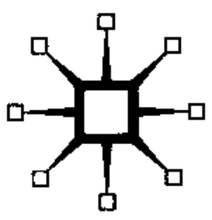
Understanding psychology as a science

An introduction to scientific and statistical inference

Zoltán Dienes
University of Sussex

palgrave
macmillan



First published 2008 by
PALGRAVE MACMILLAN
Houndmills, Basingstoke, Hampshire RG21 6XS and
175 Fifth Avenue, New York, N.Y. 10010
Companies and representatives throughout the world

PALGRAVE MACMILLAN is the global academic imprint of the Palgrave Macmillan division of St. Martin's Press, LLC and of Palgrave Macmillan Ltd. Macmillan® is a registered trademark in the United States, United Kingdom and other countries. Palgrave is a registered trademark in the European Union and other countries.

ISBN-13: 978–0–230–54231–0 paperback
ISBN-10: 0–230–54231–X paperback
ISBN-13: 978–0–230–54230–3 hardback
ISBN-10: 0–230–54230–1 hardback

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources. Logging, pulping and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

A catalogue record for this book is available from the British Library.

A catalog record for this book is available from the Library of Congress.

10 9 8 7 6 5 4 3 2 1
17 16 15 14 13 12 11 10 09 08

Printed and bound in China

4 Bayes and the probability of hypotheses

We saw in Chapter 3 that interpreting probability as *objective* probability, namely a long-run relative frequency, meant developing statistical tools that only allow inferences consistent with that meaning, namely inferences about long-run relative frequencies. Classic (Neyman–Pearson) statistics can tell you the long-run relative frequency of different types of errors. Classic statistics do not tell you the probability of any hypothesis being true. We also saw that, in contrast, many people mistakenly believe that significance values do tell them the probability of hypotheses. This common belief leads in practice to erroneous decisions in reviewing studies and conducting research (Oakes, 1986).

An alternative approach to statistics is to start with what *Bayesians* say are people's natural intuitions: People apparently want statistics to tell them the probability of their hypothesis being right. In this chapter, we will define *subjective* probability (subjective degree of conviction in a hypothesis) and then see how we can develop statistical tools to inform us how we should change our conviction in a hypothesis in the light of experimental data. Many aspects of classic statistical inference are simplified in this Bayesian approach. We shall review some surprising consequences of such an approach. We will finish by considering how well the classic and Bayesian approaches really are consistent with or conflict with our intuitions about appropriate research decisions.

Amongst users of statistics, like psychologists, Neyman–Pearson statistics is the unchallenged establishment view. But few users of statistics are even aware of the names of the approaches to statistical inference, let alone the conceptual issues, and thus have scarcely made an informed choice. Amongst statisticians and philosophers, there is a raging debate concerning what we are even trying to do with statistics (let alone how we do it), and as yet no clear intellectual victor (despite protests to the contrary). In this chapter and the next, we consider the main rivals to the orthodoxy. Maybe you personally will find that, after considering the arguments, you favour one side of the dispute more than the others.

Subjective probability

In everyday life, we often say things like 'It will probably snow tomorrow', 'There are even odds that the next colour will be red', 'Uzbekistans is most likely to win the match', and 'Baker's theory of sperm competition is probably true'. As you now know, none of these statements are legitimate statements about objective probability (long-run relative frequency). But the everyday use of 'probable' does not respect the strict demands of objective probability. In the everyday use, we are quite willing to talk about how probable single events or hypotheses are. *Subjective* or *personal* probability is the degree of conviction we have in a hypothesis. Given this meaning, probabilities are in the mind, not in the world. If you say 'It is highly probable it will snow tomorrow', you are making a statement about how strongly you believe that it will snow tomorrow. No expert on the weather can tell you that you are wrong in assigning a high personal probability to snow tomorrow. Although the weather expert knows about weather patterns, she does not know better than you yourself about

what state your mind is in. She might change your mind by giving you more information, but your statement about how strongly you believed (at that point in time) that it will snow tomorrow still stands as a true statement.

The initial problem to address in making use of subjective probabilities is how to assign a precise number to how probable you think a proposition is. Let's use a number between 0 and 1, where 0 means zero probability, there is no chance that the statement is true; and 1 means you are certain that the statement is true. But if you are neither certain that the statement is false (probability = 0) nor certain that it is true (probability = 1), what number between 0 and 1 should you choose? A solution is to see how much money you would be willing to bet on the statement.

To determine your personal conviction in the statement 'it will rain here tomorrow', statement (1) below, choose either that statement or statement (2) below. For whichever statement you have chosen, if it turns out to be true, I will pay you £10:

1. It will rain here tomorrow.

   I have a bag with one red chip and one blue chip in it. I shake it up, close my eyes and draw a chip:
2. I draw the red one.

   If you chose the first statement, then your probability that it will rain tomorrow is greater than 0.5; otherwise it is less. We can narrow down your probability more precisely by choosing a bag with different proportions of red chips in it. For example, let us say you chose option 1 above. Then, we could ask you to choose between 1 again and:

   I have a bag with three red chips and one blue. I shake it up, close my eyes and draw a chip:
3. I draw the red one.

   You can choose 1 or 3. If your chosen scenario turns out to be true, I will pay you £10. Which do you choose?

If you choose 1 again, then your personal probability is greater than 0.75; otherwise it is between 0.5 and 0.75. You can specify your probability as precisely as you think it is worth by giving yourself repeated choices between the proposition in question 1 and a bag with different proportion of red chips in. (The Appendix provides a procedure for honing in one's subjective convictions in a single gamble.) The initial personal probability that you assign to any theory is just up to you. Reach deep inside your soul and see what you are willing to bet.

Sometimes it is useful to express your personal convictions in terms of *odds* rather than probabilities:

$$\text{odds(theory is true)} = \text{probability(theory is true)}/\text{probability(theory is false)}.^{1}$$

For example, if your personal probability that the theory is true is 0.5, your odds in favour of the theory are $0.5/(1-0.5) = 1$, or as we say 1:1 (1 to 1), or even odds. If your personal probability is 0.75, your odds in favour of the theory is $0.75/(1-0.75) = 3:1$ (3 to 1). Conversely, if your personal probability is 0.25, your odds are 1:3.

Now let us stipulate something that does not come naturally to people: These numbers we get from deep inside us must obey the axioms of probability. This is the stipulation that ensures the way we change our personal probability in a theory is coherent and rational.

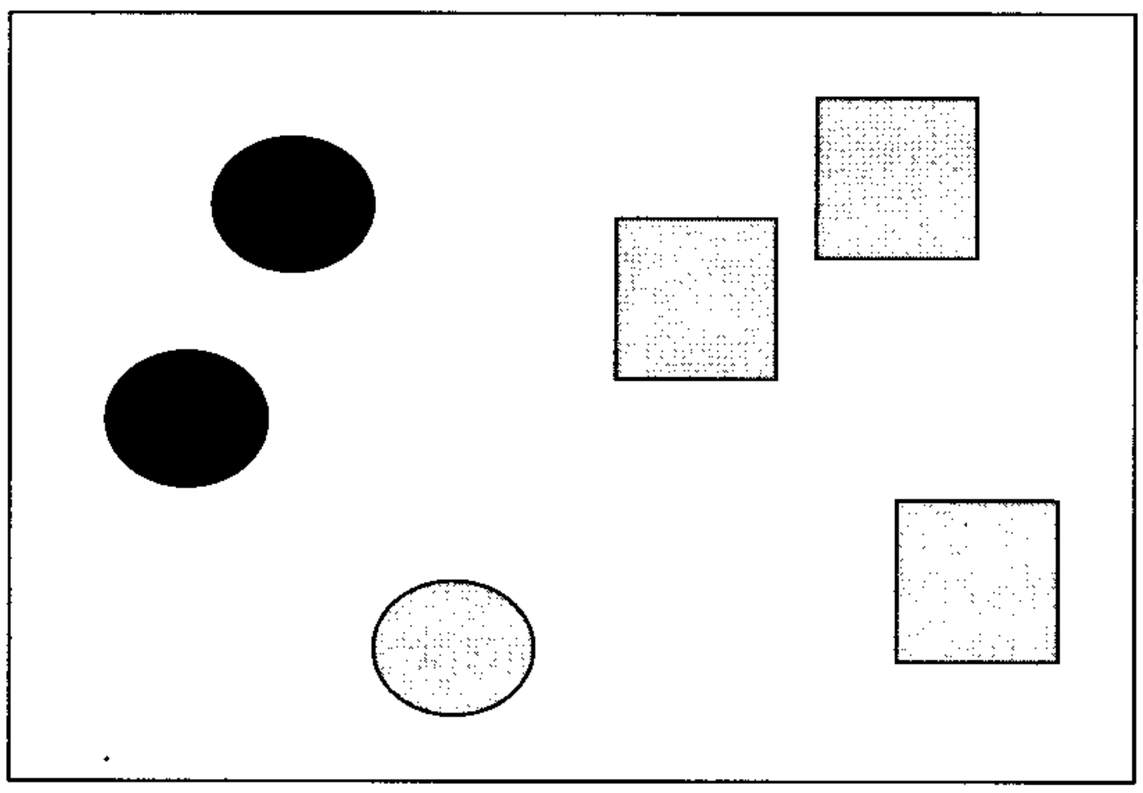
1. Hence probability(theory is true) = odds/(odds + 1).

People's intuitions about how to change probabilities in the light of new information are notoriously bad (see e.g. Sutherland, 1994). This is where the statistician comes in and forces us to be disciplined.

At first sight, making our personal probabilities obey the axioms of probability seems just like common sense. There are only a few axioms, each more-or-less self-evidently reasonable. Two axioms effectively set limits on what values probabilities can take; namely, all probabilities will lie between 0 and 1, inclusive. The next asserts $P(A \text{ or } B) = P(A) + P(B)$, if A and B are mutually exclusive. For example, imagine rolling a die; it can come up as '1' or '2' or '3' and so on up to '6'. Each of these possibilities is mutually exclusive: if '1' comes up then a '2' does not. Let my personal probability for a '1' coming up be a 1/6 and my personal probability for a '2' coming up also be 1/6. The axiom asserts that my personal probability P(getting '1' OR getting a '2'), the probability of getting either a '1' or a '2' on a roll of the die, should be $P(\text{getting '1'}) + P(\text{getting '2'}) = 1/6 + 1/6 = 1/3$. Finally, there is an axiom that says, $P(A \text{ and } B) = P(A) \times P(B|A) \cdot P(B|A)$ is the probability of B given A, which just means assuming A is the case, then what is the probability of B? Box 4.1 illustrates this final axiom.

It is surely not too much to ask of our personal probabilities that in updating, they do no more and no less than obey these axioms.

Box 4.1 $P(L \text{ and } C) = P(L) \times P(C|L)$



Imagine you have a box with the above objects in, circles and squares which are light or dark. You shake it and repeatedly put your hand in, to draw out an object. You regard each object as equally likely to be drawn. What is the probability that you will draw something that is both light and a circle, $P(L \text{ AND } C)$?

The probability of drawing a circle is a half ($P(C) = 1/2$). That is, you expect that half of the draws will be circles. Given that you have drawn a circle, the probability that it is light is a third ($P(L|C) = 1/3$). That is, a third of those half of draws that are circles are expected to be light circles: $P(L \text{ and } C) = P(C) \times P(L|C) = 1/2 \times 1/3 = 1/6$.

One can think of this the other way round as well. The probability of drawing a light object is 4/6 ($P(L) = 4/6$). Given that you have drawn a light object, the probability that it is a circle is a quarter ($P(C|L) = 1/4$). That is, a quarter of those 4/6 of draws that are light will be light circles: $P(L \text{ and } C) = P(L) \times P(C|L) = 4/6 \times 1/4 = 1/6$.

Bayes' theorem

Thomas Bayes (1702–1761), a nonconformist minister, was a fellow of the Royal Society, despite the fact that he had no published papers in his lifetime, at least none under his

name. (The good old days.) After the reverend Bayes' death, a friend of his, Richard Price, found a manuscript among Bayes' papers, and considered it of such importance that he presented it to the Royal Society on behalf of his friend in 1764. Bayes had worked on the problem of how one may obtain the probability of a hypothesis given some data, that is $P(H|D)$.

Bayes' theorem is easy to derive from first principles using the axiom in Box 4.1. If you ever forget Bayes' theorem, you can always derive it for yourself. Consider hypothesis H and data D. We have

$$P(H \text{ and } D) = P(D) \times P(H|D) \tag{4.1}$$

and

$$P(H \text{ and } D) = P(H) \times P(D|H) \tag{4.2}$$

Equations (4.1) and (4.2) are just statements of the axiom in Box 4.1.
We see the right hand sides of (1) and (2) are equal, so

$$P(D) \times P(H|D) = P(H) \times P(D|H)$$

Moving $P(D)$ to the other side

$$P(H|D) = P(D|H) \times P(H)/P(D) \tag{4.3}$$

Equation (4.3) is one version of Bayes' theorem. It tells you how to go from one conditional probability to its inverse. We can simplify Equation (4.3) if we are interested in comparing the probability of different hypotheses given the *same* data D. Then $P(D)$ is just a constant for all these comparisons. So we have

$$P(H|D) \text{ is proportional to } P(D|H) \times P(H) \tag{4.4}$$

Equation (4.4) is another version of Bayes' theorem.

$P(H)$ is called the *prior*. It is how probable you thought the hypothesis was prior to collecting data. This is your personal subjective probability and its value is completely up to you. $P(H|D)$, the probability of the hypothesis given the data, is called the *posterior*. 'Posterior' means literally 'coming after'. The posterior is how probable your hypothesis is to you, after you have collected data. $P(D|H)$ is the probability of obtaining the data, given your hypothesis; this is called the *likelihood* of the hypothesis. (Strictly, because the posterior is merely *proportional* to the likelihood times the prior, the likelihood is defined as anything proportional to $P(D|H)$). In words, Equation (4.4) says that

your posterior is proportional to the likelihood times the prior.

This is the mantra of Bayesian statistics. It tells us how we can update our prior probability in a hypothesis given some data. Your prior can be up to you; but having settled on it, the posterior is determined by the axioms of probability. From the Bayesian perspective, scientific inference consists precisely in updating one's personal conviction in a hypothesis in the light of data.[2]

2. Thus, if you think back to Chapter 1, you see Bayesians apparently are *inductivists,* engaging in what Hume and Popper argued does not exist: inductive logic, using specific observations to progressively confirm general statements. Bayesians do not see the Bayesian part of what they do as violating Hume's arguments though; Bayes is just a piece of valid mathematics, so it only makes explicit what was already implicit in your existing beliefs. In order to calculate probabilities, you need a model of the world. If you already believe the world is described by a certain type of model, Bayes will tell you how to update probabilities for different variants of that model.

The likelihood

According to Bayes' theorem, if you want to update your personal probability in a hypothesis, the likelihood tells you *everything* you need to know about the data (Edwards, 1972). All support for a hypothesis provided by the data is captured by the likelihood. Posterior is proportional to *likelihood* times prior. The notion that all the information relevant to inference contained in data is provided by the likelihood is called the *likelihood principle*. At first sight, the likelihood principle may seem like a truism, something that just follows from the axioms of probability. In fact, it is controversial. While Bayesian inference respects it (as does likelihood inference, considered the next chapter), Neyman–Pearson inference violates it, as we will see later (and discuss further in the next chapter). So what does a likelihood look like exactly?

Imagine we are interested in how men respond to people telling them about their problems. Let us say a man can respond in one of two ways: He can offer a solution to the friend or he can offer no solutions, but provide empathy. We call the first type of man a solver and the second type an empathizer. Gray (2002) suggested that men habitually offer solutions when women describe problems, and this gets the man in deep trouble because women are looking for empathy! Our research question is what proportion of men in a population are solvers? We tell five men our problems. All five men suggest solutions to our problems; they are solvers. Our *data* are the fact that five out of the five men in the sample were solvers.

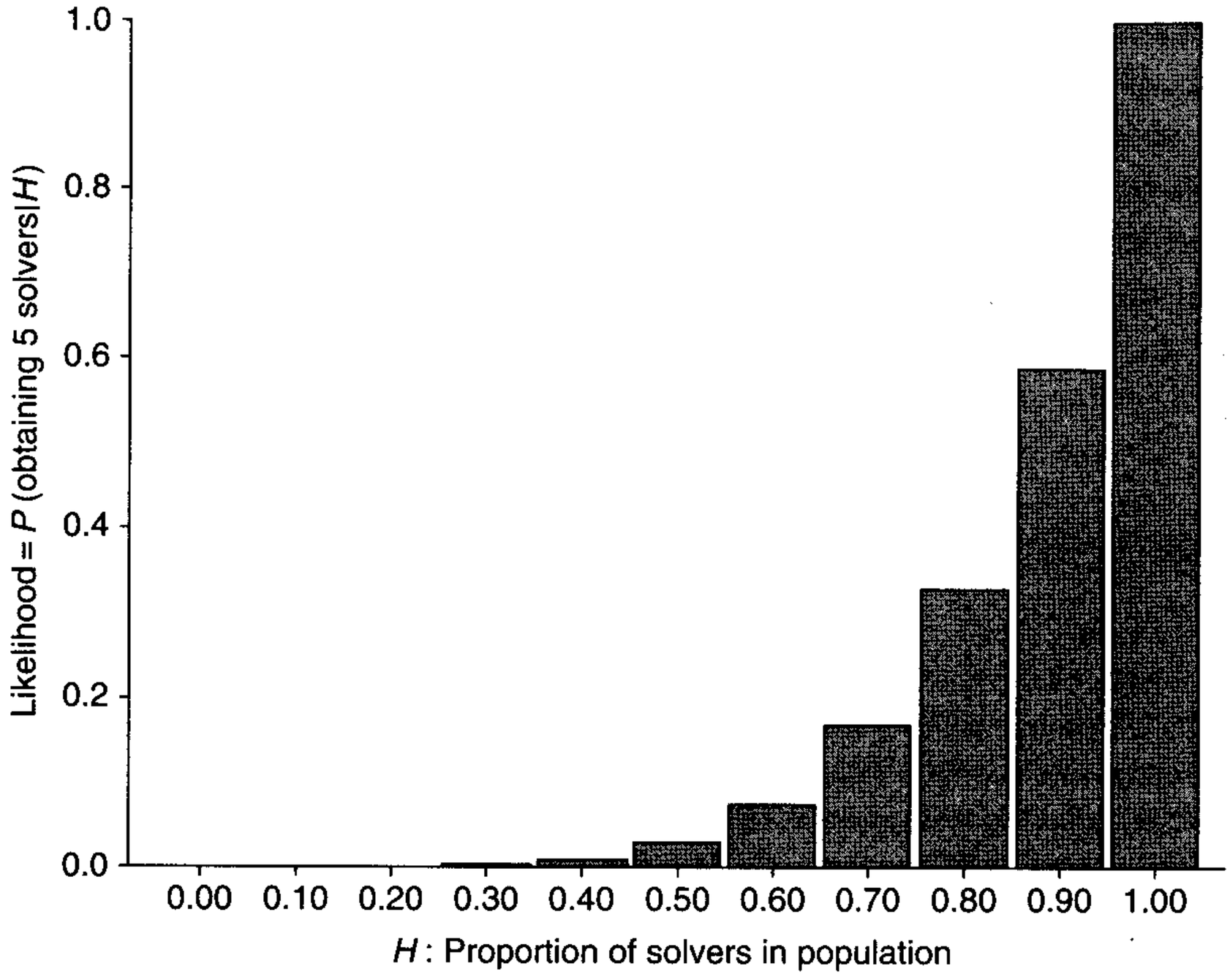
The likelihood is the probability of obtaining these data given a hypothesis. One hypothesis is that the proportion of men who are solvers in our population is 0.1. The likelihood, $P(D|H) = P(\text{obtaining 5 solvers} \mid \text{proportion of solvers} = 0.1)$, is $0.1^5 = 0.000001$. Another hypothesis is that the proportion of men who are solvers is 0.5. Likelihood $= P(\text{obtaining 5 solvers} \mid \text{proportion of solvers is } 0.5) = 0.5^5 = 0.03125$. Figure 4.1 plots the likelihood against different hypotheses. The data could be obtained given many different population proportions, but the data are more probable for some population proportions than others. The data are most probable for a population proportion of 1. The hypothesis that 'the population proportion is 1' has the highest likelihood.

Notice a possible misinterpretation. In everyday speech, saying that a hypothesis has the highest likelihood is the same as saying it has the highest probability. But for statisticians, the two are not the same. The probability of the hypothesis in the light of our data is $P(H|D)$, which is our posterior. The likelihood of the hypothesis is the probability of the data given the hypothesis, $P(D|H)$. We can use the likelihood to obtain our posterior, but they are not the same. Just because a hypothesis has the highest likelihood, it does not mean you will assign it the highest posterior probability. The fact that a hypothesis has the highest likelihood means the data support that hypothesis most. If the prior probabilities for each hypothesis were the same, then the hypothesis with the highest likelihood will have the highest posterior probability. But the prior probabilities may mean that the hypothesis with the greatest support from the data, that is with the highest likelihood, does not have the highest posterior probability.

In the previous chapter, we talked about testing whether a new drug changes blood pressure. We have a group of people treated with the drug, and they have a mean blood pressure of M_d; and we have a placebo group treated only with saline, and they have a mean blood pressure of M_p. The difference in blood pressure between the groups is thus $M_d - M_p$.

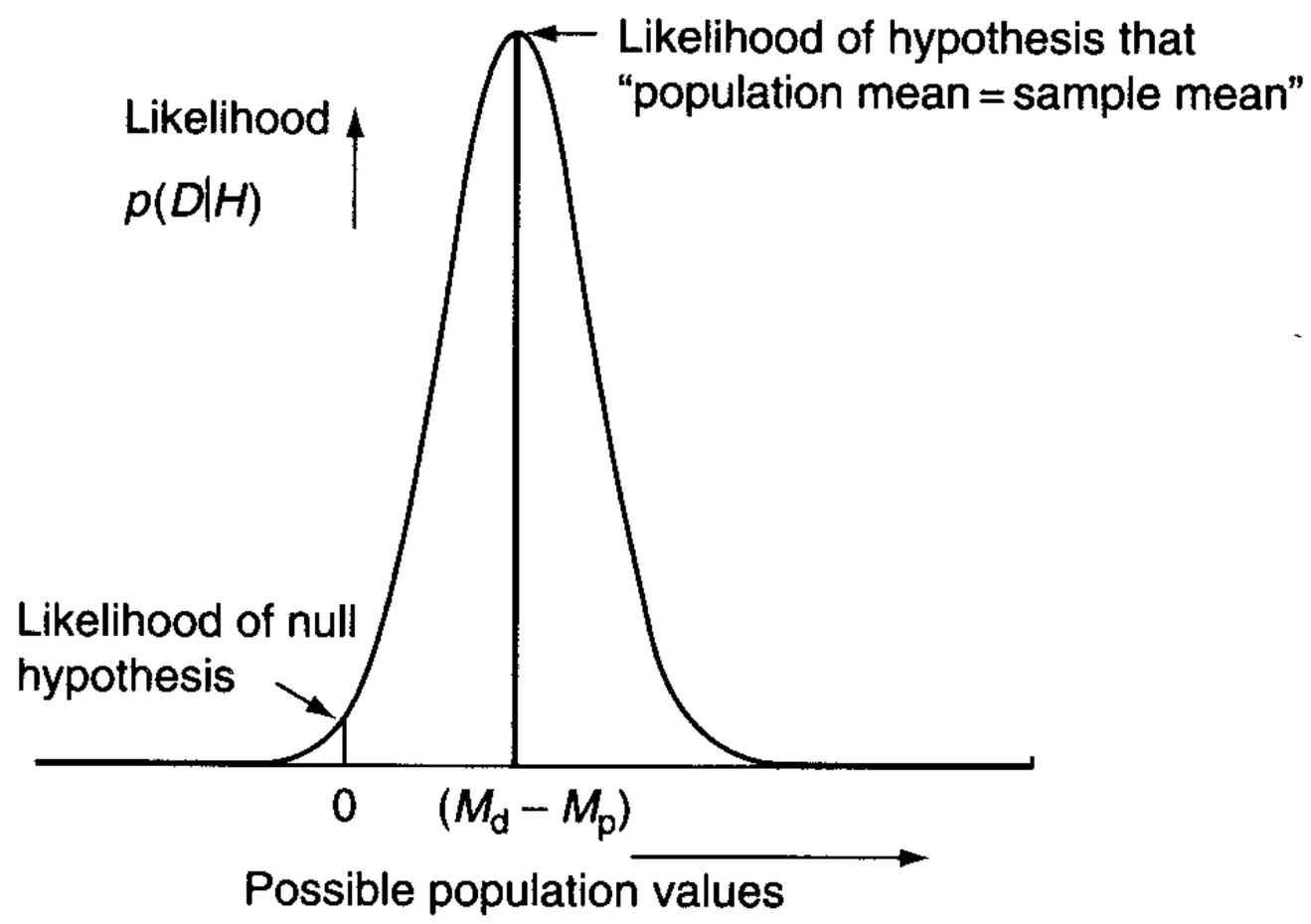
The observed sample (drug–placebo) mean could be obtained if the actual population (drug–placebo) mean was of exactly the same value as the sample mean. But the observed sample mean could also be obtained from a population with a mean slightly different from the sample, or indeed from a population with a mean greatly different from the sample. However,

Figure 4.1



Likelihood function for the data that five out of five men are solvers.

Figure 4.2



Likelihood as a function of difference in population means.

if the population mean were very different from the sample mean, it is not very probable that the population would generate sample means around the value of our sample mean. Figure 4.2 illustrates how likely it would be to obtain the sample mean for different possible population means. Each different possible population mean on the horizontal axis is a different hypothesis, H. The height of the curve is highest when the hypothetical population mean is the sample mean, $(M_d - M_p)$. The height of the curve when the hypothetical population mean is zero is the likelihood of the null hypothesis that the population mean is zero.

If the dependent variable can be assumed to vary continuously (as blood pressure does) – that is, the values do not come in steps – then its distribution is properly called a *probability*

Box 4.2 Probability density versus probability

In general, if a variable varies continuously, the probability distribution of that variable is properly called a probability density distribution function. Blood pressure varies continuously and so its distribution is a probability density distribution. Why invent a special name – probability density – for continuous variables? If the variable really varies continuously, then it can take an infinite number of values. If all values had some positive probability, then the probability of one value or another being true would be

$p(\text{value} = 1 \text{ OR value} = 2.3 \text{ OR value} = 5 \text{ OR } \ldots) = p(\text{value} = 1) + p(\text{value} = 2.3) + p(\text{value} = 5) + \ldots$ for all the uncountable infinite number of values

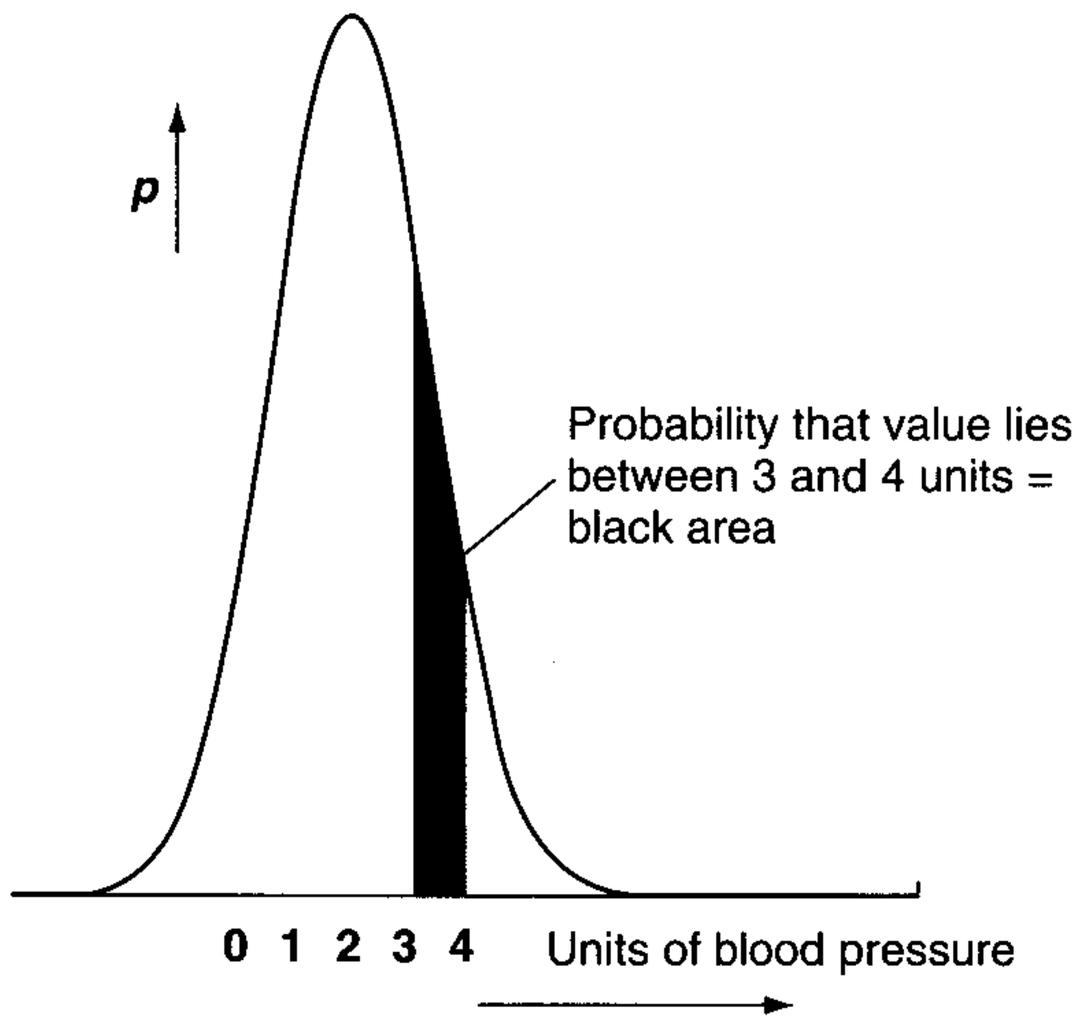
which would be infinite!

But we know from the axioms of probability that probability cannot be greater than one. Thus, for these infinite number of values, we cannot assign a positive probability to each precise value! But we CAN assign a probability to an interval: that is, we can say there is some probability that the true value of blood pressure change lies in the interval 3–4 units of blood pressure. The required probability is the area under the curve between 3 and 4 units, as shown in the figure. A probability density distribution tells you how probable it is that the variable takes a value in any interval.

The probability that the variable has a value in one or other of the intervals is the sum of the areas for each interval. The sum is just the area under the whole curve. Thus, so long as the area under the curve is 1, our probabilities will behave properly!

Note the area under the curve for just one value of blood pressure change, for example for a value of precisely 2 units of blood pressure, is zero. So the probability of the change being *exactly* 2 units is zero. Similarly, the probability of there being exactly no change is also zero.

By convention, $P(A)$ is the probability of A and $p(A)$ is the probability density of A.



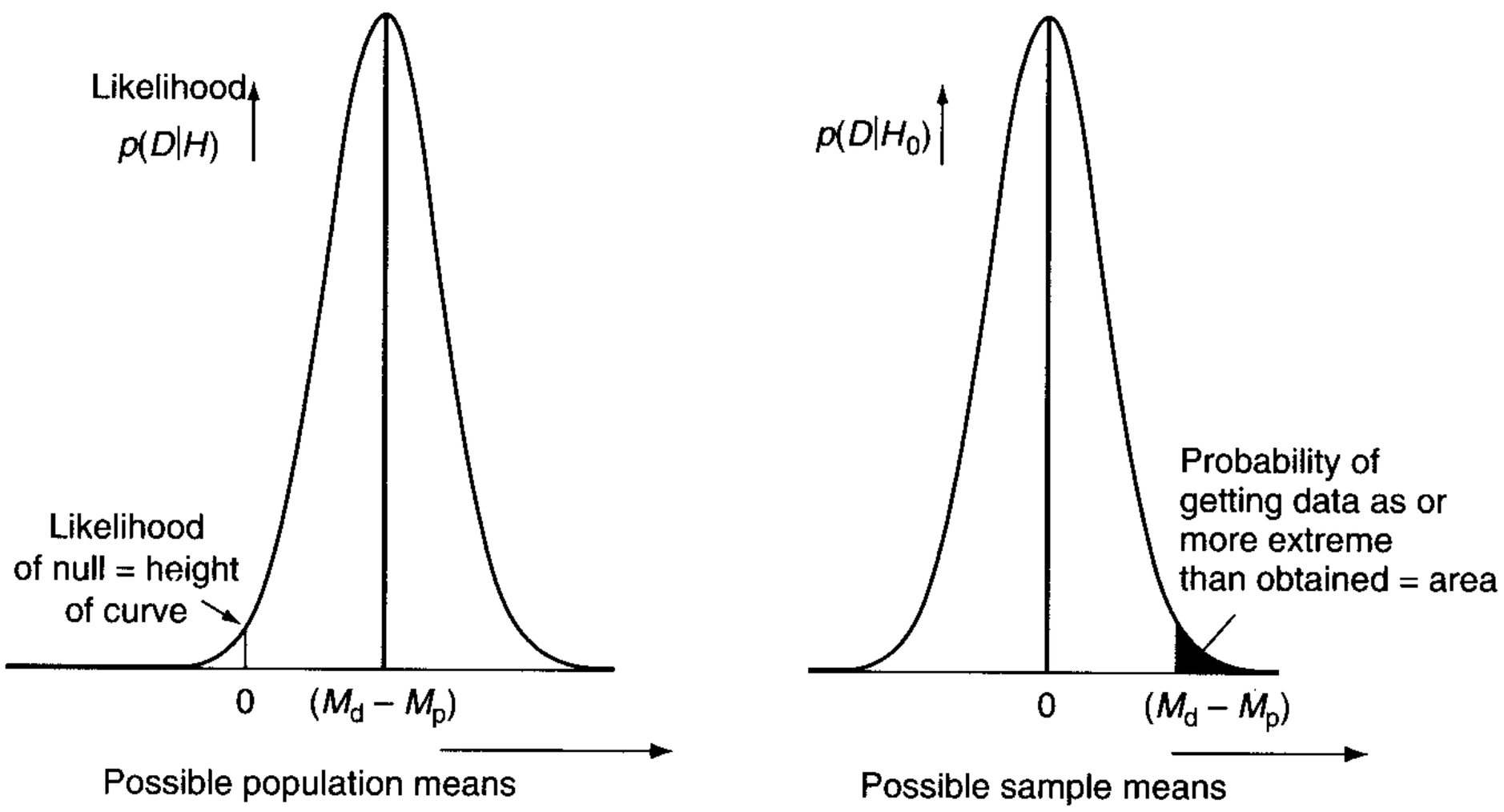
density distribution. See Box 4.2 for an explanation of probability density. A likelihood could be (or be proportional to) a probability density as well as a probability.

In significance testing, we calculate a form of $P(D|H)$. But the $P(D|H)$ used in significance testing is conceptually very different from the likelihood, the $p(D|H)$ we are dealing with here. The 'p-value' in significance testing is the probability of rejecting the null, given the null is really true: P(reject null $|H_0$). Another way of saying the same thing is that the p-value is the probability of obtaining data as extreme or more extreme as we obtained, given the null hypothesis is true: P(obtaining data as extreme or more extreme than $D|H_0$). In calculating a significance value, we hold fixed the hypothesis under consideration – namely H_0 – and we vary the data we might have obtained (we ask, what is the probability – or probability density – of obtaining this sample mean or this sample mean or this sample mean . . . given H_0). The likelihood is p(obtaining exactly this $D|H$), where H is free to vary; but the D considered is always exactly the data obtained.

Figure 4.3 illustrates the differences. For the curve on the left, what varies along the horizontal axis is different possible population values. Each value corresponds to a different H; it is a hypothesis about what the true population value might be. For the curve on the right, what varies along the horizontal axis is different possible sample values, that is different possible data. On the left, the likelihood of the null is the height of the curve corresponding to the null hypothesis. On the right, we take a corresponding part of the curve (the height of the curve above $M_d - M_p$ is the same value as the likelihood); however, we are not interested in the height of the curve at this point but the area under the curve beyond this point. The area represents the probability of obtaining data as large as the obtained mean or larger given the null hypothesis.

Likelihood analysis regards the data as fixed but the hypothesis can vary. Significance testing regards the hypothesis as fixed but the data can vary. In calculating the likelihood, we are interested in the height of the curve for each hypothesis. In significance testing, we are interested in the 'tail area', the area under the curve beyond the obtained data. This area is the probability

Figure 4.3



The likelihood in Bayes versus significance testing.

of obtaining our data OR data more extreme – data that we did not observe but might have. Likelihood reflects just what the data were (the curve is plotted only for the actually obtained data); significance tests, using tail areas, reflect what might have happened but did not (the data might have been more extreme, but were not). In significance testing, we make a black and white decision: Is the tail area smaller than our preset alpha value (e.g. 5%) or not? In contrast, likelihoods give a continuous graded measure of support for different hypotheses.

In significance testing, tail areas are calculated in order to determine long-run error rates. The aim of classic statistics is to come up with a procedure for making decisions that is reliable, which is to say that the procedure has known controlled long-run error rates. To decide the long-run error rates, we need to define a collective (see Chapter 3). What exactly is it that defines the procedure that constitutes an element of the collective? A procedure could be performing one *t*-test. The elements of the collective are individual acts of performing a *t*-test, and the error probabilities can be controlled by rejecting the null when the tail area is less than the preset alpha, as discussed above. But if the procedure is conducting a set of five *t*-tests, the elements of the collective are sets of five *t*-tests performed at once (often called a 'family' of *t*-tests), and we want to control the error of rejecting any one of the five possible null hypotheses when they are true (we want to control 'family-wise error rate'). Now we need to adjust the tail area. For example, if we used the Bonferroni adjustment, we would multiply the tail area of each test by five, and only reject the null for that test if this adjusted area was less than our preset alpha. (See Chapter 3 for discussion.) The tail error needs to be adjusted also according to the stopping rule. If we had a different stopping rule, we might have stopped at a different time. In classic statistics, we need to take into account what else we might have done: performed one test or five? When else might we have stopped? Further we need to know whether the test is post hoc or planned (what came first – the explanation or the data)? The likelihood – the probability (or probability density) of obtaining exactly these data given a hypothesis – is clearly the same whatever other tests you might have done, whether you decide to stop now or carry on collecting data, and whatever the timing of your explanation (before or after the data). The insensitivity of the likelihood to what other tests you might have done, to stopping rules, and to the timing of the explanation is a profound philosophical and practical difference between Bayesian and classical statistics which we will discuss in detail later. The sensitivity of classical statistics to multiple testing, stopping rules, and timing of explanation is how classical statistics violates the likelihood principle: In these ways, classical statistics regards more aspects of the data than just the likelihood as relevant to inference.

Bayesian Analysis

Bayes' theorem says that posterior is proportional to likelihood times prior. We can use this in two ways when dealing with real psychological data. First, we can calculate a credibility interval, which is the Bayesian equivalent of a confidence interval. Second, we can calculate how to adjust our odds in favour of a theory we are testing over the null hypothesis in the light of our experimental data (the 'Bayes factor'), which is the Bayesian equivalent of null hypothesis testing. We discuss each in turn.

Credibility intervals

We wish to test the extent to which 1 g of a new drug can change blood pressure. Each possible value of population change in blood pressure is a hypothesis. You need to decide

what your prior probability density is for each of these hypotheses. Before we have collected data, presumably you have some, albeit vague, idea of what sort of changes in blood pressure are relatively more probable than others. Let us say a normal distribution would not violate the shape of your prior too much, that is you think certain values are reasonably probable and more extreme values less probable in a symmetric way. The value you think is most probable defines the centre (or mean) of your prior distribution. The spread in your values – the standard deviation – can be assigned by remembering: You should think that plus or minus one standard deviation from the mean has a 68% probability of including the actual population value; and you should think that plus or minus two standard deviations has a 95% probability of including the actual population value. You should be virtually certain that plus or minus three standard deviations includes the true population value. If the standard deviation is infinite (or just very large compared to what is practically possible), you think all population values are equally likely. This is called a 'flat prior' or 'uniform prior'. You have NO idea what the population value is likely to be. That is OK, if that is how you feel. Remember there are no 'right' answers: This is YOUR prior! In sum, in choosing a prior decide (a) whether your prior can be approximated by a normal distribution and if so (b) what the mean of this distribution is (call it M_0) and (c) its standard deviation (call it S_0).

Conduct the following exercise before you continue reading. Consider a research question you are interested in. The research will probably involve estimating the value of a population parameter; for example, the difference in means between one condition and another. Construct your prior probability distribution for the parameter in question. See Box 4.3 to help you.

Box 4.3 An example in constructing a prior

Often as a psychologist you will be interested in the mean difference between two conditions. For example, you may be interested in people's self-esteem on a 0–6 scale after seeing an advert with a skinny model compared with after seeing an advert with a model of average body shape. You may believe that on average, self-esteem will be lower after the advert with a skinny model than the average model. Label changes in that direction as positive. What size change do you consider most likely? One thing to bear in mind is that a 0–6 scale does not allow much room for change. If people normally have a self-esteem in the middle of the scale (say a 3), and assuming that seeing the average model does not change the viewer's self-esteem, then the most self-esteem can be lowered by on average is 3 points. That is the very most, and *everyone* would have to reduce maximally for the population change to be 3. So even average changes of 1 or 2 points would be reasonably large. To determine quite what change is most likely, you might like to consider previous similar studies. In any case, your final guess is your personal guess and completely up to you.

You feel optimistic about the strength of your manipulation and you pick 1 as the mean of your prior. Next consider how uncertain you are about the population mean, that is the standard deviation of your prior. Ask yourself what is the probability that the population mean could be less than zero, that is that the average change goes in the opposite direction. You might think that the probability of the change being negative is only a few percent. If we can call it 2.5% without doing your intuitions an injustice, then the difference between the mean and zero is two standard deviations (the normal curve has 2.5% of its area less than two standard deviations below the mean). With a mean of 1, it follows the standard deviation for the prior is 0.5. *It is not that you think 2.5% of people will score in the negative direction.* It is that you think the probability of the population mean being less than zero is 2.5%. You may think that the probability the population mean being less than zero is zero, but also think that a good proportion of people in the population would score below zero.

Box 4.3 *continued*

You need to check if you are really happy with that standard deviation. If we go two standard deviations above the mean ($1 + 2 \times 0.5 = 2$), you should believe there is a 2.5% probability that the true population mean is above 2. If that does not seem reasonable, you might consider revising either the mean or the standard deviation of your prior. Or you might consider revising the idea that your prior can be represented by a normal distribution (e.g. your prior may not drop off symmetrically about the mean). But see if you can produce a reasonable fit with a normal before giving it up; the exact details of your prior will not be important if you have a decent amount of data, so the aim is only to see if you can capture your prior intuitions without doing them gross violation.

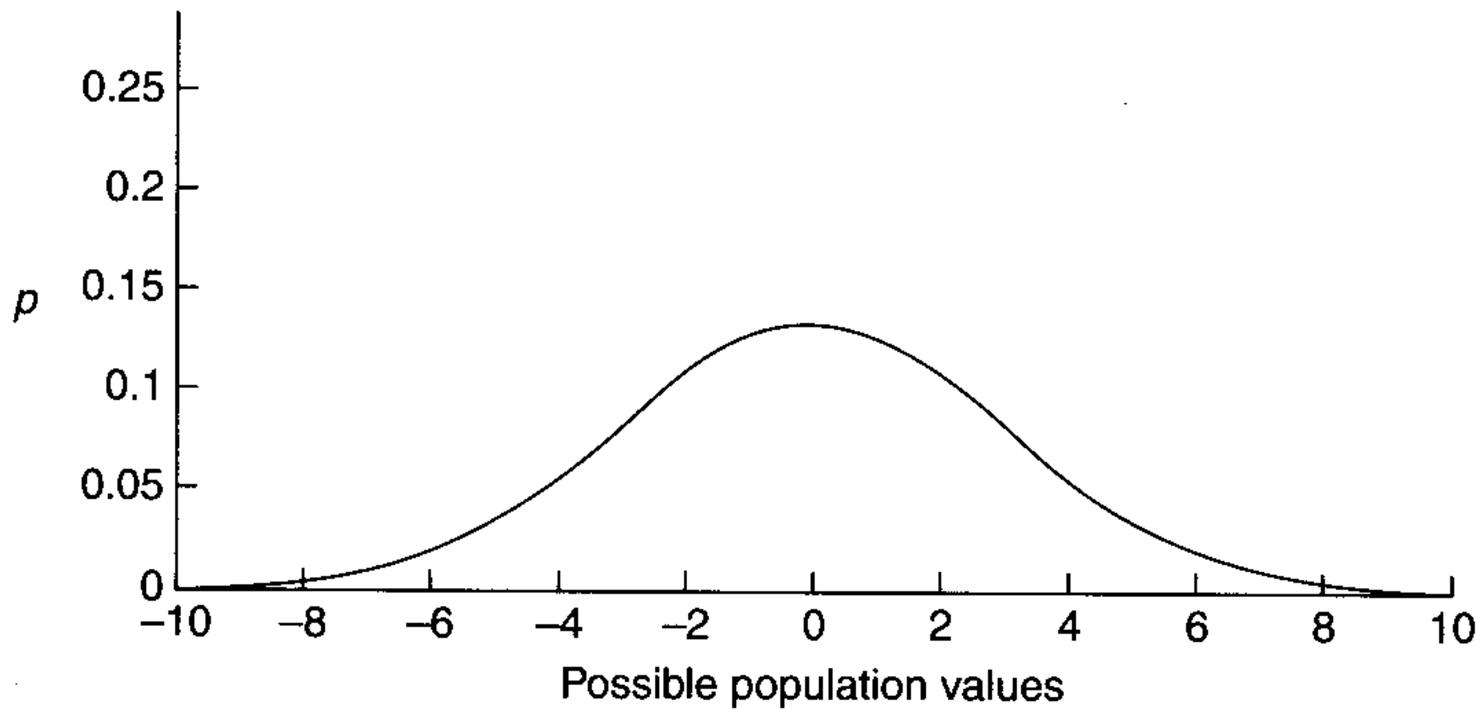
You might think that the probability of the population mean being less than zero is more like 10%. Normal tables show that 10% of the area of a normal curve lies below 1.28 standard deviations below the mean. With a mean of 1 for your prior, it follows that 1.28 SDs = 1, so one SD = 1/1.28 = 0.78 units. Using that figure as a working value, see if you are happy with its consequences. For reference, 20% of the area of a normal lies below 0.84 standard deviations below the mean, and 30% of the area below 0.52 standard deviations below the mean. So if you thought there was a 30% probability of the population mean being below zero, you know there are 0.52 standard deviations between 0 and your mean. Hence, one standard deviation = mean/0.52.

If you settle on a mean of 1 and an SD of 0.78, there is 10% probability of the population mean being below zero. Conversely, there is a 90% probability of the population mean being above zero. *Thus, there is a zero probability of the population mean being exactly zero.* Of course, the probability of the mean being exactly zero is always zero, no matter what your mean and standard deviations are. (There is a finite probability of the mean being a small interval around zero. For example, for the mean and SD chosen, the prior probability of the population mean being within 0.1 scale points around zero is 0.09. Can you work this out yourself using normal tables?)

Sketching a prior for your experiment is an excellent exercise to carry out before you run your experiment, whether or not you accept Bayesian statistics. Thinking hard about effect sizes is important for any school of statistical inference, but sadly a process often neglected.

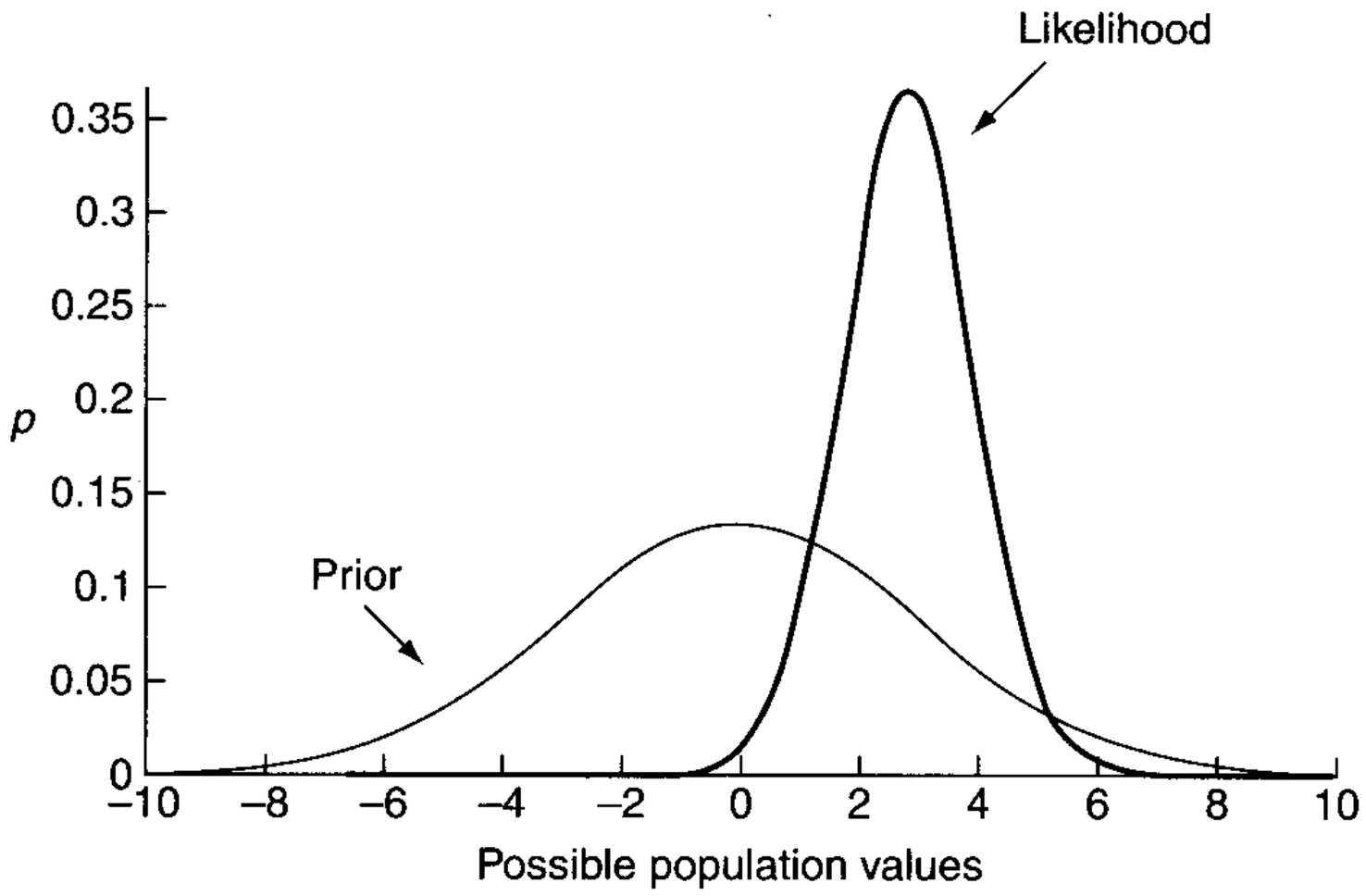
Figure 4.4 shows a possible prior for the effectiveness of a new drug. The units are whatever you measure effectiveness of the drug in; for example, it could be millimetres of mercury (mmHG) as a measure of blood pressure change. This prior shows that you think a mean of zero effect is most plausible ($M_0 = 0$) and you are virtually certain that the true effect, whatever it is, lies between −9 and +9 units (here $S_0 = 3$).

Figure 4.4



A possible prior.

Figure 4.5



Prior and likelihood.

Figure 4.4 shows what you felt before you collected data. You collect data from a normally distributed population. Your sample has a mean of 2.8 and a standard error of 1.09. Assuming your data are based on 30 or more observations, you can represent your likelihood as a normal distribution with a mean of 2.8 and a standard deviation of 1.09, as shown in Figure 4.5. See Box 4.4 for calculating your likelihood distribution.

Box 4.4 Determining the likelihood

Imagine each subject took part in two conditions; your dependent variable is continuous and roughly normally distributed. For each subject, calculate the difference between the two conditions. The mean for your likelihood function is the mean of these difference scores. The standard deviation of the likelihood function is the standard *error* of the difference scores. If the difference scores have a standard deviation S, then the standard error is $S/\sqrt{n}$, where n is the number of subjects. S is the standard deviation of the *difference* scores, not of the scores in any one condition. You might wish to perform a Bayesian analysis on data reported in a paper comparing two conditions using a within-subjects design (i.e. each subject participated in each condition). The tables in a paper typically report the standard deviation of the scores in each condition but do not give you what you want: the standard error of the difference scores. There is a trick you can use to get the information though. The paper may report a t-test for the difference. $t =$ (mean difference)/(standard error of difference). Thus, standard error of difference = (mean difference)/t.

If each subject took part in just one of two conditions, the standard deviation of the likelihood function is still the standard error of the difference. But now we cannot take a difference score for each subject because each subject participated in just one condition. Let the conditions have standard deviations S_1 and S_2, and also have n_1 and n_2 subjects.

The standard error of the difference is now given by $\sqrt{(S_1^2/n_1 + S_2^2/n_2)}$. Alternatively, the paper may have reported a t-test for the difference. $t =$ (mean difference)/(standard error of difference), as before. Thus, standard error of difference = (mean difference)/t, as before.

Box 4.4 *continued*

We are considering examples of mean differences where the means can be considered (roughly) normally distributed because the equations turn out to be simple for Bayesian analyses in this case. But bear in mind Bayesian procedures can be used for all the common situations for which researchers need statistics, from the simple to the complex, with new solutions coming out all the time. The introductory textbook by Berry (1996) explains how to construct priors, likelihoods and posteriors for data involving proportions (as well as for means from roughly normal distributions). McCarthy (2007) illustrates the use of free software (winBUGS) for the Bayesian analysis of many designs, from the simple to the very complex.

We have considered your prior uncertainty in the mean of the population distribution, but we have not mentioned your uncertainty in its standard deviation. Technically, the equations we will use in Box 4.5 assume the population standard deviation is known, which is rather unlikely. When the standard deviation is unknown (only estimated by your sample), Berry recommends a correction when $n < 30$. Transform the standard deviation, S, of scores in each group by an amount $S' = S(1 + 20/n^2)$ and use S' in the above equations for calculating likelihood. The WINbugs software described by McCarthy allows a more principled Bayesian solution: One can construct priors for both mean and standard deviation, and these are jointly used in determining the posteriors for both mean and standard deviation.

The equations given in this chapter will cover you in many cases though; namely, wherever a roughly normal distribution is involved or a t-test would be used using orthodox statistics. For example, Pearson correlations can be transformed to be normal using Fisher's transform, r', given in Box 3.7. In that case, the likelihood function has a mean equal to the observed r' and a standard deviation given by the SE given in the box.

Box 4.5 Formulae for normal posterior

Mean of prior $= M_0$

Mean of sample $= M_d$

Standard deviation of prior $= S_0$

Standard error of sample $=$ SE

Precision of prior $= c_0 = 1/S_0^2$

Precision of sample $= c_d = 1/\text{SE}^2$

Posterior precision $c_1 = c_0 + c_d$

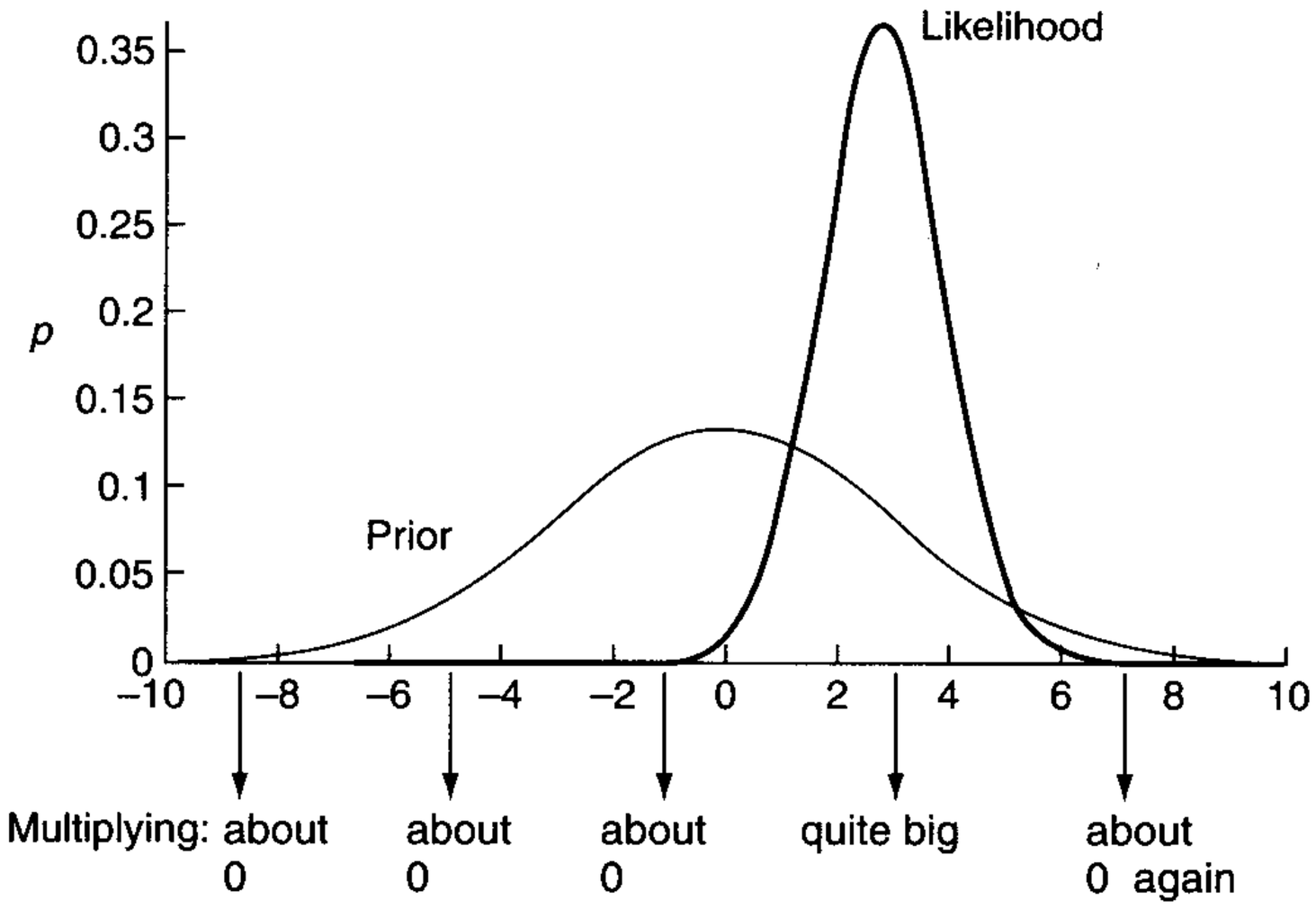
Posterior mean $M_1 = (c_0/c_1)M_0 + (c_d/c_1)M_d$

Posterior standard deviation, $S_1 =$ square root$(1/c_1)$

Bayes tells us that $p(H|D)$ is proportional to $p(D|H) \times p(H)$; that is that the posterior is given by the likelihood times the prior. Each value on the horizontal axis of Figure 4.5 is an H, a hypothesis about the population value of the drug's effectiveness. For each H, we need to multiply the likelihood by the prior to get the posterior for that H, as shown in Figure 4.6. (You might worry about what the constant of proportionality should be. In fact, we need not worry about it. We just make sure the area under the posterior distribution is equal to one.)

Notice how similar the posterior is to the likelihood. For a reasonably diffuse prior (i.e. one representing fairly vague prior opinions), the posterior is dominated by the likelihood, i.e. by the data. If you started with a flat or uniform prior (you have no opinion concerning which values are most likely), the posterior would be identical to the likelihood. Further, even if people started with very different priors, if you collect enough data, as long as the priors were smooth and allowed *some* non-negligible probability in the region of the

Figure 4.6



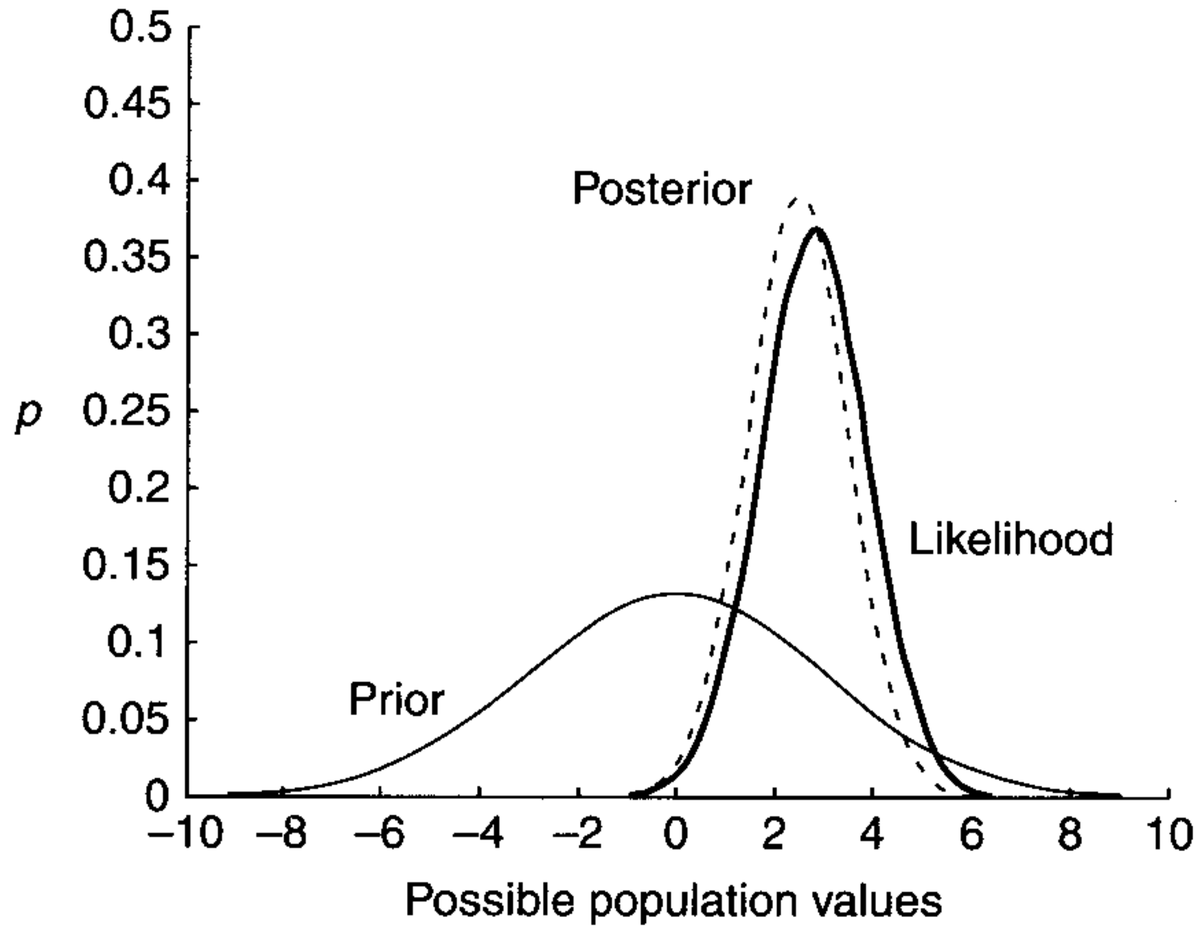
Multiplying likelihood by prior.

true population value, the posteriors, being dominated by the likelihood, would come to be very similar (Edwards et al., 1963, discuss the conditions more precisely). In this sense, Bayesian statistics emphasizes the objective nature of science: Different starting opinions will be brought together by sufficient data. The likelihood, representing the data, comes to dominate conclusions.

It turns out that if the prior and likelihood are normal, the posterior is also normal. There also turn out to be simple formulae for determining the mean and standard deviation of the posterior given the mean and standard deviation of the prior and likelihood. One doesn't have to literally go through and multiply prior and likelihood values for every population value. Box 4.5 gives the formulae.

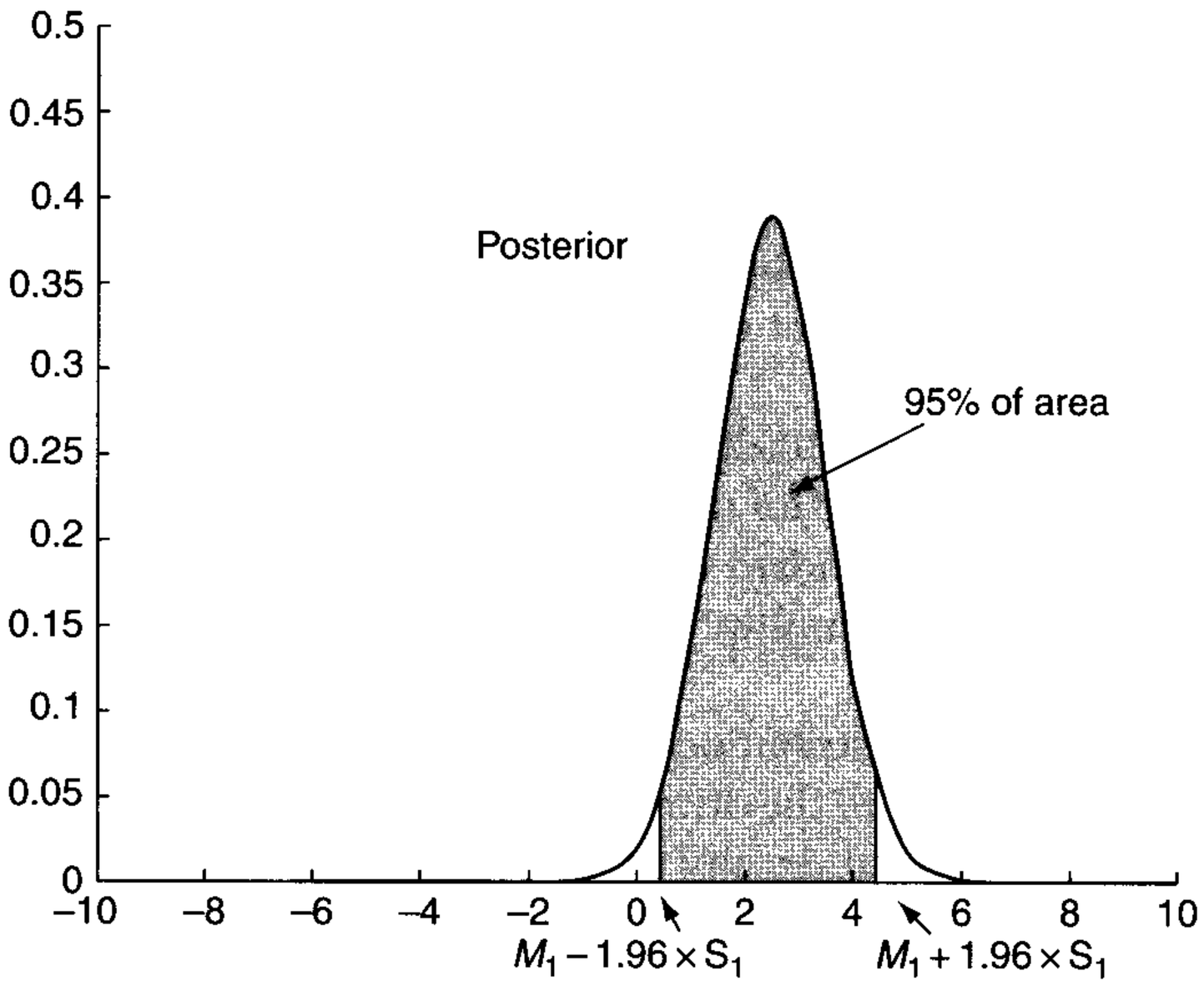
Having found the posterior distribution, you have really found out all you need to know. It can be convenient to summarize the distribution in terms of a *credibility interval* (also called a probability interval or a highest density region, or HDR). For example, we could find the range of drug effects which have a 95% probability of including the true drug effect. We generally stipulate that this is centred on the mean of the posterior distribution. You will remember that in a normal distribution, the mean plus or minus 1.96 standard deviations includes 95% of the area. Thus, if the standard deviation of the posterior is S_1, and the mean is M_1, then going from $(M_1 - 1.96 \times S_1)$ to $(M_1 + 1.96 \times S_1)$ defines the 95% credibility interval. The posterior in Figures 4.7 and 4.8 has a mean of 2.47 and a standard deviation of 1.10 (confirm these values using the formulae in Box 4.5). The 95% credibility interval thus runs from 0.5 to 4.5 units. That is, there is a 95% probability that the true effect of the drug lies in the interval from 0.5 to 4.5. Our prior had a mean of 0 and a standard deviation of 3. A 95% credibility interval based only on the prior runs from $(0 - 1.96 \times 3)$ to $(0 + 1.96 \times 3)$, that is, from -5.9 to $+5.9$. Because we have collected data, we have gained in precision: as we have just noted, with the posterior the interval runs from 0.5 to 4.5.

Figure 4.7



The posterior distribution that results from the prior and likelihood we have been considering.

Figure 4.8



The 95% credibility interval

If you needed information more precise than this, you could simply collect more data until the credibility interval had shrunk to a sufficiently small precision. How precise you want the answer to be is just up to you; you can collect data until you are satisfied, and there is no need to define in advance how precise that has to be. Just stop when you know enough for your purposes at that point in time. For example, if after some more participants had been run your 95% credibility interval went from 2.9 to 3.2, you might decide to

stop: A drug that changes blood pressure by around +3 units on average will do the job you want.

If the priors were flat, the 95% credibility interval would be numerically the same interval as the 95% confidence interval of Neyman–Pearson. But the meanings of the intervals are very different. The confidence interval is associated with an objective probability: If you repeated your experiment an indefinite number of times, the interval calculated each time would include the true population value 95% of the time. But you cannot make any claim about how likely THIS confidence interval is to enclose the true population mean. You cannot really be 95% confident that the population value lies in the 95% confidence interval. But as Savage (1962, p. 98) said in criticism of the Neyman–Pearson approach 'The only use I know of for a confidence interval is to have confidence in it!' The Bayesian credibility interval allows you to be confident in it.

Further, there is a profound practical difference between confidence and credibility intervals. The confidence interval will have to be adjusted according to how many other tests you conducted, under what conditions you planned to stop collecting data, and whether the test was planned or post hoc. The credibility interval is unaffected by all these things. On the other hand, the credibility interval IS affected by any prior information you had (because such information will change your prior distribution) whereas the confidence interval is not. All these respective differences are seen as strengths by each side; we will discuss the contrasting intuitions later.

The Bayes factor

There is no such thing as significance testing in Bayesian statistics. All one often has to do as a Bayesian statistician is determine posterior distributions. However, sometimes people like to compare the probability of their experimental theory to the probability of the null hypothesis. We can do this with the 'Bayes factor', the logic of which is now described. It is the Bayesian equivalent of null hypothesis or significance testing. Notice in the example used for calculating a credibility interval, the prior and posterior were both probability *density* distributions. Because the dependent variable plotted horizontally was continuous, the distributions only allow probabilities to be assigned to intervals. Any single point, such as a blood pressure change of exactly 2.3331564 units, had a probability of zero. Thus, the hypothesis that the population blood pressure change was exactly zero also has a probability of zero. That is, the null hypothesis both before and after data collection had a probability of zero. In such a context, it does not make much sense to talk about accepting or rejecting the null hypothesis. And typically in real research contexts, it indeed makes no sense to think of an independent variable having absolutely no effect on a dependent variable, or a correlation between two psychological variables being exactly zero, whatever the truth of one's favourite theory. Even given your theory is false, there are bound to be other reasons for why there is some at least tiny relation between the variables. But sometimes we might want to assign some finite non-zero probability to a particular hypothesis, like the null hypothesis, and see how that probability changes as we collect data. This is what the Bayes factor allows us to do.

Bayes says that $P(H|D)$ is proportional to $P(D|H) \times P(H)$. Thus to consider two particular hypotheses, your experimental hypothesis H_1 and the null H_0, we have

$$P(H_1|D) \text{ is proportional to } P(D|H_1) \times P(H_1) \tag{4.5}$$

$$P(H_0|D) \text{ is proportional to } P(D|H_0) \times P(H_0) \tag{4.6}$$

Dividing (4.5) by (4.6),

$$P(H_1|D)/P(H_0|D) = P(D|H_1)/P(D|H_0) \times P(H_1)/P(H_0)$$

$$\text{Posterior odds} = \text{likelihood ratio} \times \text{prior odds}$$

The likelihood ratio is (in this case) called the Bayes factor B in favour of the experimental hypothesis. Whatever your prior odds were in favour of the experimental hypothesis over the null, after data collection multiply those odds by B to get your posterior odds. If B is greater than 1, your data support the experimental hypothesis over the null; if B is less than 1, your data support the null over the experimental hypothesis. If B was about 1, then your experiment was not sensitive.[3] You did not run enough participants, so the data do not distinguish your experimental hypothesis from the null. (Notice how you automatically get a notion of how sensitive your experiment was. Contrast just relying on p-values in significance testing.) The Bayes factor gives the means of adjusting your odds in a continuous way; you are not being asked to make a black and white decision. Arguably, in reality we typically do not make black and white decisions accepting or rejecting hypotheses; we let the data more or less strongly nudge our strength of belief one way or the other. If we really made black and white decisions, there would be little need to replicate experiments exactly.

Example of calculating the Bayes factor

We will now consider an example with real data. Some years ago, I ran a series of experiments to test the theory of morphic resonance postulated by English biologist Rupert Sheldrake. Morphic resonance will be a useful example to contrast Bayesian with classical analyses because people start with wildly different prior odds in favour of Sheldrake's theory. Morphic resonance is an idea so radically in conflict with existing scientific paradigms, some scientists are like Kuhnian Normal scientists and react in abject horror to Sheldrake's theory; on the other hand, a few, acting in a more Popperian manner, note with interest it is a bold conjecture that could be tested.

According to the theory of morphic resonance, any system by assuming a particular shape becomes associated with a 'morphic field'. The morphic field then plays a causal role in the development and maintenance of future systems. 'Morphic' means shape; the effect of the field is to guide future systems to take similar shapes.

The subsequent systems 'resonate' with previous systems through the fields. The effect is stronger the more similar the future system is to the system that generated the field; and the effect is stronger the more times a form has been assumed by previous similar systems. The effect occurs at all levels of organization, for example from crystallization of chemical substances to brain patterns. In terms of brains, for example, if a group of rats have solved a certain maze, future rats should find the same maze easier because they can resonate with the brains of the previous successful rats, a type of rodent ESP (Sheldrake, 1981; 1988).

Repetition priming is a phenomenon well known to experimental psychologists: People identify a stimulus more quickly or more accurately with repeated presentation of the stimulus. This can be shown in, for example, the lexical decision task, in which people

3. This statement is typically true but not always. If your theory predicts just one possible population value, and the true population value lies *exactly* in between the null and the value predicted by theory, the Bayes factor tends to 1 as you collect more and more observations!

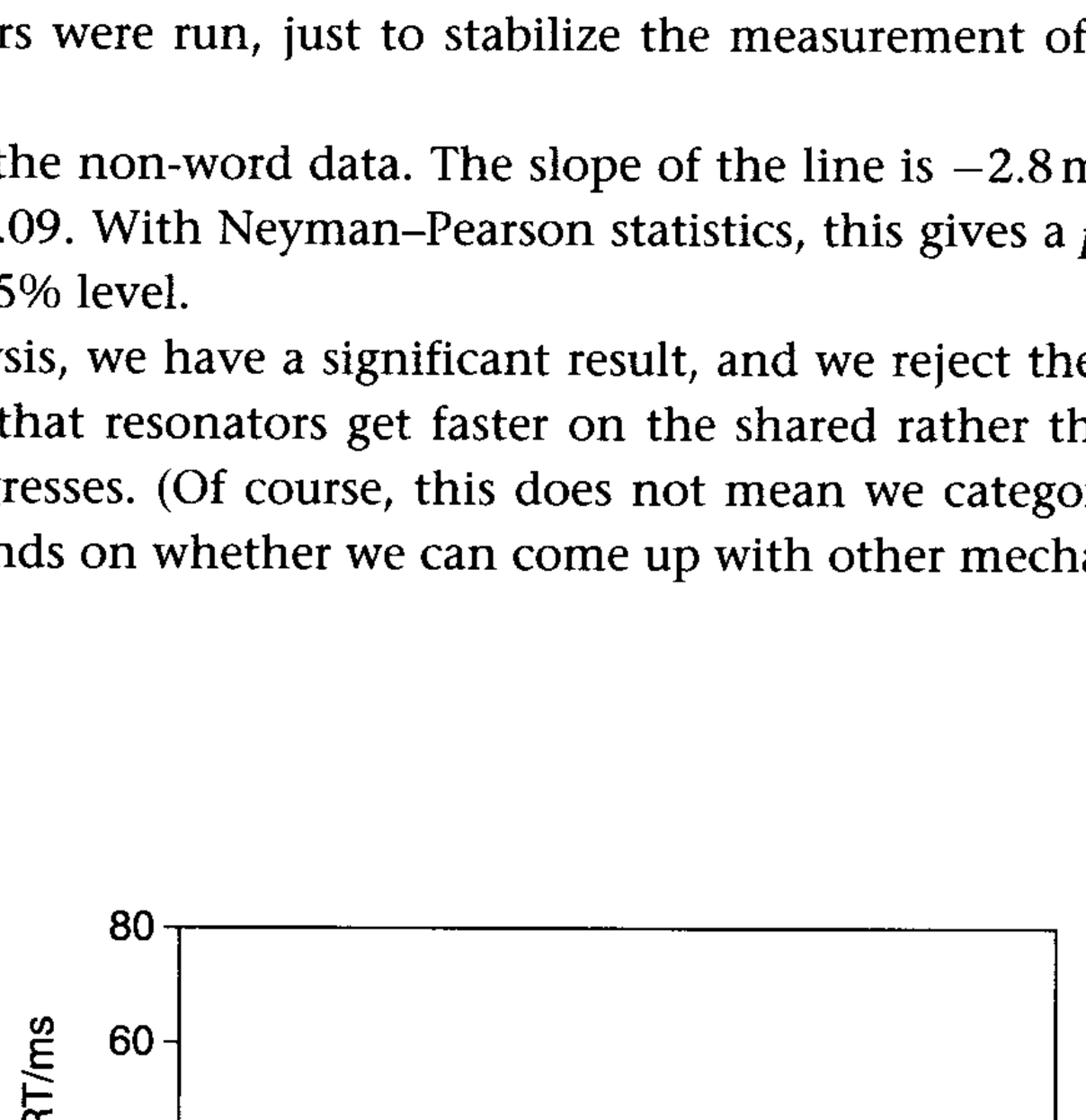
have to decide whether a presented letter string makes a meaningful English word or not, with the letters in the presented order. People are faster to make a lexical decision when a letter string is repeated. The theory of morphic resonance predicts that repetition priming should occur: The person will resonate with themselves in the past facilitating future performance. Of course, other theories also predict repetition priming should occur. But there is a unique prediction of morphic resonance not made by other theories favoured by psychologists; namely, that there should be repetition priming between separate people (i.e. a form of ESP) (see box 4.6).

I ran the following experiment in 1989. One set of words and non-words was called the shared set; another set was called the unique set. The first, 10th, 20th, 30th, ... and 100th participant performed a lexical decision task on both the unique and shared sets. These participants were called 'resonators'. The intervening participants, the 'boosters', were just exposed to the shared set. Thus, over successive participants, the shared stimuli were receiving morphic resonance at 10 times the rate as the unique stimuli. Thus, resonators should get progressively faster on the shared rather than unique stimuli. (At the end of the experiment, some extra resonators were run, just to stabilize the measurement of the morphic field at that point in time.)

Figure 4.9 shows the non-word data. The slope of the line is −2.8 ms per resonator, with a standard error of 1.09. With Neyman–Pearson statistics, this gives a p of 0.018, significant at the conventional 5% level.

On a classic analysis, we have a significant result, and we reject the null hypothesis. We categorically accept that resonators get faster on the shared rather than unique stimuli as the experiment progresses. (Of course, this does not mean we categorically accept *morphic resonance*. That depends on whether we can come up with other mechanisms for the effect.)

Figure 4.9



Non-word data.

Now consider the Bayesian analysis. If 'MR' stands for the theory that morphic resonance is true:

$$P(MR|D)/P(H_0|D) = p(D|MR)/p(D|H_0) \times P(MR)/P(H_0)$$

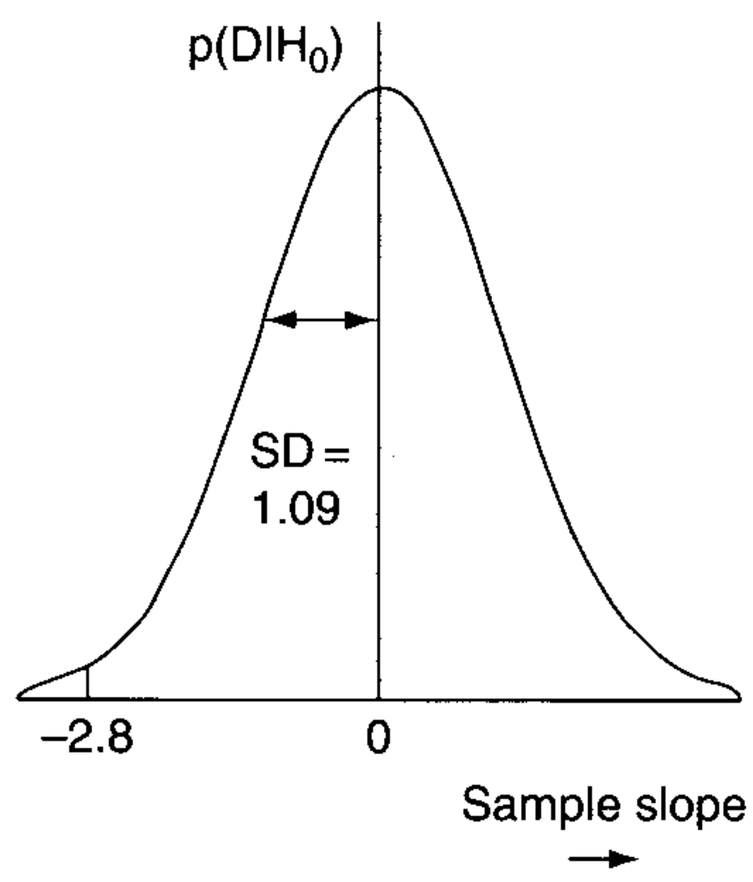
$$\text{Posterior odds} = \text{Bayes factor, } B \times \text{Prior odds}$$

Before we continue, remind yourself of your personal prior odds in favour of morphic resonance you determined previously.

To calculate the Bayes factor, we need to determine $p(D|H_0)$ and $p(D|MR)$. H_0 is that the population slope is zero. $p(D|H_0)$ can be obtained by plotting the probability density of getting different slopes in a sample given the null hypothesis is true, as illustrated in Figure 4.10. For these data, we will assume a normal distribution. Naturally, the sample mean with the highest probability density is zero, the hypothetical population mean. The standard deviation is the standard error (the standard error in the study was 1.09). $p(D|H_0)$ is just the height of this normal curve for a slope of -2.8 (see Figure 4.12). The height here is 0.014. That is, $p(D|H_0) = 0.014$. In sum, to determine $p(D|H_0)$ decide if a normal distribution is adequate for the shape of $p(D|H_0)$. Next you need to know your sample mean and its standard error. If you use the program called 'Bayes factor' given at the end of the chapter, enter the sample mean and standard error, it will tell you $p(D|H_0)$.

The difficult part is to determine $p(D|\text{MR})$. Morphic resonance is consistent with a number of population slopes; in fact, at first blush, any slope < 0. The theory says that resonators will get faster on shared relative to unique words, but the theory does not predict a precise value (in this respect, its quantitative vagueness, it is like almost all theories in psychology). But in fact, morphic resonance does not allow any slope < 0. The between-subject priming (shown in Figure 4.9) must be less than the priming shown by a person resonating with themselves (because people are more similar to themselves than to other people). In fact, boosters in the study saw each stimulus three times, so we know what the effect of repetition was within a person. People sped up by 20 ms with a repetition. So the speed up in reaction time from one resonator to the next certainly could not be more than 10 people $\times$ three repetitions per

Figure 4.10



The probability of the data (reported in the text) given the null hypothesis.

Figure 4.11

Simplest assumption for p(population slope|MR).

person × 20 ms per repetition = 600 ms. The slope predicted by morphic resonance cannot be steeper than −600 ms per resonator.

Now assume that given we accept morphic resonance is true, we have no preference whatsoever for thinking any of the slopes in the range from 0 to −600 ms (exclusive) are more likely than any others (see Figure 4.11). This is an implausible assumption, but we will assume it now to see the consequences. Shortly, we will make a more plausible assumption about how likely different slopes are, assuming morphic resonance is true.

To go from p(population slope|MR) to p(observing a slope|MR), i.e. $p(D|\text{MR})$, we need to smear each point on the graph in Figure 4.13 by the standard error of the sample. Figure 4.10 illustrates p(observing a slope| null hypothesis) where the null only allows one population value, namely a slope of zero. The data are distributed around this point. Just so, assuming morphic resonance, each population slope can produce a sample slope somewhat lower or higher than the population value. Thus, a population slope of −599 ms could produce a sample slope of −601 ms. A population slope of −599 ms is unlikely to produce a sample slope more than about two standard errors away. Because the standard error of the sample is about 1 ms, this smearing is tiny; we are unlikely to see sample slopes lower than about −601 ms. In this case, $p(D|\text{MR})$ looks pretty much like p(population slope|MR).

Assume the distribution in Figure 4.11, and smearing with a standard error of 1.09 to obtain $p(D|\text{MR})$. Do not worry about how exactly this is done, the provided program will do it for you; we will come to what you enter in the program for your own data later. Since the distribution is very long (600 ms long), that is, so many values are possible, the probability of observing a slope in any one 1 ms interval, e.g. 2–3 ms, is actually very small. The actual sample value was 2.8 ms. p(observing slope = 2.8 ms| this model of morphic resonance) = 0.0017. We determined above that $p(D|H_0)$ was 0.014. So Bayes factor = $p(D|M)/p(D|H_0)$ = 0.0017/0.014 = 0.12.

Remember posterior odds = Bayes factor × prior odds. Our Bayes factor is 0.12. This Bayes factor means data should REDUCE your confidence in morphic resonance and INCREASE your confidence in the null hypothesis! Contrast Neyman–Pearson in this case: $p = 0.018$, so we reject the null hypothesis! This is one of the counter-intuitive findings of Bayesian analysis. Sometimes a significant result on a Neyman–Pearson analysis, meaning one should categorically reject the null hypothesis on that approach, actually means one should be more confident in the null hypothesis because of the data! (Not all the time, just sometimes.) Remember (no matter what anyone tells you) Neyman–Pearson analyses do not directly license assigning any degree of confidence to one conclusion rather than another. If you want

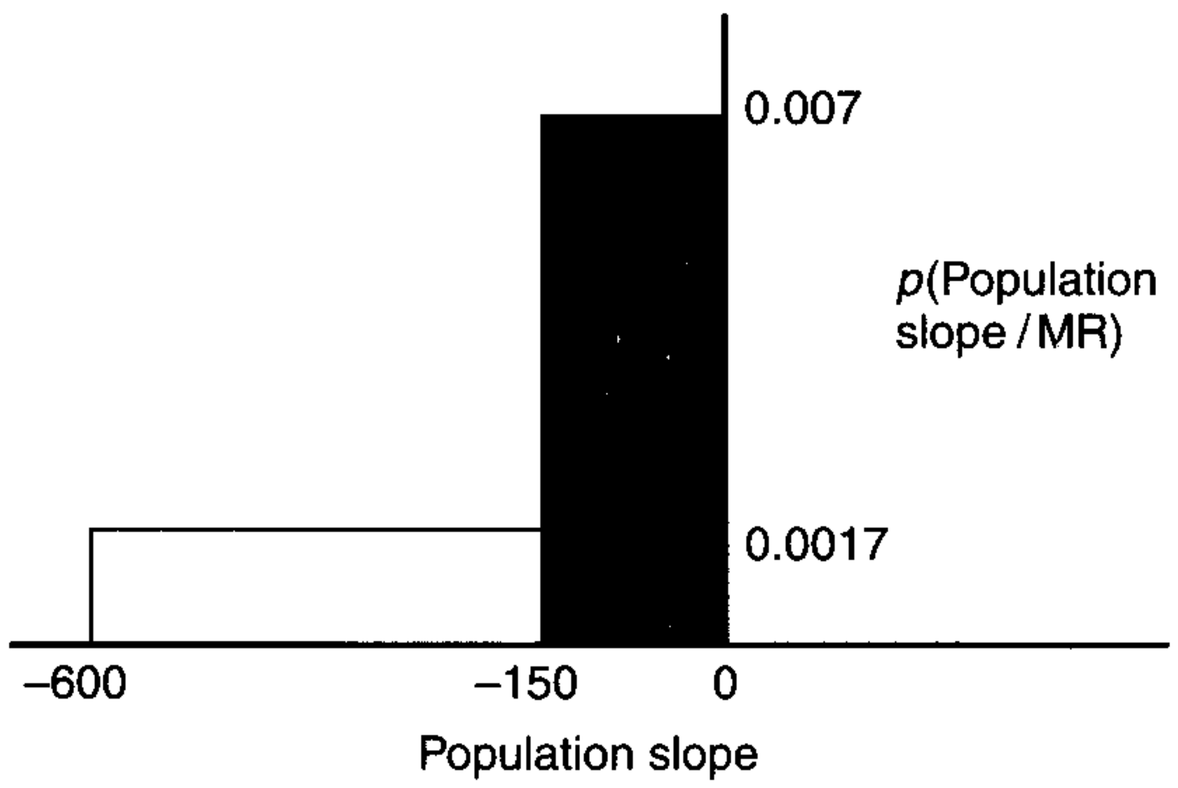
to know how confident you should be in different statistical conclusions, you need to deal with statistics designed to address what your confidence should be, namely Bayesian statistics. The differences between the approaches are not just a matter of philosophical quibbles; they can produce radically different research decisions. Neyman–Pearson may license a certain behaviour, but Bayesian analyses can tell you how confident you can be in your hypothesis.

Why does the Bayesian analysis indicate we should have more confidence in the null hypothesis? Because the theory we tested allowed a large range of outcomes; it was vague. The moral is that on a Bayesian analysis, a significant result may lead one to prefer the null hypothesis even more rather than a poorly specified theory. To see the logic of this, contrast the distributions *p*(slope|MR) in Figure 4.12. One of the distributions is the one considered so far, 600 ms long. But let us say for some reason we decide morphic resonance would not allow slopes greater than 150 ms (the filled-in distribution). This account of morphic resonance is more precise; it rules out more. The distribution is a quarter as long. Now remember that the area under a probability density distribution has to be one. The filled-in distribution therefore has to be higher. Each slope for the filled-in distribution has a higher probability density than the other distribution because the theory is more precise.

Evidence most supports the theory that most strongly predicted it. A vague theory might allow some obtained data but the data hardly supports that theory if the theory would have allowed virtually any data. That is the intuition that our Bayes factor of 0.12 illustrated: For a very vague theory, evidence supports it less than the null, even when the probability density of the data assuming the null is low. The probability density of the data is even lower assuming the vague theory (see Figure 4.13, which puts Figures 4.10 and 4.11 on the same scale). By embedding this intuition in the heart of its techniques, Bayesian analyses punishes those with vague theories, whereas significance testing does not.

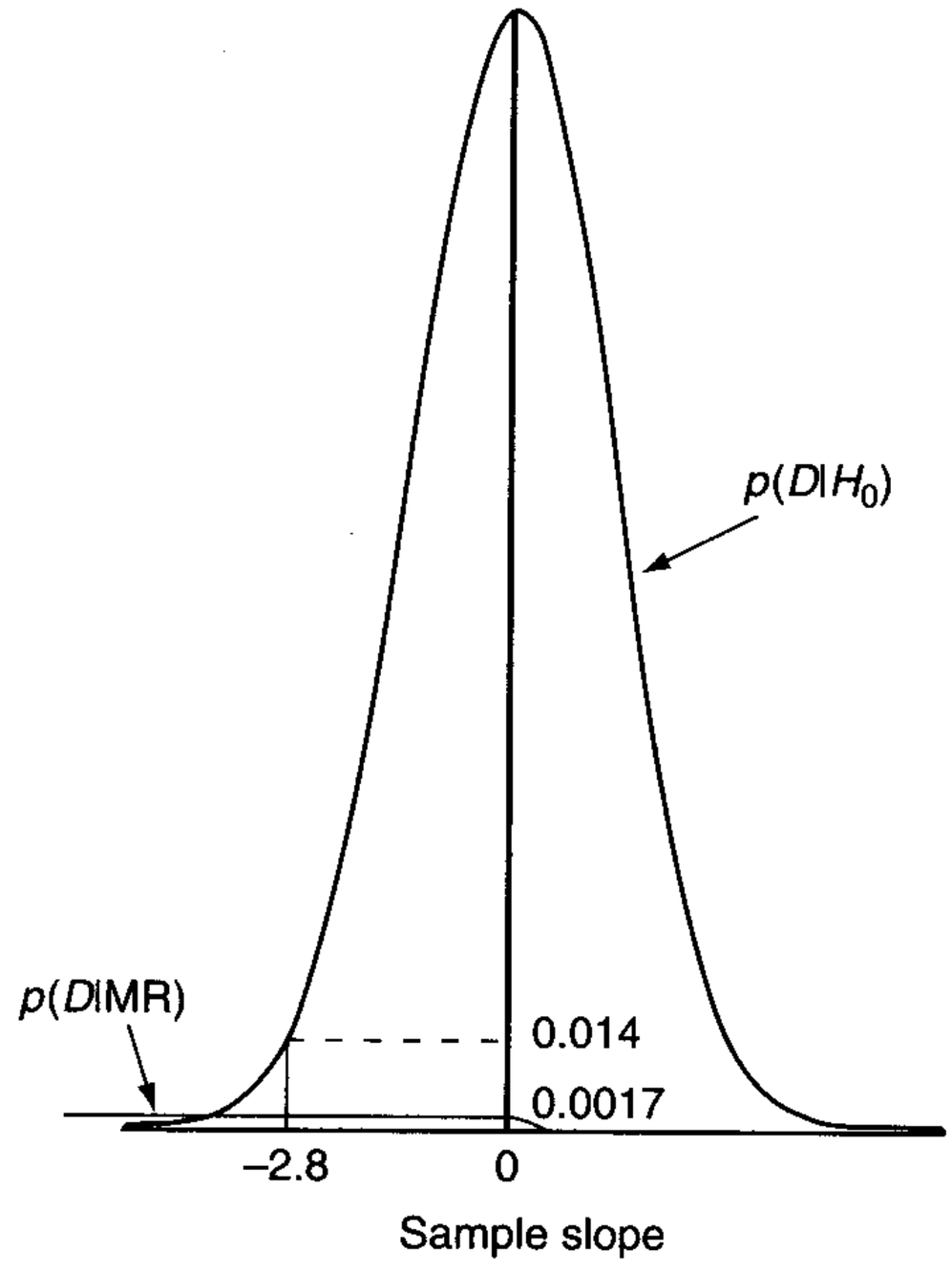
Our example thus far has done an injustice to morphic resonance. The theory is not so poorly specified as in this example. The assumption that morphic resonance allows all slope values between 0 and 600 ms per resonator equally is implausible. Sheldrake (1988) discusses the application of morphic resonance to data on learning and memory. The most striking example taken to be a case of morphic resonance is McDougal (1938), in which successive generations of rats learn a choice task with progressively fewer errors, despite the

Figure 4.12



Contrasting theories of different precision.

Figure 4.13

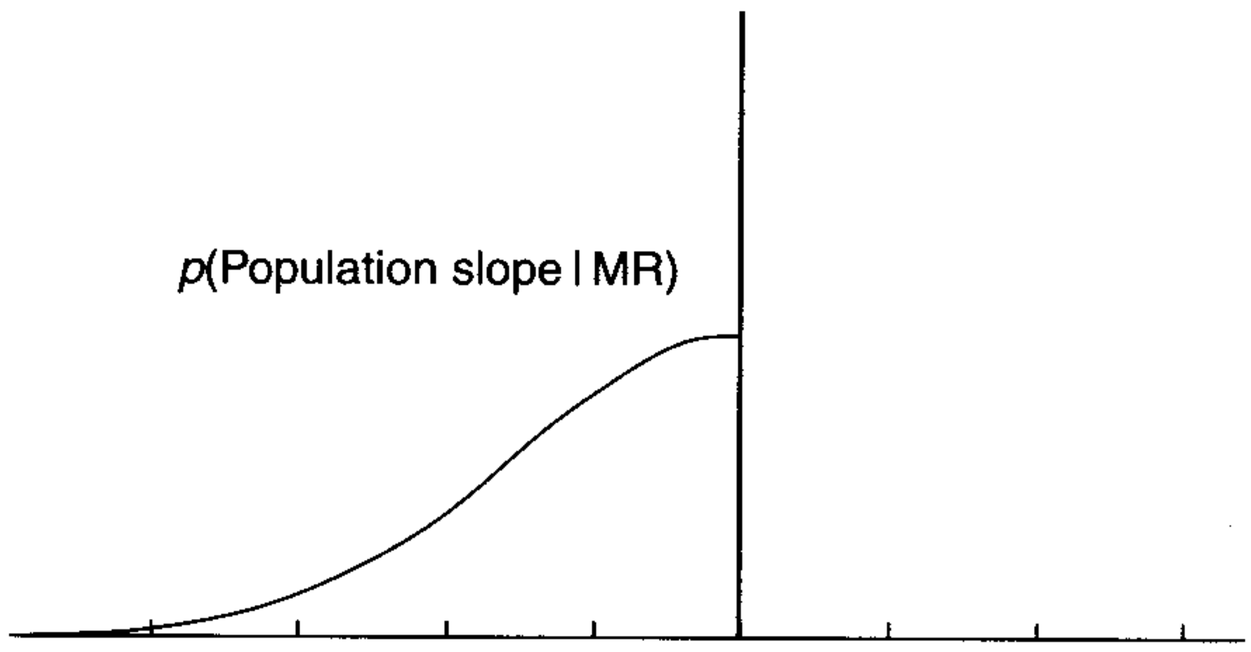


Likelihood of null and morphic resonance: Combining Figures 4.10 and 4.11.

fact McDougal selected the worse performers in each generation to sire the next. After a mere 500 rats have learned the task, average errors dropped from 215 to 27 per rat (McDougal, 1938, p. 331). That is, it took about 500 rats for between-rat resonance to produce almost the learning achieved by each rat in its lifetime (resonating with itself). The theory of morphic resonance, by linking disparate domains, can gain precision by using the data in one domain to constrain expectations in the other. Thus, it seems rather unlikely in our study that the between-participant effect would be much more than 1/500th of the within-participant effect, i.e. our population slope is not very likely to be much greater than 600 ms/500 = about 1 ms per resonator. Even then, a rectangular distribution does not capture our intuitions very well; presumably, the probability of large effects is smaller than small effects. On balance, effects smaller than 1 ms per resonator seem more likely than effects greater than 1 ms per resonator. One way of representing *p*(population slope|MR) is as a normal distribution centred on zero, and with half of it removed, as shown in Figure 4.14. If we are 95% sure that the true population slope allowed by morphic resonance is not more (in absolute terms) than 4 ms per resonator, then we can assign a standard deviation to the normal curve we are using of 2 ms.

Now morphic resonance, by allowing itself to be reasonably constrained in the domain of human repetition priming by knowledge we have about the application of the theory to other domains, is a more precise theory. A theory in having generality can gain in precision. The gain in precision is reflected in the Bayes factor. With the new assumptions concerning the theory, the Bayes factor is 12. Whatever your odds were in favour of morphic resonance before the data were collected, you should multiply them by 12 in the light of these data (given the assumptions we have made). This seems a reasonable type of conclusion to draw

Figure 4.14



A plausible shape of the slopes predicted by morphic resonance.

from the data. We are not making a black and white decision. If you thought morphic resonance was wildly implausible before the data, you will still think it implausible in the light of these data; it's just that your odds have been nudged up. If you had even odds in favour of the theory before hand, now you will think on balance it is more likely than not. It will take a lot more data before everyone agrees; and that is just as it should be.

The morphic resonance example highlights another difference between Neyman–Pearson and Bayesian analysis. Given a theory of morphic resonance which allows effects arbitrarily close to zero (as illustrated in Figure 4.14), then assuming the null hypothesis to be true, significance testing would not allow the conclusion that morphic resonance does not exist, no matter how big our sample. We can only accept the null given we have enough power to detect a minimally interesting effect size, and in this case, any effect above zero is possible. In contrast, notice how the Bayesian analysis can allow us to either increase or decrease our confidence in the null, depending on how the data turn out. A Bayesian analysis can allow us to draw inferences from data where significance testing is mute.

I ran two further studies with a similar design. One study (experiment 2) varied from the first in that there were 20 instead of 9 boosters between each resonator (so 200 boosters were run in total) and two rather than one resonators were run after each set of boosters. The other study (experiment 3) involved resonators being run not only in Sussex but also in Goetingen by Professor Suitbert Ertel; his job was to tell me which set of stimuli I had been boosting in Sussex. All results were flat as a pancake. For the same assumptions we used above to get a Bayes factor of 12 for the first experiment, the Bayes factors for the new data for non-words were 0.49 (experiment 2), 0.14 (experiment 3 Goetingen data), and 1.58 (experiment 3 , Sussex data). For these extra experiments, the overall Bayes factor for non-words is therefore: $0.49 \times 0.14 \times 1.58 = 0.11$. These new data do not rule out morphic resonance, no black and white decision need be made; the data just change our odds. For these data taken together including the first experiment, our odds in favour of morphic resonance should remain roughly unchanged (12×0.11 is about 1). If we take into account the data for the words as well as the non-words, and assume the effect for words is half that for non-words (i.e. the standard deviation is 1 ms), then the overall Bayes factor is 0.5. With further data, the Bayes factor will be driven either up or down, and at some point, the data will change which side of the decision bound in Figure 4.9 the theory falls for different people.

This discussion should have provided you with the concepts needed to consider a Bayesian analysis of your own data, if you can assume your data are roughly normally distributed. To use the provided program to calculate a Bayes factor for your data, you need to enter your sample mean and standard error. You also need to decide: Does my theory predict a rectangular distribution for the population effect (as in Figure 4.11) or a normal distribution? If rectangular, what are the limits? If a normal, what is its mean and standard deviation? If a normal, does the theory allow both positive and negative effects or only effects in one direction? Answers to these questions will enable the program to calculate a Bayes factor for you. You do not need to know how to do the calculations, just the concepts behind them.

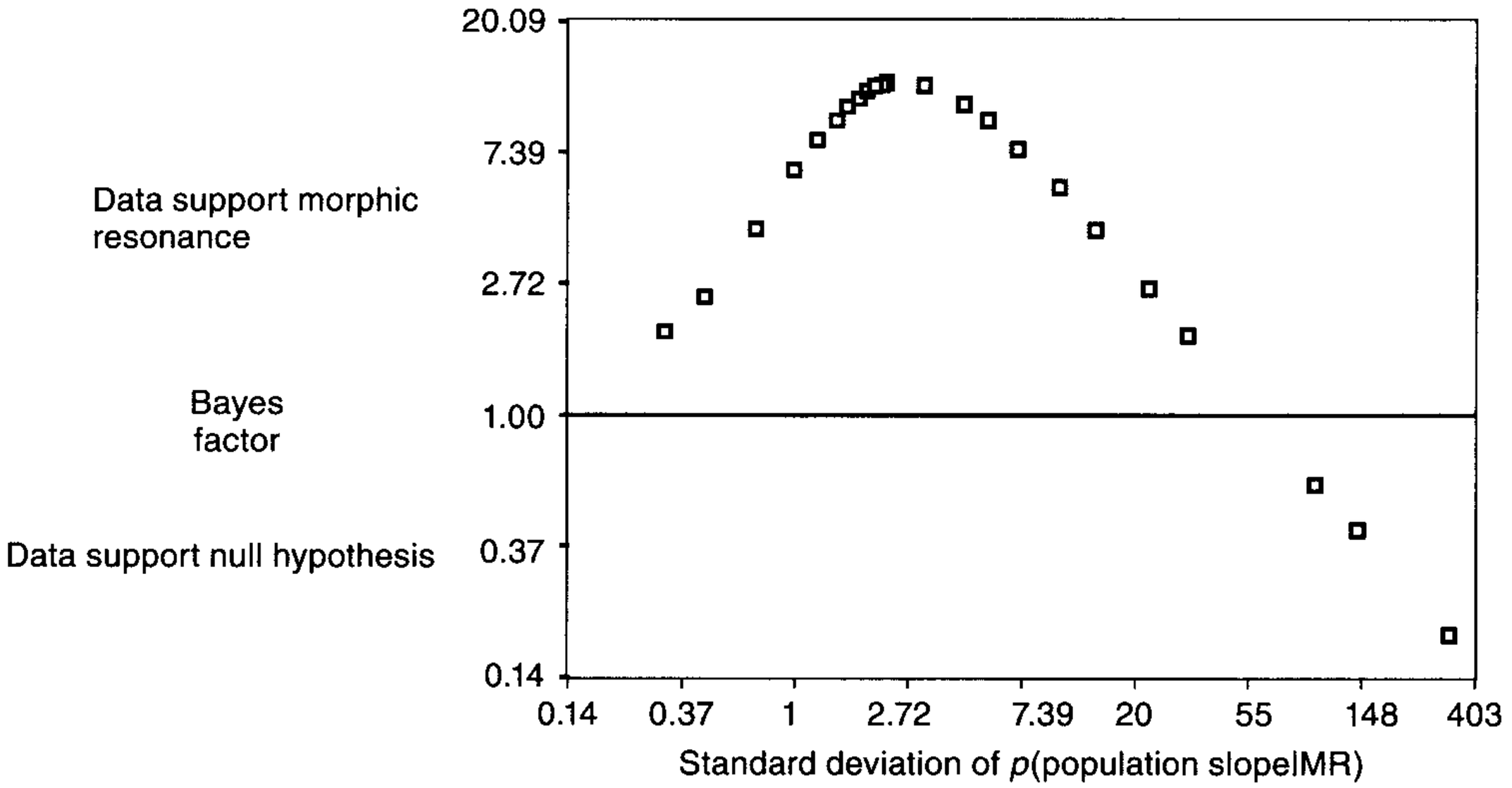
Finally, a comment on what the Bayes factor means about your personal probability in a theory, say morphic resonance. A Bayes factor tells you how much to multiply your prior odds in favour of morphic resonance *relative to the null hypothesis* in the light of data. If those were the only two theories that could be conceived of, then your personal probability of morphic resonance is given by (posterior odds)/(posterior odds + 1) where the posterior odds are obtained from the prior odds by multiplication with the Bayes factor. However, in practice other theories will exist that compete in explaining the data. Morphic resonance may have increased odds relative to the null but decreased odds relative to other theories that predict the data more strongly. The final effect of the data on the probability of morphic resonance depends on spelling out all the other relevant theories. But this is not possible; there are bound to be theories one has not yet conceived of. So it is best to treat your posterior probability as a probability conditional on the theories you have actually considered. You should reserve some of your personal probability for theories as yet undreamt of.

Summary

A Bayes factor tells you how much to multiply your prior odds in favour of your experimental theory (over the null hypothesis) in the light of data. An advantage of using a Bayes factor to analyse your experiments is that for common situation, low sensitivity shows up as a Bayes factor near 1. You are not tempted to accept the null hypothesis just because the experiment was insensitive. The Bayes factor also penalizes vague theories; in that case, data significantly different from the null may actually support the null. Further, you can combine experiments together or continue to collect data until the Bayes factor for all the data is extreme enough for you to make your point.

A disadvantage of the Bayes factor is shown by the somewhat arbitrary way we settled on p(population effect|MR) and hence on $p(D|\text{MR})$. $P(D|\text{MR})$ reflected not only the data but also our subjective judgments about what is likely, given both morphic resonance and other things we know. In that sense, $p(D|\text{MR})$ did not reflect just the data and the theory of morphic resonance per se; so it was not a true likelihood. Unfortunately, a logical analysis of morphic resonance, or almost any theory in psychology, together with relevant past studies, does not yield a definitive unique distribution for p(population effect|theory). One solution is to report how the Bayes factor varied with assumptions, for example, how it varied with different assumed standard deviations for the p(population effect|theory) distribution. Figure 4.14 shows how the Bayes factor varies for non-word data for the first experiment. People can then choose the Bayes factor based on the assumptions that they endorse. Most simply, one could report just the maximum value of Bayes factor the analysis allows and the associated assumptions (in this case, the peak in Figure 4.15 is a Bayes factor of 12 for a standard deviation of p(population effect|theory) of around 3).

Figure 4.15



Relation of Bayes factor to distribution of *p*(population slope|MR) (Note: The axes use 'log scales': Each labelled point is 2.72 times higher than the previous point, and 2.72 is an arbitrary number for this purpose.).

Stopping rules

Likelihoods (and hence posterior distributions, credibility intervals and Bayes factors) are insensitive to many stopping rules. For example, in the previous chapter, we considered a hypothetical study to determine the proportion of women who experienced G Spot orgasm. One could decide to stop collecting data when a certain sample size has been reached (e.g. 30 women in total) or alternatively one could decide in advance to stop when a certain number of women who experienced G Spot orgasm had been reached (say, 6). Even if in both cases one ended up with six women claiming to experience G Spot orgasm out of 30 asked in total, on a classic Neyman–Pearson analysis, one estimates the population proportion differently in the two cases (6/30 vs 5/29). A Bayesian analysis produces the same answer in both cases. In both cases, what is important is the exact data obtained in asking successive women. They answer: no, no, no, yes, no and so on for 30 answers. Those are the exact data. The probability of obtaining those data given a hypothesis, for example that the population proportion = 0.3, is the *same* regardless of the stopping rule. The probability is $0.7 \times 0.7 \times 0.7 \times 0.3 \times 0.7 \times \ldots$ and so on for 30 answers for both stopping rules. The likelihood is unaffected by which stopping rule is used; thus, the posterior, credibility interval and Bayes factor are unaffected as well.

Remember we cannot use a stopping rule like 'Stop when my *t*-test is significant at the 5% level' in the Neyman–Pearson approach. For example, consider a one-sample *t*-test, which is mean/SE (for the null hypothesis that population mean = 0). For a reasonable number of subjects, the critical value of *t* is 2; that is, the test is significant at the 5% level when the mean is about two standard errors from zero. As more subjects are run, the standard error gets smaller and smaller, and the sample mean will vary more and more tightly around zero. Even given the null is really true, sooner or later the sample mean will randomly sneak out to two standard errors from zero. At this point, the sample mean may be very very close to

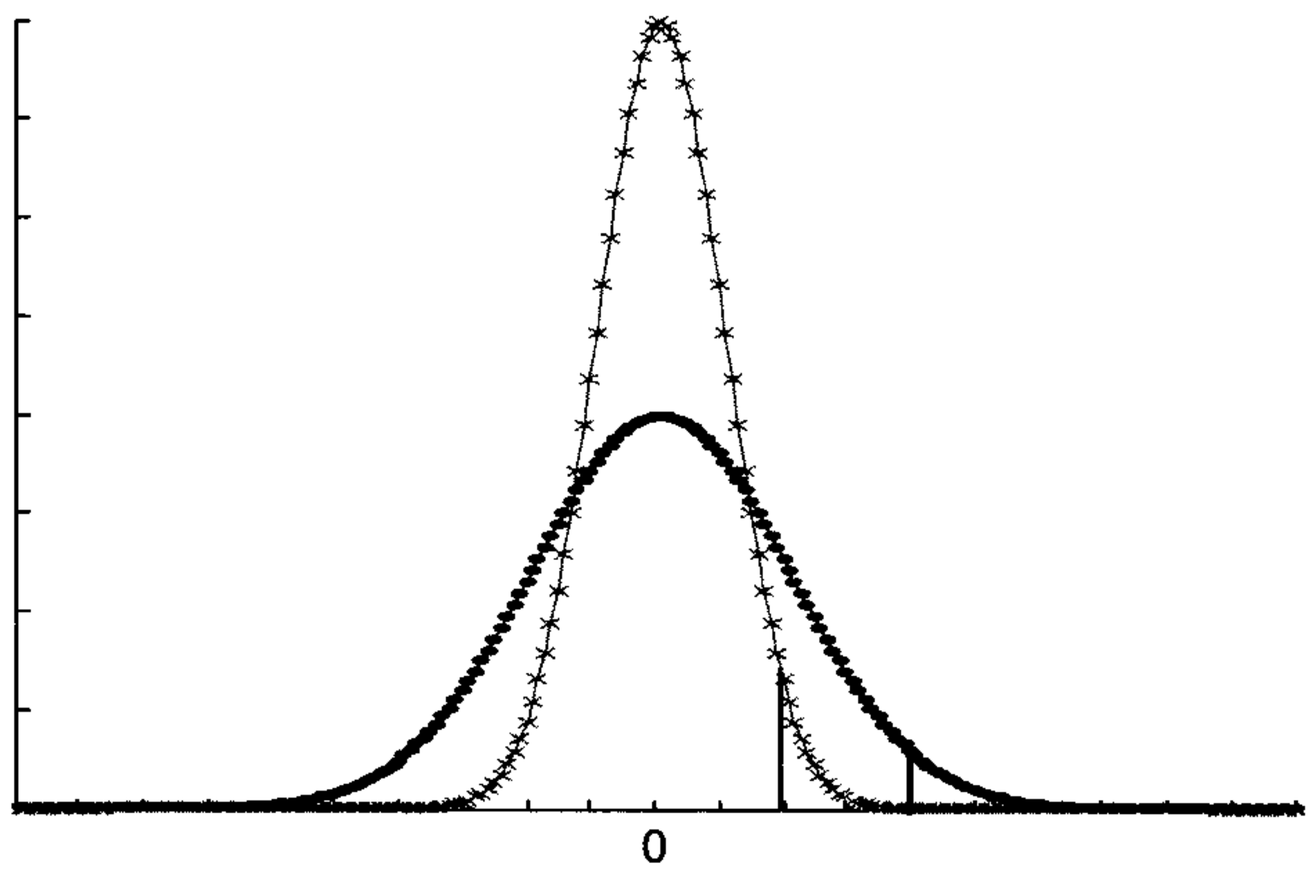
zero, but because the standard error is so small, we still get a *t*-value of 2. Sooner or later that is guaranteed to happen. (In fact, sooner or later the mean will sneak out to any number of standard errors from zero as you wish; you just have to wait long enough.) As we shall see, the Bayes factor behaves very differently.

The Bayes factor is the likelihood for the alternative theory (e.g. morphic resonance), $p(D|\text{MR})$, divided by the likelihood for the null, $p(D|\text{null})$. As more and more subjects are run and the sample standard error shrinks, and, assuming the null is true, the sample mean will hover around zero more and more closely. But $p(D|\text{MR})$ remains pretty constant as more subjects are run, given that the theory allows a range of population values covering the area around zero. The probability density of these values will not change just because you run more subjects. Now, what about $p(D|\text{null})$? As more subjects are run, the standard error shrinks, so the $p(D|\text{null})$ distribution has to become more peaked, higher around zero, in order that the area under the curve remain unity. This is illustrated in Figure 4.16, which shows $p(D|\text{null})$.

In Figure 4.16, the curve of crosses corresponds to when four times as many subjects are run as in the curve of filled circles (say $4n$ vs n subjects). Thus, the $4n$ curve has a smaller standard error than the n curve (half the size in fact). Compare the height of the line (i.e. $p(D|\text{null})$ for a mean two standard errors away from zero for the $4n$ curve with the n curve. It is higher for the $4n$ curve.

Assuming the null is true, as more data are collected, the mean will vary around zero, mostly keeping within about two standard errors from zero, but sometimes reaching two standard errors or beyond. When the mean reaches two standard errors out, $p(D|\text{null})$ will be higher after many subjects are run rather than after only a few. This is true for any number of standard errors out. That is, as more subjects are run, $p(D|\text{null})$ is expected to just keep increasing. As $p(D|\text{null})$ increases, so the Bayes factor decreases. Assuming the null is true, with increasing subjects the *t*-value does a random walk around zero, going up and down randomly; the Bayes factor is, by contrast, driven closer and closer to zero with increasing

Figure 4.16



$p(D|\text{null})$ for data collected with n subjects (filled circles) and $4n$ subjects (crosses). The tall vertical line indicates the height of the $4n$ curve two standard deviations out and the small vertical line indicates the height of the n curve two standard deviations out. Notice the lines are of different heights.

subjects. There is no guarantee that the Bayes factor will ever exceed a certain value, like 4, even after an infinite number of subjects. If you run 'until I get a significant result', you are guaranteed to, sooner or later, even assuming the null is true. Neyman–Pearson statistics are sensitive to the stopping rule; Bayesian statistics are relatively insensitive. That is a distinct advantage of the Bayesian approach. All that matters is the data you actually obtained, not what intentions you may or may not have had about when to stop. In commenting on the insensitivity of Bayesian analyses to the stopping rule, Savage (1962, p. 72) said 'it is impossible to be sure of sampling until the data justifies an unjustifiable conclusion, just as it is impossible to build a perpetual motion machine. After all, whatever we [i.e. classic and Bayesian statisticians] may disagree about, we are surely agreed Bayes' theorem is true where it applies'.

The relative insensitivity of the likelihood to stopping rules gives Bayesian statistics an advantage over conventional significance testing. If journals ever adopted, say Bayes factors as the standard statistical tool, they would probably introduce a criterion for publication like 'The Bayes factor should be 4 for supporting a theory or 1/4 for confirming a null'.[4] This provides a perfectly good stopping rule 'Stop when the Bayes factor is 4 or 1/4'. If you collected 30 participants and were not quite there yet, just collect some more. On the other hand, if you collected 30 participants and the results were not quite significant, you could not collect some more until you made it at the 5% level; yet the current system makes the former practice almost inevitably wide spread. If you had collected 30 subjects and just missed out, binning all that work would be a tragic waste, and where is the virtue in that. The current system invites cheating.

Running until the confidence interval is a preset width also avoids common wicked temptations (compare with the previous chapter). If you decided to run until the 95% confidence interval excluded zero, you are guaranteed to always succeed even if the null is true, for exactly the same reasons just noted above. So running until a confidence interval excludes zero is not a valid stopping rule for Neyman–Pearson statistics. Similarly, if you decided to run until the 95% *credibility* interval excluded zero, you are also guaranteed to succeed. The implication for Bayesian inference is different than for classic inference, however (contrast Mayo, 1996). With the Neyman–Pearson approach the exclusion of zero from the confidence interval corresponds to a decision to reject the null hypothesis. With the Bayesian approach, given the prior and posterior are probability density functions, the null before and after any amount of data collection has zero probability. We are entitled to use the posterior to calculate the probability that the true population value lies in any *interval*. If the true population value is zero, then the bulk of the area of our posterior will become contained in an interval closer and closer to zero, no matter what the stopping rule.

Weakness of the Bayesian approach

The strengths of Bayesian analyses are also its weaknesses. Its strengths lie in directly addressing how we should update our personal probabilities and in the insensitivity of the statistical conclusions to various apparently arbitrary circumstances. As we will see, these two points define its greatest weaknesses.

First, are our subjective convictions really susceptible to the assignment of precise numbers? And are they really the sorts of things that do or should follow the axioms of probability?

4. Jeffreys (1961, p. 432) suggested that Bayes factors between 1 and 3 were 'barely worth mentioning'; those between 3 and 10 were 'substantial'; those above 10 strong.

Should papers worry about the strength of our convictions in their result sections, or just the objective reasons for why someone might change their opinions? Our convictions can seem like will-o'-the wisps. A small comment by someone, an analogy, a random thought, spotting a connection or a stylistic incompatibility in what you believe can shift the plausibility of a theory completely in the absence of any data. If you are uncomfortable with trying to assign precise numbers to your convictions, one solution is just to focus on the likelihood. After all it is the likelihood that tells you everything you need to know about the relative support the data gives different hypotheses. People who only use likelihoods are called likelihood theorists. It gives you many of the advantages of Bayesian statistics and avoids any issues concerning nailing down slippery subjective probabilities (see next chapter and Oakes, 1986, for further discussion; and Royall, 1997, for a sustained defense of likelihood inference). On the other hand, a Bayesian might ask a likelihood theorist, if you are going to go part way down the Bayesian path, is it not half hearted not to go all the way? Given that the likelihood will come to dominate the posterior, assuming that the likelihood is only somewhat more precise than the prior, a reasonable approximation to most people's beliefs in many scientific cases can be obtained by assuming a uniform or uninformative prior so that the likelihood dominates completely (people who by default use uninformative priors are called 'objective Bayesians'). In the next chapter, we will discuss uniform priors and consider likelihood inference in more detail.

The second (at least apparent) weakness of Bayesian statistics is that they are not guaranteed to control Type I and Type II error probabilities (Mayo, 1996), especially with multiple testing. Whether this is a weakness is at the heart of the debate between Bayesian and classical approaches. I will motivate intuitions on both sides and leave you to decide if there is a weakness.

Neyman–Pearson statistics are designed to control Type I and Type II error probabilities in an optimal way. Any other method for making acceptance and rejection of decisions will not be so optimal. Bayesian statistics do not give you black and white decisions; but life forces such decisions on us. At some point, we have to act on our knowledge one way or another. A journal editor may say he will only publish papers with Bayes factors above 4 or less than ¼. Now we have a decision routine and we can ask about its long-term Type I and Type II error rates.

Imagine 10 measures of early toilet training are correlated with 10 measures of adult personality. Out of these 100 correlations, three are found to be significant at the normal 5% level: that is, for these three, their 95% confidence intervals exclude zero. A Neyman–Pearson user of statistics would say: 'One expects about 5 to be significant by chance alone; these are weak data and do not lead one to accept any hypothesis about toilet training affecting personality.' How might a Bayesian proceed? If we are really interested in evaluating point null hypotheses, we can calculate Bayes factors for each of the 100 hypotheses. I ran a simulation calculating 1000 Bayes factors by sampling 1000 times from a population with a mean of zero.[5] The alternative hypothesis in each case allowed positive and negative population values symmetrically (normal distribution centred on zero). Of course, some Bayes factors will exceed a given threshold by chance alone. For example, 78 of these 1000 Bayes factors,

5. Each sample was a random draw from a normal population with a mean of zero and a standard deviation of 0.1. The prior for the alternative hypothesis had a mean of zero and a standard deviation of 0.2. (If we had determined the maximum Bayes factor for each sample allowed by varying the standard deviation of the prior independently for each sample to find the optimal one, the proportion of Bayes factors above 4 would be a bit higher.) The standard deviation of sample means depends on the number of data points in each sample. If we had collected more data in each sample so that each sample was drawn from a population with a standard deviation of .01, then the error rate decreases: For example, the proportion of Bayes factors above 2 in a simulation of 1000 studies was only 4.4%.

that is, 7.8%, were above 2. Further, 2.8% were above 4. Thus in 100 tests, we might expect 2 or 3 Bayes factors to be above 4 by chance alone (this is not a fixed error rate; it depends on the exact alternative hypothesis considered). According to the Bayesian, in interpreting any one of these Bayes factors, unlike the Neyman–Pearson statistician interpreting significance tests, we would take no account of the fact that we ran 100 tests in total. Mayo (1996) presents this as a case against Bayesians: Bayesians are highly likely to generate support for experimental hypotheses when the null is true.

A Bayesian does not ignore all the other 97 results in evaluating a grand theory concerning toilet training and personality. For example, if a Freudian theory predicted ALL tested relationships, its Bayes factor in the light of the 100 correlations would be very low! BUT, the critic continues, the Bayesian can still run a lot of tests and pick out one result for which the Bayes factor is above some threshold. SHOULD not one's confidence in the alternative hypothesis supported by this one test be reduced because of all the other tests that were done (albeit they were tests of different hypotheses)? But why should what else you may or may not have done matter? The other tests were not, as such, relevant to the one hypothesis under consideration; the other tests tested different hypotheses. How can they matter?

If you were an editor would you publish the paper because there was 'substantial' support (e.g. Bayes factor = 4) for one of the 100 specific hypotheses? What if the author just reported testing that one hypothesis and constructed a plausible theory for it to put in his introduction. He did not even mention running the other tests. According to the Bayesian, there is nothing wrong with that (assuming the other tests really did not bear directly on the theory). According to classical statistics, that is clearly cheating. What do you think?

We can motivate the intuition that multiple testing without correction involves cheating more strongly with another example. This will also enable us to see how the Bayesian can respond in detail – and show that while failing to correct for multiple testing is cheating when using classical statistics, our intuitions may not see it quite that way when Bayesian methods are used!

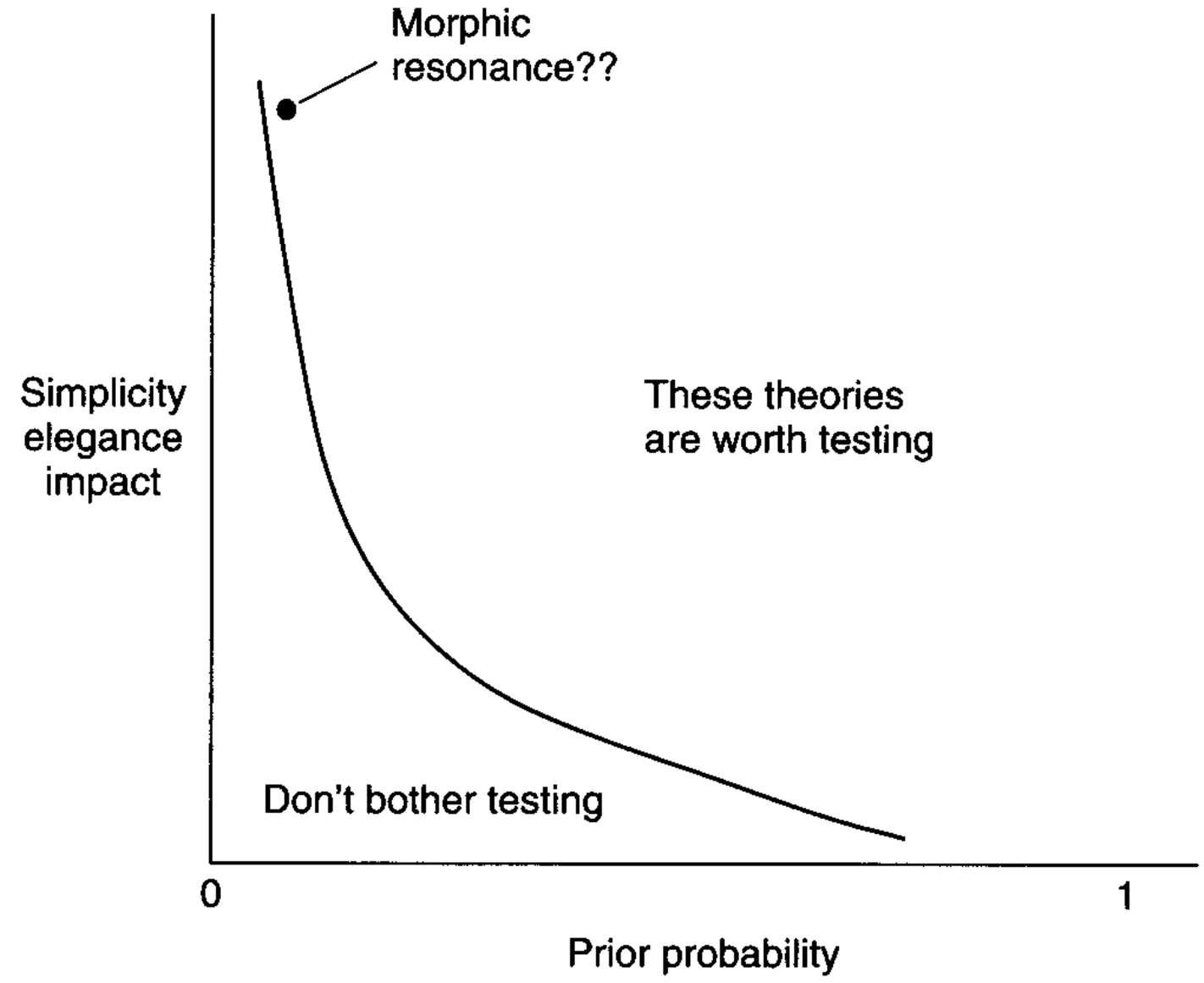
Box 4.6 Which theory to test

In a *Nature* editorial, John Maddox (1981) said of Sheldrake's first book that 'the book is the best candidate for burning there has been in many years ... Sheldrake's argument is pseudo-science ... Hypotheses can be dignified as theories only if all aspects of them can be tested.' While this last comment is naïve (contrast Popper 1934, who recognized good science will include metaphysical elements), Lewis Wolpert (1984) argued '... it is possible to hold absurd theories which are testable, but that does not make them science. Consider the hypothesis that the poetic Muse resides in tiny particles contained in meat. This could be tested by seeing if eating more hamburgers improved one's poetry.'

One way of capturing the notion that not all testable theories are actually worth testing is in terms of the prior probability of the theory, as illustrated in Figure 4.17. One wants to test theories that will have high impact, that are simple, and have other desirable characteristics. For example, the theory that my pen will drop to the ground if I let go of it now will have no impact on anyone if found to be true. On the other hand, it will not be worth putting in effort to test a high-impact theory if you find it outrageously implausible. Morphic resonance is a theory that if confirmed would have an enormous impact in all areas of science. But some, like Wolpert, find it is so implausible it is not worth testing. Despite arguments (e.g. by Popper) that all theories have essentially a zero probability (after all, only a miniscule fraction of all theories are true), we certainly treat some theories as more probable than others, as the morphic resonance debate demonstrates. We take into account a theory's plausibility in deciding whether to pursue it. Figure 4.17 represents a space personal to each individual; a plot of impact versus prior probability. Theories will pepper this space. The line represents a decision bound. Theories that are sufficiently implausible or

Box 4.6 *continued*

Figure 4.17



Which theory to test.

sufficiently trivial are not worth testing; theories with enough joint plausibility and potential impact are worth testing. Remember this is a personal space, a space each individual configures for themselves. Only you can say what your odds are in favour of different theories, and where your decision bound is. What are your subjective odds in favour of morphic resonance? Where would morphic resonance be for you in the space of Figure 4.17?

A Tibetan monastery wishes to find the reincarnation of a recently departed lama. The monks will test a child's ability to pick a favourite object of the old lama amongst a collection of similar objects. They put the old lama's walking stick in a collection of 20 other walking sticks of other monks from the monastery. Field testing at a nearby school (which the monks are sure does not contain the reincarnation) shows that the probability of a child picking the old lama's stick from the collection of 21 sticks is indeed 1/21. Now they have their test. If a given candidate child does pick the old lama's stick, then $p = 1/21 < 0.05$. That is, using the classical approach, if a candidate child passes the test, we can reject the null hypothesis that the child chose at random. Assuming we are happy we have excluded other possible hypotheses (like the testers gave conscious or unconscious cues), we can then accept the hypothesis that the child is the reincarnation. The old lama had said before he died that he would be reborn in a certain area. Various omens narrows the search down to 21 candidate children in that area. The monks test all the children and one child passes the test. Can the monks conclude that this child is the reincarnation? No. The test controlled Type I error to a satisfactory degree only when conducted as a single one-off test. With 21 tests, the probability of at least one of the tests giving a positive result assuming all children chose at random is $1 - (20/21)^{21} = 0.64$. That is, the probability of a Type I error for the

tests considered as a family is 0.64. Clearly, the family of tests does not constitute a reliable decision-making procedure; it cannot legitimately be used for indicating that one or other child in a set of 21 children is the reincarnation. This example illustrates the obvious need to take into account multiple testing. If one tests enough children, sooner or later a child will pass the test by chance alone. Yet, as we have discussed, the Bayesian approach does not have any corrections for multiple testing. Does not this sort of example therefore show the Bayesian approach *must* be wrong?

The Bayesian disagrees. A null hypothesis is that none of the children were reincarnations and all chose randomly. The likelihood of this null hypothesis is the probability of observing the exact sequence of data obtained given the null is true, likelihood(null) = $(20/21)^{20}(1/21)$. Let us say that the 10th child was the one who chose the stick. Assuming that the reincarnation will definitely choose the stick, the likelihood of the hypothesis that the 10th child is the reincarnation, likelihood(10th), is $(20/21)^{20} \times 1 = (20/21)^{20}$. Thus, whatever your prior odds were on the 10th child being the reincarnation against the null hypothesis, they should be increased by a factor likelihood(10th)/likelihood(null) in the light of these data, that is, by a factor of 21. 'Ha!' the classical statistician smirks, 'you have manufactured evidence out of thin air! By ignoring the issue of multiple testing, you find strong evidence in favour of a child being the reincarnation just because you tested many children!' Indeed, the fact that we have tested all the other children does not change the value of the Bayes factor for this child – it stands at 21.

But the Bayesian just patiently continues. The likelihood of the hypothesis that the first child is the reincarnation is zero because if he had been the reincarnation, he would have chosen the old lama's stick, and likewise for each of the 20 children who did not choose the stick. The probabilities of any one of them being the reincarnation go to zero. We have also just seen the probability of the 10th child being the reincarnation increases by a factor of 21. You can think of it as the probability associated with each of the 20 children who did not choose the stick being passed onto the probability associated with the child who did choose the stick. That is, the probability of *one or other* of the children being the reincarnation does not change one wit in the light of these data. Let us say before we collected the data the probability of one or other of the children being the reincarnation was π (read as 'pi'). Thus, the probability of the null, that none of the children were the reincarnation, was $(1 - \pi)$. If the prior probabilities of each of the 21 children being the reincarnation were equal, they would have been each $\pi/21$. After collecting the data, 20 of these probabilities go to 0, and the remaining one goes to π. They still sum to π. The posterior probability of the null is still $(1 - \pi)$. If you were convinced before collecting the data that the null was false, then you could choose the reincarnation with confidence (he is the 10th child); conversely, if you were highly confident in the null before collecting the data, you should be every bit as confident afterwards. And if you think about it, this is just as it should be. The Bayesian answer does not need to correct for multiple testing, the Bayesian explains, because if an answer is already right, it does not need to be corrected.

The example illustrates how the Bayesian can consider families of tests. In fact, as you run more tests of true nulls, the posterior probability of all nulls being true will typically increase (compared to the prior) because while the probability of some nulls decrease, most will increase (Box 4.7 considers an example). And it may be that we feel multiple tests need correction precisely because we sense that the probability of all nulls being true should increase as we collect more data (test more true nulls). The Bayesian approach respects this intuition.

Box 4.7 A Bayesian analysis of a family of tests

You gather together 20 famous psychic superstars, all of whom claim to have a particular paranormal ability. You test the psychic who claims special ability to see what other people see by having a sender look at a picture and the psychic choose one of 20 pictures. Similar tests are devised for each psychic such that the probability of passing by chance alone is 1/20. The psychics do not claim to be able to respond 100% of the time; but given the test conditions, and the claimed extent of their abilities, you work out that the probability of passing given they have psychic abilities is 0.90 (coincidentally the same for each person). So for each person we have a test with alpha = 0.05 and power = 0.90. We conduct 20 tests so we would expect one psychic to pass the test by chance alone. Thus, on the classical account, finding one pass and 19 fails provides no grounds for asserting any paranormal activity happened. What does a Bayesian analysis say?

Let us say it was the 15th psychic who passed her test. Your prior odds on that psychic having powers against the null should be changed by the ratio of the likelihoods, that is by $0.90/0.05 = 18$. We have strong evidence that this psychic has paranormal powers. If we regard the fact that 19 other psychics were tested is irrelevant to the claim that *this* psychic has powers, then no account is taken of the 19 failures. The strong evidence stands as it is. Again, as in the reincarnation case, it seems we can capitalize on chance (testing lots of people) in *creating* evidence. As a whole the situation is exactly as we would expect by chance, yet by using Bayes we have strong evidence that one person is a psychic superstar. Surely Bayes has got it wrong!

But the reason we wish to treat the set of tests as a family in this case is precisely because we regard the tests as addressing in some sense the same issue. We do not like to take one test in isolation because all tests bear on the issue. For a psychic that failed, the likelihood of the null is 0.95 and the likelihood of the hypothesis that they have powers is 0.10. Whatever your prior odds were on the null, they should be increased by a factor of $0.95/0.10 = 9.5$ for this psychic. So whatever your prior odds were on the null that *none* of the psychics had powers, they should be increased in the light of these data by an amount $9.5^{19}(1/18) = 2 \times 10^{17}$. The data provide astronomically high support for the null over the hypothesis that they all had powers. In fact, the data also provide support for the full null over the hypothesis that psychic powers exist and only 1/20 of psychic superstars really have such powers.

In sum, when one views the family of tests as a whole, one test passing out of 20 increases the probability of the complete null being true. This analysis shows that it is both true that your odds against the null for the one individual test passed should increase and that your odds on all the nulls jointly being true should increase. This may at first sound contradictory but in fact the statements are consistent, as the Bayesian analysis shows. This analysis may explain the apparently contradictory pull of our intuitions in dealing with cases of multiple testing. On the one hand, we recognize that classical statisticians are right to insist that when thinking of a set of tests as a family, the total number of tests is important. On the other hand, many users of classical statistics do not always correct for the number of tests because they feel evidence for a hypothesis is still evidence for that hypothesis regardless of being part of a family (as just one example, consider the common practice of placing asterisks in a correlation matrix to indicate $p < 0.05$, with no correction for multiple testing). The Bayesian analysis respects both intuitions. In our example, your probability for the complete null should increase yet your probability for the one particular null should decrease.

Our intuitive need for corrections for multiple testing in some circumstances may sometimes reflect additional implicit assumptions that can be analyzed in a Bayesian way. In a study that seems to be data dredging (e.g. analysing the correlation of religiosity with 20 other variables with no strongly motivating theory given) the presence of many weak results may change our prior probability that the procedure used for selecting variables was good at filtering out null hypotheses. In this case, a large number of weak results may affect our posterior probability of any given null being true: We may increase our probability that the variables were selected in a thoughtless way.

Finally, bear in mind that issues to do with multiple testing often arise when the prior probability of the nulls is very high. Data dredging means precisely one expects most nulls to be true. In motivating corrections for multiple testing, the classical statistician asks us to imagine all nulls are true, that is, we are certain that the nulls are true. Of course, if we are certain the nulls are true before collecting the data, a Bayesian analysis indicates we should be certain they are true afterwards as well. Even if we are not completely certain in a null, an only moderate Bayes factor may still not leave the posteriors for the nulls high enough for us to want to treat them as true.

Is this enough to assuage you that the problem of multiple testing has been dealt with?

Using the Bayesian approach to critically evaluate a research paper

From reading the introduction to the paper, identify the key main prediction the paper makes that follows from the substantial theory. For now we will stick to a prediction that could be tested with a *t*-test. Sketch your personal prior probability distribution for the effect predicted, assuming a normal distribution does not violate your intuitions. From the data, determine the likelihood function and hence, using your personal prior, the posterior probability function. Make a good sketch of your posterior. Consider the authors theory. The author may just predict that there should be some difference between conditions, but he does not specify what size difference. Both your prior and posterior will give a 100% probability for there being *some* difference: In most cases, the prediction of there being some difference is so obvious it scarcely counts as a prediction at all. The author is probably being lazy. He may really mean there will be a difference larger than some minimal amount. If the author does not state what that minimal amount is, estimate it yourself based on the sort of effect typically found in that literature. Let us say the minimal interesting effect is more than 5 ms either aside of zero. In your prior how much area is under the curve from −5 to +5 ms? That is, your prior probability for there being an effect so small, the author's theory is rendered false or irrelevant. What is your posterior probability for the effect lying between −5 and +5 ms? Has it increased or decreased? If the probability has increased, the data do not support the author's theory; if decreased the data support it. No categorical decisions in accepting or rejecting any hypotheses need be made, however. How do your conclusions compare with the author's?

Another approach is to use the Bayes' factor. Assume you believe the author's theory absolutely. Sketch your prior again, assuming the theory (call it theory A for Author's theory) – this is p(population effect|theory A). In the previous paragraph, we assigned zero probability to any particular effect size, including zero, because you assumed a normal for your prior. Do you wish to assign a non-zero probability to the null hypothesis of no effect? If you do, you can calculate the Bayes' factor using the provided program. The Bayes' factor tells you how much to increase your probability in the theory over the null given the data. How does this conclusion compare to the author's?

Now whether not you are willing to assign a non-zero probability to the null, you can use the Bayes' factor to compare the author's theory to the substantial theory which the author or you consider to be the main competitor. Identify this theory (call it theory B), assume it absolutely and sketch the prior assuming this theory: p(population effect|theory B). Calculate the Bayes' factor for this theory. When you do this, the program tells you what '`Likelihoodtheory`' is. Divide the likelihood for theory A (which you will have obtained by following the exercise in the last paragraph) by the likelihood for theory B. This is the

Bayes' factor for the author's favoured theory over the main competitor. By how much does the data support the author's theory over the competitor theory? How does this compare to the author's own conclusions?

Sometimes a Bayesian analysis will support the instincts of an author in interpreting Neyman–Pearson statistics as a way of updating their personal convictions. But sometimes the Bayesian analysis will give very different answers because only the Bayesian approach requires personal probabilities be updated coherently, that is, according to the axioms of probability. Often people should be penalized more severely for the vagueness of their theories than they realize.

Summary: Neyman–Pearson versus Bayes

Table 4.1 summarizes the contrasts between Neyman–Pearson and Bayesian statistics. In classic statistics, it matters in what context a test was done (as part of a family of 100 other tests, 5 other tests, or just done by itself?); in Bayesian statistics, the support for a hypothesis depends *only* on the data directly relevant to that hypothesis. In classic statistics, it matters whether you invent your explanation for an effect before conducting the test or afterwards (planned vs post hoc tests). In Bayesian statistics, it is irrelevant whether you invented the hypothesis on Wednesday or Friday, the *timing* of an explanation is irrelevant to how good an explanation it is of the data, and to how much the data support it. Contrast the common idea, for example supported by Popper and Lakatos (see Chapters 1 and 2), that the novelty of a prediction is important for how much confirmation supports a theory. From the Bayesian perspective, timing and novelty are clearly irrelevant for the magnitude of the likelihood, and the likelihood tells you everything you need to know about the support the data has for a theory. Therefore, the timing of data relative to explanation is irrelevant.

Table 4.1 Neyman–Pearson versus Bayes

	Meaning of probability	**Aim**	**Inference**	**Long-run error rates**	**Sensitive to**
Neyman–Pearson	Objective frequencies	Provide a reliable decision procedure with controlled long-term error rates	Black and white decisions	Controlled	Stopping rule; what counts as the family of tests; timing of explanation relative to data
Bayes	Subjective	Indicate how prior probabilities should be changed by data	Continuous degree of posterior belief	Not guaranteed to be controlled	Prior opinion

In sum, in the Neyman–Pearson approach, it is assumed that what the scientist wants is a generally reliable procedure for accepting and rejecting hypotheses, a procedure with known and controlled long-term error rates. According to the Bayesian, the scientist wants to know the relative amount of support data provided for different hypotheses so she knows how to adjust her convictions.

What do you want from your statistics?

We established in the previous chapter that you were probably an unwitting closet Bayesian. Now you know the issues, do you want to come out of the closest? Or do you want to renounce your old tacit beliefs and live a reformed life?

Before you draw any final conclusions, consider the final major school of inference, likelihood inference, discussed in the next chapter.

Review and discussion questions

1. Define the key terms of probability, probability density, prior probability, likelihood, posterior probability, credibility interval, Bayes factor.
2. State the likelihood principle. In what ways does Neyman–Pearson hypothesis testing violate the likelihood principle?
3. Think of some null hypotheses that have been tested in a journal article you read recently. To what extent does it make sense to give them a non-zero prior probability?
4. How does a credibility interval differ from a confidence interval? Calculate both for the same set of data and compare and contrast the conclusions that follow from each.
5. Discuss whether the Neyman–Pearson or Bayesian approach gives more objective conclusions.
6. Discuss whether it makes sense to correct for multiple testing in the Bayesian approach.

Further reading

One of the only textbooks of statistics providing a Bayesian perspective suitable for people with the mathematical abilities of the average psychology undergraduate is Berry (1996). Berry provides the equivalent of a first-year undergraduate course in statistics for psychologists, building up step by step to determining credibility intervals for normal data and also for proportions. Bayes factors are not considered.

For an excellent introduction to both objective and subjective probabilities, and the logic of statistical inference that follows from each, see Hacking (2001).

The standard philosophical treatment advocating Bayes with unshakable conviction is Howson and Urbach (1989, see especially Chapters 1, 10 and 11). A defence of Neyman–Pearson and criticism of Bayes is provided by Mayo (1996), especially pp. 69–101 and Chapters 9 and 10. Howard, Maxwell and Fleming (2000) is a readable short comparison of Bayes with Neyman–Pearson. Oakes (1986) is an excellent overview of different approaches to statistics (though unfortunately now out of print). If you want to explore Bayesian analyses for a range of research designs, McCarthy (2007) takes the reader through how to use the free online software WinBUGS which allows great flexibility in data analysis.

Finally, if you have some mathematical background, even if just very good high school level, I strongly recommend Jaynes (2003), for a forceful argument for the 'objective Bayesian' approach.

Matlab program for calculating Bayes factors

To use the program to calculate a Bayes factor for your data you need to enter your sample mean and standard error. You also need to decide: Does my theory predict a rectangular distribution for the population effect (as in Figure 4.11) or a normal distribution? If rectangular, what are the limits? If a normal, what is its mean and standard deviation? If a normal, does the theory allow both positive and negative effects or only effects in one direction? Answers to these questions will enable the program to calculate: The likelihood of the obtained data given your theory, the likelihood of the obtained data given the null and the Bayes factor.

Notes

1. If the theory allows only effects in one direction, and you want to specify a normal for the distribution of the population effect given the theory, enter zero as the mean of the normal. The program will also presume that the theory predicts positive effects in this case.
2. The null hypothesis is assumed to be that the population mean is zero.
3. It is assumed your likelihood can be represented by a normal distribution (and the prior by normal or rectangular). The adequacy of the results depends on the plausibility of that assumption. Note you can test hypotheses relevant to the contrasts described in Box 4.8.

Box 4.8 What about ANOVA? (for readers already exposed to the analysis of variance)

The statistical tool psychologists use most often is analysis of variance (ANOVA). This box will briefly indicate a possible Bayesian approach to various ANOVA designs. In the Neyman–Pearson approach, if there are more than two conditions, an F test can be conducted to answer the question of whether any one condition is different from any other. I hope it strikes you that this question is not really the question you typically want answered. What you want to know is by how much does *this* condition differ from *that* condition, or by how much does the average of these two conditions differ from the average of those three, or some other equally specific question. The only reason one conducts an overall F test is to protect against the inflated Type I error rate that results from multiple testing. With three groups, there are three t-tests that could be conducted comparing each group with each other. In classical statistics, one has to worry about the fact that three tests are conducted at once. So a typical procedure is to first conduct an overall ('omnibus') F, testing if one can reject the null that all group means are equal. If and only if that is significant does one continue with a post hoc test for pair-wise comparisons. But from a Bayesian point of view, one can adjust ones prior beliefs in the size of any difference without regard for multiple testing. There is no need for the omnibus F, one can just get straight down to asking the questions one is interested in.

Given a set of independent variables, a specific question (including a main effect or interaction) can generally be formulated as a contrast, i.e. a difference between averages of groups, as described in Box 3.7. Comparing group 1 with group 2 is a contrast. Comparing the average of groups 1 and 2 with the average of groups 3 and 4 is a contrast. If the design is completely repeated measures (all subjects have a score for all conditions), then you can already perform a Bayesian analysis on any contrast that interests you.

For each subject, calculate the contrast of interest (e.g. average of conditions 1 and 2 minus the average of conditions 3 and 4). Now for each subject you have a number, the mean of which is the mean of your likelihood and the standard error of which is the standard deviation of your likelihood. So you can determine prior, likelihood and posterior just as we have done in Boxes 4.3–4.5. Perform as many contrasts in this way as address questions you are interested in.

Box 4.8 *continued*

For a between-subjects design, we can use the formulae in Box 3.7. Say we have four groups with means m_1, m_2, m_3 and m_4. We can represent our contrast as a set of numbers, a_i. For example, the difference between groups 1 and 2 is a contrast, $C = (1)m_1 + (-1)m_2 + (0)m_3 + (0)m_4 = m_1 - m_2$. In this case, $a_1 = 1$, $a_2 = -1$, $a_3 = 0$ and $a_4 = 0$. If we wanted to compare the average of groups 1 and 3 with the average of groups 2 and 4, we have the contrast $C = (0.5)m_1 + (0.5)m_3 + (-0.5)m_2 + (-0.5)m_4 = \frac{1}{2}(m_1 + m_3) - \frac{1}{2}(m_2 + m_4)$. In this case, $a_1 = 0.5$, $a_2 = -0.5$, $a_3 = 0.5$ and $a_4 = -0.5$. For roughly normally distributed data within each group, C is roughly normally distributed with standard error given by $\sqrt{(\sum a_i^2)}SD_p/\sqrt{n}$ where n is the number of subjects in each group and $SD_p = \sqrt{(1/4(SD_1^2 + SD_2^2 + SD_3^2 + SD_4^2))}$. Now you can work out priors, likelihoods and posteriors for any contrast of interest in a between-subjects design.

As a Bayesian, you can kiss ANOVA goodbye forever. Do you feel sad?

A compiled version will be available at
http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/

```
normaly = @(mn, variance, x) 2.718283^(- (x - mn)*(x -mn)/(2*variance))/
realsqrt(2*pi*variance);

        sd = input('What is the sample standard error? ');
        sd2 = sd*sd;
        obtained = input('What is the sample mean? ');

        uniform = input('is the distribution of p(population value|theory)
 uniform? 1= yes 0=no ');

      if uniform == 0
           meanoftheory = input('What is the mean of p(population value|
theory)? ');
           sdtheory = input('What is the standard deviation of p(population
value|theory)? ');
           omega = sdtheory*sdtheory;
           tail = input('is the distribution one-tailed or two-tailed?
(1/2) ');
      end

      if uniform == 1
           lower = input('What is the lower bound? ');
           upper = input('What is the upper bound? ');
      end

area = 0;
if uniform == 1
    theta = lower;
else theta = meanoftheory - 5*(omega)^0.5;
end
if uniform == 1
    incr = (upper- lower)/2000;
else incr = (omega)^0.5/200;
end
```

```
for A = -1000:1000
      theta = theta + incr;
      if uniform == 1
          dist_theta = 0;
          if and(theta >= lower, theta <= upper)
              dist_theta = 1/(upper-lower);
          end
      else %distribution is normal
           if tail == 2
              dist_theta = normaly(meanoftheory, omega, theta);
          else
              dist_theta = 0;
              if theta >0
                  dist_theta = 2*normaly(meanoftheory, omega, theta);
              end
          end
      end

      height = dist_theta * normaly(theta, sd2, obtained);
      %p(population value=theta|theory)*p(data|theta)
      area = area + height*incr; %integrating the above over theta
end

Likelihoodtheory = area
Likelihoodnull = normaly(0, sd2, obtained)
Bayesfactor = Likelihoodtheory/Likelihoodnull
```

Appendix Getting personal odds directly

Say somebody is willing to pay you 1 unit of money if a theory is found to be true. If you wish to take the bet, you must reciprocate by paying them a specified amount if the theory is found to be false. What is the maximum specified amount which you are just willing to pay? This maximum amount is called your odds in favour of the theory. I will assume that you are a betting person and you will always bet if you think the bet either favourable or indeed simply fair. Consider the theory that the next toss of this coin will be heads. I will pay you a pound if the next toss is heads. Will you play with me if I want 50p if the next toss is tails? Most people would, but there is not a right answer; it is up to you. Will you play with me if I want 90p if the next toss is tails? £1? £1.50? Assuming the highest amount you would just be willing to accept paying is £1, then your personal odds in favour of next toss being heads is 1 (i.e. 1:1, 50:50, or even odds).

You can convert odds to probability by bearing in mind that

$$\text{odds(theory is true)} = \text{probability(theory is true)}/\text{probability(theory is false)}.$$

Hence

$$\text{probability(theory is true)} = \text{odds}/(\text{odds}+1).$$

Thus, in this case, with odds of 1, probability(next toss is heads) = $1/(1+1) = 0.5$.

Now consider the theory that there is a two-headed winged monster behind my office door. I will pay you a pound if we open the door and find a monster. Will you play if I want 50p if there is no monster? No? How about 25p then? 0p? Assuming you picked 0 as the highest amount, your odds in favour of the theory being true are 0. So your personal subjective probability of the theory being true is also 0.

Now consider the theory it will snow tomorrow. I will pay you a pound if it does snow tomorrow. Will you play if I want you to pay me 50p if it does not snow? If you think it is very likely to snow, this is a very good bet for you. Chances are it will snow, I pay you a pound and you pay me nothing. Will you play if I want a pound if it does not snow? That is still a good bet if you think it is very likely to snow. Will you play if I want £2 if it does not snow? This may be your cut-off point. In that case your odds are 2 (in other words, 2 to 1), and your personal probability that it will snow tomorrow is $2/(2+1) = 0.67$.

Finally, to consider an extreme, consider the theory that there will be traffic in London tomorrow. I will pay you a pound if there is traffic. Will you play if I want you to pay me £1 if there is no traffic? Presumably, you will find this an attractive proposition, a good way of fleecing some idiot out of a pound. How about if I want £10 if there is no traffic? Presumably, you are so certain that there will be traffic in London that you will still unflinchingly accept the game. In fact, you would presumably be willing to play the game even if you have to pay some arbitrarily high number if there is no traffic; your personal odds in favour of the theory that there will be traffic in London tomorrow are very large. Thus, your personal probability that there will be traffic in London tomorrow is (a very large number)/(a very large number +1), which is very close to 1.