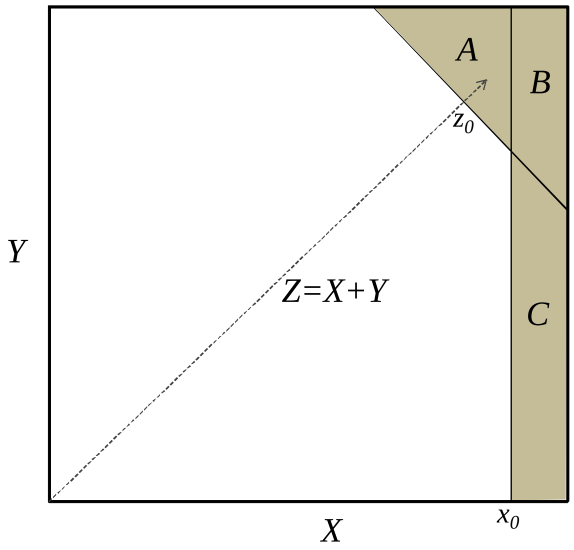


Figure 4.2 Diagram to show subadditivity of expected shortfall

VaR$_\alpha(X)$+VaR$_\alpha(Y)$ $\geq$VaR$_\alpha(X+Y)$. It turns out that expected shortfall *is* a coherent risk measure. There are four properties to check: monotonicity; positive homogeneity; translation invariance and subadditivity (see section 3.3). The fact that value at risk has the first three properties means that we can use (4.2) to show that expected shortfall also has these properties, so

$$\text{If } X \leq Y \text{ then } \mathrm{ES}_\alpha(X) \leq \mathrm{ES}_\alpha(Y);$$
$$\mathrm{ES}_\alpha(bX) = b\mathrm{ES}_\alpha(X);$$
$$\mathrm{ES}_\alpha(X+c) = \mathrm{ES}_\alpha(X) + c.$$

Now we will explain why expected shortfall is also subadditive. The key observation is that expected shortfall for X is an average of the *highest* values of X that can occur, where the events generating these values have a given probability $1-\alpha$. If we choose a different set of events which has the same probability $1-\alpha$ and look at the average of the values of X that occur under these events, then the value must be lower than the expected shortfall. We can represent this in a diagram as shown in Figure 4.2.

Here we take both X and Y as having finite ranges so that the diagram is easier to draw. We define a new random variable $Z = X + Y$. The regions B and C involve the highest possible values of the loss variable X . Suppose that they in total have a probability of $1-\alpha$ so that the point shown as x_0 will be at VaR$_\alpha(X)$. Now suppose that the combined A and B regions are where the highest values of Z occur and A and C have the same probability. Then A and B will also have a total probability of $1-\alpha$ and the value $Z = z_0$, which marks the lower boundary of this region, will be at VaR$_\alpha(Z)$. Now notice that taking the expected value of X over the region B and C and comparing it with the expected value of X over the region

A and B involves changing a set of events where $X > x_0$ to another set of events (with the same probability) where $X < x_0$. Hence

$$E(X|\ X > VaR_\alpha(X)\ ) \geq E(X|\ Z > VaR_\alpha(Z)\ ).$$

Exactly the same argument can be used to show that

$$E(Y|\ Y > VaR_\alpha(Y)\ ) \geq E(Y|\ Z > VaR_\alpha(Z)\ ).$$

Then we add these two inequalities together to establish subadditivity:

$$\begin{aligned} \text{ES}_\alpha(X) + \text{ES}_\alpha(Y) &= E(X|\ X > VaR_\alpha(X)\ ) + E(Y|\ Y > VaR_\alpha(Y)\ ) \\ &\geq E(X|\ Z > VaR_\alpha(Z)\ ) + E(Y|\ Z > VaR_\alpha(Z)\ ) \\ &= E(Z|\ Z > VaR_\alpha(Z)\ ) = \text{ES}_\alpha(Z). \end{aligned}$$

A more formal proof of this result can be found in Acerbi and Tasche, 2002.

The good theoretical properties of expected shortfall alongside its relative simplicity have made it more and more popular as a way of keeping track of risk.

4.2 Heavy-tailed distributions

When dealing with risk we are most often interested in random variables that have the possibility (at least theoretically) of infinitely large values. Like the normal distribution they do not have a finite range. Obviously in practice there will most often be a finite limit on the distribution: for example if dealing with a loss distribution there will be a maximum loss determined by our company's ability to avoid bankruptcy. But it is usually more revealing to model the losses as though there were no maximum limit. In this context it makes sense to look at the shape of the tail of the distribution as it goes to infinity.

The normal distribution gives a natural point of comparison for other distributions. A distribution is called 'heavy tailed' if it has more weight in the tails than a normal distribution. But we have to stop and think about what we might mean by such a statement. Because the standard deviation of a normal can be set to whatever we like it is always possible to find a normal distribution which has a zero mean but a large probability (anything up to 0.5) of being greater than some given value. So to make sense of a heavy tailed distribution we need to think of the behavior of the distribution across a whole range of values and not just at a single point.

One way to understand the behavior of the tail of a distribution is to ask how quickly the cdf of the distribution approaches one. A good way to do this is to consider the product $(1 - F(x))x^k$ for some power k. The first term $1 - F(x)$ will get closer and closer to zero as x increases, while the second term x^k will get larger and larger, so we can ask which will win? Does the product go to zero or go to infinity? We can guess that as we make k larger there will be some point k_0 at which the x^k term starts to dominate. For $k < k_0$ the product will approach zero and for $k > k_0$ the product will go to infinity. If the tail is heavy then there are high chances of seeing a large value and that means that $F(x)$ approaches 1 only slowly. So we will not be able to multiply by such a large value of x^k and still get the product going to zero. Hence a heavy tail is associated with a low value of k_0.

But what if there is no single value of k_0 that lies between the region where $(1 - F(x))x^k$ goes to infinity and the region where it goes to zero? In fact this is a very unlikely occurrence, sufficiently unlikely that we can really ignore the possibility. Lets try to see why this is so. We start by being more precise about what it means for a function $G(x)$ not to go to infinity and not to go to zero. Not going to infinity means that there is some value M such that $G(x) < M$ for an infinite sequence of x values $x_1, x_2, ...$ Thus no matter how large we take x we can always find a larger point where $G(x) < M.$ This is the logical negative of a statement that $G(x)$ tends to infinity which is equivalent to saying that for every M, $G(x)$ will end up above M and stay there. In the other direction $G(x)$ not going to zero means there is some value of m such that $G(x) > m$ for an infinite sequence of x values $x_1, x_2, ...$ Now suppose that this happens for $G(x) = (1 - F(x))x^k$ for all k in a range $[k_0 - \delta, k_0 + \delta]$. Now if

$$(1 - F(x_i))x_i^{k_0+\delta} < M \text{ for } x_i \to \infty,$$

then $(1 - F(x_i))x_i^{k_0} < Mx_i^{-\delta}$ and the right hand side goes to zero as x_i gets larger. In the same way from

$$(1 - F(x_i))x_i^{k_0-\delta} > m \text{ for } x_i \to \infty,$$

we have $(1 - F(x_i))x_i^{k_0} > mx_i^{\delta}$ and the right hand side goes to infinity as x_i gets larger. So at one and the same time there are x_i sequences where $(1 - F(x_i))x_i^{k_0}$ goes to infinity and where it goes to zero. The only way this can happen is for the function $(1 - F(x_i))x_i^{k_0}$ to oscillate up and down with the peaks being larger and larger and the troughs getting closer and closer to zero. We could construct such a function if we tried, but it is not something we would expect to occur in practice.

When we can define a value of k_0 in this way with $(1 - F(x))x^k$ going to either zero or infinity according as k is either below or above k_0, then we say that the distribution has a *tail index* of k_0. So for example a tail index of 2 is roughly equivalent to the statement that $1 - F(x)$ goes to zero in the same way as $1/x^2 = x^{-2}$. But the way we have defined this as a dividing point between two regimes is more precise. Also we don't have to say whether the function $L(x) = (1 - F(x))x^2$ itself goes to zero or infinity. In fact either of these options is possible, but the function $L(x)$ must not go to infinity or to zero too quickly. There is a specific condition required for this: $L(x)$ must be *slowly varying*, meaning that $\lim_{x\to\infty} L(tx)/L(x) = 1$ for any value of $t > 0$. So for example, taking $t = 2$, doubling the value of x cannot (in the limit of large x) look like applying any particular multiplier other than 1. Notice that if $L(x) = kx^{\beta}$ then $\lim_{x\to\infty} L(tx)/L(x) = t^{\beta}$ and so this can only equal 1 (and L be slowly varying) if $\beta = 0$.

There are some complications here that we don't want to get sucked into. The condition that $(1 - F(x))x^{\alpha}$ is slowly varying is actually a stronger condition than saying that the exponent α marks the dividing point between functions approaching zero and functions approaching infinity. (We can see this by observing that a periodic function like $2 + \sin x$ is not slowly varying but will be dominated by x^{ε} for even tiny values of ε.) The slowly varying condition is the one that is required to prove the more complex extreme value results that we give later on, even though the way we have defined a tail index is a bit simpler.

The variance of a random variable X is defined as $E(X^2) - E(X)^2$ but for distributions with a heavy tail this may not be defined (it may be infinite). Assuming that the mean $E(X)$ exists, we just need to check that the second moment $E(X^2)$ exists. When X has a density function f and a cdf F, we can write

$$E(X^2) = \lim_{R\to\infty} \int_{-R}^{R} f(x)x^2 dx.$$

Here we have written the upper and lower limits of the integral as $-R$ and R (rather than infinity) because the integral from $-\infty$ to ∞ is only defined when the limit as $R \to \infty$ exists, and the question of existence or not is precisely what we are interested in. Now choosing an arbitrary point u and integrating in the range above u (noting that in this range $x \geq u$) we get the inequality

$$\int_{u}^{R} f(x)x^2 dx \geq \int_{u}^{R} f(x)u^2 dx = (F(R) - F(u))u^2.$$

Letting R go to infinity shows that $\lim_{R\to\infty} \int_u^R f(x)x^2 dx \geq (1 - F(u))u^2$.

Now consider a distribution with a tail index α strictly less than 2. Then $(1 - F(u))u^2$ approaches infinity. Hence for any large number M we can choose a u with $(1 - F(u))u^2 > M$ and so, for this u, $\lim_{R\to\infty} \int_u^R f(x)x^2 dx > M$. The integral over the whole range $-R$ to R has to be larger than this, i.e.

$$\lim_{R\to\infty} \int_{-R}^{R} f(x)x^2 dx > M.$$

But since M can be chosen to be any number we like, this limit as $R \to \infty$ cannot exist, i.e. $E(X^2) = \infty$.

More generally we can use this argument to show that if a distribution has a tail index of α then the moments $E(X^k)$ will not exist for any $k > \alpha$. The smaller the tail index the fatter the tails of the distribution and the more likely it is that the moments will not exist. The are a whole set of distributions for which the tail index is infinite and all moments exist. These are the distributions like the normal for which the probability in the tail $(1 - F(x))$ approaches 0 faster than any power of x.

4.2.1 Estimating the tail index

Given a set of data points we may want to estimate how quickly the tail of the distribution goes to zero to find out whether or not we are dealing with a heavy tailed distribution. For example this is an important question to answer if we need to estimate an expected shortfall measure from the data. Later in the chapter we will give a more detailed discussion of the way that the estimation of a risk measure can be carried out using extreme value theory, but here we want to give a more elementary approach to estimating tail indexes.

Suppose that a distribution has a tail index of α then for large values of x we expect the cdf to be given approximately by

$$F(x) = 1 - kx^{-\alpha}$$

for some constant k. If this is true then

$$\log(1 - F(x)) = \log(kx^{-\alpha}) = \log(k) - \alpha \log(x)$$

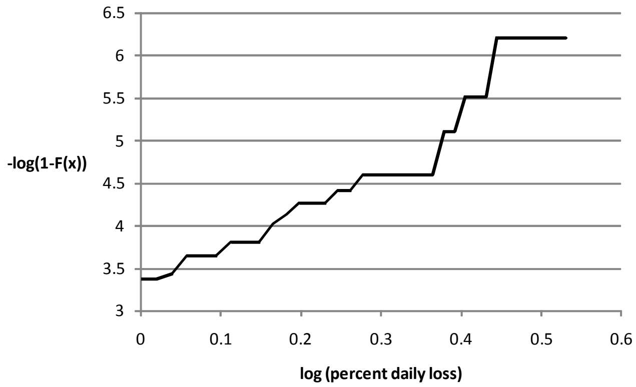


Figure 4.3 Tail data for percent daily loss in exchange rate GBP/USD

Since $1 - F(x)$ is between 0 and 1, and the log of a number less than 1 is negative, it can be helpful to multiply this equation by -1 to get

$$-\log(1 - F(x)) = -\log(k) + \alpha \log(x)$$

This shows that if we plot $-\log(1 - F(x))$ against $\log(x)$ then we should get a straight line with a slope of α.

Given a set of data points it is simple to form an estimate of $1 - F(x)$ by looking at the proportion of the points in the sample above the value x. Hence the procedure is to plot minus the log of this estimate against the log of x to estimate the tail index. To illustrate this Figure 4.3 shows a plot of this form for the exchange rate between the British piound and US dollar for a 500 day period starting in May 2010. The data is for the daily percentage change in the closing price. We are interested in losses if pounds are purchased (pound moving down relative to the dollar). During this period the greatest loss was 1.73 % on 13 May 2011. The log-log graph shown degenerates into steps at the right hand end as there are fewer and fewer points. The slope is relatively consistent at a value of around 4. This is clear cut evidence of heavy tailed behavior. Carrying out the same exercise with data drawn from a normal distribution would usually lead to slopes of 50 or more (typically large data sets allow us to move further into the tail and that suggests higher and higher slopes in a log-log graph of this sort).

4.3 Limiting distributions for the maximum

In this section we look at problems where the maximum of a number of values is of interest. Suppose that we are concerned to predict the maximum of N different random variables X_1, X_2, $...X_N$ all with the same distribution. We know that if all the X_i are independent and each have the cdf $F(x)$, then the distribution of their maximum which we call $F_{\max}$, has cdf $(F(x))^N$. When N is large it is only the tail of the original distribution that matters, since any value of x where $F(x)$ is substantially below 1 will automatically have $(F(x))^N$ very small.

This makes sense: when we take the maximum over a large number of draws from a random variable we are pretty much bound to see a value in the right hand tail of the distribution, and so the behavior of the distribution in the tail is the only thing that counts.

Lets start with a simple example. Suppose that F is a distribution where, in the tail of the distribution, $1 - F(x) = kx^{-\alpha}$ and we define $a_N = (kN)^{1/\alpha}$, then

$$\begin{aligned} F_{\max}(a_N x) &= (F(a_N x))^N \\ &= \left(1 - \frac{N(1 - F(a_N x))}{N}\right)^N \\ &= \left(1 - \frac{Nk(a_N x)^{-\alpha})}{N}\right)^N \\ &= \left(1 - \frac{x^{-\alpha}}{N}\right)^N . \end{aligned}$$

Now it is well-known that

$$\exp(x) = \lim_{N\to\infty} (1 + x/N)^N,$$

so when N is large we have $F_{\max}(a_N x)$ is approximately $\exp(-x^{-\alpha})$.

Thus if we take the maximum of say 50 independent identically distributed random variables where $1 - F(x)$ goes to zero like $x^{-\alpha}$ then the distribution is approximately a scaled version of one with distribution function $\exp(-x^{-\alpha})$, where the shape stays the same but is multiplied by $a_{50} = (50k)^{1/\alpha}$.

We can prove that the same kind of behavior occurs in a more general setting. Suppose that $(1 - F(t))t^\alpha$ is a slowly varying function of t (this implies that α is the tail index of F but is a slightly stronger claim) Thus if

$$z_N(x) = \frac{(1 - F(a_N x))a_N^\alpha x^\alpha}{(1 - F(a_N))a_N^\alpha} = \frac{(1 - F(a_N x))x^\alpha}{(1 - F(a_N))},$$

then for any fixed x, $z_N(x) \to 1$ as $N \to \infty$. Now we set $a_N = F^{-1}(1 - 1/N)$ which is the $(N - 1)/N$ quantile for the distribution. Now we can use the same argument we used above to show that

$$\begin{aligned} F_{\max}(a_N x) &= \left(1 - \frac{N(1 - F(a_N x))}{N}\right)^N \\ &= \left(1 - \frac{N z_N(x)(1 - F(a_N))x^{-\alpha}}{N}\right)^N \\ &= \left(1 - \frac{z_N(x)x^{-\alpha}}{N}\right)^N . \end{aligned}$$

So for any $\varepsilon > 0$ we will have $1 - \varepsilon < z_N < 1 + \varepsilon$ for N large enough. Hence

$$\lim_{N\to\infty} \left(1 - \frac{(1-\varepsilon)x^{-\alpha}}{N}\right)^N > \lim_{N\to\infty} F_{\max}(a_N x) > \lim_{N\to\infty} \left(1 - \frac{(1+\varepsilon)x^{-\alpha}}{N}\right)^N .$$

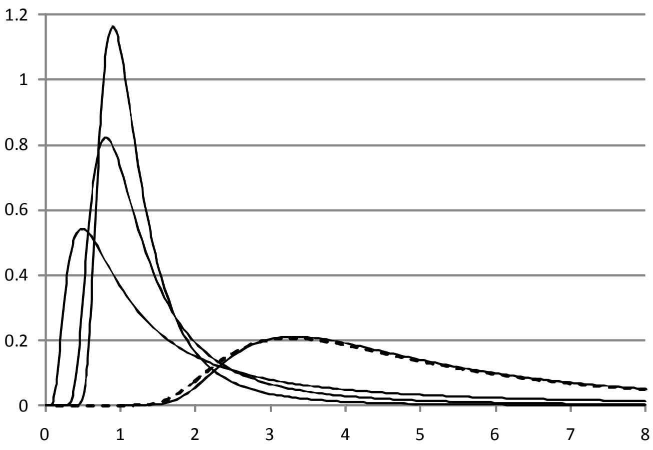


Figure 4.4 Frechet densities for different values of α

Since the left and right hand sides can be made as close as we like to $\exp(-x^{-\alpha})$ we have established that

$$\lim_{N\to\infty} F_{\max}(a_N x) = \exp(-x^{-\alpha}).$$

Thus the tail index alone determines the shape of the distribution, and this has form $F(x) = \exp(-x^{-\alpha})$, but this gets scaled by the $(N-1)/N$ quantile for the distribution, and thus by an amount that depends on N and the exact behavior in the tail of the distribution. The distribution with cdf $F(x) = \exp(-x^{-\alpha})$ and density $f(x) = \alpha x^{-\alpha-1}\exp(-x^{-\alpha})$ is called the Fréchet distribution after the very eminent French mathematician Maurice Fréchet. Figure 4.4 shows the density of the Fréchet distribution for different values of α. The three curves on the left of the figure show the density function when $\alpha = 1$, $\alpha = 2$ and $\alpha = 3$ (with higher values of α giving higher peaks). Shown on the same graph is the distribution of the maximum of 16 draws from a distribution with $F(x) = 1 - x^{-2}$, for $x > 1$. This should approximately match the Fréchet with $\alpha = 2$ scaled by an amount $16^{1/2} = 4$ and this comparison is also shown (the Fréchet is shown dashed). The match here is extremely good in the right hand tail, especially given the relatively small size of $N = 16$.

We can try the same approach with tails that approach zero faster than any polynomial (they have infinite tail index). As an example consider the exponential distribution, $F(x) =$

$1 - e^{-x}$ (and $f(x) = e^{-x}$) for $x > 0$. Then

$$\begin{aligned}
F_{\max}(\log(N) + x) &= (F(\log(N) + x))^N \\
&= \left(1 - \frac{N(1 - F(\log(N) + x))}{N}\right)^N \\
&= \left(1 - \frac{N \exp(-\log(N) - x))}{N}\right)^N \\
&= \left(1 - \frac{\exp(-x)}{N}\right)^N .
\end{aligned}$$

So

$$\lim_{N \to \infty} F_{\max}(\log(kN) + x) = \exp(-e^{-x})$$

Again we have a distribution that is reached in the limit but this time it is reached by shifting rather than by scaling. In fact the distribution where $F(x) = \exp(-e^{-x})$ is called the Gumbel distribution. This discussion shows that the maximum of say, 50 draws from the exponential distribution where $F(x) = 1 - e^{-x}$ has a distribution that is approximately Gumbel shifted by an amount $\ln(50) = 3.91$.

Now consider a second distribution where $F(x) = 1 - e^{-x^2}$(this has density $f(x) = 2xe^{-x^2}$). Define a_N as before by the $(N-1)/N$ quantile for the distribution, so we choose a_N so that $\exp(-a_N^2) = 1/N$, i.e. $a_N = \sqrt{\log N}$. Let

$$b_N = 1/(Nf(a_N)),$$

where f is the density function for F. Now

$$f(a_N) = 2a_N \exp(-a_N^2) = \frac{2}{N}\sqrt{\log N},$$

and hence $b_N = 1/(2\sqrt{\log N})$. So

$$\begin{aligned}
F_{\max}(a_N + b_N x) &= \left(1 - \frac{N(1 - F(a_N + x b_N))}{N}\right)^N \\
&= \left(1 - \frac{N \exp(-(a_N + x b_N)^2)}{N}\right)^N \\
&= \left(1 - \frac{N \exp(-(\sqrt{\log N} + x/(2\sqrt{\log N}))^2)}{N}\right)^N \\
&= \left(1 - \frac{N \exp(-\log N - x - x^2/(4\log N))}{N}\right)^N \\
&= \left(1 - \frac{z_N(x)\exp(-x)}{N}\right)^N ,
\end{aligned}$$

where $z_N(x) = \exp(-x^2/(4\log N))$. Observe that for fixed x, $z_N(x)$ will approach 1 as $N \to \infty$. Now following a similar argument to that we used before, we have

$$\lim_{N \to \infty} F_{\max}(a_N + x b_N) = \lim_{N \to \infty} \left(1 - \frac{\exp(-x)}{N}\right)^N = \exp(-e^{-x}).$$

We end up with exactly the same limit distribution as for the exponential case and this is rather surprising. We would not expect to reach the same limit distribution for two such different tail behaviors: there is an enormous difference between e^{-x} and e^{-x^2} (when $x = 5$ we get $e^{-5} = 6.7 \times 10^{-3}$ and $e^{-25} = 1.4 \times 10^{-11}$). Another interesting aspect is that the multiplier b_N goes to zero as $N \to \infty$. This corresponds to a bunching up of the distribution that does not occur with the exponential. So here there is a qualitative difference between the two cases. With the thinner tail associated with e^{-x^2} we find that we can be more and more accurate with our prediction of what the maximum value of N draws from the distribution will be.

The two examples we have given are particular instances of a more general result called the Fisher-Tippett theorem that we discuss below: this shows that for very many different distributions with an infinite tail index (essentially exponential decay) the Gumbel distribution occurs as the limit when considering maxima of repeated draws. For example a normal distribution also has this property. We show in Figure 4.5 the distribution of the maximum of 50 draws from a normal distribution compared with the Gumbel distribution with the appropriate shift and scaling applied. As we have seen the shift and scaling depend on the specifics of the distribution. For the normal we have $F_{\max}(a_N + b_N x)$ is approximately Gumbel with $a_N = \Phi^{-1}(1 - 1/N)$ and $b_N = 1/(N\varphi(a_N))$ where we have used the usual notation for the density, $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, and cdf, $\Phi(x) = \int_{-\infty}^{x} \varphi(s)ds$, of the normal distribution. When $N = 50$ we get $a_N = \Phi^{-1}(0.98) = 2.054$ and $b_N = 0.02/\varphi(2.054) = 0.413$. Thus $F_{\max}(2.054 + 0.413x)$ is approximately $\exp(-e^{-x})$. Figure 4.5 shows this reversed with the density of the maximum of 50 normal random variates compared with the density function for a Gumbel distribution $\exp\left(-\exp\left(-\frac{(y-2.054)}{0.413}\right)\right)$. This is not a particularly good match, which shows that convergence in the case of the normal distribution is rather slow. Using different scaling constants a_N and b_N can improve this slightly, but no matter what values are used the approximation will not be exact. We have chosen to use quite simple expressions for the two sequences we need ($a_N = F^{-1}(1 - 1/N)$ and $b_N = 1/(Nf(a_N))$), but these are not unique - different choices of these sequences can give the same result in the limit.

Now we want to move towards a more formal description of the Fisher-Tippett theorem. Notice that when we say that the distribution of maxima approaches a limiting distribution, what we mean is that the distribution is obtained by a scaling and shifting procedure (with this scaling and shifting depending on the number N of draws from the original distribution). So anything that is a scaled and shifted version of the limiting distribution could be used instead. We use the terminology of types: two distributions F_1 and F_2 are of the same *type* if they can be obtained from each other by scaling and shifting, i.e. if $F_1(x) = F_2(a + bx)$ for the right choice of a and b. For example any normal distribution, no matter what its mean and standard deviation is of the same type.

We say that the distribution F is in the *maximum domain of attraction* of a distribution H if there are sequences a_N and b_N with $(F(a_N + b_N x))^N \to H(x)$ for every x. We write this as $F \in MDA(H)$. This is really a statement about types of distribution: if $\widetilde{F}$ is a distribution of the same type as F, and $\widetilde{H}$ is of the same type as H, then $F \in MDA(H)$ and $\widetilde{F} \in MDA(\widetilde{H})$ are exactly equivalent statements. The terminology here can be misleading - the maximum domain of attraction sounds like it is the largest domain of attraction in some sense, but what is meant is that it is the domain of attraction under a maximum operator. Now we are ready

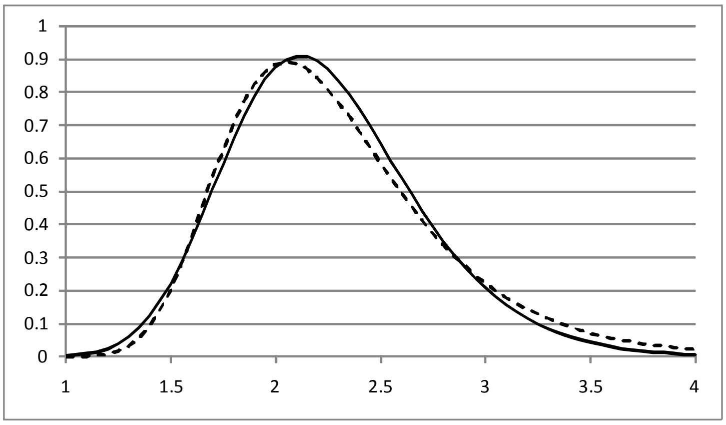


Figure 4.5 Gumbel compared with maximum from 50 normal random variates

to state the Fisher-Tippett theorem in the form it is usually given.

Theorem 4.1. If F is in $MDA(H)$ for some distribution H (and H is not concentrated on a single point) then there is a parameter ξ for which H is the same type as

$$H_\xi(x) = \begin{cases} \exp(-(1+\xi x)^{-1/\xi}) & \text{for } \xi \neq 0, \\ \exp(-e^{-x}) & \text{for } \xi = 0. \end{cases}$$

The formula for H_ξ defines a set of distributions usually called *generalized extreme value* (or GEV) distributions. The formula also implies the range of x values that apply, since we need $H_\xi(x)$ increasing from 0 to 1. When $\xi < 0$ then we need $x \leq -1/\xi$ and when $\xi > 0$ we need $x \geq 1/\xi$.

We have already met two examples of the generalized extreme value distributions. The $\xi = 0$ case is the Gumbel distribution. When $\xi > 0$ we can set $\alpha = 1/\xi$ and get a distribution function $H(x) = \exp(-(1 + x/\alpha)^{-\alpha})$ which is of the same type as the Fréchet distribution $\exp(-x^{-\alpha})$. The reason for giving this result in the more complicated form, with ξ rather than α, is that it makes it clearer that the $\xi = 0$ case can be reached in the limit as $\xi \to 0$ from either above or below. In both cases we are just using our standard observation that $(1 + x/n)^n$ approaches e^x as $n \to \infty$.

We need to say a little more about the case with $\xi < 0$. In this case the GEV distribution has a fixed right hand end point. and is of the type of a Weibull distribution. A Weibull is normally given as having a cdf $H(x) = \exp(-(-x)^\alpha)$ for $x < 0$ and $\alpha > 0$, but it is easy to see that this is the same type as the GEV with $\xi < 0$. These distributions arise when the original distribution from which draws are made has a fixed right hand end point $x_{\max}$, so obviously the maximum can never be larger than $x_{\max}$. It can be shown that in this case the behavior of the function $F(x)$ as x approaches $x_{\max}$ will determine the parameter ξ. If for example $F(x) = 1 - (x_{\max} - x)^2$ then the limiting distribution is Weibull with $\alpha = 2$ which converts to a GEV with $\xi = -1/2$. To make this come out in the form of a slowly varying function (which is all about how the function behaves as it goes to infinity) we need to do some manipulation and work with the inverse of the difference between x and $x_{\max}$. If

$$(1 - F(x_{\max} - (1/z)))z^{\alpha}$$

is a slowly varying function then the the maximum of N draws from the distribution is approximately (after scaling and shifting) a Weibull with parameter α (or GEV with parameter $-1/\alpha$).

The Fisher-Tippett theorem as it stands doesn't say how likely it is that there will be a limiting distribution for $F_{\max}$, it just specifies the form of that distribution if it occurs. But in fact it is very difficult to find examples where some limiting distribution does not occur. Remember that we allow scaling and shifting, so it is only the shape of the distribution that is important. We could for example use scaling and shifting to fix the fifth percentile and ninety fifth percentile points. Then we can watch what happens to the other quantile points as N increases. If they all converge then we will have a limiting distribution. If we imagine that there is no limiting distribution then it is hard to see how we could ever cycle between different shapes, but if there is no cycling then the quantile points which are constrained to be between the 5'th and 95'th percentile are bound to converge. So it is only at the extremes that things could go wrong, e.g. by having tail behavior like x^{-k} with k getting larger and larger as N increases, but doing this in a way that does not lead to some other limiting distribution. It turns out that this will not happen for distributions that occur in practice, and this can be demonstrated by looking in more detail at conditions sufficient to guarantee that F is in $MDA(H_{\xi})$. We have already shown how a slowly varying condition is enough when there is a tail index of α (and a similar condition applies with a fixed right hand limit), but we need something more complicated to deal with the $\xi = 0$ case and we will not give any details here.

4.4 Excess distributions

In this section we will look at the distribution of the tail by considering the distribution of $X - u$ for some threshold u conditional on X being greater than u. We write this as $F_u(x)$, thus

$$F_u(x) = \Pr(X - u \le x \mid X > u), \text{ for } x > 0,$$

and this is simply the probability that the random variable is in the range u to $x + u$ given that it is greater than u, i.e. using Bayes

$$F_u(x) = \frac{F(x+u) - F(u)}{1 - F(u)}, \text{ for } x > 0.$$

An alternative expression for F_u which can be useful is

$$F_u(x) = 1 - \frac{1 - F(x+u)}{1 - F(u)}.$$

We will discover that for many distributions F there will be a limiting distribution (defined up to scaling) for F_u as u gets very large. This theory closely parallels the theory of the distribution of the maximum of N draws given in the previous section.

Suppose that $(1-F(t))t^{\alpha}$ is a slowly varying function of t (this implies that α is the tail index of F) Thus if

$$z_u(k) = \frac{(1-F(ku))u^{\alpha}k^{\alpha}}{(1-F(u))u^{\alpha}} = \frac{(1-F(ku))k^{\alpha}}{(1-F(u))},$$

then for any fixed k, $z_u(k) \to 1$ as $u \to \infty$. Thus

$$\begin{aligned} F_u(ux) &= \frac{F(ux+u)-F(u)}{1-F(u)} \\ &= 1 - \frac{1-F((1+x)u)}{1-F(u)} \\ &= 1-(1+x)^{-\alpha} z_u(1+x). \end{aligned}$$

So as $u \to \infty$ we have

$$F_u(ux) \to 1-(1+x)^{-\alpha} \text{ for } x > 0. \tag{4.3}$$

This is a Pareto distribution. In general we say that X is distributed as a Pareto distribution Pa(α, κ) if $F(x) = 1-(\kappa/(\kappa+x))^{\alpha}$, so in our case scaling by u gives a Pareto with $\kappa = 1$. We can rewrite this to move the scaling factor into the Pareto parameters by noting that

$$\lim_{u\to\infty} \left(F_u(x) - 1 - (1+\frac{x}{u})^{-\alpha}\right) = 0 \text{ for } x > 0$$

Now consider a second distribution with a smaller tail where $F(x) = 1-e^{-x^2}$. Thus

$$\begin{aligned} F_u(\frac{x}{2u}) &= 1 - \frac{1-F(\frac{x}{2u}+u)}{1-F(u)} \\ &= 1 - \frac{\exp(-(\frac{x}{2u}+u)^2)}{\exp(-u^2)} \\ &= 1-\exp(-(\frac{x}{2u}+u)^2)\exp(u^2) \\ &= 1-\exp(-\frac{x^2}{4u^2})\exp(-x). \end{aligned}$$

Since $\exp(-\frac{x^2}{4u^2})$ approaches 1 as $u \to \infty$ we see that $F_u(\frac{x}{2u}) \to 1-\exp(-x)$ as $u \to \infty$. Thus in this case scaling by $1/(2u)$ gives an exponential distribution. Again we can, if we like, rewrite this to show that

$$\lim_{u\to\infty} (F_u(x) - 1 - \exp(-2ux)) = 0 \text{ for } x > 0.$$

It turns out that the right choice of scaling constant will achieve an exponential distribution in the limit for any reasonably simple distribution with an infinite tail index.

Not only is this reminiscent of our earlier discussion of the distribution of the maximum of N draws from the same distribution, it turns out that the conditions for convergence are also exactly the same. Again we can summarize the final position in a single theorem.

Theorem 4.2. If F is in $MDA_\xi(H)$ then there is a function $\beta(u)$ with

$$\lim_{u\to\infty} F_u(\beta(u)x) = \begin{cases} 1-(1+\xi x)^{-1/\xi} & \text{for } \xi \neq 0, \\ 1-e^{-x} & \text{for } \xi = 0. \end{cases}$$

With this formulation we obtain the exponential distribution in the limit as $\xi \to 0$, so the distribution shape changes in a way that is continuous with ξ. However, since there is an arbitrary $\beta(u)$ involved we could scale x by a further factor of ξ and get back to the form we gave earlier, i.e. with $1-(1+x)^{-\alpha}$ as the limit. For $\xi \geq 0$ the limiting distribution is defined over the range $x \geq 0$ and when $\xi < 0$ we require $0 \leq x \leq -1/\xi$.

It is interesting to compare this result with the Fisher-Tippett theorem. There we work with a maximum over N draws, and look at the limit as $N \to \infty$, while here we look at an excess over u and look at a limit as $u \to \infty$. In the Fisher-Tippett theorem the limiting distribution has the form $\exp(-Z)$ where the cdf moves from 0 to 1 as the expression Z moves from ∞ to 0, while here the limiting distribution is simply $1-Z$ and is defined over the the range where Z moves from 1 to 0. Moreover in the Fisher-Tippett theorem we may require both scaling and shifting (i.e. both a_N and b_N non zero) whereas here we only need to scale by $\beta(u)$.

To use this result we will fix on a large value of u and then approximate $F_u(x)$ with the *generalized Pareto distribution* (GPD) which has the following distribution function:

$$G_{\xi,\beta}(x) = \begin{cases} 1-(1+\xi x/\beta)^{-1/\xi} & \text{for } \xi \neq 0 \\ 1-\exp(-x/\beta) & \text{for } \xi = 0 \end{cases}, \tag{4.4}$$

where $\beta > 0$. When $\xi \geq 0$ we require $x \geq 0$ and when $\xi < 0$ we require $0 \leq x \leq -\beta/\xi$. The parameter ξ captures the shape and β simply acts as a scaling constant.

This scaling constant will grow linearly with u. If we compare Theorem 4.2 with the expression (4.3) we can see that $\beta(u) = u\xi$ in the case that $\xi > 0$ and the tail index is $1/\xi$. But in fact the same kind of linear scaling takes place even if $\xi \leq 0$. The result is that if $F_u(x) = G_{\xi,\beta}(x)$ for some particular ξ and β then when $v > u$ we have $F_v(x) = G_{\xi,\beta'}(x)$ where $\beta' = \beta + \xi(v-u)$. In other words, increasing the threshold by an amount Δ leaves the shape parameter ξ unchanged but increases the scaling parameter by an amount $\xi\Delta$.

We see this since

$$\begin{aligned} 1-F_v(x) &= \frac{1-F(x+v)}{1-F(v)} = \frac{1-F(x+v)}{1-F(u)}\frac{1-F(u)}{1-F(v)} \\ &= \frac{1-F(u+x+v-u)}{1-F(u)}\frac{1-F(u)}{1-F(u+v-u)} \\ &= \frac{1-F_u(x+v-u)}{1-F_u(v-u)} \\ &= \frac{1-G_{\xi,\beta}(x+v-u)}{1-G_{\xi,\beta}(v-u)} \end{aligned}$$

In the case where $\xi \neq 0$ we get

$$1-F_v(x) = \frac{(1+\xi(x+v-u)/\beta)^{-1/\xi}}{(1+\xi(v-u)/\beta)^{-1/\xi}} = \left(1+\frac{\xi x}{\beta+\xi(v-u)}\right)^{-1/\xi},$$

and in the case where $\xi = 0$ we get

$$1 - F_v(x) = \frac{\exp((x + v - u)/\beta)}{\exp((v - u)/\beta)} = \exp(x/\beta).$$

When $\xi = 0$ increasing the threshold makes no difference to the excess distribution, it stays as an exponential with the same parameter. This is the well-known memoryless property of the exponential distribution: if the time between events is an exponential distribution then knowing that nothing has happened so far makes no difference to the distribution of the time to the next event.

The theory gives the distribution of the excess, but the first thing we are likely to be interested in is the mean value of the excess over u, i.e. $E(X - u|X > u)$. To calculate the mean excess if $F_u(x) = G_{\xi,\beta}(x)$ we can start by taking the derivative of (4.4) to get the density function for the GPD: it is $1/\beta(1 + \xi x/\beta)^{-1/\xi-1}$. Hence the mean of the this distribution is given by the limit for large R of

$$\begin{aligned}
&\int_0^R (x/\beta)(1 + \xi x/\beta)^{-1/\xi-1}dx \\
&= \left(\beta/\xi^2\right) \int_1^{1+\xi R/\beta} (s-1)(s)^{-1/\xi-1}ds \\
&= (\beta/\xi)\left[s^{-\frac{1}{\xi}}\frac{(1-\xi-s)}{1-\xi}\right]_1^{1+\xi R/\beta} \\
&= (\beta/\xi)\left(\frac{\xi}{1-\xi} - (1+\xi R/\beta)^{-\frac{1}{\xi}}\frac{(\xi+\xi R/\beta)}{1-\xi}\right) \\
&= \frac{\beta}{1-\xi}\left(1 - (1+\xi R/\beta)^{-\frac{1}{\xi}}(1+R/\beta)\right)
\end{aligned}$$

where we have used the substitution $s = 1 + \xi x/\beta$ so $ds = (\xi/\beta)dx$ and the indefinite integral

$$\int (s-1)(s)^{-1/\xi-1}ds = \frac{\xi\,(1-\xi-s)}{1-\xi}s^{-\frac{1}{\xi}}.$$

If $\xi < 1$ then $(1+\xi R/\beta)^{-\frac{1}{\xi}}(1+R/\beta) < (1+R/\beta)^{1-\frac{1}{\xi}}$ and thus the term involving R goes to zero. Hence when X is distributed as $G_{\xi,\beta}$ its mean value is

$$E(X) = \beta/(1-\xi).$$

The same formula holds when $\xi = 0$ since the exponential distribution with density $(1/\beta)\exp(-x/\beta)$ has mean β.

In the case that $\xi \geq 1$ the term involving R goes to infinity and the mean does not exist. This is exactly what we would expect: the tail index is $1/\xi$ and if this index is less than 1 then the mean does not exist.

4.5 Estimation using extreme value theory

An understanding of the tail of a distribution is particularly valuable if we want to estimate probabilities associated with events that we have not seen, or have only seen rarely. The idea here is to estimate the shape and scale parameters from the data and then use these to estimate the probability of interest. Of course we will need to assume that there is sufficient consistency in the tail behavior that the theory we have discussed applies, but this is a caveat that needs to be born in mind whatever estimation we carry out. Estimation is always in some sense a deduction about what happens at a point where we do not have data, and this deduction is bound to make use of some modelling assumptions.

This takes us back to the opening scenario of this chapter, in which the requirement is to estimate a quantile of a distribution of weekly requirements for a particular drug. In that example Maria wanted to know the x value which would be exceeded on only one week in 156. This is extrapolation well beyond the 104 data points that he has observed, and so any estimate he makes is bound to be very uncertain. Nevertheless the ideas we are exploring here allow us to use quite unrestrictive assumptions to say something about what happens well out in the tail. We don't need to use a particular model of the distribution (whether that be normal, or exponential or something else) since the estimates we make will apply to all these distributions and many more. All we need to assume is that there is enough regularity in the behavior in the tail - the existence of a specific tail index would be enough.

The idea of using assumptions on reasonably regular behavior in order to make extrapolations is not too different in kind from what we do when we interpolate (i.e. when we estimate behavior between existing data points). Extrapolation makes an assumption that the behavior of the distribution does not change suddenly for extreme values, while making estimates from within the range of values that we have observed makes an assumption that there is no sudden change of behavior within that range. The difference is that when we extrapolate there is by definition nothing in the data itself which can warn us about changes, whereas a big change in the distribution at a point within the range of data we observe might be detected by looking at the data itself. This might seem like an important distinction, but in practice we would need quite a lot of data to spot such a change. For the problem that Maria faces even four or five years of data would give much the same issue. If for some reason there is a change in behavior of the distribution of weekly requirements that occurs at the 99% point, then looking at empirical data where only two or three data points are expected to be larger than this value will never reveal what is going on. So having double the amount of data will certainly help in making a good estimate, but there will still be a need to assume some regularity in the data.

In this section we will introduce a step by step method that can be effective in estimating Value at Risk or Expected Shortfall or other risk measures. However it is important to realize that this is just one of several methods that might be employed. Particularly for time series data there are good arguments for using a more sophisticated approach that takes account of changes in volatility over time (as often happens with financial data). Nevertheless when the correlations between data points are not too strong and when there is no reason to expect changes in the underlying distribution over time the method we describe can be very effective.

The first step is to choose a threshold u. The idea is to approximate the distribution of losses above this threshold level using a generalized Pareto distribution. We need to make u reasonably large so that the extreme value theory will apply. However since we will end up estimating the parameters of the GPD from the data that occurs above the threshold we must ensure that there are enough data points to do this. It is not so hard to choose a reasonable value of the threshold given a large data set, but it can be difficult if the data set is small. The approach we will illustrate uses a sample mean excess plot. For any threshold value v we look

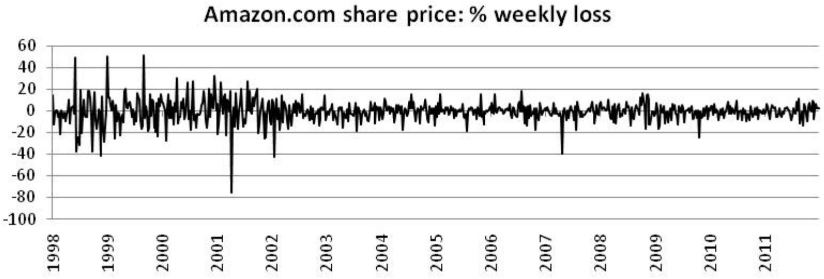


Figure 4.6 Weekly losses/gains for Amazon.com stock over period 1998 to 2011

at the average amount by which the data points exceed this threshold, averaging just over the data points larger than v. The sample mean excess is an estimate of the true value, and as we have already seen this will be $\beta/(1-\xi)$ if the excess distribution is a GPD $G_{\xi,\beta}$. The discussion in the previous section also demonstrated that once we have chosen a threshold large enough for the GPD approximation to be accurate then we expect that increasing the threshold by an amount Δ leaves ξ unchanged but increases β by an amount $\xi\Delta$. Thus we expect that in the tail of the distribution the mean excess will be linear as a function of the threshold with a slope of $\xi/(1-\xi)$ provided that $\xi < 1$.

Hence if we plot the sample mean excess for different thresholds we should get a rough straight line for values of the threshold large enough for the GPD approximation to apply but small enough for there to be enough points above the threshold for the sample estimate of the mean to be accurate. It is reasonable to carry out the estimation procedure with a value of u that is near the start of this final straight line section if it exists.

To illustrate this we look at some stock price data: Figure 4.6 shows the percentage weekly loss (negative values are gains) for the share price of Amazon.com over the 14 year period from the start of 1998 to the end of 2011. There are 730 weekly data points from which we calculate 729 percentage losses. Obviously there was significantly higher volatility in the early part of this period, but we will ignore this aspect of the data in our estimation procedure. Figure 4.7 shows the mean excess data as the threshold varies. We can see that for thresholds in the range 0 to 2% the mean excess stays around 6% loss, but for thresholds larger than a 2% loss the mean excess starts to rise. It does this in a fairly erratic way, but there is a definite upward trend. This suggests that we should use a value of 2% loss for u.

Having fixed a value of u, the second step in our procedure is to estimate the parameters ξ and β in the GPD distribution that applies above u. We can do this using a maximum likelihood approach. This is a fairly standard method but in case you are not familiar with it we lay out the details here.

Suppose that we have M observed values and we ask what is the probability that we would observe these values if the distribution were really GPD $G_{\xi,\beta}$ and each observation was chosen independently? Clearly getting any exact set of values is zero but we can consider the *likelihood* defined as the probability density of the overall distribution of M possible values

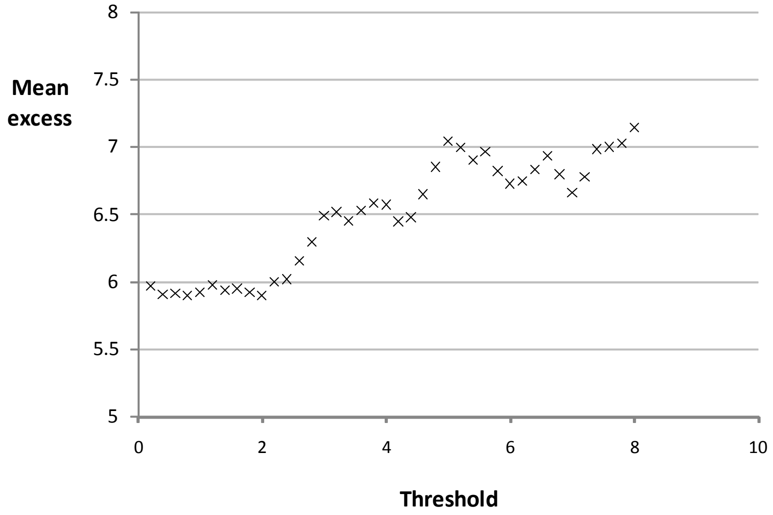


Figure 4.7 Mean excess plot for Amazon stock price data

evaluated at the set of observations we observe. Another way to think of the likelihood is as the limit of the probability of getting points in small intervals around those we actually observe, but normalized to allow for the effect of the interval sizes. For example suppose there are just 3 points that are above the level u and the amounts by which they exceed u are Y_1, Y_2 and Y_3. Since the density of the GPD is $(1/\beta)(1+\xi x/\beta)^{-1/\xi-1}$ the density of the joint distribution assuming independence is the product of this, i.e. the likelihood is

$$(1/\beta^3)(1+\xi Y_1/\beta)^{-1/\xi-1}(1+\xi Y_2/\beta)^{-1/\xi-1}(1+\xi Y_3/\beta)^{-1/\xi-1}$$

The idea is now to look for the values of the two parameters which make this as large as possible. The product form here is awkward to deal with particularly when we have many more than three observations. But maximizing likelihood will end up at the same choice of ξ and β as maximizing any increasing function of likelihood, such as the square of the likelihood. The increasing function which works best is to maximize the log likelihood in order to turn the product into a sum. So we maximize

$$\log\left((1/\beta^3)(1+\xi Y_1/\beta)^{-1/\xi-1}(1+\xi Y_2/\beta)^{-1/\xi-1}(1+\xi Y_3/\beta)^{-1/\xi-1}\right)$$
$$= -3\log\beta + (-1/\xi-1)\left(\log(1+\xi Y_1/\beta)+\log(1+\xi Y_2/\beta)+\log(1+\xi Y_3/\beta)\right).$$

More generally with M observed excess values Y_1, Y_2,... Y_M we make the estimate of ξ and β by choosing the values $\widehat{\xi}$ and $\widehat{\beta}$ which maximize

$$-M\log\beta - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{M}\log\left(1+\frac{\xi Y_i}{\beta}\right).$$

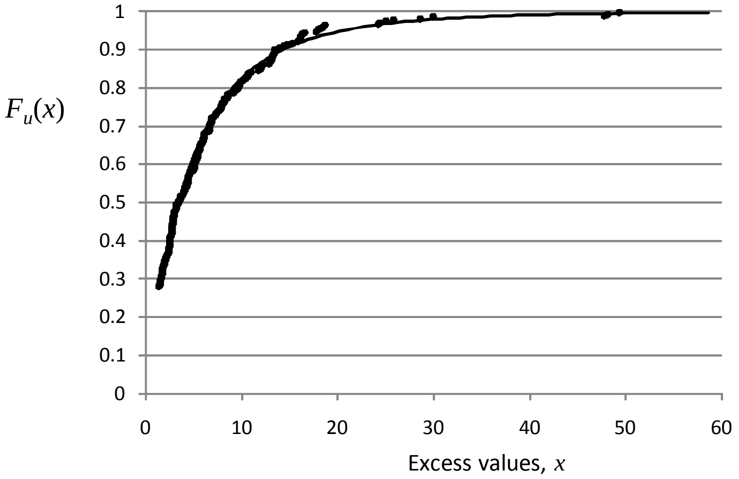


Figure 4.8 The estimated GPD distribution compared with empirical data for the Amazon.com % weekly loss with a threshold of 2

subject to the constraints $\beta > 0$ and $1 + \xi Y_i/\beta > 0$ for all Y_i.

Returning to the Amazon.com data we have 264 weeks in which the loss was greater than 2%. The log likelihood expression is maximized by taking $\widehat{\xi} = 0.27$ and $\widehat{\beta} = 4.51$. This corresponds to a tail index value of $1/\widehat{\xi} = 3.7$. Figure 4.8 shows how well the fitted cdf fits the data for the 264 data points above the threshold. We show the estimated value of $F_u(x)$ as a line and the empirical data as points. For example the 10th largest loss is 20.73 and this creates a point where $F(254/264) = 20.73$.

The third and final step is to use the fitted distribution to estimate values of interest. We will consider estimating both value at risk and expected shortfall. First consider value at risk. $\text{VaR}_\alpha(X)$ is the x such that $F(x) = \alpha$. If we are working with a fixed threshold u and $x > u$ then

$$\begin{aligned} F(x) &= 1 - \Pr(X > x) \\ &= 1 - \Pr(X > u)\Pr(X > x \mid X > u) \\ &= 1 - (1 - F(u))(1 - F_u(x - u)) \end{aligned}$$

If we assume that the F_u distribution is $G_{\xi,\beta}$ with the estimated values of $\widehat{\xi}$ and $\widehat{\beta}$ then we have

$$\alpha = 1 - (1 - F(u))(1 + \widehat{\xi}(x - u)/\widehat{\beta})^{-1/\widehat{\xi}}.$$

Hence to get the required value of x we invert this and obtain the following estimate

$$\text{VaR}_\alpha(X) = u + \frac{\widehat{\beta}}{\widehat{\xi}}\left(\left(\frac{1-\alpha}{1-F(u)}\right)^{-\widehat{\xi}} - 1\right)$$

The expected shortfall at a level α is $ES_\alpha(X) = E(X|\, X >VaR_\alpha(X)\,)$. To calculate this we can set $v =$VaR$_\alpha(X)$ and use our previous calculation that in the tail of the distribution for $v > u$ the excess distribution is $F_v(x) = G_{\xi,\beta'}(x)$ where $\beta' = \beta + \xi(v - u)$. Moreover we know that the mean value of a distribution $G_{\xi,\beta}$ is $\beta/(1-\xi)$. Hence the expected shortfall is

$$\begin{aligned} E(X|\, X > \text{VaR}_\alpha(X)) &= \text{VaR}_\alpha(X) + E(\text{excess over VaR}_\alpha(X)) \\ &= \text{VaR}_\alpha(X) + \frac{\beta_V}{1-\widehat{\xi}} \end{aligned}$$

where β_V is the estimated β value for the tail above VaR$_\alpha(X)$ so $\beta_V = \widehat{\beta} + \widehat{\xi}(\text{VaR}_\alpha(X) - u)$. Thus

$$\begin{aligned} \text{ES}_\alpha(X) &= \text{VaR}_\alpha(X) + \frac{\widehat{\beta} + \widehat{\xi}(\text{VaR}_\alpha(X) - u)}{1-\widehat{\xi}} \\ &= \frac{\text{VaR}_\alpha(X) + \widehat{\beta} - \widehat{\xi}u}{1-\widehat{\xi}} \end{aligned}$$

We can calculate the estimated 99% VaR and Expected shortfall values for the Amazon.com data in exactly this way. We use our earlier estimates of $\widehat{\xi} = 0.27$ and $\widehat{\beta} = 4.51$ with the threshold value of $u = 2$ to get

$$\text{VaR}_{0.99}(X) = 2 + \frac{4.51}{0.27}\left(\left(\frac{0.01}{(264/729)}\right)^{-0.27} - 1\right) = 29.3.$$

In this calculation we have used the proportion of observations over the threshold of 2 as an estimate of $(1 - F(u))$. The expected shortfall is therefore estimated by

$$\text{ES}_{0.99}(X) = \frac{29.3 + 4.51 - 0.27 \times 2}{1 - 0.27} = 45.6.$$

Notes

This chapter has the most sophisticated analysis of any part of this book, but at the same time it covers ground where much more could be said.

The book by Embrechts, Kluppelberg and Mikosch is a good source for the main theoretical ideas of extreme value theory and contains proofs of the two theorems we quote. The paper by De Haan also gives an accessible proof of the Generalized Extreme Value result and provides the sequence of normalizing constants $a_N = F^{-1}(1 - 1/N)$ and $b_N = 1/(Nf(a_N))$, that have the advantage of being relatively simple to use in examples. These references are also provide more specific conditions for a distribution to be in $MDA(H_0)$ (this will be guaranteed if $\lim_{x\to\infty} f'(x)(1 - F(x))/(f(x)^2 = -1$). The book by McNeil, Frey and Embrechts also gives a comprehensive introduction to many different techniques in this area.

References

Carlo Acerbi and Dirk Tasche, 2002, On the coherence of expected shortfall, *Journal of Banking & Finance*, Volume 26, pp. 1487-1503.

Laurens De Haan, 1976, Sample extremes: an elementary introduction, *Statsitica Neerlandica*, Vol 30, pp. 161-172.

Paul Embrechts, Claudia Kluppelberg and Thomas Mikosch, 1997, *Modelling Extremal Events for Insurance and Finance*, Springer.

Alexander McNeil, Rudiger Frey and Paul Embrechts, 2005, *Quantitative Risk Management*, Princeton University Press.

Exercises

4.1. (Expected shortfall for a normal distribution)

Calculate the expected shortfall at the 99% level for a normal distribution with mean 0 and standard deviation 1 using the fact that

$$\int_v^{\infty} x\exp(-x^2/2)dx = \exp(-v^2/2).$$

Hence estimated the expected shortfall at the 99% level if losses are distributed with mean $-\$10000$ and standard deviation $\$3000$.

4.2. (Expected shortfall is subadditive)

Use the formula you calculated in Exercise 4.1 to see what happens to the 99% expected shortfall if the losses in one project have a distribution with mean $-\$10000$ and standard deviation $\$3000$ and the losses in another project have a distribution with mean $-\$5000$ and standard deviation $\$1500$. Assuming that both distributions are normal, and the projects are independent, calculate the 99% expected shortfall for the sum of the two projects and hence check subadditivity in this case.

4.3. (Bound on expected shortfall)

Suppose that the density function f of the distribution for the loss random variable X is decreasing above the 0.95 percentile. Show, by considering the shape of the cdf function F that VaR$_u(X)$ is a convex function of u for $u > 0.95$ (i.e. has a slope increasing with u). Use a sketch to convince yourself that the average value of a convex function over an interval is greater than the value half way along the interval (this can also be proved formally). Finally use (4.2) to show that

$$ES_{0.95}(X) \geq VaR_{0.975}(X).$$

4.4. (Comparing upper and lower tails for exchange rate)

Figure 4.3 is generated using the data in spreadsheet *BRM_ch4_ExchangeRate*. Use this same data to carry out a similar analysis for movements in exhnage rate in the opposite direction (a drop in value of teh US$ in comparison with the pound). Do you think a tail index of about 4 is also appropriate in this case?

4.5. (Tail behaviour in a mixture of normals)

Suppose that we model the cost of gold at some fixed time in the future (say 10 January 2020) as given by a normal distribution with mean μ and standard deviation σ. Our idea is that there will be an average value that gold has in 2020 but that the price will fluctuate around that value. We do not know what either of these numbers will be, but we think that μ will be about $\$1500$ per oz and we think that the daily volatility which is measured by σ will be about $\$100$. More precisely we represent our uncertainty about μ by saying that μ is drawn from a normal distribution with mean 1500 and standard deviation 100 and we represent our uncertainty about σ by saying that σ is drawn from a normal distribution with mean 100 and standard deviation 20. Use the spreadsheet model *BRM_ch4_MixtureOfNormals* to explore the way that mixtures of normal distributions impact on tail behavior.

4.6. (GEV distribution)

A company is concerned about the possible bad publicity arising out of a guarantee made on its web site (“We will fix your router related problem in less than 20 minutes or we will give you free internet for a year”). Assume that the manager looks at the data on the 20 working days to assess for each day the maximum time that a router related problem took to fix. This list of maximum times has a mean of 12 minutes. On each day there were between 25 and 30 customer enquiries of this sort made. Assume that the time required is heavy tailed with a tail index of 5, and hence determine the distribution of daily maximum times. Use this distribution to estimate the probability that the guarantee will be broken on a given day and hence the expected number of days before this occurs.

4.7. (Estimating parameters from mean excess figures)

An analyst is looking at data on fee costs from winding up businesses after firm banckruptcy events. He has data on 900 such events and he calculates the mean excess figures using thresholds of \$10 million and \$20 million. There are 50 events with total fee costs greater than \$10 million with an average for those 50 of \$19 million (giving a mean excess of \$9 million) and there are 15 events with total fee costs in excess of \$20 million with an average for those 15 of \$32 million (giving a mean excess of \$12 million). Estimate the values of β and ξ for the Generalized Pareto Distribution for the excess above \$25 million.

4.8. (Mean excess plot when means are not defined)

Generate some data from a distribution with a value of $\xi = 1.2$ by using the cell formula =1/(1-RAND())^1.2 in a spreadsheet (with such a low tail index the mean of the distribution will not exist). Check what happens to the mean excess plot in this case. You should find that it seems surprisingly well-behaved with a straight line feel through most of the range. Can you explain what is going on? This shows the value of checking the fit obtained from the Maximum Likelihood estimator in the way that is done in Figure 4.8.

4.9. (Estimating VaR and ES using extreme value theory)

Use the process described for the Amazon stock data to estimate the 99% VaR and 99% ES for daily losses on the S&P 500 index. The data is given in the spreadsheet model *BRM_ch4_S&P500*

5

Making Decisions Under Uncertainty

Do we want stable prices?

Rico Tasker is in charge of fresh fruit and vegetable sales at a large retail chain. An important product for Rico is tomatoes. The price the retailer pays is fixed at wholesale fruit markets and varies according to the weather and the season. The retailer makes a markup of between 60% and 72% on sales with an average markup of 66%. These high markups are needed to cover the cost of storage, handling and other retail expenses. Any tomatoes not sold after 5 days are thrown away and on average this amounts to 10% of the tomatoes bought. The average wholesale price of tomatoes is $3 per kilo. After discussions with the particular grower who currently provides about three quarters of the tomatoes that the retailer sells, Rico goes to his boss, Suvi Marshall, with a proposal that the grower be offered a fixed price of $3 per kilo throughout the year and that the retailer sell the product with a price promise at $4.99 a kilo. This guarantees the same markup ($1.99/3 = 0.663$) and by working with a single grower with fixed prices the whole process will be simpler. The grower has agreed to meet all the retailer's requirements up to a limit given by the grower's entire output in any week. Rico argues that, in addition to reducing management costs, making this choice will remove part of the risk faced by the retailer by eliminating volatility in price.

"But what happens when there are a large number of tomatoes and everyone else has low prices," says Suvi "Won't that make our sales much lower?"

"Yes, I guess there will be a drop, but many people come to our shop for all their fruit and vegetables" Rico says. "so we will still have healthy sales. Besides we should make more sales at times when there is a shortage when everyone else has higher prices."

Suvi is still not entirely convinced. "What about the minimum level of supply from the grower?" she asks. "If there is a general shortage don't we normally sell most of what we have anyway? So will we really sell more in those periods?"

On the other hand Suvi is quite attracted by the idea of a price promise seeing the marketing potential of a guarantee that prices will be fixed for a full year. But this is a commitment that could lead to a bad outcome. Sometimes weather patterns persist for a long time. What if there was six months of high availability and low wholesale prices? Or six months of shortage? Moreover she knows that the actual profit made depends critically on the amount that goes to waste. If the 10% average was to creep up to 12% or 13% it would make a big difference.

Business Risk Management Edward Anderson

Overall she faces a difficult decision: paradoxically it is uncertainty about outcomes that makes this hard, even though the proposed change is designed to reduce uncertainty.

5.1 Decisions, states and outcomes

In the previous chapters we were concerned with understanding (and measuring) the characteristics of risk in terms of probabilities and the consequences in terms of costs. Now we turn to the question of how a manager should behave in a risky environment. In Chapter 6 we will focus on how individuals actually behave in a risky environment. But we start with a normative, rather than descriptive, view: given the need to make a decision between alternatives, each of which carries risks, how should a manager make this decision?

It is helpful to distinguish carefully between the things that we can control, these are the *decisions* we take, and the things that happen that are outside of our control, these are the *events* that occur.

A decision is actually a choice between possible actions. If only one thing can be done, then there is no decision to be made. In this chapter we will focus on decisions made at a single point in time and that makes things simpler (In Chapter 7 we look in more detail at dynamic problems where a succession of decisions need to be made.) A decision could involve the choice of a variable, for example we might decide how high to build a sea wall, or how much inventory of some product to hold. In these cases there are effectively an infinite number of possible choices. But we will concentrate on the situation in which there are only a finite set of possible actions. This will make our discussion much simpler and in practice an entirely free choice of a variable can usually be approximated through giving a large enough menu of choices.

We treat events as random: we may have knowledge of the likelihood of different events, but we cannot forecast exactly what will happen. We refer to the uncertain events or the values of uncertain variables as the *state* and this is unknown to the decision maker at the time that the choice of an action is made. The list of all possible states that may occur is called the *state space*. The state captures all of the uncertainty involved in the decision problem. For example suppose that we want to model a situation in which we invest $1000 in a stock on Wednesday if its price is higher at the close on Tuesday than it was at the close on Monday. We will need to decide what to do if the closing prices on the two days are the same: suppose that we toss a coin to decide whether to invest in this case. We might decide to model this by saying that the states are the difference in prices between Monday and Tuesday and there are three actions: invest, not invest and toss a coin. But this leaves some uncertainty out of the state description, and instead we need to include both the change in price and the coin toss result within the description of the state.

One way to think of the states is to imagine a game in which one of the players is the decision maker and the other player is ‘nature’. Both players get a chance to move: the decision maker chooses his actions and nature chooses hers. In this view the state is simply what nature chooses to do.

The action we take and the random events that occur together determine what happens to us: this is the *outcome* or consequence of taking a particular decision. The outcome will often contain several dimensions: for example an investment decision will lead to a dividend income stream as well as a final portfolio value; a decision to abandon a new product launch will lead to consequences not only for profit but also for reputation; a decision to relocate a

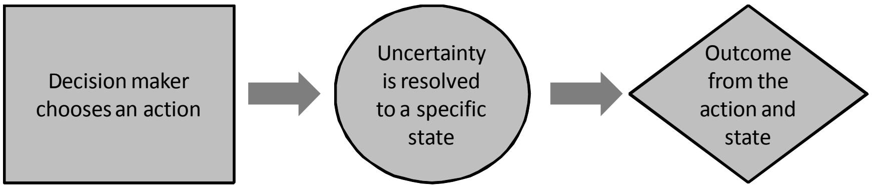

Decision maker chooses an action
Uncertainty is resolved to a specific state
Outcome from the action and state

Figure 5.1 Framework for actions, states and outcomes

manufacturing facility will lead to both direct costs and more indirect costs associated with travel to work times for employees. The outcome needs to take into account everything that can have an impact on the decision we make.

We can summarize this by saying that our framework has the three components shown in Figure 5.1 :

- A decision maker who chooses between a set of available actions.
- The possible states that can occur.
- The outcome that is reached as a result of the combination of the choice of action by the decision maker and the state that occurs.

All this is relatively straightforward, but a word of warning is in order. The terminology that we have used is not quite universal. Sometimes the word ‘outcome’ is used simply to refer to the particular state that occurs rather than the consequence of a decision. Also some people use the word ‘state’ to refer to the information available to a decision maker at a certain point in time (this happens often when dealing with problems which evolve over time). So, for example, we might follow the amount of money that a gambler has after a number of different gambles at the roulette wheel and refer to this as the state at time t. This is really the state of the gambler and not a state of nature. When the term ‘state’ is used in this way it will be determined by the random states of nature as well as the decisions made by the gambler; so in our terminology this is more of an outcome than a state.

Once we have set in place the framework of decisions and states together with the outcomes that arise from every combination of action and state, we need two further components to complete our decision model. We need to know the probability that different states occur and we need to know the value that we place on different outcomes.

Throughout this book we have been happy to talk about probabilities for events (i.e. probabilities for subsets of the state space). In doing so we have sidestepped what is really a topic of considerable debate. It may be obvious what we mean when we say that the probability of rolling a six with a dice is $1/6$, but few real events can have their probabilities calculated so simply. We may say that the probability that the US has a recession (two quarters of negative growth) at some point in the next 10 years is 60%. But this is imply an informed guess: either this event happens or it doesn’t and the chances are greatly influenced by many many factors both within and outside the US (for example policy decisions by governments around the world). Moreover there are factors that go beyond a simple political calculation such as climate change, natural disasters, terrorist action and wars (e.g. through changes in the price of gas and oil). However if we have to make a business decision that

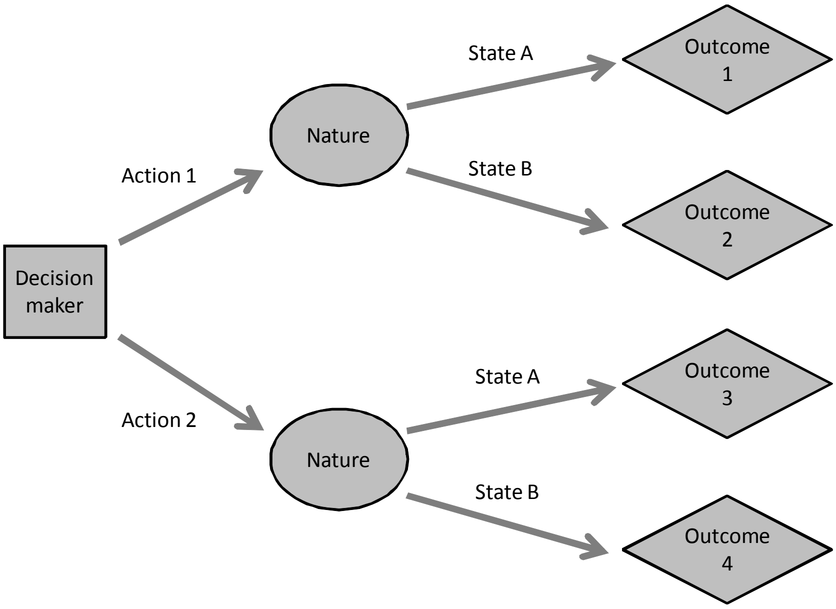


Figure 5.2 A decision tree with two actions and two states

is impacted by the state of the economy in the future, then there is an implication that we need to take account of the possibility of a recession. This leads us to make some sort of subjective judgement of this probability, and even if we don't write down probability values (or include them in spreadsheets) if a business decision has outcomes that depend on this uncertain event then the probabilities are likely to be taken account of in our decision process in some implicit way - in which case it is probably better to have them out in the open and subject to debate. Here we will proceed on the assumption that the decision maker has an agreed set of probabilities for the events involved in the decision problem. There are some alternative approaches to decision problems that can be used when the uncertainties are large enough that working with a specific subjective probability is dangerous: these techniques of robust optimization are discussed in Chapter 9.

Finally, and critically, our decision model has to have a well defined way of comparing different outcomes. In fact we need to go beyond the comparison of simple outcomes; one choice of action might be certain to lead to outcome A, while an alternative is equally likely to lead to either outcome B or outcome C. If C is preferable to A, but A is preferable to B then making a decision between the two possible actions will be hard. This is the central problem that we address in this chapter: "How can a decision maker choose between different actions when each possible choice leads not to a single outcome but to a range of outcomes with different probabilities?"

Another way to represent a situation like this is to draw a decision tree with different paths in the tree indicating different decisions and different states that can occur. This has been done for a simple problem in Figure 5.2. In this case there are just two choices of action for the decision maker and two states that can occur. This gives a combination of 4 different outcomes. Often we will write probabilities against the different states so that we can see the likelihood of different outcomes.

In the model we have described the arrows suggest a movement in time as well. It makes sense to start at the point where a choice of action is made, since our whole interest is in the best choice for the decision maker. So any uncertainty in the state of nature that is resolved before the decision point is no longer relevant. The state space needs to deal with all the uncertainty about the state of nature which will be resolved *after* the decision is made. Often the choice of action determines what happens, and thus the probabilities and possible states evolve differently depending on the action that is taken (we will see this in some of the examples that we look at later). From a conceptual point of view this makes things more complicated since it introduces a dependence between actions and states (rather than the interaction between the two only occurring with the outcome), but from a practical point of view there is no problem with drawing an appropriate decision tree and making the required calculations.

5.2 Expected Utility Theory

5.2.1 Maximizing expected profit

We want to start with the simplest situation so let us suppose that the only thing which matters for our business is the overall profit; so that means that there is a single dollar number associated with each possible outcome and we do not need to allow for any of the less quantifiable aspects of a decision. Sometimes even if there are other factors to consider we can price these to give a final result expressed in dollars. For example if we are considering moving a call centre operation overseas then the lower standard of spoken English could lead to our customers being less satisfied and this needs to be taken account of in our decision process. We need to ask ourselves what is the real cost of this lower level of service - perhaps we should think about how much we would need to spend on improvements on other parts of our operations in order to make our customers as happy overall with our service under the new arrangements with an overseas call centre as they are now. In any event if we can convert the service issues into a dollar number then we will have a better basis for making this decision.

The first proposal we might make, and perhaps the simplest choice for a manager, is to maximize expected profit. This is entirely reasonable. Given a number of possible courses of action a manager calculates the expected profit for each and then chooses the action with the highest expected profit value.

Consider Matchstock Enterprises who are considering investing in a new tunneling machine at a cost of $520,000 new. This is required for a specific job that Matchstock have taken on and will be sold in 2 years time for $450,000. A similar second hand machine is available right now at $450,000, and will be able to do the job satisfactorily. After two years this (now 4 year old) machine could be sold for $400,000. The main difference is in reliability and what happens if there is a breakdown. With a new machine parts will be covered throughout the first two years, leaving only the labour costs and the cost of lost working time. Matchstock have taken expert advice and believe that with a new machine they will have a 0.25 chance of a single breakdown and a 0.05 chance of 2 breakdowns

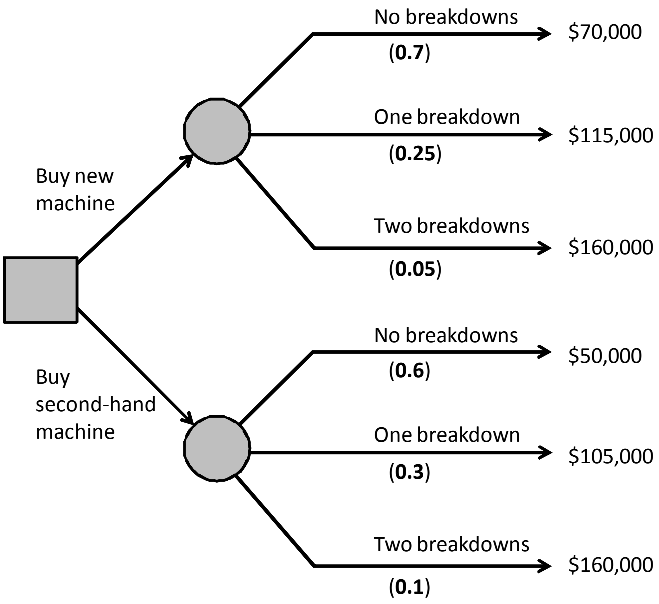


Figure 5.3 The possible outcomes and their probabilities for Matchstock Enterprises

during the two year project (with a probability of 0.7 of no breakdowns) and each breakdown will cost $45,000 in total. With a second hand machine they believe that there will be a 0.3 probability of a single breakdown and a 0.1 probability of 2 breakdowns (with 0.6 probability of no breakdowns). In this case each breakdown is estimated to cost $55,000.

Figure 5.3 shows how this is represented as a decision tree. In comparison with Figure 5.2 we have shrunk the nodes and put information (including probabilities) on the arrows. In this case we can calculate the total tunneling machine related cost for each of the possible outcomes and this is shown in the right hand column of the Figure. For example buying a new machine and getting two breakdowns leads to a cost of $70,000 from the loss in value of the machine over two years plus a cost of $45,000 incurred twice giving $160,000 in total. These total costs are taken off the profit to give a final profit, and so maximizing expected profit is equivalent to minimizing expected cost. We have taken no account in this of the ‘time value of money’ so there is no discounting of costs.

What should Matchstock do to maximize its expected profit? Buying a new tunneling machine incurs expected costs of $0.7 \times 70 + 0.25 \times 115 + 0.05 \times 160$ (in $\$1000$s), which works out to 85.75 thousand. Buying a second hand machine gives expected costs of $0.6 \times 50 + 0.3 \times 105 + 0.1 \times 160 = 77.5$ thousand. So choosing the second hand tunneling machine has lower costs and will maximize the expected profit.

This methodology could also be applied to the decision facing Suvi Marshall in the tomato purchasing scenario we gave at the beginning of the chapter. The uncertainty relates to the weather and hence the conditions of shortage or surplus. Looking at the problem in this way will encourage Suvi to make a more detailed investigation of likely waste figures alongside sales estimates for different scenarios. Where there is lack of information that too can be included as uncertainty in the analysis. In this case it seems unlikely that the decision tree analysis will produce an unequivocal recommendation to go with the fixed price scheme or not, but it will certainly help in establishing the critical parameters for this decision.

5.2.2 Utility

We want to introduce the idea that utility is more useful than money in comparing different options. A starting point is to show that most people, especially when dealing with personal decisions rather than decisions they take as a manager, will not just maximize expected profit. For example suppose that you wish to choose between the following two options:

Choice A: With probability 0.5 gain $\$1000$; and with probability 0.5 lose $\$800$,
Choice B: With certainty gain $\$99$.

In this case choice A has an expected profit of $0.5(1000) + 0.5(-800) = \$100$ and this is greater than the profit from choice B (a certain profit of $\$99$). So a decision maker maximizing expected profit will definitely choose A in preference to B.

However facing exactly this choice in practice most people would have no hesitation in choosing B. This is not a misunderstanding of the choices available, or a failure to do the simple arithmetic. Instead it represents a reaction to the unpleasantness of having to hand over $800 if a coin toss goes the wrong way. Given that there is only a $1 difference in the expectations the great majority of people will opt for the certainty of a $99 payoff. If we reflect on what is happening here then we can see that our choice depends both on our current financial circumstances and our taste for gambling.

This leads us to define an individual ‘utility’ function that determines how valuable (or how costly) it would be for us to gain (or lose) different amounts of money. In some form this can be traced back to the cousins Nicolas and Daniel Bernoulli who considered how players should behave in games of chance. A Swiss mathematician called Cramer in writing to Nicolas Bernoulli in 1728 talked of a utility function in the following way

> “The mathematicians estimate money in proportion to its quantity, and men of good sense in proportion to the usage that they may make of it.”

The idea here is that we may value $\$2000$ as being worth to us less than twice as much as $\$1000$ - it all depends on how we are likely to use the money.

One of the problems or paradoxes that had exercised the Bernoulli cousins is a game in which we are offered a prize that depends on tossing coins. If the first toss comes up heads then we win a dollar and the game finishes. If the first toss is a tail and the next toss is a head then we win two dollars and the game finishes. But if the first two tosses are tails and the next is heads then we win $\$8$ and the game finishes. More generally if we have n tails tossed and the $n + 1$’th toss is a head we will win $\$2^n$. If this is the arrangement, how much would we be prepared to pay to enter this game? Most people might pay $\$5$ or may be $\$10$ but no more. If we use an expected profit calculation then we should be happy to pay any amount less than the expected prize value in the game, so we need to calculate this expectation. Suppose that we play just three rounds. It is easy to see that we have a $1/2$ probability of getting $\$1$; a $1/4$ probability of getting $\$2$ and a $1/8$ probability of getting $\$4$. Our expected winnings

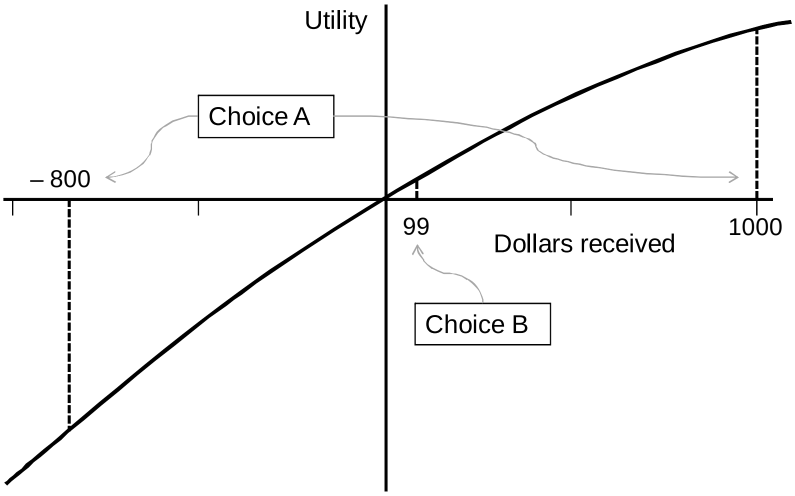


Figure 5.4 Comparing choices A and B using a utility function

are $(1/2) + (1/4) \times 2 + (1/8) \times 4 = 3/2$. But if we play on for more tosses we can see that each extra toss adds a term like $(1/2) \times (1/2^n) \times 2^n = 1/2$. After 50 tosses our expected winnings would be $\$25$. As we keep playing the sums of money become exponentially large but the probabilities of winning these amounts become tiny. Overall it is clear that the expected value is infinite.

We might well object that the amounts of money here are crazy - after 30 throws without a head we are set to receive $\$2^{30}$ which is more than a billion dollars. But the resolution that Daniel Bernoulli gave to this paradox (usually called the Petersburg paradox) rests on the idea of utility. For each extra toss the amount won doubles but the probability of winning this amount halves. From an expected profit viewpoint the contribution of each term is the same (and equal to 0.5 in the way we have done this calculation above). But Bernoulli argued that even if having $\$2$ was twice as valuable as $\$1$, receiving a prize of $\$200$ million dollars was not twice as valuable as receiving a prize of $\$100$ million. For most people $\$100$ million is such a large amount (enough to live in luxury without having to work) that it would be foolish to think that the additional benefit they receive from the extra $\$100$ million is of the same value as the first $\$100$ million. But this is the implication of a pure expected profit calculation; we should be indifferent between receiving $\$100$ million for sure and tossing a coin and getting $\$200$ million only if we win.

So these arguments lead us to a theory of decisions based on utilities. Instead of expected profit we need to look at expected utility, and the theory is called Expected Utility Theory (EUT). In EUT in order to compare the two choices A and B above we need to know the utility function that we have for different possible gains and losses. Figure 5.4 shows a possible utility function. On the positive side it curves downwards so that gaining $\$1000$ is less than twice as good as gaining $\$500$. On the negative side it also curves downwards so that losing $\$1000$ is more than twice as bad as losing $\$500$.

Looking at Figure 5.4 it is easy to see that, in this case, the negative utility at $-\$800$ is bigger than the positive utility at $\$1000$. Since choice A gives equal chance to these two possibilities, the expected utility for A will be negative. Using EUT we would not choose A even if the alternative was to receive nothing. Thus we certainly prefer choice B with a guaranteed gain.

Worked Example 5.1

Suppose that an investor is deciding between two investment funds: a growth fund and a stable fund. In the past three years there has been one bad year in which the stable fund lost 5% and the growth fund lost 25%; one good year in which the stable fund gained 15 % and the growth fund gained 30%; and one medium year in which the stable fund gained 5 % and the growth fund gained 10%. The investor has no way of knowing what will happen next year, but by looking at the relative probabilities of different outcomes she estimates that bad years occur three years in ten, good years occur three years in ten, and medium years occur four years in ten. Assume that the investor's utility for x thousand dollars is given by $\log(x - 1)$. Given she has $\$100,000$ to invest, which option has the highest expected utility?

Solution

Table 5.1 shows for each fund option the outcomes, the utilities (calculated as log to base e) and the expected values. For example the expected utility for the growth fund is calculated from $0.3 \times 4.317 + 0.4 \times 4.701 + 0.3 \times 4.867 = 4.636$.

Table 5.1: Different outcomes for the investment choice of Example 5.1

	Bad year	Medium year	Good year	Expected value
Probability	0.3	0.4	0.3	
Stable fund:				
Value	95	105	115	105
Utility	log(95) = 4.554	log(105) = 4.654	log(115) = 4.745	4.651
Growth fund:				
Value	75	110	130	106.25
Utility	log(75) = 4.317	log(110) = 4.701	log(130) = 4.867	4.636

We can see that, even though the growth fund has a higher expected value than the stable fund, the expected utility is a little lower than for the stable fund. So we can conclude that, with this utility function, the investor should choose the stable fund.

5.2.3 Expected Utility Theory from axioms

Expected Utility Theory is a powerful way to think about making decisions in a risky environment: and it represents a rational approach when maximizing expected profits is inappropriate. But it is interesting to ask whether there are other formulations we might use. For example we might consider some way in which we take account directly of the variance of the dollar outcomes rather than doing this indirectly through the shape of the utility function. Is our risk taking propensity something that should be considered over and above the utility function? It is a remarkable fact that Expected Utility Theory can be derived from three (very reasonable) axioms for choices: ordering; continuity and independence. The consequence is that, if we accept the axioms, then we do not need to consider any more complex decision algorithm.

Before embarking on a description of the axioms we need to introduce some notation and terminology. We will use the term *prospect* to describe a choice with a whole set of possible outcomes each with a probability (like the choices A and B in the previous example). Each prospect is a single (circle) node in the decision tree, representing a single choice for the decision maker.

More formally we assume that a prospect has a finite set of outcomes. A prospect q involves an outcome x_1 with probability p_1, an outcome x_2 with probability p_2,$\ldots$ and an outcome x_n with probability p_n. It is helpful shorthand to write a prospect as $(x_1, p_1;\ x_2, p_2;\ ...\ x_n, p_n)$ so that each outcome is followed by its probability. Often we will leave out a zero outcome so that the prospect $(\$100, 0.3;\ -\$200, 0.2)$ is taken as meaning a 0.3 probability of receiving $\$100$, a 0.2 probability of losing $\$200$, and a 0.5 probability of getting nothing.

If the set of consequences are simply dollar amounts then a prospect is just a discrete probability distribution on dollar outcomes. There are other terms that are used instead of 'prospect', (for example some authors use the terminology of lotteries) but we choose this term because it is conventional in the area of behavioral decision theory that we discuss in the next chapter. In this chapter all the distributions are discrete, we will not discuss any kind of 'continuous prospect'.

We use the symbol $\succcurlyeq$ to indicate preference between two prospects. Thus we say $q \succcurlyeq r$ when we mean that q is preferred to r. The way this is written, as a 'weak' inequality, is deliberate. It means that q might be chosen if both options are available, it does not mean that q is always chosen when both options are available. Thus it incorporates the possibility that the decision maker is indifferent between them. In fact if we are indifferent between q and r then both $q \succcurlyeq r$ and $r \succcurlyeq q$.

Now we introduce the three axioms. In each case we might want to consider how reasonable they are - if we accept all of them then we have to accept Expected Utility Theory as giving a complete description of the way that a rational decision maker should behave.

Axiom 1: Ordering

The ordering axiom really has two parts: 'completeness' and 'transitivity'. *Completeness* entails that for all choices q, r: either $q \succcurlyeq r$ or $r \succcurlyeq q$ or both. So that the decision maker has a consistent way of making decisions between prospects.

Transitivity requires that for all prospects q, r, and s: if $q \succcurlyeq r$ and $r \succcurlyeq s$, then $q \succcurlyeq s$. This again seems entirely obvious. If a decision maker prefers q to r, but finds r preferable to s, then it is hard to see how it could be wrong to choose q in preference to s.

Axiom 2: Continuity

Continuity requires that for all prospects q, r, and s where $q \succcurlyeq r$ and $r \succcurlyeq s$: we can find a probability p such that we are indifferent between r and the (compound) prospect which has q with probability p, and s with probability $1 - p$. i.e. we have $r \succcurlyeq (q, p;\ s, 1 - p)$ and $(q, p;\ s, 1 - p) \succcurlyeq r$.

Axiom 3: Independence

Independence requires that for all prospects q, r, and s: if $q \succcurlyeq r$ then $(q, p;\, s, 1-p) \succcurlyeq (r, p;\, s, 1-p)$, for all p. In other words, knowing that we prefer q to r, we should still prefer an option with q rather than r even if there is some fixed chance of a different prospect, s occurring.

If all three of the axioms hold, then it can be proved that preferences can be obtained from expected utilities for some "utility" function $u(\cdot)$ defined on the set of outcomes. We make this more precise as follows.

Theorem 5.1. (von Neumann and Morgenstern 1947) Suppose that a preference relation $\succcurlyeq$ on the set of all prospects satisfies the axioms of ordering, continuity and independence. Then there is a utility function, u, defined on all the possible outcomes, and a utility function U on prospects derived from u by

$$U(x_1, p_1;\, x_2, p_2;\, ...;\, x_n, p_n) = \sum_{i=1}^{n} p_i u(x_i)$$

with the property that for any prospects q and r, $U(q) \geq U(r)$ if and only $q \succcurlyeq r$. Moreover any utility function v on outcomes which has this property must be a positive linear transformation of u (i.e. $v(x) = a + bu(x)$ for a certain a and b with $b > 0$).

We give an idea of how the proof of this works in the section below. But for now we want to look in more detail at what the axioms claim.

We consider the continuity axiom. This supposes that there is some attractive option q: we might think of a prize of a 100 day around the world cruise, and a less attractive option s: no holiday prize at all, and finally an intermediate option r: a prize of a 14 day cruise around the Mediterranean. We don't have to worry about the fact that some people who hate spending time at sea might find the 100 day option q worse than the 14 day option r. In this thought experiment we are dealing with an individual who definitely has $q \succcurlyeq r$ and $r \succcurlyeq s$. Continuity as its name suggests is really all about watching what happens to the preferences when the chance of the better outcome is slowly increased. We suppose that there is a lottery in which there is a probability p of winning the prize of a 100 day round the world cruise, but if we don't win that prize we get nothing. This is the prospect $(q, p;\, s, 1-p)$. If p starts at zero then we cannot win the prize in the lottery and we would therefore prefer to have the result r than a ticket in the lottery. On the other hand by the time p reaches 1 the lottery is no longer a lottery; it always produces the prize. So with $p = 1$ a lottery ticket will be preferred to r. The axiom simply states that there must be an intermediate value of p at which the decision maker is indifferent between having a ticket for the lottery and taking the prize of r, a 14-day cruise. This is just the value of p at which we swap from preferring the outcome r to preferring the lottery ticket.

This axiom could perhaps fail if there was an infinitely bad outcome, so however small the probability of it occurring the prospect automatically becomes very bad. Of course death is the ultimate bad outcome here, and it is sometimes argued that this would make the continuity axiom break down. If q is getting a single dollar; r is getting nothing, but s is dieing, then obviously $q \succcurlyeq r$ and $r \succcurlyeq s$. But does it therefore follow that there will be some probability p

very close to 1 where we are indifferent between (A) getting nothing and (B) getting a dollar with probability p but losing our life otherwise?

Against this objection we may observe that we often in practice operate in exactly this way. A coffee costs $4 on this side of the street, but just as good a coffee costs only $3 on the other side of the street. However if we cross the street for the cheaper coffee aren't we running some tiny risk of being killed by a reckless driver? Gilboa observes that whether or not we cross the street may depend on how the decision is put to us:

> ... If you stop me and say "What are you doing?Are you nuts to risk your life this way? Think of what could happen! Think of your family!" I will cave in and give up [the dollar saving].

This demonstrates the importance of 'framing' which we discuss a little further in the next chapter: it would be foolish to assume that a real decision maker always makes the same decision when faced with the same two options.

Next we consider the axiom of independence. One way to think of this is to see it as related to decisions made ahead of time. If we prefer q to r, the 100 day cruise to the 14 day option then this should still be true even if a third option may end up occurring. In this example if s is the option of no prize at all, the axiom states that if given a straight choice between prizes q and r we prefer q, then given a lottery with say one in a 1000 chance of winning we would prefer to be holding a ticket for the otter with q as the prize than holding a ticket for the lottery with r as the prize; and this remains true no matter what the probability of winning provided it is the same for the two lotteries. If we play the lottery and lose then there is no decision to be made. If we play the lottery and win and are then offered a choice between the prizes q and r we have already said that we would choose q. There would thus be something odd about choosing the r lottery, if the q lottery were available. It would amount to a kind of 'dynamic inconsistency' where we make a different decision ahead of time to the decision we make when it comes to the final choice. Of course this sort of inconsistency does sometimes occur in our choices, but we might prefer to be more consistent and the independence axiom is simply a method to enforce that. In the next chapter we will explore in much more detail the way that the axiom of independence may fail for real decision makers facing real choices.

5.2.4 A sketch proof of the theorem

We will not give a complete proof of Theorem 5.1, but (if you have the mathematical interest) it is worthwhile looking at the most important steps in such a proof.

To do this we need to manipulate prospects and the first thing to note is that when prospects occur inside other prospects we can expand them in order to get to a single list of outcomes and their probabilities. For example if q_1 is the prospect which has $100 with probability 0.5 and $50 with probability 0.5 and $q_2 = (\$100, 0.5;\, q_1, 0.5)$ then we can expand the prospect q_1 to get

$$q_2 = (\$100, 0.5;\, 0.5(\$100, 0.5;\, \$50, 0.5)) = (\$100, 0.75; \$50, 0.25).$$

We suppose that we have a finite set of prospects which implies that there are a finite set of outcomes amongst which different prospects distribute their probabilities. We will work in stages.

Step 1 The ordering axiom can be used to establish a best and a worst outcome amongst the set of outcomes. We do this by taking the outcomes one at a time and comparing them with the best from amongst the outcomes already considered. The overall best must have been compared directly with every outcome which came after it in the order and by transitivity will also be better than everything that came before (we will not try to give any details). Here we are dealing with outcomes rather than prospects but an outcome is essentially the same as the prospect that assigns a probability 1 to that outcome. The same procedure works to find the worst outcome as well. We call the best outcome x^* and the worst x_*.

Step 2 Now we assign the utility value of 0 to x_* and a utility value of 1 to x^*. For any other outcome x in the list we assign a utility $u(x)$ equal to the probability p such that we are indifferent between the prospect $(x, 1)$ and the prospect $(x^*, p;\, x_*, 1-p)$ (using the continuity axiom)

Step 3 The independence axiom allows a free choice of the probability p and the exact alternative chosen s. Suppose that we are given numbers α and β with $\alpha \geq \beta$. We choose $p = \alpha - \beta$ and $s = (x^*, \beta/(1-p);\, x_*, (1-\beta/(1-p)))$ and substitute these values into the independence axiom Thus since $x^* \succcurlyeq x_*$ we get

$$(x^*, \alpha-\beta;\, x^*, \beta;\, x_*, 1-\alpha) \succcurlyeq (x_*, \alpha-\beta;\, x^*, \beta;\, x_*, 1-\alpha),$$

which simplifies to

$$(x^*, \alpha;\, x_*, 1-\alpha) \succcurlyeq (x^*, \beta;\, x_*, 1-\beta). \tag{5.1}$$

This is a useful intermediate result: If we form a prospect from two outcomes them increasing the probability of the better of the two makes the prospect more attractive.

Step 4 Now we show that the utility of the outcomes matches the preference ordering between them. Suppose that $u(x) \geq u(y)$, then from (5.1)

$$(x^*, u(x);\, x_*, 1-u(x)) \succcurlyeq (x^*, u(y);\, x_*, 1-u(y)).$$

But if we look back at how $u(x)$ and $u(y)$ are defined, we see that this is equivalent to

$$(x, 1) \succcurlyeq (y, 1).$$

Step 5 Now suppose that we have a prospect $(x, \alpha; y, 1-\alpha)$: we want to show that this has utility $\alpha u(x) + (1-\alpha)u(y)$ in other words we want to show indifference between $(x, \alpha; y, 1-\alpha)$ and $(x^*, \alpha u(x) + (1-\alpha)u(y);\, x_*, 1 - \alpha u(x) - (1-\alpha)u(y))$ We do this by observing that

$$(x, 1) \succcurlyeq (x^*, u(x);\, x_*, 1-u(x))$$

and hence

$$(x, \alpha; y, (1-\alpha)) \succcurlyeq (x^*, \alpha u(x);\, x_*, \alpha - \alpha u(x);\, y, (1-\alpha)). \tag{5.2}$$

But because we also have

$$(y, 1) \succcurlyeq (x^*, u(y); x_*, 1 - u(y)),$$

we can take the prospect on the left hand side and obtain the following from the independence axiom

$$\begin{aligned}&(y, (1 - a); x^*, au(x); x_*, a - au(x)) \\ &\quad\succcurlyeq (x^*, (1 - a)u(y); x_*, (1 - a)(1 - u(y)); x^*, au(x); x_*, a - au(x)). \qquad (5.3)\end{aligned}$$

The prospect on the right hand side can be simplified to

$$(x^*, \alpha u(x) + (1 - \alpha)u(y); x_*, 1 - \alpha u(x) - (1 - \alpha)u(y)).$$

Thus we can combine (5.2) and (5.3) by transitivity to obtain the relationship:

$$(x, \alpha; y, 1 - \alpha) \succcurlyeq (x^*, \alpha u(x) + (1 - \alpha)u(y); x_*, 1 - \alpha u(x) - (1 - \alpha)u(y)).$$

We can repeat all of this with the preference orders reversed to show

$$(x^*, \alpha u(x) + (1 - \alpha)u(y); x_*, 1 - \alpha u(x) - (1 - \alpha)u(y)) \succcurlyeq (x, \alpha; y, 1 - \alpha),$$

which finally establishes the indifference we require

There are a number of things we need to do in order to fill in the gaps here. First the result holds without any restriction on there being just a finite set of possible outcomes or prospects. This requires us to start with a finite set of prospects and then to add another set which lie outside this range (say they are all worse than the worst outcome in the first set) and stitch together the two utility functions we generate.

Second we have not fully included all the components of the argument we need in step 4 which shows that an inequality in utilities for outcomes translates into a preference order. We need to show that the same thing is true for prospects and we also need an if and only if argument.

Also in step 5 we have demonstrated what we want for a prospect with just two alternatives - we need to extend this to prospects with any number of alternatives.

Finally we have not dealt with the uniqueness of the utility function (up to positive linear transformations) - the theorem will not allow us to, for example, square all the utility values. This would leave the ordering of individual outcomes unchanged, but we would lose the ability to get the utility of a prospect as the probability weighted combination of the individual outcome utilities.

5.2.5 What shape is the utility function?

Where outcomes are monetary then the utility function is simply a real valued function defined on the real line. It is helpful to use the terminology of convex and concave functions. A convex utility function has the property that its slope is increasing (or to put it another way it has a second derivative which is non-negative). Another way to characterize a convex function is to say that a straight line joining two points on the graph of the function curve can never go below the function. This can be put into mathematical form by saying that a function $u(\cdot)$ is convex if for any p between 0 and 1,

$$p_1u(x_1) + (1 - p_1)u(x_2) \geq u(p_1x_1 + (1 - p_1)x_2).$$

The left hand side is the height of a point a proportion $1 - p$ along the straight line between points $(x_1, u(x_1))$ and $(x_2, u(x_2))$ and the right hand side is the point on the curve itself at this x value. This property of convex functions can be generalized to any number of points. So a convex function u has the property that

$$\sum_{j=1}^{n} p_j u(x_j) \geq u\left(\sum_{j=1}^{n} p_j x_j\right),$$

if the p_j are nonnegative and $\sum_{j=1}^{n} p_j = 1$.

The connection with expected utility is that a convex utility function implies risk seeking behavior. If there is a prospect having a probability p_1 of achieving x_1 and a probability $(1 - p_1)$ of achieving x_2, then the expected utility is $p_1u(x_1) + (1 - p_1)u(x_2)$ which we prefer to (its value is greater than) the utility $u(p_1x_1 + (1 - p_1)x_2)$ that we obtain from the expected result. Thus under this condition it will always be preferable to choose a prospect involving uncertainty, rather than having the expected outcome with certainty.

The reverse is true for a concave utility function. A concave function is one where a straight line joining two points on the graph of the function curve can never go above the function. In this case having the expected outcome $\sum p_j x_j$ with certainty is always preferable to facing the gamble involved in the uncertain prospect. In other words a concave utility function like the one shown in Figure **??** implies risk averse behavior.

Of course we can also have utility functions which are convex in some areas and concave in others. For example, suppose that the utility for a wealth of x measured in \$100,000 units is $\sqrt{x} - 0.9\log(x + 1)$. Though it is not obvious, this function turns out to be positive for positive wealth: we draw it in Figure 5.5 for x in the range 0 to 3. We can see that the curve is concave for x below about 1. In fact it is convex for values above that.

Now suppose that the individual currently has wealth \$100,000 corresponding to $x = 1$. There is a small risk of a fire destroying \$75,000 worth of property. This happens next year with probability $1/1000$ and the insurance company charges \$100 to fully insure against this loss. Should the individual take out insurance? The expected loss is only $\$50000 \times (1/1000) = \$75$ so the insurance company is making quite a lot of money at this premium level. However we can do the calculation from the point of view of the individual. Current utility is $\sqrt{1} - 0.9\log(2) = 0.37617$. The expected utility if we take out the insurance can be calculated as the utility we have after paying the insurance premium $= \sqrt{1 - 0.001} - 0.9\log(2 - 0.001) = 0.376\,12$. The expected utility if we do not take out insurance is

$$0.001u(25000) + 0.999u(100000) = 0.001 \times 0.299171 + 0.999 \times 0.37617 = 0.37609$$

This is less than the expected utility if we do insure, and so insurance makes sense.

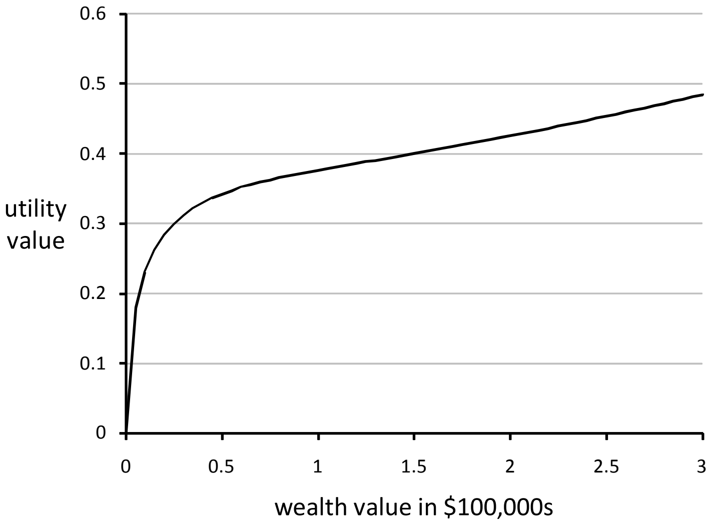


Figure 5.5 A curve showing utilities for different wealth values, using the formula $u(x) = \sqrt{x} - 0.9\ln(1+x)$

Now we look at a different choice offered to the same individual. There is an opportunity to enter a lottery where a single ticket costs $500 but there is a one in a thousand chance of winning a prize worth $500500 (i.e. we get $500,000 and also the price of our ticket back). We can do the sums again. Not buying the lottery ticket leaves us at the current utility of 0.37617, if we do buy the ticket we have a 0.001 probability of winning $500,000 and a 0.999 chance of losing $500. This gives a final expected utility of

$$\begin{aligned}&0.001u(600000) + 0.999u(99500)\\&= 0.001 \times (\sqrt{6} - 0.9\log(7)) + 0.999 \times (\sqrt{0.995} - 0.9\log(1.995))\\&= 0.37624\end{aligned}$$

So entering the lottery gives a slightly higher expected utility. The same individual is risk averse on losses (enough to buy a rather expensive insurance product), but risk seeking enough on gains to enter a lottery.

This particular utility function has a derivative $(1/2)x^{-0.5} - 0.9/(x+1)$ which approaches zero as x gets large. If we plot the utility function for much larger values of x we can see that it is actually concave (the second derivative becomes negative for $x \geq 8.163$). Hence the function moves from concave to convex and back to concave.

5.2.6 Expected utility when probabilities are subjective

The development of von Neumann Morgenstern style expected utility theory is fundamental to our understanding of decision making under uncertainty, but it is in a way a lopsided development. The assumption is that utilities are unknown (they are deduced from the choices made between prospects) but that probabilities are known precisely.

As we pointed out earlier decisions often involve choices that are impacted by events that we can have no control over and where even the idea of a probability needs to be treated carefully. It is interesting to ask how a decision maker might behave if she was not prepared to specify fixed probabilities for different events. After all in everyday life we routinely make decisions without stopping to ask about probabilities. So if a manager makes a decision without consciously thinking about probabilities, but at the same time is entirely rational and thoughtful about those decisions, can we deduce that there is some consistent decision framework that doesn't use probabilities?

The answer to this question is a qualified 'No'. In an important piece of work by Leonard 'Jimmie' Savage published in 1954 it is shown that if choices satisfy some reasonable seeming axioms then the decision maker must act as though he were assigning a subjective probability to each of the possible outcomes that can arise from a choice he makes, as well as a utility function on those outcomes, and then decide between choices on the basis of maximizing expected utility as determined by the subjective probabilities.

Putting all this into a solid theory requires a great deal of care. In Savage's model a decision maker has a choice between different actions and these actions will determine the outcomes that go along with the states. Formally we list all possible states as a set S and an action is treated as a function taking states to outcomes, where X is the set of outcomes. The set of states has to resolve all uncertainty, so that the action simply specifies what outcome occurs in each of the possible states. This way of thinking does not fit well with a decision tree framework, where we imagine taking the decision first followed by the uncertainty being resolved to lead to a particular outcome. In some ways it is like reversing this process: we think of the uncertainty being resolved to a single state and then the decisions map each state to an outcome. The two ways of thinking are not really so different; in each approach a decision and a state of nature together produce an outcome. The reason for proceeding with the more complex idea of actions as maps from states to outcomes is that we will need to make comparisons between all possible actions. That is we need to be able to imagine an action that specifies particular outcomes for each possible state of nature, without restricting in any way which outcomes go with which states. Having once imagined such an action, we then need to be able to compare it with any other imagined action and answer the question: Which would be preferable?

To get a flavour of the axioms we will describe three of them (there are seven in total).

Axiom P1 This states that there is a weak order on the actions. So for actions $f : S \to X$ and $g : S \to X$ either $f \succcurlyeq g$ or $g \succcurlyeq f$ or both, and this relationship is transitive.

Axiom P2 This axiom shows that in comparing actions we only care about the states where the two actions produce a different result. So if $f \succcurlyeq g$ and there is a set $A \subset S$ with $f(s) = g(s)$ for all states s in A, then the preference between f and g is determined by what is happening for $s \notin A$. More precisely we can say that if $f'(s) = f(s)$ for $s \notin A$ and $g'(s) = g(s)$ for $s \notin A$ and $f(s) = g(s) = h(s)$ for $s \in A$ then $f' \succcurlyeq g'$.

Before giving the third axiom we need two preliminaries. First we need to be able to make a comparison between outcomes rather than actions. This is simple enough, we say that for outcomes $x, y \in X$, $x \succcurlyeq y$ if the action that takes every state to x is preferred to the action that takes every state to y. If we know we are going to end up with x under some action f and we know we are going to end up with y under the action g then it no longer matters what is happening with the states.

Second we need to define a property of a set of states (i.e. an event) $A \subset S$ which amounts to saying there is a non-negligible possibility that one of the states in A occurs. The most natural way to describe this is to say that the probability of A is greater than zero, but with the Savage theory we do not have probabilities to work with. Instead we say that the event A is *null* if for any f and g that differ only on A these two actions are equivalent (both $f \succcurlyeq g$ and $g \succcurlyeq f$). In other words what happens for null events makes no difference to our preferences.

Axiom P3 This states that if outcome x is preferred to outcome y, and two actions f and g differ only on a set of states A which is not null and moreover on A, f produces x and g produces y then f must be preferable to g. Actually the axiom says more than this since it also says that the reverse implication is true (equivalently we require that if x is strictly preferred to y then f is strictly preferred to g). Formally we can write this as follows. For an event A that is not null, if $f(s) = x$ for $s \in A$, $g(s) = y$ for $s \in A$, and $f(s) = g(s) = h(s)$ for $s \notin A$, then

$$x \succcurlyeq y \text{ if and only if } f \succcurlyeq g$$

At this point you may well feel that we have stretched our minds enough without the need to go into further details. The remaining four axioms are a mixed bunch. Axiom P4 relates to a situation where outcome x is strictly preferred to y, and outcome z is strictly preferred to w. The axiom states that if the action which delivers x on A and y otherwise is preferred to the action which delivers x on B and y otherwise, then the action which delivers z on A and w otherwise is preferred to the action which delivers z on B and w otherwise. This is more or less the same as saying that A happens more often then B, but of course we cannot use this language since we do not have a notion of probability yet defined. Axiom P5 simply states that there must be two actions which are not equivalent to each other. Axiom P6 is related to continuity and implies that we can always partition the set of states S sufficiently finely that a change on just one component of the partition leaves a strict preference unchanged. This is quite a strong assumption and it can only work if there is an infinite state space S, and each individual state s in S is null. Often this is not true for problems of interest, but we can make it true by expanding our state space to consider some infinite set of (irrelevant) other events occurring in parallel with our original states. The final axiom P7 is only needed when the outcome set X is infinite, and we will not give a description of it.

To state Savage's theorem we have to understand what might be meant by an assignment of probabilities to all the states in S. This requires a measure μ which assigns a probability to every event $A \subset S$. The measure has to be finitely additive (so $\mu(A \cup B) = \mu(A) + \mu(B)$ if A and B are disjoint), and it also has to be non-atomic in the sense that if there is an event A with $\mu(A) > 0$ and we choose any r a number between 0 and $\mu(A)$, then we can find a subset $B \subset A$ with $\mu(B) = r$. Once we have a measure μ on the states S and a scalar function defined on the states then we can evaluate the expectation of that function by writing its integral with respect to μ. Thus finally we are ready to state the theorem.

Theorem 5.2 (Savage) When X is finite, the relationship $\succcurlyeq$ satisfies axioms P1 to P6 if and only if there exists a non-atomic finitely additive probability measure μ on S and a

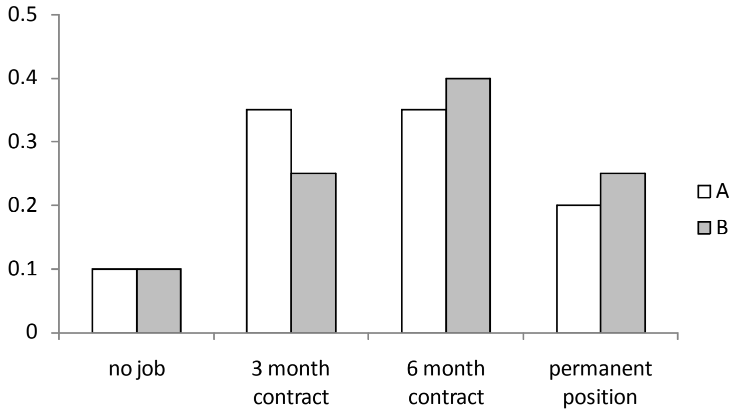


Figure 5.6 Risk profile for the two options for Raj

non-constant function $u(x)$, for $x \in X$ such that for any actions f and g

$$f \succcurlyeq g \text{ if and only if } \int_S u(f(s))d\mu(s) \geq \int_S u(g(s))d\mu(s)$$

Furthermore, in this case μ is unique and u is unique up to a positive linear transformation.

5.3 Stochastic dominance and risk profiles

Suppose that we order the outcomes for a prospect from the worst x_1 to the best x_n. This could be done even if the outcomes do not have monetary values. We can then draw a risk profile and use this to compare two different prospects. For example suppose that a short term contract employee Raj currently has a contract position for 6 months. If Raj does nothing he believes that there is a 10% chance of his job not being renewed at the end of the contract period, a 35% chance of the job being renewed for another 3 months, a 35% chance of the job being renewed for another 6 months and a 20% chance of the job being made permanent. What happens will depend both on Raj's performance and the trading performance of the company. Raj still has 2 months of his existing contract to run, but believes he has some skills that will help in getting a permanent position and that might be overlooked in the normal process. So he is considering going to his boss and making the case for a permanent appointment straight away. He believes that this will increase the chance of his being made permanent to 25% and in this case the probabilities of the other outcomes are 10% job not renewed; 25% job renewed for 3 months; 40% job renewed for 6 months. What should Raj do? The risk profile for these two choices are shown in Figure 5.6.

It is not obvious how to make a comparison between the actions A-'do nothing' and B-'ask for permanent position' from these risk profiles alone. The complication here is that we have not specified how much more valuable a permanent appoint is than a 6 months one. And we also have to consider the relationship of a 6 month to a 3 month appointment; all

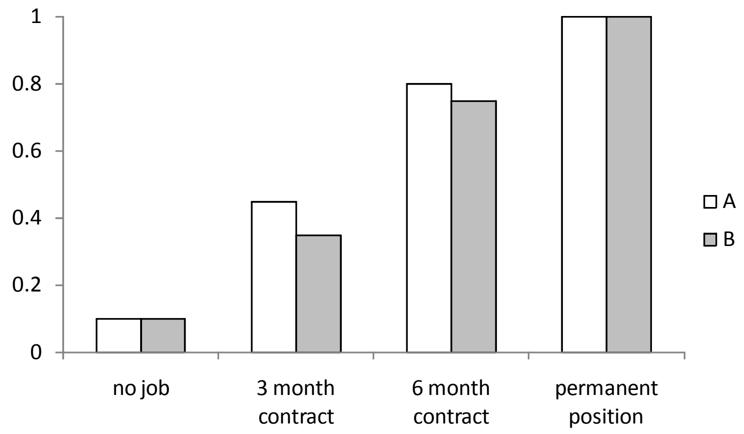


Figure 5.7 Cumulative risk profile for the two options for Raj

that we know is that there is a preference order: permanent is best, then 6 months is better than 3 months and losing the job is worst. The right approach here is to draw a cumulative risk profile as is shown in Figure 5.7. Here it becomes clear that action A has a cumulative risk (of a worse outcome) that is always higher or equal to the cumulative risk for action B; there is a kind of dominance between the two. Later we will show that this implies that it is better for Raj to choose option A and make the case for a permanent appointment, no matter what the exact utility values he places on the different outcomes. But first we need to more carefully define what is meant by stochastic dominance.

Given two prospects q and r we can take the combined set of all possible outcomes and put them in order of increasing value or utility. Notice that there is no loss of generality in comparing two prospects in assuming they have the same set of possible outcomes, and if one of the prospects does not include one of the outcomes we simply assign zero probability to that outcome.

We say that q stochastically dominates r if when we take the combined set of outcomes with probabilities q_i for q and r_i for r (and where q_i and r_i relate to the same outcome x_i), the following inequalities all hold

$$\sum_{j=m}^{n} q_j \geq \sum_{j=m}^{n} r_j \text{ for } m = 2, 3..., n \tag{5.4}$$

with at least one of these inequalities being strict. Each of these sums goes from an outcome m through all the better outcomes up to the best. Thus prospect q gives higher probabilities to the higher x_j which are preferable.

There is an alternative definition which uses the fact that the sum of all the q_j is 1 (since they are probabilities). Using this we see that (5.4) can be rewritten

$$1 - \sum_{j=1}^{m-1} q_j \geq 1 - \sum_{j=1}^{m-1} r_j \text{ for } m = 2, 3, ...n.$$

This can be rewritten as (check that you can follow the argument which involves swapping the sums to the other side of the inequality and setting $l = m - 1$)

$$\sum_{j=1}^{l} r_j \geq \sum_{j=1}^{l} q_j \text{ for } l = 1, 2, ...n - 1. \tag{5.5}$$

This is the inequality that we can see holds in the case of Raj's choice, from observing the cumulative risk profile of actions A and B in Figure 5.7.

We can apply the same logic when there are monetary outcomes. For example the prospect $A = (\$100, 0.2; \$160, 0.2, \$200, 0.6)$ stochastically dominates the prospect $B = (\$100, 0.3; \$120, 0.1; \$180, 0.1; \$200, 0.5)$. We can see this from the Table 5.2 which gives for each prospect the same set of \$ values and both probabilities and sums of the form (5.4). Each element in the column that shows the sum for A is greater than the corresponding element in the sum for B.

Table 5.2. Comparison of prospects A and B

Dollar amount	A probabilities	Sum for A	B probabilities	sum for B
\$100	0.2	1	0.3	1
\$120	0	0.8	0.1	0.7
\$160	0.2	0.8	0	0.6
\$180	0	0.6	0.1	0.6
\$200	0.6	0.6	0.5	0.5

Again we can if we wish look at the sums of probabilities of receiving up to a certain amount and use the inequalities (5.5). More generally if the random variables X and Y are defined, then we say that X stochastically dominates Y if the cumulative distribution functions F_X and F_Y, for X and Y, satisfy $F_X(x) \leq F_Y(x)$ for every x with strict inequality for some x. This definition of stochastic dominance does not need the random variabels to be defined over just a finite set of outcomes, since we can apply it equally well when X and Y are continuous rather than discrete random variables.

Under Expected Utility Theory if the prospect q stochastically dominates the prospect r then q will be preferred to r no matter what utility is given to the individual outcomes. the only requireemnt is that the utilities are strictly increasing, so $u(x_1) < u(x_2) < ... < u(x_n)$.

The result can be established as follows:

$$\begin{aligned} U(q) - U(r) &= \sum_{j=1}^{n} (q_j - r_j) u(x_j) \\ &= u(x_1) \sum_{j=1}^{n} (q_j - r_j) + (u(x_2) - u(x_1)) \sum_{j=2}^{n} (q_j - r_j) \\ &\quad + (u(x_3) - u(x_2)) \sum_{j=3}^{n} (q_j - r_j) + ... + (u(x_n) - u(x_{n-1}))(q_n - r_n) \\ &> 0. \end{aligned}$$

Here we have broken up a simple sum into a set of separate sums: to see that this works look at the multiplier for $u(x_1)$. We get $\sum_{j=1}^{n}(q_j - r_j)$ from the first term and $-\sum_{j=2}^{n}(q_j - r_j)$ from the second, so that everything cancels except the $(q_1 - r_1)$ that we want. The same idea applies to $u(x_2)$ which appears in the second term and the third, and so on for each of the $u(x_i)$. The final inequality follows because each term is non-negative and at least one is strictly positive. So we have shown that $U(q) > U(r)$ and if q has the higher expected utility then it will be preferred.

The reverse of this statement is also true: If a prospect q does not stochastically dominate a prospect r then there is some assignment of utility values to outcomes that makes r preferable to q. We can make this statement more precise: if there is any value of m for which $\sum_{j=m}^{n} q_j < \sum_{j=m}^{n} r_j$ (no matter what happens with all the other sums) then there is a choice of $u(x_1), u(x_2), ...u(x_n)$ which makes r preferable to q and still has $u(x_1) < u(x_2) < ... < u(x_n)$. We do this by setting $u(x_i) = i\delta$, for $i = 1, 2, ..., m-1$ and $u(x_i) = 1 + i\delta$, for $i = m, m+1, ..., n$. Then

$$
\begin{aligned}
U(q) - U(r) &= u(x_1)\sum_{j=1}^{n}(q_j - r_j) + (u(x_2) - u(x_1))\sum_{j=2}^{n}(q_j - r_j) \\
&\quad + (u(x_3) - u(x_2))\sum_{j=3}^{n}(q_j - r_j) + ... + (u(x_n) - u(x_{n-1}))(q_n - r_n) \\
&= \delta\left(\sum_{j=1}^{n}(q_j - r_j) + \sum_{j=2}^{n}(q_j - r_j) + ...\sum_{j=2}^{n}(q_j - r_j)\right) + \sum_{j=m}^{n}(q_j - r_j)
\end{aligned}
$$

This final expression arises by noting that the difference between successive $u(x_i)$ values is always δ except for the one occasion when it is $1 + \delta$ and this gives rise to the final term in the expression. Now it is clear that for small enough δ the last term dominates and hence $U(q) - U(r) < 0$ and prospect r is preferred to prospect q.

5.4 Risk decisions for managers

Many of the examples we have given in this chapter have focussed on an individual making a choice with implications for the individual. Now we want to turn to decisions taken by companies and specifically the managers or boards of those companies. Our starting point is to ask what the utility function will look like for business decisions rather than personal ones. The von Neumann Morgenstern result suggests that there should be an underlying utility function, but what is it? There are a number of issues to consider.

Managers and shareholders

We need to begin by thinking about who makes decisions and what they may be aiming to achieve. It is usual to think about management as acting in the best interest of shareholders who are the owners of the company, but in reality we have complex systems of corporate governance with a board of directors given the responsibility of hiring and monitoring top management in order to safeguard the interests of the shareholders.

Most shareholders have the opportunity to diversify their holdings across multiple firms. There are exceptions to this when shareholders wish to maintain control or influence by holding a significant fraction of the shares, for example in companies controlled by family interests. Usually, however, the majority of shares are owned by institutions or individual investors who are well diversified and consequently are likely to see only small proportional changes in their overall wealth from changes in the value of a single firm. From this we can deduce that shareholders will be risk neutral in their view of the actions of an individual firm.

To see why this is so, consider a firm that can pay $\$k$ million to take a risky action that delivers either nothing or $\$1$ million each with probability 0.5. If a shareholder with a proportion δ of the firm's equity was to make the decision on what is a fair value of k then the shareholder with current wealth W and utility function u should compare $u(W)$ (manager does nothing) with $0.5u(W - \delta k) + 0.5u(W + \delta(1 - k))$ (manager takes risky action). The shareholder operating on an expected utility basis would want the manager to take the risk provided that

$$0.5u(W - \delta k) + 0.5u(W + \delta(1 - k)) > u(W)$$

But for small δ we can approximate the left hand side of this expression using the derivative of u at W (which we write $u'(W)$) to get

$$\begin{aligned} &0.5u(W) - 0.5\delta k u'(W) + 0.5u(W) + 0.5\delta(1 - k)u'(W) \\ &= u(W) + 0.5\delta(1 - 2k)u'(W). \end{aligned}$$

Hence for any k less than 0.5 this is a worthwhile investment from the shareholder perspective, but it is not worthwhile if k is greater than 0.5. We end up with the risk neutral value put on the investment.

As an aside we need to observe that this argument should not be seen as suggesting that investors are completely indifferent to risk. The capital asset pricing model (CAPM) explores how the component of an asset's variance that cannot be diversified away (its β value or systematic risk) is reflected in the price of the asset. But the type of management decision we are considering here, that is idiosyncratic to this particular firm, would not appear in β.

But if diversification makes shareholders risk neutral the same cannot be said for the managers. A manager is committed to the firm in a way that the investor is not and the success of a manager's career is tied to the performance of the company. Moreover a senior manager may well hold stock options which also give her a direct interest in the company share price. This can be expected to lead to risk averse behavior by a manager and this differs from the risk neutral behavior that would be preferred by investors. So there is a potential *agency* problem where managers who in theory acting as agents of the owners are in fact subject to different incentives.

A single view of risk

In many cases it is convenient to view a company as a single entity; perhaps this is connected to the legal fiction of a company as an individual. But in practice it is clear that there will be differences between the type of actions and choices made by one manager over against another within the same company. This corresponds to the different personalities involved: one manager may be particularly risk averse by nature, whereas her colleague at the next desk is a natural gambler. When this happens there is a danger of inefficiency, since one manager might pay a risk premium to achieve some measure of certainty, only for this to be negated at the firm level by the relatively risky behavior of a second manager. The Dennis Weatherstone approach at JP Morgan (discussed at the start of Chapter 3) certainly attempts to bring the entire company under a single risk umbrella. The same ideas also lie behind the idea that firms need to determine an appropriate risk appetite for the firm as a whole (one of the tenets of Enterprise Risk Management) which implies that this idea can be discussed and agreed within the top management team. In practice, however, it is not so easy to obtain a uniform approach to risk across the whole of a company.

One problem with the attempt to have a single level of risk appetite for the firm as a whole is that if levels of risk appetite are related to the shape of the utility function then they may depend on factors like the size of the 'bet' and the current company cash reserves. This explains why it is so hard to give a simple definition of risk appetite: it cannot just be seen as a point on a scale moving from risk averse through to risk seeking.

In Chapter 1 we commented on the way that those involved in trading operations often take a different view of risk than the back office and it is common to have different levels of risk preference at different levels in a hierarchy, or in different departments. At some level this can be traced back not only to different individual risk preferences (as we see in the caricature of a trader as being a fast living young man burning out on the adrenaline rush of millions of dollars riding on instant decisions) but also to differences in reward structures, either explicitly through bonuses or implicitly through what is valued within the culture of a work group.

All of this will make us wary of assigning a single level of risk appetite to a company as a whole. It is rarely as simple as that. The complications that arise when dealing with this issue make the 'quick and dirty' approach of measuring Value at Risk and using this on a company wide basis seem more attractive.

Risk of insolvency.

Whether or not managers are risk averse for most of the time, they certainly become so if there is a threat of insolvency. This would suggest that there is a concave utility function dropping sharply as the solvency of the company becomes an issue. In fact once a company enters what is called the *zone of insolvency* (i.e. when insolvency is a real possibility) then the board will (or should) change the way that it behaves and the decision processes that are used. Company law may involve individuals on the board becoming personally responsible if the business fails, so it is not surprising that if insolvency is a possibility then the board will act promptly. One aspect of this is that if a company is insolvent then the creditors have a higher claim on the company assets than the shareholders, and so for a company that is in the zone of insolvency the directors should not act in a way that would prejudice the interests of the creditors over against the equity holders if insolvency occurs.

What will the utility look like as a function of total net assets (assets minus liabilities) for a company that faces the possibility of insolvency? The discussion here suggests that the utility function might look something like the curve shown in Figure 5.8. Once the firm goes out of business then the degree of insolvency (the extent of the company debts) will determine how much the creditors receive. Managers will have some interest in seeing creditors given a reasonable deal but the slope of the utility function in this region will be relatively flat. At first sight a utility function as shown in the figure suggests the possibility of risk seeking behavior. To take a simplified example suppose that a company with current net assets of just $\$100,000$ if allowed to trade normally produces a 10% chance of net assets moving to $-\$0.5$ million, a 70% chance of net assets remaining at $\$100,000$ and a 20% chance of net assets increasing to $\$0.4$ million. Thus the expected value of the net assets stays at $\$100,000$.

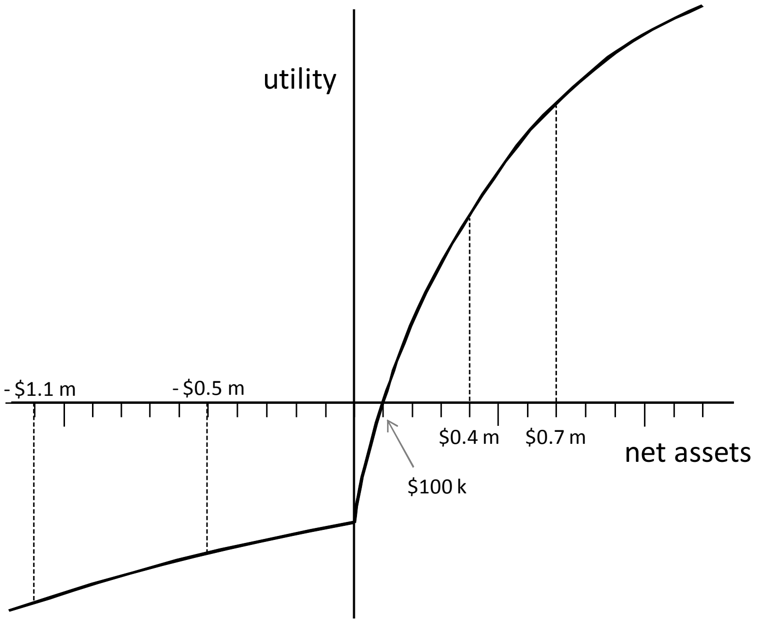


Figure 5.8 The utility function for a company facing insolvency

Now suppose that some more aggressive trading strategy could increase the size of both the gains and the losses while leaving the expected value unchanged. Specifically the aggressive strategy gives a 10% chance of net assets moving to $-\$1.1$ million, a 70% chance of net assets remaining at $\$100,000$ and a 20% chance of net assets increasing to $\$0.7$ million. Given the shape of the utility function shown in Figure 5.8 the aggressive trading strategy is easily seen to have a higher expected utility, since the gain in utility in moving from $\$0.4$ to $\$0.7$ million exceeds the loss in utility in moving from $-\$0.5$ to $-\$1.1$ million and occurs with greater probability. The corner in the utility function produces a region of convexity and hence risk seeking behavior.

In practice, however, the scenario sketched above is unlikely to happen since it supposes that the aggressive strategy can allow a much higher net deficit to be created, whereas the company should cease trading as soon as the net assets become zero. Moreover the directors would carry significant personal risk that the aggressive trading strategy would be found to be improper, whereupon they would individually carry some liability for the debts.

Notes

The book by Peter Bernstein gives more information about the development of the idea of utility by Daniel Bernoulli and others. The decision tree ideas that we present are quite standard and can be found in any textbook on Decision Theory.

An excellent book which goes quite deeply into the different frameworks that underlie the use of Expected Utility Theory is ‘Theory of Decision under Uncertainty’ by Itzhak

Gilboa. This is a book which addresses some of the important philosophical and conceptual underpinnings of decision theory, as well as being an entertaining read. Note though that he gives a slightly different form of the continuity axiom in describing the von Neumann-Morgenstern theory. This is arguably a weaker assumption, but slightly increases the complexity of understanding what is going on.

The discussion of Savage's theorem is drawn from Gilboa's book. This theory is designed around a situation with a rich state space (infinite with every individual element null). There are alternative approaches that need a rich outcome space, but a simpler state space, and Wakker (2010) gives such a development.

References

Peter Bernstein, *Against the Gods*, Wiley,1996.

Itzhak Gilboa, *Theory of Decision under Uncertainty*, Cambridge University Press, 2009.

Peter Wakker, *Prospect Theory for Risk and Ambiguity*, Cambridge University Press, 2010.

Exercises

5.1. (Making the right bid)

Xsteel is bidding to supply corrugated steel sheet for a major construction project. It knows that Yco is likely to be the only other serious bidder. There is a single price variable and the lower price bidder will win, and if both prices are the same then other factors will determine the winning bid with Xsteel and Yco equally likely to win. Xsteel believes that Yco will make an offer of either \$800 or \$810 or \$820 or \$830 per ton with each of these possibilities equally likely. Xsteel has a production cost of \$790 per ton and only bids at multiples of \$10 per ton are possible. What price bid will maximize Xsteel's expected profit?

5.2. (EUT and a business venture)

James has \$1000 which he wants to invest for a year. He can put the money into a savings account which pays an interest rate of 4 percent. His friend Kate asks him to invest the money in a business venture for which she needs exactly \$1000. However Kate's business will fail with probability 0.3 and if this happens James will get nothing. On the other hand if the business succeeds it will make \$2000 and this money can be used to repay the loan to James. The time taken to repay the money if the venture succeeds will be one year.

(a) Using EUT and assuming that James is risk-neutral (his utility is linear), how much would Kate have to repay James in order to convince James to lend her the money?

(b) Assume James is risk-averse with concave utility function $u(x) = \sqrt{x}$. How much would Kate have to repay James if the business venture succeeds in order to convince James to lend her the money?

5.3. (Calculating utilities from choices)

A researcher is trying to understand the utility function of an entrepreneur with a company worth $\$1$ million. He does this by describing various potential ventures and asking the manager whether she would take on this venture under the terms described. By changing the probabilities assigned to success and failure he finds three ventures where the manager is indifferent between taking them or not. The probabilities on different outcomes are given in Table 5.3.

Table 5.3. Outcomes for different ventures

Outcome:	Lose \$0.5 million	Gain \$0.5 million	Gain \$1 million	Gain \$1.5 million
	Probabilities:			
Venture A	0.4	0.6	0	0
Venture B	0.6	0	0.4	0
Venture C	0.7	0	0	0.3

Use this information to estimate the utilities on the different firm values: \$1.5 million, \$2 million and \$2.5 million assuming that the utility for firm value \$0.5 million is 1 and the utility for firm value \$1 million is 2. (These two values can be set arbitrarily because utilities are only defined from choices up to a positive linear transformation: see theorem 5.1). Sketch a possible utility function for the entrepreneur.

5.4. (Stochastic dominance and negative prospects)

Show that if a prospect $A = (x_1, p_1; x_2, p_2; ...x_n, p_n)$ stochastically dominates the prospect $B = (y_1, q_1; y_2, q_2; ...y_n, q_n)$, then the prospect $-B = (-y_1, q_1; -y_2, q_2; ... - y_n, q_n)$ stochastically dominates the prospect $-A = (-x_1, p_1; -x_2, p_2; ... - x_n, p_n)$.

5.5. (Stochastic dominance and normal distributions)

The profits from sales of a product depend on demand which follows a normal distribution. The demand in week 1 has a distribution with mean 1000 and standard deviation 100. The demand in week 2 has mean 1010. .

(a) Suppose that the standard deviation of demand in week 2 is 95. Explain why the profit in week 2 stochastically dominates the profit in week 1, and this result does not depend on the exact relationship between profit and sales.

(b) Show that if demand in week 2 has standard deviation of 105 then the profit in week 2 will not stochastically dominate the profit in week 1.

(c) Now suppose that demand in week 3 drops: it has a normal distribution with mean 200 and standard deviation 100, where any negative demand is simply set to zero (for parts (a) and (b) the possibility of negative demand can be ignored, but with the lower mean it cannot). Sketch the cumulative distribution function for demand in week 3. Suppose that demand in week 4 has mean 205 and standard deviation μ. Find the largest value of μ for which the demand in week 4 will stochastically dominate the demand in week 3.

5.6. (Failure of stochastic dominance)

Consider the following two prospects: $q = (\$100, 0.1;\ \$300, 0.2;\ \$400, 0.2;\ \$700, 0.3;\ \$900, 0.2)$ and $r = (\$300, 0.3; \$500, 0.3; \$700, 0.2; \$900, 0.1; \$1000, 0.1)$. Show that neither stochastically dominates the other and find a set of utility assignments where q is preferred and a set of utility assignments where r is preferred. Your utility values should satisfy the requirement that having more money gives strictly higher utility.

6
Understanding Risk Behavior

The economics of extended warranties

A consumer who buys either an electronic item (like a TV or laptop) or a white goods item (like a washing machine) will inevitably be offered an extended warranty. The technical term for this is an 'Extended Service Contract' or ESC. This will take the manufacturer's warranty of perhaps one year and extend it to say three years from the date of purchase. The consumer pays some additional cost for the peace of mind of knowing that they will not have to face an expensive repair. The sums of money are not small; for example an ESC on a laptop costing $600 could cost $100. This is an enormous business worth billions of dollars a year and it can be very profitable for retailers who charge a generous margin on top of the cost that they pay to the warranty provider. There are reports that margins can be 50% or more, and that electronics retailers can earn half of their total profits from extended warranty sales.

The consumer is facing a decision where it is hard to estimate the probabilities involved and also hard to estimate the costs that may be incurred. There is some chance of a problem that can be fixed quite cheaply, but it is also possible that the item will fail completely. But, if there is so much money being made by the suppliers of the warranty, then it suggests that the ESC is a bad idea on an expected cost basis. Nevertheless, Expected Utility Theory could explain this as a transaction involving a risk averse consumer paying for the insurance provided by a more risk neutral service provider.

Many of the experts, and every advice column, say that buying an ESC is a bad choice on the part of the consumer. In 2006 there was a full page advert placed in many newspapers by *Consumer Reports*, a respected publication, saying simply "Dear shopper, Despite what the salesperson says, you don't need an extended warranty. Yours truly, Consumer Reports". The argument is made that, even assuming a relatively high level of risk aversion on the part of the consumer, it is still hard to justify the high cost of the ESC for many products.

On this basis one might expect that customers who had purchased extended warranties would be unhappy, but not a bit of it. A survey reported in Warranty Week (an online newsletter for people working in this area) in March 2012 asked those who had bought extended warranties in the past whether they would do so again: 49% said "Yes", 48 % said "Perhaps depending on product and pricing" and only 3% said "No". Consumers continue to purchase extended warranties and seem happy to do so.

Business Risk Management Edward Anderson

6.1 Why Decision Theory fails

In the previous chapter we discussed Expected Utility Theory and how it can be used to make decisions in a risky or uncertain environment. We showed how working with utilities on outcomes and simply making choices that maximize the expected utility value is a strong normative decision theory. It describes the way that individuals should make decisions. In fact the case for EUT seems unassailable. It can be derived from axioms which seem entirely reasonable. It has all the right properties - such as always preferring a choice that stochastically dominates another. It also enables an elegant understanding of risk averse or risk seeking behavior depending on the concavity or convexity of the utility function. However there is now a great body of evidence to show that EUT does not do well in predicting the *actual* choices that people make. In fact as we will show individuals deviate from EUT in consistent ways.

In understanding what might be wrong with Expected Utility Theory as a predictor of individual choice, we need to question the individual components of the theory.

6.1.1 The meaning of utility

Perhaps the most fundamental idea in the decision theory we have presented is that utility is defined simply by the outcomes for an individual. This allows us to say that A is preferred to B, and imply by this that A is always preferred to B. One problem here is that I may make one choice today and another tomorrow, so that there is inconsistency in an individual's choices. This might simply be because I am more or less indifferent between the choices available. Which cereal do I choose for breakfast? Where in the train do I choose to sit? A lack of consistency here is no real problem for the theory; but what if I am not indifferent between two choices, and have a clear preference for one over the other? Does it therefore follow that I will make the same choice on another occasion? Perhaps not, because every decision is made in a particular context. For example a choice to buy a product or not will depend on the way that the product is described (psychologists talk of this as 'framing') and this can make a difference - perhaps the product has a particularly persuasive salesman. Moreover the decision I make on this occasion cannot be divorced from the other things that are going on in my life: a fright that I receive while driving my car today will make me more inclined to purchase life insurance tomorrow than I was yesterday.

A separate problem that we should address is the connection between the utility of an outcome and the satisfaction that I achieve from it. In crude terms I can be expected to choose the action that will bring me the greatest enjoyment, but we have only to make this statement to recognize that it is a dramatic over-simplification.

- I may be altruistic and choose an action that benefits a friend, family member or society at large. So at the least we can say that satisfaction or contentment is complex and does not just involve our own pleasures.
- I may postpone doing something that I know I will enjoy. From the point of view of the decision I face right now the cup of coffee will be well worth its price, but I know that I can enjoy the cup of coffee later in the morning and so decide to wait. Or perhaps I am saving for my retirement and I decide not to purchase the expensive holiday that I want because it will eat into those savings. Part of what is going on in these examples is that there is enjoyment to be had simply in the anticipation of pleasure.

- The action I choose may be determined by who I perceive myself to be, and how I wish to be seen by others. For example I want to be seen as responsible and restrained and so, tempting though it may be, I do not buy the convertible sports car that I would really enjoy.

These observations all demonstrate that we cannot simply define utility on the basis of immediate personal enjoyment.

In addition to these complexities we can also observe that the utility of an outcome is often determined in part by a comparison with other outcomes. There are three important aspects to this.

- When assessing an outcome we often compare it with *the outcomes achieved by others*. For example an employee will judge her salary not only by its size but also by how it compares with her colleagues and peers. If others do better than me then I will feel worse. As Gore Vidal put it "It is not enough to succeed. Others must fail." This idea is important when we think about the concept of a Pareto improving outcome, in which all parties do at least as well as they did before, and some people do strictly better. It may seem as though that must be a good thing, but life is more complicated. A boss who arbitrarily giving a bonus of $500 to half of his employees and nothing to the others may produce a Pareto improving outcome for the workforce, but there will certainly be some unhappy people (and imagine what would happen if male employees got the bonus and the women did not!).
- When assessing an outcome we often compare it with *how things were before*. This means that the utility of an outcome can depend on the route by which it arrived. Two people go into a casino and spend the night gambling: one begins by losing some money and then over the course of the evening wins it back to emerge $500 ahead, while the other has some early successes and at one point has made a profit of $5000 before continuing to gamble less successfully losing 90% of his winnings to finish with a profit of $500. Then the person who has slipped from a potential profit of $5000 to a profit of $500 is likely to feel much less happy about the outcome than the other.
- When assessing an outcome we often compare it with *our expectations*. In a similar way to that in which people compare their current state with their previous state, the expectations of outcomes also play a role. An employee who last year had a bonus of $5000 and expects a similar bonus this year, will feel much less positive on receiving this bonus, than an employee who expects a bonus of $2000 and instead receives $5000.

Our discussion so far has demonstrated that utility cannot be understood without taking account of many different factors: our happiness depends both on context and comparison, and the choices we make are not just about our own immediate pleasure. But we can still rescue the idea of a definite utility for each possible outcome, we just need to understand this utility more broadly. For example we can say that we receive utility from seeing friends do well; from believing that we are thought well of by others; from thinking that we have done better than others; from experiencing an improvement in our circumstances; and from the pleasure of a surprise in relation to our expectations.

The von Neumann Morgenstern Theorem deduces utility from decisions not the other way around. And so nothing about the difficulties of constructing utilities from looking at the properties of outcomes necessarily undercuts this theory.

6.1.2 Bounded rationality

If Expected Utility Theory holds then it suggests that good decision makers (who make consistent and thoughtful choices) should be investing in estimating both the utilities of different outcomes and the probabilities that they may occur. When this has been done then a rational individual will carry out a computation of the expected value of different choices in order to decide between them. Though we can construct artificial examples in which these calculations are relatively easy to carry out, to do this in practice is far more difficult. In most cases there are enough potential outcomes to make even listing them a challenge, let alone evaluating utilities for them all. And what methods can be used to estimate probabilities for these outcomes? Finally there is a non-trivial computation of expectations to be carried out. This also does not chime well with what we know in practice, since we make most decisions without recourse to a spreadsheet or calculator.

Herbert Simon called into question whether we can expect so much from decision makers and called this “bounded rationality”. There is not only the question of computational capability, but also whether the time and expense involved is likely to be compensated for by a better final decision. But what is the alternative for making decisions? If a full scale computation of expected utilities does not take place then we need to investigate the mental processes that occur instead.

There has been a large amount of academic research into the processes that people use to make decisions. There are a variety of different heuristics and biases that come into play, for the most part subconsciously. Kahneman (2003) describes two systems of decision making or cognition, that he describes as ‘reasoning’ and ‘intuition’. Reasoning is carried out deliberately and requires effort, whereas intuition takes place spontaneously and requires no effort. Most of the time we operate intuitively, with our reasoning capacity acting as a monitor or restraint on our intuitive actions and choices. The reasoning component in decisions requires effort, of which there is only a limited overall capacity, and for that reason it can be interrupted if another reasoning task arises. Kahneman points out that this is what happens when a driver halts a conversation in order to carry out a difficult manoeuvre: the driving decisions temporarily move from being intuitive to requiring reasoning.

But even when decision making is done within a ‘reasoning mode’ simplifying heuristics and biases will come into play. For example:

- When faced with a complex choice involving many alternatives, decision makers tend to eliminate some choices quickly using relatively small amounts of information, and only when the choice set is reduced to a small size (maybe two only) does the decision maker attempt to compare on the basis of all available information.
- Decision options or outcomes that have greater saliency or greater accessibility will be given greater weight. Accessibility here refers to the ease with which something can be brought to mind and may be determined by recent experience or the description of the outcome or option. Saliency refers to the extent that an item is distinctive or prominent.

These heuristics imply that the framing of a decision will have a significant effect on the choice made.

- Decision makers will usually accept the formulation that is given relatively passively, and are unlikely to make their own framework of evaluation for different options. In particular whatever choice is presented as the default option has a greater likelihood of being selected.

6.1.3 Inconsistent choices under uncertainty

In our discussion so far we have reduced the range of circumstances in which we can expect that Expected Utility Theory will apply. We need to assume that outcomes are simple so that utilities can be evaluated easily. We need to ensure that the decision is taken through reasoning (and applying mental effort) rather than being carried out in an intuitive fashion. Finally we must have a simple arrangement with well defined probabilities in order to avoid the constraints of bounded rationality.

It is remarkable that, even working within these limitations, we can find examples where decision makers make choices that are not consistent with *any* choice of utility function, and hence demonstrate a deviation from Expected Utility Theory. The first such observation was made by Maurice Allais in a 1953 article in Econometrica ("The behavior of rational man in risky situations - A critique of the axioms and postulates of the American School"). Allais demonstrates that the axiom of independence may not hold in practice. Or to put it another way, the axiom may not hold in a descriptive rather than normative theory of decision making.

Next we give three examples from amongst many that could be given In each case people are asked to make a choice between two options (or to be more precise to say which of two options they would prefer if it was offered). This is Decision 1. Then people are asked to make a choice between a different pair of options (Decision 2). The choices are constructed in such a way that certain decisions are inconsistent with any set of utility values. The experiments are repeated with many individuals to demonstrate a consistent pattern in the way that people make decisions. Taken together these examples provide a very convincing case that an alternative to expected utility theory is needed if we want to do a good job of explaining how individuals actually make choices when faced with uncertainty.

Example 6.1

Consider the following two experiments. In Decision 1 participants are asked to choose between the two prospects A1 and B1 described as follows:

A1: gain \$2,500 with prob. 0.33; gain \$2,400 with prob. 0.66; 0 with prob. 0.01.
B1: gain \$2,400 with certainty.

The experiment shows that more than 80% of people choose B1. Under the assumptions of EUT we can convert this into a statement about the utilities of the various sums of money involved. We deduce that, for most people,

$$u(2400) > 0.33u(2500) + 0.66u(2400)$$

since we can assume $u(0) = 0$. This can be simplified to

$$0.34u(2400) > 0.33u(2500).$$

In Decision 2 participants are asked to choose between the two prospects C1 and D1 described as follows

C1: gain $2,500 with probability 0.33; 0 with probability 0.67,
D1: gain $2,400 with probability 0.34; 0 with probability 0.66,

and in this case more than 80% of people choose C1. But, again using $u(0) = 0$, this implies that

$$0.33u(2500) > 0.34u(2400)$$

i.e. the exact reverse inequality to that we just derived.

Try checking what your own choices would be in these two different decisions. For most people having the inconsistency pointed out to them does not alter the choices they would make. In the Allais paradox prospects C1 and D1 are obtained from prospects A1 and B1 simply by eliminating a 0.66 chance of winning $2400 from both prospects. This change produces a greater reduction in desirability when it turns a sure gain to a probable one, rather than when both the original and the reduced prospects are uncertain. There is something about the sure gain of prospect B1 that makes it particularly attractive, and this leads to a violation of the independence axiom.

Example 6.2

The same type of phenomenon appears in the following two experiments. In Decision 1 participants are asked to choose between the two prospects A2 and B2 described as follows

A2: gain of $4000 with probability 0.8; 0 with probability 0.2,
B2: gain of $3000 with certainty.

The majority of people (80%) choose B2. For these people we can deduce that $u(3000) > 0.8u(4000)$. In Decision 2 the two prospects are

C2: gain of $4000 with probability 0.2; 0 with probability 0.8,
D2: gain of $3000 with probability 0.25; 0 with probability 0.75.

Then the majority of people (65%) choose C2. We can deduce that for these people $0.25u(3000) < 0.2u(4000)$ which is the reverse of the inequality derived from the first experiment. This is another example of people preferring certainty.

In this example the choice of B2 in preference to A2 is an example of risk aversion. If individuals were risk neutral then A2 with a higher expected value (of $3200) would be preferred. The preference for certainty here reflects a concave utility function. The problem for EUT is that the two different decisions imply different amounts of concavity (a greater degree of concavity for Decision 1 than for Decision 2).

Example 6.3

The exact opposite of these results are found when losses are involved rather than gains. In this case in Decision 1 participants are asked to choose between the two prospects A3 and B3 described as follows

A3: loss of $4000 with probability 0.8; 0 with probability 0.2,
B3: loss of $3000 with certainty.

The great majority of people (90%) choose A3. For these people we can deduce that $u(-3000) < 0.8u(4000)$. In Decision 2 the two prospects are

C3:	loss of \$4000 with probability 0.2;	0 with probability 0.8,
D3:	loss of \$3000 with probability 0.25;	0 with probability 0.75.

Then the majority of people (58%) choose D3, which leads to the inequality: $0.25u(-3000) < 0.2u(4000)$, that amounts to the exact opposite of the deduction from Decision 1.

Since the great majority of people choose A3 in preference to B3, even though the expected loss under A3 is greater (at \$3200) we can deduce that people are risk seeking over negative gains (i.e. losses), where they will gamble to give themselves a chance of avoiding a loss. This implies a utility function that is convex in this area. The difficulty in this example is that there is a greater degree of convexity in Decision 1 than in Decision 2.

6.1.4 Problems from scaling utility functions

Our final example of the way that Expected Utility Theory can fail is taken from Rabin and Thaler 2001. It demonstrates that problems arise unless we deal with changes in wealth rather than absolute values. Suppose that you are offered a choice to gamble with a 50% chance of winning \$100 and a 50% chance of losing \$90. Most people would reject this bet independently of the size of their bank balance. Following expected utility theory, if W is their current wealth then this implies that they prefer W to equal chances of being at $(W - 90)$ or $(W + 100)$. With a utility function u this gives

$$u(W) > 0.5u(W - 90) + 0.5u(W + 100).$$

Hence (multiplying through by 2 and rearranging)

$$u(W) - u(W - 90) > u(W + 100) - u(W). \tag{6.1}$$

This shows that u is concave over the interval, but we can be more specific. As Figure 6.1 shows, the derivative of u at $W + 100$ is less than the slope of the straight line joining the points on the curve at W and $W + 100$, i.e.

$$u'(W + 100) < (u(W + 100) - u(W))/100.$$

In the same way the derivative of u at $W - 90$ is more than the slope of the straight line joining the points on the curve at $W - 90$ and W, i.e.

$$u'(W - 90) > (u(W) - u(W - 90))\,/90.$$

Putting these observations together with (6.1) shows that

$$u'(W - 90) > (10/9)u'(W + 100).$$

Now since the W in this inequality is arbitrary, we can deduce that

$$\begin{aligned} u'(W) &> (10/9)u'(W + 190) \\ &> (10/9)^2 u'(W + 380) \\ &> (10/9)^n u'(W + 190n). \end{aligned}$$

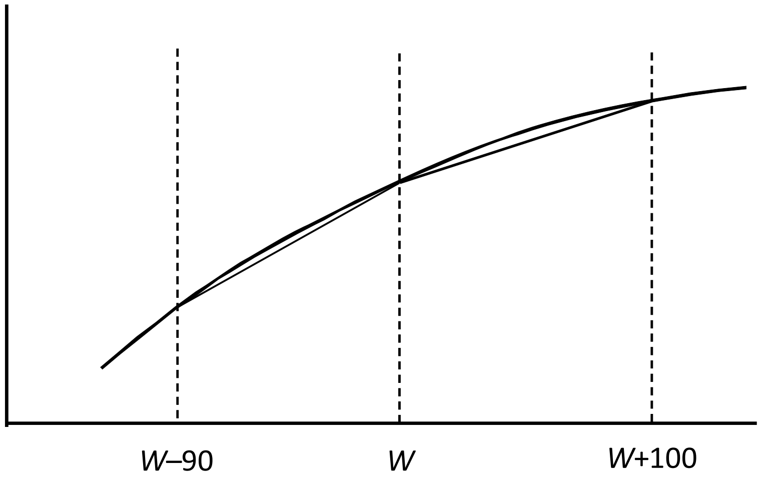


Figure 6.1 Comparing straight line segments to the utility function

What has happened here is that, in effect, we have applied the inequality arising from not gambling at different points along the curve and then stitched the inequalities together to say something about the way that the slope decreases over a much longer interval. When $n = 50$ this gives

$$u'(W) > 194u'(W + 9500).$$

The slope of the utility function simply tells us an extra dollar would be worth to us, and so this inequality says that the value of an extra dollar is almost 200 times less if you are $9500 dollars wealthier. This does not seem believable: it is quite reasonable to suppose that increasing wealth makes someone value additional wealth less, but this cannot happen to the extent that this calculation predicts. From this we can see that uniformly applying the consequences of rejecting a small gamble implies unbelievable consequences when scaled up.

6.2 Prospect Theory

A number of people have worked on different ways of explaining deviations from Expected Utility Theory. We will describe just one of these theories, called prospect theory, developed by Daniel Kahneman and Amos Tversky. This theory is built on many of the ideas that preceded it and even if some would argue for a different approach the main ingredients of these theories are similar. So we will not lose much by concentrating on a version of prospect theory. Moreover Kahneman and Tversky's work (summarized in their papers of 1979 and 1992) is the single most important contribution in this area and they were awarded the Nobel prize for their work in 2002.

Our starting point is to make three observations about the way people make decisions.

Using a reference point.

One weakness in Expected Utility Theory is that for consistency it must apply to the total wealth of an individual. And yet there seems little evidence that people take much account of their bank balance or housing equity when considering small scale financial decisions. Obviously this principle will depend to some extent on individual circumstances: when facing bankruptcy then indeed the absolute wealth may be the focus of attention. But in the normal course of events (say in deciding whether or not to take up an extended warranty offer that involves paying out money now for additional security in the future) our total wealth is not a big factor.

Instead of thinking about the total value of all their assets and using that in a calculation of utilities, people tend to compare possible outcomes against some benchmark in their mind. We have already said that our feeling about outcomes may depend on how well we do in comparison with others around us, or in comparison with an expectation we have formed. But when decisions are made that may involve gains or losses, then the current position becomes the normal starting point. Decision makers focus on changes rather than on absolute values.

The way in which a reference point is constructed will depend on the exact circumstances of the decision. There is an opportunity for framing to lead a decision maker towards a particular reference point. We will concentrate on simple prospects without looking at any of the contextual factors that can come into play in practice. In these cases it seems that a prospect in which every outcome involves a gain will be evaluated by taking the lowest gain as certain and using this lowest gain as a reference point. In much the same way in evaluating a prospect where every outcome produces a loss, most people take the smallest loss as certain and evaluate the resulting prospect from that point.

Avoiding losses if possible

The existence of a reference point opens up the possibility of different behavior on one side of the reference point than the other, and this is exactly what we find. People dislike losses in an absolute way rather than in a way that corresponds to simply looking at a reduction in utility as wealth decreases. For example we can look at gambles in which there are equal chances of losing X or gaining Y. For this to be attractive most people want Y to be about twice as big as X. Thinking in utility terms would suggest that as X and Y get smaller decision makers should be more inclined to accept the gamble provided that Y (the gain) is larger than X (the loss). In practice however this doesn't seem to happen until X and Y are made so small as to be immaterial.

A good way to describe this is to say that individuals have a value function that is kinked around the reference point, as is indicated in Figure 6.2. This makes sense of a lot of observations that we can make about people's behavior. For example it has often been observed that decision makers seem to prefer the status quo. One explanation is that in considering an alternative this is seen as providing with some probability a loss of wealth and with some probability a gain of wealth when compared with the status quo. Then the loss aversion effect means that gains or the probability of gains need to be disproportionately high in order to make a change worthwhile.

Giving too much weight to small probabilities

Most people are not good at understanding intuitively the properties of very small probabilities. For example we are likely to have a good idea of what it means to leave home late and risk missing the train. We can readily make decisions like whether or not to go back and pick up a forgotten umbrella, given the chance that this will make us miss

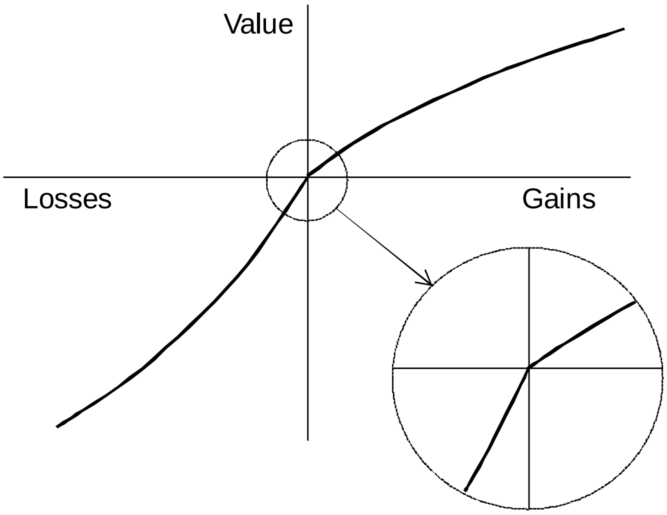


Figure 6.2 The shape of the value function in prospect theory

the train. In essence this is a choice with different uncertainties (the chance of rain; the chance of missing the train) and different possible outcomes (getting wet; being late), but the probabilities involved are not too tiny (say greater than 5%). But when we deal with much smaller probabilities then we are less likely to make effective intuitive judgements. For example if we suddenly recall that there is a back window unlocked, do we then go back and lock it at the risk of missing the train? Here the potential loss if the house is burgled is much greater than just getting wet, but the probability is very small (burglary is attempted; the burglar finds the unlocked window; and the burglar would have gone away if all the windows were secure). Of course we will make a decision quickly (since there is that train to catch) but the quality of this decision may not be as good.

Experiments show that a small probability of a large gain is given a higher value than we might expect. Suppose that we compare two lotteries: A has one chance in 100,000 of delivering a large prize, and B has one chance in 10,000 of delivering the same large prize. It is easy to see that a ticket for A is worth only one tenth as much as a ticket for B. But since people have a hard time conceptualizing what a chance of one in 100,000 really means, the tickets for A will actually be valued at much more than that.

The same thing happens in reverse with probabilities that are nearly 1, so that the event is nearly certain. Then the chance of the event *not* happening seems to be inflated in people's minds. The result is that a near certainty of a gain of a large gain seems less attractive than we might expect. For example suppose that I am offered a prize of $6000 unless a 6 is thrown twice running with a fair dice (i.e. the $6000 prize is received with a probability $35/36$ and nothing is received with a probability $1/36$). But there is an alternative which is to take a prize of $5000 for sure. Most people opt for the certainty of the $5000 prize. This is because we tend to over-weight the small probability of ending with nothing and feeling foolish, and this is the same sort of behavior that is described in Example 6.1 above.

6.2.1 Decision weights and subjective values

Kahneman and Tversky set out their first version of Prospect Theory in 1979 and then updated it later. We will start by discussing "Version 1" of prospect theory before going on to the full-blown cumulative version in the next section. Essentially prospect theory is constructed out of the three observations above.

Since people seem to consistently overweight small probabilities it makes sense to define a function π which converts probabilities into decision weights. Thus $\pi(p)$ is defined for $0 \leq p \leq 1$ and we will assume that $\pi(0) = 0$ and $\pi(1) = 1$.

We also need to define a subjective value $v(x)$ for each change in outcome (gain or loss). We have $v(0) = 0$. This is a little like a utility function, but it applies to changes in wealth, rather than to absolute values of wealth.

Then in the simplest case consider a prospect where a gain of x occurs with probability p and a loss of $-y$ occurs with probability q. Here there is no change with probability $1 - p - q$. We form a value function for the prospect (in comparison with the reference point of no change) by replacing probability with weights according to the decision weight function π and using the values of individual gains and losses in the same way that utilities of outcomes are used in EUT. So we get the prospect value as

$$V(x,p;y,q) = \pi(p)v(x) + \pi(q)v(y) \tag{6.2}$$

Note that since $v(0) = 0$ this does not appear in the expression for V.

However we also need to capture the observation that if all outcomes are gains, then the lowest gain is treated as certain (and similarly for all losses). Hence if $x > y > 0$ and $p + q = 1$ then this is perceived as being equivalent to a gain y with subjective value $v(y)$ but with a probability p (with decision weight $\pi(p)$) that the gain of y will be replaced by a gain of x. This gives

$$V(x,p;y,q) = v(y) + \pi(p)[v(x) - v(y)]. \tag{6.3}$$

This differs from (6.2) unless $\pi(1 - p) = 1 - \pi(p)$ in which case the two expressions are the same.

The situation for losses is similar. When $x < y < 0$,

$$V(x,p;y,q) = v(y) + \pi(p)[v(x) - v(y)].$$

If the decision weights were equal to probabilities so that $\pi(p) = p$ then these expressions just revert to EUT. Prospect theory implies that decisions will generally break the rules for a rational decision maker (with inconsistencies of some sort). Sometimes if a decision maker has these anomalies pointed out then he will adjust his preferences to avoid being inconsistent. But if the decision maker does not discover that his preferences violate appropriate decision rules then the anomalies implied by prospect theory will occur. Indeed when there is just one decision to be made (rather than a whole series) and the results are personal to the decision maker so that real gains and losses are involved, then a decision maker is unlikely to be concerned about breaking (say) the independence axiom of EUT. In this case people are likely to follow the predictions of prospect theory even when they take time to consider their decisions carefully.

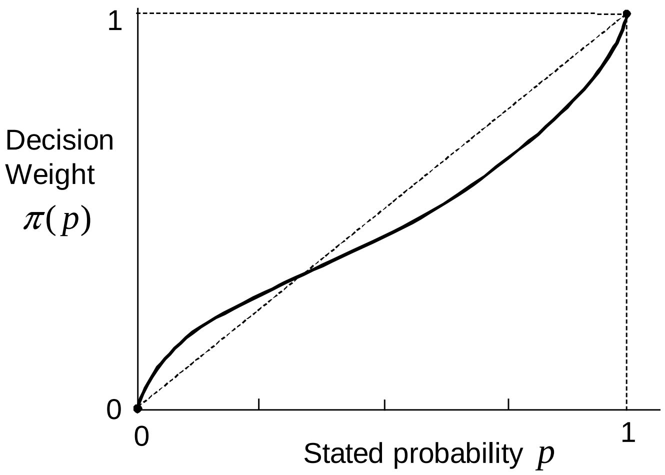


Figure 6.3 A typical decision weight function

To understand the implications of this theory in more detail, we need to look at the functions v and π. We have already seen in Figure 6.2 roughly how the subjective value function v behaves for many people.

We can also ask how the π function behaves. Again this will depend on the individual but the general shape is shown in Figure 6.3.

Many experiments have been carried out to understand what this function looks like, and they suggest a lot of uncertainty or variation in π at both zero and one (and possible discontinuities). People are simply not very good at evaluating probabilities near zero; either they get ignored, or they are given too much weight.

The shape of the function reflects the characteristics that we discussed above. First relatively small probabilities are overweighted. If we know that something happens 10% of the time we behave much as we 'should' behave under a utility model if this event was much more likely (say 15% or 20%). Secondly very high probabilities are under-weighted (which ties in with the preference for certainty that we have already observed). We treat an event which occurs 99% of the time, and so is nearly certain to occur, almost as if it happens only 95% of the time.

Now we return to Example 6.2 and ask whether it is consistent with this theory. In this example the majority of people chose \$3000 with certainty over the option of \$4000 with probability 0.8. Thus $(\$3000, 1.0) \succcurlyeq (\$4000, 0.8)$, from which we deduce that

$$\pi(0.8)v(4000) < v(3000).$$

But the example also has $(\$4000, 0.2) \succcurlyeq (\$3000, 0.25)$, from which we deduce that

$$\pi(0.2)v(4000) > \pi(0.25)v(3000).$$

Suppose we write $k = v(3000)/v(4000)$. Then these two inequalities can be rewritten

$$\pi(0.8) < k < \pi(0.2)/\pi(0.25). \tag{6.4}$$

If we look at Figure 6.3 we can see that for this decision weight function π is below the diagonal at 0.8 so

$$\pi(0.8) < 0.8.$$

Also π is concave in the region 0 to 0.25. This implies that $\pi(0.2)$ lies above the straight line joining $\pi(0)$ to $\pi(0.25)$, i.e.

$$\pi(0.2) > 0.8\pi(0.25),$$

from which (6.4) will follow if we set $k = 0.8$.

We expect the value function to be concave for gains and this implies that

$$v(3000) > 0.75v(4000)$$

and hence a value of $k = 0.8$ is just what we might expect. Hence a detailed analysis of this example is completely in line with the shape of the decision weight function and the value function.

6.2.2 Lotteries and insurance

It is instructive to look at the sales of lottery tickets from the perspective of prospect theory. The majority of people are prepared to buy tickets in a fair lottery. So for example they prefer the prospect $(\$5000, 0.001)$, which is the lottery, to $(\$5, 1)$ which is the price of the ticket (and here has the same expected value). This implies that $\pi(0.001)v(5000) > v(5)$. But if the value function is concave for gains, which is what we would expect, then $v(5)/v(5000) > 0.001$. Combining these two inequalities we get $\pi(0.001) > 0.001$. More generally a preparedness to engage in lotteries supports the idea that $\pi(p) > p$ for small p.

A full explanation of why individuals are often prepared to gamble in lotteries involves a second important factor and that is the pleasure in anticipating the possibility of a win even when this does not, in the end, materialize. Th elottery ticket is as much about purchasing a day dream as it about purcahsing a small probability of the big prize.

The same underlying idea occurs with insurance. A homeowner insuring his property is essentially preferring the prospect of a certain small cost to a much larger cost which occurs with a small probability. So $(-\$5, 1) > (-\$5000, 0.001)$. Since we expect the subjective value function to be convex for losses then we can use the same idea as in our discussion of lottery tickets to show that this also implies $\pi(p) > p$ for small p.

Again there is a second effect that relates to the way that an individual feels over the lifetime of an insurance policy. This can be regarded as ‘peace of mind’: the knowledge that with an insurance policy is in place we don’t need to worry about the possibility of a calamity.

6.2.3 A power law for values

One of the observations that emerges from experiments is that, to a first approximation, a prospect that is preferred to another is still preferred if all the amounts of money involved are multiplied by a constant. This is a property that occurs when the subjective value function follows a ‘power law’ i.e. we have $v(x) = \gamma x^\alpha$ for some γ and α. Since we expect the value function to be concave we need a negative value for α, but that does not effect the argument below.

In this case if we are indifferent between the prospect (x, p) and the prospect (y, q) then $\pi(p)v(x) = \pi(q)v(y)$ and hence

$$\pi(p)\gamma x^\alpha = \pi(q)\gamma y^\alpha.$$

Thus we have

$$\pi(p)v(kx) = \pi(p)\gamma k^\alpha x^\alpha = \pi(q)\gamma k^\alpha y^\alpha = \pi(q)v(ky),$$

showing that we are still indifferent between these prospects when we have multiplied both the outcomes by a factor of k.

Moreover we can show this in the other direction. Suppose that multiplying values by k makes no difference to two prospects that are equivalent. Hence $\pi(p)v(x) = \pi(q)v(y)$ implies $\pi(p)v(kx) = \pi(q)v(ky)$. Thus

$$v(kx)/v(x) = v(ky)/v(y). \tag{6.5}$$

The value y here is arbitrary: for different values of y we simply choose a different values of q so that we remain indifferent between (y, q) and (x, p). So we can write h_k for the value of the ratio (6.5) and $v(x)$ satisfies an equation of the form

$$v(kx) = h_k v(x) \text{ for all } x > 0.$$

Now we define a function g by

$$g(w) = \log(v(e^w))$$

(This is a bit like plotting v on log-log graph paper). Notice that for any k

$$g(w + \log k) = \log(v(e^w k)) = \log(h_k v(e^w)) = \log(h_k) + g(w),$$

so the function g must have what we can call a ‘constant increase’ property: in other words $g(x + A) - g(x)$ depends only on A (and not on x). This implies that g is linear, since if g does not have a constant slope then we can find a point x and distance δ where $g(x) - g(x - \delta) \neq g(x + \delta) - g(x)$ (we just take x somewhere with non-zero second derivative and choose δ small enough). But this contradicts the constant increase property of g. So g must be linear, and we can write it as $g(w) = a + bw$ for some choice of a and b.

Hence $\log(v(e^w)) = a + bw$ and so

$$v(e^w) = e^a (e^w)^b,$$

which is the form

$$v(z) = \gamma z^\alpha,$$

with $\gamma = e^a$ and $\alpha = b$. Thus we have established that only power law functions can be used as value functions if we want to preserve ordering of prospects when values are multiplied by a constant.

6.3 Cumulative prospect theory

The prospect theory that we have developed works well when prospects have just two outcomes, but we can get into difficulties if there are more than two outcomes. It turns out that with 'Version 1' of prospect theory it is possible for very similar prospects to end up with very different values. For example compare the prospects:

$$\begin{aligned} &AA : (\$100, 0.05; \$101, 0.05; \$102, 0.05; \$103, 0.05) \\ &BB : (\$103, 0.2). \end{aligned}$$

We expect BB to be preferred since it stochastically dominates AA. Whatever we say about risk aversion or overweighting of small probabilities, it is hard to imagine a decision maker selecting AA in preference to BB. Now if we just weight values with decision weights we get

$$\begin{aligned} V(AA) &= \pi(0.05)\,(v(100) + v(101) + v(102) + v(103)) \\ &\simeq 4\pi(0.05)v(103). \end{aligned}$$

Notice that there is a 0.8 probability of getting nothing and so we are not in the position of a sure gain to apply (6.3).

However our previous discussion of prospect theory shows that we would expect $\pi(0.05)$ to be much greater than $\pi(0.2)/4$. Thus using prospect theory version 1, implies that the prospect AA has a substantially larger value than BB.

To fix this problem we need to introduce cumulative prospect theory. This makes the expressions much harder to write down, but in essence this 'Version 2' of prospect theory is only slightly more complicated.

We start by defining two different decision weight functions; one will apply to positive outcomes and one to negative. We call these decision weight functions $w^+(p)$ and $w^-(p)$ and they are defined on probabilities p. These weight function are similar to π and, as with π we will assume that $w^+(0) = w^-(0) = 0$ and $w^+(1) = w^-(1) = 1$.

We will deal with positive outcomes first. The approach is to order the potential outcomes and apply to each a weight given by the increment in the w^+ function, if $0 < y < x$ then

$$V(x, p; y, q) = [w^+(p + q) - w^+(p)]v(y) + w^+(p)v(x).$$

Notice that with just two outcomes (i.e. no option of zero change) this reverts to the previous version. Thus if $0 < x_1 < x_2$ and $p_1 + p_2 = 1$, then

$$\begin{aligned} V &= v(x_1)[w^+(p_1 + p_2) - w^+(p_2)] + v(x_2)w^+(p_2) \\ &= v(x_1) + w^+(p_2)[v(x_2) - v(x_1)], \end{aligned}$$

using the fact that $w^+(p_1 + p_2) = w^+(1) = 1$.

More generally, for prospects where outcomes are 0 or x_i with $0 < x_1 < ... < x_n$ and there is probability p_i of outcome x_i, then

$$V = \sum \pi_i^+ v(x_i),$$

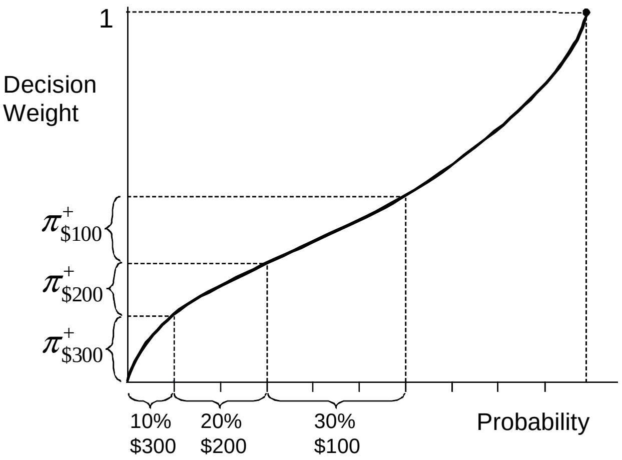


Figure 6.4 Calculating decision weights using the incremental method of cumulative prospect theory.

with

$$\pi_i^+ = w^+(p_i + \ldots + p_n) - w^+(p_{i+1} + \ldots + p_n)$$

and

$$\pi_n^+ = w^+(p_n).$$

Figure 6.4 illustrates the way that this incremental calculation is carried out for a prospect in which there is 40% chance of a gain of nothing, a 30% chance of a gain of $100, a 20% chance of a gain of $200 and a 10% chance of a gain of $300. Thus the prospect is $(\$100, 0.3; \$200, 0.2; \$300, 0.1)$. In order to calculate probability weights for each of the outcomes we divide the probability axis into regions of the appropriate length starting with the highest gain of $\$300$. Then the π^+ values are read off from the increments in the w^+ function values.

Notice that weight assigned to the first outcome of $300 is proportionally higher in relation to the probabilities than the weight allocated to a middle outcome of $100 (remember that the worst outcome is $0). In fact the worst outcome, especially if it has low probability, is also given a higher weighting. These facts follow from the higher slope of the decision weight curve at the two ends of the interval. A key characteristic of prospect theory is that it gives higher weights to relatively unlikely extreme outcomes (either large gains or near zero gains) and this is also true for losses.

The definitions for negative outcomes are similar. Suppose a prospect has outcomes of 0 or x_i with $0 > x_1 > ... > x_n$ and there is probability p_i of outcome x_i. Then

$$V = \sum \pi_i^- v(x_i),$$

with $\pi_i^- = w^-(p_i + \ldots + p_n) - w^-(p_{i+1} + \ldots + p_n)$ and $\pi_n^- = w^-(p_n)$.

If f is a prospect with both positive and negative outcomes then we let f^+ be f with all negative elements set to zero, and we let f^- be f with all positive elements set to zero. Then

$$V(f) = V(f^+) + V(f^-)$$

(remember that $v(0) = 0$ so the extra zero value outcomes in f^+ and f^- do not change the value of V).

This is most easily understood by looking at an example: If the outcomes of a prospect are $-\$5$, $-\$3$, $-\$1$, $\$2$, $\$4$, $\$6$ each with probability $1/6$, then

$$f^{\,+} = (\$0, 1/2; \$2, 1/6; \$4, 1/6; \$6, 1/6)$$

$$f^{\,-} = (-\$5, 1/6; -\$3, 1/6, -\$1, 1/6; \$0, 1/2).$$

So

$$\begin{aligned} V(f) &= v(2)[w^+(1/2) - w^+(1/3)] + v(4)[w^+(1/3) - w^+(1/6)] + v(6)w^+(1/6) \\ &\quad + v(-1)[w^-(1/2) - w^-(1/3)] + v(-3)[w^-(1/3) - w^-(1/6)] + v(-5)w^-(1/6). \end{aligned}$$

Now we return to the example we discussed above. Prospect AA will have value

$$\begin{aligned} V(AA) &= v(100)(w^+(0.2) - w^+(0.15)) + v(101)(w^+(0.15) - w^+(0.1)) \\ &\quad + v(102)(w^+(0.1) - w^+(0.05)) + v(103)w^+(0.05)), \end{aligned}$$

and it is easy to see that provided $v(100)$ is close to $v(101), v\,(102)$ and $v(103)$ then $V(AA)$ is approximately equal to $v(100)w^+(0.2) = V(BB)$.

6.3.1 The link to stochastic dominance

We commented already that in the example involving prospects AA and BB, the second prospect stochastically dominates the first and thus we would like to have BB to be preferred no matter what value function is used. The use of cumulative prospect theory will guarantee that this happens. To demonstrate this we will just consider prospects with all positive values. $0 < x_1 < ... < x_n$ and assume that for prospect P there is probability p_i of outcome x_i, and for prospect Q there is probability q_i of outcome x_i.

Remember that P stochastically dominates Q if

$$\sum_{i=m}^{n} p_i \geq \sum_{i=m}^{n} q_i, \text{ for } m = 2, 3, ...n \tag{6.6}$$

and there is strict inequality for at least one m. We want to calculate the π_i^+ values for P and the corresponding values for Q, that we will write as ρ_i^+. Thus

$$\pi_i^+ = w^+(p_i + \ldots + p_n) - w^+(p_{i+1} + \ldots + p_n),\ \pi_n^+ = w^+(p_n),$$

$$\rho_i^+ = w^+(q_i + \ldots + q_n) - w^+(q_{i+1} + \ldots + q_n),\ \rho_n^+ = w^+(q_n).$$

We want to show that $V(P) > V(Q)$. Now

$$\begin{aligned} V(P) &= \sum \pi_i^+ v(x_i) \\ &= \sum_{i=1}^{n-1} \left(w^+(p_i + \ldots + p_n) - w^+(p_{i+1} + \ldots + p_n)\right) v(x_i) + w^+(p_n)v(x_n) \\ &= v(x_1) + \sum_{i=2}^{n} w^+(p_i + \ldots + p_n)(v(x_i) - v(x_{i-1})) \end{aligned}$$

where we have used the fact that $w^+(p_1 + \ldots + p_n) = w^+(1) = 1$and gathered together the terms with the same w^+ value. In the same way

$$V(Q) = v(x_1) + \sum_{i=2}^{n} w^+(q_i + \ldots + q_n)(v(x_i) - v(x_{i-1})).$$

Because of our ordering for the x_i we know that $v(x_i) - v(x_{i-1}) > 0$, for $i = 2, ..., n$. Thus we can deduce from (6.6) that each term in this expansion for $V(P)$ is greater than the corresponding term in $V(Q)$ with strict inequality for at least one term. And hence $V(P) > V(Q)$ as we wanted.

6.3.2 Applying prospect theory

Tversky and Kahneman (1992) suggest some functional forms for the functions v and w^+, w^- and estimate the parameters for these functions for a group of experimental subjects (students). They propose that the value functions for both gains and losses follow a power law and hence

$$\begin{aligned} v(x) &= x^\alpha \text{ if } x \geq 0 \\ &= -\lambda(-x)^\beta \text{ if } x < 0. \end{aligned}$$

Note that we can normalize so there is no need of a constant multiplier for the value function for positive x. Moreover the properties of the power law mean that we don't need to specify the units of money involved here.

The functional forms that Tversky and Kahneman propose for the decision weight functions are:

$$\begin{aligned} w^+(p) &= \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{1/\gamma}}, \\ w^-(p) &= \frac{p^\delta}{(p^\delta + (1-p)^\delta)^{1/\delta}}. \end{aligned}$$

Tversky and Kahneman also give estimates for the various parameters:

$$\begin{aligned} &\alpha = \beta = 0.88; \\ &\lambda = 2.25; \\ &\gamma = 0.61; \;\; \delta = 0.69. \end{aligned}$$

We will call these the TK parameter values. They are median values obtained when estimates are made separately for each individual experimental subject. The values implied for w^+ and w^- are given in Table 6.1. Notice that even though this table gives three decimal places, the figures are estimated on the basis of limited set of decisions made in a laboratory setting and should be taken as only a rough guide to the behavior of any given individual.

Table 6.1. Median decision weight values found by Tversky and Kahneman.

p	$w^+(p)$	$w^-(p)$	p	$w^+(p)$	$w^-(p)$
0.05	0.132	0.111	0.55	0.447	0.486
0.10	0.186	0.170	0.60	0.474	0.518
0.15	0.227	0.217	0.65	0.503	0.552
0.20	0.261	0.257	0.70	0.534	0.588
0.25	0.291	0.294	0.75	0.568	0.626
0.30	0.318	0.328	0.80	0.607	0.669
0.35	0.345	0.360	0.85	0.654	0.717
0.40	0.370	0.392	0.90	0.712	0.775
0.45	0.395	0.423	0.95	0.793	0.850
0.50	0.421	0.454	1	1	1

Worked Example 6.1

Use cumulative prospect theory with the TK parameter values to determine which of the following three prospects is preferable:

$$\begin{aligned}
&A : (\$100, 0.5; \$200, 0.4; \$300, 0.1),\\
&B : (\$-100, 0.1; \$300, 0.5; \$800, 0.1),\\
&C : (\$150, 1).
\end{aligned}$$

Solution

Prospect A has all gains and we get

$$\begin{aligned}
V(A) &= \big(w^+(1) - w^+(0.5)\big)\,v(100) + \big(w^+(0.5) - w^+(0.1)\big)\,v(200) + w^+(0.1)v(300)\\
&= (1 - 0.421) \times 100^{0.88} + (0.421 - 0.186) \times 200^{0.88} + 0.186 \times 300^{0.88}\\
&= 86.35.
\end{aligned}$$

For prospect B we have one loss and two gains and we get

$$\begin{aligned}
V(B) &= w^-(0.1)v(-100) + \big(w^+(0.6) - w^+(0.1)\big)\,v(300) + w^+(0.1)v(800)\\
&= 0.170 \times (-2.25) \times 100^{0.88} + (0.474 - 0.186) \times 300^{0.88} + 0.186 \times 800^{0.88}\\
&= 88.28.
\end{aligned}$$

For prospect C there is no uncertainty and we simply have

$$V(C) = v(150) = 150^{0.88} = 82.22.$$

Overall we see that the highest value is achieved by prospect B and using these parameters would lead to B being chosen.

6.3.3 Why prospect theory does not always predict well

Earlier we mentioned the fact that there are many competing theories in this area. And though the general predictions made by cumulative prospect theory are correct, some aspects of the theory are the subject of considerable debate. The primary problem with prospect theory as a description of the way that decisions are taken, is that it can be quite hard to calculate the value of a prospect, and it seems hard to imagine that this is a good match with the way that individuals actually make decisions (even in cases where decisions are thought about carefully with the decision maker operating within a 'reasoning' rather than intuitive framework). An example is the observation we made earlier that in a choice between prospects outcomes with higher saliency will usually be given greater weight. To some extent this is reflected within prospect theory by the use of current wealth as a reference point - extreme changes will be more salient. But it is likely that other aspects of the choice will also have an impact on saliency.

In 'Version 1' of prospect theory, as propounded in Kahneman and Tversky's 1979 paper, there was a greater role for heuristics applied by the decision maker in order to simplify the decision to be made (for example eliminating dominated choices). In moving to version 2 these preliminary 'editing' steps were dropped; in essence they were made unnecessary by the use of a rank-based cumulative weighting function. This has the great advantage of making the predicted choice quite definite (once the parameters have been selected) whereas any theory that puts more emphasis on the processes used by a decision maker is likely to lead to cases where the prediction is less clear cut (for example depending on the order in which some initial editing processes are carried out). However the price to be paid is that there are many situations in which the neatness of cumulative prospect theory does not seem to capture the whole picture.

There are a number of other reasons why we should be cautious in using prospect theory to predict individual behavior:

- Different individuals will have different patterns of behavior, involving different degrees of loss aversion, different degrees of risk aversion etc. In other words even accepting the core assumptions of prospect theory still leaves the question of what the parameters should be for an individual. Moreover we should not assume that an individual always operates with a particular decision approach - perhaps our experiences over the past few hours can have an impact on the amount of risk we will choose.
- There will be a random element in the way that choices are made, especially when they are perceived as being quite similar in overall value. We recognize this in our won choice sometimes: "Its six of one and half a dozen of the other". In an experimental setting it is quite common for the same individual to make different choices between exactly the same pair of prospects when these are presented at different times.
- The decision weight functions w do not appear to work well with probability values that are near zero or near 1. Individuals are particularly poor at making choices when faced with probabilities that are of the order of 0.01 or smaller (or 0.99 or greater). This is where inconsistencies are most likely to arise.
- The particular functional forms chosen for w and v are to some extent arbitrary. Different functional forms have been suggested and may result in different predictions.

- Individuals may change their preferences between options over time as a result of repeated exposure to the same choices, or even discussion with colleagues. This issue is complex, though, since if at the outset a decision maker knows that she will be faced with a whole sequence of similar choices or gambles that are perceived as being grouped together, then her response is likely to be different to the choice made when she is faced with a single gamble.
- Individuals may make different decisions depending on how the choice situation is framed. Since people typically look at changes in wealth with respect to a reference point, framing may occur through the suggestion of a reference point that is different to zero. For example a choice may be presented as follows: “Would you rather have $30 for sure or be given a 50% chance of winning $80?” Exactly the same decision problem can be framed be telling someone they have won $30 and then asking “Do you want to enter a competition where there is a 50% chance of losing $30 and a 50% chance of winning $50?” In this second framing of the choice the reference point has become $+\$30$. Loss aversion will ensure that fewer people will accept the gamble when they have the $30 as the reference point than in the first framing.

6.4 Decisions with ambiguity

In this section we will discuss the way that individuals take decisions when there is ambiguity in relation to the probabilities involved. The theory of Savage discussed in Chapter 5 implies that in many cases consistent decision makers will be working as though there were a subjective probability associated with any particular event. But at this point we are more interested in describing the way that decisions are made in practice.

The examples we have dealt with so far have all ducked the question of where the probabilities come from. We have assumed that some oracle (the psychologist conducting the experiments) announces for us the probabilities of particular events, allowing us to write a prospect as a string of values and probabilities. Or perhaps the probabilities are generated by throwing dice or tossing coins. This is what Nassim Taleb calls the ‘ludic fallacy’ - the belief that real decisions are well-represented by the decisions faced when playing simple games of chance. But as we have pointed out in earlier chapters actual decisions usually involve probabilities that we can guess at, but not know for sure. This happens when the probability arises because of our uncertainty about the way that the world will change in the future.

For example, suppose we ask “What is the probability that the price of oil will be above $150 a barrel a year from now?” A business decision may well depend on the likelihood that we assign to this event, so we may be forced to give an answer, either explicitly, or implicitly by the decision that we make.

The prediction of the price of oil is amenable to all sorts of economic analysis, so a decision maker will have some knowledge of this probability, but many business decisions need to be made with very little knowledge of the probabilities involved. For example we might be interested in the probability that a competing company decides to pull out of a particular marketplace, or the chance of a new pharmaceutical compound under development will prove both safe and effective as a treatment for Alzheimer’s disease. In these types of decisions the extent of our ignorance is greater and we are likely to be forced into looking at statistics for similar cases in the past (if we can find them).

A good example of the way that a lack of knowledge about exact probabilities has an impact on our decisions is provided by the Ellsberg paradox. Suppose that you have two urns each containing 100 balls. In the first urn there are exactly 50 black balls and 50 white balls, but you have no knowledge of the number of balls of different colours in the second urn. Now you are offered a prize of $100 if you draw out a white ball from one of the two urns, but you only get one attempt and so you need to choose which urn. What would you do? It turns out that a large proportion of people will choose the first urn. In a sense there is less uncertainty associated with the first urn where we know that there will be a 50% chance of winning; if we choose the second urn then we have no knowledge at all about the probability of winning, which might even be zero if every ball in the second urn is black.

The preference for the first urn remains in true for a different problem in which the prize of $100 is given if you can correctly predict the color of the first ball drawn. In this formulation first the decision maker selects a color and then she chooses one of the two urns, and hence there is no possibility of the composition of the second urn being somehow made unfavorable.

In order to explain the theoretical problem that this creates more clearly, suppose that the you are the decision maker presented with two urns and you are offered a prize of $100 if you draw out a white ball from one of the two urns. You are likely to choose the first urn; perhaps you win and perhaps you do not. Then you put the ball back into the urn and shake it up to remix all the balls. Next you are offered a second chance to win a prize of $100, but this time you get the prize if you draw out a black ball. What would you choose? Most people still have a clear preference for the first urn with the known composition of 50 balls of each color. After all at this point the second urn is untouched, so it is hard to see why a change of color would change the preference decision.

The first decision is between a prospect $(\$100, 0.5)$ and a prospect $(\$100, p_W)$ where p_W is the (subjective) probability of a draw of a white ball from the second urn. The preference for the first urn implies that $p_W < 0.5$. This is obvious and is also implied by the prospect valuation which has $\pi(0.5)v(\$100) > \pi(p_W)v(\$100)$. But if p_W is less than 0.5 then p_B, the subjective probability of a black ball being drawn from the second urn, must be greater than 0.5. Hence when we reach the second choice it is rational to prefer the second urn. This is the crux of the paradox, which can only be resolved if decision makers were indifferent between the two urns.

This is an example of a situation in which uncertainty about the probabilities in the second urn makes us reluctant to choose it. Sometimes this type of uncertainty is called ‘ambiguity’ and the effect we see in the Ellsberg paradox is called ambiguity aversion. This is an extreme example of a kind of second order uncertainty when the probability is itself a random variable. We will return to thinking about these types of problems in our discussion of robust optimization in Chapter 8.

Much more rarely it is possible to observe an ambiguity preference rather than ambiguity aversion. Ellsberg has suggested an example in which there are two urns: the first has 1000 balls numbered 1 to 1000 and the second has 1000 balls each with a number between 1 and 1000, but numbers may occur more than once and we have no information on which numbers have been used. So for example we might have 500 balls marked 17 and the other 500 balls marked with randomly selected numbers between 200 and 800. Now we are asked to write down a number between 1 and 1000 and then draw out a ball from one of the two urns. If the number on the ball matches the number we have written then we win a prize. In this decision scenario people are quite likely to choose the second (ambiguous) urn. The basic structure here is the same as for the Ellsberg paradox but we are dealing with probabilities of 1 in 1000 rather than 1 in 2.

6.5 How Managers Treat Risk

In this final section we will look at the impact of the psychology of risk on management decisions. One of the key observations of prospect theory is that individuals make judgements based on a change in their wealth, rather than looking at utilities associated with different values of their total wealth after the decision and its consequences. From a manager's perspective this is not what shareholders would like; a more rational decision would see maximizing total profit as the real aim, independently of the route taken in getting there.

This is related to what happens when we have multiple opportunities to gamble. For example the gamble involving gaining $100 or losing $90 might not seem attractive, but if we knew it was offered many times over then its attractiveness changes. For example with 4 repetitions there is a $1/16$ chance of losing $360, a $1/4$ chance of losing $170, a $3/8$ chance of gaining $20, a $1/4$ chance of gaining $210 and a $1/16$ chance of gaining $400. This still might not be enough to encourage everyone to accept the package of gambles, but it is certainly closer to being attractive than a gamble played just once.

Or we might take another example for which many people are more or less indifferent to accepting the gamble, which is when the loss is roughly twice as large as the gain; say we have a half chance of losing $100 and a half chance of gaining $200. But with two repetitions of this gamble we end with a $1/4$ chance of gaining $400, a $1/2$ chance of gaining $100 and a $1/4$ chance of losing $200. Faced with this prospect most people give it a positive value. Multiple opportunities to gamble with a positive expected outcome lead to a greater and greater chance of a good outcome.

It seems, however, that as decision makers we are hard-wired to look just at the immediate outcome of the choice facing us, rather than seeing this as one element in a sequence of choices. As a result we pay attention just to the change arising from the current decision more or less independent of the results of previous decisions. We can say that decision makers are *myopic* (a technical term for being 'shortsighted' in a decision making sense).

This leads to relatively high risk aversion for small gambles as well as for large gambles. This is not rational because small gambles are likely to recur. They may not recur in exactly the same form, but over a period of time there are almost certain to be risky opportunities available to a decision maker, that correspond to small gambles. A decision maker who consistently takes these small gambles where they have a definite positive expected value will end up ahead over time.

Much of our discussion so far has been framed around individual decisions on prospects - we might suppose (or hope) that manager's decisions are in some way 'better'. Managers make decisions in contexts that often involve a whole team of people and that are subject to significant scrutiny. However there seems little evidence that managers make better corporate decisions than individuals make personal decisions.

The first observation to make is that a manager's decisions are taken within a personal context. A manager's actions are not simply about achieving the best result for the company. In addition a manager will be asking herself "What will this do for my career?" or "Will this be good for my stock options?"

A second observation is that manager's own ideas about their roles have an impact on their behavior. We may see managers as dealing with uncontrollable risks in a way that accepts these (negative) possible outcomes because they are compensated by significant chances of gain. But this is not the way that managers view themselves and their roles (March and Shapira, 1987). Instead managers are likely to view risk as a challenge to be overcome by the exercise of skill and choice. They may accept the possibility of failure in the abstract, but tend to see themselves not as 'gamblers' but as careful and determined agents, exercising a good measure of control both over people and events. The net result is that managers are often much more risk averse than we would expect. Consciously or not, many managers believe that risks should be hedged or avoided if they are doing their jobs well.

What is the remedy for this 'narrow framing' that looks only at a single decision and ends up being unnecessarily risk averse? Kahneman and Lovallo suggest that this bias can be helped by doing more to encourage managers to see individual decisions as one of a sequence (perhaps by doing evaluations less frequently) and also by encouraging an attitude that "you win a few and you lose a few", because it suggests that the outcomes of a set of separable decisions should be aggregated before evaluation.

Even if an observer may see managers as taking significant risks, managers themselves perceive those risks as small - the difference is explained by the way that managers habitually underestimate the degree of uncertainty that they face. Kahneman and Lovallo describe this in terms of 'bold forecasts'. Why do managers so frequently take an optimistic view and underestimate the potential for negative outcomes?

This is an example of a more general phenomenon which is the near universal tendency to be more confident in our estimates than we should be. Even when we are sophisticated enough to understand that any estimate or forecast is really about a distribution rather than a single number, we still tend towards giving too much weight to our best guess, or to put it another way we use distributions that do not have sufficient probability in the tails.

There may be many factors at work but one important aspect of this bias relates to what Kahneman and Lovallo describe as the 'inside view' (which to some extent mirrors the narrow framing bias we mentioned above). When faced with an uncertain future and the need to forecast, our natural instinct is to consider very carefully all the specifics of the situation, bring to bear our understanding of the potential causal chains and then determine what seems the most likely outcome. The problem with this approach is that there are often simply too many possible ways in which events may unfold for us to comprehend them all. It may well be that the chain of events that leads to a project completion on time is the most likely amongst all possibilities, but if there are many thousands of possible reasons for delay each happening with a small probability then it may well be the case that significant project delay becomes a near certainty.

The remedy for this situation is to step back from considering the specifics of what may happen in detail, but instead to understand the situation in a more statistical sense. In contrast to an 'inside view' we could describe this as an 'outside view'. Are there a set of roughly equivalent circumstances from which a manager can learn more of the likely range of outcomes? Sometimes this is reasonably straightforward - for example in predicting the box office takings for a four person comedy drama playing a short season in Sydney and Melbourne, it is natural to look to information about similar shows in the past. Often however it requires care to find the right group of comparator situations. There is a lot of evidence that adopting an outside view is more likely to lead to good predictions, but it is surprisingly rare in practice. As Kahneman and Lovallo explain:

> "The natural way to think about a problem is to bring to bear all one knows about it, with special attention to its unique features. The intellectual detour into the statistics of related cases is seldom chosen spontaneously. Indeed, the relevance of the outside view is sometimes explicitly denied: physicians and lawyers often argue against the application of statistical reasoning to particular cases. In these instances, the preference for the inside view almost bears a moral character. The inside view is valued as a serious attempt to come to grips with the complexities of the unique case at hand, and the outside view is rejected for relying on crude analogy from superficially similar instances."

The specific issue here is related to well-understood characteristics of optimism. In general people are optimistic when it comes to evaluating their own abilities (so for example a large majority of people regard themselves as above average drivers); they are optimistic about future events and plans; and finally they are optimistic about their ability to control what happens. In general this is a positive characteristic and optimism in this form is associated with mental health (Taylor and Brown 1988). This does not mean, however, that it is a positive characteristic when practiced by managers facing important decisions on the future of their organizations.

Notes

There is an enormous amount that has been written on behavioral decision theory and behavioral economics. The book by Wilkinson gives an accessible introduction to this field, and the paper by Starmer (2000) gives a more detailed discussion of much of the literature in this area. Peter Wakker's book gives a very thorough and technical treatment of all aspects of prospect theory, but this is a difficult read for the non-expert. The material presented in this chapter has drawn heavily on the papers by Kahneman and Tversky (1979), Tversky and Kahneman (1992), and Kahneman and Lovallo (1992). Kahneman's Nobel prize lecture is also an easy introduction to this area (see Kahneman 2003).

For the discussion of ambiguity in decision making I have drawn on the paper by Einhorn and Hogarth (1986).

References

Maurice Allais (1953) The behavior of rational man in risky situations - A critique of the axioms and postulates of the American School, *Econometrica*, Vol 21, pp. 503–546.

Hillel Einhorn and Robin Hogarth, 1986, Decision making under ambiguity, *The Journal of Business*, Vol 59, pp. 225-250

Daniel Kahneman 2003, Maps of bounded rationality: Psychology for behavioral eonomics, *American Economic Review*, Vol. 93, pp. 1449-1475.

Daniel Kahneman and Dan Lovallo, (1993) Timid Choices and Bold Forecasts: A cognitive perspective on risk taking, *Management Science*, Vol. 39, No. 1, pp. 17–31.

Daniel Kahneman and Amos Tversky, (1979) Prospect Theory: An analysis of decision under risk, *Econometrica*, Vol. 47, No. 2, pp. 263–292.

James G. March and Zur Shapira, (1987) Managerial perspectives on risk and risk taking, *Management Science*, Vol. 33, No. 11, pp. 1404–1418.

Matthew Rabin and Richard H. Thaler, (2001) Anomalies: Risk Aversion, *Journal of Economic Perspectives*, Vol. 15, No. 1, pp. 219–232.

Chris Starmer, (2000) Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk, *Journal of Economic Literature*, Vol. 38 pp. 332–382.

Shelley Taylor and Jonathan Brown, (1988) Illusion and Well-Being: A Social Psychological Perspective on Mental Health, *Psychological Bulletin*, Vol 103, pp.193–210.

Amos Tversky and Daniel Kahneman, (1992) Advances in Prospect Theory: Cumulative representation of uncertainty, *Journal of Risk and Uncertainty*, Vol 5, pp. 297–323.

Nick Wilkinson, 2008, An Introduction to Behavioural Economics, Palgrave Macmillan.

Peter Wakker, 2010, Prospect Theory for Risk and Ambiguity, Cambridge Univrsity Press.

Exercises

6.1. (Explanation of Example 6.1)

The text shows how the behavior shown in Example 6.2 is exactly what one would expect from prospect theory. Carry out the same kind of analysis to explain the behavior of Example 6.1 (try to use a general argument rather than evaluating the prospects for particular values of the parameters).

6.2. (Prospect theory when gains turn to losses)

Suppose that $w^+(p) = w^-(p)$ and the value function has the property that

$$v(x) = -\lambda v(-x) \text{ for } x < 0.$$

Show that if a prospect $A = (x_1, p_1; x_2, p_2; ...x_n, p_n)$ is preferred to prospect $B = (y_1, q_1; y_2, q_2; ...y_n, q_n)$, and all the x_i and y_i are positive, then prospect $-B = (-y_1, q_1; -y_2, q_2; ... - y_n, q_n)$ is preferred to $-A = (-x_1, p_1; -x_2, p_2; ... - x_n, p_n)$.

6.3. (Probabilistic insurance)

Kahneman and Tversky carried out an experiment with 95 Stanford students in which the students were presented with the following problem:

> "Suppose you consider the possibility of insuring some property against damage (e.g. fire or theft). After examining the risks and the premium, you find that you have no clear preference between the options of purchasing insurance or leaving the property uninsured. It is then called to your attention that the insurance company offers a new program called probabilistic insurance. In this program you pay half of the regular premium. In the case of damage, there is a 50 per cent chance that you pay the other half of the premium and the insurance company covers all losses; and there is a 50 per cent chance that you get your insurance premium back and suffer all the losses. For example, if the accident falls on an odd day of the month, you pay the other half of the premium and the insurance company covers all losses; but if the accident occurs on an even day of the month, you get your insurance premium back and suffer all the losses.
>
> Remember that the premium is such that you find this insurance is barely worth its cost. Under these circumstances, would you purchase probabilistic insurance?"

In this experiment 80% of the students answered 'No'. Show that this is inconsistent with Expected Utility Theory with a concave utility function because probabilistic insurance gives a strictly higher utility than standard insurance. You can do this by writing the result of probabilistic insurance in the form $(1 - p)u(W - (z/2)) + (p/2)u(W - z) + (p/2)u(W - K)$ and then use the fact that for a concave function u we have

$$u\left(W - \frac{1}{2}z\right) \geq \frac{1}{2}u(W) + \frac{1}{2}u(W - z).$$

6.4. (Exponent in power law)

Students are asked to decide between two choices, Option A and Option B.

- Option A: Get \$1 with probability 95% and get \$381 with probability 5%
- Option B: Get \$20 for sure.

Most students prefer option B. Next the students are presented with the same two choices, but with \$300 added to all the outcomes, i.e.

- Option C: Get \$301 with probability 95% and get \$681 with probability 5%
- Option D: Get \$320 for sure.

Many students then switch and decide that they prefer the risky option. Show that, with the standard TK parameters, prospect theory will not predict this switch, but that for individuals where the power law exponent α is reduced to 0.6 then we will see a switch under prospect theory. (This experiment is reported in Bordalo, P., N. Gennaioli, N., Shleifer, A., 2011. Salience Theory of Choice Under Risk, NBER Working Paper #16387 who provide an alternative explanation for these observations).

6.5. (Laptop warranties

A company selling laptops offers an extended warranty on its products. For one laptop model costing \$900 the standard warranty is for one year and the extended warranty covers a further two years at a cost of \$75.

(a) Suppose that the probability of a breakdown in the extended warranty period is 0.1 and that a customer who does not take the warranty faces a cost of \$500 if this happens. Suppose that a customer's choices can be described using prospect theory with the TK parameters. Determine whether the customer is likely to take up the extended warranty.

(b) Now suppose that if a breakdown occurs and the warranty has not been taken up, then the laptop is equally likely to require replacement at a cost of \$900 or a simple repair at a cost of \$100 (so the expected cost is \$500). Calculate the value of the relevant prospect in this case and hence whether the customer will take up the warranty.

6.6. (Splitting prospects can change choices)

(a) Use prospect theory with $\alpha = \beta = 0.9$; $\lambda = 2$ and the standard decision weight functions given in Table 6.1 to calculate the values given to the following four prospects in order to predict which will be chosen.

$$A : (-\$100, 0.5; \$1000, 0.5)$$
$$B : (\$1000, 0.4)$$
$$C : (\$200, 0.3; \$300, 0.3; \$550, 0.4)$$
$$D : (\$340 \text{ with certainty.})$$

(b) Now suppose that prospects A, B and C are constructed in two stages, with \$340 received first and then gambles presented (so A is replaced with D followed by $(-\$440, 0.5; \$660, 0.5)$ and similarly for B and C). Show that none of the second stage gambles will be chosen.

6.7 (Risk seeking when value function is concave for losses)

For some individuals the general pattern is reversed and instead of the value function v being convex for losses it is either straight or mildly concave. Nevertheless the decision weight function may still lead to risk seeking behavior (where a risky option with the same expected value is preferred to a certain outcome). Find an example where this occurs and a certain loss of \$100 is less attractive than a gamble having the same expected value. You should use the decision weights w^- given in Table 6.1 and take

$$v(x) = x^{0.9} \text{ for } x > 0$$
$$v(x) = -2(-x)^{1.1} \text{ for } x < 0.$$

7

Stochastic Optimization

Maximizing profit from pumped storage

A pumped storage facility operates by pumping water up to a high reservoir when power is cheap and then letting that water flow out of the reservoir through hydro electric generators in order to generate power when it is expensive. There are inefficiencies so the water, once pumped up hill, can never deliver as much energy from letting it flow through the turbines as was needed to pump it in the first place. Nevertheless the overall exercise is worthwhile because the actual cost of electricity at night is so much lower than it is during the day. So cheap electricity can be used to fill the high reservoir during the night and then that power can be released during the day when prices are high. With increasing amounts of wind power which often delivers energy peaks at a time when demand is not high, the opportunities and need for pumped storage is greater than ever.

The Raccoon Mountain Pumped-Storage Plant is a good example of this sort of operation. It was built in the 1970s and is owned and operated by the Tennessee Valley Authority (TVA). The reservoir on Raccoon Mountain is more than 500 acres in size (200 hectares). When water is being pumped up this reservoir it can be filled in 28 hours. When electricity demand is high, water is released and can generate up to 1600 MW per hour. If the reservoir is full it would take 22 hours for it to empty if it was run continuously.

To decide how the pumped storage plant is to be operated each day is divided into different periods. Peak is for 7 hours, shoulder for 8 hours and off peak for 9 hours. Typically the reservoir is pumped in the offpeak hours during the night and is full at the end of that time, then the reservoir is run down during the 7 hours of peak and high shoulder demand (9 hours of pumping will give a volume of water sufficient for 7 hours of generation) But when demand and prices are high the reservoir can be used for a longer period so that the water level at the end of the day is significantly lower and cannot be completely replenished during the night; this leads to a slow drop in the reservoir level over successive days. Eventually the reservoir reaches a stage of completely emptying during the day, and in the morning the only water is that which was pumped up over night. This sets the constraint on the amount of power available on any day (independently of how high the electricity price might be).

The problem facing the plant operator is whether to take advantage of high prices today by running the generators for longer. And if the price is high enough to make it worthwhile to generate power during the shoulder period, how much power should be generated? The longer the generator is run, the lower the resulting water level will be and the less opportunity there will be to benefit if high prices occur again tomorrow.

Business Risk Management Edward Anderson

7.1 Introduction to Stochastic Optimization

In this chapter we will use the methods of optimization to determine how to make good decisions in a stochastic environment. Whereas in the previous chapter we primarily looked at decisions made with a clear cut set of options each of which contains a small number of possibilities with associated probabilities and consequences, in this chapter we will consider more complex problems. There will be a need to set up a model of what is happening (containing stochastic elements) and then to analyze this model. Because of this additional complexity our focus will switch back to the 'normative' rather than the descriptive. We ask what managers should do, rather than what they will do. For most of the models that we deal with in this chapter it is necessary to carry out an analysis using some computational tool like a spreadsheet.

An important difference exists between stochastic optimization problems where we get just one opportunity to make a decision and problems where we have an opportunity to make decisions at different points in time. For example in an operations environment in which demand for our product is uncertain we may either have a single decision at the start of the season on how much of a particular fashion item to make, or we may have a decision at the start of the season together with an opportunity to make more half way through the season when we have some information on how sales are going. In this chapter we will concentrate on models in which the stochastic elements have known probability distribution.

Before going on with our discussion of stochastic optimization we need to review some elementary ideas and notation that we use for optimization problems.

7.1.1 A review of optimization

An optimization problem is one in which we need to choose some *decision variables* in order to maximize an *objective function* subject to some *constraints* on the variables specifying which values are possible. These are the three critical components: variables, objective and constraints. If the variables are n real numbers that need to be chosen then we can think of the problem in the way that is shown in Figure 7.1. The decision variables are x and y, the set X is the set of feasible points defined by the constraints of the problem, and the objective function is shown by its contour lines. The optimal point is on the boundary of the set X and maximizes the objective function (sometimes optimization problems involve minimizing the objective, sometimes they involve maximizing the objective).

A first categorization of optimization problems distinguishes between problems in which there is only one local optimum and problems for which there are many possible local optima, and we need to compare them to find the best. From the point of view of Figure 7.1 the good property of having only one possible maximum arises from the nature of the objective function and the feasible set. The feasible set is convex (which means that the line joining two points in X can never go outside of X) and the objective function is concave (which implies that it looks like a hill: a straight line joining any two points on the surface defined by the objective function never gets above it.) With these properties there will just be one local maximum (which is also a global maximum). On the other hand if we have a minimization problem rather than a maximization problem, then we want the objective function to be convex, instead of concave.

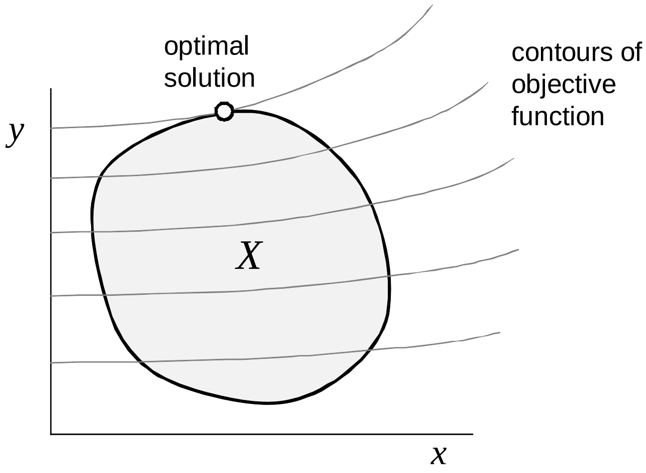


Figure 7.1 A diagram of a maximization problem

A convex maximization problem (with a concave objective function and a convex feasible set) may have its optimum on the boundary or in the interior of X. The Figure 7.2 shows a maximization problem of this sort with an interior maximum. The objective function for this problem ($P1$) is

$$P1 : \text{maximize } 4 - (x-1)^2 - (y-1)^2$$

with constraints:

$$\begin{aligned} 2y - (x-1)^2 &\geq 1 \\ 3y + 2(2x-3)^2 &\leq 6 \end{aligned}$$

The first constraint defines the lower boundary of the feasible region, and the second defines the upper boundary. The shaded region shows the feasible points.

The point which maximizes the objective is $x = 1$, $y = 1$ and this is inside the feasible region X. But to find the point that minimizes the objective function is harder. The contours of the objective function are shown with dashed lines. From this we can see that there are two local minima on the right of the Figure (and one more on the left). A local minimum is a point that is lower than any feasible points close to it. Small changes in the decision variables may lead to an infeasible point if one of the constraints is broken, but at a local minimum small changes that retain feasibility can only make the objective function larger. To find the global minimum we need to compare the objective function values at different local minima. In this example the higher point on the smooth part of the boundary is the global minimum, as we can see from the contour lines.

If a problem has a single local optimum (and hence this is also a global optimum) then it is much easier to find this numerically. We can start with an initial feasible point (some set of decision variables that satisfy all the constraints) and then consider changes that improve the objective function without breaking any of the constraints. This gives a new and improved

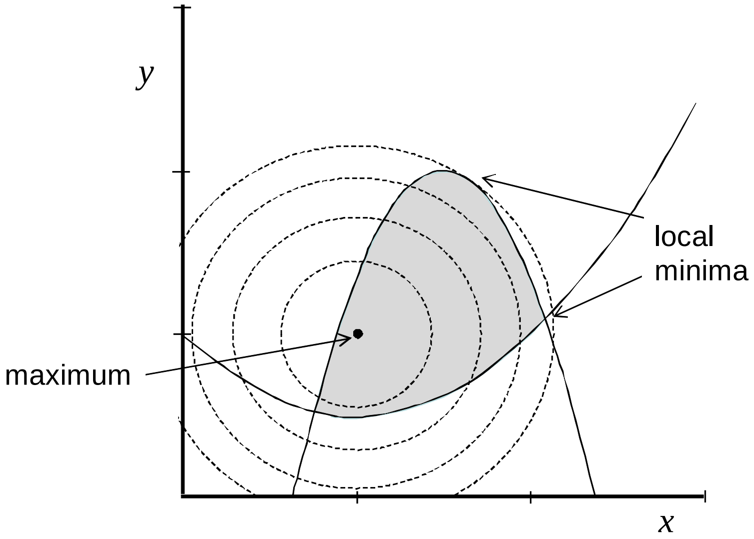


Figure 7.2 This example has a single maximum but more than one local minima.

feasible point and we repeat the procedure of searching for a change that improves the objective. Eventually we reach a local optimum when no further change is possible.

When there are multiple local optima then the problem becomes much harder. We can find a single local optimum using the approach of repeated improvements, but when we have discovered this we will not know how many other local optima there are. One idea is to use the repeated improvement approach again, but to begin at a different starting point: this could either lead us to the same local optima, or a different one. But for a complex problem we might try hundreds of different starting points, and still not be absolutely sure that we have found every local optima.

To learn about optimization it is important to try out the ideas in practice. There are a great many pieces of software available that are suited for different types of optimization problem. As a tool for solving simple optimization problems, the Solver add-in for Excel is quite satisfactory. It is a good exercise to solve the problem of minimizing the objective function for $P1$ using Solver. This has been done in the spreadsheet BRMch7-P1.xlsx. Try starting the optimization process at different points: these need not be feasible points, Solver will start by finding a feasible solution before improving it. You will find that the three different local minima can all be obtained depending on what initial values Solver is given.

An important special class of optimization problem, called a *linear program*, occurs when the constraints and the objective function are all linear. This will produce a well behaved problem with just one local optimum, since the set of feasible points is convex and the objective is both convex and concave. At first sight one might think that this produces a trivial problem, but in practice for problems with a large number of variables and a large number of constraints a solution requires a computer. An example of a linear program is

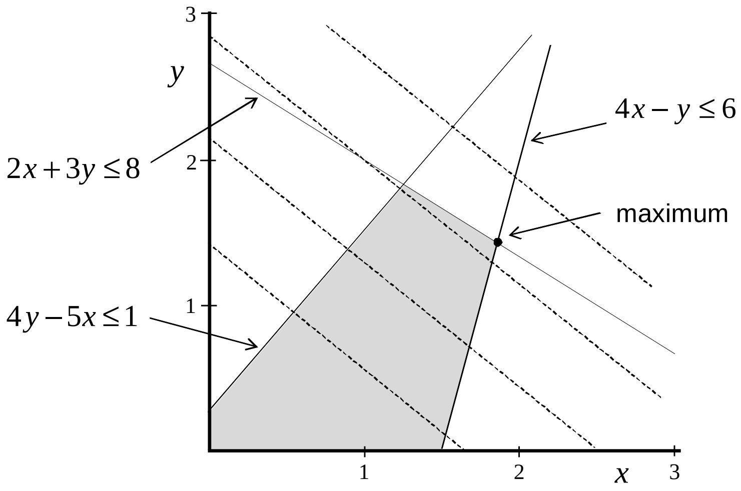


Figure 7.3 The feasible set and the optimal solution for the problem $P2$

given as problem $P2$ below.

$$P2 : \text{maximize } 6x + 7y + 2$$

with constraints:

$$\begin{aligned} 2x + 3y &\leq 8, \\ 4x - y &\leq 6, \\ 4y - 5x &\leq 1, \\ x &\geq 0, \\ y &\geq 0. \end{aligned}$$

Figure 7.3 shows the feasible points for $P2$. The dashed lines are contours of the objective function and the maximum occurs at the point shown ($x = 1.857$ and $y = 1.429$).

Linear programs can be solved for very large scale problems using special purpose software. Excel Solver has a setting under options for 'Assume Linear Model' and another setting for 'Assume Non-Negative' and these can be used for problems like $P2$. The file BRMch7-P2.xlsx gives a spreadsheet for this example.

An example of a two stage problem

In order to illustrate some critical ideas we start by analyzing a simple model problem. The Parthenon Oil Company (POC) sells fuel oil for home heating and uses a planning horizon related to the winter selling season. Most demand takes place in the months of October to

March. POC has storage tanks for fuel oil and buys it from oil companies that sell at a market price, that fluctuates month by month. In any month POC can supply customer orders either from its storage or by buying on the market. Customer contracts stipulate that POC will respond quickly to customer demand. To make things simple we consider the problem facing POC in February, near the end of its season. Suppose we start February with no oil in storage. In February POC buys oil from the market, delivers some to its customers right away and puts the rest in storage for the following month. Then in March the company can supply from storage or buy from the market. We need to decide x_1, how much oil to purchase in February, and x_2, how much to purchase in March.

The decision depends (amongst other things) on the price of oil in February and March, the storage cost, and the demand in each period. Suppose that it costs $\$5$ to store a barrel of oil for a month. With this information the problem can be modelled as a simple linear optimization problem with the objective to minimize overall cost. In practice the price and demand in March will be uncertain. Suppose that demand in February is 1000 barrels and the price is $160. Moreover we think that March can have one of three equally likely weather scenarios: normal, cold, or very cold. Cold weather means more demand for oil and at the same time the price that Parthenon pays for the oil will increase. The demand and price data for the three scenarios is given in Table 7.1.

Table 7.1: Data for three possible March scenarios for Parthenon Oil

Scenario	Probability	Oil Cost ($)	Demand (units)
Normal	1/3	160	1000
Cold	1/3	164	1200
Very Cold	1/3	174	1400

We write d for the demand in March and c for the cost in March. These are the things that are unknown at the point when Parthenon Oil makes a decision on how much to buy in February. Since demand in February is 1000 we know that the amount in storage at the end of February is $x_1 - 1000$ for which Parthenon pays $5 per unit. Thus we want to minimize total costs

$$\text{Minimize} \quad 160x_1 + 5(x_1 - 1000) + cx_2$$

subject to the constraints:

$$\begin{array}{ll} x_1 \geq 1000 & \text{(there is enough for February demand),} \\ x_1 - 1000 + x_2 \geq d & \text{(there is enough for March demand),} \\ x_2 \geq 0 & \text{(we cannot sell back to the market if we have too much).} \end{array}$$

In this problem we will find out what d and c are before we need to determine x_2, but we need to choose x_1 before we know d and c.

If we know in advance which of the three scenarios will occur then we can solve a linear program to find the optimal solution (this is in the spreadsheet BRMch7-Parthenon1.xlsx). We can calculate that if March is normal we should take $x_1 = 1000, x_2 = 1000$; if March is cold we should take $x_1 = 1000$, $x_2 = 1200$. Finally if March is very cold then it is worthwhile to buy all the oil we need in February, and $x_1 = 2400$, $x_2 = 0$. But the problem we face involves making a choice of x_1 right now. Since there is a two thirds chance of normal or cold weather, and under both these scenarios a purchase quantity of 1000 is optimal

- perhaps that is the best choice of x_1, and we can determine x_2 after we find out the demand in March.

However a different approach can produce a different answer. One common method in these circumstances is to deal with average behavior, rather than looking at individual scenarios. Often this is coupled with a sensitivity analysis in which we consider how sensitive the optimal decision is to changes in the parameters. If we find that it is sensitive, then we may consider a range of possible solutions corresponding to the range of values that we expect, but if we find that it is relatively insensitive then we may simply stick with the 'average' behavior. For the POC problem each of the three scenarios is equally likely. The average values are $c = 166$ and $d = 1200$. With these values we can solve the linear program and discover that the optimal solution $x_1 = 2200$ and $x_2 = 0$. So if we use average figures we should buy ahead in February leaving 1200 in storage at the end of the month.

Both these approaches are flawed and the proper way to approach this decision is to set up an optimization problem and embed within this problem the correct structure for the decisions we will take. Here we must allow the choice of different values for x_2 depending on the scenario. We call these x_{2A}, x_{2B}, and x_{2C}. Similarly we use the notation that the scenario demands and costs are given by d_A, d_B, and d_C; c_A, c_B, and c_C. Then we can formulate the problem as

$$\text{Minimize} \quad 160x_1 + 5(x_1 - 1000) + (c_A x_{2A} + c_B x_{2B} + c_C x_{2C})/3 \tag{7.1}$$

subject to the constraints:

$x_1 \geq 1000$ (there is enough for February demand)
$x_1 - 1000 + x_{2A} \geq d_A$ (there is enough for March demand for each scenario)
$x_1 - 1000 + x_{2B} \geq d_B$
$x_1 - 1000 + x_{2C} \geq d_C$
$x_{2A} \geq 0, x_{2B} \geq 0, x_{2C} \geq 0$ (purchases are all non-negative)

This linear program is set up in the spreadsheet BRMch7-Parthenon2.xlsx. Using Solver shows that the optimal solution has $x_1 = 2000$ meaning that there is 1000 in store at the start of March. No more is purchased under scenario A, 200 more is purchased under scenario B and 400 more is purchased under scenario C.

The Parthenon Oil Company example should serve as a warning against two common approaches adopted by planners faced with uncertainty. One option starts by computing an optimal solution for each scenario separately, then to compare these solutions and choose the decision that is best for as many of these scenarios as possible. The candidate solutions for the POC problem are then to store either 0 or 1400 units of fuel for the next stage. The optimal policy (as delivered by the stochastic program) is to store 1000 units. This does not correspond to the optimal solution in any of the scenarios. A second common approach is to take the average value as the prediction and solve the problem without a stochastic component, but for the Parthenon Oil Company problem this gives the wrong choice of x_1 and too much oil stored.

Understanding the structure of recourse problems

The Parthenon Oil Company example is a type of problem which is quite common. A first stage decision needs to be made, then as time goes by the uncertainty in the problem is resolved (For Parthenon the weather in March becomes known) and finally a second decision is made, which will depend on the first decision and the outcome of the uncertain event.

This is often called a *two-stage stochastic problem with recourse*. The word recourse here refers to what needs to be done at the second stage as a result of the random event.

Thus there is a nested structure to the problem. At the second stage we need to choose a decision variable y (if there is more than one variable involved then y will be a vector). The variable y will be chosen to minimize costs knowing both the first stage decision, x say, and the outcome of the random event. We will write ξ for the random variable capturing the randomness. In the Parthenon example, y is x_2, the amount ordered in March, and ξ is the weather in March and can take only three different values.

In order to complete the formal description of the problem note that costs occur at both the first and second stage. In the first stage they depend only on the first stage decision , so we can write this as $C_1(x)$, but at the second stage costs depend both on first and second stage decisions and also on the random outcome ξ. Thus the second stage costs are given by a function $C_2(x, y, \xi)$.

We write $Q(x, \xi)$ for the lowest cost at the second stage, given x and ξ. Thus

$$Q(x, \xi) = \min_{y \in Y} \{C_2(x, y, \xi)\},$$

where Y is the feasible set for the second stage decision (which might also depend on x and ξ though we don't show this in the notation).

The first stage decision is to choose a value of x that minimizes the expected overall costs assuming that the second stage decision is taken optimally. For a particular realization of the random variable ξ the total cost for a choice x is $C_1(x) + Q(x, \xi)$. Thus the stochastic problem we wish to solve is

$$\min_{x} \{C_1(x) + E_\xi[Q(x, \xi)]\},$$

where we write $E_\xi[\cdot]$ for the expectation with respect to the random variable ξ.

In the Parthenon example we can include the cost of storing oil in either the first stage or the second stage. Suppose that we take it as a first stage cost, then

$$C_1(x_1) = 160x_1 + 5(x_1 - 1000),$$
$$C_2(x_1, x_2, \xi) = c_\xi x_2,$$

where ξ takes the values A, B, or C. Thus

$$Q(x_1, \xi) = \min_{x_2} \{c_\xi x_2 : x_2 \geq d_\xi + x_1 - 1000, x_2 \geq 0\}.$$

Since c_ξ (c_A, c_B or c_C) is positive the minimum in Q occurs at the lowest possible value of x_2. This is obvious; each x_2 value should be made as small as possible. It will be set so as to just meet the demand in March (allowing for stored oil), or it will be zero if the stored oil is sufficient on its own to meet demand. So

$$x_{2A} = \max(0, d_A - x_1 + 1000),$$
$$x_{2B} = \max(0, d_B - x_1 + 1000),$$
$$x_{2C} = \max(0, d_C - x_1 + 1000).$$

This means that we have

$$Q(x_1, \xi) = c_\xi \max(0, d_\xi - x_1 + 1000).$$

So finally the expression we need to minimize (over x_1) is

$$160x_1 + 5(x_1 - 1000) + (1/3)c_A \max(0, d_A - x_1 + 1000) \\ + (1/3)c_B \max(0, d_B - x_1 + 1000) + (1/3)c_C \max(0, d_C - x_1 + 1000).$$

Thus we have transformed the problem into one where there is a single decision variable x_1 and a complex objective function. This can be solved by evaluating the objective function for different values of the decision variable $x_1$and there is no need to solve a linear programming problem.

The original linear program does the first and second stage minimizations in one go. It is easier to formulate the problem in this way, rather than working out the solution of the second stage optimization explicitly, and the eventual solutions reached will be the same.

Ordering with stochastic demand

A common problem that firms face is to determine a purchase quantity to meet demand when that demand is uncertain. For example a retailer selling fashion garments may need to order these well in advance of the selling season. If demand is more than the amount ordered then the retailer sells out of that item, but if demand is less than the amount ordered then there is left over stock at the end of the season. Usually this will be marked down to sell, sometimes to below cost price. The right amount to order depends on the distribution of demand and the difference between the amount of money made when selling at full price and the amount of money lost when selling in the end of season sale.

We formulate this as a stochastic optimization model. A decision is made by the retailer on how many items of a particular product to buy at a price of $\$c$ each. Each item is sold at a price $\$p$ during the selling season, and any left over at the end of the season are marked down to a price of $\$s$ in order to be sold in an end of season sale.

The difficulty here is that the demand is stochastic. If x items are purchased by the retailer and demand turns out to be D then the retailer will make a profit of $(p - c)x$ if x is less than D so that all the product are sold, and a profit of $pD + s(x - D) - cx$ if $x > D$ so that some items are marked down at the end of the selling season.

In this problem there is only a single decision to make rather than two stages, so the function $Q(x, \xi)$ (the lowest cost at the second stage, given the first stage decision x and a particular value of the random variable ξ) is replaced by the retailer profit given an order quantity x (the first stage decision) and a particular value of the random demand D, and we change from minimizing costs to maximizing profits. We write this profit function as $G(x, D)$ which is given by

$$G(x, D) = (p - c)x - (p - s)\max(x - D, 0).$$

(You can check that this expression is right for the two cases that x is either less than or more than D). For a fixed value of D this has a maximum at $x = D$ as we would expect (the retailer buys exactly the number of items that it will sell).

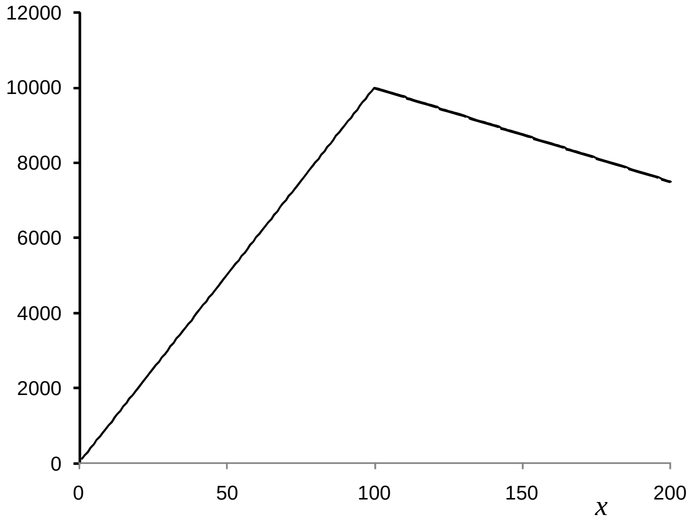


Figure 7.4 A plot of $G(x, D)$ for $D = 100$

Now let's consider a specific example. Troy Fashions Ltd (TFL)sells a type of winter coat for \$200 and works with a clothing manufacturer who can supply at \$100 per item. Winter coats that are not sold at the end of the season will be marked down to \$75. Thus $c = \$100$, $p = \$200$, and $s = \$75$. Figure 7.4 shows the graph of the profit function $G(x, D)$ when $D = 100$.

Suppose that Troy Fashions knows the distribution of the demand D. What is the best choice of order quantity x?We will assume that Troy Fashions is risk neutral and simply wishes to maximize the expected profit. Suppose that the demand D can take values between 0 and M and has a density function f, then the maximum expected profit is given by

$$\max_x E_D\left[G(x, D)\right] = \max_x \int_0^M G(x, z) f(z) dz.$$

To find the best choice of x we can look for a value so that the derivative of the expected profit is zero:

$$\frac{d}{dx} E_D\left[G(x, D)\right] = \frac{d}{dx} \int_0^M G(x, z) f(z) dz = 0.$$

It happens that this derivative is particularly easy to evaluate. There is a well know theorem in mathematics (the Leibniz integral rule) about how to take the derivative of an integral. Since the limits of the integral do not depend on x

$$\frac{d}{dx} \int_0^M G(x, z) f(z) dz = \int_0^M \frac{\partial}{\partial x} G(x, z) f(z) dz = 0.$$

Now the derivative of G with respect to x depends on the value of the demand, as we can see from Figure 7.4. In fact

$$\frac{\partial}{\partial x}G(x,z) = p - c \text{ if } z > x,$$
$$\frac{\partial}{\partial x}G(x,z) = s - c \text{ if } z < x.$$

So

$$\begin{aligned}\int_0^M \frac{\partial}{\partial x}G(x,z)f(z)dz &= (s-c)\int_0^x f(z)dz + (p-c)\int_x^M f(z)dz \\ &= (s-c)F(x) + (p-c)(1-F(x))\end{aligned}$$

where we write $F(x)$ for the cumulative distribution function for the demand. Setting this to zero gives

$$(p-c) = (p-s)F(x).$$

Thus x should be chosen so that

$$F(x) = \frac{p-c}{p-s}. \tag{7.2}$$

Now we can return to Troy Fashions where $c = \$100, p = \$200, s = \$75$. Suppose that the demand is uniformly distributed between 0 and 200, so that $F(x) = x/200$ for $0 \le x \le 200$. Hence substituting into (7.2), the optimal choice of order quantity, say x^*, is given by

$$\frac{x^*}{200} = \frac{p-c}{p-s} = \frac{100}{125} = 0.8.$$

Hence $x^* = 200 \times 0.8 = 160$, and TFL should order 160 of this coat at the beginning of the season.

With this distribution of demand we can calculate the expected profit, $E_D\left[G(x,D)\right]$, as a function of x:

$$\begin{aligned}\frac{1}{200}\int_0^{200} G(x,z)dz &= \frac{1}{200}\int_0^{200}\left[(p-c)x - (p-s)\max(x-z,0)\right]dz \\ &= \frac{(p-c)x}{200}\int_0^{200} dz - \frac{(p-s)}{200}\int_0^x (x-z)dz \\ &= (p-c)x - \frac{(p-s)}{200}\left[(xz - \frac{z^2}{2})\right]_0^x \\ &= (p-c)x - \frac{(p-s)}{200}\frac{x^2}{2}.\end{aligned}$$

This is plotted in Figure 7.5 for the values $c = 100$, $p = 200$, $s = 75$ and this confirms that 160 is indeed a maximum for this function and achieves an expected profit of \$8000 for TFL from this item.

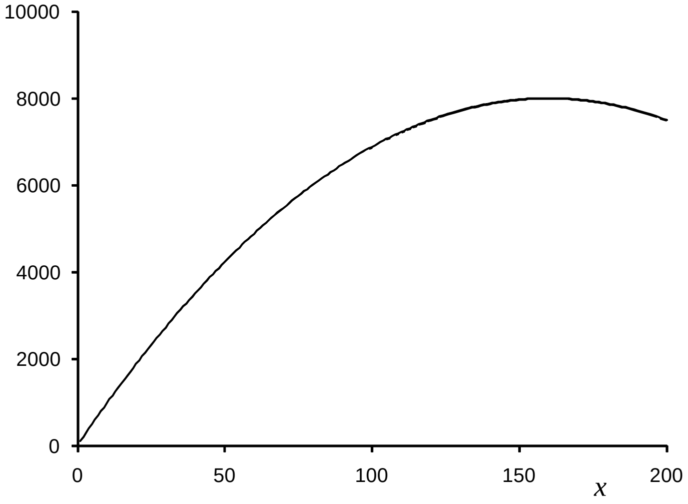


Figure 7.5 The expected profit as a function of x

7.2 Choosing scenarios

The two examples of Parthenon Oil and Troy Fashions have both been relatively easy to solve. In practice we are unlikely to have a problem with the kind of closed form solution that exists for Troy Fashions, or the small set of scenarios that exist for Parthenon Oil. A more typical problem combines the features we have seen in these two problems; it will have a stochastic element with a continuous distribution like Troy Fashion's demand, but it will also have the more complex multi-stage structure of Parthenon Oil.

For these more complex problems we need to do something different and the best approach is to generate a set of different scenarios in order to make estimates of expected profit with different decisions. In this section we will introduce Monte-Carlo simulation as a way of generating scenarios and in the next section we will show how these ideas can be applied to more complex multistage problems.

Now we return to TFL. Suppose that instead of carrying out a complete optimization we look instead for an optimal solution when each of four demand scenarios is equally likely: demand is either 30, 80, 120, or 170. The problem becomes

$$\max_x 0.25G(x, 30) + 0.25G(x, 80) + 0.25G(x, 120) + 0.25G(x, 170)$$

with

$$G(x, D) = 100x - 125 \max(x - D, 0).$$

So the problem can be written

$$\max_x [100x - \frac{125}{4}(\max(x - 30, 0) + \max(x - 80, 0) + \max(x - 120, 0) + \max(x - 170, 0))].$$

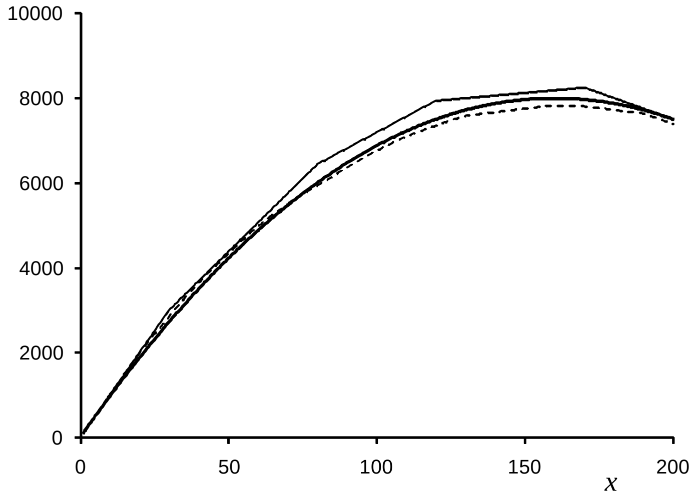


Figure 7.6 Using scenarios to approximate the expected value of profit

Figure 7.6 shows the exact objective and this approximate objective (the thinner line with 5 segments).

By adding more scenarios and giving them equal weights we will end up with better and better approximations. One way to generate scenarios is to choose them randomly. (The idea of generating scenarios randomly is called Monte Carlo sampling and we will have more to say about this in the next section.) Here any value of demand between 0 and 200 is supposed to be equally likely. So we choose 15 random numbers between 0 and 200 (rounded to two decimal places): 110.59, 57.42, 168.24, 17.57, 98.77, 190.82, 130.29, 42.81, 188.80, 35.12, 158.13, 24.18, 72.81, 128.20, 62.61. Then we construct the appropriate objective function:

$$100x - \frac{125}{15}(\max(x - 110.59, 0) + \max(x - 57.42, 0) + ... + \max(x - 62.61, 0)).$$

This is the dashed line in Figure 7.6.

7.2.1 How to carry out Monte Carlo simulation

The idea of a Monte Carlo simulation is an important one in modelling risk. Even if we cannot write down explicit expressions to evaluate the expected performance of a decision, we can almost always make an estimate using a simulation. It is worth thinking carefully about the way that a Monte Carlo simulation is carried out.

The idea is to select the variables in a way that matches the actual stochastic process through which they will be determined. Then an average over a large number of scenarios each drawn with the appropriate probability distribution will tend to the true expected value. This is Monte-Carlo sampling. For problems of stochastic optimization there is often a

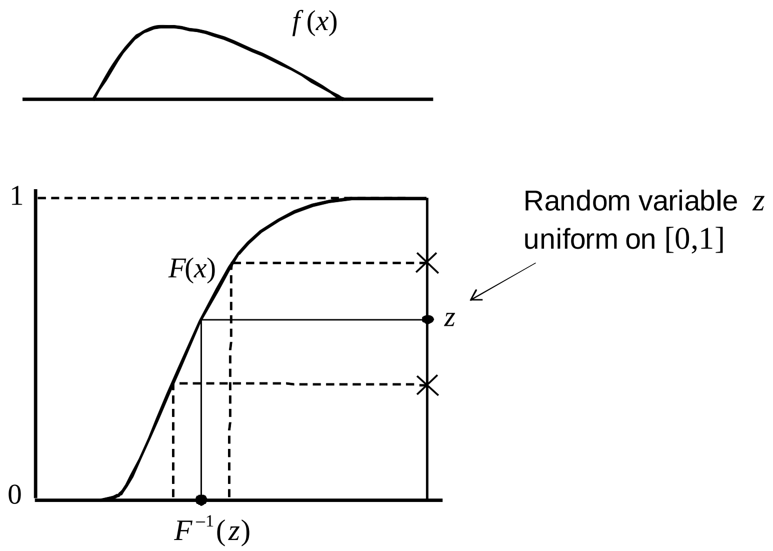


Figure 7.7 A uniform random variable z is transformed into a given distribution through $F^{-1}(z)$

structure involving a stochastic evolution over time with the random variables occurring at one stage feeding into what happens at the next stage and there will also be decision variables that need to be chosen. We will explain how this works out in the next section. For the moment we want to focus on the generation of random scenarios through Monte-Carlo simulation.

There are many excellent spreadsheet-based programs that allow Monte-Carlo sampling to be carried out very simply and automatically. Rather than base this chapter of this book on one of these programs, we will instead use the simplest possible approach using just what is available within Excel itself. This makes things a bit more cumbersome with larger spreadsheets (as we shall see) but has the advantage of being straightforward and transparent.

In order to create random scenarios from within a spreadsheet we will use the RAND function. This function is written RAND() (and takes no arguments). It returns a uniformly distributed random number greater than or equal to 0 and less than 1. A new random number is returned every time the worksheet is calculated or F9 pressed. So to get a demand with a uniform distribution between 0 and 100 we can put the formula =RAND()*100 in a cell.

In practice it is rare that we want to generate scenarios using a uniform distribution, so what we need is a way of getting from a uniform distribution to some other distribution. This is achieved using the inverse transform method as is illustrated in Figure 7.7.

Suppose that a required density function f is given, with associated CDF of F. We start by generating a random number z that is uniform between 0 and 1 and then transform that as shown in the diagram to the point y where $F(y) = z$. If the density f is positive in its range then F will be strictly increasing and there will be a single point y defined by this procedure. Formally we can say that y is given by the inverse of F applied to the number z and write this as $y = F^{-1}(z)$.

Figure 7.7 suggests that this process will be more likely to produce a number where the function F has a large slope, i.e. where the density function has a high value. This is exactly what we want. Now we establish more formally that the inverse transform method will produce a random variable with the right distribution function.

Suppose that X has a uniform distribution on $[0, 1]$ and the random variable Y is given by $F^{-1}(X)$. We can get a sample from Y by taking a sample x from the distribution X and then choosing a value y so that $F(y) = x$. To find the distribution of y notice that

$$\begin{aligned}\Pr(y \leq a) &= \Pr(F(y) \leq F(a))\\ &= \Pr(x \leq F(a))\\ &= F(a).\end{aligned}$$

The first equality is because F is an increasing function, the rest follows from the way that y is chosen and the fact that x is a sample from a uniform distribution. Thus for any value a, the probability that y is less than a matches what it would be if y had the distribution F . Hence we have shown that y has the distribution we want.

Worked Example 7.1

We want to generate a scenario in which the demand has a distribution with density $f(x) = 2 - x$ on the range $0.5 \leq x \leq 1.5$. What formula should be put in the spreadsheet cell to produce a sample from this distribution?

Solution

We need to start by finding the CDF for this density function. The function F is zero below 0.5, F is 1 above 1.5 and between these values we have

$$\begin{aligned}F(x) &= \int_{0.5}^{x} f(z)dz\\ &= \int_{0.5}^{x} (2-z)dz = \left[2x - \frac{x^2}{2}\right]_{0.5}^{x}\\ &= 2x - \frac{x^2}{2} - (1 - \frac{1}{8})\end{aligned}$$

We can check that this takes the value 1 when $x = 1.5$. We have $2 \times (3/2) - (9/8) - (7/8) = 1$ as required.

Given a value z that is uniform we generate a new value from the inverse of this function. Hence we need to find the value of x which solves

$$2x - \frac{x^2}{2} - \frac{7}{8} = z.$$

Thus

$$x^2 - 4x + \frac{7}{4} + 2z = 0$$

and

$$x = 2 \pm \frac{1}{2}\sqrt{9 - 8z}.$$

Now we need to decide whether the higher or lower root is correct; should we take the plus or minus in this expression? We need values of x between 0.5 and 1.5, so we need the minus sign (i.e. the lower root). Hence we reach the following expression for the spreadsheet cell:

```
=2-0.5*SQRT(9-8*RAND())
```

We will return to the example of Troy Fashions, but now we will suppose that the distribution of demand is normal with a mean of $\mu = 100$ and a standard deviation of $\sigma = 20$. We can use equation (7.2) to determine the optimal order quantity which satisfies $F(x) = 0.8$, which leads to a value $x = 117$. This number can be obtained from tables of the standard normal distribution that give a z value of 0.84 to achieve a probability of 0.8. Thus the optimal order quantity is

$$x = \mu + z\sigma = 100 + 0.84 \times 20 = 116.8$$

and we round this up to 117.

We will illustrate the Monte-Carlo simulation approach by using this method to estimate the expected profit with this choice of order quantity. The idea is to average over different scenarios for the demand with the scenarios drawn from the desired population. The spreadsheet BRMch7-TroyFashions.xlsx carries out 1000 random draws from a normal distribution for demand in order to estimate the average profit. Each row in the spreadsheet represents a different randomly drawn demand scenario. Have a look at the cell `B4` which generates one of the random demands: `=NORMINV(RAND(),100,20)`. The function `NORMINV`(y, μ, σ) finds the x value for which a normal distribution with specified mean μ and standard deviation σ achieves the probability y of being less than x. It is precisely the inverse of the CDF for a normal distribution and so is the function we need to generate demands with a normal distribution.

The average profit is around $\$9300$ Even taking the average over 1000 repetitions, which might be expected to be close to the real expected value, there is still a lot of variation (try pressing `F9` repeatedly, which recalculates all the random numbers, and you should see that the average profit jumps around a lot going from less than $\$9200$ to more than $\$9400$).

7.2.2 Alternatives to Monte-Carlo

In the Troy Fashions example it is surprising how variable the estimation of expected profit is even with 1000 scenarios analyzed. There are a number of reasons for this: essentially the randomness in the sample often produces a clustering of the sample in demand regions where there are either higher than average or lower than average profits. There is an alternative approach which is to spread out the sample points in a regular way. This has been done on the right hand side of the spreadsheet BRMch7-TroyFashions.xlsx. Instead of making an estimate of the mean by taking 1000 random numbers between 0 and 1 and using these to generate 1000 demand values, this calculation takes just 98 values 0.01, 0.02, 0.03, ..., 0.98, 0.99 and these values are then used, instead of random numbers, to generate a set of demand values and the expected profit is estimated by averaging these. The inverse transform method is being used here as well to ensure the right distribution of values for the demand.

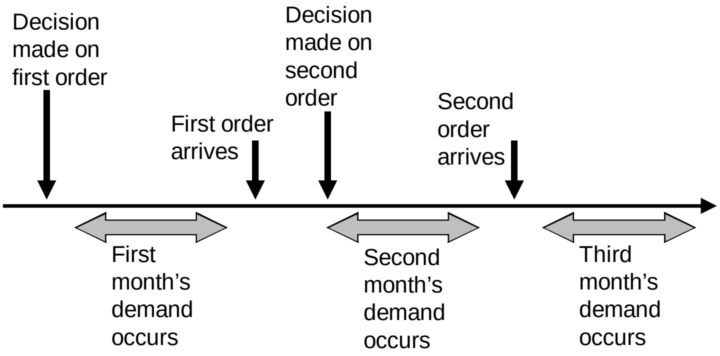

Decision made on first order
First order arrives
Decision made on second order
Second order arrives
First month's demand occurs
Second month's demand occurs
Third month's demand occurs

Figure 7.8 Timeline for Gardenia Patio Supplies

This gives a far better estimate than a Monte Carlo approach and this approach should always be used when there is just one or two random variable involved in the simulation.

However the Monte Carlo method really comes into its own when there are many different random variables each with a distribution (or with a joint distribution on them all). Suppose that there are 3 independent random variables involved in the calculation. The approach of taken evenly spread random numbers now requires 1000 repetitions just to get 10 different values for each of the three variables. In essence we are evaluating the expected profit by averaging the results that occur in a 3-dimensional grid. As the number of variables increases it gets harder and harder to make a grid based approach work. For example if there are 15 different random variables each of which is uniformly distributed on $(0, 1)$ then we might sample these by letting each random variable take 3 values: say 0.25, 0.5 and 0.75. But with 15 variables there are $3^{15} = 14, 348, 907$ scenarios which is far to many to allow an exact solution. Using a Monte Carlo method may be about the only practical way to proceed. In fact it works reasonably well in most cases with the accuracy of the solution determined by the number of scenarios independently of the number of random variables.

7.3 Multi-stage stochastic optimization

To illustrate our discussion we will consider a simple multi-stage example problem.

Suppose that Gardenia Patio Supplies (GPS) has a planning horizon of $T = 3$ months. At the beginning of each month GPS places an order for garden chairs from its supplier and this is delivered at the beginning of the next month. Demand occurs during the month, and if this demand is more than the available inventory then customers will go elsewhere (so that the excess demand is lost). The time line is shown in Figure 7.8.

We assume that demand in each month has a normal distribution with mean 50 and standard deviation 10 and demand in successive months is independent. Garden chairs are bought at $100 each and sold at $140. To hold a chair over from one month to the next is expensive - it costs $10. Suppose we have in inventory an amount $y_1 = 50$ at the beginning of the first period. The first decision is to choose x_1, the number of chairs ordered in the first month. If demand in the first month is D_1 then we sell $\min(y_1, D_1)$ and hold over an amount $\max(y_1 - D_1, 0)$. We begin the second period with an amount $y_2 = \max(y_1 - D_1, 0) + x_1$ and the whole process repeats. We allow for the cost of holding over inventory at the end of the three months but otherwise do not consider any further costs.

A spreadsheet Monte-Carlo simulation has been set up in BRMch7-Gardenia.xlsx, with each row of the spreadsheet representing a different scenario and each scenario involving three different stochastic demands. There are 1000 different scenarios and the spreadsheet has been set up with an initial inventory of 50 and an order of 55 at the start of week 1 and an order of 40 at the start of week 2.

The contents of cell B6 are `=ROUND(NORMINV(RAND(),Mu,Sigma),0)`. The function `NORMINV` has the role of converting the uniformly distributed random number `RAND()` into a normal distribution with mean from the cell with name `Mu` and standard deviation from the cell with name `Sigma`. The function `ROUND(.,0)` rounds this to the nearest integer (customers buy complete chairs not just parts of them!). The same formula is repeated for the other monthly demand in columns E and H. The sales in each month (columns C, F and I) are just given by the minimum of the inventory at the start of the month and the demand. The inventory at the start of month 2 in cell `D6` is `=Inv_S+Order1-C6` which is the initial inventory (from the cell named `Inv_S`) plus the month 1 order (from the cell named `Order1`) minus the month 1 sales from cell `C6`. Column G contains a similar formula.

Column J gives the costs consisting of the total cost of the products ordered at \$100 per chair and the cost of holding stock over at $\$10 \times$ (starting inventory $-$ sales) for each of the three months. Profits are obtained from three months of $\$140 \times$ (sales)$-$ total costs.

Notice that the profit figures for different scenarios vary wildly. Even after taking the average over 1000 repetitions, which might be expected to be close to the real expected value, there is still a lot of variation (try pressing `F9` and seeing what happens to the average profit.)

Now we ask what are the best values of the month 1 and month 2 orders. This problem is a little similar to the Parthenon Oil Company example and we could set it up as a linear program with decision variables x_1 and x_2 being the two orders made. The easiest thing to do is to use Solver directly to maximize the average profit calculation in BRMch7-Gardenia.xlsx. But in order to do so we have to fix the demand values to produce a fixed set of 1000 scenarios against which we will evaluate changes in the order quantity. Using the "paste values" function this has been done in the second worksheet in BRMch7-Gardenia.xlsx. Try using Solver to find the best choice of the two orders. It turns out that with this set of scenarios the best choice is to set $x_1 = 49$ and $x_2 = 36$.

There is a hidden difficulty here that Solver deals with 'out of sight' and that is the use of the functions like $\min$(inventory,demand) to calculate the sales. This introduces corners into the functions (i.e. places where the derivatives jump) and this in turn make it much harder to solve the optimization problem.

There are ways to set up the problem that avoid these non-smooth functions. In general an optimization problem in which terms like $\min(x, y)$ appear, but that still has a convex feasible set and (if we are maximizing) a concave objective function, can always be replaced by a smooth version of the same thing. We simply replace a constraint of the form $A \leq \min(x, y)$ with two constraints: $A \leq x$ and $A \leq y$. Note that if we have a constraint like $A \geq \min(x, y)$ then the feasible region will no longer be convex and we lose the property of the problem having only one local optimum. If the objective involves maximizing $\min(x, y)$ then we create a new variable v and then we maximize v subject to constraints $v \le x$ and $v \le y$. In doing these manipulations it helps to remember the rules of operating with $\min$ and $\max$.

$$\begin{aligned}
\max(x, y) &= -\min(-x, -y) \\
a \min(x, y) &= \min(ax, ay) \text{ if } a \ge 0 \\
\min(\min(x, y), z) &= \min(x, y, z) \\
z + \min(x, y) &= \min(z + x, z + y).
\end{aligned}$$

7.3.1 Non-anticipatory constraints

Our discussion of the GPS example so far fails to treat the problem correctly since it forces us to set both x_1 and x_2 at the beginning. It is important to realize that, at the time that the value of x_2 is chosen (the order placed at the start of month 2) the company will already have information on the demand during the first month. If there has been high demand - leading to zero inventory held over - then it makes sense to order more, but if there has been low demand and there are relatively high levels of inventory at the start of month 2 then it will be better to order less.

Thus we need to set up the optimization problem paying careful attention to the exact information that can be used in any decision. A formulation that forces us to choose x_2 at the start gives too little flexibility, but it is easy to make the mistake of a formulation which allows too much flexibility.

Consider taking just the first three scenarios that are listed on the second sheet of workbook BRMch7-Gardenia.xlsx. Thus the demand values for the three scenarios chosen are given by the following table:

Scenario:	A	B	C
d_1	60	41	52
d_2	48	58	36
d_3	34	66	53

A natural formulation is to choose different values of x_2 for different scenarios. Scenario A with a high value of the first month demand can then have a higher value of x_2 than scenario B where the first month's demand is only 41. This is the setup shown in the third sheet of the work book. Try using Solver to find the best choice of the 4 different variables x_1 and the values of x_2 for the three different scenarios (x_{2A}, x_{2B}, x_{2C}). You should find that the optimal values are $x_1 = 49$, $x_{2A} = 33$, $x_{2B} = 66$, $x_{2C} = 40$. Instead of scenario A getting a large order in the second month, it has a small order. The reason is that the second month order is really only required for the third month;s demand. It is the low value of d_3 in scenario A that makes it optimal to order a small amount in the second month.

We can see that by allowing the value of the variable x_2 to depend on the scenario A, B or C then in effect we allow x_2 to be affected not only by d_1, but also by d_2 and d_3 which is information not available at the time when the decision is made. In selecting a particular scenario we are making a selection of future variables as well. So implicit in the procedure we have used, is that the decision at the end of month 1 depends on events that have not yet

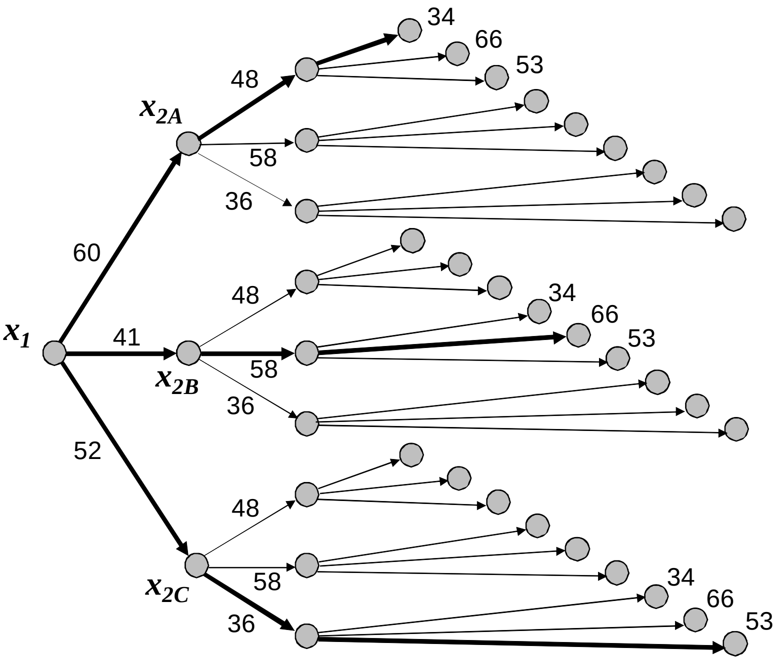


Figure 7.9 An example of a scenario tree

occurred. Not being able to look ahead when we make decisions is called a *non-anticipatory constraint.*

A formulation of a stochastic optimization problem might include a specific non-anticipatory constraint forcing decisions that are made with the same information available to be the same. But this still leaves the possibility of a wrong formulation as we saw in the example with just three scenarios. Usually it is safer to build the non-anticipatory constraint more directly into the structure of the problem.

In order to correctly solve the GPS problem we need to work with a *scenario tree* such as we show in Figure 7.9, rather than just a set of scenarios. The scenario tree is a type of decision tree, as was introduced in Chapter 5 (but since the decision variables here are continuous they do not contribute arcs in this tree). In this Figure the scenarios are built up of demand realizations and for each month three possible demands are shown. Of course this dramatically over simplifies what is actually possible, since in each month any demand realization from around 30 to around 70 is quite possible.

The bold arrows in the Figure show the demand numbers that occur in the scenarios A, B and C that are now just three out of a possible 27 different scenarios. Using this set of 27 scenarios we can construct a more appropriate model in which x_2 is set differently at the three initial nodes according to whether demand has been 60, 41, or 52.

In constructing a scenario tree like this, the more accurately the stochastic component is represented, the more scenario branches there will be at each stage. This can lead to enormous trees and hence great computational difficulties in finding optimal solutions. One option is to reduce the accuracy of the model for steps further into the future by reducing the number of branches at higher levels of the tree.

There has also been a great deal of research on how problems of this sort can be effectively solved numerically. This research is well-beyond our scope in this book, but we can sketch a couple of ideas that are useful. One idea is to decompose the problem into separate subproblems for each of the first stage outcomes. These would be the three subtrees in Figure 7.9. If we guess a value for x_1 then it is quite easy to find optimal values for the different values of x_{2A}, x_{2B}, x_{2C}. The solution procedure will also usually generate sensitivity information (especially when the problem is linear), so that for each subtree we can test the consequence of small changes up or down in x_1. This information can be used to find a change in x_1 that produces an overall improvement in expected profit. In the linear case this idea can be transformed into a process of generating additional constraints in the master problem that can be effective numerically.

Another idea that is important from a practical perspective is to use scenario trees where different scenarios are given different probabilities of occurring rather than being equally weighted. This idea is related to the way that Monte Carlo simulations are generated - it is often possible to get better estimates of the critical quantities by ensuring that the set of samples drawn from the random distributions have particular properties.

There are two other issues that we should mention in relation to multi-stage stochastic optimization. First we note that it is usually just the first stage decision which is of interest. In practice we can expect a stochastic optimization problem to be solved on a rolling basis. So in the Gardenia Patio Supplies example the problem is solved and an order decided for month 1. Then at the end of the month when demand for that month is known the problem can be solved again, but this time pushing one more month out into the future.

The second point to make is that there is a close relationship between this type of problem and that which can be solved using dynamic programming. The dynamic programming approach involves looking more closely at what might influence the decisions made at any point in time. In this way the decision is seen as a function of the *state* at time t . For example in the GPS example the decision on what to order at the beginning of the second month can only depend on the amount of inventory carried over from the first month, if we assume that the demand we observe in each month is independent of the previous demands. This is because at the time the decision is made we can ignore the costs and profits already achieved and look at optimizing the remaining profit. This (optimal) remaining profit can only depend on the state of the system - which just means the current inventory level. If we can find a way to formulate the problem so that at each stage decisions are a function of the state at that stage, then incorporating this into the solution procedure through some type of Dynamic Programming recursion will usually be worthwhile. .

7.4 Chance constraints

So far we have assumed that the problem can be posed as minimizing (or maximizing) the expected value of an objective function. We can use exactly the same approach to deal with a case where the decision maker is risk averse. In this case we simply define a concave utility function for the decision maker and incorporate the utility function into the objective function. But there are occasions when a different approach is valuable.

We suppose that the decision maker is concerned with risk and in particular wishes to avoid bad outcomes. If there is a certain level of loss that is unacceptable then one option is to maximize expected profit as before but to insist that any solution chosen avoids the possibility of a loss greater than the chosen value. So for example if we are solving a recourse problem of the following form

$$\min\{C_1(x) + E_\xi[Q(x, \xi)]\},$$

then we could add a constraint

$$C_1(x) + Q(x, \xi) < M \text{ for all } \xi.$$

Since this is a minimization problem it is large values of the costs given by $Q(x, \xi)$ that are to be avoided. However in many problems it is not necessary to entirely avoid the possibility of large costs, but just to ensure that it is very unlikely that a large cost occurs. Thus we end up with a constraint of the form

$$\Pr\{C_1(x) + Q(x, \xi) > M\} \leq \alpha,$$

which is called a *chance constraint*.

We will discuss a version of the problem where we maximize the expected profit $E_\xi[\Pi(x, \xi)]$ from a decision x with stochastic behavior described by the random variable ξ. Then the equivalent chance constraint can be written

$$\Pr\{\Pi(x, \xi) < -M\} \leq \alpha.$$

Notice that we are still using the same objective function, but just with an added constraint. So if $\alpha = 0.01$ and $M = \$1,000,000$ then we can express the chance constraint in words by saying that we maximize expected profit subject to the condition that the decision x does not give more than a 1% chance of losing a million or more.

This is closely related to the use of value at risk (VaR) discussed in Chapter 3. For example suppose that a company wishes to maximize expected profit but must operate under risk constraints that impose a limit of \$500,000 on absolute 95% VaR. Then this can be stated as a problem with a chance constraint

$$\text{maximize } E_\xi[\Pi(x, \xi)]$$

$$\text{subject to } \Pr\{\Pi(x, \xi) < -500,000\} \leq 0.05.$$

As an example of this type of problem we return to the portfolio optimization problem we introduced in Chapter 2. Suppose there are two investments, both having a normal distribution for the profits after 1 year. A \$1000 invested in the first investment returns an expected profit of \$1000 with a standard deviation of \$400, while the same amount invested in the second investment gives an expected profit of \$600 with a standard deviation of \$200. A natural stochastic optimization problem with a chance constraint is to suppose that we have \$1000 to invest and wish to maximize our expected return subject to the condition that the probability of losing money is less than 0.5% say. Alternatively we can express this by saying that the absolute 99.5% Value at Risk is less than \$0

Suppose we invest an amount $1000w_1$ in the first investment and $1000w_2$ in the second. If we assume that the performance of the investments are independent, then the profits earned follow a normal distribution with mean $1000w_1 + 600w_2$, so the problem can be written

$$\text{maximize } 1000w_1 + 600w_2$$

$$\begin{aligned} \text{subject to} \quad & \Pr\{w_1X_1 + w_2X_2 < 0\} \le 0.005 \\ & w_1 + w_2 = 1, \\ & w_1 \ge 0, w_2 \ge 0. \end{aligned}$$

where X_1 and X_2 are the random variables giving the individual investment returns. The variance of the total return is given by

$$(400w_1)^2 + (200w_2)^2$$

and the standard deviation is given by the square root of this.

The probability of making a loss can be calculated from the z value giving the number of standard deviations that the mean is above zero. We have

$$z = \frac{1000w_1 + 600w_2}{\sqrt{(400w_1)^2 + (200w_2)^2}}.$$

We can use tables of the normal distribution or the NORMINV function in a spreadsheet to show that we need $z \le 2.5758$ in order to ensure that the probability of a value less than zero is no more than 0.005. Thus the constraint becomes

$$1000w_1 + 600w_2 \le 2.5758\sqrt{(400w_1)^2 + (200w_2)^2}.$$

We can divide through by 100 and square both sides of this inequality to show that the problem can be written

$$\begin{aligned} \text{maximize} \quad & 1000w_1 + 600w_2, \\ \text{subject to} \quad & (10w_1 + 6w_2)^2 \le (2.5758)^2 \left(16{w_1}^2 + 4{w_2}^2\right), \\ & w_1 + w_2 = 1, \\ & w_1 \ge 0, w_2 \ge 0. \end{aligned}$$

At the optimum the inequality will hold with equality and, since we can substitute using $w_2 = 1 - w_1$, we end up with

$$a\left(16{w_1}^2 + 4(1 - w_1)^2\right) - (10w_1 + 6(1 - w_1))^2 = 0$$

where $a = (2.5758)^2 = 6.6349$. Multiplying this out we get

$$(20a - 16)w_1^2 - (48 + 8a)w_1 + 4a - 36 = 0.$$

The maximum is achieved at the higher of the two roots of this quadratic equation. So

$$\begin{aligned} w_1 &= \frac{1}{5a - 4}\left(a + 6 + \sqrt{a\,(61 - 4a)}\right) \\ &= \frac{1}{29.1745}\left(12.6349 + \sqrt{228.6413}\right) \\ &= 0.95137 \end{aligned}$$

giving a split of \$951 in the first investment with the remaining \$49 in the second investment. In this case any weighting with less than \$951 in the first investment will achieve the chance constraint of a less than 0.5% probability of a loss.

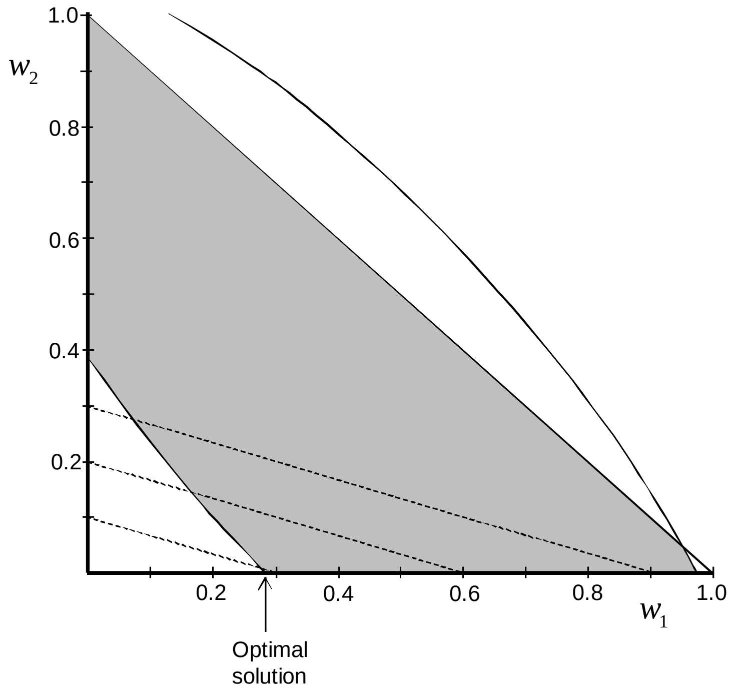


Figure 7.10 Optimizing a three-investment portfolio with a Value at Risk constraint

Now we extend this to three stocks by adding a third investment with mean profit of $1200 and standard deviation 600. The problem becomes

$$\begin{aligned} &\text{maximize} \;\; 1000w_1 + 600w_2 + 1200w_3, \\ \text{subject to} \quad &(10w_1 + 6w_2 + 12w_3)^2 \leq (2.5758)^2 \left(16{w_1}^2 + 4{w_2}^2 + 36w_3^2\right), \\ &w_1 + w_2 + w_3 = 1, \\ &w_1 \geq 0, w_2 \geq 0, w_3 \geq 0. \end{aligned}$$

We can substitute for $w_3 = 1 - w_1 - w_2$ in order to reformulate this as an optimization problem over a two dimensional region. This is shown in Figure 7.10.

The objective function becomes

$$1000w_1 + 600w_2 + 1200(1 - w_1 - w_2) = 1200 - 600w_2 - 200w_1$$

and the dashed lines in the Figure show contours of this. The feasible region is shown shaded. The straight line upper boundary arises from the constraint $w_3 \geq 0$, that translates to $w_1 + w_2 \leq 1$. The curved lines are given by the chance constraint involving value at Risk, in fact they are part of a large ellipse since this constraint is quadratic in w_1, w_2. The optimal solution is $w_1 = 0.2854$, $w_2 = 0$, $w_3 = 0.7146$

Notes

The Parthenon Oil Company example is loosely based on an example that appears on the wiki pages at the NEOS site (http://wiki.mcs.anl.gov/NEOS/index.php/Stochastic Programming). The NEOS Server is a project run by the University of Wisconsin - Madison that allows anyone to submit optimization problems to state-of-the-art optimization solvers.

The problem of determining the order amount for a fashion good is a classic operations management problem. It is usually called the newsvendor problem, since it was first formulated in the context of a shop selling newspapers. Newspapers left unsold at the end of the day are returned to the publisher, and a decision is needed on the number of newspapers to be ordered by the newsvendor given an uncertain daily demand. The discussion of this problem is based on the tutorial paper by Shapiro and Philpott.

7.4.1 References

Alexander Shapiro and Andy Philpott, (2007) A Tutorial on Stochastic Programming, Manuscript available at http://www2.isye.gatech.edu/ashapiro/publications.html.

Exercises

1. (Parthenon Oil with risk averse behavior)

Suppose that a utility function $u(x) = \sqrt{x}$ is used by the management of the Parthenon Oil Company. Make the appropriate adjustment to the spreadsheet BRMch7-Parthenon1.xls to allow for this utility function. Note

(a) This utility function is defined on the total profit that Parthenon makes (which needs to be positive for this utility function to make sense) - you should assume that all oil is sold at a price of $200 per barrel.

(b) You should assume that POC is maximizing expected utility so you need to average the utility of the three scenarios. The problem becomes non-linear so you need to select the options in Solver appropriately.

Does the optimal choice of the oil purchase in February (x_1) change in any way? (Since the utility function is undefined when POC makes a loss you will need to ensure that the starting point for the optimization has every scenario profitable)

2**. (Olympic Property)**

Olympic Property is a company that purchases office space to lease. It specializes in medium size office suites in CBD locations in Sydney. The office space purchased this year may be leased to potential customers at the current rate per square metre or it can be held for lease in the future. All leases are for a period of five years. Both the demand for office space next year, the price that it may command and the cost of buying additional office space next year are uncertain. At the end of next year any office space still unleased can be sold. Olympic calculates that the cost of new office space means that properties purchased this year will cost $450 per square metre per year to finance the purchase, this cost will be paid each year for the period of the loan, with interest rates fixed for the first six years of the loan. This year the going rate for Olympic office suites is $500 per square metre per year. Demand this year is estimated to be 15,000 square metres.

Demand for new leases next year will depend on the economic climate. Olympic now needs to determine how much office space to purchase this year given that prices are expected to rise next year. Olympic has developed 3 scenarios for next year as follows

Scenario	A (pessimistic)	B (median)	C (optimistic)
Probability	0.2	0.4	0.4
Demand for new leases	12,000	15,000	18,000
Rental price (per sq m per yr)	520	560	580
Cost of financing new suites (per sq m per year)	500	540	560

(a) Formulate this as an optimization problem. You should ensure that the objective function includes the profit to be made from existing leases held at the end of the second year for the remaining term of those leases. For simplicity you should assume that unleased property at the end of the second year is sold at a price that exactly covers the outstanding loan for that property.

One way to do this is to let $z_i = \max(0, x_1 - d_1 - d_{2i})$ where x_1 is the space purchased in the first year, d_1 is the first year's demand and d_{2i} is the second year's demand under scenario i. So z_i is the property purchased in the first year still unlet at the end of the second year (if any). Explain why under scenario i the total profit is

$$= 5d_1(500-450) + 5(x_1 - d_1 - z_{2i})(r_i - 450) + 5(d_{2i} - x_1 + d_1 + z_{2i})(r_i - c_{2i}) \\ - 450(x_1 - d_1) - 450z_{2i}$$

where r_i is the rental price under scenario i and c_{2i} is the cost of financing purchases in year 2.

(b) Use a spreadsheet model to solve this optimization problem.

3. **(Generating random variables)**
Suppose that demand is stochastic and has a density function as follows

$$f(x) = 0 \text{ for } x \leq 10$$
$$f(x) = (x/2) - 5 \text{ for } x \in (10, 11)$$
$$f(x) = 0.5 \text{ for } x \in [11, 12]$$
$$f(x) = 6.5 - (x/2) \text{ for } x \in (12, 13)$$
$$f(x) = 0 \text{ for } x \geq 13$$

Use the method discussed in this chapter to generate 5 random samples from this distribution using the following random samples from the uniform distribution on $[0, 1]$: 0.543, 0.860, 0.172, 0.476, 0.789.

4. **(Non-anticipatory constraints in GPS example)**
Consider the problem solved in GPS-example3.xls for which $x_{2A} = 50$, $x_{2B} = 45$, and $x_{2C} = 40$. Explain why a non-anticipatory solution to this problem would have $x_{2B} = x_{2C}$.

8

Robust Optimization

Managing risk by gaming

Roger Riley is the CEO of Safety First Corporation (SFC), and certainly takes his company's name to heart. For SFC the key uncertainty relates to demand for their various safety related products and the uncertainty about manufacturing volume that arises because of the stringent checking that takes place across all their product lines. Sometimes they will reject as much as 10% of a week's production on quality grounds. Roger has an unusual approach to risk management working closely with Wendy Morris as his Chief Risk Officer. A big part of Wendy's role is to dream up possible scenarios in terms of demand and manufacturing yield that will cause difficulties. This is not as easy as it sounds because manufacturing decisions can be wrong in both directions: producing too much of a product that doesn't sell well will lead to scrap, and making too little of a product that does sell well will mean rush manufacturing orders using expensive overtime and this can also lead to the comapny losing money.

Roger has always had a pessimistic streak and he sees the risk management process as a game between himself and Wendy. He will come up with a production plan, and then Wendy will play a kind of 'Murphy's law' role to generate a set of demand and yields that are believable, but designed to cause maximum difficulties with the production plan that Roger has chosen. Then Roger and Wendy together use a simple planning spreadsheet to figure out how much money SFC would make (or lose) in this worst case scenario. The next step is for Roger to adjust the manufacturing quantities to try to improve the overall performance of SFC, but each new set of manufacturing decisions is Wendy's cue to redesign the worst case scenario, to try to ensure that SFC does badly. Often Roger and Wendy go through 5 or 6 iterations before Roger decides that he doesn't need to try any more variations.

Before they set out on this process Roger and Wendy have to jointly agree the boundaries within which Wendy can choose the relevant numbers, as well as agreeing the estimates of the costs involved to feed into the planning spreadsheet. Now they have this procedure well-established and both of them enjoy the challenge of playing the game. Wendy says that it appeals to some malicious instinct in her, and Roger is convinced that the production plans he eventually comes up with are robust: "These production plans may seem very conservative, but I know that after Wendy has attempted to blow them up, then the plans are not going to be thrown out by an unexpected set of manufacturing and demand data - and that is worth a lot to me".

Business Risk Management Edward Anderson

8.1 True uncertainty: beyond probabilities

Now it is time to return to a topic that we have touched on earlier. In almost all of our discussions of risk we have taken for granted a notion of probability. The risks we take are associated with the losses we may incur and the probabilities associated with those events. Sometimes we can be confident of the probabilities involved ("What is the probability that an Ace is drawn when we choose at random from a full pack of cards?"). Sometimes the probabilities are deduced from looking at the frequency with which something has happened in the past ("What is the probability that a person selected at random in New York city is left handed"). And sometimes we make a subjective judgement on the basis of our experience perhaps putting together what we know from different spheres ("What is the probability that the price of gold will climb over the next two years?").

One of the great cynics of the twentieth century, Frank Knight, would caution us against coming up with a specific number when looking at the probability of a future event. Knight taught in Chicago from 1928 till he died in 1972 at the age of 87. But the idea that he is most remembered for comes from his Ph.D. dissertation of 1916. Knight argues for the existence of a kind of uncertainty that is not amenable to measurement through probabilities:

> "Uncertainty must be taken in a sense radically distinct from the familiar notion of Risk, from which it has never been properly separated.... The essential fact is that 'risk' means in some cases a quantity susceptible of measurement, while at other times it is something distinctly not of this character; and there are far-reaching and crucial differences in the bearings of the phenomena depending on which of the two is really present and operating"

Lord Kelvin made a famous remark about the importance of measurement claiming that if you cannot measure something then "your knowledge is of a meagre and unsatisfactory kind", Knight, thought that economists and other social scientists had taken Kelvin's statement to mean "If you cannot measure, measure anyhow." He was scathing about those he saw as trying to turn economics into a science like Physics, based on the rational behavior of all the economic actors.

For Knight the whole of life was full of examples of individuals making judgements about future events and often the individual could nominate some degree of confidence in this judgement, and yet to talk of the *probability* of a particular judgement being correct is "meaningless and fatally misleading".

Keynes writing in 1937 also stressed the difference between what can be calculated as a probability, and the uncertainty that prevails over something like the obsolescence of a new invention, "About these matters there is no scientific basis on which to form any calculable probability whatever. We simply do not know. Nevertheless the necessity for action and for decision compels us as practical men to do our best to overlook this awkward fact..."

Faced with the necessity of making decisions when there is uncertainty there are two broad approaches: The first is to push hard for at least a subjective assessment of probabilities even under conditions of Knightian uncertainty where we are naturally uncomfortable to provide these. The idea is that if we will in the end make some decision, then the decision we take will imply some range of values for the missing probabilities. Logically it seems preferable to have our uncertainty translated into a subjective probability of some sort so that it can feed into the decision we need to take, rather than have it emerge as a kind of by-product of the decision we end up making. Looked at from this angle the question becomes one of finding a way to dig down to the underlying and perhaps unconscious beliefs of an individual regarding the probabilities of different events. There has been much work done on the best way to elicit the beliefs of decision makers both on the values of different outcomes and on the probabilities of those outcomes.

There is a second approach that seeks to limit the damage from a bad decision rather than fully optimize some specific objective function. This is called *Robust Optimization* and is the approach we will explore in this chapter. One motivation is that we are always inclined to overestimate our certainty and a robust optimization approach will avoid this being too painful. Bernstein quotes G. K. Chesterton as saying that life... "looks just a little more mathematical and regular than it is; its exactitude is obvious, but its inexactitude is hidden; its wildness lies in wait." With robust optimization we focus on dealing with this 'wildness'.

8.2 Avoiding disaster when there is uncertainty

A robust decision is one which will not turn out to have disastrous results. Some aspects of the problem setup are uncertain and there is at least the possibility of a very bad outcome: the idea is to eliminate or minimize this possibility. A focus on the bad results make it important to know the range of values that some uncertain quantity may take, and these range statements will, in a sense, replace more precise statements about probabilities.

In many cases we are dealing with multiple uncertain variables. So we need to decide whether to specify ranges for each variable independently or whether to look at the combination of values in determining the range.

The first problem we consider is one where the coefficients in the constraints are uncertain. For example suppose that a manufacturing company MRB Ltd needs to meet an order for 10000 units of product A and has two factories that it can use, but there are different costs and efficiencies involved. Factory 1 has higher labour costs of \$30 per hour while factory 2 has labour costs of \$26 per hour. However the machinery in factory 1 is more reliable: in an hour 130 units of product A can be produced in factory 1, but in the same time only 110 units can be produced in factory 2. We can formulate MRB's decision as an optimization problem of minimizing costs subject to meeting the order. If x_1 and x_2 are the hours used in factory 1 and factory 2, respectively, then we want to

$$\begin{array}{ll} \text{minimize} & 30x_1 + 26x_2, \\ \text{subject to} & 130x_1 + 110x_2 \geq 10000, \\ & x_1 \geq 0, x_2 \geq 0. \end{array}$$

But if we have to determine a schedule in advance then it is critical that we are able to meet the order so we may want to be safer. How confident are we that the production rate will be exactly as we have forecasted? Machines can break down, personnel can change and obviously the numbers 130 and 110 may not be exactly right. So we would be better to solve a problem where we ask that a constraint like $(130 - \Delta_1)x_1 + (110 - \Delta_2)x_2 \geq 10000$ is satisfied for some appropriately chosen values of Δ_1 and Δ_2.

We can easily imagine more complicated versions of the same problem. For example suppose that MRB also has to meet an order for 5000 of product B and for this product the production rates in factory 1 and 2 are 90 per hour and 80 per hour respectively. Moreover there is a constraint on the time available with each factory having only 90 hours available prior to the order delivery deadline. Then writing y_1 and y_2 for the hours used on product B in the two factories, the overall problem of minimizing costs becomes

$$\begin{array}{ll} \text{minimize} & 30(x_1 + y_1) + 26(x_2 + y_2) \\ \text{subject to} & (130 - \Delta_1)x_1 + (110 - \Delta_2)x_2 \geq 10000 \\ & (90 - \Delta_3)y_1 + (80 - \Delta_4)y_2 \geq 5000 \\ & x_1 + y_1 \leq 90 \\ & x_2 + y_2 \leq 90 \\ & x_1 \geq 0, x_2 \geq 0, y_1 \geq 0, y_2 \geq 0 \end{array}$$

Now we have four safety factors $\Delta_1, \Delta_2, \Delta_3, \Delta_4$ to choose. It would be easy to drown in the detail of this kind of example and it is helpful to take a step backwards.

A good way to think of this problem is to see it as an optimization problem in which the coefficients appearing in the constraints are uncertain. Our aim is to optimize an objective subject to meeting the constraints for any actual values of the coefficients that may occur. The set of possible coefficient values is clearly critical and for the moment we suppose that we can identify this set. The form of the general (robust linear programming) problem with just two variables and two constraints is

$$\begin{array}{lll} RLP: & \text{maximize} & c_1x_1 + c_2x_2 \\ & \text{subject to} & a_{11}x_1 + a_{12}x_2 \leq b_1 \text{ for all } (a_{11}, a_{12}) \text{ in } A_1 \\ & & a_{21}x_1 + a_{22}x_2 \leq b_2 \text{ for all } (a_{21}, a_{22}) \text{ in } A_2 \\ & & x_1 \geq 0, x_2 \geq 0 \end{array}$$

and this can obviously be extended to any number of variables and constraints. Notice that the choice to make this a maximization problem with '$\leq$' constraints is fairly arbitrary – we can always convert a maximization problem to a minimization one by looking at the negative of the objective, and we can change the inequalities around by multiplying through by -1.

The first thing to notice here is that we can deal with this on a constraint by constraint basis. There might be some complex interaction between the values of a_{11} and a_{12} for the first constraint and the values of a_{21} and a_{22} for the second constraint leading to a combined uncertainty set $(a_{11}, a_{12}, a_{21}, a_{22}) \in A$, but since we need to have the both constraints satisfied for every possible set of parameters these interactions will not make any difference in the end, and so we can split the set A into separate components for each constraint. The decision variables x_1 and x_2 must satisfy the first constraint for any possible values of a_{11} and a_{12} that appear as a pair in A, and must also satisfy the second constraint for any possible values of a_{21} and a_{22}.

The nature of the solution to this problem depends on the structure of the uncertainty sets involved. Consider a single constraint of the form

$$a_1x_1 + a_2x_2 \leq b \text{ for all } (a_1, a_2) \in A.$$

Each element of the set A produces a different constraint and all of them must be satisfied by (x_1, x_2).

The situation is illustrated by Figure 8.1 in which we show the feasible region for the constraints

$$(2 + z_1)x_1 + (3 + z_2)x_2 \leq 3, \text{ for all } (z_1, z_2) \in Z = \{z_1 \geq 0, z_2 \geq 0, z_1 + z_2 \leq 1\}.$$

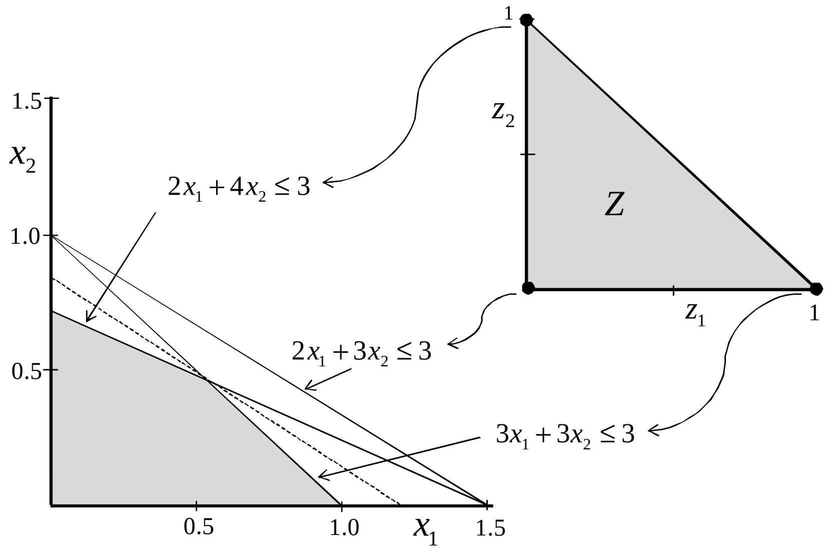


Figure 8.1 The feasible region for the constraint $2x_1 + 3x_2 \le 3$ when the coefficients are subject to the perturbations in Z

The set Z here is the set of deviations possible to the base values $a_1 = 2$ and $a_2 = 3$ in order to reach the set of allowable coefficients A.

The figure shows that the overall feasible set is obtained from looking at the constraints generated by the two corner points in Z, i.e the points where $z_1 = 0, z_2 = 1$ and where $z_1 = 1, z_2 = 0$. The third corner point at $z_1 = z_2 = 0$ gives the base constraint which is higher than the others and does not contribute to defining the feasible region. All the other points in Z generate constraints which will be satisfied within the feasible (shaded) region. The dashed line is an example of a constraint generated by one such point in Z. The fact that this goes through the point $x_1 = 0.5$, $x_2 = 0.5$ in fact indicates that it comes from a point somewhere along the top boundary of Z where $z_1 + z_2 = 1$.

The corners of the feasible region tell us all we need to know, and this property actually holds more generally. If the uncertainty set A associated with a particular constraint is a polytope with a set of k corners, then we can replace the single constraint with k copies defined by the corner points of A. This will be exactly the same as asking for the constraint to hold at all points in A. We can see this by noting that all of the points in A can be obtained as (convex) combinations of the corner points. In other words if $(a_1^{(j)}, a_2^{(j)}, ...a_m^{(j)})$ is the j'th corner point of A where $j = 1, 2, ...k$, then any point $(a_1, a_2, ...a_m) \in A$ can be obtained from some set of non-negative multipliers λ_1, λ_2, ..., λ_k with $\sum_{j=1}^{k} \lambda_j = 1$ applied to the corners, and $a_i = \sum_{j=1}^{k} \lambda_j a_i^{(j)}$ for $i = 1, 2, ..., m$.

Now suppose that a point $(x_1, x_2, ...x_m)$ satisfies each of the constraints at the corners of A, so

$$a_1^{(j)} x_1 + a_2^{(j)} x_2 + ...a_m^{(j)} x_m \le b \text{ for } j = 1, 2, ..., k.$$

Then

$$\begin{aligned}
&a_1x_1 + a_2x_2 + ...a_mx_m \\
&= \sum_{j=1}^{k} \lambda_j a_1^{(j)} x_1 + \sum_{j=1}^{k} \lambda_j a_2^{(j)} x_2 + ... \sum_{j=1}^{k} \lambda_j a_m^{(j)} x_m \\
&= \sum_{j=1}^{k} \lambda_j \left(a_1^{(j)} x_1 + a_2^{(j)} x_2 + ...a_m^{(j)} x_m \right) \\
&\leq \sum_{j=1}^{k} \lambda_j b = b,
\end{aligned}$$

and so the point $(x_1, x_2, ...x_m)$ also satisfies the arbitrary constraint $(a_1, a_2, ...a_m)$ picked from somewhere inside A.

Thus we can take a general problem like RLP above and simply replace the first constraint by a set of copies derived from the corner points of A_1, and similarly replace the second constraint by a set of copies derived from the corner points of A_2, and so on. This increases the size of the problem, but it retains the same structure and in the case of a linear program it is still easy to solve.

8.2.1 Using information on the uncertainty set

We have established the principle that linear programs with polyhedral uncertainty sets for the coefficients remain as linear programs. But to make this a more useful approach in practice it is convenient to work more directly with the constraints that define the polyhedral set A rather than with the corners (or "extreme" points) of A. When the dimension of the set A increases the number of corner points can get quite large even for simple constraints.

To do this we need to take a short mathematical detour into the duality theory that is associated with linear programs. Perhaps you might want to just take on trust the final result we are leading up to, but the properties of the dual linear program are both surprising and beautiful, and there is no harm in spending a little while looking at this area. The duality result we need is quite general, but to make it easier to read we will describe the result for a problem with just two variables x_1, x_2 and three constraints. The duality theorem for linear programs states that the value of the (primal) linear program given by

$$\begin{aligned}
LP: \quad & \text{maximize} && g_1x_1 + g_2x_2 \\
& \text{subject to} && d_{11}x_1 + d_{12}x_2 \leq h_1, \\
& && d_{21}x_1 + d_{22}x_2 \leq h_2, \\
& && d_{31}x_1 + d_{32}x_2 \leq h_3, \\
& && x_1 \geq 0, x_2 \geq 0,
\end{aligned}$$

is the same as the value of the (dual) linear program

$$\begin{aligned}
DLP: \quad & \text{minimize} && h_1y_1 + h_2y_2 + h_3y_3 \\
& \text{subject to} && d_{11}y_1 + d_{21}y_2 + d_{31}y_3 \geq g_1, \\
& && d_{12}y_1 + d_{22}y_2 + d_{32}y_3 \geq g_2, \\
& && y_1 \geq 0, y_2 \geq 0, y_3 \geq 0.
\end{aligned}$$

If you have never seen this before you need to stop and look carefully to see what has happened in moving from one problem to the other. The dual linear program has a constraint for each variable in the original (primal) LP, and it has a variable for each constraint in the primal problem. Also the coefficients in the objective function get translated into the constraint right hand sides and vice versa. Not only have the variables and constraints swapped places, but we have changed a maximization problem with "$\leq$" constraints into a minimization problem with "$\geq$" constraints. In both problems all the variables are constrained to be positive. In fact the duality relation still works if a variable is not constrained to be positive (a "free" variable), in this case the corresponding constraint has to be an equality. Thus for example if the primal problem did not have the constraint $x_2 \geq 0$, then the second constraint in the dual would become

$$d_{12}y_1 + d_{22}y_2 + d_{32}y_3 = g_2.$$

Notice what we are saying here: the minimum value of the objective function in the dual DLP is equal to the maximum value of the objective in the primal LP. It is interesting to try and see why these two problems have the same value. You can try for example putting actual numbers in instead of all the algebra to check that the result really does hold. But be warned that the reason for the duality result being true is quite deep (it comes down to a 'separating' hyperplane argument). It is easy enough to show that the minimum in DLP is greater than the maximum in LP, but to show that these values are the same is quite a bit harder.

There are other forms of dual that can be written down, but this is the form that is easiest to remember and what we have said here will be enough for the result we want to derive. What is the connection with our robust optimization problem?

Suppose that in our problem RLP the uncertainty set A_1 is a polytope that is defined by saying that the values a_{11} and a_{12} satisfy a set of constraints

$$\begin{aligned}
&d_{11}a_{11} + d_{12}a_{12} \leq h_1,\\
&d_{21}a_{11} + d_{22}a_{12} \leq h_2,\\
&d_{31}a_{11} + d_{32}a_{12} \leq h_3,\\
&a_{11} \geq 0, a_{12} \geq 0.
\end{aligned}$$

Then we can rewrite the first constraint of RLP that $a_{11}x_1 + a_{12}x_2 \leq b_1$ for all (a_{11}, a_{12}) in A_1 as saying that the maximum value that $a_{11}x_1 + a_{12}x_2$ can take for $(a_{11}, a_{12}) \in A_1$ is less than b_1. And then this can be expressed by saying that the solution of the linear program

$$\begin{array}{ll}
\text{maximize} & a_{11}x_1 + a_{12}x_2\\
\text{subject to} & d_{11}a_{11} + d_{12}a_{12} \leq h_1,\\
& d_{21}a_{11} + d_{22}a_{12} \leq h_2,\\
& d_{31}a_{11} + d_{32}a_{12} \leq h_3,\\
& a_{11} \geq 0, a_{12} \geq 0.
\end{array}$$

should be less than or equal to b_1. From our duality result this is exactly the same as saying that the solution of the dual linear program

$$\begin{array}{lll}
DLP1: & \text{minimize} & h_1y_1 + h_2y_2 + h_3y_3\\
& \text{subject to} & d_{11}y_1 + d_{21}y_2 + d_{31}y_3 \geq x_1,\\
& & d_{12}y_1 + d_{22}y_2 + d_{32}y_3 \geq x_2,\\
& & y_1 \geq 0, y_2 \geq 0, y_3 \geq 0.
\end{array}$$

should be less than or equal to b_1. Think about this statement carefully and you can see that it amounts to saying that a pair of values x_1 and x_2 will satisfy the first constraint of RLP with its uncertainty set A_1 if and only if there are some values $y_1 \geq 0, y_2 \geq 0, y_3 \geq 0$ with

$$\begin{aligned} h_1y_1 + h_2y_2 + h_3y_3 &\leq b_1, \\ d_{11}y_1 + d_{21}y_2 + d_{31}y_3 &\geq x_1, \\ d_{12}y_1 + d_{22}y_2 + d_{32}y_3 &\geq x_2. \end{aligned}$$

Thus we can take our original problem and convert the constraint with an uncertainty set into three constraints if we also add three new variables. In general if the uncertainty set for a constraint is defined by m constraints on the coefficients and there are n coefficients that are uncertain (going with the n original variables) then we will need m new variables and $n+1$ constraints to represent the original constraint with its uncertainty set.

Later we will give an example, but first it will be helpful to say more about how we can construct an uncertainty set.

8.2.2 Uncertainty set with a budget of uncertainty

It is frequently possible to give a range of possible values for an uncertain parameter. So for a parameter a_1 it often happens that we do not know its exact value but we are confident that it will lie in a certain range. It is convenient to take the midpoint of the range as a kind of base value $\overline{a}_1$ and then define δ_1 as the distance from $\overline{a}_1$ to the two bounds. Hence the uncertainty set for a_1 is given by $\overline{a}_1 - \delta_1 \leq a_1 \leq \overline{a}_1 + \delta_1$.

When a constraint contains a number of different uncertain parameters $a_1, a_2, ...a_n$ say, then this will determine a combined uncertainty set A by simply asking for each a_i to satisfy a range constraint $\overline{a}_i - \delta_i \leq a_i \leq \overline{a}_i + \delta_i$. With this arrangement the set A becomes an n-dimensional rectangular shape centred on $(\overline{a}_1, \overline{a}_2, ..., \overline{a}_n)$.

But this is in practice an extremely conservative uncertainty set, since we allow all the uncertain parameters to take their extreme values at the same time. Unless they are highly correlated variables it makes sense to be more conservative in the choice of the individual interval lengths δ_i and compensate for this by being less conservative in the points where a lot of parameters are close to their extreme values at the same time.

This leads to defining a ‘budget of uncertainty’ B. If there are 10 variables and we have a budget of uncertainty of 5 this would mean that the sum of the ratios $\left|a_i - \overline{a}_i\right|/\delta_i$ is less than 5. This might be achieved, for example, by having 5 variables at $\overline{a}_i + \delta_i$ and the other 5 at $\overline{a}_i$, or by having 5 variables at $\overline{a}_1 + \delta_1/2$ and 5 variables at $\overline{a}_1 - \delta_1/2$. To see what this looks like for a specific example Figure 8.2 shows the situation when there are three uncertain parameters and compares what happens when $B = 2.5$ and $B = 1.5$. The central point in this diagram is given by the vector of base values $(\overline{a}_1, \overline{a}_2, \overline{a}_3)$.

Next we investigate the nature of the adjusted linear program that we reach when there is an uncertainty set of this form. We start by looking at a simple case where there is a constraint $a_1x_1 + a_2x_2 \leq b$ for all $(a_1, a_2) \in A$ where A is defined as the a_1, a_2 satisfying

$$a_1 = \overline{a}_1 + z_1\delta_1, a_2 = \overline{a}_2 + z_2\delta_2,$$

$$\text{with } |z_1| \leq 1,\ |z_2| \leq 1, \text{ and } |z_1| + |z_2| \leq B$$

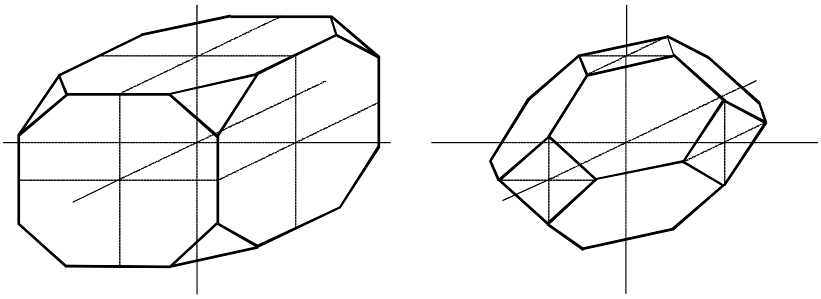


Figure 8.2 There are three variables: the left hand diagram shows a budget of uncertainty of 2.5, and the right hand diagram shows a budget of uncertainty of 1.5.

If the original constraint holds for all $(a_1, a_2) \in A$ then we can reformulate the constraint by saying that the following linear program with decision variables a_1 and a_2 has a value no greater than b:

$$\begin{array}{ll} \text{minimize} & a_1x_1 + a_2x_2 \\ \text{subject to} & a_i - \delta_i u_i + \delta_i v_i = \overline{a}_i,\ i = 1,2, \\ & u_i + v_i \leq 1,\ i = 1,2, \\ & (u_1 + v_1) + (u_2 + v_2) \leq B, \\ & u_1 \geq 0, u_2 \geq 0, v_1 \geq 0, v_2 \geq 0. \end{array}$$

Here we have written u_1 and v_1 for the positive and negative parts of z_1 (i.e. $u_1 = \max(z_1, 0)$ and $v_1 = \max(-z_1, 0)$). This means that $z_1 = u_1 - v_1$ and $|z_1| = u_1 + v_1$. There is a trick here since defining the variables in this way means that only one of them is non-zero, whereas the linear program as formulated could have both $u_1 > 0$ and $v_1 > 0$. However any solution in which both variables are non-zero can be replaced by one in which the same quantity is subtracted from both u_1 and v_1 to make the smaller of the two equal to zero. The equality constraint will still be satisfied and the inequality constraints also still work since $u_1 + v_1$ is reduced. Similarly u_2 and v_2 are the positive and negative parts of z_2.

Thus using duality we can show that the following linear program has a value no greater than b:

$$\begin{array}{ll} \text{minimize} & \overline{a}_1y_1 + \overline{a}_2y_2 + w_1 + w_2 + Bt \\ \text{subject to} & y_1 = x_1, \\ & y_2 = x_2, \\ & -\delta_1 y_1 + w_1 + t \geq 0, \\ & -\delta_2 y_2 + w_2 + t \geq 0, \\ & \delta_1 y_1 + w_1 + t \geq 0, \\ & \delta_2 y_2 + w_2 + t \geq 0, \\ & w_i \geq 0, t \geq 0. \end{array}$$

We can use the first two constraints here to substitute x_1 and x_2 for y_1 and y_2. We can also combine the two constraints involving w_1: if $w_1 + t$ is greater than both $\delta_1 x_1$ and $-\delta_1 x_1$ then we have $w_1 + t \geq \delta_1 |x_1|$. Thus we reach the following optimization problem that has a value $\leq b$.

$$\begin{array}{ll}\text{minimize} & \overline{a}_1x_1+\overline{a}_2x_2+w_1+w_2+Bt \\ \text{subject to} & w_1+t\geq\delta_1\left|x_1\right|, \\ & w_2+t\geq\delta_2\left|x_2\right|, \\ & w_i\geq 0, t\geq 0.\end{array}$$

With the formulation in this form it is easier to see how a general problem can be formulated. Suppose that there are n decision variables $x_1, x_2, ...x_n$ and the original problem is to maximize $c_1x_1+...+c_nx_n$ subject to some constraints one of which has the form $a_1x_1+a_2x_2+...+a_nx_n\leq b$ for all coefficients $(a_1, a_2, ...a_n)\in A$ where

$$A=\{(\overline{a}_1+z_1\delta_1, \overline{a}_2+z_2\delta_2, ..., \overline{a}_n+z_n\delta_n)\}$$
$$\text{for } |z_i|\leq 1, i=1,2,...,n \text{ and } |z_1|+|z_2|+...|z_n|\leq B.$$

Then $n+1$ new variables $t, w_1, w_2, ..., w_n$ are added to the problem, each new variable being constrained to be non-negative and the constraint in question can be replaced by $n+1$ new constraints

$$\overline{a}_1x_1+\overline{a}_2x_2+...+\overline{a}_nx_n+w_1+w_2+...+w_n+Bt\leq b, \quad (8.1)$$

$$w_i+t\geq\delta_i\left|x_i\right|, i=1,2,...,n. \quad (8.2)$$

To get back to a linear program we simply replace each of the constraints $w_i+t\geq\delta_i\left|x_i\right|$ with two constraints $w_i+t\geq\delta_ix_i$ and $w_i+t\geq-\delta_ix_i$.

It is easy to check that if there are variables x_i, w_i and t satisfying these conditions then the original inequality with b is satisfied for all $(a_1, a_2, ...a_n)\in A$. Since then

$$\begin{aligned}& a_1x_1+a_2x_2+...+a_nx_n \\ &=\overline{a}_1x_1+\overline{a}_2x_2+...+\overline{a}_nx_n+z_1\delta_1x_1+...+z_n\delta_nx_n \\ &\leq\overline{a}_1x_1+\overline{a}_2x_2+...+\overline{a}_nx_n+\left|z_1\right|(w_1+t)+...+\left|z_n\right|(w_n+t) \\ &\leq\overline{a}_1x_1+\overline{a}_2x_2+...+\overline{a}_nx_n+w_1+...+w_n+Bt \\ &\leq b.\end{aligned}$$

Here we used inequalities like $z_1\delta_1x_1\leq\left|z_1\right|\delta_1\left|x_1\right|$ and also made use of the fact that $\left|z_i\right|\leq 1$ and $\sum\left|z_i\right|\leq B$. This is one direction of the duality argument, but it is much harder to go the other way around and show that the new set of constraints are no more restrictive than the original set.

Worked example 8.1

Avignon Imports has to determine the order to place for products A, B and C. The entire order for the three products will require delivery together and transport constraints imply that the total weight of the shipment is less than 5000 kg. Product A, B and C all weigh 5 kg per unit but there is uncertainty about the way that the products will be packed and hence the weight of packaging that the suppliers will use. For A and C this is estimated at 0.2 kg per unit but this is a guess and it is thought that figures between 0.1 kg and 0.3 kg are possible. Product B is more complicated and the packaging is estimated at 0.5 kg per unit, with figures between 0.2 kg and 0.8 kg being possible. All of the items are supplied at a cost of $100 per unit. After importing them Avignon Imports will auction the products. The expected price to be achieved by selling Product A is $200 per unit with a possible variation up and down of $50. The expected price for product B is $205 with a possible variation up or down of $60 per item and the expected price for product C is $195 with a possible variation of $70. There is a requirement that the company make a profit of at least $50,000 from the transaction. Avignon Imports wishes to maximize its expected profit subject to the constraints on transport weight and minimum profit achieved. Formulate this as a robust optimization problem using a budget of uncertainty of $B = 2$ for both the constraints and solve the problem in a spreadsheet.

Solution

Let x_A, x_B, and x_C be the amounts ordered for the three products. Write a_A, a_B, and a_C for the weight per unit and s_A, s_B, and s_C for the sale price per unit. The expected profit is $100, $105 and $80 for the three products and so we have a robust optimization problem of

$$\begin{array}{ll} \text{maximize} & 100x_A + 105x_B + 95x_C \\ \text{subject to} & a_Ax_A + a_Bx_B + a_Cx_C \leq 5000 \text{ for } (a_A, a_B, a_C) \in A, \\ & -s_Ax_A - s_Bx_B - s_Cx_C \leq -50000 \text{ for } (s_A, s_B, s_C) \in S, \\ & x_A \geq 0, x_B \geq 0, x_C \geq 0, \end{array}$$

where the uncertainty sets are

$$A = \{(a_A, a_B, a_C) : a_A = 5.2 + 0.1z_A, a_B = 5.5 + 0.3z_B, a_C = 5.2 + 0.1z_C \\ \text{for } (z_A, z_B, z_C) \in Z\},$$

$$S = \{(s_A, s_B, s_C) : s_A = 100 + 50q_A, s_B = 105 + 60q_B, s_C = 95 + 70q_C \\ \text{for } (q_A, q_B, q_C) \in Z\},$$

where

$$Z = \{(z_A, z_B, z_C) : |z_A| \leq 1, |z_B| \leq 1, |z_C| \leq 1, |z_A| + |z_B| + |z_C| \leq 2\}.$$

Notice that the S constraint on minimum profit has been multiplied by -1 to bring it into standard form. The fact that both budgets of uncertainty are the same means we can use a single set Z for the two different constraints. Now we can use the rules we developed earlier to add constraints as in (8.1) and (8.2); the resulting formulation has each of the existing constraints replaced by 4 new ones (together with four new variables). This gives

$$\begin{array}{ll} \text{maximize} & 100x_A + 105x_B + 95x_C \\ \text{subject to} & 5.2x_A + 5.5x_B + 5.2x_C + w_A + w_B + w_C + 2t_1 \leq 5000, \\ & w_A + t_1 \geq 0.1x_A, \\ & w_B + t_1 \geq 0.3x_B, \\ & w_C + t_1 \geq 0.1x_C, \\ & -100x_A - 105x_B - 95x_C + u_A + u_B + u_C + 2t_2 \leq -50000, \\ & u_A + t_2 \geq 50x_A, \\ & u_B + t_2 \geq 60x_B, \\ & u_C + t_2 \geq 70x_C, \\ & \text{and all variables non-negative.} \end{array}$$

In this formulation we have been able to change $|x_A|$, $|x_B|$ and $|x_C|$ into x_A, x_B and x_C since these are all positive.

The solution to this problem is given in the spreadsheet BRMch8-Avignon.xlsx. We obtain

$$x_A = 785.38, x_B = 81.65, x_C = 69.98,$$

$$w_A = 54.04, t_1 = 24.49, u_A = 34370.14, t_2 = 4898.91,$$

and other variables are zero. In practice we would need to round the variables to whole numbers (or, better, use an optimization procedure that searches amongst integer solutions).

8.2.3 *Analyzing the safety margin*

We can work out how likely we are to exceed the budget of uncertainty when each uncertain parameter is symmetric and independent. However rather than simply asking about the probability that the budget of uncertainty is exceeded we are more interested in the probability that the optimal solution that we reach fails to satisfy the constraints of the problem. To be more precise we suppose that the value of B is used to define an uncertainty set A and then we want to know what is the probability that the optimal solution using this uncertainty set will turn out to be infeasible, assuming that the original ranges are accurate. It is possible that the actual values of the uncertain parameters lie in those parts of the n-dimensional cube cut off by the budget of uncertainty constraint, and this is a question about how likely it is that this will lead to infeasibility.

Suppose that x^* is an optimal solution to the problem with the budget of uncertainty in place. We know that the solution x^* satisfies the constraint with any choice of z_i satisfying the budget of uncertainty and we will make a specific choice for z.

We do this by first reordering the variables so that the highest values of $\delta_i \left|x_i^*\right|$ come first and then we choose $z_i = 1$ or -1 according to the sign of x_i^* for the first $L = \lfloor B \rfloor$ of these variables and $z_i = B - L$ or $-B + L$ for the next (again in order to match the sign of x_{L+1}^*). All other values of z_i are set to zero. With this choice of z_i we will have $\sum_{i=1}^n |z_i| = B$ and the coefficients $a_1, a_2, ..., a_n$ will lie in the defined uncertainty set A. Hence the constraint will be satisfied and we can deduce

$$\sum \overline{a}_i x_i^* + \sum_{i=1}^{L} \delta_i \left|x_i^*\right| + (B - L)\delta_{L+1} \left|x_{L+1}^*\right| \leq b. \tag{8.3}$$

Now suppose that the constraint does not hold at x^* for some set of a_i values in the ranges given. Thus there is a set of z_i values for which

$$\sum \overline{a}_i x_i^* + \sum z_i \delta_i \left|x_i^*\right| > b,$$

and hence, from (8.3),

$$\sum \overline{a}_i x_i^* + \sum z_i \delta_i \left|x_i^*\right| > \sum \overline{a}_i x_i^* + \sum_{i=1}^{L} \delta_i \left|x_i^*\right| + (B - L)\delta_{L+1} \left|x_{L+1}^*\right|.$$

This inequality can be rewritten

$$\sum_{i=L+1}^{n} z_i \delta_i \left|x_i^*\right| > \sum_{i=1}^{L} (1 - z_i)\delta_i \left|x_i^*\right| + (B - L)\delta_{L+1} \left|x_{L+1}^*\right|.$$

Because of the ordering of the $\delta_i |x_i^*|$ (and using the fact that $1 - z_i \geq 0$) this implies

$$\sum_{i=L+1}^{n} z_i \delta_i |x_i^*| \geq \delta_{L+1} \left| x_{L+1}^* \right| \left(\sum_{i=1}^{L} (1 - z_i) + (B - L) \right).$$

So

$$\sum_{i=1}^{L} z_i + \sum_{i=L+1}^{n} z_i \frac{\delta_i |x_i^*|}{\delta_{L+1} \left| x_{L+1}^* \right|} > B.$$

This inequality has the form

$$\sum_{i=1}^{n} z_i h_i > B \tag{8.4}$$

with $0 \leq h_i \leq 1$ for all i.

Now we ask what is the probability that the z_i values make the constraint not satisfied (if each z_i is chosen in a way that is independent and symmetric around 0)? Since the inequality (8.4) is satisfied if the constraint is broken the probability of this inequality holding must be greater than the probability that the constraint is broken.

We can use the central limit theorem to produce a bound on this probability for large n. The random variable $\sum_{i=1}^{n} z_i h_i$ has mean zero (since each z_i has zero mean) and variance $V = \sum_{i=1}^{n} h_i^2 V_i$ where V_i is the variance of the variable z_i. Since z_i lies in the range -1 to 1 it cannot have variance larger than 1 and h_i^2 is also less than 1. Hence the variance of $V \leq n$. Finally we have

$$\begin{aligned} \Pr\left(\sum_{i=1}^{n} z_i h_i > B \right) &\approx \Pr(N(0, \sqrt{V}) > B) \\ &\leq 1 - \Phi\left(\frac{B}{\sqrt{n}} \right). \end{aligned}$$

Thus we have established that for large n and any symmetric distribution of coefficient errors around the base levels, provided these errors are independent, using a budget of uncertainty B in solving the optimization problem will give a probability of the constraint being broken at the optimal solution of no more than $1 - \Phi\left(B/\sqrt{n}\right)$.

8.3 Robust optimization and the minimax approach

An important type of uncertainty relates to the objective function in the minimization. In this context the classical stochastic optimization would look at the expected value of the objective under some model describing the probabilities of different parameters in the objective function. But we are interested in an environment in which there is no known distribution for these parameters, at the most we simply have a range of possible values. The objective function parameters belong to an uncertainty set A.

In this context it is natural to consider a 'minimax' approach which assumes the worst and makes decisions under which the 'worst' will not be too bad. This approach is the same as that which arises from rewriting the problem with an extra variable representing the objective function. Thus a standard optimization problem may be written

Maximize $f(x)$ subject to $x \in X$

where f is the objective function and X is the feasible set defined by the constraints. This standard problem can be rewritten adding an unconstrained (scalar) variable v as:

$$\begin{array}{ll} \text{Maximize} & v \\ \text{subject to} & v \leq f(x), \\ & x \in X. \end{array}$$

Then an uncertainty in the objective function is translated into an uncertainty in the constraints as was dealt with in the previous section.

In the case that the objective function is linear we have $f(x) = c_1x_1 + c_2x_2 + ... + c_nx_n$ and we know that $(c_1, c_2, ..., c_n)$ lies in a given uncertainty set A. So the problem becomes

$$\begin{array}{lll} \text{PZ:} & \text{Maximize} & v \\ & \text{subject to} & v - c_1x_1 - c_2x_2 - ... - c_nx_n \leq 0 \text{ for all } (c_1, c_2, ..., c_n) \in A, \\ & & x \in X. \end{array}$$

This formulation allows all the machinery introduced in the previous section to be applied. Note that by asking for the constraint to apply for all choices of coefficient in the set A, we end up with a value v that is equal to the smallest value of the objective for possible $(c_1, c_2, ..., c_n) \in A$. Thus the formulation PZ is equivalent to

$$\begin{array}{ll} \text{Maximize} & \min_{(c_1,c_2,...,c_n)\in A} c_1x_1 + c_2x_2 + ... + c_nx_n \\ \text{subject to} & x \in X. \end{array}$$

We can see that this with this formulation we are maximizing the objective subject to the most pessimistic assumptions on the values of the uncertain parameters. We can think of this as a game between us and an opponent, just the sort of game that we saw Roger Riley playing in the scenario at the start of this chapter. We choose the values of the decision variables $x_1, x_2, ..., x_n$ and then the other player chooses the values of $c_1, c_2, ..., c_n$. But our opponent here is simply malicious: their aim is to give us the worst possible outcome. Sometimes people talk of playing against 'nature' though this implies a rather paranoid view of the world! We will simply regard this problem as one of guaranteeing a reasonable outcome whatever nature throws at us.

More generally we can think of a profit function Π that depends not only on our actions x but also the values of some uncertain parameters given by the vector a and the only information we have is that $a \in A$, the 'uncertainty set' for the problem. Then the best we can do in guaranteeing a certain level of profit is to maximize the minimum value of the profit for $a \in A$, i.e. we solve

$$\max_{x\in X} \left\{ \min_{a\in A} \Pi(x, a) \right\}. \tag{8.5}$$

An important case of this problem is when the profit function Π is a concave function of a for each value of x and A is a polytope with corners $a^{(1)}, a^{(2)}, ...a^{(k)}$. Each of these corners is itself a vector so we have $(a_1^{(j)}, a_2^{(j)}, ...a_m^{(j)})$ being the j'th corner point of A where $j = 1, 2, ...k$.

In this case the equivalent to the formulation PZ becomes

$$\begin{array}{lll} \text{PZ1:} & \text{Maximize} & v \\ & \text{subject to} & v - \Pi(x, a) \leq 0 \text{ for all } (a_1, a_2, ...a_m) \in A, \\ & & x \in X. \end{array}$$

Using the same approach we used before, the next step is to replace the single constraint with k copies: one for each of the extreme points of A. We get the following

$$\begin{array}{lll} \text{PZ2:} & \text{Maximize} & v \\ & \text{subject to} & v - \Pi(x, a^{(1)}) \leq 0, \\ & & v - \Pi(x, a^{(2)}) \leq 0, \\ & & ... \\ & & v - \Pi(x, a^{(k)}) \leq 0, \\ & & x \in X. \end{array}$$

Clearly if v and x are feasible for PZ1 they must satisfy the constraints for all the corner points of A and hence are feasible for PZ2. We can use our assumption on the concavity of Π to show that if each of the constraints of PZ2 are satisfied then so will the constraint of PZ1. As we discussed earlier any point in $(a_1, a_2, ...a_m) \in A$ can be obtained from some set of non-negative multipliers λ_1, λ_2, ..., λ_k with $\sum_{j=1}^{k} \lambda_j = 1$ and $a_i = \sum_{j=1}^{k} \lambda_j a_i^{(j)}$ for $i = 1, 2, ..., m$. In other words the set A can be obtained from the convex combinations of its extreme points $a^{(1)}, a^{(2)}, ...a^{(k)}$. Now using the concavity of $\Pi(x, a)$ as a function of a for fixed x we can deduce

$$\begin{aligned} v - \Pi(x, a) &= v - \Pi(x, \sum_{j=1}^{k} \lambda_j a^{(j)}) \\ &\leq \sum_{j=1}^{k} \lambda_j v - \sum_{j=1}^{k} \lambda_j \Pi(x, a^{(j)}) \\ &= \sum_{j=1}^{k} \lambda_j \left(v - \Pi(x, a^{(j)}) \right) \\ &\leq 0. \end{aligned}$$

So the end result is that when the profit function is concave in the uncertain parameter and the uncertainty set is a polytope we can replace the problem (8.5) with the optimization problem PZ2.

Worked example 8.2

To see how this works on an example consider Sentinel Enterprises who sell tablet computers and e-readers. They have a new product launch of the ‘FlexReader’ for which there has been considerable advertising. They have advance orders for 5000 FlexReaders. The advance order customers have been given a secure code which they can use to make an online purchase two weeks before the FlexReaders are available at retail stores. FlexReaders come in two screen sizes (large and small) - and due to a mistake the advance order customers were not asked which of these they wanted. To make matters worse the manufacturers have been experiencing problems with meeting the launch date and the result will be an extra cost for FlexReaders available for advance purchase. Sentinel will pay $550 for the large screen format and $520 for the small screen format for FlexReaders available in time for advance purchase, while FlexReaders delivered two weeks later will cost $70 less. The large FlexReaders sell for $640, the small ones for $590. Customers who have placed an advance order but cannot get their preferred format are likely not to purchase at all. FlexReaders that Sentinel gets delivered early that are not needed for the advance order customers will simply be sold later. How many of the two different readers should Sentinel order?

Solution

In this problem if there was a known distribution for the preferences of customers between large and small formats and a known distribution of launch demand then it would be possible to solve the problem using the techniques of stochastic optimization we have discussed earlier. In the absence of this information then a robust optimization approach could be used. This will have the effect of maximizing the worst case profits.

We want to determine the order size x_L for large FlexReaders and x_S for small FlexReaders to be delivered at launch date. If the total advance orders are split as d_L of large and d_S of small, then the advance order sales are $z_L = \min(x_L, d_L)$ and $z_S = \min(x_S, d_S)$ with a profit made on these sales of $90z_L + 70z_S$. However there is an extra cost of \$70 for any FlexReaders purchased early that are not needed for the advance purchase customers. Thus the profit term we want to maximize is given by

$$\begin{aligned}&90z_L + 70z_S - 70(x_L - z_L + x_S - z_S)\\ &= 160z_L + 140z_S - 70(x_L + x_S).\end{aligned}$$

Hence to maximize the profit we want to choose x_L and x_S to

$$\text{maximize } 160\min(x_L, d_L) + 140\min(x_S, d_S) - 70(x_L + x_S),$$

but nature will choose d_L and d_S from the uncertainty set

$$A = \{(d_L, d_S) : d_L \geq 0, d_S \geq 0, d_L + d_S = 5000\}.$$

We can formulate this as the optimization problem

$$\begin{array}{lll}\text{PS:} & \text{Maximize} & v\\ & \text{subject to} & v - 160\min(x_L, d_L) - 140\min(x_S, d_S) + 70(x_L + x_S) \leq 0 \text{ for all } (d_L, d_S) \in A,\\ & & x_L \geq 0, x_S \geq 0.\end{array}$$

Since the minimum operators in the objective function are concave this satisfies the conditions we need to replace the uncertainty set with copies of the constraint at the two extreme points of A; these are $d_L = 5000, d_S = 0$ and $d_L = 0, d_S = 5000$.

We can assume that $x_L \leq 5000, x_S \leq 5000$ since there can be no reason to order more than the maximum demand. Then the constraints at the two extreme points can be simplified and we obtain

$$\begin{array}{ll}\text{Maximize} & v\\ \text{subject to} & v - 160x_L + 70(x_L + x_S) \leq 0,\\ & v - 140x_S + 70(x_L + x_S) \leq 0,\\ & 0 \leq x_L \leq 5000, 0 \leq x_S \leq 5000.\end{array}$$

The spreadsheet BRMch8-Sentinel.xlsx is set up for the solution of this problem The solution turns out to be $x_L = 4375$ and $x_S = 5000$ which gives a v value of \$43750. In other words with these values of x_L and x_S a profit of at least \$43750 will be made and this is the best ‘guaranteed’ profit there can be.

Distributionally Robust Optimization

Up to this point we have been considering a situation where the uncertainty relates to particular numbers within the problem statement. Rather than assume that we know the distribution of those parameters we have assumed simply that we know an uncertainty set to which they belong. There are many cases, however, in which we know more than the range of values that a parameter may take but less than its complete distribution. When this happens it makes sense to use a *distributionally robust* model in which we specify, not a set of points, but a set of distributions as the uncertainty set. An example occurs if we are confident that the for a specific commodity tomorrow's price follows a normal distribution with a mean the same as today's price, but we are uncertain about the standard deviation.

In this case we will write the uncertainty set using a script 'A' ($\mathcal{A}$) to remind us that it is a set of distributions. If ξ is the parameter in question and $\Pi(x, \xi)$ is the profit using decision variables ξ and we knew the distribution of ξ then (if we are risk neutral) we would want to maximize the expectation of $\Pi(x, \xi)$. Since we will be considering changes of distribution it is helpful to write this expectation in terms of the distribution. We use the notation $E_F[\Pi(x, \xi)]$ to mean the expectation of $\Pi(x, \xi)$ when ξ has the distribution F.

We will concentrate on the problem where the uncertainty occurs in the objective rather than in the constraints. Then the distributionally robust optimization problem is to find the best expected profit that is guaranteed if we only know that the distribution of ξ lies in an uncertainty set $\mathcal{A}$. We can write this as

$$\max_x \left\{ \min_{F \in \mathcal{A}} E_F[\Pi(x, \xi)] \right\}.$$

If the uncertainty in the distribution can be represented by a small set of defined parameters then this problem can be brought back to a 'point based' robust formulation, by working out the value of the expectation in terms of the parameter values. In the example about tomorrow's commodity price we might bound the possible standard deviations between $\sigma_{\min}$ and $\sigma_{\max}$ and then $\mathcal{A} = \{N(0, \sigma) : \sigma_{\min} \leq \sigma \leq \sigma_{\max}\}$. Now suppose that we can calculate the expected profit achieved for a given value of standard deviation σ, and decision variables x, say this is $\overline{\Pi}(x, \sigma)$. Then we can rewrite the distributionally robust optimization problem in the form

$$\max_x \left\{ \min_{\sigma \in A} \overline{\Pi}(x, \sigma) \right\},$$

where $A = \{\sigma : \sigma_{\min} \leq \sigma \leq \sigma_{\max}\}$.

In the distributionally robust optimization problem, if the uncertainty set $\mathcal{A}$ includes distributions that have the extreme behavior of putting all the probability weight on a single value of ξ then we can assume that nature will choose one of these extreme distributions. The reason is simple, the expected value of Π under the distribution F must be greater than the minimum value that it could take, in other words if the density of F is f and this is non-zero on the range $[a, b]$ then

$$\begin{aligned} E_F[\Pi(x, \xi)] &= \int_a^b \Pi(x, s) f(s) ds \\ &\geq \int_a^b \left(\min_{a \leq z \leq b} \Pi(x, z) \right) f(s) ds = \min_{a \leq z \leq b} \Pi(x, z) \end{aligned}$$

We write δ_z for the distribution that puts all its weight at the single point z (sometimes this is called a Dirac delta distribution). Also we write $R(\mathcal{A})$ for the set of all values that may occur under a distribution in $\mathcal{A}$. Then in the special case that for every value of $z \in R(\mathcal{A})$ then δ_z is also in $\mathcal{A}$, we can deduce that

$$\max_x \left\{ \min_{F \in \mathcal{A}} E_F \left[\Pi(x, \xi) \right] \right\} = \max_x \left\{ \min_{z \in R(\mathcal{A})} \Pi(x, z) \right\}$$

so we are back to a pointwise robust optimization problem.

In some contexts it is natural to consider an uncertainty set $\mathcal{A}$ consisting of all distributions which are unimodal - the densities increase to a maximum and then decrease. For example when considering the distribution of demand for a product that has an uncertain relationship to the weather (at sufficient distance into the future that weather forecasts are not much use) we may be comfortable restricting the distribution to a unimodal one - even though almost nothing else is known.

We will consider an uncertainty set which consists of all unimodal functions defined on a range but we add the condition that the mode is known in advance. We can transform the problem so that the mode is zero. For example we may judge that the most likely value for the price of oil in a week's time is the price today, and we regard this price as stochastic with a distribution which is unimodal. We can see that the distribution is obtained by taking a unimodal distribution with mean zero and adding to it today's oil price.

A key observation is that any unimodal distribution that has support in a range $[-a, b]$ and mode 0 can be obtained by first choosing a number from a particular distribution G on $[-a, b]$ and then multiplying by a number drawn randomly from the interval $[0, 1]$. We can say that any unimodal distribution can be obtained as the product of independent samples from G and a from $U(0, 1)$ the uniform distribution on $(0, 1)$. This result is called Khintchine's Theorem. Another way to put this is to say that we can obtain any unimodal distribution from a mixture between uniform distributions each of which has a range either of the form $[0, x]$ or of the form $[-x, 0]$. Figure 8.3 demonstrates this by considering a distribution where the density function is an increasing step function for $x < 0$ and a decreasing step function for $x > 0$. We have shown how the density splits into horizontal rectangles; each represents a separate uniform distribution between the horizontal endpoints of that block (either a range $[-x, 0]$ or a range $[0, x]$). Suppose that we select each rectangle with a probability equal to its area (these sum to 1 since they equal the integral of the original density function f) and then sample from within the given rectangle uniformly within its horizontal range. It is not hard to see that the probability of ending up at any point matches that from the original distribution.

Suppose that we are considering the inner minimization. This is nature's problem, given a choice of x made by the decision maker how should the distribution of ξ be chosen to minimize the expected value of $\Pi(x, \xi)$? The available distributions are unimodal (with mode 0). Whatever distribution is chosen, an alternative to get the same result is to split this into its uniform (horizontal rectangle) components as in Figure 8.3 and then choose one of these with the appropriate probability. Thus if, for example, the distribution F was composed from three uniform distributions U_A, U_B and U_C with probabilities p_A, p_B and p_C then

$$E_F[\Pi(x, \xi)] = p_A E_{U_A}[\Pi(x, \xi)] + p_B E_{U_B}[\Pi(x, \xi)] + p_C E_{U_C}[\Pi(x, \xi)].$$

This is a convex combination of the expectations taken over the three uniform distributions, and so one of the three must have a value of $E_F[\Pi(x, \xi)]$ or lower. If all three had values

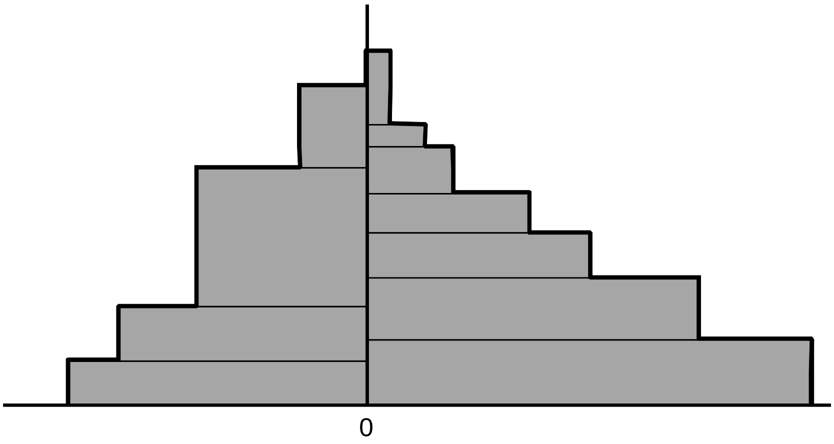


Figure 8.3 An example of a unimodal distribution as a mixture between uniform distributions

greater than $E_F[\Pi(x, \xi)]$ then the result of choosing between them with certain probabilities would also be greater than $E_F[\Pi(x, \xi)]$.

This is an example of the kind of argument we saw earlier: when minimizing a linear function over a polytope, we can just consider the extreme points of the polytope. In the same way if we are minimizing an expectation over a set of distributions $\mathcal{A}$ we can just consider the extreme points of the set $\mathcal{A}$.

In fact we can make this whole argument in a more abstract and general way. The set $\mathcal{A}$ of unimodal distributions with mode 0 and support in the range $-a$ to a is itself a convex set, since if we take a convex combination of two such distributions the result is still a unimodal distribution with mode 0. Moreover the uniform distributions on $[0, x]$ or $[-x, 0]$ for $0 \leq x \leq a$ are the extreme points of $\mathcal{A}$ (since they cannot be obtained through a convex combination of two other distributions in $\mathcal{A}$.) When minimizing a function that is linear on distributions, we only need to consider the extreme points (i.e. these uniform distributions).

We can go further in the case that $\Pi(x, \xi)$ is a concave function of ξ. In this case we can say that the minimum of $E[\Pi(x, \xi)]$ over distributions $F \in \mathcal{A} = \{$unimodal distributions with support in a range $[-a, b]$ and mode $0\}$ is attained either when F is uniform on $(-a, 0)$ or when F is uniform on $(0, b)$. We establish this result below.

8.3.1 When profits are concave in the uncertain variable

The key to understanding the simplification we can make if f is a concave function is to recognize that if we define

$$g(w) = \frac{1}{w - w_0} \int_{w_0}^{w} f(u)du,$$

so that $g(w)$ is the average value of f over the range (w_0, w) for $w \in (w_0, w_U)$, then g is also a concave function. This is enough to show that the minimum of g occur at one of the

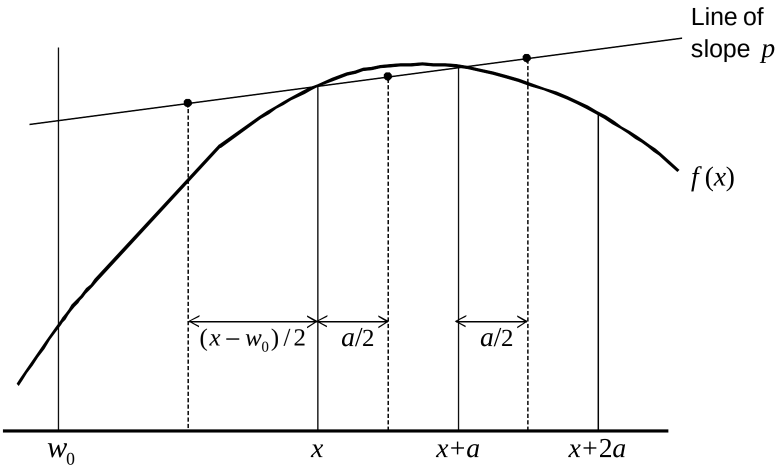


Figure 8.4 Diagram to illustrate the argument showing that the average up to x of a concave function is concave.

end points, i.e. at $w = w_U$ or at $w = w_0$. For this to make sense we need to define g at the limiting value w_0 and we simply take $g(w_0) = f(w_0)$ so that g is continuous.

To show that g is concave we will show that for any x and a the value of $g(x + a)$ is greater than $(g(x) + g(x + 2a))/2$. We write $\overline{A} = g(x)$, the average of f over (w_0, x), $\overline{B}$ for the average value of f over $(x, x + a)$ and $\overline{C}$ for the average value of f over $(x + a, x + 2a)$. We use the concavity of f to establish some relationships between these three averages. Specifically we let p be the slope of the straight line that goes through the f curve at x and $x + a$, i.e. $p = (f(x + a) - f(x))/a$. Since f lies above this line on the interval $(x, x + a)$ the value of $\overline{B}$ is greater than the point on this line at $x + (a/2)$. Also since this line is above f in the intervals (w_0, x) and $(x + a, x + 2a)$, the value of $\overline{A}$ is less than the point on this line at $x - (x - w_0)/2$ and the value of $\overline{C}$ is less than the point on this line at $x + (3a/2)$. All this is illustrated in Figure 8.4.

This means that the following inequalities hold

$$\overline{B} \geq \overline{A} + (\frac{x}{2} + \frac{a}{2})p$$

$$\overline{C} \leq \overline{B} + ap$$

Hence

$$\begin{aligned}\frac{g(x)}{2}+\frac{g(x+2a)}{2} &= \frac{\overline{A}}{2}+\frac{x\overline{A}+a\overline{B}+a\overline{C}}{2\,(x+2a)}\\ &\leq \frac{2x\overline{A}+2a\overline{A}+2a\overline{B}+a^2p}{2\,(x+2a)}\\ &= \frac{(x+a)^2\,\overline{A}+a\,(x+a)\,\overline{B}+a^2\,(x+a)\,p/2}{(x+2a)\,(x+a)}\\ &\leq \frac{(x+a)^2\,\overline{A}+a\,(x+a)\,\overline{B}+a^2\left(\overline{B}-\overline{A}\right)}{(x+2a)\,(x+a)}\\ &= \frac{x\,(x+2a)\,\overline{A}+a\,(x+2a)\,\overline{B}}{(x+2a)\,(x+a)}=\frac{x\overline{A}+a\overline{B}}{(x+a)}=g(x+a).\end{aligned}$$

Thus we have established that the average function is concave. Exactly the same argument can be used if the interval is from w to w_0 when $w < w_0$. In fact the argument can be turned around to show that when f is convex then g will also be convex.

When we use this result in the context of a distributionally robust optimization over unimodal functions then we wish to consider averages over ranges like (w, w_0) as well as averages over (w_0, w). Thus it is natural to extend the definition of $g(w)$ to w values below w_0, so

$$g(w)=\frac{1}{w_0-w}\int_w^{w_0} f(u)du,$$

for $w < w_0$. As we noted before we have $g(w_0) = f(w_0)$ and g is continuous over the whole range where f is defined.

Exactly the same arguments apply to this left hand mirror image and so g is a concave function for $w < w_0$. Moreover we can show that g is concave over the whole range. From what we know already all we need to establish is that there isn't a corner in g at w_0 with an increase in slope. But the concavity of f implies

$$f(w_0) \geq (1/2)f(w_0-y)+(1/2)f(w_0+y)$$

Integrating this inequality for values of y between 0 and δ implies

$$\delta f(w_0) \geq (1/2)\int_{w_0-\delta}^{w_0} f(u)du+(1/2)\int_{w_0}^{w_0+\delta} f(u)du$$

Dividing by δ we have $g(w_0) \geq (1/2)g(w_0-\delta)+(1/2)g(w_0+\delta)$.

The stronger result on the concavity of g over the whole range simply means that the minimum occurs at one of the two endpoints and not in the middle at w_0. So finally we have established that when we wish to minimize the expected value of a concave function $f(x)$ over choices of distribution which are unimodal, have support in (w_L, w_U) and have their mode at w_0 then we need only consider two options: the uniform distribution on (w_L, w_0) and the uniform distribution on (w_0, w_U). We show how to use this result to calculate a robust solution in the worked example below.

Worked example 8.3

Toulouse Medical Devices (TMD) needs to order heart monitors to meet an uncertain demand. TMD is uncertain about the distribution of demand but believes that the distribution is unimodal with lowest level of demand being zero and the highest being 200.The most likely value of demand is 100. Units cost $4000, and are sold at $10000. There are costs associated both with having unsold units and with not being able to meet the demand. TMD estimates that if they order x and demand is d then these 'mismatch' costs are $500(x-d)^2$. Thus the profit function (in $1000's) is

$$\Pi(x,d) = 10\min(x,d) - 4x - 0.05(x-d)^2 \tag{8.6}$$

TMD realizes that in some cases they may make a loss, but wishes to maximize their expected profit given the worst possible distribution of demand.

Solution

The profit function (8.6) is a concave function of d once x is fixed, since the three components are all concave. From the result above we need only to evaluate the expected profit for two particular distributions of d: either d is uniform on $(0,100)$ or d is uniform on $(100,200)$.

$$E_{U(100,200)}[\Pi(x,d)] = \frac{1}{100}\int_{100}^{200}(10\min(x,u) - 4x - \frac{1}{20}(x-u)^2)du$$

Clearly the value of this integral depends on the value of x. If $x \le 100$

$$\begin{aligned}
E_{U(100,200)}[\Pi(x,d)] &= \frac{1}{100}\int_{100}^{200}(6x - \frac{1}{20}(x-u)^2)du \\
&= 6x + \frac{1}{100}\left[\frac{1}{20}(x-u)^3/3\right]_{100}^{200} \\
&= 6x + \frac{1}{6000}((x-200)^3 - (x-100)^3)
\end{aligned}$$

If $x > 100$

$$\begin{aligned}
E_{U(100,200)}[\Pi(x,d)] &= \frac{1}{100}\int_{100}^{x}(10u - 4x)du + \frac{1}{100}\int_{x}^{200} 6x du - \frac{1}{100}\int_{100}^{200}\frac{1}{20}(x-u)^2)du \\
&= -4x\frac{x-100}{100} + \frac{1}{100}\left[5u^2\right]_{100}^{x} + 6x\frac{200-x}{100} + \frac{1}{2000}\left[(x-u)^3/3\right]_{x}^{200} \\
&= \frac{x}{10}(160-x) + \frac{1}{20}(x^2 - 100^2) + \frac{1}{6000}\left((x-200)^3 - (x-100)^3\right)
\end{aligned}$$

The other distribution can be evaluated similarly

$$E_{U(0,100)}[\Pi(x,d)] = \frac{1}{100}\int_{0}^{100}(10\min(x,u) - 4x - \frac{1}{20}(x-u)^2)du$$

If $x \leq 100$

$$\begin{aligned} E_{U(0,100)}[\Pi(x,d)] &= \frac{1}{100}\int_0^x (10u-4x)du + \frac{1}{100}\int_x^{100} 6xdu - \frac{1}{2000}\int_0^{100}(x-u)^2)du \\ &= -4x\frac{x}{100} + \frac{1}{100}\left[5u^2\right]_0^x + 6x\frac{100-x}{100} + \frac{1}{2000}\left[(x-u)^3/3\right]_0^{100} \\ &= \frac{1}{10}x\,(60-x) + \frac{x^2}{20} + \frac{1}{6000}\left((x-100)^3 - x^3\right) \end{aligned}$$

and finally if $x > 100$

$$\begin{aligned} E_{U(0,100)}[\Pi(x,d)] &= \frac{1}{100}\int_0^{100}(10u-4x)du - \frac{1}{100}\int_0^{100}\frac{1}{20}(x-u)^2)du \\ &= -4x + \frac{1}{100}\left[5u^2\right]_0^{100} + \frac{1}{2000}\left[(x-u)^3/3\right]_0^{100} \\ &= -4x + 500 + \frac{1}{6000}\left((x-100)^3 - x^3\right) \end{aligned}$$

The spreadsheet BRMch8-Toulouse.xlsx shows the values of $E_{U(0,100)}[\Pi(x,d)]$ and $E_{U(100,200)}[\Pi(x,d)]$ as x varies, and these are also shown in Figure (8.5). The robust optimum value for x is the one that maximizes the minimum profit. The optimum (integer) order size is $x = 73$ which guarantees a minimum expected profit of $99,883. We can see that for many x values a negative expected profit is possible if nature deals us a bad hand in the choice of demand distribution.

8.3.2 Notes

The material on Knightian uncertainty is mainly taken from the book by Bernstein. The whole area of robust optimization has excited a great deal of interest in the last few years and there are many papers that deal with different aspects of robust optimization. The review article by Bertsimas, Brown and Caraminis (2011) gives a good introduction to the very extensive literature on robust optimization. Our treatment here has been elementary and focused on relatively small scale problems with simple uncertainty sets. The discussion in Section 8.2 on budgets of uncertainty arises from work by Dimitris Bertsimas and co authors (Bertsimas and Sim, 2004). In that section we use duality properties to establish the exact problem to solve - this approach through duality can be extended to a whole variety of more complex robust optimization problems.

In the same way our treatment of distributionally robust optimization can be extended in many ways. The uncertainty set we have looked at in most detail, unimodal functions with a known mode, is particularly simple to analyze. Some related theory in a more general context of multidimensional distributions can be found in Shapiro (2006).

There has also been a great deal of work that looks at efficient computation of robust optimal solutions for a variety of problems - for example problems with dynamic characteristics matching the discussion of stochastic optimization that we gave in Chapter 7. For a much more comprehensive discussion of all of this material the reader is recommended to consult the book by Ben Tal, El Ghaoui and Nemrovski (2009).

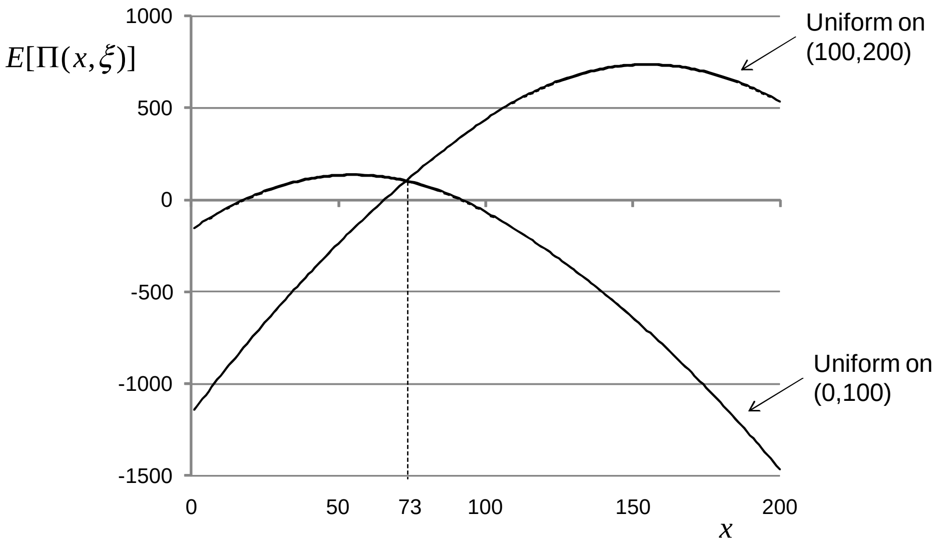


Figure 8.5 The expected profits for TMD as a function of x for the two extreme uniform distrubutions of demand.

8.3.3 References

Aharon Ben-Tal, Laurent El Ghaoui, Arkadi Nemrivoski, 2009, Robust Optimization, Princeton University Press.

Peter Bernstein,1996, *Against the Gods*, Wiley.

Alexander Shapiro, 2006, Worst-case distribution analysis of stochastic programs, *Mathematical Programming Ser. B*, Vol 107, 91-96.

Dimitris Bertsimas, David Brown, Constantine Caramanis, 2011, Theory and applications of robust optimization, *SIAM Review*, Vol 53, 464-501.

Dimitris Bertsimas and Melvyn Sim, 2004, The price of robustness, *Operations Research*, Vol 52, 35-53.

Exercises

8.1. (Impact of the budget of uncertainty)

Solve the Avignon Imports example without a budget of uncertainty (equivalent to setting $B = 3$) and compare the objective functions with $B = 2$ and $B = 3$ to see how much the expected profit is increased by taking the less conservative approach.

8.2. (Robust optimization for Sentinel)

In the Sentinel example use the spreadsheet to show that if the selling prices are reduced to \$600 and \$550 for the large and small formats, then the robust optimal solution has $x_L = x_S = 0$. Explain why this happens.

8.3. (Acropolis Rentals)

Acropolis Rentals has a fleet of 100 cars that it rents by the day. It is considering investing in GPS systems for these cars and will charge a premium of \$4 per day for hire of the GPS systems. Each of the GPS systems will cost \$500 to purchase and also requires the fitting of a secure holder with a cost of \$250 per car. Acropolis sells its cars after 500 days and this is also the effective lifetime of the GPS system. So a car with system installed has no extra value after this period in comparison with a car without the system installed. Once a Acropolis advertises this service it will be expensive in good will not to be able to provide it Acropolis reckons that there is a cost of \$10 when a customer is not able to have the system and requests it.

(a) Set this up as a robust optimization problem with the assumptions: (A) all cars are rented every day. (B) Acropolis has to make a decision at the start of the 500 days on how many GPS systems to install and cannot install any more over the course of the 500 day period. (C) The same proportion of customers request the GPS system each day (but Acropolis has no way of predicting this proportion).

(b) Show that if Acropolis makes a decision on how many systems to install and then the value of p, the proportion of customers wanting a GPS system, is chosen so that Acropolis makes least money, then either p will be chosen at 0 or at 1.

(c) Use the observation in (b) to solve the robust optimization problem.

8.4. (Toulouse Medical Devices)

In the Toulouse Medical Devices example, suppose that TMD know that the demand distribution is unimodal but are not able to say what the value of the mode is (with other aspects of the problem staying the same). Explain why their decision reduces to an ordinary (pointwise) robust optimization problem and calculate the best choice of order x.

9

Real Options

Commitment anxiety: good or bad?

“Why be in such a rush” is Darae’s comment when she hears what her boss Kandice is proposing. Kandice is the CEO of Analytics Enterprises an animation company working primarily on short advertisements and music videos and Darae is the Financial Director. For more than a year they have been looking for an opportunity to start a new division of their company working in the computer games area. Now a small company, Eckmet Ltd, specializing in computer games has approached them seeking an injection of capital. Eckmet’s main asset apart from its 10 employees are the rights to 20% of the sales for a new computer game that they have been working on, due to launch in 3 months time. The first payments under this contract are due in 6 months, but there is great uncertainty as to how successful the game will be.

Both Kandice and Darae agree that linking up with Eckmet is a good idea. Kandice is all for buying a controlling interest in Eckmet straight away. But Darae has been looking at the possibility of buying a smaller stake in the company, with the intention of taking a larger stake only if the new computer game does well. Sometimes a new game can take a while to catch on so Analytics would want to have an option on further share purchases for at least a 2 year period. Kandice has argued that this is even more expensive - a 20% stake in Eckmet would cost $2 million (with an option for a further 40% to be purchased at any time in the next two years at a cost of $4 million), while $4.8 million would give them a 60% stake right away.

Kandice cannot see the sense in this: why set yourself up to pay a total of $6 million for something that you can get right away for $4.8 million? But Darae is concerned about betting such a large sum of money on the success of one computer game, if that goes badly the investment in Eckmet will not be worth much. There must be a benefit in delaying making this commitment, but is it worth the extra price that they will end up paying?

9.1 Introduction to real options

In most cases increased uncertainty in outcomes is regarded as undesirable. Our discussion of the way that individuals make decisions in Chapter 6 confirms this to be the case for individuals when dealing with uncertain gains: we would rather have $80,000 for sure than an 80% chance of $100,000 and a 20% chance of nothing. We also met the same idea in

Business Risk Management Edward Anderson

Chapter 2 discussing portfolio theory, where we expect higher returns when the risk (i.e. the variance) is greater.

However there are circumstances where higher volatility or higher variance is beneficial. This happens whenever there is an implied option, with the effect of limiting any ‘downside’ associated with the variation. This beneficial effect is not as the result of a particular preference structure used by the decision maker: it happens when the decision maker is risk neutral.

At first sight it seems odd that an increase in variance might make things better, and it may be simplest to understand this using an example. Suppose that you are considering investing in one of two biotechnology companies both involved in similar Research and Development work. Company A typically produces about one new design idea every three months and on average the potential profitability of these product ideas, if they were to be put into production, has a mean of \$30,000 per year with a standard deviation of \$20,000 a year. (There is the possibility of losing money if the less successful product designs were ever put into production, so these designs are simply shelved). The second company, Company B, also produces one new design idea every three months with mean product profitability of \$30,000, but the standard deviation is higher at \$40,000 a year. Which is the better company to invest in?

It is not hard to see that Company B, with the *higher variance* of returns, represents the better investment. Product ideas that lose money or make very little will never be put into production. So the less successful designs have value zero, rather than a negative value. Hence there is a gain from the larger positive variations in profitability that is not offset by the larger negative variations. The idea is shown by Figure 9.1 in which a distribution of product profitability for the two companies is given with the assumption that unprofitable products have value zero. The bar at zero represent the probability of either of the profits taking a value zero. The mean value for the product profitability in the low variance case is \$31052 and for the high variance (dashed line) is \$33991. The greater the proportion of the distribution that is cut off, the higher the expected value for profit. And if we were to consider an even smaller standard deviation then we would find almost no probability of a negative profitability – leading to an expected profit equal to the mean of \$30000. The smaller the variance the worse the expected profit.

This is the phenomenon that we will explore in this chapter. It is important for managers to have a good understanding of why holding an option (in this case not to develop a product) makes increasing variability suddenly valuable. There are simple tools to carry out the calculations needed to put dollar values on this and this is what we turn to next.

9.2 Real option calculations using loss functions

Suppose that the outcome of some venture or investment is uncertain but there is a guarantee that the result will not be lower than a given value a. Perhaps we have the option of not going ahead with the venture in which case the outcome can never be less than zero, or perhaps we can always sell our investment for a given amount a since we hold a sell option for this amount (we will come back to a fuller discussion of financial options later). In any case if X is a random variable giving the profit in the absence of the option, then our expected profit, with the option to always make an amount a, will be given by

$$E(\max(X, a)) = E(\max(X - a, 0)) + a.$$

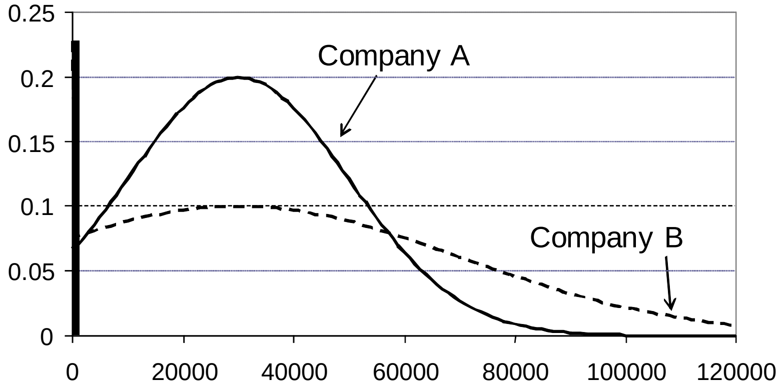


Figure 9.1 Profit distribution truncated at zero gives different results depending on the variance

This expression is reminiscent of the expected shortfall, which we met in chapter 4, and there is indeed a relation between the two. But the definition of expected shortfall is different because it involves an expectation conditional on $X > a$ (where a is the Value at Risk). We can write

$$E(\max(X, a)) = \Pr(X > a)E(X|X > a) + (1 - \Pr(X > a))a.$$

Suppose that X is a loss random variable and set a to the value VaR_α. Since $\Pr(X > VaR_\alpha) = 1 - \alpha$, then

$$E(\max(X, VaR_\alpha)) = (1 - \alpha)ES_\alpha + \alpha VaR_\alpha.$$

It will be useful to carry out calculations with a formula for $E(\max(X - a, 0))$ for different distributions for the random variable X. The simplest case is when X is uniform over a range (b, c) with $b < a < c$. Then

$$\begin{aligned} E(\max(X - a, 0)) &= \int_b^c \max(u - a, 0)\frac{1}{c - b}du \\ &= \frac{1}{c - b}\int_a^c (u - a)\, du \\ &= \frac{1}{c - b}\left(\frac{c^2}{2} - ca - \frac{a^2}{2} + a^2\right) \\ &= \frac{(c - a)^2}{2(c - b)}. \end{aligned}$$

In the same way we have

$$\begin{aligned}E(\max(a-X,0)) &= \int_b^c \max(a-u,0)\frac{1}{c-b}du\\ &= \frac{1}{c-b}\int_b^a (a-u)\,du\\ &= \frac{1}{c-b}\left(a^2-\frac{a^2}{2}-ba+\frac{b^2}{2}\right)\\ &= \frac{(a-b)^2}{2(c-b)}.\end{aligned}$$

Worked example 9.1

A salvage company has been asked to consider carrying out a difficult deep water salvage operation. The salvage value of the boat is \$3 million (20% of a \$15 million insured value), there will be a cost of \$300,000 to carry out a preliminary investigation which will determine the actual cost of the salvage operation. This cost is estimated to be between \$2 million and \$4 million and the company regards any amount between these numbers as equally likely. Obviously if the cost is found to be too large then the company will not go ahead with the salvage operation, but they will not recover the \$300,000 preliminary costs. What is the expected value of this project to the salvage company?

Solution

Once the \$300,000 has been spent it is a question of whether or not to go ahead with the salvage and this will be worthwhile if the cost is no more than \$3 million. Hence the expected profit in millions is

$$-0.3+E(\max(3-X,0))$$

where X is uniform on the interval between 2 and 4 (million). Using the formula above we have $a=3, b=2, c=4$ and an expected profit of

$$-0.3+\frac{(3-2)^2}{2(4-2)}=-0.05.$$

Hence this is a project that is not worth going ahead with. We can get the same result without using the formula since the numbers are all quite easy. 50% of the time the costs are more than \$3 million and the salvage is not worth doing, giving a loss of 0.3 million. On the other if the salvage is worthwhile (as happens 50% of the time), then the costs are uniformly distributed on the range 2 million to 3 million, giving an average cost of the salvage of \$2.5 million, meaning a profit of just 0.2 million after paying the up-front costs. With equal chances of a \$300,000 loss or a \$200,000 profit, the salvage company should walk away from this deal.

Next we will calculate the equivalent formulae for the normal distribution. We start by deriving two formulae:

$$\int_{-\infty}^{a} x\varphi_{\mu,\sigma}(x)dx = \mu\Phi_{\mu,\sigma}(a)-\sigma^2\varphi_{\mu,\sigma}(a), \tag{9.1}$$

$$\int_{a}^{\infty} x\varphi_{\mu,\sigma}(x)dx = \mu(1-\Phi_{\mu,\sigma}(a))+\sigma^2\varphi_{\mu,\sigma}(a), \tag{9.2}$$

where $\varphi_{\mu,\sigma}$ is the density function for the normal distribution with mean μ and standard deviation σ, and $\Phi_{\mu,\sigma}$ is the corresponding cumulative normal distribution function. The density of a normal distribution is given by the following expression:

$$d\Phi_{\mu,\sigma}(x)/dx = \varphi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Here we use the notation $\exp(x)$ to mean e^x.

The first formula (9.1) is derived by noting

$$d\varphi(x)/dx = \left(-\frac{(x-\mu)}{\sigma^2}\right)\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = -\frac{(x-\mu)\varphi(x)}{\sigma^2}$$

and so

$$\begin{aligned}\varphi(a) &= \int_{-\infty}^{a} \left[d\varphi(x)/dx\right] dx = \int_{-\infty}^{a} -\frac{(x-\mu)\varphi(x)}{\sigma^2} dx \\ &= \frac{\mu}{\sigma^2}\Phi(a) - \frac{1}{\sigma^2}\int_{-\infty}^{a} x\varphi(x)dx.\end{aligned}$$

We get the formula (9.1) simply by rearranging this equation.

The formula (9.2) comes from

$$\mu = \int_{-\infty}^{\infty} x\varphi(x)dx = \int_{-\infty}^{a} x\varphi(x)dx + \int_{a}^{\infty} x\varphi_{\mu,\sigma}(x)dx$$

and hence

$$\int_{a}^{\infty} x\varphi_{\mu,\sigma}(x)dx = \mu - \int_{-\infty}^{a} x\varphi(x)dx$$

and substituting from (9.1) gives the result we are looking for.

Now we suppose that the random variable X has an $N(\mu,\sigma)$ distribution and we derive the value of $E\left[\max(a-X,0)\right]$. We have

$$\begin{aligned}E\left[\max(a-X,0)\right] &= \int_{-\infty}^{a} (a-x)\varphi_{\mu,\sigma}(x)dx \\ &= a\int_{-\infty}^{a} \varphi_{\mu,\sigma}(x)dx - \int_{-\infty}^{a} x\varphi_{\mu,\sigma}(x)dx,\end{aligned}$$

and hence

$$E\left[\max(a-X,0)\right] = (a-\mu)\Phi_{\mu,\sigma}(a) + \sigma^2\varphi_{\mu,\sigma}(a). \qquad (9.3)$$

Similarly we have

$$\begin{aligned}E\left[\max(X-a,0)\right] &= \int_{a}^{\infty} (x-a)\varphi_{\mu,\sigma}(x)dx \\ &= \int_{a}^{\infty} x\varphi_{\mu,\sigma}(x)dx - a\int_{a}^{\infty} \varphi_{\mu,\sigma}(x)dx \\ &= \int_{a}^{\infty} x\varphi_{\mu,\sigma}(x)dx - a(1-\Phi_{\mu,\sigma}(a)),\end{aligned}$$

and so

$$E\left[\max(X-a,0)\right] = (\mu - a)(1-\Phi_{\mu,\sigma}(a)) + \sigma^2\varphi_{\mu,\sigma}(a). \tag{9.4}$$

Now we can return to the example of section 9.1 and see where the numbers come from. The expected profit from company A in \$1000's is given by $E(\max(X,0))$ when X has a $N(30,20)$ distribution. Setting $a=0$ in (9.4) we have

$$\begin{aligned} E(\max(X,0)) &= 30(1-\Phi_{30,20}(0)) + 20^2\varphi_{30,20}(0) \\ &= 30(1-\Phi_{30,20}(0)) + 400\frac{1}{\sqrt{800\pi}}\exp\left(-\frac{900}{800}\right). \end{aligned}$$

The first term can be evaluated from tables of the normal distribution but it is simplest just to use a spreadsheet to evaluate the formula

=30*(1-NORMDIST(0,30,20,1))+400*NORMDIST(0,30,20,0).

This uses the NORMDIST(x,μ,σ,) which returns either the cumulative distribution function $\Phi_{30,20}(x)$ or the density function $\varphi_{30,20}(x)$ according as the final argument is 1 or 0. Using this formula gives the value $\$30586.14$. This can then be repeated with $\sigma = 40$ to obtain the profit figure $\$35246.68$

Another way to think about what is going on in this example is to use the language of options. Suppose we are considering the expected benefit from a single product idea for Company B. We can regard the value of a single product idea as an option to purchase the resulting product patent. An option is (to use the standard phrase) ‘the right but not the obligation’ to purchase the product patent. If, once the idea is fully worked out and becomes an actual product, it turns out that the product is a loss maker then we will not choose to exercise our option: in other words we will not take the product to market. The key idea here is that though we pay money up front, there is a point later on, when we have more information, at which we will make a decision whether or not to go ahead.

9.3 An investment example

Applying a real option approach to an investment decision will usually involve putting some money into a project that only later delivers returns. In practice we need to do these calculations taking account of an appropriate discounting of returns over time. In other words we need to add the real options into a net present value (NPV) calculation. To see the sort of calculations involved we consider a simple example.

Foxtrot Developments is considering the purchase of a block of land for development. If purchased at the start of this year (year 1) development approvals are expected to be completed by the end of next year (year 2) and building will take a full year to complete, so that the property can be let from the beginning of year 4. Currently the building on the land is let and generates an income of \$50,000 a year. If the building is developed for commercial purposes Foxtrot calculates that it will bring an income of \$200,000 a year. However there is a question about the building costs in year 3. Building costs typically vary randomly from year to year. Foxtrot estimates that if the building were constructed now the costs would be \$2,800,000. The question is how much is the maximum that Foxtrot should pay for this development?

Suppose first that Foxtrot calculates this number by simply guessing that building costs stay the same as they are now. Then the income stream it receives is as follows (all sums are in $1000's and we have worked in constant 2010 dollar values - so income streams are shown as having a constant value even if they could be expected to increase with inflation)

	year 1	year 2	year 3	year 4	year 5	terminal value
cash flow	50	50	-2800	200	200	4200
× discount factor	1	$1/1.05$	$1/1.05^2$	$1/1.05^3$	$1/1.05^4$	$1/1.05^5$
= present value	50	47.6	-2539.7	172.8	164.5	3290.8

We have chosen to assume that all monies are paid or received at the start of the year The terminal value here is the value in year 6 of an income stream of $200,000 a year discounted at a rate of 5% using the formula:

$$x + x/(1+r) + x/(1+r)^2 + x/(1+r)^3 + ... = \frac{(1+r)}{r}x,$$

which with a discount rate of 5% gives $(1.05/0.05)x = 21x$. We can add the present values together to get a value for the project of

$$50 + 47.6 - 2539.7 + 172.8 + 164.5 + 3290.8 = 1186.$$

This figure needs to be compared with the value if there was no development of the site. In this case the value would be equivalent to an income stream of $50,000 a year which under these conditions has an NPV of $1050,000. The relatively high building costs have cancelled out most of the benefit of the additional rental income.

However by taking account of the uncertainty in building costs, together with the option value of the possibility of not going ahead with the development, then the actual value of the investment to Foxtrot can be seen to be higher.

At the beginning of year 3 a decision will be made on whether or not to build. If Foxtrot goes to tender and obtains a bid of x then the choice is between continuing with $50,000 a year or paying x this year and then receiving $200,000 a year It is not hard to work out a value of x so that we are indifferent between the two options - meaning that any higher value for the building costs would lead to Foxtrot not going ahead. At the beginning of year 3 the choice is between a present value of $21 \times 50,000 = \$1050,000$ and a present value of

$$-x + \$4200,000/(1.05) = \$4,000,000 - x.$$

These are equal when $x = \$2,950,000$.

This tells us how the decision will be made: Foxtrot will build only if the price is less than $2,950,000. The remaining piece of the jigsaw puzzle is an estimate of the volatility of building costs. Suppose that Foxtrot believes that building costs are equally likely to move up or down from the current value of $2.8 million and the result is normally distributed with a mean of $2.8 million and a standard deviation of $200,000. The expected current value of the investment in $1000s is given by the expected value of

$$50 + 47.6 + \frac{1}{1.05^2}\max(1050, 4000 - x).$$

Thus the expectation is

$$97.6 + \frac{1}{1.05^2}\left[1050 + E(\max(0, 2950 - x))\right].$$

It is worth pausing at this point to consider this formula. Once we know that the breakeven point on going ahead is at $x = 2950$ then we know that any higher value of x will just deliver the baseline NPV of 1050. On the other hand having a lower value of x is just like getting the baseline NPV of 1050 plus a payment in year three of the saving $\$2950 - x$.

Now we can use our earlier formula (9.3) to show

$$\begin{aligned} E(\max(0, 2950 - x)) &= 150\Phi_{\mu,\sigma}(2950) + 200^2\varphi_{\mu,\sigma}(2950) \\ &= 176.23. \end{aligned}$$

So the final valuation is

$$= 97.6 + \frac{1}{1.05^2}(1050 + 176.23) = 1209.827,$$

or \$1,209,827 which is \$23,827 more than the previous figure. So in this example recognizing the real option is worth an additional 2% in the value of the investment.

9.4 The connection with financial options

In the financial world options come in two varieties: a call option gives the right (but not the obligation) to purchase an underlying financial instrument at some point in the future at a given ‘strike’ or ‘exercise’ price. If the date of the exercise of the option is fixed it is called a European option, if the option can be exercised at any point up to the expiry date of the option it is called an American option. A put option is similar except that it gives the right (but not the obligation) to sell the underlying financial instrument at a given price.

<*Switch to a US stock with 2013 data*> For example when these notes were written a share in Woolworths on the ASX was trading at \$26.44 and an (American) call option with a strike price of \$28 to be exercised by 28 October 2010 is selling for \$0.155; a call option at \$27 sells for \$0.415; and a put option at \$27 sells for \$1.325. We can review what those numbers mean: For an outlay of \$0.155 an investor gets the chance to buy the stock for \$28 on or before 28/10/10. If the stock’s value on that date is less than \$28 the option is valueless, but if Woolworths is selling for \$28.50 then the option is worth \$0.50, and if the price is higher the option will be worth even more. A put option is the reverse: for an outlay of \$1.325 an investor gets the opportunity to sell the stock at \$27. If the stock price is actually above that level on 28/10/10 then the put option is valueless, but if for example the Woolworths share price returns to its current value of \$26.44 then the put option will be worth \$0.56. As well as buying put or call options at a whole lot of different exercise prices and a number of different expiry dates, there is also the option to sell these options (sometimes described as ‘writing’ an option). So there are a very large number of different positions that an investor may take, and investors will often decide to hold a portfolio of options in a stock in order to tailor the profile of possible gains or losses that they could experience.

A risk averse investor may want to purchase a significant shareholding in Woolworth’s shares and at the same time buy put options at an exercise price of say \$25 to provide a type of insurance against a large drop in the price of the shares. If the same investor was to sell a call option at a higher exercise price say $29 then they would limit their possibility of a large gain, but the money they receive for the call option could be put towards the cost of buying the put option. Notice however that there is nothing to stop an investor from buying and selling the options without actually holding any shares. In fact this is the norm. The option contract may specify physical delivery (of say the shares in Woolworths) at the point of settlement, but often this will not actually take place and instead the option will be cancelled out by buying a covering position. Some option contracts (for example those where the underlying security is an index, like the ASX index) specify cash settlement so that payments are made but no stocks delivered. Usually it is best, rather than thinking of a call option in terms of the right to buy Woolworths stock, to consider the option as a financial contract involving an agreement for the seller to pay the buyer the difference between the Woolworths stock price and the strike price if this is in the right direction.

Now we return to the example of the previous section. If we think of building costs as like a stock price, then a low value is good for Foxtrot - the lower the building cost is the more valuable the investment is, but once the building cost goes above 2.95 million then it no longer matters what the price is since the building will not be worthwhile. So the investment has the characteristics of a put option with an exercise price of 2.95 million. Suppose we take as a baseline case the project when the building cost is 2.95 million, which makes Foxtrot indifferent between going ahead with the building or not. Then any lower cost is equivalent to Foxtrot receiving the difference in year 3, but any higher cost leaves things as they are. So this matches the put option where we have the right to sell at the exercise price so that when the price falls below the exercise price we can buy at one price and sell at a higher price making a profit of the difference. Figure 9.2 shows how the value of a put or call option depends on the underlying share price at the time of exercise and the exercise price.

In the Foxtrot example we can see the purchase as having two parts: first we buy the property in its undeveloped state with a certain value and in addition we buy a put option on the building price index. This is a European option to be exercised at the beginning of year 3 with an exercise price of 2.95 million for the building. Obviously if the building price index were on a square metre basis then the exercise price is divided by the size of the building.

A final question relates to the size of the option to be purchased. Here we need to look at the slope of the put option line in Figure 2 and compare it with the amount of increased profit if the price drops below 2.95 million. In regards to the additional value of the option *at the time of its exercise* every dollar less is a dollar more earned for Foxtrot. Thus the equivalence is with a put option for the full value of the building cost. Notice that in these calculations the complexities associated with taking a one year gap in earnings and then replacing an annual sum of $50,000 with an annual sum of $200,000 are all dealt with within the single number of 2.95 million which is the exercise price of the option.

We have given a direct approach to valuing an option when the uncertainty can be represented as a Normal distribution with a known mean and variance at the date of exercise of the option. However the most famous approach to valuing options is the Black-Scholes formula. This gives the price of a European option in terms of 5 quantities:

a. the underlying stock price now, S_0,

b. the volatility in the stock price, σ,

c. the time till the exercise of the option, T,

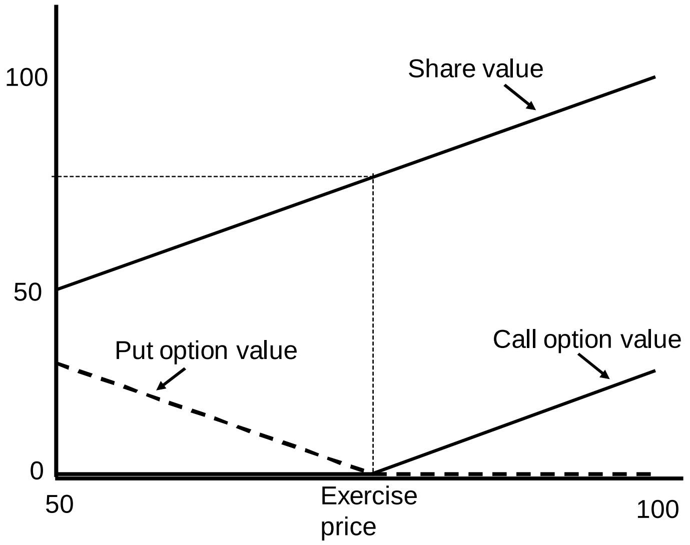


Figure 9.2 The payments for put and call options as a function of the share price

d. the exercise price, K,

e. the discount rate to be applied (risk free rate of return) r.

When there is no dividend yield the Black-Scholes formula gives the price of European call option (the right to buy the stock at price K at time T) as

$$S_0 \Phi_{0,1}(d_+) - e^{-rT} K \Phi_{0,1}(d_-)$$

where

$$d_+ = \frac{1}{\sigma\sqrt{T}} \left[\log\left(\frac{S_0}{K}\right) + \left(r + \frac{\sigma^2}{2}\right) T \right],$$
$$d_- = \frac{1}{\sigma\sqrt{T}} \left[\log\left(\frac{S_0}{K}\right) + \left(r - \frac{\sigma^2}{2}\right) T \right].$$

The first three variables (S_0, σ and T) will determine the distribution of stock prices at the exercise date.

One big difference between the Black-Scholes approach and the examples we have given for real options is that the financial markets have stock prices moving in a multiplicative way - so rather than a $100 share price being equally likely to move to $110 or $90 (say) a movements up by a factor 1.1 would imply an equally likely movement down by $1/1.1$ so that an increase from $100 to $110 and a decrease from $100 to $$90.91$ are equally likely. This multiplicative behavior means that it is the log of the share price that is likely to exhibit a normal distribution rather than the share price itself.

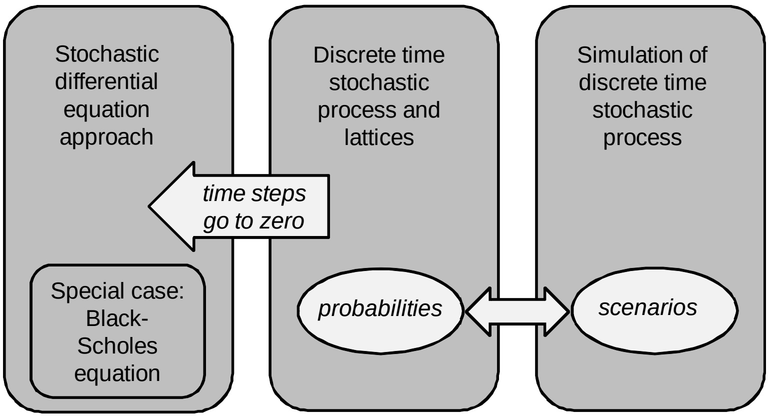

Stochastic differential equation approach
Discrete time stochastic process and lattices
Simulation of discrete time stochastic process
time steps go to zero
Special case: Black-Scholes equation
probabilities
scenarios

Figure 9.3 Different approaches to valuing real options

However the Black-Scholes formula is not derived by just looking at an expected value at the exercise time. Instead the approach is more sophisticated and involves constructing a synthetic risk-free instrument that must then match the risk-free rate of return by a 'no arbitrage' argument. This allows the calculation to take place without consideration of a drift over time: the volatility alone is enough to work things out. The idea is that if the option relates to an asset that is traded then the price for the asset now is not independent of the expected future behavior of the asset price: if the price now gets out of line with what will happen in the future then someone will seize the opportunity to buy or sell and make money from the trade. In other words we cannot separate out the drift in asset value and the risk free rate of return and have these two things independently chosen.

There are a range of different approaches to valuing real options: we can use the Black-Scholes formula; we can solve a set of stochastic differential equations with appropriate boundary conditions; we can carry out an evaluation using a type of binomial (or trinomial) lattice; or we can use a Monte Carlo simulation approach. This is illustrated in Figure 9.3. A stochastic differential equations approach arises from assuming some form of Brownian motion or related stochastic process for the uncertain prices (or returns). The Black-Scholes formula is a special case available for traded assets under certain assumptions. But in any case the stochastic differential equations can be regarded as the limit of a discrete time stochastic process as the time increments get smaller and smaller. Lattice calculations work directly with these discrete time stochastic processes and allow extra flexibility in the modelling by using a discrete state as well. The idea is to calculate the probabilities of being at different states arranged on a two-dimensional lattice with time on one dimension and price on the other. Finally the Monte Carlo approach replaces a calculation of probabilities in the lattice approach by a simulation. This is much simpler and may well be more useful in practice.

From the point of view of our discussion here we will stick to the simplest case and discuss an approach using Monte Carlo simulation in the next section

Finally it is worth commenting on some of the advantages in making a connection between real options and financial options:

- The business world now contains many people with a familiarity with options: there is no difficulty in understanding an option approach and it can be helpful in appreciating the financial structure of a potential investment (as well as explaining why a greater variation can lead to a more profitable outcome).
- With a map from the investment into a set of options it becomes possible to use well-established techniques and software for valuing options. We have already mentioned the Black-Scholes formula that applies for European options when the price movements can be modelled as a multiplicative Brownian. But there are other techniques in common use by those who need to value different types of options. American options are usually valued using some sort of simulation process. In this example the option was European but different types of real options scenarios can give rise to either European or American options. We will show how the Monte Carlo simulation technique can be used in a real options framework in the next section.
- If the underlying uncertainty relates to a traded instrument (like a commodity price) then the option value may not be something which relies on calculation - perhaps it can be found simply by looking at the prices in the market place.
- In the event of an actual market place for the option implied in the investment there is also the possibility of buying or selling options to entirely cancel the uncertainty in relation to the exercise of the implicit option. There are a number of building price indices available in different parts of the world. To the best of my knowledge there are no traded options in any of these indices. However, for the sake of an illustration suppose there were. Then Foxtrot would have the option of selling a put option of the appropriate quantity, exercise price and maturity at the same time as buying the property. Doing this carefully could end up with approximately the same net present value independently of whether or not the re-building goes ahead. Variations in net present value would only occur if the building quotes came in very different from what would be expected based on movements in the building price index.

9.5 Using Monte-Carlo simulation to value real options

Real options are all about taking advantage of the opportunity to delay a decision until more information is available, and often the decision is based on something that varies in a stochastic way. It is natural to use a Monte Carlo approach in this context and this will give us the flexibility to represent different aspects of what is going on. An advantage of using this method is that we can easily see the distribution of possible outcomes, so that we can obtain risk measures like Value at Risk from the same calculations.

We will give a simple example of how this can work in practice. Suppose Quickstep Oil is considering the development of an shale oil resource project with high costs where the decision whether or not to go ahead is related to the price of oil in the future. The first decision is whether Quickstep should buy rights to the resource. Quickstep will then need to build some substantial additional processing capacity and infrastructure, but this could be delayed until prices are higher than their current value. It will take a year to build the plant (with build costs spread over this period). Once in operation there is a large (fixed) cost per year of operation together with substantial costs that depend on the volume of oil produced

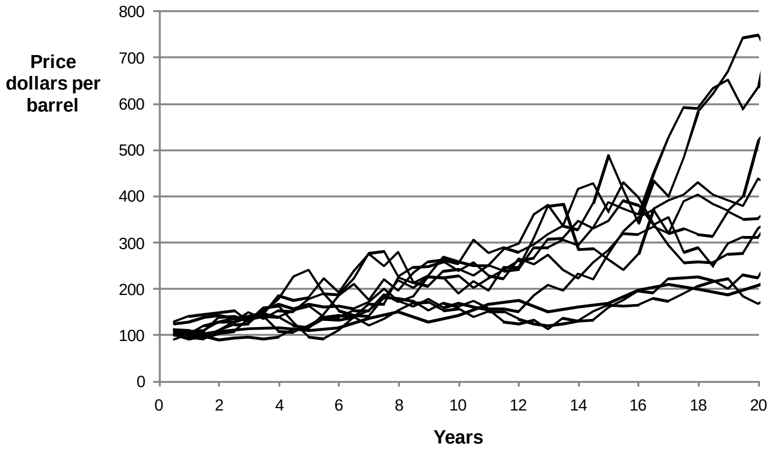


Figure 9.4 A sample of 10 price series for the Quickstep example

We suppose that the price of crude oil is modelled as a stochastic process which involves an underlying price which follows a geometric Brownian motion with drift upwards, but that in addition this is subject to fluctuations due to short term global supply and demand changes that we model through a mean reverting process. Writing w_t for the log of the price then $w_t = y_t + \theta_t$ where y_t represents the underlying level: $y_t = y_{t-1} + \alpha + \varepsilon_t$ and θ_t is mean reverting (to zero) $\theta_t = \beta\theta_{t-1} + \delta_t$ where $\beta < 1$. Whether or not this is a good model for oil prices, it does illustrate the way that the Monte Carlo methodology is not restricted in the type of price processes that can be considered. In fact it is not easy to decide on the right model for oil prices over a long horizon: prices seemed to be very stable prior to the early 2000s, then they began to steadily climb over the period since around 2002 reaching more than $130 a barrel in the middle of 2008, before dropping to below $50 at the end of that year, since when they have steadily climbed again.

We assume that the current oil price is $100 per barrel. We will work with a 6 month time unit. We have set $\alpha = 0.04$, $\beta = 0.75$ and ε_t is normal with mean zero and standard deviation 0.07, and δ_t is normal with mean zero and standard deviation 0.09. Figure 9.4 shows a sample of 10 price realizations over a 20 year period given these parameters.

The cost of securing the site properly and decommissioning some existing plant (works that need to be carried out immediately on purchase) is $25 million. The cost of building the plant (which will take a year) is $400 million. Once in operation the plant can produce 2 million barrels per year. The fixed operating cost per year is $80 million and the (variable) production costs are $70 per barrel. Quickstep uses a discount rate for capital projects (based on its weighted average cost of capital) of 10%, but we will work with constant 2013 dollar values and this makes it appropriate to use a lower value for the discount rate of 7%.

We will assume that the reserves are sufficient for the plant to run for 20 years. We will ignore what happens after the end of the plant operation (residual value of production equipment, cost of winding up the site)

Given this arrangement the spreadsheet BRMch9-Quickstep.xlsx gives the Monte Carlo simulation. This is a very unwieldy spreadsheet since it includes 1000 simulations, each one over a period of 40 years. Each simulation involves four rows: the two time series y_t and θ_t used to generate the prices, the prices themselves and the cash flow line which involves some messy formulae that have the effect of ensuring that the production plant starts to be built when the price reaches some specified price threshold and then switches off 20 years later.

Because of the delay in building the plant the earliest that the plant can begin operation is 18 months after purchase. The average price at that point will be above the figure of $110 per barrel that makes the whole thing economic (during 18 months the log price will increase by $3a = 0.12$ implying an increase in the oil price by a factor of $\exp(0.12) = 1.1275$). A starting point is to use the Monte Carlo simulation to calculate the expected net present value of the project given that there is no delay in building the plant – this can be achieved by setting the price threshold to something lower than 100. The result will depend on the specific random numbers that happen to come up in the simulation, and even with 1000 scenarios there is quite a bit of variation (which we can see by pressing “F9” to get another ‘draw’ of the random numbers). The overall average net present value is about $3190 million, but a set of 1000 individual scenarios can have an average value anywhere in a range from around $3110 million to around $3250 million.

The next step is to try to take account of the flexibility that is available to Quickstep: if the price of oil drops it makes sense to wait before committing to spend $400 million on building the plant. A conservative approach would be to wait till the price gets above the breakeven level of $110 before starting. This will not be a guarantee against a later price drop, but because there is usually an upward drift in the price process, the chance of losing money is certainly reduced. The Monte Carlo simulation gives an overall average net present value of about $3550 million, but a set of 1000 individual scenarios can have an average value ranging from around $3430 million to $3620 million. The average improvement in NPV arising from the flexibility to delay starting construction is about $365 million in this example. Figure 9.5 shows a comparison between the cumulative distribution of net present value, obtained from two sets of simulations — one with no delay and one with a price threshold of 110.

9.6 Some potential problems with the use of real options

When real options were first discussed there was considerable excitement about their application as a strategic tool in evaluating investments and other management decisions. It would be fair to say that the use of real options theory to produce ‘hard numbers’ for valuations, or investment decisions, has been less common than many were predicting.

Bowman and Moskowitz discuss some of the reasons for the limited use of real options theory by managers. The most straight forward way to use real options is simply to make use of an option valuation tool like the Black-Scholes formula. However this is likely to involve an assumption that the underlying stock price follows a lognormal distribution, which may well be an inappropriate assumption for a strategic option. Broadly speaking when the uncertainty relates to prices and dollar values then movements up or down tend to be multiplicative leading to log normal distributions in the limit (through an application of the central limit theorem), but when uncrtainty relates to something else (for example sales of a new short life-cycle product) then this is unlikely to be well modeled using a lognormal distribution.

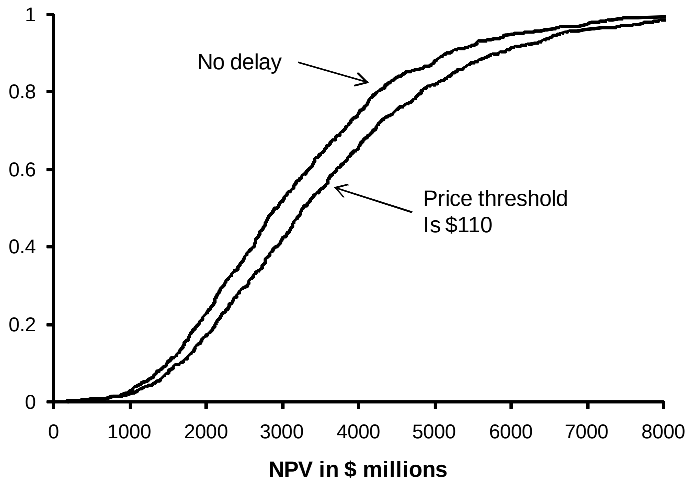


Figure 9.5 Improved NPV risk profile for Quickstep given option to delay

A second problem with the use of 'off the shelf' option valuation tools is that they assume tradeable assets where the possibility of arbitrage has a big effect on how prices behave. Thus for example with an exchange-traded option stock prices are easily observable and an option holder can readily buy or sell shares at this price to realize the profit from (or to cut the loss on) an option position. In contrast, for real options, the analogous stock price is often very hard to ascertain and it may also be hard to trade at the price implied.

There are also problems with the time to expiration. For strategic real options, there is often no set time to expiration. For example, a research project could be extended for a longer period of time, and an investment in a new product distribution system indefinitely retains the option to add additional products.

So there are a number of problems with applying a straightforward option analysis and it may be better to think of building a more advanced and customized option valuation model. But this brings with it some dangers. Creating such a model is a technical challenge that will take it out of the hands of the managers who will rely on its results. Moreover the complexity of the options approach can also make it difficult to find errors in the analysis, or to spot overly ambitious assumptions used by optimistic project champions.

This list of difficulties explains why we have given quite a bit of attention to relatively unsophisticated approaches like Monte Carlo simulation. It is important also to recognize that much of the value of an options analysis will be at the stage of project design. By bearing in mind the timing of decisions particularly those of an options nature (e.g. the decision to go ahead with, or to defer some expenditure) we may well be able to create additional value. In understanding these situations a very accurate numerical model may well not be necessary. The model inaccuracy in these situations may well be quite small. As Bowman and Moskowitz point out: "Whereas small deviations are worth fortunes in financial markets, they are fairly inconsequential in product markets."

9.6.1 Notes

Our approach to real options has focussed on the fundamentals of how to use flexibility and the calculations that are needed in specific examples to evaluate projects where there is flexibility. Because we think there are problems with its application in a real options environment, we have given rather little attention to the Black-Scholes equation which has sometimes formed the basis for these valuations (see for example Luehrman 1998). The approach we take is broadly in line with the recommendations in Copeland and Tufano (2004) and also with the approach proposed in the book on flexibility in design by de Neufville and Scholtes (2011).

We have only given a very brief introduction to this area and there is much more to be said. A paper which gives a careful and helpful treatment of some of the issues that need to be dealt with in practice is that by Smith and McCardle (1999). There are many books dealing with real options ranging from the original discussion of Dixit and Pindyck (1994) to more recent books by Mun (2005) and Guthrie (2009).

References

Bowman EH and Moskowitz GT 2001 Real options analysis and strategic decision making. *Organization Science* **12**, 772–777.

Copeland T and Tufano P 2004 A real-world way to manage real options. *Harvard Business Review*. **82**(3), 90–99.

Dixit A and Pindyck R 1994 *Investment under Uncertainty*. Princeton University Press.

Guthrie G 2009 *Real Options in Theory and Practice.* Oxford University Press.

Luehrman T 1998 Investment oportunities as real options: Getting started on the numbers, *Harvard Business Review*. July-August 1998 3–15.

Mun J 2005 *Real options analysis: Tools and techniques for valuing strategic investment decisions.* Wiley, 2nd edn.

De Neufville R and Scholtes S 2011 *Flexibility in Engineering Design.* MIT Press.

Smith J and McCardle K 1999 Options in the real world: Lessons learned in evaluating oil and gas investments, *Operations Research* **47** (1) 1–15.

Exercises

9.1 (Pop concerts)

Two investments are available in pop concerts. One of them involves paying an up front sum of $\$10,000$ and then receiving 10% of the net ticket sales which are uncertain but are expected to be $\$120,000$ with a standard deviation of $\$20,000$. The other venture is more speculative and will fund a series of three shows. The expected ticket sales are $\$230,000$ with a standard deviation of $\$80,000$. The fixed cost to put on the shows is $\$190,000$. 19 investors have all been asked to put in $\$10,000$ to cover these fixed costs. There is a chance that the shows make less than $\$171,000$, in which case it has been agreed that each investor will receive back $9000 and any final shortfall will be met by the producers of the show. If the shows achieve net ticket sales of more than $$\$171,000$ then each of the 19 investors will receive $9000 dollars back plus one twentieth of the profit over and above $\$171,000$ (so at the point where the ticket sales are $\$191,000$ the investors will receive all of their $\$10,000$ stake back). The remaining one twentieth share will be paid to the producers. Calculate the expected value of both investments. Which investment is preferred?

9.2 (Gold extraction)

Charleston Mining Company is considering buying a licence allowing the extraction of gold from mine ‘tailings’ for a period of 3 years. Extraction is an expensive process and only worthwhile if the gold price is high enough. The operation costs an average of $1000 to extract 1 kg of gold. It is expected that the operation will produce a total of 500 kg per month. The price of gold is volatile and is currently $1200 per kg. If the price drops below $1000 per kg the operation will simply be stopped until the price rises above that threshold.

(a) If the price of gold in January next year is estimated to have a normal distribution with mean $1200, and standard deviation $200 what is the expected revenue from the operation for January?

(a) What options purchase would match the cash flows for the operation in January next year? (If this was repeated for each month of the licence it would give a route to valuing the licence without needing to estimate volatility in gold prices.)

9.3. (Trade shows)

To develop a new product Tango Electronics must spend $12 million in 2011 and $15 million in 2012. If the firm is the first on the market with this product it will earn $60 million in 2013. If the firm is not first in the market it will do no more than cover its costs. The firm believes it has a 50% chance of being first on the market. From 2014 onwards there will be a number of other firms that enter this market and though Tango may well continue with production it expects to do no more than cover its costs. The firm’s cost of capital is 12% and you should use this figure to discount future earnings.

(a) Should the firm begin developing the product?

(b) Now suppose that there will be a trade show on 1 January 2012, when all the potential market entrants will show their products. After the trade show Tango will make a new estimate of its probability of being first on the market. Assume that at this point it can correctly state its probability of being first in the market, and further assume that this probability estimate is equally likely to take any value between 0 and 1. Should the firm begin developing the product?

9.4. (SambaPharm)

A company has the option to buy a small firm called SambaPharm whose main asset is a patent on a pharmaceutical product currently undergoing clinical testing. Testing will take a further 3 years at a cost of $60,000 a year. The long term profitability of the drug will depend on the results of these clinical trials. The best existing treatments are effective for 40% of patients. The final sales of the product are related to the number of patients for whom it is effective. It will just break even if it is equally as effective as the best current drug. But for every additional 10 percent of patients for which it is effective the net annual income (after production costs) will increase by $50,000. The best guess is that the effectiveness is equally likely to be at any value between 0 and 80%. Hence with a probability of 0.5 the drug will be found to be less effective than the best existing treatment and will not be put into production. Once in production the drug will have an estimated 5 year life before the next variety of this pharmaceutical family appears and profits are reduced to zero (or close to zero). But for a period of 5 years, starting in year 4 immediately after clinical testing, the profitability of the drug is expected to be stable. Taking account of the implied real option, what is the value of the company holding this patent if future profits are discounted at a rate of 8% per year?

10

Credit Risk

Credit ratings track a firm in hard times

Liz Claiborne was an American fashion designer who was born in Brussels in 1929. In 1976 she was one of the cofounders of Liz Claiborne Inc. which made it into the Fortune 500 in 1986. She was the first woman to be the CEO of a Fortune 500 company, and had a big influence on the way that fashion is sold insisting that the clothes in her collection be grouped together in the store, rather than Liz Claiborne skirts being put together with other skirts and in a separate place to shirts. Claiborne retired from active management in 1989 and died in 2007.

Liz Claiborne Inc was generating 2 billion dollars in annual sales in the early 1990s and expanded through acquisitions through the 1990s and early 2000s, buying Lucky Brand Jeans, Mexx and Juicy Couture. But then they hit problems; perhaps there were too many acquisitions, but there were also difficulties managing relationships with the big retailers Macey’s and J. C. Penney. William McComb took over as CEO in 2006 and in early 2007 the stock price peaked at $46.64, but there were clear problems to deal with: an ageing set of customers and a headline brand that was in decline. Things then got much worse with losses starting in the last quarter of 2007 and continuing right through to 2012. The stock price dropped dramatically, going lower than $2 in 2009 at a point when the company was laying off workers and closing distribution centres. In 2011 the company made a substantial loss (of $172 million) but the share price continued its slow recovery and in 2012 (as this is being written) there is talk of a possible buyout at $20 a share.

Meanwhile the company has for a long time had a credit rating from Standard and Poors. Looking at the history since 2000 we can see that Liz Claiborne Inc was rated at investment grade (BBB in S&P terminology) until 3 June 2008, when it was cut to BB which is below investment level (commonly called ‘junk’). At the time S&P cited significantly higher debt levels and a challenging retail environment for this regrading. Then on 17 August 2009 Liz Claiborne’s rating was cut again to B, and in March 2010 it was cut to CC. This rating means that S&P regards the company as ‘highly vulnerable’ and is an enormous red flag. The final indignity occurred on 11 April 2011 when the company was judged to have made a selective default. In some senses when the S&P team chose to regard a particular tender offer refinancing as “distressed” and equivalent to a selective default, this was a technical decision, and a long way from a standard default on debt. In fact the very next day the company was re-rated back to a B. Since then the company (which is itself rebranding as Fifth & Pacific Cos.) is in much better shape: some brands have been sold to pay back debt and the company has concentrated on its three main brands of Kate Spade, Juicy Couture and Lucky Brand.

For a company like Liz Claiborne that needs to borrow money, credit ratings are essential. The aim of the credit rating agencies is to give an indication of the chances that the company will be unable to meet repayments that are due. But to what extent does a credit rating give information that is different to the stock market valuation? And how reliable are the ratings?

10.1 Introduction to Credit Risk

In this chapter we will give an introduction to credit risk, spending most time on consumer credit, since this is the area most likely to be relevant to managers working outside the financial sector. Of course we may well have an interest in credit risk as it affects us as individuals, since our own circumstances and credit history will influence our ability to obtain credit and the interest rates that we pay.

As we discussed in Chapter 1, credit risk refers to the possibility that a legally enforceable contract may become worthless (or at least substantially reduced in value) because the counterparty defaults and goes out of business. In the case of an individual, an outstanding debt may be uncollectable even without the individual concerned becoming bankrupt, and this too would be classified as credit risk.

Two aspects of credit risk will drive our discussion in this chapter. First credit risk has a yes-no characteristic, with either a default or not. Discussion of the tails of distributions and a range of possible outcomes are no longer very relevant. Second credit risk is about what happens to another party and so we have less information than for our own businesses. We need to work harder to get the maximum value out of whatever information we do have. We will want to look at what happens over time (Is a company moving to a less stable situation?) and we will want to make use of any indirect pointers that are available (Is the fact that this individual has taken out 4 new credit cards over the last six months a bad sign?).

The Basel II framework is designed for banks. In this context credit risk will relate to a large number of different kinds of loans - both loans to businesses and loans to individuals. A large component of the lending to individuals is through mortgages. In this case the value of the property itself provides some security for the bank, but there can still be a significant credit risk if property prices fall sufficiently far that the outstanding debt becomes larger than the house value (as happened in the US with sub-prime mortgages).

From a corporate perspective credit risk is related to credit ratings that are given by one of three major credit rating agencies: Standard and Poors, Moodys, and Fitch. When entering into a contract with, and especially when lending money to a firm that has a lower rating and hence a higher risk of default, it will be appropriate to pause and think carefully. At the very least it will be wise to ask for a higher rate of interest on loans to firms where there is a higher perceived risk. By limiting the contractual arrangements with firms that have low ratings managers can limit the credit risk they take on.

Each credit rating agency has its own terminology and codes, but the codes for Standard and Poors are as follows:

Investment grade

AAA: The best quality borrowers, reliable and stable (many of them governments).

AA: Very strong capacity to meet financial commitments, a slightly higher risk than AAA.

A: Strong capacity to meet financial commitments, but could be susceptible to bad economic conditions.

BBB: Medium class borrowers, with an adequate capacity to meet commitments. Satisfactory at the moment.

Non-Investment grade

BB: Not vulnerable in the near term but facing major ongoing uncertainties.

B: Currently has the capacity to meet commitments, but vulnerable to adverse conditions.

CCC: Currently vulnerable and dependent on favorable business and economic conditions.

CC: Currently highly vulnerable, very speculative bonds.

C: Virtually bankrupt, but payments of financial commitments are continued.

Moody's system includes codes Aaa, Aa, A, Baa etc., but is broadly similar. The credit ratings agencies failed spectacularly at the start of the global financial crisis where they continued to rate certain CDOs (collateralized debt obligation) at the highest level shortly before they were announced to be 'toxic'. However given that none of the agencies did conspicuously better than the others, and given the important role that the agencies play in the financial system, they have continued much as before.

10.2 Using credit scores for credit risk

Credit scores are drawn up with the express intention of giving guidance on the risk of a company defaulting on debt. The precise methodology that is used will vary: when assessing the risk associated with government bonds the agency will take a different approach than when making a corporate rating. For corporate risk the process is roughly as follows. A company that wishes to borrow money by issuing bonds will ask one of the ratings agencies to give a rating, and it will pay for this service. In fact the two largest agencies, Standard and Poors and Moodys will rate all large corporate bonds issued in the U.S., whether asked to or not. The issuer may in any case wish to pay the fee and embark on a more serious engagement with the rating agency to avoid a situation where the rating takes place without complete knowledge of the facts. In fact institutional investors will prefer to have the bond issue rated by more than one agency.

The rating agency will consider two things: business risk and financial risk. Assessing business risk involves considering trading conditions, the outlook for the industry and the quality of the management team. Financial risk is assessed in a more quantitative way using accounting data looking at profitability, debt levels, and the financial strength and flexibility of the firm. Following a visit to the company by the analysts involved, a committee will meet to consider the analysts' recommendations and vote on the final grade assigned. The company is informed of the recommendation and the commentary on it that will be published and then given a final chance to provide additional information (or to ask for company confidential information to be removed from the commentary) before the rating is confirmed and published. If a company is issuing bonds with different characteristics then these are rated with reference to the overall company rating, though senior debt, with priority rights

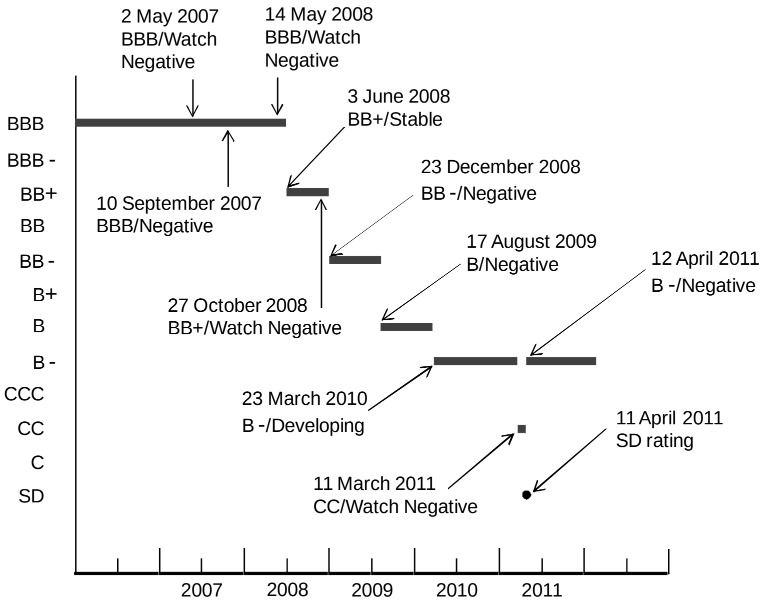


Figure 10.1 Timeline of announcements by Standard and Poors of credit ratings for Liz Claiborne Inc

for payment, is likely to be given a notch higher rating than subsidiary debt. In the case of highly structured products like CDOs, where debt from various sources has been combined and sliced up, then the process is more complex, since the overall level of risk involves looking at the constituent parts of the structured product.

At that point the process moves to a surveillance mode where a separate team within the ratings agency are tasked with keeping an eye on the developments within the company in order that there can be a timely change in the rating (either up or down) if circumstances warrant it. The actual ratings are issued with plus or minus signs attached for all the grades between AA and CCC. Moreover Standard and Poors give what they describe as an outlook statement indicating the direction of change if they anticipate a change as likely in the next one to two years (with a status of ‘developing’ if a change is likely but it could be in either direction). If the company enters a phase where the agency believes there is a significant chance of change in a shorter time frame of three months or so, then the company is placed on ‘credit watch’. We can see how this played out in the Liz Claiborne example in the time line shown in Figure 10.1. Notice that there were announcements in May 2007, September 2007 and May 2008 that reflected the possibility of a drop in rating as problems piled up for the company, but the first downgrade did not occur till June 2008.

Credit rating agencies perform a vital task in the market for capital. They have expertise in the task of evaluating risk and are independent of individual companies. The additional information that they provide to investors and market participants will ultimately mean that companies can raise money at lower costs since they provide an efficient way for investors

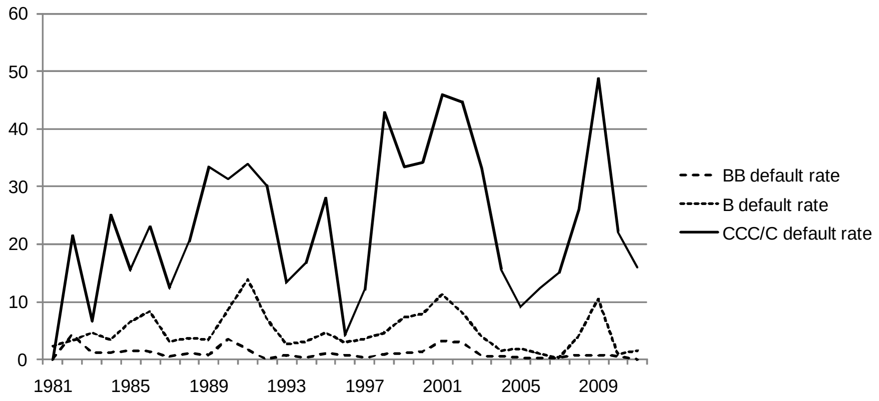


Figure 10.2 The percentage of companies defaulting during the year for three different (non-investment) grades

to reduce their uncertainty. At the same time they have an important function for regulators who want to limit the risk that financial companies can take and can use ratings within the rules they set up.

The probability of a default obviously varies with overall macro-economic factors: the global credit crunch of 2008–2009 led to many more defaults occurring. The ratings agencies do not set out to give ratings aligned with particular probabilities of default, since this would require wholesale re-ratings as the economic climate changes from year to year. Instead ratings agencies are concerned with relative probabilities: an investment grade company is less likely to default than one with a grade of BB, which in turn is less likely to default than a company with grade B and so on. Figure 10.2 shows how the percentage of defaults per year varies over time for the three non-investment grades given by Standard and Poors. As we would expect the default rates from C grades are higher than from B grades which in turn are higher than from BB grades. But it is also interesting to see how much variation there is. In years like 2005 or 1996 even a C grade bond had less than a 10% chance of default, but in 2001 and 2009 entering the year with a C grade would mean more than a 45% chance of default before the end of the year.

10.2.1 A Markov chain analysis of defaults

In order to understand the risks associated with a particular grade we must specify the time horizon involved. There is very little chance that an AA rated company will default this year, but in 5 years time that same company may have slipped to a BB rating and with that the chance of default will have increased substantially. Standard and Poors publish a report that includes the probabilities of making the various transitions that are possible. Table 10.1 shows the average transition rates over a thirty year period. Thus for example on average 8.7 % of firms rated AAA drop to AA during the course of a year. The table includes a column NR for not rated. Sometimes a firm will drop out of the ratings shortly before defaulting, but equally

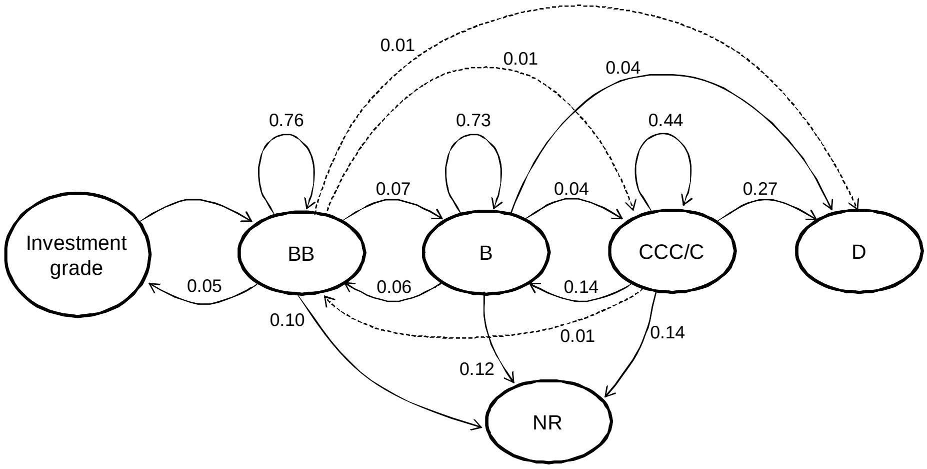


Figure 10.3 Annual probabilities of transitions between non-investment grades. Transitions with probability less than 0.01 are not shown, and transitions with probability less than 0.02 are shown dashed.

it may simply be a decision by the company that a rating is no longer necessary. The rating agency will track companies that drop out mid year and then default during that year, so that the figures on annual defaults are correct.

Table 10.1. Global corporate average transition rates for the period 1981-2011 (%)
One year

	AAA	AA	A	BBB	BB	B	CCC/C	D	NR
AAA	87.2	8.7	0.5	0.1	0.1	0	0.1	0	3.4
AA	0.6	86.3	8.3	0.5	0.1	0.1	0	0	4.1
A	0	1.9	87.3	5.4	0.4	0.2	0	0.1	4.7
BBB	0	0	3.6	84.9	3.9	0.6	0.2	0.2	6.4
BB	0	0	0.2	5.2	75.9	7.2	0.8	0.9	9.8
B	0	0	0.1	0.2	5.6	73.4	4.4	4.5	11.7
CCC/C	0	0	0.2	0.3	0.8	13.7	43.9	26.8	14.4

Sources: Standard & Poor's Global Fixed Income Research and Standard & Poor's CreditPro®.

It is confusing to draw a diagram showing all the transitions that are possible – in Figure 10.3 we have shown just the annual transitions between the 'non-investment' grades. The transitions with a probability of less than 0.01 have been omitted and those with a probability of less than 0.02 are shown dashed.

Given information on the chances of having a default or making a transition in one year, what does this imply about the chance of default over a three or five year time horizon? One simple way to approach this problem is to analyze a Markov chain model of the process. The Markov assumption is that changes in rating only depend on the current score. A company that has been AA for 15 years is no more or less likely to move to A than a company that only achieved AA status last year. And a company that in successive years has dropped from A to BBB to BB is no more or less likely to move down again to B than a company that has moved in the opposite direction, having been rated B last year but just moved up to BB.

One of the basic facts of a Markov chain is that the n'th power of the transition matrix gives the probability of making a transition in n steps. To illustrate this lets try to analyze the probability that a company starting at state B will have a default over a two year period using the information in Figure 10.3. We will write $p(X,Y)$ for the transition probability from some state X to a state Y. Then we can look at all the possible two step paths from B to D. We can move there in one year; or we can stay at B in the first year and move to D in the second year; or we can move to C in the first year and move from C to D in the following year, and so on. We get

$$\begin{aligned}
&\Pr(B \text{ to } D \text{ in 2 years}) \\
&= p(B,D) + p(B,B)p(B,D) + p(B,C)p(C,D) + p(B,BB)p(BB,C) \qquad (10.1)\\
&= 0.04 + 0.73 \times 0.04 + 0.04 \times 0.27 + 0.06 \times 0.01 = 0.08.
\end{aligned}$$

Since we are interested in the probability of a default at some time in the next two years once we reach D the calculations can finish. This makes it sensible to define $p(D,D) = 1$ and then we can see that the formula has the form

$$\Pr(x \text{ to } y \text{ in 2 steps}) = \sum_i p(x,i)p(i,y), \qquad (10.2)$$

where we sum over i being any of the states we can get to. In the equation (10.1) we have left out the terms like $p(B,NR)$ where the probability on the second step is zero. Suppose we take P to be the matrix shown in Table 10.1 augmented by two rows for D and NR where in each case the probability of staying at the same state is 100%. Thus P is a square matrix and it will be convenient to divide each element by 100 to express the transition probabilities directly rather than as percentages. Then we can write the element in the i'th row and j'th column as p_{ij}: it is the probability of making a jump from stat i to state j. The rules of matrix multiplication tell us that the element in the i'th row and j'th column of $P \times P = P^2$ is given by $\sum_k p_{ik}p_{kj}$ which exactly matches the formula (10.2). Hence P^2 simply gives all the two step transition probabilities, and P^3 gives the three step transition probabilities and so on.

Since Standard and Poors also report on the three year transitions we can compare our default rate predictions using the Markov chain model with actual behavior. The full Markov chain comparisons are given in the spreadsheet BRMch10-Markov with the array function MMULT used to carry out the matrix product. Figure 10.4 compares the actual and predicted three and five year default rates. We can see from this that the Markov model does not do a great job of predicting these rates. For example starting at a BB grade the Markov assumption predicts 3 and 5 year rates of 3.5% and 4.9% respectively, while the actual figures reported by Standard and Poors are significantly higher at 5.0% and 9.2% respectively.

The Markov chain model is a good way to think about credit rating movements, but is a big simplification of the real behavior. We can identify a number of possible reasons for the poor three and five year predictions from the Markov model.

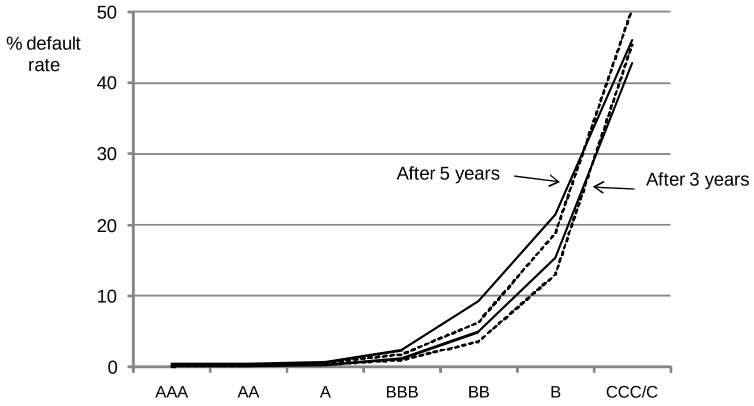


Figure 10.4 Actual (solid line) and predicted (dashed line) default rates from different starting grades over three and five years

Grouping together of states. Actual ratings are given with plus and minus signs giving more states in total. When a Markov chain has states grouped together it no longer behaves as a Markov chain. For example if companies typically move slowly through ratings from say $BB+$ to BB to $BB-$ then knowing that the company has been in the grouping BB for some time may increase the chance of it being at the lower level $BB-$, thus increasing the chance of a jump to level B and breaking the Markov assumption.

Different types of companies behave differently. Suppose that different types of companies exist and they follow different Markov chains. This doesn't necessarily mean that one type of company is more risky than another (at the same rating); perhaps it is characteristic of companies in the financial sector to have a higher risk of immediate default and that is balanced by companies in the non-financial sector having a higher risk of moving to the next notch down. In any case the existence of different types of company can mess up the Markov assumption. Exercise 10.2 gives an example of this.

Bad years cluster together. We have already observed that default rates (and more generally downgrades) vary substantially from year to year. This means that a more appropriate model might be one in which the probability of any transition depends on the year in question. We should then replace equations like (10.2) with

$$\Pr(x \text{ at time } n \text{ to } y \text{ at time } n+2) = \sum_i p_n(x,i)p_{n+1}(i,y)$$

where the subscript represents the time period. If the probability of a downgrade in one year is positively correlated with the probability of a downgrade in the next, then this can increase the overall probability of two successive downgrades when compared with the case when probabilities in one year are independent of those in the next year.

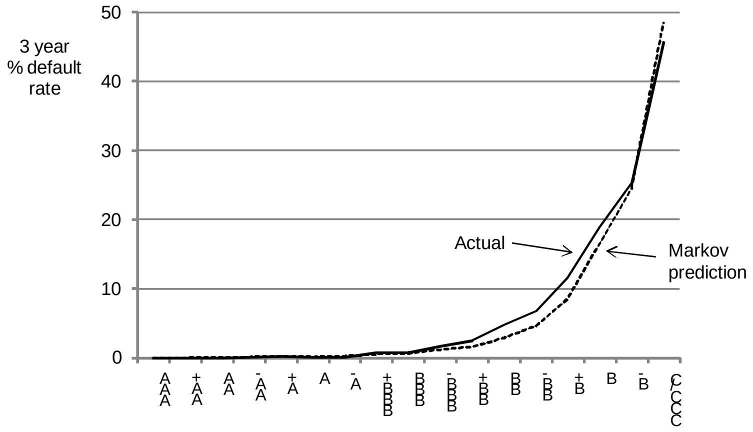


Figure 10.5 Three-year default predictions for non-financial companies compared with actual average figures.

Subjective decisions by the ratings agency. There has been much discussion of the extent to which subjective factors may play a role in credit ratings. Clearly the agencies themselves believe that their ratings are objective, but questions have often been raised as to whether initial ratings may be too generous since there is an unconscious desire by the agency to win the business of new debt issuers. Also there may be a reluctance to issue a downgrade too quickly, with the agency waiting until it is certain that it is justified. This might be particularly the case when the issuer's debt conditions are tied to the grading; then a downgrade from BBB to BB could trigger a requirement for faster debt repayment, and this in turn could cause the company to get into further difficulties. In the other direction, agencies may be deliberately conservative in their ratings, holding back on an upgrade that would be appropriate. This type of behavior by the agencies could explain a misleading forecast from the Markov assumption.

We can at least partially address the first two issues by using Standard and Poors data broken down into more exact ratings and distinguishing between non-financial, insurance and financial companies. The second spreadsheet in the workbook BRMch10-Markov compares the results for three year default rates arising from a Markov model restricted to non-financial firms and the actually observed rates . The results are shown in Figure 10.5 and it seems that the Markov predictions are a little more accurate, but still involve substantial errors, for example the Markov model predicts a 3 year default rate starting at BB as 2.8% whereas the actual observed figure is 4.7%.

10.3 Consumer credit

In the remainder of this chapter we will focus on consumer credit. Here the rating is made for an individual rather than a company. The first large scale applications of automated ways of assessing credit were driven by the credit card explosion that happened in the 1960s and 70s.

This led to lenders looking at credit histories and credit bureaus were set up with the aim of pooling data from different lenders so that a consumer who failed to pay off a store card, for example, would find that information was made available to other potential lenders (perhaps a car finance company).

Different credit bureaus (sometimes called credit reference agencies) operate in different countries: the largest such company in Australia is Veda Advantage. Credit bureaus collect together credit information from various sources and can provide a credit report on an individual when requested. Around the world millions of these are issued every day (in the US alone more than 2 million a day) and the process is simply an automated interrogation of a database. The data held by credit bureaus varies from country to country and is affected by data protection laws. In the US a huge amount of information is kept while in some European countries it is limited to mainly publicly available information (for example court records of bankruptcy). Usually a debt remains on the record only until it is repaid, or until a specified time limit has passed (perhaps 10 years). In many countries including Australia consumers have the right to receive a free copy of their own credit record.

One widely used technique is to look at applicants for loans and try to judge on the basis of their characteristics how likely they are to default. The measure of default that is traditionally used is the likelihood that an applicant will go 90 days overdue on their payments within the next 12 months. However it does not matter so much what this measure is: in the end a ranking of individuals occurs and the least attractive (that is the most likely to default) are refused credit or funneled into a different type of credit arrangement.

The credit scoring method is straightforward. A lender is interested in whether or not to extend credit to an individual and in order to make this decision a number of variables are checked (such as age, and whether the individual rents or owns their home). These variables are used to produce a "score" and this is used to predict the likelihood that the individual defaults on the loan. The scoring method is to use a scorecard which simply adds together score components for various attribute that the individual possesses. For example the fact that the credit applicant has lived at the same address for more than 5 years might be worth 15 points. The development of this scorecard is based on the credit histories of many thousands of other people. <*Check US and UK versions*>If you are interested in how your own score might come out you can look at the website http://www.checkmyfile.com.au/ which offers a free test calculation.

The objective of the lender is simply to make this decision as accurately as possible. Any information which can legally be used and which has a bearing on the credit worthiness of an individual will come into play What is illegal? It is not permitted to discriminate on the basis of race, gender, sexuality or religion, so these questions cannot be asked. However it is fine to consider an individual's postcode when carrying out the check. The rules on age and marital status are more complex and the law in Australia <*Check US and UK versions*> limits the extent to which credit can be denied by a bank on the basis of age or marital status.

10.3.1 Probability, odds and log-odds

For a particular type of individual we can use previous data to predict the probability that they are '*good*', i.e. that they will repay the loan. Write p_i for this probability. An alternative is to look at the odds of being 'good' as opposed to 'bad'. This might be familiar from a betting context: if we say that the odds of a horse winning are 2 to 1 then it is a statement that the horse is twice as likely to win as not. The odds o_i are simply the probability of being good divided by the probability of being bad, i.e.

$$o_i = \frac{p_i}{1 - p_i}.$$

The probability of being good varies between 0 and 1. But the odds of being good can vary from 0 to any positive value.. It can also be useful to look at the log odds defined as

$$\log_e(o_i) = \log_e \left(\frac{p_i}{1 - p_i} \right).$$

Log odds can take any value both positive and negative. Notice that both odds and log odds are increasing functions of the probability.

Suppose that we are considering extending credit to an individual. The profit from this credit may be the profit we make on a product that will not be sold unless we offer credit. There is however a loss, usually much larger than the profit, that will accrue if the individual does not repay the debt. Suppose we write L for the loss and R for the profit. Then the expected value to us of the loan is

$$p_i R - (1 - p_i) L.$$

This is positive (meaning we should go ahead with the loan) if

$$p_i > \frac{L}{L + R},$$

or equivalently if $o_i > (L/R)$, a condition that can also be written in terms of the log odds:

$$\log_e(o_i) > \log_e \left(\frac{L}{R} \right).$$

Consider a hypothetical example which we will call Bank of Sydney. The data for this example are given in the spreadsheet BRMch10-BankofSydney. There are 1200 customers who have been loaned money on a short term basis and most have kept up with loan repayments (these are called ‘good’ or G. Some have not and anyone who falls a total of more than 90 days behind in repayments is classified as ‘bad’ or B. Besides data on age at time of loan agreement classified as above, Bank of Sydney keeps data on whether the individuals were owners, renters or some other classification in respect to their home and also whether they have a credit card or not.

The Bank of Sydney data can be presented in various ways. Table 10.2 shows the number of good and bad individuals in each of the 24 different sub-categories obtained from ‘credit card status (2)’ × ‘housing status (3)’ × ‘age bracket (4)’

Table 10.2: Bank of Sydney data: Goods and Bads

	under 30	30-39	40-49	over 50
Owner with credit card	$G=59$ $B=5$	$G=111$ $B=9$	$G=118$ $B=5$	$G=232$ $B=11$
Renter with credit card	$G=47$ $B=10$	$G=16$ $B=5$	$G=22$ $B=4$	$G=64$ $B=18$
Other with credit card	$G=63$ $B=6$	$G=21$ $B=2$	$G=16$ $B=3$	$G=91$ $B=5$
Owner without credit card	$G=19$ $B=2$	$G=13$ $B=3$	$G=44$ $B=4$	$G=31$ $B=2$
Renter without credit card	$G=18$ $B=9$	$G=14$ $B=10$	$G=5$ $B=1$	$G=10$ $B=3$
Other without credit card	$G=12$ $B=2$	$G=26$ $B=8$	$G=6$ $B=2$	$G=12$ $B=1$

In this data there are a total of 1070 Goods and 130 Bads.

We define the *population odds* as

$$o_{\text{Pop}} = \Pr(G)/\Pr(B).$$

This is the odds of an individual being good if they are chosen at random from the whole population. For the Bank of Sydney the population odds are $o_{\text{Pop}} = 1070/130 = 8.23$.

If we look at a single category of individual we find that the odds vary from this. For example given that the individual is aged under 30 the first column in the Bank of Sydney data shows that the odds are

$$\frac{59+47+63+19+18+12}{5+10+6+2+9+2} = \frac{218}{34} = 6.412.$$

So borrowers in this category are much more likely to be bad than the population as a whole.

One way to think about this is to observe that out of 1070 good individuals, 218 are in this age bracket and out of 130 bad individuals 34 are in this age bracket. In other words

$$\Pr(\text{age} < 30|G) = 218/1070,$$

$$\Pr(\text{age} < 30|B) = 34/130.$$

This leads to the definition of the *information odds,* I_A, for a category A as

$$I_A = \frac{\Pr(A|G)}{\Pr(A|B)}.$$

In this case the information odds for ‘age under 30’ is

$$I_{age<30} = \frac{218/1070}{34/130} = \frac{218}{34} \times \frac{130}{1070} = 0.779.$$

Clearly multiplying the information odds for a category by the population odds gives the odds for the category, i.e.

$$o_A = \frac{\Pr(G|A)}{\Pr(B|A)} = \frac{\Pr(A|G)}{\Pr(A|B)} \times \frac{\Pr(G)}{\Pr(B)} = I_A \times o_{\text{Pop}}.$$

We have used Bayes rule here $\Pr(G|A) = \Pr(A|G)\Pr(G)/\Pr(A)$ and cancelled the two terms $\Pr(A)$. Thus the information odds provides a kind of modifier for the population odds to get to the odds in a particular category.

The *Weight of Evidence* (WoE) for a category is just the natural logarithm of the information odds for that category:

$$w_A = \ln(I_A) = \ln\left(\frac{\Pr(A|G)}{\Pr(A|B)}\right),$$

so the weight of evidence for ‘age under 30’ is

$$w_{age<30} = \ln\left(\frac{218}{34} \times \frac{130}{1070}\right) = \ln(0.779) = -0.2497.$$

It is natural to make this definition because it enables us to find the log odds for a category simply by adding the weight of evidence and the log odds for the population:

$$\begin{aligned}\ln(o_A) &= \ln(I_A \times o_{\text{Pop}}) = \ln(I_A) + \ln(o_{\text{Pop}})\\ &= w_A + \ln(o_{\text{Pop}}).\end{aligned}$$

If the behavior under different attributes is independent, then the weights of evidence can be added together for different attributes to find the log odds for an individual with a number of different attributes. This is a slightly surprising result. To see why it is true consider an individual with two attributes A_1 and A_2.

We wish to calculate the odds that this individual is good, i.e. we want to find $\Pr(G|A_1, A_2)/\Pr(B|A_1, A_2)$. We know from Bayes rule that,

$$\Pr(G \mid A_1, A_2) = \Pr(A_1, A_2 \mid G)\frac{\Pr(G)}{\Pr(A_1, A_2)}.$$

If A_1 and A_2 are independent (so that information about one attribute does not tell us anything about the other) then we have $\Pr(A_1, A_2 \mid B) = \Pr(A_1|B) \times \Pr(A_2|B)$ and $\Pr(A_1, A_2 \mid G) = \Pr(A_1|G) \times \Pr(A_2|G)$. If this holds then the odds for an individual in the category of A_1 and A_2 are

$$\begin{aligned}\frac{\Pr(G \mid A_1, A_2)}{\Pr(B \mid A_1, A_2)} &= \frac{\Pr(A_1, A_2 \mid G)}{\Pr(A_1, A_2 \mid B)}\frac{\Pr(G)}{\Pr(B)}\\ &= \frac{\Pr(A_1|G)}{\Pr(A_1|B)} \times \frac{\Pr(A_2|G)}{\Pr(A_2|B)} \times \frac{\Pr(G)}{\Pr(B)}.\end{aligned}$$

The same expression can be extended to any number of terms. So the log odds given $A_1, A_2, ..., A_n$ is

$$\ln\left(\frac{\Pr(G \mid A_1, A_2, ..., A_n)}{\Pr(B \mid A_1, A_2, ..., A_n)}\right) = w_1 + w_2 + ... + w_n + \ln(o_{\text{Pop}}),$$

where o_{Pop} is the population odds and w_i is the weight of evidence for the characteristic A_i.

A scorecard has a very simple structure: it associates every category with a score and then adds the scores for an individual together to get a final score. For example we might have a scorecard as follows:

Attribute	
Age < 30	-25
Age $30 - 39$	-42
Age $40 - 49$	30
Age ≥ 50	29
Owns home	62
Rents home	-92
Other	3
Has credit card	23
No credit card	-61
Constant	211

Then the score for an individual who is an owner, with a credit card and is age 42 is $30 + 62 + 23 + 211 = 326$, whereas a 28 year-old who lives at home and has no credit card has a score $-25 + 3 - 61 + 211 = 128$.

This particular scorecard has been calculated from the Bank of Sydney data using what can be called a naive Bayes approach - we have simply taken the weight of evidence for each category and multiplied by 100. The constant term is the log of the population odds. We can check how effective this is by using the scorecard to predict the odds that a 42 year old homeowner with a credit card is ‘good’. Since the score for this individual is 326, the predicted log odds is 3.26. Hence the odds are predicted to be

$$e^{3.26} = 26.05.$$

In this category there are 118 good individuals and 5 bad so the actual odds are 23.6 to 1. In the same way we can look at the odds of a 28 year-old who lives at home and has no credit card being a ‘good’. The score is 128 corresponding to a log odds of 1.28 which means that the odds are predicted to be

$$e^{1.28} = 3.60.$$

The actual results in this category are 19 good individuals and 2 bad individuals, giving odds of 9.5 to 1. Thus though the prediction in the first case is reasonable, in the second case the predictions is not very good. Essentially the assumption on the independence of attributes is too strong for this data. A better way to make these prediction is to use Logistic Regression, which is the subject of the next section.

10.4 Logistic Regression

In this section we will show how logistic regression works: this is the most common way to construct a scorecard. But we start with the fundamental question of how we should predict probabilities. Ordinary regression predicts a dependent variable y on the basis of the observation of a set of explanatory variables x_j, $j = 1, 2, ..., m$. The linear form of this

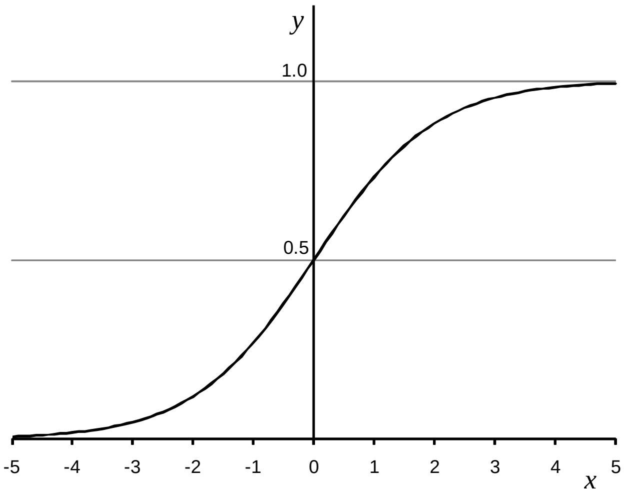


Figure 10.6 A graph of the logistic function $y = e^x/(1 + e^x)$.

prediction can be written

$$\begin{aligned} y &= \beta_0 + \sum \beta_i x_i \\ &= \boldsymbol{\beta} \cdot \mathbf{x} \end{aligned}$$

where we use a boldface letter to indicate a vector: $\mathbf{x}$ =$(x_0, x_1..., x_m)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)$ and we set $x_0 = 1$. We then estimate the β's by looking at previous data to find a good fit. We cannot use the same approach here because we want to ensure that the predictions y are all between 0 and 1. The logistic (or logit) approach is to use a non linear transformation to get from x_i values to a probability prediction p. We set

$$p = \frac{e^{\boldsymbol{\beta}\cdot\mathbf{x}}}{1 + e^{\boldsymbol{\beta}\cdot\mathbf{x}}}$$

If $\boldsymbol{\beta} \cdot \mathbf{x}$ gets very large then p approaches 1, if $\boldsymbol{\beta} \cdot \mathbf{x} = \mathbf{0}$ then $p = 1/2$ and if $\boldsymbol{\beta} \cdot \mathbf{x}$ is a large negative number then p approaches 0. Figure 10.6 shows what this function looks like (and there are other functions we might choose which would produce similar results.)

One advantage of the logistic function is that it fits beautifully with our previous definition of log odds. We have:

$$1 - p = 1 - \frac{e^{\boldsymbol{\beta}\cdot\mathbf{x}}}{1 + e^{\boldsymbol{\beta}\cdot\mathbf{x}}} = \frac{1}{1 + e^{\boldsymbol{\beta}\cdot\mathbf{x}}}.$$

So

$$\text{odds } = \frac{p}{1-p} = \frac{e^{\boldsymbol{\beta}\cdot\mathbf{x}}}{1 + e^{\boldsymbol{\beta}\cdot\mathbf{x}}} \frac{1 + e^{\boldsymbol{\beta}\cdot\mathbf{x}}}{1} = e^{\boldsymbol{\beta}\cdot\mathbf{x}},$$

and thus

$$\log_e \left(\frac{p}{1-p} \right) = \log_e (\text{odds}) = \log_e \left(e^{\boldsymbol{\beta}\cdot\mathbf{x}} \right) = \boldsymbol{\beta} \cdot \mathbf{x}.$$

Thus if the actual probabilities of ‘good’ or ‘bad’ are given by a logistic model derived from an underlying linear function $\boldsymbol{\beta} \cdot \mathbf{x}$, then the log odds become a linear function of the observations x_i.

How can we estimate the values $\beta_0, \beta_1, ..., \beta_m$? In an ordinary regression we try to minimize the error term looking at the prediction against what actually happened. But for our logit model the prediction is a probability and what actually happened is either a good result or a bad result (which we can think of it as either a 1 or a 0). In this context it is more natural to use a maximum likelihood estimator. So we choose the values of β which maximize the probability of our set of observations (which are all 0 or 1).

To see how to work this out, we start from one observation say y_1 with a prediction of its probability of p_1. So the likelihood of observing y_1 is p_1 when $y_1 = 1$ and $(1 - p_1)$ when $y_1 = 0$ which we can see is

$$(p_1)^{y_1}(1 - p_1)^{(1-y_1)}.$$

With two observations y_1, y_2 with respective probabilities p_1, p_2 the probability of observing say $y_1 = 1$ and $y_2 = 0$ is $p_1(1 - p_2)$ or

$$(p_1)^{y_1}(1 - p_1)^{(1-y_1)}(p_2)^{y_2}(1 - p_2)^{(1-y_2)}.$$

More generally we can take n observations and build up a big product having $2n$ terms:

$$\prod_{i=1}^{n}(p_i)^{y_i}(1 - p_i)^{(1-y_i)}. \tag{10.3}$$

This is the probability of observing $y_1, y_2, ...y_n$ if the probabilities are really $p_1, p_2...p_n$.

If the p_i come from a logit model then

$$p_i = \frac{e^{\boldsymbol{\beta} \cdot \mathbf{x}^i}}{1 + e^{\boldsymbol{\beta} \cdot \mathbf{x}^i}}, \tag{10.4}$$

where we have written $x^i = (x^i_0, x^i_1..., x^i_m)$ for the independent variables associated with the i’th observation. So the problem of finding a maximum likelihood estimator from a given set of data $(y_1, y_2, ...y_n)$ is equivalent to finding a set of β values that maximizes the product expression (10.3) where the p_i are given by (10.4).

But if we want to maximize an expression we can also maximize the log of that expression. So rather than maximize the likelihood, we maximize the log likelihood

$$\begin{aligned} L &= \log\left(\prod_{i=1}^{n}(p_i)^{y_i}(1 - p_i)^{(1-y_i)}\right) \\ &= \sum_{i=1}^{n} y_i \log(p_i) + \sum_{i=1}^{n}(1 - y_i)\log(1 - p_i). \end{aligned}$$

Since the p_i values are given by logistic functions, this can be simplified considerably:

$$\begin{aligned}
L &= \sum_{i=1}^{n} y_i \log\left(\frac{e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}}{1+e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}}\right) + \sum_{i=1}^{n}(1-y_i)\log\left(\frac{1}{1+e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}}\right) \\
&= \sum_{i=1}^{n} y_i \left(\log\left(e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}\right) - \log(1+e^{\boldsymbol{\beta}\cdot\mathbf{x}^i})\right) - \sum_{i=1}^{n}(1-y_i)\log\left(1+e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}\right) \\
&= \sum_{i=1}^{n} y_i \log\left(e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}\right) - \sum_{i=1}^{n}\log\left(1+e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}\right) \\
&= \sum_{i=1}^{n} y_i \left(\boldsymbol{\beta}\cdot\mathbf{x}^i\right) - \sum_{i=1}^{n}\log\left(1+e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}\right)
\end{aligned}$$

To maximize L we take derivatives with respect to β_j. Now, since

$$\frac{d}{dz}\log(f(z)) = \frac{f'(z)}{f(z)} \text{ and } \frac{d}{dz}e^{za+b} = ae^{za+b},$$

we get

$$\begin{aligned}
\frac{\partial L}{\partial \beta_j} &= \sum_{i=1}^{n} y_i x_j^i - \sum_{i=1}^{n}\frac{x_j^i e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}}{1+e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}},\ j=1,2,...,m, \\
\frac{\partial L}{\partial \beta_0} &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\frac{e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}}{1+e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}}.
\end{aligned}$$

To find the maximum we set all these derivatives to zero and solve the simultaneous equations

$$\begin{aligned}
\sum_{i=1}^{n} x_j^i\left(y_i - \frac{e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}}{1+e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}}\right) &= 0,\ j=1,2,...,m, \\
\sum_{i=1}^{n}\left(y_i - \frac{e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}}{1+e^{\boldsymbol{\beta}\cdot\mathbf{x}^i}}\right) &= 0.
\end{aligned}$$

These are $m+1$ non-linear equations to be solved for the $m+1$ values $\beta_0, \beta_1, ..., \beta_n$. It is not difficult to solve the equations using some iterative method (like Newton-Raphson). It is this solution that is provided by software carrying out logistic regression: the estimate of the β values is the one that maximizes the chance of observing the pattern of 1's and 0's that are observed.

10.5 Scorecards and categorical data

So far our discussion of logistic regression has not made any assumptions about the form of the explanatory variables x. In practice the most important case is when the explanatory variables are categorical. Sometimes this follows from the nature of the data (for example does an individual rent their home or not). But even when this is not true most credit scoring creates categorical data by assigning individuals to categories. The most obvious example is a variable like age: this is most naturally treated as a continuous variable but credit scoring would normally determine certain age brackets and assign individuals to just one of these. The same holds true for income levels.

In this situation we use dummy variables for each category. So for example if we are using age brackets: (A) less than 30, (B) 30 to 39, (C) 40 to 49,(D) 50 or more, then an individual of age 33 has $x_{ageA} = 0$, $x_{ageB} = 1$, $x_{ageC} = 0$, $x_{ageD} = 0$. Since $x_{ageA} = 1 - x_{ageB} - x_{ageC} - x_{ageD}$ these categorical variables automatically give rise to a problem of 'collinearity' in the regression. There is no extra information in the last category.

With such a categorization of the explanatory variables it is normal to have a number of individuals in each of the possible categories. To illustrate this consider the Bank of Sydney example we gave earlier.

A logistic regression considers the 1200 individual observations and estimates the β values associated with the six variables $x_{own} =$'owner', $x_{rent} =$'renter', $x_{cc} =$'credit card holder', $x_{ageB} =$'age 30-39', $x_{ageC} =$'age 40-49' and $x_{ageD} =$'age over 50'. The three other variables 'other', 'no credit card' and 'age under 30' are linear combinations of other explanatory variables and do not add anything to the regression, so they are left out.

All the individuals in the top left box of the table have $x_{own} = 1$, $x_{rent} = 0$, $x_{cc} = 1$, $x_{ageB} = 0$, $x_{ageC} = 0$ and $x_{ageD} = 0$. Thus for each of these individuals the scalar product $\boldsymbol{\beta} \cdot \mathbf{x} = \beta_{own} + \beta_{cc}$ which is also the log odds. The predicted probability of being good for these 64 individuals is

$$p = \frac{e^{\beta_{own}+\beta_{cc}}}{1 + e^{\beta_{own}+\beta_{cc}}}.$$

We can present the same Bank of Sydney data either in terms of odds or log odds and this is done in Tables 10.3 and 10.4.

Table 10.3: Bank of Sydney data: Odds

	under 30	30-39	40-49	over 50
Owner with credit card	11.80	12.33	23.60	21.09
Renter with credit card	4.70	3.20	5.50	3.56
Other with credit card	10.50	10.50	5.33	18.20
Owner without credit card	9.50	4.33	11.00	15.50
Renter without credit card	2.00	1.40	5.00	3.33
Other without credit card	6.00	3.25	3.00	12.00

Table 10.4: Bank of Sydney data: Log Odds

	under 30	30-39	40-49	over 50
Owner with credit card	2.47	2.51	3.16	3.05
Renter with credit card	1.55	1.16	1.70	1.27
Other with credit card	2.35	2.35	1.67	2.90
Owner without credit card	2.25	1.47	2.40	2.74
Renter without credit card	0.69	0.34	1.61	1.20
Other without credit card	1.79	1.18	1.10	2.48

This data is also available in the spreadsheet BRMch10-BankofSydney. The second sheet in the workbook contains data that is on an individual level with each of the 1200 individuals having a row in the spreadsheet. The columns here leave out the category 'other' (which can be deduced as applying to someone who is neither an owner or a renter) and the category age-30-or-less (which can be deduced as applying to individuals not being identified as in another age bracket).

The data has been used to run a logistic regression (using the free software ‘Gretl’) and part of the output is shown below.

Model 1: Logit, using observations 1-1200
Dependent variable: Good

	coefficient	std. error
const	1.72256	0.287793
creditcard	0.628808	0.211980
owner	0.493278	0.261849
renter	−0.983520	0.249570
age 30 to 39	−0.369558	0.271993
age 40 to 49	0.134251	0.319367
age 50 or more	0.204348	0.259001

Thus we have the values of the β variables as follows:

$$\begin{aligned}
&\beta_{cc} = 0.629\\
&\beta_{own} = 0.493\\
&\beta_{rent} = -0.984\\
&\beta_{ageB} = -0.370\\
&\beta_{ageC} = 0.134\\
&\beta_{ageD} = 0.204\\
&\beta_0 = 1.723
\end{aligned}$$

where β_0 is the constant term. Thus we have the following maximum likelihood estimator:

$$\begin{aligned}
&\log_e\left(\frac{p}{1-p}\right)\\
&= 1.723 + 0.629x_{cc} + 0.493x_{own} - 0.984x_{rent} - 0.370x_{ageB} + 0.134x_{ageC} + 0.204x_{ageD}.
\end{aligned}$$

We can use these values to produce a log odds table for each of the 24 categories and this is shown in Table 10.5.

Table 10.5: Bank of Sydney data: Logit Predicted Log Odds

	under 30	30-39	40-49	over 50
Owner with credit card	2.84	2.48	2.98	3.05
Renter with credit card	1.37	1.00	1.50	1.57
Other with credit card	2.35	1.98	2.49	2.56
Owner without credit card	2.22	1.85	2.35	2.42
Renter without credit card	0.74	0.37	0.87	0.94
Other without credit card	1.72	1.35	1.86	1.93

Table 10.4 can be compared with Table 10.5 to see how well this approach works. The beauty of logistic regression is that it works well when there are many more types of characteristic than the three in this example (credit card, housing type, age bracket). If we

measure individuals through their answers to say 10 different questions, then even if each question is a simple yes/no we will still end up with more than a 1000 cells. This means that there will be many cells with no individual and many more cells where there are no ‘bad’ individuals. But we still need a way to evaluate the credit risk from a new customer who may belong to one of the cells where we don’t have many previous customers to compare to.

There is an alternative ‘quick and dirty’ approach to the problem. Our model presupposes that the log odds are a linear combination of the explanatory variables, so one option is to estimate this linear combination directly using an ordinary least squares regression to estimate the log odds scores. In practice this method works pretty well and has the advantage that it can be used without access to a data analysis tool that includes a logistic regression component (for example we can use this ordinary least squares regression approach by applying the ‘Data Analysis’ add in that comes with Excel.) If this is done with the Bank of Sydney data we get

$$\log_e\left(\frac{p}{1-p}\right)$$
$$= 1.650 + 0.575x_{cc} + 0.527x_{own} - 0.788x_{rent} - 0.349x_{ageB} + 0.090x_{ageC} + 0.424x_{ageD}.$$

There are some points to be born in mind, however, in using this approach:

- A least squares fit to the log odds will not be equivalent to a direct estimation involving probabilities of individuals being good or bad and so the estimation will not be completely correct.
- The method gives equal weight to all categories, while a logistic regression automatically correctly weights the evidence from categories containing different numbers of individuals. We will expect to see that for classifications in which there are a large number of individuals the log odds predictions from a logistic regression are better than for an ordinary least squares approach. This is true for the Bank of Sydney data where the log odds prediction from the logistic regression is exactly right for the most common type of customer (over 50, home owner with a credit card).
- The least squares regression approach will fail if the log odds cannot be defined for a category because there are no ‘bad’ individuals in that category. In this case the log odds becomes infinite, but there is no problem for logistic regression which simply looks at individual results. Sometimes we add 0.5 to both good and bad numbers to get an adjusted log odds value. So for example if a certain combination has only 10 individuals, all of them ‘Good’ we might replace an infinite log odds value with $\log_e(10.5/0.5) = 3.0445$ (Note that this is still substantially more than the log odds value if 9 of the 10 individuals were ‘good’, which is $\ln(9/1) = 2.1972$).

Building a scorecard

We show how to construct a scorecard using the logistic regression results. Using the same approach we saw earlier with the naive Bayes scorecard we take the regression estimates β, multiply by 100 and round to get the scorecard labelled as Scorecard 1 in Table 10.5. We can make an adjustment to Scorecard 1 by taking a fixed amount off the constant term and adding it to each category in a mutually exclusive and exhaustive set. For example if we take 135 away from the constant term but add 98 to each of owner, renter and other, and if in addition we add 37 to each age category we get Scorecard 2. This gives precisely the same score to each individual as scorecard 1 but has the advantage that all the numbers are positive (which is normally the way that score cards are constructed).

Table 10.5: Adjusting a scorecard to have desirable characteristics

Attribute	Scorecard 1	Scorecard 2	Scorecard 3
Age < 30	0	37	97
Age $30 - 39$	-37	0	70
Age $40 - 49$	13	50	106
Age ≥ 50	20	57	111
Owns home	49	147	107
Rents home	-98	0	0
Other	0	98	71
Has credit card	63	94	68
No credit card	0	0	0
Constant	172	37	0

In both Scorecard 1 and Scorecard 2 a score of 100 corresponds to log odds of 1 which is an odds of $e = 2.72$ and every additional 100 adds 1 to the log odds which is like multiplying the odds by a factor of $e = 2.72$. When constructing a score card we may want to use a linear transformation (multiplying by a constant b and adding another constant c). We can choose the linear transformation (i.e. choose b and c) in order to anchor the score so that a particular score translates into a particular odds and at the same time arrange for a particular increment of score to correspond to multiplying the odds by a certain constant. For example we might decide that a score of 200 should correspond to odds of 10 to 1 and each increment of 50 should correspond to multiplying the odds by a factor of 2.

It is easy to make this change. We take the initial score as s (so $s = 100\ln(\Pr(G)/\Pr(B))$). We have two conditions to satisfy. First at odds of 10 to 1 we have $s = 100\ln(10) = 230.26$ whereas we would like a score of 200. Second, multiplying the odds by 2 will increase s by $100\ln(2) = 69$ whereas we would like to increase the score by 50. So if the new score is $\widetilde{s} = bs + c$, then multiplying the odds by 2 will increase $\widetilde{s}$ by $69b$ and the score at odds of 10 to 1 will be $(230.26)b + c$. For the conditions we require we want

$$(230.26)b + c = 200,$$
$$69b = 50.$$

Hence $b = 50/69 = 0.725$ and $c = 200 - 0.725 \times 230.26 = 33.06$ and we will applying this transformation to Scorecard 2. To achieve the constant added just once to the score we should add it to the constant term and not elsewhere, and multiply all the other scores by 0.725. But to make the scorecard easier to apply we take the additional step of adding the constant term, which is now $37 + 33.06 = 70.06$, to each of the age categories so that we can drop the constant term. After rounding this gives the Scorecard 3 shown in the Table.

Other scoring applications

Having done all this work to determine credit scoring procedures it is worth pointing out that exactly the same techniques can be used in another common management problem which is the targeting of promotions. This is a kind of reverse to the credit scoring problem. We no longer want to identify the people who are likely to be bad in order to avoid giving them credit. Instead we want to pick out the people who are more likely to respond positively to a promotion in order to justify the costs of a mailout targeted to them.

The idea of modelling the behavior in terms of a scorecard built up from different categories is still valuable. Both the credit scoring problem and the promotion targeting problem share the characteristic that quite a small proportion of the sample are in the category of responding (or in the category of a bad debt). From an estimation point of view the consequence is that a small (absolute) number of individuals in any particular combination of categories will end up responding. This is the reason for using an indirect logistic regression procedure rather than simply taking the odds we observe in a single cell in the table and using this to predict the log odds for this cell. Because of the small numbers of responding individuals in some cells, just treating each cell on its own is a poor way to proceed. We are likely to get a better result by using the logistic regression model.

In order to avoid a negative log odds score, we need to deal with odds that are greater than 1. So in this promotion targeting problem we define the odds as:

$$\frac{\text{probability of not responding}}{\text{probability of responding}}$$

With this change everything goes through as before, except that in applying the scorecard we select individuals to mail with low values of the score corresponding to a relatively high probability of responding. Exercise 10.5 is an example of this kind of problem.

Notes

In the discussion of credit ratings agencies I have made extensive use of the information provided by Standard and Poors. Table 10.1 is taken from Table 21 in Standard and Poors 2011 Annual Global Corporate Default Study And Rating Transitions (Publication date: 21 March 2012). Figure 10.2 graphs the changes over time shown in Table 3 of that report. Moreover the spreadsheet BRMch10-credit default.xlsx contains material taken from Tables 21, 59, 61, 62 and 64 of the S&P 2011 Annual Global Corporate Default Study. This information is used to compare the predictions from a Markov assumption and the actual behavior seen which forms the basis for Figures 10.4 and 10.5. The time line for Liz Claiborne given in Figure 10.1 is taken from information given in Table 8 in the Standard and Poors 2011 Default Synopses (Publication date: 21 Mar 2012).

There is much to be said about the way that accounting data can be used to infer credit risk. This work goes back to a well-known model proposed by Altman in 1968 and a review of this literature is given by Altman and Saunders (1998). The use of techniques like logsitic regression in this context provides a link between the corporate and cosnumer level credit risk. Another strand in the assessment of corporate risk relies on seeing the value of the firm evolving according to a stochastic process. Then information on the volatility of the process and the upward drift in value can be translated into a statement about probability of default in a period of T years. This approach was originally proposed by Merton (1974) and since then has been adapted by KMV Corporation, acquired by Moody's in 2002. For more on this appraoch see Chapter 16 in Culp (2001) or the paper by Bharath and Shumway (2008).

The data provided for the Bank of Sydney example is based in part on a similar set of simplified hypothetical data given by Thomas (2009) (he calls it Bank of Southampton data). The book by Thomas gives a much more detailed treatment of the way that consumer credit models operate and is a good place to start if you want to go deeper into this material. In the discussion we mention the problems of carrying out an ols regression on log odds if categories have zero bad individuals. It may be overly optimistic to say that logistic regression avoids this problem. There is quite a literature on the way that the logistic regression maximum likelihood estimates are biased for small samples (see e.g. Firth, 1993 and the references there). One simple approach to reducing this bias is to add 0.5 to both the 'good' and 'bad' cells.

References

Altman E.I and Saunders A. 1998. Credit risk measurement: developments over the last 20 years. *Journal of Banking and Finance*. **21** 1721–1742.

Bharath S.T. and Shumway T. 2008. Forecasting default with the Merton distance to default model. *Review of Financial Studies*, **21** 1339–1369.

Culp C. 2001. *The risk management process: Business strategy and tactics*, Wiley.

Firth D. 1993. Bias reduction of maximum likelihood estimates, *Biometrika*, **80**, 27–38.

Merton R.C. 1974. On the pricing of corporate debt: The risk structure of interest rates, *Journal of Finance*, **29** 449–470

Thomas L. 2009. *Consumer Credit Models*, Oxford University Press,. Oxford.

Exercises

10.1. (Markov groupings)

Verify the claim that grouping states together can destroy the Markov assumption. Suppose that there are 4 states A, B, C, D. Once either A or D is reached there is no change possible. From B there is a 10% chance of moving to A and a 20% chance of moving to C, and otherwise there is no change. Similarly from C there is a 10% chance of moving to B and a 20% chance of moving to D, and otherwise there is no change. New companies arrive at B in such a way we expect the same number of companies in B as there are companies in C. Calculate the three year probability of reaching D knowing that we are equally likely to start in either B or C, and compare this with the estimate made if we group together the states B and C.

10.2. (Markov types)

Verify the claim that different types of firms each following a Markov chain can produce non-Markov behaviour in aggregate. Suppose that there are 4 states A, B, C, D. Once either A or D is reached there is no change possible. Type X firms behave as in Exercise 10.1, i.e starting from B after 1 year there is a 10% chance of moving to A and a 20% chance of moving to C. Similarly from C there is a 10% chance of moving to B and a 20% chance of moving to D. Type Y companies are the same except that they change state twice as often, i.e. from B there is a 20% chance of moving to A and a 40% chance of moving to C. Similarly from C there is a 20% chance of moving to B and a 40% chance of moving to D. New companies arrive at B in such a way that we expect to have N companies of type X in B, N companies of type Y in B, N companies of type X in C, and N companies of type Y in C. Calculate the probability of moving to D in two steps from B for a Markov chain which matches the observed annual transitions and compare this with the true probability.

10.3. (Octophone)

Octophone is a Canadian mobile phone company that keeps data on its customers and rates them according to whether they fail to make scheduled contract payments or not in the first year of the contract term. The data available on application are age bracket, whether they have a credit card and whether they have had a mobile phone contract before (either with Octophone or another company). The results from 2000 of last year's applicants living in Montreal are given in Table 10.6. In this table there are 1850 Goods and 150 Bads. Calculate the weights of evidence for the different attributes involved and use these to construct a (naive) scorecard.

Table 10.6 Octophone data: Goods and Bad

	age 18-21	age 22-29	age 30-45	46 and over
Previous phone, credit card	$G = 150$ $B = 6$	$G = 256$ $B = 8$	$G = 312$ $B = 9$	$G = 250$ $B = 7$
Previous phone, no credit card	$G = 114$ $B = 10$	$G = 123$ $B = 13$	$G = 92$ $B = 9$	$G = 91$ $B = 12$
No previous phone, credit card	$G = 99$ $B = 12$	$G = 182$ $B = 6$	$G = 59$ $B = 11$	$G = 45$ $B = 8$
No previous phone, no credit card	$G = 22$ $B = 9$	$G = 26$ $B = 14$	$G = 13$ $B = 7$	$G = 16$ $B = 9$

10.4. (Octophone with contract costs)

Octophone sells a variety of contracts but they can be classified on a dollars per month basis into three categories low cost less than \$30. medium cost: between \$30 and \$40 and high cost: more than \$40.

(a) Of the 2000 in the sample of question 1 there are a total of 800 low cost contracts, 600 medium cost contracts and 600 high cost contracts. The 150 Bads are distributed with 40 on low cost contracts, 40 on medium cost contracts and 70 on high cost contracts. Calculate the new weights of evidence including this additional information.

(b) A logistic regression is carried out on this data and produces the coefficients shown below. Here the attributes age 45+; no credit card; no previous phone; and high cost contract have all been omitted because of colinearity. Calculate a scorecard using this data.

Model3: Logit, using observations 1-2000

Dependent variable: Good

	Coefficient
const	0.194171
age18_21	0.135372
age22_29	0.401779
age30_45	0.00415712
credit_card	1.38715
prev_phone	1.45171
low_cost	0.913791
med_cost	0.427606

(c) Use scaling to adjust the scorecard from (b) so that it has the following properties: (a) there is no constant term; (b) a score of 500 represents odds of 100 to 1; and (c) an increase in the scaled score of 100 points represents the odds being multiplied by 10. (so a score of 600 represents odds of 1000 to 1).

(d) It is calculated that a mobile phone contract achieves an average profit of \$15 per year, while every customer who fails to make scheduled payments will on average cost \$100 (including cost of follow up, the average uncollected debt, and the replacement value of those handsets that are not recovered, after allowing for the average payments already made). For the scorecard in (c), what is the cutoff score where Octophone would not go ahead with the contract?

(e) Suppose that further breakdown of the figures suggests that the profits and losses are both dependent on the size of the contract as follows:

	low cost	medium cost	high cost
Profit per year	$10	$15	$20
Average cost if bad	$80	$100	$160

What would you recommend to Octophone?

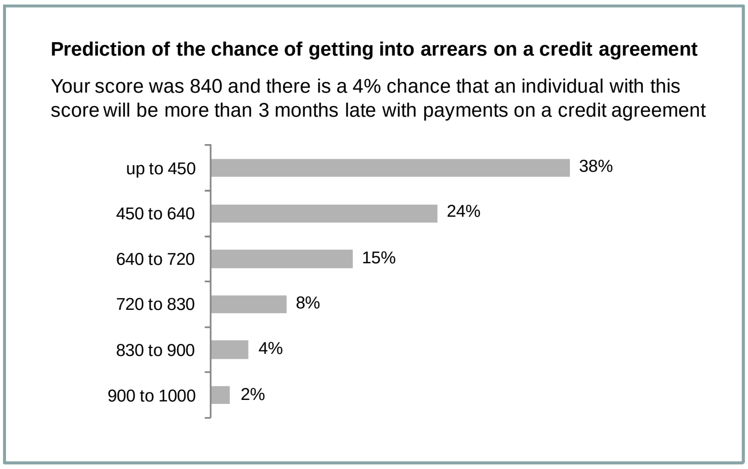
10.5. (Coral Skincare)

Coral Skincare sells beauty products over the internet. Every 6 months Coral runs a campaign to advertise one of its new products and to try to increase its customer base.. This involves sending a free sample through the mail to potential customers drawn from a mailing list. The mailing list which Coral purchases includes certain attributes of the recipients. For example Coral could pick out just females below the age of 30 who have a driving licence and live alone. Coral uses the data from the last promotion to guide it as to the kind of potential customers to include in the mail out. As is common with this sort of exercise only a small proportion of the recipients of a free sample will go on to make an online purchase. However Coral has found that a single online purchase is usually followed by others . It estimates that the profits earned from a single new online customer are approximately $150 on average. The free sample costs a total of $2.50 per recipient including mailing costs.

Table 10.7 gives the results from the last mail out to 10,000 potential customers. From this 310 new online customers were achieved which was regarded as a good result by Coral. In this table the P results are those who purchased on line, the N results are those who did not make an online purchase.

Table 10.7 Results from Coral Skincare mail out

	under 30	30-39	40-49	over 50
Household of 1, full-time work	$N = 456$ $P = 34$	$N = 412$ $P = 16$	$N = 386$ $P = 8$	$N = 386$ $P = 15$
Household of 1, not full-time work	$N = 453$ $P = 24$	$N = 389$ $P = 12$	$N = 375$ $P = 6$	$N = 395$ $P = 12$
Household of 2, full-time work	$N = 463$ $P = 24$	$N = 409$ $P = 15$	$N = 387$ $P = 10$	$N = 363$ $P = 15$
Household of 2, not full-time work	$N = 460$ $P = 18$	$N = 401$ $P = 13$	$N = 373$ $P = 5$	$N = 371$ $P = 18$
Household of $>$ 3, full-time work	$N = 448$ $P = 12$	$N = 407$ $P = 6$	$N = 405$ $P = 6$	$N = 375$ $P = 9$
Household of $>$3, not full-time work	$N = 412$ $P = 12$	$N = 404$ $P = 8$	$N = 377$ $P = 4$	$N = 383$ $P = 8$



Prediction of the chance of getting into arrears on a credit agreement

Your score was 840 and there is a 4% chance that an individual with this score will be more than 3 months late with payments on a credit agreement

Figure 10.7 Graphic from website giving free credit scores

(a) Use this data (also available on the Excel spreadsheet Coral Skincare.xls) to work out the log odds for each category and then run an ordinary least squares regression to determine a set of β values to use.

(b) Develop a scoring rule to enable Coral Skincare to decide who to include in its mail out.

10.6. (Consistent with log odds)

Suppose that Figure 10.7 is produced by a web site offering a free credit score evaluation. Are the numbers shown consistent with a log odds scoring scheme?