3 Neyman, Pearson and hypothesis testing

In this chapter we will consider the standard logic of statistical inference. That is the logic underlying all the statistics you see in the professional journals of psychology and most other disciplines that regularly use statistics. It is also the logic underlying the statistics taught in almost all introductory statistics courses for social or natural scientists. This logic is called the Neyman–Pearson approach, though few users of it would recognize the name.

Statistics are often taught to scientists in a peculiarly uninspiring cook-book style. Thus, many users of statistics would be surprised to learn that the underlying (Neyman–Pearson) logic is both highly controversial, frequently attacked (and defended) by statisticians and philosophers, and even more frequently misunderstood by experienced researchers (and many writers of statistics textbooks). The result of these misunderstandings makes a real practical difference to what research is conducted and how it is interpreted. It makes a real practical difference not just to researchers but to anybody who is a consumer of research. It makes a difference, therefore, to just about everybody.

The material of this chapter illustrates the relevance of understanding core philosophical issues. Fisher (1935, p. 2) said, 'The statistician cannot excuse himself from the duty of getting his head clear on the principles of scientific inference, but equally no other thinking person can avoid a like obligation.'

A common retort to the use of statistics is 'there are lies – damned lies – and statistics' (a phrase apparently coined by Leonard Courtney in 1895[1]). The latter category certainly includes untruths based on the misuse of statistics *but also* valuable knowledge that could only be gained by statistical inference. The misuse of statistics does not make statistics bad; it makes it important for people to understand how to use statistics. But what is the appropriate way of making statistical inferences?

In this chapter we will clarify the common confusions concerning the Neyman–Pearson logic and discuss common objections. The issues will be especially highlighted by showing alternative ways statistics could be done – the Bayesian and likelihood approaches – described in the next two chapters. We will go through key arguments motivating each approach.

This chapter presumes that you have had some exposure to significance testing. You should understand the following concepts before reading the chapter: standard deviation, standard error, null hypothesis, distribution, normal distribution, population, sample and significance. The issues we will talk about apply to any significance test; we will use the *t*-test as an example and I will presume you understand what a *t*-test is and have used one before. If you have not come across these concepts before, consult an elementary statistics book; for example, Wright (2002). If you have come across these concepts before, you might like to revise the logic of a *t*-test, from whichever book you have, before continuing.

To limber up with, consider the example in Box 3.1 and answer the questions. We will return to the answers later in the chapter.

1. See http://www.york.ac.uk/depts/maths/histstat/lies.htm. In 1924 the phrase was mistakenly attributed to Disraeli by Mark Twain.

Box 3.1 Limbering up

You compare the means of your control and experimental groups (20 subjects in each sample). Your result is $t(38) = 2.7$, $p = 0.01$. Please decide for each of the statements below whether it is 'true' or 'false'. Keep a record of your decisions.

(i) You have absolutely disproved the null hypothesis (that there is no difference between the population means).
(ii) You have found the probability of the null hypothesis being true.
(iii) You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
(iv) You can deduce the probability of the experimental hypothesis being true.
(v) You know that if you decided to reject the null hypothesis, the probability that you are making the wrong decision.
(vi) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

From Oakes (1986)

Fisher, Neyman and Pearson

Many of the techniques and concepts we use in statistics come from the British genius, Sir Ronald Fisher (1890–1962). At the beginning of last century he coined the terms 'null hypothesis' and 'significance', urged the systematic distinction between sample and population, introduced degrees of freedom into statistics, suggested a p of 0.05 as an arbitrary but convenient level to judge a result significant (a convention that has since become sacrosanct), urged random assignment to conditions, and proposed many techniques, including the analysis of variance. These are just some of his achievements you may recognize; the true extent of his accomplishments runs even deeper. Fisher created much of what we users of statistics now recognize as statistical practice.

In 1947, Fisher gave this advice to young scientists on BBC radio:

> A scientific career is peculiar in some ways. Its raison d'être is the increase of natural knowledge. Occasionally, therefore, an increase in natural knowledge occurs. But this is tactless, and feelings are hurt. For in some degree it is inevitable that views previously expounded are shown to be either obsolete or false. Most people, I think, can recognise this and take it in good part if what they have been teaching for 10 years or so comes to need a little revision; but some undoubtedly take it hard, as a blow to their amour proper, or even as an invasion of the territory they have come to think of exclusively their own, and they must react with the same ferocity as we can see in the robins and chaffinches these spring days when they resent an intrusion into their little territories. I do not think anything can be done about it. It is inherent in the nature of the profession; but a young scientist must be warned and advised that when he has a jewel to offer for the enrichment of mankind some certainly will wish to turn and rend him
>
> (quoted in Salsburg, 2002, p. 51)

Despite Fisher's enormous contributions, it was the Polish mathematician Jerzy Neyman (1894–1981) and the British statistician Egon Pearson (1895–1980) who provided a firm consistent logical basis to hypothesis testing and statistical inference, starting with a series of papers in the 1920s and 1930s. It was not a jewel that Fisher appreciated, and nor did he revise any of his lectures in the light of it. Nonetheless, the work of Neyman and Pearson transformed the field of mathematical statistics and defined the logic that journal editors in psychology, medicine, biology and other disciplines came to demand of the papers they publish.

Egon's father, Karl Pearson (as in the Pearson correlation coefficient), held the Galton Chair of Statistics at University College London. When Karl retired, his department split into two, with Egon Pearson becoming Head of the Department of Applied Statistics, and Fisher, for some years, Head of the Department of Eugenics. Despite being in the same building, Fisher avoided any contact with Egon Pearson. Neyman, briefly a lecturer at University College London after leaving Poland, left to set up a statistical laboratory at University of California at Berkeley in 1938, a lab that would become under Neyman's leadership one of the most prestigious in the world. Fisher made repeated attempts in print to rend Neyman, but it was Neyman's philosophy of statistical inference that largely prevailed.

Probability

We start by considering what probability means. The meaning of probability we choose determines what we can do with statistics, as will become clear over the course of this chapter and next. The proper way of interpreting probability remains controversial, so there is still debate over what can be achieved with statistics. The Neyman–Pearson approach follows from one particular interpretation of probability; the Bayesian approach considered in the next chapter follows from another.

Interpretations often start with a set of axioms that probabilities must follow. The first axioms were formulated by Christiaan Huygens in 1657, though these considered probability in only one narrow application, gambling. The Russian polymath Andrei Kolmogorov (1903–1887) put the mathematical field of probability into respectable order in 1933 by showing that the key mathematical results concerning probability follow from a simple set of axioms. Box 3.2 lists a standard set of probability axioms.

Box 3.2 The axioms of probability

For a set of events A, B, ...

1. $P(A) \geq 0$
   All probabilities must be greater than or equal to zero.
2. If S refers to one or other of the events happening then $P(S)=1$
   The probability of at least one event happening is 1.
3. $P(A \text{ or } B)=P(A) + P(B)$ if A and B are mutually exclusive.
   A and B are mutually exclusive if they cannot both happen.
4. $P(A \text{ and } B)=P(A) * P(B|A)$

where $P(B|A)$ is 'the probability of B given A'. Axiom (4) is often presented as a definition of $P(B|A)$ rather than an axiom as such. The next chapter will explain and make use of axiom (4).

Probabilities are things that follow the axioms of probability. But what sorts of things? How could the abstract mathematical entity picked out by the axioms make contact with the real world? We will broadly distinguish two interpretations of probability: the subjective and the objective. According to the subjective interpretation, a probability is a degree of conviction in a belief. For example, if I say it will probably snow tomorrow, it means I have a certain conviction in my belief that it will snow. (The "events" in Box 3.2 are statements I believe in to various degrees. For example, $P(A)$ is the extent to which I believe in statement A.) By contrast, objective interpretations locate probability not in the mind but in the world.

Sometimes people say we only need to use probabilities in situations where we are ignorant; if we could know enough to make exact predictions, we would not need to talk in terms of probabilities. That is, probabilities just reflect our subjective states, our states of knowledge or ignorance. If I knew the exact state of nature today in nanoscopic detail, I would not need to talk about the probability of it snowing tomorrow. I would just talk about whether it would snow or not. If I knew the precise velocity and position of every gas molecule in a box, I would not need to calculate the probability of a molecule hitting a wall to work out the gas pressure: I would just work it out from summing all the molecules actually hitting it. Popper (1982) argued that such subjective interpretations of probability did not apply in science. The statistical laws producing gas pressure, or governing why gas will escape the box when opened but not spontaneously return to it, are nothing to do with our knowledge or ignorance of the locations of particular molecules. It is irrelevant whether or not we know what every molecule is doing; the statistical laws apply regardless. According to the objective interpretation of probability, probabilities are not in the mind but in the world. It is an objective fact of the world that gas molecules have a certain probability of hitting a wall, or escaping through an opening. Such objective probabilities exist independently of our states of knowledge. Probabilities are to be *discovered* by examining the world, not by reflecting on what we know or on how much we believe. If I want to know whether a coin has a probability of $\frac{1}{2}$ of producing heads, I need to examine the coin and how it is tossed, and hopefully discover the relevant probability.

The most influential objective interpretation of probability is the long-run relative frequency interpretation of von Mises (1928/1957). A probability is a relative frequency. For example, if I toss a coin, the probability of heads coming up is the proportion of times it produces heads. But it cannot be the proportion of times it produces heads in any finite number of tosses. If I toss the coin 10 times and it lands heads 7 times, the probability of a head is not therefore 0.7. A fair coin could easily produce 7 heads in 10 tosses. The relative frequency must refer therefore to a hypothetical infinite number of tosses. The hypothetical infinite set of tosses (or events, more generally) is called the *reference class* or *collective*. The collective might be, for example, the set of all potential tosses of this coin using a certain tossing mechanism – for example, my right arm, given the command sent to it to throw the coin. Because the long-run relative frequency is a property of all the events in the collective, it follows that a probability applies to a collective, not to any single event. That is, I cannot talk about the probability of the next toss being a head. The next toss, as a singular event, does not have a probability; only the collective of tosses has a probability. The next toss is just heads or tails.

One way to see clearly that probabilities do not apply to singular events is to note that a single event could be a member of different collectives. Take a particular coin toss. It lands heads. On the one hand, that event is part of a collective defined by my right arm repeatedly given the instruction 'toss the coin!'. For this collective, the precise way my arm moves on each trial is variable, and the probability of heads is $\frac{1}{2}$. On the one hand, that same toss

is part of another collective involving events with a more precise description: My right arm moving in a very precise way, with wind conditions being very precisely so, such that every toss results in heads. For this collective, the long-run relative frequency of heads is 1. The event, part of both collectives, does not have a long-run relative frequency. A singular event does not have a probability, only collectives do.

Consider another example. What is the probability I will get cancer? If the reference class or collective is the set of all men, we can look at the proportion of men with cancer. If the collective is men who do not smoke, we can look at the proportion of non-smoking males who get cancer. I am part of both collectives, but each has a different probability associated with it. Objective probabilities do not apply to single cases. They also do not apply to the truth of hypotheses. A hypothesis is simply true or false, just as a single event either occurs or does not. A hypothesis is not a collective, it therefore does not have an objective probability.

Data and hypotheses

Let us symbolize some data by D and a hypothesis by H. We can talk about $P(D|H)$, the probability of obtaining some data given a hypothesis; for example P('getting 5 threes in 25 rolls of a die'|'I have a fair die'). We can set up a relevant collective. Each event is 'throwing a fair die 25 times and observing the number of threes'. That is one event. Consider a hypothetical collective of an infinite number of such events. We can then determine the proportion of such events in which the number of threes is 5. That is a meaningful probability we can calculate. However, we *cannot* talk about $P(H|D)$, for example P('I have a fair die'|'I obtained 5 threes in 25 rolls'), the probability that the hypothesis that I have a fair die is true, given I obtained 5 threes in 25 rolls. What is the collective? There is none. The hypothesis is simply true or false.

$P(H|D)$ is the inverse[2] of the conditional probability $p(D|H)$. Inverting conditional probabilities makes a big difference. For example, consider

$$P(\text{'dying within two years'}|\text{'head bitten off by shark'}) = 1$$

If you take all the people who have had their head bitten off, clean off, by a shark, and then see if they are dead 2 years later, you will find that they all are. The probability is 1.

$$P(\text{'head was bitten off by shark'}|\text{'died in the last two years'}) \sim 0$$

By contrast, if you went to all the places where the dead are kept, the morgue, the graveyard, bottom of the sea and so on, dug up all the bodies of people who had died in the last 2 years, and found how many were missing heads because of a shark bite, you would find practically none were. The probability is very close to zero.

In general, $P(A|B)$ can have a very different value from $P(B|A)$. If ever you feel doubtful about that, consider the shark example.

If you know $P(D|H)$, it does not mean you know what $P(H|D)$ is. There are two reasons for this. One is that inverse conditional probabilities can have very different values. The other is that, in any case, it is meaningless to assign an objective probability to a hypothesis.

2. Really it is just one sort of inverse; taking the reciprocal of $p(D|H)$ would be another, for example.

Hypothesis testing: α

Neyman and Pearson subscribed to an objective relative frequency interpretation of probability. Thus, statistics cannot tell us how much to believe a certain hypothesis. What we can do, according to Neyman and Pearson, is set up decision rules for certain behaviours – accepting or rejecting hypotheses – such that in following those rules in the long run we will not often be wrong. We can work out what the error rates are for certain decision procedures and we can choose procedures that control the long-run error rates at acceptable levels.

The decision rules work by setting up two contrasting hypotheses. One, H_0, is called the null hypothesis. For example, H_0 could be μ_1 (population mean blood pressure given drug) $= \mu_2$ (population mean blood pressure given placebo). The alternative hypothesis is symbolized as H_1 and could be $\mu_1 < (\mu_2 - 10)$ (i.e. that the drug reduces blood pressure by at least 10 units).[3] The null hypothesis need not be the hypothesis of no difference (e.g. $\mu_1 = \mu_2$ or in other words that $\mu_1 - \mu_2 = 0$), it could be a hypothesis concerning a specific difference (e.g. $\mu_1 - \mu_2 = 5$) or a band of differences (e.g. $\mu_1 > \mu_2$). The alternative could also be a hypothesis of some specific difference as well as being a band. The only difference between the null and the alternative is that the null is the one most costly to reject falsely. For example, a drug company may find it more costly to bring a drug forward for further testing (for toxicity, etc.) if the drug is actually useless than to fail to detect that the drug is useful.[4] In that case, the null would be the hypothesis of no difference.

There is a convention in statistics that 'parameters', which are properties of *populations*, are symbolized with Greek letters, like μ (mu). 'Statistics', which are summaries of *sample* measurements, are symbolized by Roman letters, like M. M could be, for example, the mean of your sample. The null and alternative hypotheses are about population values (parameters); they are not about particular samples. We wish to use our samples to make inferences about the population. Make sure in formulating your null and alternative hypotheses you do not refer to sample properties, just population properties.

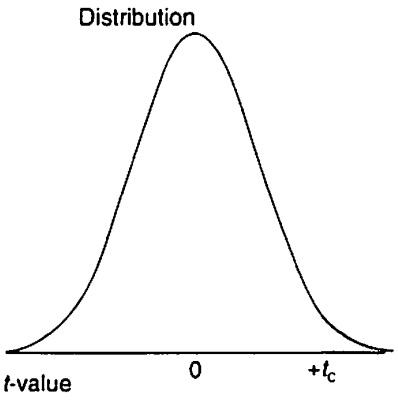
We collect data on the blood pressure of people given either a drug or a placebo. Let us call the sample mean blood pressure with drug M_d and the sample mean blood pressure with placebo M_p. Call the standard error of difference SE. Thus, we have $t = (M_d - M_p)/SE$.

In order to calculate any probabilities, we need a collective or reference class. A collective can be constructed by assuming H_0, and imagining an infinite number of experiments, calculating t each time. Each t is an event in the collective. The distribution of the infinite number of ts in the collective can be plotted, as in figure 3.1.

Next we work out a 'rejection region'; that is, values of t so extreme (as extreme or more extreme than a critical value, t_c) that the probability of obtaining a t in that region is equal to some specified value, α (alpha). If our obtained t is as extreme or more extreme than the critical value, we reject H_0. Figure 3.1 illustrates the probability distribution of t. The shaded area is the rejection region. The area under the whole curve is 1 and the area of the rejection region is α; this just means the probability of obtaining a t in the rejection region is α if the null hypothesis is true. You may know α by another name; it is just the level of significance

3. '<' means 'less than'. That should be easy to remember: The more narrow end is always on the side of the smaller element.
4. This cost–benefit analysis is unlikely; normally a drug company would regard as more important that they did not miss a new effective drug rather than they brought a useless one forward to the next stage of testing. In this case, the hypothesis of 'no difference' should NOT be H_0 by the strict logic of the Neyman–Pearson approach. But which way round one names the hypotheses does not really matter in the end, so long as the relative error rates are controlled appropriately. So if H_0 is the hypothesis of no difference, in what follows β should actually be kept below α if the drug company regards Type II errors as worse than Type I errors. Type I, Type II and α and β are explained in the text that follows.

Figure 3.1



The probability distribution of t. The shaded area is a predetermined rejection region.

we set in advance. 'Significance' is Fisher's term; Neyman preferred the term 'size of the test' (the bigger the shaded area, the bigger the 'size' of the test). The term 'size' never caught on, so we will use the name 'significance', which is now universally used (though maybe Neyman and Pearson introduced the term 'size' to escape the Fisherian connotations of the term 'significance'. We will return to the difference between Fisher and Neyman–Pearson later). By convention α is typically set at 0.05. Assuming the null hypothesis is true, if we run an indefinite number of experiments following this decision procedure, in the long run we will reject the null hypothesis in 5% (i.e. α) of the experiments.

Put another way, for a given experiment we can calculate $p = P$('getting t as extreme or more extreme than obtained'$|H_0$), which is a form of $P(D|H)$. This p is the 'p-value' or simply 'p' in statistical computer output. If p is less than α, the level of significance we have decided in advance (say 0.05), we reject H_0. By following this rule, we know in the long run that when H_0 is actually true, we will conclude it false only 5% of the time. In this procedure, the p-value has no meaning in itself; it is just a convenient mechanical procedure for accepting or rejecting a hypothesis, given the current widespread use of computer output. Such output produces p-values as a matter of course.

α is an objective probability, a relative long-run frequency. It is the proportion of errors of a certain type we will make in the long run, if we follow the above procedure and the null hypothesis is in fact true. Conversely, neither α nor our calculated p tells us how probable the null hypothesis is; they are not $P(H|D)$, as we will now illustrate in a concrete example.

I have a coin I know is biased such that P(heads) = 0.6. Imagine I have some very reliable way of knowing it is biased. Maybe the metallurgist who made it is an expert in making coins biased to a specified amount and he has tested it on a million trials to be sure. He gives it to me. I throw it six times and it lands heads three times. We test a null hypothesis H_0 that P(head) = 0.5. We will set α as 0.05 as per convention. Our p-value is P('getting data as extreme or more extreme as three heads'$|H_0$). Any possible outcome is more extreme than the one we obtained; getting four heads out of six is more extreme than getting three out of six. Getting four tails is more extreme than getting three tails out of six. In fact, all possible outcomes are as extreme or more extreme than the one we obtained. Thus, $p = 1$, greater than 0.05. Non-significant! But that does not change our conviction that the coin is biased – we know it is biased. Indeed, getting three heads out of six tosses is pretty likely, given a

coin with a P(heads) of 0.6. The probability of the null hypothesis (i.e. that the coin is fair) is not 1. Our p-value, a form of $P(D|H_0)$ (and equal to 1 in this case), does not tell us $P(H_0|D)$ (we know H_0 to be false; that is, our subjective probability $P(H_0|D)$ is 0). The p-value does not tell us the probability of the null being true. (If your statistics textbook tells you otherwise – and many do – I suggest you get a new statistics textbook.)

Maybe when the metallurgist gave me the coin, and I threw it six times, it landed heads four times. That is a sample proportion of 4/6 or 0.67, closer to P(heads) of 0.6, the true value, than 0.5, the value we happen to be treating as the null. Our p-value, that is the probability of obtaining a sample proportion as extreme or more extreme than 0.67, given the null hypothesis is true, can be calculated to be 0.69. p is greater than 0.05, so the result is once again non-significant! The null is not rejected by our decision procedure. But clearly, the probability of the null hypothesis is not 0.69. We know the null is false. We just did not throw the coin enough times to have a powerful test.

Our procedure tells us our long-term error rates BUT it does not tell us which particular hypotheses are true or false or assign any of the hypotheses a probability. Remember objective probabilities are properties of a collective, not an individual event. We cannot assign an objective probability to THIS experiment, and we especially cannot assign one to any hypothesis tested in the experiment. All we know are our long-run error rates.

Hypothesis testing: α and β

α is the long-term error rate for one type of error: Saying the null is false when it is true. But there are two ways of making an error with the decision procedure. Table 3.1 shows the situation. We divide the world into two: Either the null hypothesis is true or it is false. We make a binary decision: Accept the null (and reject the alternative) hypothesis or reject the null (and accept the alternative). We are right when the null is true and we accept it and when the null is false and we reject it. A Type I error is when the null is true and we reject it. In the long run, when the null is true, we will make a Type I error in α proportion of our decisions. We can also make an error by accepting the null when it is false. This is called a Type II error and is illustrated by the coin tossing example just given. In that example, the coin was biased, but the results were non-significant. In the long run, when the null is false, the proportion of times we nonetheless accept it is labelled β (beta). That is, $\alpha = P$(rejecting $H_0|H_0$) and $\beta = P$(accepting $H_0|H_0$ false). Both α and β should be controlled at acceptable levels.

Table 3.2 illustrates an imaginary case in which 4000 true null hypotheses and 1000 false null hypotheses are tested. In practice, of course, we never know how many of the null hypotheses we test are true and how many are false. What α rate and what β rate does the table illustrate?

Table 3.1 Two ways of making an error

	State of world	
Decision	**H_0 true**	**H_0 false**
Accept H_0		Type II error
Reject H_0	Type I error	

Table 3.2 A world in which 80 % of the nulls tested are in fact true

	State of world	
Decision	**H_0 true**	**H_0 false**
Accept H_0	3800	500
Reject H_0	200	500
Totals	4000	1000

Sometimes significance or α is defined simply as 'the probability of a Type I error', but this is wrong. α is specifically the probability (long-run relative frequency) of a Type I error *when the null hypothesis is true.* So to work out α from Table 3.2, we must consider only the cases when the null hypothesis was true. The proportion of Type I errors shown when the null hypothesis is true is 200/4000 = 0.05. That is, the researchers used the conventional 5 % level for all their tests. Another meaning of 'the probability of a Type I error' could be the probability of a Type I error *when we have rejected the null.* This probability is determined by considering only the cases where we have rejected the null (i.e. the 200 + 500 = 700 cases shown in the "*Reject* H_0" row). The proportion of Type I errors that occur when we have rejected the null is 200/700 = 29 %. This illustrates that strictly using a significance level of 5 % does not guarantee that only 5 % of all published significant results are in error (Oakes, 1986)!

The β illustrated in Table 3.2 is 500/1000 = 0.5. Notice how β can be very different from α. Controlling one does not mean you have controlled the other. *Power* is defined as $1-\beta$. That is, power is $P(\text{reject } H_0 | H_0 \text{ false})$, the probability of detecting an effect, given an effect really exists in the population.

In the Neyman–Pearson approach, one decides on acceptable α and β levels before an experiment is run. By convention, α is generally set at 0.05. If one thinks Type II errors are as important as Type I, one would also set β at 0.05 as well (i.e. power = 0.95). If one thought Type II errors were not quite as important, one could set β at, for example, 0.10. How does one control β at the level you wish?

In order to control β, you need to:

(1) Estimate the size of effect (e.g. mean difference) you think is interesting, given your theory is true. For example, if a drug company is testing different drugs for their ability to reduce blood pressure, they may not be interested in a drug (/dose) that reduced blood pressure by only 1 mmHg (millimeter of mercury). They need to detect drug doses that reduce blood pressure by at least 5 mmHg. The drug example is a practical one; we can say we want to detect an effect of at least 5 mmHg because we know what use we want to put the drug to. If you are testing a psychological theory, the theory may not seem to say what size effect is minimally interesting. The theory might just seem to say one condition will be faster than another. One way of getting a fix on what size difference would be meaningful is to see what size difference has in the past been explained by this theory in other contexts or by similar theories. What would you think if you found one condition differed from the other by only 1 ms? If previous applications of this or similar theories were for effects of 10–15 ms, then other things being equal, the theory should predict an effect of about 10 ms in this new context. You may decide a difference a little less than 10 ms would be acceptable. If a 5 ms effect would be *just* acceptable to you as the size of difference you would expect on the theory, then that is your minimal meaningful difference needed for a power analysis. Our

alternative hypothesis becomes $\mu_1 - \mu_2 > 5$ ms (the null is $\mu_1 - \mu_2 = 0$ ms). In general, the bigger the minimal effect you are interested in detecting, then the more power you have for detecting it, other things being equal.

(2) Estimate the amount of noise your data will have. The probability of detecting an effect depends not just on how big the effect is, but how much noise there is through which you are trying to detect the effect. In a between-participants design, for example, the noise is provided by the within-group variance. The greater the population variance, the less the power you have. You can use the variances from past similar studies to estimate the population variance in your case. Or you can run a pilot study with your procedures to get an estimate.

Having determined (1) and (2) above, you can use standard statistics textbooks to tell you how many participants you need to run to keep β at 0.05 (equivalently, to keep power at 0.95) (e.g. Howell, 2001, contains a chapter on how to calculate power and determine sample sizes for t-tests; see Murphy and Myors, 2004, for power analysis for many different types of design). The more participants you run, the greater your power. Typically people are astonished at how many participants they need to achieve decent power, say 0.95. A typical reaction is to reduce the power of the study first to maybe 0.90, then finding the number of participants still too much, reducing further to 0.80, and settling for that.

Power in practice

Most studies do not systematically use power calculations to determine the number of participants. But they should. Strict application of the Neyman–Pearson logic means setting the risks of both Type I and II errors (α and β) in advance. Many researchers are extremely worried about Type I errors but allow Type II errors to go uncontrolled. Ignoring the systematic control of Type II errors leads to inappropriate judgments about what results mean and what research should be done next, as we will discuss next. First have a go at the exercise in Box 3.3; then the exercise in Box 3.4 (both based on Oakes, 1986).

Box 3.3 Picking a sample size

Smith and Jones, working in New York, USA, publish an experiment on a new method for reducing prejudice, with 20 participants in each of two groups, experimental and control. They obtain a significant difference in prejudice scores between the two groups, significant by t-test, $p = 0.02$.

You are based in Sussex, UK, and decide to follow up their work. Before adding modifications to their procedure, you initially attempt as exact a replication as you can in Sussex.

How many subjects should you run?

Box 3.4 Interpreting null results

Presume that like Smith and Jones you ran 20 participants in each group. You obtain an insignificant result in the same direction, $t = 1.24$ ($p = 0.22$)

Box 3.4 *continued*

Should you

(a) try to find an explanation for the difference between the two studies?
(b) regard the Smith and Jones result as now thrown into doubt: you should reduce your confidence in the effectiveness of their method for overcoming prejudice?
(c) run more participants? (how many?)

Many researchers in this sort of situation instinctively respond with both reactions (a) and (b). But assuming the population effect was estimated exactly by the American study and the within-group variance was exactly as estimated by the American study, the power of replicating with the same number of subjects as the original study was only 0.67. That is, even if the same effect was there in the Sussex as American population, the decision procedure would fail to detect it fully a 1/3 of the time. With such a large probability, failing to replicate is uninformative. There is no justification for either trying to invent explanations for the difference in outcomes of the two studies or for thinking that the original findings have been thrown into doubt. (On the second point, note that if the data from the two studies were combined, and analysed together, we have $t(78) = 2.59$, $p = 0.011$.)

For this example, the power provided by different sample sizes are as follows:

Power	*N* per group
0.67	20
0.8	26
0.9	37
0.95	44

A sensible number of participants to use in a replication of this particular study would be around 40. That is, given the estimates of the original study, if you used 40 participants and accepted the null whenever your t-value was less than the critical value, you would only falsely accept the null about 5% of the time.

You are now in a position to consider what would happen if an original study obtained a significant result, with $p = 0.05$ (or 0.0499 if you like). They used an N of 20, and you replicate the experiment with a sample also of 20. Assuming *there really is* an effect of exactly the size found in the original experiment and the population variance had been exactly estimated by the original study, what is the probability that you will also obtain a significant result?

If in your sample you also happen – miraculously – to obtain a mean sample difference exactly equal to the population one and a sample variance exactly equal to the population one, your obtained t-value would be exactly equal to the critical value, just as the original team obtained. You have exactly even odds that your mean difference will be a little bit more rather than a little bit less than the population mean difference. If your mean difference is a little bit more, your t-value will be slightly higher than the critical value; but if your mean difference is a little bit less, your obtained t-value will be a little bit less than the critical value. Similarly, you have about even odds that your variance will be a little bit more or a little bit less than the population value. If your variance is a little bit more than the population value, your t-value will be a little bit less than the critical value; but if your variance is a little bit less than the population value, your t-value will be a little bit more than the critical value.

Box 3.5 Reviewing a set of studies

You read a review of studies looking at whether meditation reduces depression. One hundred studies have been run and 50 are significant in the right direction and the remainder are non-significant.

What should you conclude?

If the null hypothesis were true, how many studies would be significant? How many significant in the right direction?

You are sitting on a knife edge. You have a 50% probability of obtaining a significant result. That is, you have a power of 50%.

In reviewing the 100 studies mentioned in Box 3.5, some reviewers will be tempted to think that the result is thrown in doubt by all the non-significant studies. How often have you seen an author disputing an effect because she can count as many null findings as significant findings? But if the null hypothesis were true one would expect 5% of studies to be significant. That is what a significance of 5% means. There can be a 'file drawer problem'; that is, not all non-significant results may become generally known about, because the experimenter thinks, 'I cannot publish this; I will just file it in my cabinet'. Even so, if the null hypothesis were true, one would expect an equal number of studies showing a significant effect in one direction as the other. Finding 50 studies significant in the right direction (and none in the wrong direction) is highly unlikely to happen by chance alone. The pattern shown in Box 3.5 is just what you would expect if an effect really existed and the average power of the studies was 50% – a power sadly typical of much research.

The process of combining groups of studies together to obtain overall tests of significance (or to estimate values or calculate confidence intervals) is called *meta-analysis* (see Rosenthal, 1993, for an introduction, and Hunter and Schmidt, 2004, for a thorough overview). A convenient and easy technique you can use when you have run a series of studies addressing the same question with the same dependent variable is create a data file with all the experiments in; then test for the effect using all the data.[5] You might have six studies all non-significant but when combined together they are significant! (On reflection that should not be very surprising. You can put it the other way round. Imagine you had a study that was *just* significant at the 0.05 level. If you split it into six equal parts, would you expect any part to be significant on its own?) A set of null results does not mean you should accept the null; they may indicate that you should reject the null. To take a simple example, let us say for each of these six non-significant studies the effect was in the right direction. The probability of obtaining six effects in the right direction assuming either direction is equally likely is $1/2^6 = 1/64$ on a one-tailed test, or 1/32 on a two-tailed test, $p < 0.05$ in both cases. That is, those six non-significant studies allow one to reject the null!

The situations we have just considered should convince you of the importance of power. To summarize, if your study has low power, getting a null result tells you nothing in itself. You would expect a null result whether or not the null hypothesis was true. In the Neyman–Pearson approach, you set power at a high level in designing the experiment, before you run it. *Then you are entitled to accept the null hypothesis when you obtain a null result.* In following

5 You should also put in 'experiment' as a factor and test if the effect interacts with experiment.

this procedure you will make errors at a small controlled rate, a rate you have decided in advance is acceptable to you.

Once you understand power, the need for considering it is self-evident; it may seem strange that so many researchers seem so willful in disregarding it. Why do people ignore power, when it has been part of the orthodox logic for over 70 years? Oakes (1986) suggests that it is because many people interpret the p-value as telling them about the probability of the null (and logically hence the alternative) hypothesis. Bayesian statisticians have developed techniques for actually assigning probabilities to hypotheses in coherent ways. Many people interpret significance levels in a Bayesian way, and a Bayesian has no need for the concept of power. Once I know the probability of my hypothesis being true, what else do I need to know? But Neyman–Pearson methods are not Bayesian, and confusing the two leads to mistakes in conducting research. In the next chapter we will discuss the Bayesian approach.

Now we will return to the questions in Box 3.1, here repeated as Box 3.6. Have a go at the questions again and compare your new answers with your original answers.

Box 3.6 Revisiting Box 1

You compare the means of your control and experimental groups (20 subjects in each sample).

Your result is $t(38) = 2.7$, $p = 0.01$. Please decide for each of the statements below whether it is 'true' or 'false'. Keep a record of your decisions.

(i) You have absolutely disproved the null hypothesis (that there is no difference between the population means).
(ii) You have found the probability of the null hypothesis being true.
(iii) You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
(iv) You can deduce the probability of the experimental hypothesis being true.
(v) You know that if you decided to reject the null hypothesis, the probability that you are making the wrong decision.
(vi) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99 % of occasions.

From Oakes (1986)

Statistics never allows absolute proof or disproof, so most people are happy to reject – correctly – options (i) and (iii). Options (ii) and (iv) may have been initially tempting, but hopefully you see now that as they refer to the probability of hypotheses they cannot be correct (an objective probability refers to a collective of events, not the truth of a hypothesis). Option (v) is a sneaky one and often catches people out; but notice it refers to the probability of a single decision being correct; thus, option (v) cannot be correct (an objective probability does not refer to a single event). Many people, initially, think option (vi) is right. But option (vi) is a description of power, not significance. As Oakes (1986) comments, no wonder people ignore power if they think they already have it in the notion of significance! In fact, none of the options are correct. Oakes put these questions to 70 researchers with at least 2 years of research experience, and including full professors. Only two researchers showed they had a sound understanding of the concept of statistical significance. I dare say not too much has changed in the two decades since Oakes conducted his research.

Sensitivity

Fisher talked loosely about the sensitivity of an experiment; Neyman and Pearson gave precision to one way of specifying sensitivity, namely power. Also both Fisher and Neyman introduced a notion of confidence intervals (which is Neyman's term; his term – and logic for them – stuck), discussed later. Sensitivity can be determined in three ways: power, confidence intervals and finding an effect significantly different from another reference one. We will discuss confidence intervals later. Here is an important methodological maxim which is often ignored: *Whenever you find a null result and it is interesting to you that the result is null, you should always indicate the sensitivity of your analysis.* If in getting a null result you fail to replicate someone's study, you must provide some indication of the sensitivity of your study. For example, you could calculate the power of your procedure for detecting the size of effect the original authors found.[6] Let us say, you find a significant effect of ginseng on the reaction time of men, and in a subsequent study using only women you find a null result. From this pattern of result you cannot conclude that ginseng affects men and women differently. Maybe you had a power of 50% for detecting the effect of ginseng and so found an effect in one study but not another, a success rate which is just to be expected. If, however, you showed that the effect of ginseng on men was significantly different from the effect on women, you have demonstrated a differential effect of ginseng. Similarly, if you found a correlation between two variables was significant in one condition but not another that does not mean the two correlations can be regarded as different. Perhaps you had low power. If, however, a test of the difference between the correlations was significant, you can conclude that the correlations differed. Further, before you could accept the null result in the condition that was null, you would need to show you had appropriate power to pick a minimally interesting effect.

Stopping rules

The conditions under which you will stop collecting data for a study define the *stopping rule* you use. A common stopping rule is to run as many participants as is traditional in that area. The trouble with this rule is that it does not guarantee that power is controlled at any particular value. Another stopping rule one might follow is to continue running until the test is significant. The problem with this rule is that even assuming the null hypothesis is true, you are guaranteed to obtain a 'significant' result eventually if you are prepared to continue collecting data long enough! That may sound astonishing, but it is true. So this procedure, although having a power of 1, also has an α of 1! You are guaranteed to obtain a 'significant' result eventually, whether or not the null hypothesis is true. Running until you have a significant result is not a good stopping rule. You may decide to test for 'significance' with a t-test after every participant, stopping if $p < 0.05$ or until 100 participants had been reached. Then the α would actually be 0.39, not 0.05 at all.

The standard Neyman–Pearson stopping rule is to use power calculations in advance of running the study to determine how many participants should be run to control power at a

6. Some statistical packages report power for effects along with their significance. If the power is calculated assuming the population effect size and noise level that is estimated in that very sample, such power calculations are not informative. If the p-value is less than 0.05, the power will be less than 0.5. You want to calculate power based on the minimal effect size *you are interested in detecting* or maybe the effect size found by prior researchers.

predetermined level. Then run that number of subjects. Both α and β can then be controlled at known acceptable levels. Another legitimate stopping rule involves the use of confidence intervals, which we will describe later.

A development of the standard Neyman–Pearson approach is to consider sequential stopping rules. These are particularly important in clinical trials but can be used in any domain. In clinical trials it may be unethical to continue a study when there is strong evidence that one treatment is more effective than another. In sequential testing you can look at your data several times and decide to stop collecting data at any of those times and still control α at 0.05. If each time you looked you tested nominally at the 0.05 level, α would actually be greater than 0.05 because you have given yourself several chances to reject H_0 and stop. So to control α at 0.05, single tests need to be conducted at less than the 0.05 level. For example, if you decide in advance you will check five times at equal intervals, each test can be conducted at a nominal significance level of 0.016 and overall α is controlled at 0.05. Armitage, et al. (2002, pp. 615–623) provide an introduction and further references, discussing various sequential schemes that also control power.

Multiple testing

In the Neyman–Pearson approach it is essential to know the collective or reference class for which we are calculating our objective probabilities α and β. The relevant collective is defined by a testing procedure applied an indefinite number of times. For example, 'Collect data and calculate the t-value at five points in time; reject H_0 whenever any of the ts individually is greater than 2.41' is a test procedure defining a collective and we can make sure this procedure has definite values of α and β. Another collective is defined by the procedure, 'Calculate the difference between the control group and each of the three experimental groups; reject any of the three corresponding null hypotheses whenever the t is greater than 2'. If this procedure were applied an indefinite number of times where all three H_0s were true, how often would we reject one or more of the H_0? In other words, when we perform a set or 'family' of tests, how often would we make a Type I error? If we conduct one t-test, the probability that it is significant by chance alone is 0.05 if we test at the 0.05 level. If we conduct two t-tests, the probability that at least 1 is significant by chance alone is slightly less than 0.10. If we conduct three t-tests, the probability that at least one is significant by chance alone is slightly less than 0.15. In the Neyman–Pearson approach, in order to control overall Type I error ('familywise error-rate'), if we perform a number of tests we need to test each one at a stricter level of significance in order to keep overall α at 0.05. There are numerous corrections, but the easiest one to remember is Bonferroni. If you perform k tests, then conduct each individual test at the $0.05/k$ level of significance and overall α will be no higher than 0.05. So with the three tests mentioned, each would have to have a p-value less than $0.05/3 = 0.017$ to be accepted as significant.

A researcher might mainly want to look at one particular comparison, but threw in some other conditions out of curiosity while going to the effort of recruiting, running and paying participants. She may feel it unfair if the main comparison has to be tested at the 0.017 level just because she collected data on other conditions but need not have. The solution is that if you *planned* one particular comparison in advance then you can test at the 0.05 level, because that one was picked out in advance of seeing the data. But the other tests must involve a correction, like Bonferroni. A collective defined by 'Calculate three t-tests; if any one of them are significant at the 5% level, reject H_0 for that one' will involve an α of about 0.15. That is

because you can pick and choose which test to accept AFTER the data have been collected; such tests are called *post hoc* ('after the fact'). Thus, post hoc tests require a correction of the nominal significance level; single planned tests do not. Even in the post hoc case, the family of tests is decided in advance of looking at the data; that is, you cannot just look and see which three tests are most significant and choose these for your family. That would defeat the object. The family is chosen in advance.

Fisherian inference

Despite Fisher's contribution to statistics, it was the Neyman–Pearson approach that dominated statistical theory. Neyman provided an internally coherent logic for developing mathematical statistics based on objective probability as a long-run relative frequency. Fisher had a notion of probability neither clearly objective nor subjective, 'fiducial probability'; but few people could understand it. The consensus seems to be that the notion was flawed. Fisher regarded the p-value calculated as a property of the particular sample it was calculated from; in the Neyman–Pearson approach, significance level or α is a property of the testing procedure not of the sample. α is an objective probability and hence a property of a collective and not any individual event, not a particular sample. Fisher wanted significance testing to provide evidence concerning the truth of the null hypothesis; and, according to Fisher, the smaller the p-value, the stronger the evidence provided by the sample against the null hypothesis. According to Fisher, 'From a test of significance... we have a genuine measure of the confidence with which any particular opinion may be held, in view of our particular data' (Fisher, 1955, quoted in Gigerenzer, 1993, p. 318). It was Fisher who motivated the belief that, for example, a p-value less than 1% is very strong evidence against the null hypothesis, between 1 and 5% is moderate evidence, and between 5 and 10% is suggestive or marginal evidence.

On a frequentist notion of probability, this use of the p-value does not follow at all. In the Neyman–Pearson approach, the relevant probabilities α and β are the long-run error rates you decide are acceptable and so must be set in advance. If α is set at 0.05, the only meaningful claim to make about the p-value of a particular experiment is either it is less than 0.05 or it is not. If you obtain a p of 0.009 and report it as $p < 0.01$, that is misleading. You are implying that your test procedure has a long-term error rate of 1%, but in fact if the p had been 0.04, you would still reject the null; in fact, your test procedure has an error rate of 5%.

The Neyman–Pearson approach urges you to engage in a certain behaviour: Whatever your beliefs or degree of confidence, act as if you accept the null when the test statistic is outside the rejection region, and act as if you reject it if it is in the rejection region. Amazingly, the statistics tell you nothing about how confident you should be in a hypothesis nor what strength of evidence there is for different hypotheses.

Fisher found this philosophy irrelevant to the true concern of scientists, if not offensive. But it has an unassailable logical purity. If you want statistics to tell you what confidence you should have, you need to develop statistics based on the subjective notion of probability; that is what the Bayesians did, discussed in the next chapter. The procedures are quite different. Fisher tried to mix the logic of the two.

But, you may object, surely the smaller the p-value the stronger the evidence IS against the null? Perhaps Fisher's muddling the two is actually taking the sensible middle path?

An argument for *p*-values can at first blush be based on the logic Popper used for non-probabilistic theory (though this is not an argument put forward by either Fisher or Popper in the probabilistic case):

If H then NOT R
R

Not H

If a hypothesis says R will not occur, but it does occur, it follows the hypothesis has been falsified. On first blush, one may be tempted to extend the logic to the probabilistic case:

If H then probably NOT R
R

Probably not H

That is, if given the hypothesis is true it is unlikely for data to occur in the rejection region, it follows (according to the schema above) the hypothesis is unlikely to be true when data does in fact fall in the rejection region.

On second blush, as Pollard and Richardson (1987) point out, the logic in the probabilistic case is flawed. For example, consider the obvious fallacious reasoning, 'If a person is American then they are probably not a member of Congress; this person is a member of Congress; therefore this person is probably not American'.

In fact, it is hard to construct an argument for why *p*-values should be taken as strength of evidence per se (see Royall, 1997, Chapter 3). Conceptually, the *strength* of evidence for or against a hypothesis is distinct from the *probability* of obtaining such evidence; *p*-values confuse the two (we will disentangle them in Chapter 5). Part of Fisher's lasting legacy, and a tribute to his genius, is that in 1922 he *did* devise a way of measuring strength of evidence (using 'likelihoods', to be discussed in the next two chapters), an inferential logic in general inconsistent with using *p*-values, but which is the basis of both Bayesian inference and a school of inference of its own, likelihood inference (Edwards, 1972; Royall, 1997). There is no need to force *p*-values into taking the role of measuring strength of evidence, a role for which they may often give a reasonable answer, but not always. Fisher frequently used *p*-values in his examples and that is what many scientists took on board from reading Fisher.

Gigerenzer et al. (1989, Chapter 3) pointed to an analogy between Popper's ideas, in his classic *Logic der Forschung* officially published in 1935 (see Chapter One), and Fisher's *Design of Experiments*, also published in the same year. In coining the phrase 'null hypothesis', Fisher said,

> In relation to any experiment we may speak of this hypothesis as the "null hypothesis", and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis
>
> (p. 18).

Fisher vacillated somewhat about what to conclude when the null hypothesis is not rejected. Because he rejected the notion of power, there remain no clear grounds for accepting the

null when the null fails to be rejected. Perhaps one is just left unable to conclude anything. But just as Popper developed the idea of corroboration for a hypothesis that had withstood attempts at falsification, Fisher later said (in 1955) that while the null is not established it may be said to be 'strengthened'. Popper insisted that theories were corroborated only to the extent that tests were severe; in the statistical case, a null has been tested severely only when the test is powerful. By rejecting power, Fisher opened the way to all the problems of low power experiments discussed above.

Further points concerning significance tests that are often misunderstood

Significance is not a property of populations. Hypotheses are about population properties, such as means, for example that two means are equal or unequal. Significance is not a property of population means or differences. Consider the meaningless statement (often made by undergraduates but not seasoned researchers): 'The null hypothesis states that there will be no significant difference in mean reaction time between the conditions'. Hypothesis testing is then circular: a non-significant difference leads to the retention of the null hypothesis that there will be no significant difference! Type I errors are therefore impossible (Dracup, 1995)! The null hypothesis in this case should be simply that in the population mean reaction time is the same in the two conditions.

Decision rules are laid down before data are collected; we simply make black and white decisions with known risks of error. Since significance level is decided in advance, one cannot say one result is more significant than another. Even the terms 'highly significant' versus 'just significant' versus 'marginally significant' make no sense in the Neyman–Pearson approach. A result is significant or not, full stop. Having decided on a significance level of 0.05, you cannot then use 0.01 if the obtained p is, for example, 0.009. Sample size is also decided in advance. Having decided to run, say, 20 subjects, and seeing you have a not quite significant result, you cannot decide to 'top up' with 10 more and then test as if you had set out to run that amount from the start.

A more significant result does not mean a more important result, or a larger effect size. Maybe 'significant' is a bad name; but so is the alternative sometimes used, 'reliable', which implies the concept of power. Wright (2003) suggests a good term 'detected': 'A difference between the two conditions was not detected, $p > 0.05$'. If you have a sensitive enough test, an effect as small as you like can be significant to as low a level as you like. For example, if you run enough participants, a difference of 0.001 ms could be detected between two conditions. Getting a 'marginally significant' result does not mean the effect was smallish, despite what papers sometimes say. 'Women were somewhat ($p< 0.10$) faster than men in the gossip condition': The speed being 'somewhat faster' is an incorrect inference from the p-value. A large mean difference can be insignificant and a small difference significant – depending on, for example, sample size.

Confidence intervals

The Neyman–Pearson approach is not just about null hypothesis testing. Neyman also developed the concept of a confidence interval, a set of possible population values the data are consistent with. Instead of saying merely we reject one value ('I reject zero'), one reports the

set of values rejected, and the set of possible values remaining. To calculate the 95% confidence interval, find the set of all values of the dependent variable that are non-significantly different from your sample value at the 5% level. For example, I measure blood pressure difference before and after taking a drug. Sample mean difference is 4 units. That may be just significantly different (at the 5% level) from −1 units and also from +9 units. So the '95% confidence interval' runs from −1 to +9 units. All these points are non-significantly different from the sample mean. Outside this interval, out to infinity either way, all the values are significantly different (at the 5% level) from the sample mean, so can be rejected as possible population values. The points within the interval cannot be ruled out as population values. My data are consistent with a population value somewhere between −1 and +9 units. This tells me that the data are consistent both with the drug having no effect AND with the drug having an effect as large as almost 9 units. I know if my study is sensitive enough to detect the sort of effect I would be interested in. Say I am interested in whether the drug has an effect of 5 units or more. The fact that zero is in the confidence interval means the sample is non-significantly different from zero; but since it is also non-significantly different from +5 the data do not distinguish the interesting states of affairs. The drug may be effective. I would need to collect more data to find out.

Use of the confidence interval overcomes some of the problems people otherwise have with Neyman–Pearson statistics:

First, it tells you the sensitivity of your experiment directly; if the confidence interval includes the value of both the null hypothesis and the interesting values of the alternative hypothesis, the experiment was not sensitive enough to draw definitive conclusions. It tells you sensitivity in a way that is more intuitive than power. It directly tells you about effect sizes. Conversely, if the interval includes an interesting effect size and excludes zero, but also includes uninterestingly small values (e.g. in an interval going from +2 to +7, +2 units might be too small an effect to be interesting), you may wish to collect more data as well. If the lower limit of your interval is +3 units, you can act as if the population value is not lower than 3 units, and conclude that you are dealing with effects of a size that interest you. Confidence intervals allow you to act in a more fine-grained way than just rejecting zero.

Second, it turns out you can use the confidence interval to determine a useful stopping rule (see Armitage et al., 2002, p. 615): Stop collecting data when the interval is of a certain predetermined width (a width that is chosen so as to make sure you exclude, for example, either zero or the minimal interesting difference). Such a stopping rule would ensure that people do not get into situations where illegitimate stopping rules are tempting. For example, I decide an interesting drug effect would be +5 units. So I decide to collect data until the width of the interval is +4 units. Then if the interval includes zero it will exclude +5; and if it includes +5 it will exclude zero. There is no need to 'top up' with some more participants because $p = 0.06$.[7] This stopping rule is very sensible but virtually no one knows about it.

If people thought in terms of confidence intervals, rather than just significance tests, they would be more likely to draw appropriate conclusions. For example, they would immediately see if the failure to replicate a previous study was because of a lack of sensitivity. They could tell if an already published null study had sufficient sensitivity. There is a quick interval estimate one can do if a study reports a null effect but does not give a confidence interval. If the mean was 4 units and non-significantly different from zero with a *t*-test, then it was also

7. One could set oneself more demanding tasks. For example, if you wanted to either exclude drugs with an effect of +2 units or lower but accept drugs with an effect of +5 units or higher, you could run until the confidence interval had a width of 2.9 units.

non-significantly different from 8 units. The 95% confidence interval spans at least 0–8 (or, in general, if the mean is m, it spans at least 0–2m; Rosenthal, 1993). If that interval includes interesting effect sizes, the study was not sensitive enough.

Confidence intervals are a very useful way of summarizing what a set of studies as a whole are telling us. You can calculate the confidence interval on the parameter of interest (mean difference, slope, correlation, proportions, etc.) by combining the information provided in all the studies (a type of meta-analysis): 'This manipulation could change preferred meal size by between −5 and 30%'. How much more informative is that than saying 'there were six null results'? It is a tragic waste of data simply to count the number of null results. Maybe a change in meal size of, say, 20% or more is theoretically or practically important. Imagine the unjustified decision to stop pursuing the theory tested that might have followed from just counting null results.

Like all statistics in the Neyman–Pearson approach, the 95% confidence interval is interpreted in terms of an objective probability. The procedure for calculating 95% confidence intervals will produce intervals that include the true population value 95% of the time. There is no probability attached to any one calculated interval; that interval either includes the population value or it does not. There is not a 95% probability that the 95% confidence limits for a particular sample includes the true population mean. But if you acted as if the true population value is included in your interval each time you calculate a 95% confidence interval, you would be right 95% of the time (Box 3.7).

Box 3.7 Using confidence intervals

Confidence intervals can be used in many situations. Anytime your statistical software reports a standard error you can calculate a confidence interval; for example for a regression slope. Confidence intervals can also be used in standard analysis of variance (ANOVA) designs by using *contrasts*. Imagine you test the comprehension of males or females on either a romance or a science fiction story. Based on pilot data you expect the men to understand the science fiction better than the women but the women to understand the romance better than the men (who got lost after the bit where she said, 'Do you love me?'). This design is as illustrated in the table below, where the m_is are the sample means for comprehension scores. Each group also has a standard deviation, SD_i.

	Romance	Science fiction
Males	m_1	m_2
Females	m_3	m_4

A contrast is a difference between the average (or equivalently sum) of some conditions and the average (or sum) of others, which we can represent as a set of numbers, a_i. For example, the difference between m_1 and m_2, which is the effect of text type for men, is a contrast, $C = (1) \times m_1 + (-1) \times m_2 + (0) \times m_3 + (0) \times m_4 = m_1 - m_2$. In this case, $a_1 = 1$, $a_2 = -1$, $a_3 = 0$ and $a_4 = 0$.

The effect of text type for females is another contrast ($m_3 - m_4$). The main effect of gender is the difference between all the men's scores and all the female's scores, $C = (1) \times m_1 + (1) \times m_2 + (-1) \times m_3 + (-1) \times m_4 = (m_1 + m_2) - (m_3 + m_4)$. If you want to test a theory that the genders are socialized differently to understand different topics you would need to show that the difference in understanding between the text types is different for the genders. Thus the *interaction* between gender and text type, $C = (1) \times m_1 + (-1) \times m_2 + (-1) \times m_3 + (1) \times m_4 = (m_1 - m_2) - (m_3 - m_4)$.

Box 3.7 *Continued*

For example, consider the imaginary data below (with standard deviations shown in parentheses), with 20 subjects in each cell:

	Romance	Science fiction
Males	5 (1.5)	8 (2.1)
Females	9 (1.7)	7 (1.9)

The obtained size of interaction is $C = (5 - 8) - (9 - 7) = -5$. That is, female's ability to comprehend the romance rather than the science fiction was 5 comprehension points better than the male's ability to comprehend the romance rather than the science fiction. For roughly normally distributed data and equal variances within each group, C has a standard error $\sqrt{(\sum a_i^2)} \times SD_p/\sqrt{n}$, where n is the number of subjects in each group and $SD_p = \sqrt{(1/4 \times (SD_1^2 + SD_2^2 + SD_3^2 + SD_4^2))}$. For our example data, $SD_p = 1.81$, and $SE = \sqrt{(1+1+1+1)} \times 1.81/\sqrt{20} = 0.81$.

For any contrast, you can calculate a confidence interval:

$$C \pm t_{crit} \times SE$$

where t_{crit} is the critical value of t with $4(n - 1)$ degrees of freedom.

For our example, there are 76 degrees of freedom and (consulting tables of the t-distribution in Howell, 1987) the critical value of t at the 5% level is 1.99. So the confidence interval for the contrast is $(-5) \pm 1.99 \times 0.81$, which gives an interval of $(-6.6, -3.4)$. The interval excludes zero, so the interaction is significant at the 5% level. More informatively, we reject any value less than -6.6 or more than -3.4 as population values of the advantage of females over males in understanding the romance over the science fiction text. The substantial meaning of the result depends on the comprehension scale used. If it were number of key terms recalled out of 15, the interaction strikes me as substantial and the theoretical basis of the interaction well worth pursuing further. Conversely, one could imagine an interaction, though significant, spanning such a small range of effects, it was of no further interest; or an interaction, though non-significant, spanning such an interesting range of effects, more data were required to draw any practical conclusion. Whenever you have an important effect that is non-significant, calculate the confidence interval on the relevant contrast to see what range of effect sizes your data are consistent with.

In a within-subjects design, you can calculate a contrast for each subject separately. Now you end up with a single number for each subject - the contrast - and you can find the mean, C, of these numbers and their standard error ($= SD/\sqrt{n}$), and calculate the confidence interval:

$$C \pm t_{crit} \times SE$$

where t_{crit} is the critical value of t with $n - 1$ degrees of freedom.

The skill is in working out how to translate your research questions into contrasts. Once you have achieved this, you can have all the benefits of confidence intervals over null hypothesis testing in 'ANOVA designs'.

You can also calculate confidence intervals on correlations. Pearson correlations can be transformed to be roughly normal using Fisher's transform

$$r' = (0.5) \ln |(1 + r)/(1 - r)|$$

where r is the correlation, $|x|$ means take the absolute value of x ignoring its sign, and ln is the natural logarithm, a function you can find on your calculator. r' is roughly normally distributed with standard error $SE = 1/\sqrt{(n - 3)}$. The 95% confidence interval is

$$r' \pm 1.96 \times SE$$

r' can be converted back to r, $r' = (e^{2r'} - 1)/(e^{2r'} + 1)$ where $e = 2.71828$.

For example, if on 50 people you obtained a correlation of 0.30, $r' = 0.31$. $SE = 0.15$. The 95% confidence interval for r' runs from 0.02 to 0.60. These limits correspond to correlations of 0.02 and 0.54, respectively.

Criticism of the Neyman–Pearson approach

(1) *Inference consists of simple acceptance or rejection.* According to Neyman, deciding to accept a hypothesis does not mean knowing it or even just believing it to some extent. By strictly following the demands of a frequentist objective probability, statistical inference becomes 'inductive behaviour', as Neyman puts it, rather than the process of strengthening certain beliefs. It is statistical inference that is scarcely inference at all (Oakes, 1986). Do not the data indicate more than just a black and white behavioural decision? Data seem to provide *continuous* support for or against different hypotheses. Fisher tried to capture this intuition with *p*-values. He found it arbitrary that Neyman–Pearson required us to simply behave one way if $p = 0.048$ and the opposite if $p = 0.052$; and identically when $p = 0.048$ as when $p = 0.0001$. The Fisherian urge to use what appears to be useful information carried in the continuous *p*-value, as strength of evidence has been irresistible to researchers (we will consider measuring strength of evidence in Chapter 5).

Arguably, what a scientist wants to know is either how likely certain hypotheses are in the light of data (in which case Bayesian statistics should be used) or, more simply, how strong the evidence supports one hypothesis rather than another (in which case likelihood inference should be used). It is meaningless to use the tools and concepts developed in the Neyman–Pearson framework to draw inferences about the probability of hypotheses or the strength of evidence. We will discuss the alternative frameworks in the next two chapters, which allow graded conclusions, conclusions which can be expressed in terms of intervals similar to confidence intervals.

(2) *Null hypothesis testing encourages weak theorizing.* This is one of the most important criticisms of significance testing as it is generally used. Meehl (1967) contrasted how 'soft' psychology and hard sciences test their theories. In psychology the scientist typically sets up a null hypothesis of no difference between conditions, or of zero correlation between variables. When this point hypothesis is rejected, the scientist regards his own substantive theory as confirmed, or at least as having survived the test. But notice how little content the alternative hypothesis actually has; it rules out virtually nothing, just one point. The hypothesis that '$m_1 <> m_2$' has virtually zero content; the directional hypothesis ('$m_1 - m_2 > 0$') is still weak. The hard scientist, by contrast, often predicts a certain value (not just ruling out one point, 'the force will be different from zero'; not just directional, 'gravity pulls things down rather than up'; but a specific value, e.g. 'the force will be 13 units down'); when the point hypothesis lies within the margin of error of the data her substantive hypothesis has survived the test. Instead of trying to reject the null (to corroborate a theory of virtually no content), she wants to accept the null (to corroborate a highly falsifiable theory). The difference is symptomatic of the relative state of theory development in psychology and hard sciences. Psychologists are often wary of predicting the null and then accepting it; but this is precisely what they should be aiming for.

Most point null hypotheses – of no difference, no correlation and so on (so-called 'nill hypotheses') – are virtually certain to be false; it is just a matter of to what decimal place. A good theory should specify the size of effect not just that it is different from zero.

Of course, as Meehl pointed out, there are difficulties with the subject matter of psychology that make such precise theories very hard to construct. But such theories do exist even within psychology. Why are they so few? The problem may be that by training and requiring psychologists to get excited by rejecting null hypotheses, there is no incentive to think more carefully about effect sizes. If our statistical tools required at least the reporting of the range of effects the data were consistent with, it would encourage more careful theorizing, motivate

more precise theories where they were possible. The habitual use of confidence intervals instead of simple null hypothesis testing would overcome this objection. The objection can be taken as one against how the Neyman–Pearson approach is typically used (or mis-used) rather than against the approach per se. Nonetheless, after reading the next two chapters you may agree that the other approaches invite this misuse less than does the Neyman–Pearson approach.

(3) *In the Neyman–Pearson approach it is important to know the reference class – we must know what endless series of trials might have happened but never did.* This is important when considering both multiple testing and stopping rules. A particular event can be a member of various reference classes. Probabilities, including α and β, belong to reference classes, not singular events like particular experiments. Thus, a particular experiment must be assigned to a reference class. The choice of reference class can be determined by what else the experimenter did even if it was unrelated to the hypothesis under consideration (we will discuss this under multiple testing). Other than the actual experiment, the remainder of the chosen reference class includes an infinite number of experiments that never happened. Which set of experiments that never happened are included defines the reference class and hence the probabilities. It strikes some as unreasonable that what never happened should determine what is concluded about what did happen.

In the Neyman–Pearson approach, as explained above, when conducting a number of tests, there should be a correction for repeated testing. (By contrast, a Bayesian does not have to.) The data relevant to judging one hypothesis will legitimate different judgments depending on how many other hypotheses were tested. If I test one correlation by itself, I can test at the 0.05 level; if I test five correlations, each concerning a different null hypothesis, I should test at the 0.01 level (with Bonferroni correction). When do we correct for repeated testing? We do not correct for all the tests we do in a journal, a paper, or for an experiment, or even in one analysis of variance. Why not? Why should we correct in some cases and not others? The decision is basically arbitrary and tacit conventions determine practice.

Further, in multiple testing one must distinguish planned from post hoc comparisons: Was the difference strongly predicted in advance of collecting the data? In Bayesian and likelihood inference, by contrast, the evidential import of the data is independent of the timing of the explanation, a point we will return to in the next chapter. In Bayesian and likelihood inference there is no distinction between planned and post hoc comparisons. Whether this is a strength or weakness of Neyman–Pearson is controversial, depending on whether it is thought the timing of data relative to explanation should be important.

The reference class of possible outcomes by which to judge an obtained statistic is also determined by the stopping rule. In testing the efficacy of a drug for reducing blood pressure, Mary tests 30 patients and performs a t-test. It is not quite significant at the 0.05 level and so she tests 10 more patients. She cannot now perform a t-test in the normal way at the 0.05 level. Jane decides to test 40 patients and collects exactly the same data. She CAN perform a t-test at the 0.05 level. If $p = 0.049$, Jane can reject the null hypothesis at the 0.05 level; Mary cannot (and, interestingly, could not no matter how much additional data she collects[8]). *Should* the same data lead to different conclusions because of the intentions of the experimenter? Whether this is a strength or weakness of Neyman–Pearson is controversial.

8. Having tested once at the 0.05 level, the testing procedure can never have an α value of less than 0.05: The probability of one or other of two events happening is never less than the probability of one of them. So even if the calculated p-value at the second test is $p = 0.0000003$, the 'actual' p-value is greater than 0.05.

Consider another often discussed type of example. Mary and Jane wish to estimate the proportion of women in a population that report having had a G spot orgasm. Mary decides in advance to sample 30 women and count the number which report G spot orgasms. She finds six which do. The best estimate of the population proportion is 6/30.

Jane decides to count women until her sixth woman reporting a G spot orgasm. That happens to be the 30th woman. In the Neyman–pearson approach, the best estimate of population proportion is now 5/29! To explain why the same data lead to different conclusions, we need to consider the respective reference classes.

For the stopping rule 'Stop after 30 observations', the reference class is the set of all samples with $n = 30$. This set defines what else the researcher might have done (e.g. counted 5/30

"Two scientists, Peter and Paul, are collecting data and happen to agree on stopping at the same time despite having different stopping rules. Peter decided to stop after 30 subjects. Paul decided we would also have stopped at 15 subjects if the result had been significant then. They have the same data, but on the Neyman Pearson approach, whether or not to reject the null depends on which stopping rule was used – that is, on what would have happened if their stopping rules had disagreed."

or 10/30 or 22/30, etc.). If we take an infinite number of samples, we might get 5/30, 6/30, 10/30 and so on. If we average all those proportions, the expected mean is the population proportion. So in this case, unbiased estimate of population proportion is 6/30.

Now consider the stopping rule: 'Stop when have six positive outcomes'. The reference class is the set of all samples with six positives. Some will be $n = 30$, some $n = 6$, some $n = 1000$ and so on. This set defines what else the researcher might have done. In infinite number of samples we might get 6/30, 6/50, 6/243 and so on. If we averaged all of these proportions, the mean would be higher than population proportion! The final observation is known to be a positive in advance, so it is unfair to count that trial; it is not free to vary but is fixed. So one should ignore the final trial. And in fact if you averaged the proportions of all the possible samples generated by a given population proportion *ignoring the final trial*, 5/29, 5/49, 5/242 and so on, the expected mean is the population mean. In this case the unbiased estimate is 5/29!

The point about what way of averaging in the reference class is unbiased is just a mathematical fact; but why should we be concerned with which other sort of events that never happened we should average over? Why should it matter what else the experimenter might have done (but did not)? Should not what actually happened be the only thing that matters? Philosophers and statisticians have not agreed what the correct answer to this question is. What if the experimenter was not sure of her intentions of when to stop? What if two experimenters are collecting data together and they have different stopping rules in mind – do we need to know who would win if there were an argument? What if the experimenter unconsciously held one stopping rule while believing she held another? Why should what is locked in someone's head affect the objective statistics we compute?

For Bayesian and likelihood inference both sets of data lead to an estimate of the population proportion of 6/30; what matters is only what the data were. In the Neyman–Pearson approach, the same data can lead to different conclusions. Whether this is a strength or weakness of Neyman–Pearson is controversial. (These issues are discussed again in the next two chapters; hopefully at the end of those chapters you will have your own opinion.)

The limits of confidence intervals are sensitive to the stopping rule and multiple testing issues as well. In the next two chapters we consider interval estimates like a confidence interval but insensitive to stopping rule and multiple testing.

Using the Neyman–Pearson approach to critically evaluate a research article

If the article uses significance or hypothesis tests, then two hypotheses need to be specified for each test. Most papers fall down at the first hurdle because the alternative is not well specified. One hypothesis is the null which often states that there will be no effect. How big an effect would just be interesting on the theory in question? The size of a just interesting effect determines the alternative hypothesis (e.g. 'The difference is bigger than five units'). Does the paper specify how big an effect would be just interesting or instead does it just indicate vaguely that there will be some difference (or some difference in a certain direction)?

Note from the introduction section of the paper whether any specific comparisons were highlighted as the main point of the experiment. These comparisons, if few in number, can be treated as planned comparisons later. If a direction is strongly predicted at this point, one-tailed tests could be considered later.

The stopping rule should be specified. Do the authors state their stopping rule? Usually they do not, but the default assumption is that the number of subjects was planned in advance

at a fixed number and significance testing took place once at the end of data collection. Was a power calculation performed to determine the appropriate number of subjects? Are the chosen levels of α and β stated in advance? If very different numbers of subjects are used in different experiments in the paper for no apparent reason, it may be a sign that multiple significance tests were conducted, as each experiment progressed and stopping occurred when the required results were obtained. (Other genuine reasons for differing subject numbers in different experiments include whether the subjects were run in large groups, for example teaching groups, and the researchers made sure they had a minimum number of subjects but if the groups were larger than that they still tested everyone.)

When families of tests were conducted (e.g. a table of t-tests addressing a common issue, a table of correlations, a set of post hoc tests following a significant overall test) is a correction to significance levels made to control family-wise error rate? No correction is needed for planned comparisons, identified by specifically being mentioned in the introduction.

Now we come to one of the most important points. Even if minimally interesting effect sizes were not stated in advance and if power were not stated in advance, a crucial point is how the authors dealt with interesting null results. *Given a null result was obtained, did the authors give some measure of sensitivity of the test?* For example, were confidence intervals provided together with an indication of what effects would be interesting based on the theory? If there were a set of results over experiments bearing on one question, perhaps all or some of them null, was a meta-analysis conducted to determine what decision is allowed by the data as a whole? (The more data that goes into an analysis, the more sensitive it is.) In the case of a meta-analysis, did the authors also give some measure of the sensitivity of null results (e.g. confidence intervals)?

In the discussion section of the paper, were only effects treated as real that had been specifically tested and found significant? Was serious consideration given to null results as null results only when the test had been shown to have adequate sensitivity? Or did the authors mistakenly treat a very non-significant p-value as in itself a measure of their high confidence in the null under the conditions of their experiment?

Summary

The Neyman–Pearson approach defines the dominant framework within which mathematical statistics is developed and the often tacit orthodox philosophy in journals of researchers who use statistics, including psychology journals. Many if not the vast majority of published papers deviate from the orthodoxy; but in the case of dispute about correct interpretation or use of statistics in standard journal articles in psychology, appeal to the Neyman–Pearson philosophy will normally trump appeal to any other. It has its own coherent logic which all researchers and consumers of research should understand. The persistent failure to understand it has led to many mistakes in the interpretation of results and in faulty research decisions generally. A large part of the confusion probably comes from people unwittingly having a Bayesian understanding of statistics and unthinkingly believing Neyman–Pearson statistics give them Bayesian answers (Oakes, 1986). The confusions are so wide spread that there have been regular calls over many decades for psychologists to give up significance testing altogether (see e.g. Harlow et al., 1997, for an overview). Indeed, Fisher regarded significance testing as the tool one uses in only the most impoverished stages of theory development. Neyman rarely used hypothesis testing as such in his own numerous statistical

applications – instead, he often gave confidence limits around estimates of model parameters. Relatedly, there have been many calls over the decades for researchers to rely more habitually on confidence intervals. For example, the 2001 version of the Publication Manual of the American Psychological Association tells us that confidence intervals constitute 'the best reporting strategy' (p. 22). Change has been slow. But you, dear reader, can help make a start.

Review and discussion questions

1. Define the key terms of: population, null and alternative hypotheses, α, β, power, significance, p-value, stopping rule, family-wise error rate, probability.
2. What is the difference between p-value and alpha?
3. When is a null result meaningful?
4. Describe two legitimate stopping rules and two illegitimate ones.
5. Why should one correct for multiple testing?
6. What would be the advantages and disadvantages of reporting confidence intervals rather than p-values in psychology papers?

Further reading

Hacking (2001) provides an excellent clear introduction to different theories of probability and how they motivate different approaches to statistical inference, including the Neyman–Pearson approach. It assumes no previous knowledge of mathematics nor statistics. The focus is on theories of probability rather than the details of the Neyman–Pearson approach.

The material in this chapter is discussed in more detail in Oakes (1986). The latter is now out of print, but anyone in research or thinking of going into research should try to get hold of a copy. The material in the book formed Oakes' PhD thesis at University College London; his examiner was so impressed that he told Oakes to publish it. While many of the ideas in the Oakes book are around generally in the literature, Oakes (1986) is one of the best single sources available. Most of it will be accessible once you have read and understood this chapter and the next two.

The following sources are also useful for theoretical discussion: Chow (1998), Cohen (1994), Gigerenzer (e.g. 2000, 2004, Chapter 13) , Howson and Urbach (1989, Chapters five and seven), Mayo (1996, Chapter nine), and Meehl (1967).

Many stats books will have information on power and confidence intervals. See, for example, Howell, D. *Fundamental Statistics for the Behavioural Sciences*, or *Statistical Methods for Psychology*. Duxbury, any of the editions.

A collection of anecdotes concerning statistics is provided by Salsburg (2002). Finally, a highly informative and engaging account of ways of thinking about data (analysed by orthodox methods) is Abelson (1995), which also should be read by anyone engaged in research.