1 Circling the Square

It's my bad friend Kent . . . Kent works at the Central Statistics Bureau. He knows how many litres of milk Norwegians drink per annum and how often people have sex. On average that is.

Erlend Loe, *Naive. Super*

The Charisma Casualty: A Scientist in Need of an Apology and the Question He Dreads

Look at that miserable student in the corner at the party. He could be my younger self. He was doing well until she asked the dreaded question: 'What are you studying?' At such a moment what would one not give for the right to a romantic answer: 'Russian', perhaps, or 'drama'. Or a coldly cerebral one: 'philosophy' or 'mathematics' or even 'physics'. Or to pass oneself as a modern Victor Frankenstein, a genetic engineer or a biochemist. That is where the action will be in this millennium. But statistics?

The 1990 French film *Tatie Danielle*, a dark comedy about a misanthropic, manipulative and downright nasty old lady, was advertised by posters stating 'you don't know her, but she loathes you already'. Of most people one might just as well say, 'you've never studied statistics but you loathe it already'. You know already what it will involve (so many tonnes of coal mined in Silesia in 1963, so many deaths from TB in China in 1978). Well, you are wrong. It has nothing, or hardly anything, to do with that. And if you have encountered it as part of some degree course, for no scientist or social scientist escapes, then you know that it consists of a number of algorithms for carrying out tests of significance using data. Well, you are also wrong. Statistics, like Bill Shankly's football, is not just a matter of life and death: 'Son, it's much more important than that.'

Statistics Are and Statistics Is

Statistics singular, contrary to the popular perception, is not really about facts; it is about how we know, or suspect, or believe, that something is a fact. Because knowing about things involves counting and measuring them, then it is true that statistics plural are part of the concern of statistics singular, which is the science of quantitative reasoning. This science has much more in common with philosophy (in particular epistemology) than it does with accounting. Statisticians are applied philosophers. Philosophers argue how many angels can dance on the head of a pin; statisticians *count* them.

Or rather, count how many can *probably* dance. Probability is the heart of the matter, the heart of all matter if the quantum physicists can be believed. As far as the statistician is concerned this is true whether the world is strictly deterministic as Einstein believed or whether there is a residual ineluctable indeterminacy. We can predict nothing with certainty but we can predict how uncertain our predictions will be – on average, that is. Statistics is the science that tells us how.

Quacks and Squares

I want to explain how important statistics is. For example, take my own particular field of interest, pharmaceutical clinical trials: experiments on human beings to establish the effects of drugs. Why, as a statistician, do I do research in this area? I don't treat patients. I don't design drugs. I scarcely know a stethoscope from a thermometer. I have forgotten most of the chemistry I ever knew and I never studied biology. But I have successfully designed and analysed clinical trials for a living. Why should it be that the International Conference on Harmonisation's guidelines for Good Clinical Practice, the framework for the conduct of pharmaceutical trials in Europe, America and Japan, should state 'The sponsor should utilize qualified individuals (e.g. biostatisticians, clinical pharmacologists and physicians) as appropriate, throughout all stages of the trial process, from designing the protocol and CRFs and planning the analyses to analyzing and preparing interim and final clinical trial reports'?[1] We know why we need quacks but these 'squares' who go around counting things, what use are they? We don't treat patients with statistics, do we?

High Anxiety

Of course not. Suppose that you have just suffered a collapsed lung at 35 000 ft and, the cabin crew having appealed for help, a 'doctor' turns up. A PhD in statistics would be as much use as a spare statistician at a party. You damn well want the doctor to be a medic. In fact, this is precisely what happened to a lady travelling from Hong Kong to Britain in May 1995. She had fallen off a motorcycle on her way to the airport and had not realized the gravity of her injuries until airborne. Luckily for her, two resourceful physicians, Professor Angus Wallace and Dr Tom Wang, were on board.[2]

Initially distracted by the pain she was experiencing in her arm, they eventually realized that she had a more serious problem. She had, in fact, a 'tension pneumothorax' – a life-threatening condition that required immediate attention. With the help of the limited medical equipment on board plus a coat hanger and a bottle of Evian water, the two doctors performed an emergency operation to release air from her pleural cavity and restore her ability to breathe normally. The operation was a complete success and the woman recovered rapidly.

This story illustrates the very best aspects of the medical profession and why we value its members so highly. The two doctors concerned had to react quickly to a rapidly developing emergency, undertake a technical manoeuvre in which they were probably not specialized and call not only on their medical knowledge but on that of physics as well: the bottle of water was used to create a water seal. There is another evidential lesson for us here, however. We are convinced by the story that the intervention was necessary and successful. This is a very reasonable conclusion. Amongst factors that make it reasonable are that the woman's condition was worsening rapidly and that within a few minutes of the operation her condition was reversed.

A Chronic Problem

However, much of medicine is not like that. General practitioners, for example, busy and harassed as they are, typically have little chance of learning the effect of the treatments they employ. This is because most of what is done is either for chronically ill patients for whom no rapid reversal can be expected or for patients who are temporarily ill, looking for some relief or a speedier recovery and who will not report back.

Furthermore, so short is the half-life of relevance of medicine that if (s)he is middle-aged, half of what (s)he learned at university will now be regarded as outmoded, if not downright wrong.

The trouble with medical education is that it prepares doctors to learn facts, whereas really what the physician needs is a strategy for learning. The joke (not mine) is that three students are asked to memorize the telephone directory. The mathematician asks 'why?', the lawyer asks 'how long have I got?' and the medical student asks 'will the Yellow Pages also be in the exam?' This is changing, however. There is a vigorous movement for evidence-based medicine that stresses the need for doctors to remain continually in touch with developments in treatment and also to assess the evidence for such new treatment critically. Such evidence will be quantitative. Thus, doctors are going to have to learn more about statistics.

It would be wrong, however, to give the impression that there is an essential antagonism between medicine and statistics. In fact, the medical profession has made important contributions to the theory of statistics. As we shall see when we come to consider John Arbuthnot, Daniel Bernoulli and several other key figures in the history of statistics, many who contributed had had a medical education, and in the medical specialty of epidemiology many practitioners can be found who have made important contributions to statistical theory. However, on the whole it can be claimed that these contributions have arisen because the physician has come to think like a statistician: with scepticism. 'This is plausible, how might it be wrong?' could be the statistician's catchphrase. In the sections that follow, we consider some illustrative paradoxes.

A Familiar Familial Fallacy?

'Mr Brown has exactly two children. At least one of them is a boy. What is the probability that the other is a girl?' What could be simpler than that? After all, the other child either is or is not a girl. I regularly use this example on the statistics courses I give to life scientists working in the pharmaceutical industry. They all agree that the probability is one-half.

One could argue they are wrong. I haven't said that the *older* child is a boy. The child I mentioned, the boy, could be the older or the younger child. This means that Mr Brown can have one of three possible combinations of two children: both boys, elder boy and younger girl, or elder girl and younger boy, the fourth combination of two girls being excluded by what I have stated.

But of the three combinations, in two cases the other child is a girl so that the requisite probability is $^2/_3$. This is illustrated as follows.

	Possible	Possible	Possible	Excluded
Elder	♂	♂	♀	♀
Younger	♂	♀	♂	♀

This example is typical of many simple paradoxes in probability: the answer is easy to explain but nobody believes the explanation. However, the solution I have given *is* correct.

Or is it? That was spoken like a probabilist. A probabilist is a sort of mathematician. He or she deals with artificial examples and logical connections but feels no obligation to say anything about the real world. My demonstration, however, relied on the assumption that the three combinations boy–boy, boy–girl and girl–boy are equally likely and this may not be true. In particular, we may have to think carefully about what I refer to as *data filtering*. How did we get to see what we saw? The difference between a statistician and a probabilist is that the latter will define the problem so that this is true, whereas the former will consider *whether* it is true and obtain data to test its truth.

Suppose we make the following assumptions: (1) the sex ratio at birth is 50:50; (2) there is no tendency for boys or girls to run in a given family; (3) the death rates in early years of life are similar for both sexes; (4) parents do not make decisions to stop or continue having children based on the mix of sexes they already have; (5) we can ignore the problem of twins. Then the solution is reasonable. (Provided there is nothing else I have overlooked!) However, the first assumption is known to be false, as we shall see in the next chapter. The second assumption is believed to be (approximately) true but this belief is based on observation and analysis; there is nothing logically inevitable about it. The third assumption is false, although in economically developed societies the disparity in the death rates between sexes, although considerable in later life, is not great before adulthood. There is good evidence that the fourth assumption is false. The fifth is not completely ignorable, since some children are twins, some twins are identical and all identical twins are of the same sex. We now consider a data set that will help us to check our answer.

In an article in the magazine *Chance*, in 2001, Joseph Lee Rogers and Debby Doughty attempted to answer the question 'Does having boys or girls run in the family?'[3] The conclusion that they came to is that it does not, or, at least, if it does that the tendency is at best very weak. To establish this conclusion they used data from an American study: the National Longitudinal Survey of Youth (NLSY). This originally obtained a sample of more than 12 000 respondents aged 14–21 years in 1979. The NLSY sample has been followed up from time to time since. Rogers and Doughty used data obtained in 1994, by which time the respondents were aged 29–36 years and had had 15 000 children between them. The same data that they used to investigate the sex distribution of families can be used to answer our question.

Of the 6089 NLSY respondents who had had at least one child, 2444 had had exactly two children. In these 2444 families the distribution of children was: boy–boy, 582; girl–girl, 530; boy–girl, 666; and girl–boy, 666. If we exclude girl–girl, the combination that is excluded by the question, then we are left with 1914 families. Of these families, 666 + 666 = 1332 had one boy and one girl, so the proportion of families with at least one boy in which the other child is a girl is $1332/1914 \simeq 0.70$. Thus, in fact, our requisite probability is not $^2/_3$ as we previously suggested, but $^7/_{10}$ (approximately).

Or is it? We have moved from a view of probability that tries to identify equally probable cases – what is sometimes called classical probability – to one that uses relative frequencies. There are, however, several objections to using this ratio as a probability, of which two are particularly important. The first is that a little reflection shows that it is obvious that such a ratio is itself subject to chance variation. To take a simple example, even if we believe a die to be fair we would not expect that whenever we rolled the die six times we would obtain exactly one 1, 2, 3, 4, 5 & 6. The second objection is that even if this ratio is an adequate approximation to some probability, why should we accept that it is the probability that applies to Mr Brown? After all, I have not said that he is either an American citizen who was aged 14–21 in 1971 or has had children with such a person, yet this is the group from which the ratio was obtained.

The first objection might lead me to prefer a theoretical value such as the $^2/_3$ obtained by our first argument to the value of approximately $^7/_{10}$ (which is, of course, very close to it) obtained by the second. In fact, statisticians have developed a number of techniques for deciding how reasonable such a theoretical value is. We shall consider one of these in due course, but first draw attention to one further twist in the paradox.

Child's Play

I am grateful to Ged Dean for pointing out that there is another twist to this paradox. Suppose I argue like this. Let us consider Mr Brown's son and consider the other child relative to him. This is either an older brother or a younger brother or an older sister or a younger sister. In two out of the four cases it is a boy. So the probability is one-half after all.

This disagrees, of course, with the empirical evidence I presented previously but that evidence depends on the way I select the data: essentially sampling by fathers rather than by children. The former is implicit in the way the question was posed, implying sampling by father, but as no sampling process has been defined you are entitled to think differently.

To illustrate the difference, let us take an island with four two-child families, one of each of the four possible combinations: boy–boy, boy–girl, girl–boy and girl–girl. On this island it so happens that the oldest child has the name that begins with a letter earlier in the alphabet. The families are:

Fred and Pete (father Bob);
Andrew and Susan (father Charles);
Anthea and Zack (father Dave);
Beatrice and Charlotte (father Ed).

Let us choose a father at random. There are three chances out of four that it is either Bob, Charles or Dave, who each have at least one son. Given that the father chosen has at least one boy there are two chances out of three that the father is either Charles or Dave and therefore that the other child is a girl. So, there is a probability of two-thirds that the other child is a girl. This agrees with the previous solution.

Now, however, let us choose a child at random. There are four chances out of eight that it is a boy. If it is a boy, it is either Fred or Pete or Andrew or Zack. In two out of the four cases the other child is a boy.

To put it another way: given that the child we have chosen is a boy, what is the probability that the father is Bob? The answer is 'one-half '.

A Likely Tale*

We now consider the general problem of estimating a probability from data. One method is due to the great British statistician and geneticist R. A. Fisher (1890–1962) whom we shall encounter again in various chapters in this book. This is based on his idea of likelihood. What you can do in a circumstance like this, he points out, is to investigate each and every possible value for the probability from 0 to 1. You can then try each of these values in turn and see how likely the data are given the value of the probability you currently assume. The data for this purpose are that of the 1914 relevant families: in 1332 the other child was a girl and in 582 it was a boy. Let the probability in a given two-child family that the other child is a girl, where at least one child is male, be P, where, for example, P might be $^2/_3$ or $^7/_{10}$ or indeed any value we wish to investigate. Suppose that we go through the 1914 family records one by one. The probability of any given record corresponding to a mixed-sex family is P and the probability of it corresponding to a boys-only family is $(1-P)$. Suppose that we observe that the first 1332 families are mixed sex and the next 582 are boys only. The likelihood, to use Fisher's term, of this occurring is $P \times P \times P \cdots P$, where there are 1332 such terms P, multiplied by $(1-P) \times (1-P) \times (1-P) \cdots (1-P)$, where there are 582 such terms. Using the symbol L for likelihood, we may write this as

$$L = P^{1332}(1-P)^{582}.$$

Now, of course, we have not seen the data in this particular order; in fact, we know nothing about the order at all. However, the likelihood we have calculated is the same for any given order, so all we need to do is multiply it by the number of orders (sequences) in which the data could occur. This turns out to be quite unnecessary, however, since whatever the value of P, whether $^2/_3$, $^7/_{10}$ or some other value, the number of possible sequences is the same so that in each of such cases the number we would multiply L by would be the same. This number is thus irrelevant to our inferences about P and, indeed, for any two values of P the ratio of the two corresponding values of L does not depend on the number of ways in which we can obtain 1332 mixed-sex and 582 two-boy families.

It turns out that the value of P that maximizes L is that which is given by our empirical proportion, so we may write $P_{max} = 1332/1914$. We can now express the likelihood, L, of any value of P as a ratio of the likelihood L_{max} corresponding to P_{max}. This has been done and plotted against all

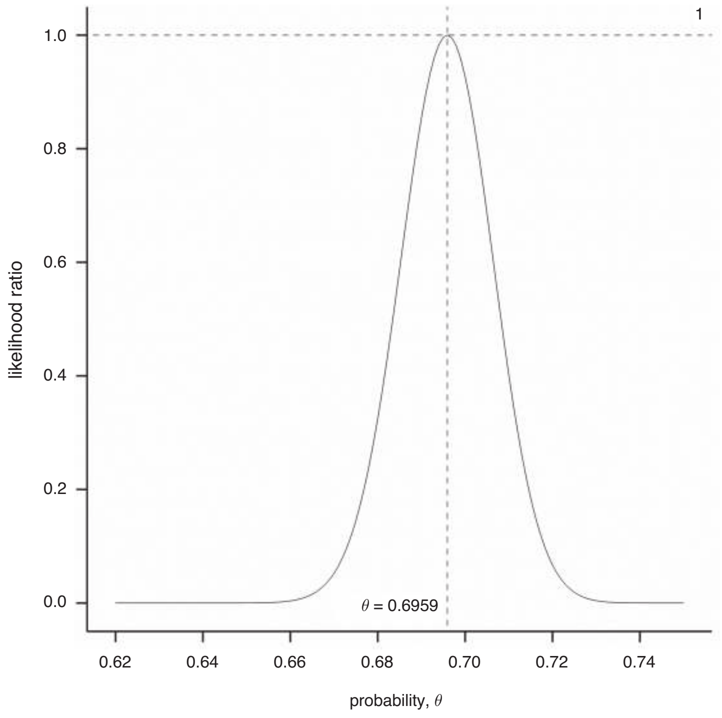

Figure 1.1 Likelihood ratio for various values of the probability of the other child being a girl given the NLSY sample data.

possible values of *P* in Figure 1.1. One can see that this ratio reaches a maximum one at the observed proportion, indicated by a solid line, and tails off rapidly either side. In fact, for our theoretical answer of $^2/_3$, indicated by the dashed line, the ratio is less than $^1/_{42}$. Thus, the observed proportion is 42 times more likely to occur if the true probability is P_{max} = 1332/1914 than if it is the theoretical value of $^2/_3$ suggested.

An Unlikely Tail?

This is all very well, but the reader will justifiably protest that the best fitting pattern will always fit the data better than some theory that issues a genuine prediction. For example, nobody would seriously maintain that the next time somebody obtains a sample of exactly 1914 persons having exactly two children, at least one of which is male, they will also observe that in 1332 cases the other is female. Another, perhaps not very different proportion would obtain and this other proportion would of course not only fit the data better than the theoretical probability of $^2/_3$, it would also fit the data better than the proportion 1332/1914 previously observed.

In fact, we have another data set with which we can check this proportion. This comes from the US Census Bureau National Interview Survey, a yearly random sample of families. Amongst the 342 018 households on which data were obtained from 1987 to 1993, there were 42 888 families with exactly two children, and 33 365 with at least one boy. The split amongst the 33 365 was boy–girl, 11 118; girl–boy, 10 913; and boy–boy, 11 334. Thus, 22 031 of the families had one boy and one girl and the proportion we require is 22 031/33 365 $\simeq$ 0.66, which is closer to the theoretical value than our previous empirical answer. This suggests that we should not be too hasty in rejecting a plausible theoretical value in favour of some apparently better-fitting alternative. How can we decide when to reject such a theoretical value?

This statistical problem of deciding when data should lead to rejection of a theory has a very long history and we shall look at attempts to solve it in the next chapter. Without entering into details, here we consider briefly the approach of significance testing which, again, is particularly associated with Fisher, although it did not originate with him. This is to imagine for the moment that the theoretical value is correct and then pose the question 'if the value is correct, how unusual are the data?'

Defining exactly what is meant by 'unusual' turns out to be extremely controversial. One line of argument suggests, however, that if we were to reject the so-called null hypothesis that the true probability is $^2/_3$, then we have done so where the observed ratio is 1332/1914, which is higher than $^2/_3$, and would be honour-bound to do so had the ratio been even higher. We thus calculate the probability of observing 1 332 or more mixed-sex families when the true probability is $^2/_3$. This sort of probability is referred to as a 'tail area' probability and, sparing the reader the details,[4] in this case it turns out to be 0.00337. However, we could argue that we would have been just as impressed by an observed proportion that was lower than the hypothesized value $^2/_3$ as by finding one that was higher, so we ought to double this probability. If we do, we obtain a value of 0.0067. This sort of probability is referred to as a '*P*-value' and is