Chapter 4
Probability

> Being a statistician means never having to say you are certain.
>
> Anon

The essence of chance

One of the definitions of statistics given in Chapter 1 was that it is the science of handling uncertainty. Since it is abundantly clear that the world is full of uncertainty, this is one reason for the ubiquity of statistical ideas and methods. The future is an unknown land and we cannot be certain about what will happen. The unexpected does occur: cars break down, we have accidents, lightning does strike, and, lest I am giving the impression that such things are always bad, people do even win lotteries. More prosaically, it is uncertain which horse will win the race or which number will come up on the throw of a die. And, at the end of it all, we cannot predict exactly how long our lives will be.

However, notwithstanding all that, one of the greatest discoveries mankind has made is that there are certain principles covering chance and uncertainty. Perhaps this seems like a contradiction in terms. Uncertain events are, by their very nature, uncertain. How, then, can there be natural laws governing such things?

One answer is that while an individual event may be uncertain and unpredictable, it is often possible to say something about collections of

events. A classic example is the tossing of a coin. While I cannot say whether a coin will come up heads or tails on a particular toss, I can say with considerable confidence that if I toss the coin many times then around half of those times it will show heads and around half tails. (I am assuming here that the coin is 'fair', and that no sleight of hand is being used when tossing it.) Another example in the same vein is whether a baby will be male or female. It is, on conception, a purely chance and unpredictable event which gender the child will become. But we know that over many births just over a half will be male.

This observable property of nature is an example of one of the laws governing uncertainty. It is called the ***law of large numbers*** because of the fact that the proportion gets closer and closer to a particular value (a half in the cases of the fair coin and of babies' gender) the more cases we consider. This law has all sorts of implications, and is one of the most powerful of statistical tools in taming, controlling, and allowing us to take advantage of uncertainty. We return to it later in this chapter, and repeatedly throughout the book.

Understanding probability

So that we can discuss matters of uncertainty and unpredictability without ambiguity, statistics, like any other scientific discipline, uses a precise language: the language of ***probability***. If this is your first exposure to the language of probability, then you should be warned that, as with one's first exposure to any new language, some effort will be required to understand it. Indeed, bearing that in mind, you might find that this chapter requires more than one reading: you might like to reread this chapter once you have reached the end of the book.

Development of the language of probability blossomed in the 17th century. Mathematicians such as Blaise Pascal, Pierre de Fermat, Christiaan Huygens, Jacob Bernoulli, and later Pierre Simon Laplace, Abraham De Moivre, Siméon-Denis Poisson, Antoine Cournot, John Venn, and others laid

its foundations. By the early 20th century, all the ideas for a solid science of probability were in place, and in 1933 the Russian mathematician Andrei Kolmogorov presented a set of axioms which provided a complete formal mathematical ***calculus*** of probability. Since then, this axiom system has been almost universally adopted.

Kolmogorov's axioms provide the machinery by which to manipulate probabilities, but they are a mathematical construction. To use this construction to make statements about the real world, it is necessary to say what the symbols in the mathematical machinery represent in that world. That is, we need to say what the mathematics 'means'.

The probability calculus assigns numbers between 0 and 1 to uncertain events to represent the probability that they will happen. A probability of 1 means that an event is certain (e.g. the probability that, if someone looked through my study window while I was writing this book, they would have seen me seated at my desk). A probability of 0 means that an event is impossible (e.g., the probability that someone will run a marathon in ten minutes). For an event that ***can*** happen but is neither certain nor impossible, a number between 0 and 1 represents its 'probability' of happening.

One way of looking at this number is that it represents the ***degree of belief*** an individual has that the event will happen. Now, different people will have more or less information relating to whether the event will happen, so different people might be expected to have different degrees of belief, that is different probabilities for the event. For this reason, this view of probability is called ***subjective*** or ***personal*** probability: it depends on who is assessing the probability. It is also clear that someone's probability might change as more information becomes available. You might start with a probability, a degree of belief, of 1/2 that a particular coin will come up heads (based on your previous experience with other tossed coins), but after observing 100 consecutive heads and no tails appear you might become suspicious and change your subjective probability that this coin will come up heads.

Tools have been developed to estimate individuals' subjective probabilities based on betting strategies, but, as with any measurement procedure, there are practical limitations on how accurately probabilities can be estimated.

A different view of the probability of an event is that it is the proportion of times the event would happen if identical circumstances were repeated an infinite number of times. The fair coin tossing example above is an illustration. We have seen that, as the coin is tossed, so the proportion of heads gets closer and closer to some specific value. This value is defined as the probability that the coin will come up heads on any single toss. Because of the role of frequencies, or counts, in defining this interpretation of probability, it is called the ***frequentist*** interpretation.

Just as with the subjective approach, there are practical limitations preventing us from finding the exact frequentist probability. Two tosses of a coin cannot really have ***completely*** identical circumstances. Some molecules will have worn from the coin in the first toss, air currents will differ, the coin will have been slightly warmed by contact with the fingers the first time. And in any case we have to stop tossing the coin sometime, so we cannot actually toss it an infinite number of times.

These two different interpretations of what is meant by probability have different properties. The subjective approach can be used to assign a probability to a unique event, something about which it makes no sense to contemplate an infinite, or even a large number of repetitions under identical circumstances. For example, it is difficult to know what to make of the suggestion of an infinite sequence of identical attempts to assassinate the next president of the USA, with some having one outcome and some another. So it seems difficult to apply the frequentist interpretation to such an event. On the other hand, the subjective approach shifts probability from being an objective property of the external world (like mass or length) to being a property of the interaction between the observer and the world.

Subjective probability is, like beauty, in the eye of the beholder. Some might feel that this is a weakness: it means that different people could draw different conclusions from the same analysis of the same data. Others might regard it as a strength: the conclusions would have been influenced by your prior knowledge.

There are yet other interpretations of probability. The 'classical' approach, for example, assumes that all events are composed of a collection of equally likely elementary events. For example, a throw of a die might produce a 1, 2, 3, 4, 5, or 6 and the symmetry of the die suggests these six outcomes are equally likely, so each has a probability of 1/6 (they must sum to 1, since it is ***certain*** that one of 1, 2, 3, 4, 5, or 6 will come up). Then, for example, the probability of getting an even number is the sum of the probabilities of each of the equally likely events of getting a 2, a 4, or a 6, and is therefore equal to 1/2. In less artificial circumstances, however, there are difficulties in deciding what these 'equally likely' events are. For example, if I want to know the probability that my morning journey to work will take less than one hour, it is not at all clear what the equally likely elementary events should be. There is no obvious symmetry in the situation, analogous to that of the die. Moreover, there is the problem of the circular content of the definition in requiring the elementary events to be 'equally likely'. We seem to be defining probability in terms of probability.

It is worth emphasizing here that all of these different interpretations of probability conform to the same axioms and are manipulated by the same mathematical machinery. It is simply the mapping to the real world which differs; the definition of what the mathematical object ***means.*** I sometimes say that the ***calculus*** is the same, but the ***theory*** is different. In statistical applications, as we will see in Chapter 5, the different interpretations can sometimes lead to different conclusions being drawn.

The laws of chance

We have already noted one law of probability, the law of large numbers. This is a law linking the mathematics of probability to empirical observations in the real world. Other laws of probability are implicit in the axioms of probability. Some very important laws involve the concept of ***independence***.

Two events are said to be independent if the occurrence of one does not affect the probability that the other will occur. The fact that a coin tossed with my left hand comes up tails rather than heads does not influence the outcome of a coin tossed with my right hand. These two coin tosses are independent. If the probability is 1/2 that the coin in my left hand will come up heads, and the probability is 1/2 that the coin in my right hand will come up heads, then the probability that both will come up heads is $1/2 \times 1/2 = 1/4$. This is easy to see since we would expect that in many repetitions of the double tossing experiment we would obtain about half of the left hand coins showing heads, and, ***amongst those***, about half of the right hand coins would show heads because the outcome of the first toss does not influence the second. Overall, then, about 1/4 of the double tosses would show two heads. Similarly, about 1/4 would show left tails, right heads, about 1/4 would show left heads, right tails, and about 1/4 would show both left and right tails.

In contrast, the probability of falling over in the street is certainly not independent of whether it has snowed; these events are ***dependent***. We saw another example of dependent events in Chapter 1: the tragic Sally Clark case of two cot deaths in the same family. When events are not independent, we cannot calculate the probability that both will happen simply by multiplying together their separate probabilities. Indeed, this was the mistake which lay at the root of the Sally Clark case. To see this, let us take the most extreme situation of events which are completely dependent: that is, when the outcome of one ***completely determines*** the outcome of the other. For example, consider a single toss of a coin, and the two events 'the coin faces heads up' and 'the coin faces tails down'. Each of these events has a probability of a half: the probability that the coin will show heads up is 1/2, and the probability that the coin will show tails down is 1/2. But they are

clearly not independent events. In fact, they are completely dependent. After all, if the first event is true (heads up) the second ***must be*** true (tails down). Because they are completely dependent, the probability that they will both occur is simply the probability that the first will occur – a probability of a half. This is not what we get if we multiply the two separate probabilities of a half together.

In general, dependence between two events means that the probability that one will occur depends on whether or not the other has occurred.

Statisticians call the probability that two events will ***both*** occur the ***joint probability*** of those two events. For example, we can speak of the joint probability that I will slip over ***and*** that it snowed. The joint probability of two events is closely related to the probability that an event will occur ***if*** another one has occurred. This is called the ***conditional probability*** – the probability that one event will occur given that we know that the other one has occurred. Thus we can talk of the conditional probability that I will slip over, ***given that*** it snowed.

The (joint) probability that both events A and B occur is simply the probability that A occurs times the (conditional) probability that B occurs given that A occurs. The (joint) probability that it snows and I slip over is the probability that it snows times the (conditional) probability that I slip over if it has snowed.

To illustrate, consider a single throw of a die, and two events. Event A is that the number showing is divisible by 2, and Event B is that the number showing is divisible by 3. The joint probability of these two events A and B is the probability that I get a number which is both divisible by 2 and is divisible by 3. This is just 1/6, since only one of the numbers 1, 2, 3, 4, 5, and 6 is divisible by both 2 and 3. Now, the conditional probability of B given A is the probability that I get a number which is divisible by 3 ***amongst those that are divisible by 2***. Well, amongst all the numbers which are

divisible by 2 (that is, amongst 2, 4, or 6) only one is divisible by 3, so this conditional probability is 1/3. Finally, the probability of event A is 1/2 (half of the numbers 1, 2, 3, 4, 5, and 6 are divisible by 2). We therefore find that the probability of A (1/2) times the (conditional) probability of B given A (1/3) is 1/6. This is the same as the joint probability of obtaining a number divisible by both 2 and 3; that is, the joint probability of events A and B both occurring.

In fact, we previously met the concept of conditional probability in Chapter 1, in the form of the Prosecutor's Fallacy. This pointed out that the probability of event A occurring given that event B had occurred was not the same as the probability of event B occurring given that event A had occurred. For example, the probability that someone who runs a major corporation can drive a car is not the same as the probability that someone who can drive a car runs a major corporation. This leads us to another very important law of probability: ***Bayes's theorem*** (or ***Bayes's rule***). Bayes's theorem allows us to relate these two conditional probabilities, the conditional probability of A given B and the conditional probability of B given A.

We have just seen that the probability that both events A and B will occur is equal to the probability that A will occur, times the (conditional) probability that B will occur given that A has occurred. But this can also be written the other way round: the probability that both events A and B will occur is also equal to the probability that B will occur times the probability that A will occur given that B has occurred. All Bayes's theorem says (though it is usually expressed in a different way) is that these are simply two alternative ways of writing the joint probability of A and B. That is, the probability of A times the probability of B given A is equal to the probability of B times the probability of A given B. Both are equal to the joint probability of A and B. In our 'car-driving corporate head' example, Bayes's theorem is equivalent to saying that the probability of running a major corporation given that you can drive a car, times the probability that you can drive a car, is equal to the probability that you can drive a car given that you are a corporate head, times the probability of being a corporate head. Both equal the joint

probability of being a corporate head ***and*** being able to drive a car.

Another law of probability says that if either one of two events can occur, but not both together, then the probability that one ***or*** the other will occur is the sum of the separate probabilities that each will occur. If I toss a coin, which obviously cannot show heads and tails simultaneously, then the probability that a head ***or*** tail will show is the sum of the probability that a head will show and the probability that a tail will show. If the coin is fair, each of these separate probabilities is a half, so that the overall probability of a head or a tail is 1. This makes sense: 1 corresponds to certainty and it is certain that a head or a tail must show (I am assuming the coin cannot end up on its edge!). Returning to our die-throwing example: the probability of getting an even number was the sum of the probabilities of getting one of 2, or 4, or 6, because none of these can occur together (and there are no other ways of getting an even number on a single throw of the die).

Random variables and their distributions

We saw, in Chapter 2, how simple summary statistics may be used to extract information from a large collection of values of some variable, condensing the collection down so that a distribution of values could be easily understood. Now, any real data set is limited in length – it can contain only a finite number of values. This finite set might be the values of ***all*** objects of the type we are considering (e.g. the scores of all major league football players in a certain year) or it might be the values of just some, a ***sample***, of the objects. We saw examples of this when we looked at survey sampling.

A sample is a subset of the complete 'population' of values. In some cases, the complete population is ill-defined, and possibly huge or even infinite, so we have no choice but to work with a sample. For example, in experiments to measure the speed of light, each time I take a measurement I expect to get a slightly different value, simply due to the inaccuracies of the measurement process. And I could, at least in principle, go on taking

measurements for ever; that is, the potential population of measurements is infinite. Since this is impossible, I must be content with a finite sample of measurements. Each of these measurements will be drawn from the population of values I could possibly have obtained. In other cases, the complete population is finite. For example, in a study of obesity amongst males in a certain town, the population is finite and, while in principle I might be able to weigh every man in the town, in practice I would probably not want to, and would work with a sample. Once again, each value in my sample is drawn from the population of possible values.

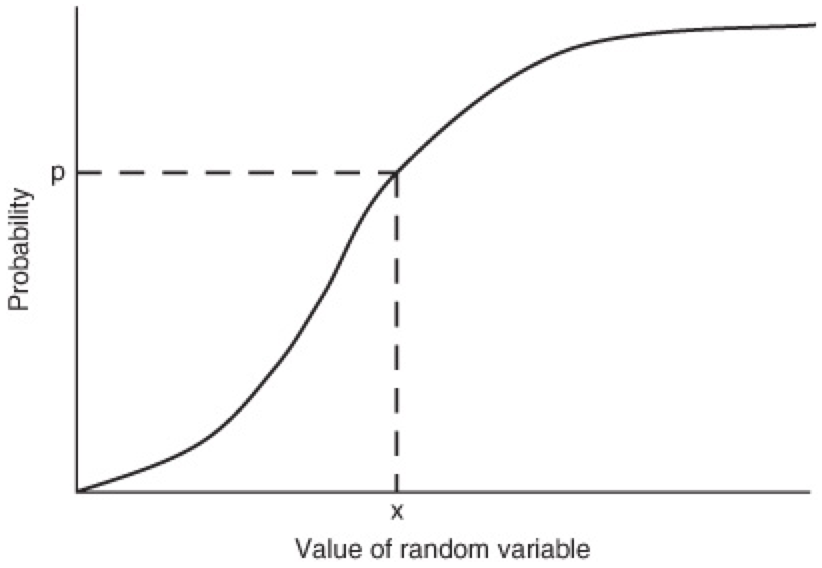
In both of these examples, all I know before I take each measurement is that it will have some value from the population of possible values. Each value will occur with some probability, but I cannot pin it down more than that, and I may not know what that probability is. I certainly cannot say exactly what value I will get in the next speed of light measurement or what will be the weight of the next man I measure. Similarly, in a throw of a die, I know that the outcome can be 1, 2, 3, 4, 5, or 6, and here I know that these are equally likely (my die is a perfect cube), but beyond that I cannot say which will come up. Like the speed and weight measurements, the outcome is random. For this reason such variables are called ***random variables***.

We have already met the concept of quantiles. For example, in the case of percentiles, the 20***th*** percentile of a distribution is the value such that 20% of the data values are smaller, the 8***th*** percentile the value such that 8% of the data values are smaller, and so on. In general, the ***k***th percentile has ***k***% of the sample values smaller than it. And we can imagine similar percentiles defined, not merely for the sample we have observed, but for the complete population of values we could have observed. If we knew the 20***th*** percentile for the complete population of values, then we would know that a value randomly taken from that population had a probability of 0.20 of being smaller than this percentile. In general, if we knew ***all*** the percentiles of a population, we would know the probability of drawing a value in the bottom 10%, or 25%, or 16%, or 98%, or any other percentage we cared to choose. In a sense, then, we would know everything about the distribution of

possible values which we could draw. We would not know what value would be drawn next, but we would know the probability that it would be in the smallest 1% of the values in the population, in the smallest 2%, and so on.

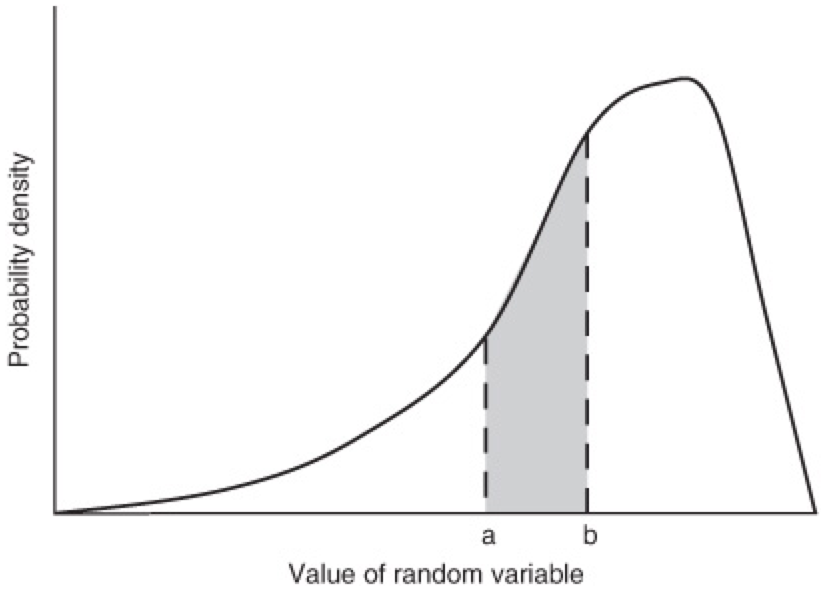
There is a name for the complete set of quantiles of a distribution. It is called the ***cumulative probability distribution***. It is a 'probability distribution' because it tells us the ***probability*** of drawing a value lower than any value we care to choose. And it is 'cumulative' because, obviously, the probability of drawing a value less than some value ***x*** gets larger the larger ***x*** is. In the example of the weights of males, if I know that the probability of choosing a man weighing less than 70kg is 1/2, then I know that the probability of choosing a man weighing less than 80kg is more than 1/2 because I can choose from all those weighing less than 70kg as well as those weighing between 70kg and 80kg. At the limit, the probability of drawing a value less than or equal to the largest value in the population is 1; it is a certain event.

This idea is illustrated in Figure 2. In this figure, the values of the random variable (think of weight) are plotted on the horizontal axis, and the probability of drawing smaller values is plotted on the vertical axis. The curve shows, for any given value of the random variable, the probability that a randomly chosen value will be smaller than this given value.



2. A cumulative probability distribution

The cumulative probability distribution of a random variable tells us the probability that a randomly chosen value will be ***less*** than any given value. An alternative way to look at things is to look at the probability that a randomly chosen value will lie ***between*** any two given values. Such probabilities are conveniently represented in terms of areas between two values under a curve of the ***density*** of the probability. For example, Figure 3, shows such a ***probability density*** curve, with the (shaded) area under the curve between points ***a*** and ***b*** giving the probability that a randomly chosen value will fall between ***a*** and ***b***. Using such a curve for the distribution of weights of men in our town, for example, we could find the probability that a randomly chosen man would lie between 70kg and 80kg, or any other pair of values, or above or below any value we wanted. In general, randomly chosen values are more likely to occur in regions where the probability is most dense; that is, where the probability density curve is highest.



3. A probability density function

Note that the total area under the curve in Figure 3 must be 1, corresponding to certainty: a randomly chosen value must have ***some*** value.

Distribution curves for random variables have various shapes. The probability that a randomly chosen woman will have a weight between 70kg and 80kg will typically not be the same as the probability that a randomly chosen man will have a weight between these two values. We might expect the curve of the distribution of women's weights to take larger values at smaller weights than does the men's curve.

Certain shapes have particular importance. There are various reasons for this. In some cases, the particular shapes, or very close approximations to them, arise in natural phenomena. In other cases, the distributions arise as consequences of the laws of probability.

Perhaps the simplest of all distributions is the ***Bernoulli distribution***. This can take only two values, one with probability ***p***, say, and the other with probability 1 – ***p***. Since it can take only two values, it is ***certain*** that one or the other value will come up, so the probabilities of these two outcomes have to sum to 1. We have already seen examples illustrating why this distribution is useful: situations with only two outcomes are very common – the coin toss, with outcomes head or tail, and births, with outcomes male or female. In these two cases, ***p*** had the value 1/2 or nearly 1/2. But a huge number of other situations arise in which there are only two possible outcomes: yes/no, good/bad, default or not, break or not, stop/go, and so on.

The ***binomial distribution*** extends the Bernoulli distribution. If we toss a coin three times, then we may obtain no, one, two, or three heads. If we have three operators in a call centre, responding independently to calls as they come in, then none, one, two, or all three may be busy at any particular moment. The binomial distribution tells us the probability that we will obtain each of those numbers, 0, 1, 2, or 3. Of course, it applies more generally, not just to the total from three events. If we toss a coin 100 times, then the binomial distribution also tells us the probabilities that we will obtain each of 0, 1, 2, …, 100 heads.

Emails arrive at my computer at random. On average, during a working morning, about (say) five an hour arrive, but the number arriving in each hour can deviate from this very substantially: sometimes ten arrive, occasionally none do. The ***Poisson distribution*** can be used to describe the probability distribution of the number of emails arriving in each hour. It can tell us the probability (if emails arrive independently and the overall rate at which they arrive is constant) that none will arrive, that one will, that two will, and so on. This differs from the binomial distribution because, at least in principle, there is no upper limit on the number which could arrive in any hour. With the 100 coin tosses, we could not observe more than 100 heads,

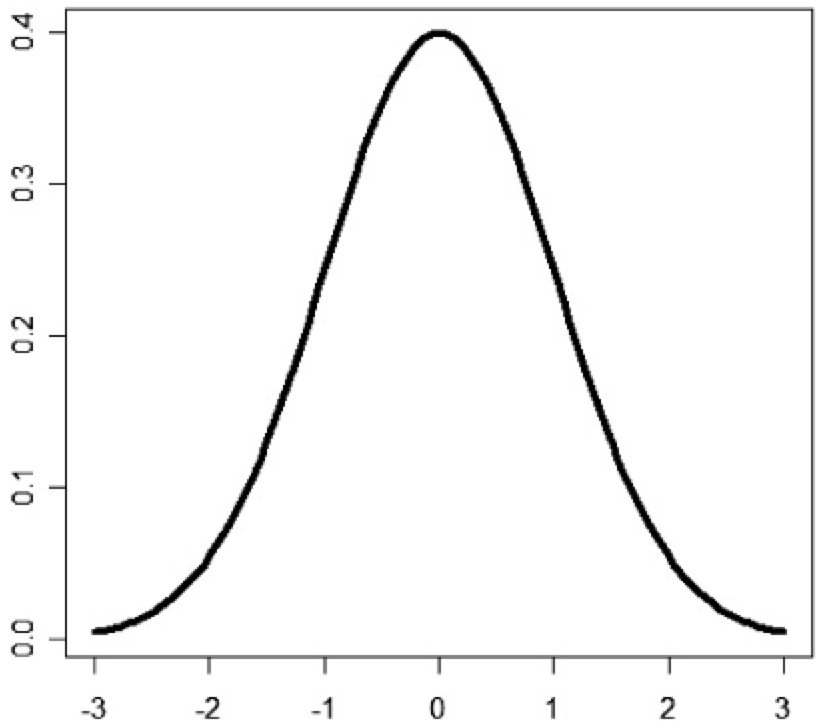
but I could (on a very bad day!) receive more than 100 emails in one hour.

So far, all the probability distributions I have described are for ***discrete*** random variables. That is, the random variables can take only certain values (two values in the Bernoulli case, counts up to the number of coin tosses/operators in the binomial case, the integers 0, 1, 2, 3, … in the Poisson case). Other random variables are ***continuous***, and can take any value from some range. Height, for example, can (subject to the accuracy of the measuring instrument) take any value within a certain range, and is not restricted to, for example, 4′, 5′, or 6′.

If a random variable can take values only within some finite interval (e.g. between 0 and 1) and if it is ***equally likely*** that it will take any of the values in that interval, then it is said to follow a ***uniform distribution***. For example, if the postman always arrives between 10am and 11am, but in a totally unpredictable way (he is as likely to arrive between 10:05 and 10:10 as in any other five minute interval, for example), the distribution of his arrival time within this interval would be uniform.

Some random variables can take any positive value; perhaps, for example, the time duration of some phenomenon. As an illustration, consider how long glass vases survive before getting broken. Glass vases do not age, so it is no more likely that a particular favourite vase will be broken in the next year, if it is 80 years old, than that it will be broken in the next year, if it is only 10 years old (all other things being equal). Contrast this with the probability that an 80-year-old human will die next year compared with the probability that a 10-year-old human will die next year. For a glass vase, if it has not been smashed by time t, then the probability that it will be smashed in the next instant is the same, whatever the value of t (again, all other things being equal). Lifetimes of glass vases are said to follow an ***exponential*** distribution. In fact, there are huge numbers of applications of exponential distributions, not merely to the lifetimes of glass vases!

Perhaps the most famous of continuous distributions is the ***normal*** or ***Gaussian distribution***. It is often loosely described in terms of its general shape: 'bell-shaped', as shown in Figure 4.



4. The normal distribution

That means that values in the middle are much more likely to occur than are values in the tails, far from the middle. The normal distribution provides a good approximation to many naturally occurring distributions. For example, the distribution of the heights of a random sample of adult men follows a roughly normal distribution.

The normal distribution also often crops up as a good model for the shape of the distribution of sample statistics (like the summary statistics described in Chapter 2) when large samples are involved. For example, suppose we

repeatedly took random samples from some distribution, and calculated the means of each of these samples. Since each sample is different, we would expect each mean to be different. That is, we would have a distribution of means. If each sample is large enough, it turns out that this distribution of the means is roughly normal.

In Chapter 2, I made the point that statistics was not simply a collection of isolated tools, but was a connected language. A similar point applies to probability distributions. Although I have introduced them individually above, the fact is that the Bernoulli distribution can be seen as a special case of the binomial distribution (it is the binomial distribution when there are only two possible outcomes). Likewise, although the mathematics showing this is beyond this book, the Poisson distribution is an extreme case of the binomial distribution, the Poisson distribution and exponential distribution form a natural pair, the binomial distribution becomes more and more similar to the normal distribution the larger the maximum number of events, and so on. They are really all part of an integrated mathematical whole.

I have described the distributions above by saying that they have different shapes. In fact, these shapes can be conveniently described. We saw that the Bernoulli distribution was characterized by a value p. This told us the probability that we would get a certain outcome. Different values of p correspond to different Bernoulli distributions. We might model the outcome of a coin toss by a Bernoulli distribution with probability of heads, p, equal to a half, and model the probability of a car crash on a single journey by a Bernoulli distribution with p equal to some very small value (I hope!). In such a situation, p is called a *parameter*.

Other distributions are also characterized by parameters, serving the same role of telling us exactly which member of a family of distributions we are talking about. To see how, let us take a step back and recall the law of large numbers. This says that if we make repeated independent observations of an event which has outcome A with probability p and outcome B with

probability 1 – ***p***, then we should expect the proportion of times outcome A is observed to get closer and closer to ***p*** the more observations we make. This property generalizes in important ways. In particular, suppose that, instead of observing an event which had only two possible outcomes, we observed an event which could take any value from a distribution on a range of values; perhaps any value in the interval [0,1], for example. Suppose that we repeatedly took sets of ***n*** measurements from such a distribution. Then the law of large numbers also tells us that we should expect the mean of the ***n*** measurements to get closer to some fixed value, the larger ***n*** is. Indeed, we can picture increasing ***n*** without limit, and in that case it makes sense to talk about the mean of an unlimited sample drawn from the distribution – and even the mean of the distribution itself. For example, using this idea we can talk about not simply the mean of 'a sample drawn from an exponential distribution', but the mean of the exponential distribution itself. And, just as different Bernoulli distributions will have different parameters ***p***, so different exponential distributions will have different means. The mean, then, is a parameter for the exponential distribution.

In an earlier example, we saw that the exponential distribution was a reasonable model for the 'lifetimes' of glass vases (under certain circumstances). Now we can imagine that we have two populations of such vases: one consisting of solid vases made of very thick glass, and the other consisting of delicate vases made of wafer-thin glass. Clearly, on average, glasses from the former population are likely to survive longer than those from the latter population. The two populations have different parameters.

We can define parameters for other distributions in a similar way: we imagine calculating the summary statistics for samples of infinite size drawn from the distributions. For example, we could imagine calculating the means of infinitely large samples drawn from members of the normal family of distributions. Things are a little more complicated here, however, because the members of this family of distributions are not uniquely identified by a single parameter. They require two parameters. In fact, the mean and standard deviation of the distributions will do. Together they serve to

uniquely identify which member of the family we are talking about.

The law of large numbers has been refined even further. Imagine drawing many sets of values from some distribution, each set being of size ***n***. For each set calculate its mean. Then the calculated means themselves are a sample from a distribution – the distribution of possible values for the mean of a sample of size ***n***. The ***Central Limit Theorem*** then tells us that the distribution of these means itself approximately follows a normal distribution, and that the approximation gets better and better the larger the value of ***n***. In fact, more than this, it also tells us that the mean of this distribution of means is identical to the mean of the overall population of values, and that the variance of the distribution of means is only $1/n$ times the size of the variance of the distribution of the overall population. This turns out to be extremely useful in statistics, because it implies that we can estimate a population mean as accurately as we like, just by taking a large enough sample (taking ***n*** large enough), with the Central Limit Theorem telling us how large a sample we must take to achieve a high probability of being that accurate. More generally, the principle that we can get better and better estimates by taking larger samples is an immensely powerful one. We already saw one way that this idea is used in practice when we looked at survey sampling in Chapter 3.

Here is another example. In astronomy, distant objects are very faint, and observations are complicated by random fluctuations in the signals. However, if we take many pictures of the same object and superimpose them, it is as if we are averaging many measurements of the same thing, each measurement drawn from the same distribution but with some extra random component. The laws of probability outlined above mean that the randomness is averaged away, leaving a clear view of the underlying signal – the astronomical object.