CHAPTER 1

The Scientific Method

Learning Objectives

After studying this chapter you should be able to:

- Describe the process called the scientific method: the way scientists plan, design, and carry out research
- Define different types of logic, hypotheses, and research designs
- Know the principles of presenting data and reporting the results of scientific research

1.1 KNOWING THINGS

What can I know?

—Immanuel Kant, philosopher

The need to know things is essential to our being in the world. Without learning we die. At the very least, we must learn how to find food and keep ourselves warm. Most people, of course, are interested in more than these basics, in developing lives which could be described as fulfilling. We endeavour to learn how to develop relationships, earn a livelihood, cope with illness, write poetry (most of it pretty terrible), and make sense of our existence. At the core of these endeavours is the belief that somewhere there is the 'truth' about how things 'really' are.

Much of the seeking after truth is based on feelings and intuition. We may 'believe' that all politicians are corrupt (based on one lot of evidence), and at the same time believe that people are inherently good (based on a different lot of evidence). Underlying these beliefs is a tacit conviction that there is truth in what we believe, even though all of our observations are not consistent. There are useful expressions like: 'It is the exception that proves the rule' to help us cope with observations that do not fit neatly into our belief systems. But fundamentally, we want to be able to 'prove' that what we believe is correct (i.e. true), and we busy ourselves collecting examples that support our point of view.

Some beliefs are easier to prove than others. Arguments rage about politics and religion, mainly because

Statistics in Nutrition and Dietetics, First Edition. Michael Nelson.

Companion website: www.wiley.com/go/nelson/statistics

the evidence which is presented in favour of one position is often seen as biased and invalid by those who hold differing or opposing points of view. Some types of observations, however, are seen as more 'objective'. In science, it is these so-called objective measures, ostensibly free from bias, that are supposed to enable us to discover truths which will help us, in a systematic way, to make progress in improving our understanding of the world and how it works. This notion may be thought to apply not only to the physical and biological sciences but also the social sciences, and even disciplines such as economics. There are 'laws' which are meant to govern the ways in which things or people or economies behave or interact. These 'laws' are developed from careful observation of systems. They may even be derived from controlled experiments in which researchers try to hold constant the many factors which can vary from one setting to another, and allow only one or two factors to vary in ways which can be measured systematically.

It is clear, however, that most of the laws which are derived are soon superseded by other laws (or truths) which are meant to provide better understanding of the ways in which the world behaves. This process of old truths being supplanted by new truths is often a source of frustration to those who seek an absolute truth which is secure and immutable. It is also a source of frustration to those who believe that science provides us with objective facts, and who cannot therefore understand why one set of 'facts' is regularly replaced by another set of 'facts' which are somehow 'more true' than the last lot. It is possible, however, to view this process of continual replacement as a truth in itself: this law states that we are unlikely[1] ever to find absolute truths or wholly objective observations, but we can work to refine our understanding and observations so that they more nearly approximate the truth (the world 'as it is'). This assumes that there is in fact an underlying truth which (for reasons which we will discuss shortly) we are unable to observe directly.[2]

Karl Popper puts it this way:

> *We can learn from our mistakes. The way in which knowledge progresses, and especially our scientific knowledge, is by unjustified (and unjustifiable) anticipations, by guesses, by tentative solutions to our problems, by* conjectures. *These conjectures are controlled by criticism; that is, by attempted* refutations, *which include severely critical tests. Criticism of our conjectures is of decisive importance: by bringing out our mistakes it makes us understand the difficulties of the problems which we are trying to solve. This is how we become better acquainted with our problem, and able to propose more mature solutions: the very refutation of a theory – that is, of any serious tentative solution to our problem – is always a step forward that takes us nearer to the truth. And this is how we can learn from our mistakes.*
>
> From *'Conjectures and Refutations. The Growth of Scientific Knowledge'* [1].

This is a very compassionate view of human scientific endeavour. It recognizes that even the simplest of measurements is likely to be flawed, and that it is only as we refine and improve our ability to make measurements that we will be able to develop laws which more closely approximate the truth. It also emphasizes a notion which the atomic physicist Heisenberg formulated in his Uncertainty Principle. The Uncertainty Principle states in general terms that as we stop a process to measure it, we change its characteristics. This is allied to the other argument which

[1]As you can see, I am already beginning to hedge my bets. I am not saying that we will never find absolute truths or wholly objective observations. I am saying that it is unlikely. *How* unlikely is the basis for another discussion.

[2]My favourite description of the world is Proposition 1 from Wittgenstein's *Tractatus Logico-Philosophicus*: 'The world is everything that is the case'. The Propositions get better and better. Have a look at: http://en.wikipedia.org/wiki/Tractatus_Logico-Philosophicus, or *Wittgenstein for Beginners* by John Heaton and Judy Groves (1994), Icon Books Ltd., if you want to get more serious.

states that the observer interacts with the measurement process. Heisenberg was talking in terms of subatomic particles, but the same problem applies when measuring diet, or blood pressure, or even more subjective things like pain or well-being. Asking someone to reflect on how they feel, and the interaction between the person doing the measuring and the subject, has the potential to change the subject's behaviour and responses. This is contrary to Newton's idea that measurement, if carried out properly, could be entirely objective. It helps to explain why the discovery of the 'truth' is a process under continual refinement and not something which can be achieved 'if only we could get the measurements right'.

Consider the question: 'What do you understand if someone says that something has been proven "scientifically"?' While we might like to apply to the demonstration of scientific proof words like 'objective', 'valid', 'reliable', 'measured', 'true', and so on, the common meanings of these words are very difficult to establish in relation to scientific investigation. For all practical purposes, *it is impossible to 'prove' that something is 'true'*. This does not allow you to go out and rob a bank, for example. An inability on the part of the prosecution to establish the precise moment of the robbery, for instance, would not excuse your action in a court of law. We cope in the world by recognizing that there is a substantial amount of inexactitude, or 'error', in all that we do and consider. This does not mean we accept a shop-keeper giving us the wrong change, or worry that a train timetable gives us information only to the nearest minute when the train might arrive before (or more commonly after) the advertised time. There are lots of 'gross' measurements that are good enough for us to plan our days without having to worry about the microdetail of all that we see and do.

For example, it is very difficult to describe an individual's 'usual' intake of vitamin C, or to relate that person's intake of vitamin C to risk of stroke. On the other hand, if we can accumulate sufficient evidence from many observations to show that increasing levels of usual vitamin C intake are associated with reduced risk of stroke (allowing for measurement error in assessing vitamin C intake and the diagnosis of particular types of stroke), it helps us to understand that we can work with imprecise observations and laws which are not immutable. Moreover, it is important (for the sake of the growth of scientific knowledge) that *any belief which we hold is formulated in a statement in such a way as to make it possible to test whether or not that statement is true*. Statements which convey great certainty about the world but which cannot be tested will do nothing to improve our scientific understanding of the world. The purpose of this book, therefore, is to learn how to design studies which allow beliefs to be tested, and how to cope with the imprecision and variation inherent in all measurements when both collecting and analyzing data.

1.2 LOGIC

In science, we rely on logic to interpret our observations. Our aim is usually to draw a conclusion about the 'truth' according to how 'strong' we think our evidence is. The type of observations we choose to collect, and how we collect them, is the focus of *research design*: 'How are we going to collect the information we need to test our belief?' The decision about whether evidence is 'strong' or 'weak' is the province of *statistics*: 'Is there good evidence that our ideas are correct?' As Sherlock Holmes put it, 'It is a capital mistake to theorise before one has data'.[3]

There are two types of logic commonly applied to experience.

1.2.1 Inductive Logic

The aim with inductive logic is to infer a general law from particular instances: arguing from the particular to the general. This type of logic is good for generating new ideas about what we think *might be* true. It is less good for *testing* ideas about what we think *is* true.

[3] *Sherlock Holmes to Watson in: The Scandal in Bohemia*

Examples of Research Designs that Depend on Inductive Logic

Case studies provide a single example of what is believed to be true. The example is so compelling by itself that it is used to infer that the particular instance described may be generally true. For example:

> *A dietitian treating a severely underweight teenage girl worked with a psychotherapist and the girl's family to create an understanding of both the physiological and psychological basis and consequences of the disordered eating, resulting in a return to normal weight within six months. The approach contained unique elements not previously combined, and could be expected to have widespread benefit for similar patients.*

A case study can be interesting and provide a powerful example. But it provides very limited evidence of the general truth of the observation.

Descriptive studies bring together evidence from a number of related observations that demonstrate repeatability in the evidence. For example:

> *In four old peoples' homes, improved dining environments using baffles to reduce noise interference and allowing more time for staff to take orders and serve meals resulted in improved nutritional status among residents after one year.*

This type of cross-sectional evidence from numerous homes is better than evidence from a single home or a case study.

The generalizable conclusion, however, depends on a number of factors that might also need to be taken into account: what was the turnover among residents – did new residents have better nutritional status when they arrived, were they younger with better appetites, did they have better hearing so that they could understand more clearly what options were available on the menu for that day, etc.? One of the difficulties with descriptive studies is that we may not always be comparing like with like. We would have to collect information to demonstrate that apart from differences in noise levels and serving times, there were no other differences which could account for the change in nutritional status. We would also want to know if the circumstances in the four selected care homes were generalizable to other care homes with a similar population of residents.

Experimental studies are designed to assess the effect of a particular influence (exposure) on a particular outcome. Other variables which might affect the outcome are assumed to be held constant (or as constant as possible) during the period of evaluation.

> *Establish if a liquid iron preparation is effective in treating anaemia.*

If an influence produces consistent effects in a chosen group of subjects, we are tempted to conclude that the same influences would have similar effects in all subjects with similar characteristics. When we evaluated the results from our observations, we would try to ensure that other factors which might affect the outcome (age, sex, dietary iron intake, dietary inhibitors of iron absorption, etc.) were taken into account.

1.2.2 Deductive Logic

Deductive logic argues from the general to the particular. This type of logic involves *a priori* reasoning. This means that we think we know the outcome of our observations or experiment even before we start. What is true generally for the population[4] will be true for each individual within the population. Here is a simple example:

> *All animals die.*
> *My dog is an animal.*
> *My dog will die.*

[4]The term 'population' is defined in Chapter 2. It is not limited to the lay definition of all people living in a country. Instead, we can *define* our 'population'. In the example above, we are talking about the population of all animals (from yeast to elephants). But we could equally well define a population as all women between 35 and 54 years of age living in London, or all GP surgeries in Liverpool. More to come in Chapter 2.

This type of logic is very powerful for testing to see if our ideas are 'true'. The logic is: *if 'a' is true, then 'b' will be the outcome*. If the evidence is robust (i.e. as good a measure as we can get, given the limitations of our measuring instruments) and shows a clear relationship, it should stand up to criticism. And as we shall see, it provides the basis for the statistical inferences based on the tests described in later chapters.

There is a problem, however. The example above about my dog is relatively simple and straightforward. We can define and measure what we mean by an 'animal', and we can define and measure what we mean by 'death'. But suppose we want to understand the impact of vitamin A supplementation on risk of morbidity and blindness from measles in children aged 1 to 5 years living in areas where vitamin A deficiency is endemic. Defining and measuring variables in complex biological systems is much harder (particularly in the field of nutrition and dietetics). It becomes harder to argue that what is true generally for the population will necessarily be true for each individual within the population. This is for two reasons. First, we cannot measure all the factors that link 'a' (vitamin A deficiency) and 'b' (morbidity and blindness from measles) with perfect accuracy. Second, individuals within a population will vary from one to the next in terms of their susceptibility to infection (for a wide range of reasons) and the consequent impact of vitamin A supplementation.

For deductive logic to operate, we have to assume that the group of subjects in whom we are conducting our study is *representative* of the population in which we are interested. (The group is usually referred to as a 'sample'. Ideas about populations and samples are discussed in detail in Chapter 2.) If the group *is* representative, then we may reasonably assume that what is true in the population should be evident in the group we are studying. There are caveats to this around the size of the sample and the accuracy of our measurements, which will be covered in Chapters 2 and 12.

Examples of Research Designs that Depend on Deductive Logic

Intervention trials are designed to prove that phenomena which are true in the population are also true in a representative sample drawn from that population.

> *Compare the relative impact of two iron preparations in the treatment of anaemia.*

This may sound similar to the statement that was made under 'Experimental Studies'. The two statements are different, however. In the intervention trial, we would try to ensure that the two groups in which we were comparing treatments were similar to each other and similar to the population from which they were drawn. In the experimental study, we chose a group of subjects, measured the exposure and outcome and other characteristics of the group, and assumed that if the outcome was true in that group, it would be true in the population with similar characteristics. These differences in approach and logic are subtle but important.

In practice, the aim of most studies is to find evidence which is generalizable to the population (or a clearly defined subgroup). The relationship between the type of logic used and the generalizability of the findings is discussed below. The limitations of inductive logic and their resolution are discussed lucidly by Popper [1, pp. 54–55].

1.3 EXPERIMENTATION AND RESEARCH DESIGN

Here is a quote from 'The Design of Experiments' by Sir Ronald Fisher [2]:

> *Men*[5] *have always been capable of some mental processes of the kind we call 'learning by experience'. Doubtless this experience was*

[5]I presume he means men and women. And children. Or 'humans'. Use of the term 'men' was common to his time of writing. Don't take offence. The point he is making is important.

often a very imperfect basis, and the reasoning processes used in interpreting it were very insecure; but there must have been in these processes a sort of embryology of knowledge, by which new knowledge was gradually produced.

Experimental observations are only experience carefully planned in advance, and designed to form a secure basis of new knowledge; that is, they are systematically related to the body of knowledge already acquired, and the results are deliberately observed, and put on record accurately.

Research usually has one of two main purposes: either to describe in as accurate and reliable a way as possible what one observes, or to test an idea about what one believes to be true. To undertake research, be it quantitative or qualitative, a systematic process of investigation is needed. This involves formulating clear ideas about the nature of the problem to be investigated, designing methods for collecting information, analyzing the data in an appropriate way, and interpreting the results.

1.3.1 A Children's Story

One of my favourite children's stories is *The Phantom Tollbooth* by Norton Juster [3], in which he brilliantly summarizes the purpose of research and statistics. This may seem unlikely, but read on.

The book tells the story of Milo, a young boy living in an apartment in New York. He is endlessly bored and someone for whom everything is a waste of time. He arrives home after school one day to find a large package sitting in the middle of the living room. (I don't know where his parents are.) He unpacks and assembles a tollbooth (he lives in America, don't forget), gets in his electric car, deposits his coin, and drives through the tollbooth into a land of fanciful characters and logical challenges.

The story is this. The princesses Rhyme and Reason have been banished, and it is his job to rescue them and restore prosperity to the Kingdom of Wisdom. He drives from Dictionopolis (where only words are important) to Digitopolis (where – you guessed it – only numbers are important) to reach the Castle in the Air, where the princesses are held captive. He shares his journey with two companions: a Watchdog named Tock who is very vigilant about paying attention to *everything* (provided he keeps himself wound up); and the Humbug, 'a large beetle-like insect dressed in a lavish coat, striped trousers, checked waistcoat, spats and a derby hat', whose favourite word is BALDERDASH – the great sceptic.

On the way to Digitopolis, the road divides into three, with an enormous sign pointing in all three directions stating clearly:

DIGITOPOLIS
5 miles
1 600 rods
8 800 Yards
26 400 ft
316 800 in
633 600 half inches

They argue about which road to take. The Humbug thinks miles are shorter, Milo thinks half-inches are quicker, and Tock is convinced that whichever road they take it will make a difference. Suddenly, from behind the sign appears an odd creature, the Dodecahedron, with a different face for each emotion for, as he says, 'here in Digitopolis everything is quite precise'. Milo asks the Dodecahedron if he can help them decide which road to take, and the Dodecahedron promptly sets them a hideous problem, the type that makes maths pupils have nightmares and makes grown men weep:

If a small car carrying three people at thirty miles an hour for ten minutes along a road five miles long at 11.35 in the morning starts at the same time as three people who have been travelling in a little automobile at twenty miles an hour for fifteen minutes on another

road and exactly twice as long as one half the distance of the other, while a dog, a bug, and a boy travel an equal distance in the same time or the same distance in an equal time along a third road in mid-October, then which one arrives first and which is the best way to go?

They each struggle to solve the problem.

'I'm not very good at problems',
admitted Milo.

'What a shame', sighed the Dodecahedron. 'They're so very useful. Why, did you know that if a beaver two feet long with a tail a foot and half long can build a dam twelve feet high and six feet wide in two days, all you would need to build the Kariba Dam is a beaver sixty-eight feet long with a fifty-one foot tail?'

'Where would you find a beaver as big as that?' grumbled the Humbug as his pencil snapped.

'I'm sure I don't know', he replied, 'but if you did, you'd certainly know what to do with him'.

'That's absurd', objected Milo, whose head was spinning from all the numbers and questions.

'That may be true', he acknowledged, 'but it's completely accurate, and as long as the answer is right, who cares if the question is wrong? If you want sense, you'll have to make it yourself'.

'All three roads arrive at the same place at the same time', interrupted Tock, who had patiently been doing the first problem.

'Correct!' shouted the Dodecahedron. 'Now you can see how important problems are. If you hadn't done this one properly, you might have gone the wrong way'.

'But if all the roads arrive at the same place at the same time, then aren't they all the right way?' asked Milo.

'Certainly not', he shouted, glaring from his most upset face. 'They're all the wrong way. Just because you have a choice, it doesn't mean that any of them has to be right'.

That is research design and statistics in a nutshell. Let me elaborate.

1.4 THE BASICS OF RESEARCH DESIGN

According to the Dodecahedron, the basic elements of research are as shown in Box 1.1:

He may be a little confused, but trust me, all the elements are there.

1.4.1 Developing the Hypothesis

The Dodecahedron: 'As long as the answer is right, who cares if the question is wrong?'

The Dodecahedron has clearly lost the plot here. Formulating the question correctly is the key starting point. If the question is wrong, no amount of experimentation or measuring will provide you with an answer.

The purpose of most research is to try and provide evidence in support of a general statement of what one believes to be true. The first step in this process is to establish a *hypothesis*. A hypothesis is a clear statement of what one believes to be true. The way in which the hypothesis is stated will also have an impact on which measurements are needed. The formulation of a clear hypothesis is the critical first step in the development of research. Even if we can't make measurements that reflect the truth, the hypothesis should always be a statement of what

BOX 1.1

The four key elements of research

Hypothesis

Design **Statistics**

Interpretation

you believe to be true. Coping with the difference between what the hypothesis says is true and what we can measure is at the heart of research design and statistics.

TIP

Your first attempts at formulating hypotheses may not be very good. Always discuss your ideas with fellow students or researchers, or your tutor, or your friendly neighbourhood statistician. Then be prepared to make changes until your hypothesis is a clear statement of what you believe to be true. It takes practice – and don't think you should be able to do it on your own, or get it right first time. The best research is collaborative, and developing a clear hypothesis is a group activity.

We can test a hypothesis using both inductive and deductive logic. Inductive logic says that if we can demonstrate that something is true in a particular individual or group, we might argue that it is true generally in the population from which the individual or group was drawn. The evidence will always be relatively weak, however, and the truth of the hypothesis hard to test. Because we started with the individual or group, rather than the population, we are less certain that the person or group that we studied is representative of the population with similar characteristics. Generalizability remains an issue.

Deductive logic requires us to draw a sample from a defined population. It argues that if the sample in which we carry out our measurements can be shown to be representative of the population, then we can generalize our findings from our sample to the population as a whole. This is a much more powerful model for testing hypotheses.

As we shall see, these distinctions become important when we consider the generalizability of our findings and how we go about testing our hypothesis.

1.4.2 Developing the 'Null' Hypothesis

In thinking about how to establish the 'truth'[6] of a hypothesis, Ronald Fisher considered a series of statements:

> **No amount of experimentation can 'prove' an *inexact* hypothesis.**

The first task is to get the question right! Formulating a hypothesis takes time. It needs to be a clear, concise statement of what we believe to be true,[7] with no ambiguity. If our aim is to evaluate the effect of a new diet on reducing cholesterol levels in serum, we need to say specifically that the new diet will 'lower' cholesterol, not simply that it will 'affect' or 'change' it. If we are comparing growth in two groups of children living in different circumstances, we need to say in which group we think growth will be better, not simply that it will be 'different' between the two groups.

The hypothesis that we formulate will determine what we choose to measure. If we take the time to discuss the formulation of our hypothesis with colleagues, we are more likely to develop a robust hypothesis and to choose the appropriate measurements. Failure to get the hypothesis right may result in the wrong measurements being taken, in which case all your efforts will be wasted. For example, if the hypothesis relates to the effect of diet on serum cholesterol, there may be a particular

[6]You will see that I keep putting the word 'truth' in single quotes. This is because although we want to test whether or not our hypothesis is true – it is, after all, a statement of what we *believe* to be true – we will never be able to collect measures that are wholly accurate. Hence, the truth is illusory, not an absolute. This is what the single quotes are intended to convey.

[7]'The term "belief" is taken to cover our critical acceptance of scientific theories – a *tentative* acceptance combined with an eagerness to revise the theory if we succeed in designing a test which it cannot pass' [1, p. 51]. It is important to get used to the idea that any 'truth' which we hope to observe is likely to be superseded by a more convincing 'truth' based on a more robust experiment or set of observations using better measuring instruments, and which takes into account some important details which we did not observe the first time.

cholesterol fraction that is altered. If this is stated clearly in the hypothesis, then we must measure the relevant cholesterol fraction in order to provide appropriate evidence to test the hypothesis.

No finite amount of experimentation can 'prove' an exact hypothesis.

Suppose that we carry out a series of four studies with different samples, and we find that in each case our hypothesis is 'proven' (our findings are consistent with our beliefs). But what do we do if in a fifth study we get a different result which does not support the hypothesis? Do we ignore the unusual finding? Do we say, 'It is the exception that proves the rule?' Do we abandon the hypothesis? What would we have done if the first study which was carried out appeared not to support our hypothesis? Would we have abandoned the hypothesis, when all the subsequent studies would have suggested that it was true?

There are no simple answers to these questions. We *can* conclude that any system that we use to evaluate a hypothesis must take into account the possibility that there may be times when our hypothesis *appears* to be false when in fact it is true (and conversely, that it may appear to be true when in fact it is false). These potentially contradictory results may arise because of sampling variations (every sample drawn from the population will be different from the next, and because of sampling variation, not every set of observations will necessarily support a true hypothesis), and because our measurements can never be 100% accurate.

A finite amount of experimentation can *disprove* an exact hypothesis.

It is easier to disprove something than prove it. If we can devise a hypothesis which is the *negation* of what we believe to be true (rather than its opposite), and then *disprove* it, we could reasonably conclude that our hypothesis was true (that what we observe, for the moment, seems to be consistent with what we believe).

This negation of the hypothesis is called the 'null' hypothesis. The ability to *refute* the null hypothesis lies at the heart of our ability to develop knowledge. A good null hypothesis, therefore, is one which can be tested and refuted. If I can refute (disprove) my null hypothesis, then I will accept my hypothesis.

> *A theory which is not refutable by any conceivable event is non-scientific. Irrefutability is not a virtue of a theory (as people often think) but a vice. [1, p. 36]*

Let us take an example. Suppose we want to know whether giving a mixture of three anti-oxidant vitamins (β-carotene, vitamin C, and vitamin E) will improve walking distance in patients with Peripheral Artery Disease (PAD), an atherosclerotic disease of the lower limbs. The hypothesis (which we denote by the symbol H_1) would be:

H_1: Giving anti-oxidant vitamins A, C, and E as a dietary supplement will improve walking distance in patients with PAD.[8]

The null hypothesis (denoted by the symbol H_0) would be:

H_0: Giving anti-oxidant vitamins A, C, and E as a dietary supplement will *not* improve walking distance in patients with PAD.

H_0 is the *negation* of H_1, suggesting that the vitamin supplements will make no difference. It is not the opposite, which would state that giving supplements *reduces* walking distance.

It is easier, in statistical terms, to set about disproving H_0. If we can show that H_0 is probably *not* true, there is a reasonable chance that our hypothesis *is* true. Box 1.2 summarizes the necessary steps. The statistical basis for taking this apparently convoluted approach will become apparent in Chapter 5.

[8]This is the hypothesis – short and sweet. The study protocol will specify which subjects we choose, the dosage of the supplement, how often it is administered over what period, how walking distance is measured in a standardized way, what we mean by 'improve,' whether we measure blood parameters to confirm compliance with the protocol, etc.

BOX 1.2

Testing the hypothesis

1. Formulate the Hypothesis (H_1)
2. Formulate the Null Hypothesis (H_0)
3. Try to disprove the Null Hypothesis

1.4.3 Hypothesis Generating Versus Hypothesis Testing

Some studies are observational rather than experimental in nature. The purpose of these studies is often to help in the generation of hypotheses by looking in the data for relationships between different subgroups and between variables. Once the relationships have been described, it may be necessary to set up a new study which is designed to test a specific hypothesis and to establish causal relationships between the variables relating to exposure and outcome. For example, Ancel Keys observed that there was a strong association between the average amount of saturated fat consumed in a country and the rate of death from coronary heart disease: the more saturated fat consumed, the higher the death rate. Numerous studies were carried out subsequently to test the hypothesis that saturated fat causes heart disease. Some repeated Ancel Keys' original design comparing values between countries, but with better use of the available data, including more countries. Other studies compared changes in saturated fat consumption over time with changes in coronary heart disease mortality. Yet other studies looked at the relationship between saturated fat consumption and risk of heart disease in individuals. Not all of the studies came to the same conclusions or supported the hypothesis. It took some time to understand why that was the case.

1.4.4 Design

The Dodecahedron: 'If you hadn't done this one properly, you might have gone the wrong way'.

When designing an experiment, you should do it in such a way that *allows* the null hypothesis to be disproved. The key is to **introduce** and **protect** a random element in the design.

Consider some research options for the study to test whether anti-oxidant vitamins supplements improve walking distance in peripheral arterial disease (PAD). In the sequence below, each of the designs has a weakness, which can be improved upon by introducing and protecting further elements of randomization.

a. Choose the first 100 patients with PAD coming into the clinic, give them the treatment, observe the outcome.
Patients may naturally improve with time, without any intervention at all. Alternatively, there may be a **placebo** *effect (patients show improvement simply as a result of having taken part in the study because they believe they are taking something that is beneficial and alter their behaviour accordingly), even if the treatment itself is ineffective.*

This is a weak observational study. Introduce a control group which receives a placebo.

b. Allocate the first 50 patients to the treatment group, the next 50 to the placebo group.
If the two groups differ by age, sex, or disease severity, this could account for apparent differences in improvement between the groups.

This is a weak experimental study. Introduce matching.

c. *Match patients in pairs for* age, sex, disease severity; assign the first patient in each pair to receive treatment, the second patient to receive a placebo.
The person assigning patients may have a subconscious preference for putting one patient first in each pair. Does the patient know which treatment they are getting?

This is a weak placebo-controlled intervention trial. Introduce randomization and blinding.

d. Allocate patients to treatment or placebo randomly within pairs. Make sure that the

researcher does not know which patient is to receive which treatment (the researcher is then said to be 'blind' to the allocation of treatment). Make sure that the patient does not know which treatment they are receiving (keep the patient 'blind' as well). This makes the study 'double blind'.

e. Conduct a placebo-controlled randomized double-blind intervention trial.

'Randomization properly carried out is the key to success.' (Sir Ronald Fisher)

Intervention studies of this type are often regarded as the most robust for testing a hypothesis. But sometimes randomized controlled trials are not ethical, especially if the exposure may be harmful (e.g. smoking, or increasing someone's saturated fatty acid intake), or it is not possible to blind either the subjects or the researchers because the treatment being given is so obviously different from the placebo. In these cases, it is possible to mimic intervention studies in samples that are measured at baseline and then followed up over months or years. These types of studies raise issues about how to deal with factors that cannot be controlled for but which might affect the outcome (e.g. in a study looking at the impact of diet on risk of cardiovascular disease, the influence of smoking and diabetes) and changes in the general environment (e.g. indoor smoking bans) that have the potential to affect all the subjects in the study. Complex statistical analysis can be used to cope with some of these design issues. The ability to test the hypothesis in a robust way remains.

1.4.5 Statistics

The Dodecahedron: 'Just because you have a choice, it doesn't mean that any of them has to be right'.

Statistical tests enable you to analyze data in order to decide whether or not it is sensible to accept your hypothesis. There are literally thousands of values that can be calculated from hundreds of tests, but unless you know which test to choose, the values that you calculate may not be appropriate or meaningful. One of the main aims of this book is to help you learn to choose the test which is right for the given research problem. Once you have decided which test is appropriate for your data, the calculation is a straightforward manipulation of numbers. It is vitally important, however, to learn which data to use, how the manipulation is carried out, and how it relates to the theoretical basis which will enable you to make the decision about the truth of your hypothesis.

Most of the time it is better to use a computer to do the computation for you. Even if you enter the values correctly and generate a meaningful outcome, the computer will not tell you if your hypothesis is true. For that, you need to know how to interpret the results of the tests.

1.4.6 Interpretation

The Dodecahedron: 'If you want sense, you'll have to make it yourself'.

Every statistical test will produce a number (the *test statistic)* which you then need to interpret. This is the last stage and often the most difficult part of statistical analysis. The final emphasis in every chapter that deals with statistical tests will be on how to interpret the test statistic. We will also look at the SPSS output to verify that the right set of values has been entered for statistical analysis.

Two concepts deserve mention here: 'Inference' and 'Acceptance'. 'Inference' implies greater or lesser strength of fact. It is usually expressed as a probability of a given result being observed. If there is a high probability that the result which you have observed is associated with the hypothesis being true, we talk about 'strong' evidence. If the observed outcome is little different from what we would expect to see if the null hypothesis were true, we talk about 'weak' or 'no' evidence.

At some point, we need to make a decision about whether to accept or reject the null hypothesis, that is, to make a statement about whether or not we

believe that the hypothesis is true. 'Acceptance' implies a cut-off point upon which action will be taken. We will discuss cut-off points in Chapter 5. It is important not to confuse political expediency (acceptance) with scientific validity (inference).

1.5 NEXT STEPS

Every year, at least one student shows up at my door, holds out an open notebook with a page full of numbers, and says, 'I've collected all this data[9] and now I don't know what to do with it'. I strongly resist the temptation to tell them to go away, or to ask why they didn't come to see me months ago. I usher them in and see what we can salvage. Usually, it is a debacle. The data collected are not suitable for testing the hypothesis; their sample is poorly defined; they don't have enough of the right types of observations; they have used different methods for collecting data at baseline and follow-up; the list goes on and on.

Box 1.3 summarizes the steps that *should* be undertaken when conducting research. Although Steps 1 and 2 are essential ('Getting the question right'), probably the most important step is Step 3, the point at which you design the research project. It is vital at this stage that you consult a statistician (as well as others who have done similar research). Be prepared to accept that your hypothesis may need modifying, and that the design that you first thought of is not perfect and would benefit from improvements. It is very unlikely that you will have got it right at your first attempt. *Be prepared to listen and to learn from your mistakes.* As I said in the Introduction to this book, statisticians may be perceived as monstrous, inhuman creatures intent only on humiliating those who come to consult them. In reality, the statistician is there to advise you concerning the likelihood of being able to prove your hypothesis, guide you in the design of the study, the choice of measurements which you intend to make, and the type of analyses you plan to undertake. Months or years of effort can be wasted if you embark on a study which is flawed in its design. Do not take the chance! Be brave! Be thick-skinned! Talk with statisticians and accept their advice. Even get a second opinion if you feel very uncertain about the advice you are given.

[9]The word 'data' is plural, by the way – she should have said 'I have collected all *these* data' – but we will come to that later.

1.6 RESEARCH DESIGN

There is a wide variety of research designs which can be used to address the many research questions that you are likely to ask. There is no strictly right or wrong answer concerning which design to use. You should recognize, however, that some designs are stronger when it comes to arguing the truth of your hypothesis. The aim in carrying out any research will always be to obtain the maximum information from a given design in relation to a particular research question, given the time and financial resources that are available.

1.6.1 Project Aims

Coming up with an interesting and useful research question will always involve reading the relevant literature (both books and journals) to explore how other people have tackled similar problems, and discussing with colleagues how best to investigate the problem at hand. Once you have done that, you can think about what it is you want to achieve in your research.

Projects have different aims. An undergraduate student project with 15 subjects carried out part-time over two months may not have much chance of establishing new findings that are statistically significant, but it will introduce the student to hypothesis formulation, designing a study, writing a research protocol, sampling, subject recruitment, data entry, computer analysis, and writing up the results. On the other hand, an MSc project carried out full-time over four months will require the same skills as the undergraduate project, but will usually involve a more detailed

BOX 1.3

Steps in undertaking research

- **Step 1**. Make observations about the world. Science doesn't happen in a vacuum.
- **Step 2**. Construct a Hypothesis. State clearly the aims and objectives of your study.

Formulate the Null Hypothesis.

- **Step 3**. Design the experiment.

This is the stage at which you should seek the advice of a statistician

regarding the hypothesis, sample selection, sample size, choice of measurements, and the type of analyses and statistical tests to be used. Failure to consult properly at this stage may mean that any work that you do may be a waste of time. Do not take that chance!

- **Step 4**. Conduct the research.
- **Step 5**. Analyze the data both observationally (do the numbers make sense?) and statistically.
- **Step 6**. Interpret the results (draw inferences) and write your report (for marking or for publication). Work that is not marked or published may just as well never have been completed.
- **Step 7**.Bask in the glory of a job well done.

consideration of design, sample size, and study power (see Chapter 12). It will also provide an opportunity to write a detailed report and to make a presentation of the findings (both for assessment), usually to an audience of postgraduate peers and their tutors. More demanding undergraduate projects may include some or all of these additional elements. For a PhD, or for funded research, all of these elements will be present, plus the requirement to write paper(s) for submission to a peer-reviewed journal and to present findings to a public audience at scientific meetings. As a professor of mine once said, 'If you haven't published the paper, you haven't done the work'.

1.6.2 Demonstrating Causality

The underlying purpose of most research is to find evidence in support of causality of the type: 'If A, then B'. Of course, we may just be interested in describing what is going on in physiological systems (what dietary factors are associated with low serum total cholesterol levels?) or in the population (are women aged 75 years and older at greater risk of osteoporosis-related fracture of the hip if they have low levels of physical activity?) More often, we want to know if there is a causal relationship between these factors (does an increased level of physical activity protect against osteoporosis-related hip fracture in women aged 75 and

BOX 1.4

Bradford Hill hierarchy of causality

• Strength of association	Is the evidence linking exposure and outcome strong? We shall see what we mean by 'strong' as we explore the different statistical tests used to evaluate associations.
• Consistency of association across studies	Are the same associations seen repeatedly in different groups or across different populations in different places and times?
• Specificity	Is there a specific link between exposure and outcome?
• Temporal association	Does A precede B? Evidence needs to show that cause (A) is followed by consequence (B). As we shall see, A and B may be associated in a cross-sectional analysis of data, but unless a clear time-sequence can be established, the evidence for causality is weak.
• Dose-response	Does increased exposure result in increased likelihood of the outcome? If fruit and vegetable consumption is protective against heart disease, can it be shown that the more fruit and vegetables are eaten, the lower the risk of disease?
• Plausible mechanism and coherence	Is there a clear physiological explanation for the observed link between A and B? What is it in fruit and vegetables that affect the factors that determine risk of heart disease? Does the new evidence fit in with what is already known? If not, why not? Are there any animal models that support evidence in humans?
• Experimental evidence	Does experimental evidence based on intervention studies support the argument for causation? Is the experimental evidence consistent across studies?
• Analogy	Are there related exposures or conditions that offer insight into the observed association?

older?). Public health recommendations to improve nutrition and nutrition-related outcomes need strong evidence of causality before they can be promoted to the general public. Confusion in the mind of the public is often caused by the media promoting a 'miracle cure' based on a single study (it makes good press but bad science). Food manufactures are often guilty of using weak evidence of causality or vague terms about 'healthiness' to promote sales of their products.[10]

We have seen earlier that the logic used to support notions of causality may be inductive or deductive. Whichever logical model is used, no single study in nutrition will provide conclusive evidence of the relationship between A and B. There is a hierarchy of evidence, first set out clearly by Bradford Hill [4, 5], which suggests that a clear picture of causality can only be built from multiple pieces of evidence (Box 1.4). Published over 50 years ago, these criteria have withstood the test of time [6].

1.6.3 Types of Study Design

The summary below provides a brief overview of some of the types of study designs available. There are many more designs, of course, that address complex issues of multiple factors influencing multiple outcomes, with corresponding statistical analysis, but these are dealt with in more advanced textbooks on research design and analysis. The list below repeats some of the material covered in Section 1.2 on logic, but goes into more detail in relation to study design.

The principle aim is to conduct studies that are free from bias and based on relevant measures of exposure and outcome so that the hypothesis can be tested effectively.

[10]The UK rules governing labelling and packaging can be found here: https://www.gov.uk/food-labelling-and-packaging/nutrition-health-claims-and-supplement-labelling. Two things are immediately apparent: there are lots of exceptions; and the EU has a lot to say.

Observational Studies

Observational studies usually focus on the characteristics or distribution of phenomena in the population that you are investigating. Such studies may analyze data at one point in time or explore time trends in the relevant variables. They may be based on observations of individuals within a sample, or they may consider the relationship between variables observed in *groups* of subjects (for example, differences in diet and disease rate between countries). They are often the basis for *hypothesis generating*, rather than *hypothesis testing*.

Case studies are reports of potentially generalizable or particularly interesting phenomena. Individually, a case study cannot provide evidence that will help you to establish the truth of your hypothesis. Consistent findings across several case studies may provide support for an idea, but cannot be used in themselves to test a hypothesis.

Descriptive studies are careful analyses of the distribution of phenomena within or between groups, or a study of relationships existing between two or more variables within a sample. Descriptive studies are often well suited to qualitative examination of a problem (e.g. examining the coping strategies used by families on low income to ensure an adequate diet for their children when other demands [like the gas bill] are competing for limited cash). But of course they also provide descriptions of quantitative observations for single variables (e.g. how much money is spent on food, fuel, etc. in families on low income), or multiple variables (e.g. how money is spent on food in relation to total income or family size). Many epidemiological studies fall into this category (see below). They are useful for understanding the possible links between phenomena, but cannot in themselves demonstrate cause and effect.

Diagnostic studies establishing the extent of variation in disease states. They are helpful when selecting subjects for a study and deciding on which endpoints may be relevant when designing a study to explore cause and effect.

Experimental and Intervention Studies

These studies are designed to create differences in exposure to a factor which is believed to influence a particular outcome, for example, the effect of consuming oat bran on serum cholesterol levels, or the effect on birth weight of introducing an energy supplement during pregnancy. The aim is usually to analyze the differences in outcome associated with the variations in exposure which have been introduced, holding constant other factors which could also affect the outcome.

These types of studies are usually prospective or longitudinal in design. Alternatively, they may make use of existing data. Depending on how subjects are selected, they may use inductive or deductive logic to draw their conclusions (see Section 1.2)

Pre-test–post-test (Figure 1.1). This is the simplest (and weakest) of the prospective experimental designs. There is one sample. It may be 'adventitious' – subjects are selected as they become available (for example, a series of patients coming into a diabetic clinic, or a series of customers using a particular food shop); or it may be 'purposive' (the sample is drawn systematically from a population using techniques that support generalization of the findings). Each individual is measured at the start of the study (the 'baseline' or 'time zero'). There is then an intervention. Each subject is measured again at the end of the study.

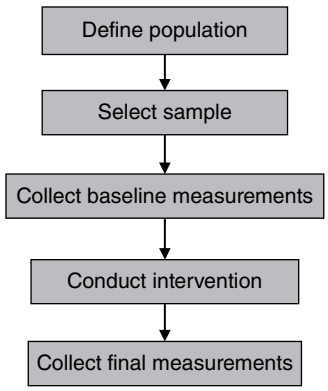

FIGURE 1.1 Pre-test–post-test design.

The weakness of this design is that there may have been a number of factors influencing the outcome of the study, but you may be aware of only some of them. An example is a study on the influence on growth of food supplements given to children. Was there bias in the way the children were recruited into the study? Did the supplement contain enough of the relevant energy or nutrients to have an impact on growth? Were the children at the right age for the supplement to have an impact? Was follow-up long enough to observe the impact of the supplement on growth? Participation in the study may in itself benefit growth (more social interaction, more stimulation from adults and peers, more food sharing, changing dietary patterns at home as a result of parental involvement in the study, etc.). It is possible that the direct effect of the supplement may represent only one of several factors that can have in impact on growth.

One sample (Figure 1.2). Here, two or more treatments are compared in one group of subjects. Each subject acts as their own control. There may be a placebo being compared with an active treatment.

The danger here is that the order of administration may have an effect on the outcome. It is therefore desirable to randomize the order of administration of the treatment and placebo (making the design more like a cross-over clinical trial – see below). Thus, for one subject, Intervention 1 will be the active treatment, and Intervention 2 the placebo; for another subject, Intervention 1 will be the placebo, and Intervention 2 the active treatment. A 'wash out' period may be required, so that any residual effect of the first treatment has time to disappear and not appear as an influence on the second treatment. This may not be possible where a psychological or attitudinal variable is being measured, because the views of a subject may be permanently influenced once they have started to participate in the study.

Two or more samples, unmatched (Figure 1.3). Similar experimental groups are given different treatments. One group acts as a control (or placebo) group.

This is a stronger experimental design than the pre-test–post-test or one-sample designs. However, the groups may differ in some characteristic which is important to the outcome (for example they may differ in age). Even if the groups are not 'matched' exactly, some attempt should be made to ensure that the groups are similar regarding variables which may be related to the outcome. For example, in a study to see if the administration of oral zinc supplements improves taste sensitivity in cancer patients, the type, degree, and severity of disease would need to be similar in both the treatment and placebo groups.

Two or more samples, matched (Figure 1.4). The participants in the study groups are matched in pairs on a subject-by-subject basis for variables such as age and gender (so-called 'confounding' variables). One group is then the control (or placebo) group, and the other group is given the active treatment(s).

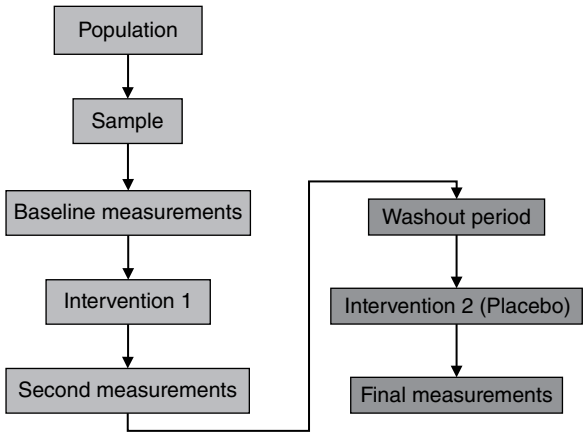

FIGURE 1.2 One-sample design.

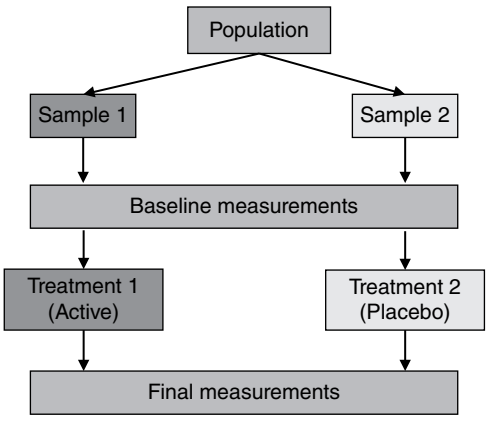

FIGURE 1.3 Two-sample (parallel) design.

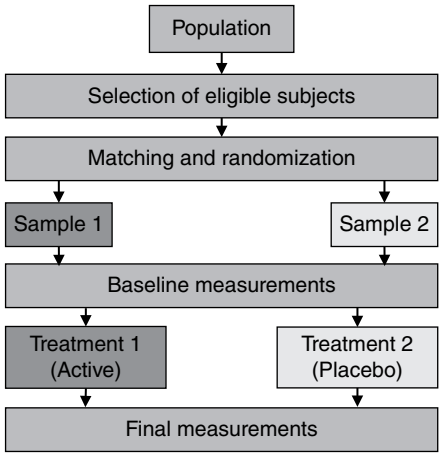

FIGURE 1.4 Two-sample (matched) design.

For example, if the aim was to test the effect of a nutrition education programme to persuade urban mothers to breast feed their infants (versus no education programme), it would be important to match on age of mother and parity (number of births). Matching could be carried out by first grouping mothers according to the number of children. Then, within each parity group, mothers could be ranked according to age. Starting with the two youngest mothers in the group with one child, one mother would be *randomly* assigned to the treatment group (the education programme), and the other would be assigned to the control group (no education programme).[11] The next two mothers would be assigned randomly to treatment or control, and so on, by parity group and age, until all the subjects have been assigned. In this way, the possible effects of age and parity would be controlled for. If there were three groups (Treatment A, Treatment B, and a control group), mothers would be matched in *triplets* according to parity and age and randomly assigned to Treatment A, Treatment B, or the control group. Using this technique, all the groups should have very similar age and parity structures, and it could be argued that age and parity therefore have a similar influence in each group. Of course, other factors such as maternal education, income, or social class might also influence outcome. The difficulty with including too many factors as matching variables is that it becomes increasingly difficult to find adequate matches for everyone in the study. Use the two or three factors that you think will be the most powerful influence on the outcome as the matching variables. Then measure all the other factors that you think might be associated with the outcome so that they can be taken into account in the final analyses (see Chapters 10 and 11).

Clinical trials. These involve the assessment of the effects of clinical interventions such as drugs or feeding programmes. They are usually carried out in a controlled setting (that is, where the subject will be unable to obtain other supplies of the drug or where all aspects of diet are controlled). The intervention is compared with a placebo.

The design and analysis of clinical trials is a science in itself [7], and there are a great many variations in design which can be adopted. The so-called Rolls Royce of clinical trials, the randomized double-blind placebo-controlled cross-over clinical trial (Figure 1.5), requires careful thought in its planning, implementation, and analysis.

Randomized controlled trials should not be embarked upon lightly! It is very demanding of time, staff, and money. Moreover, there are important limitations.

In recent years, it has been recognized that while clinical trials may be appropriate for providing proof of causality in some circumstances (e.g. understanding the impact of a new drug on disease outcomes, or a specific nutritional intervention), there may not always be equivalent, controlled circumstances that exist in the real world (e.g. promotion of five-a-day consumption in a sample versus a control group). The generalizability of findings may therefore be limited when it comes to saying whether or not a particular intervention is likely to be of benefit to individuals or

[11]Randomization is usually carried out by computer. You could do it equally well by simply tossing a coin, but if the coin is not tossed in exactly the same way for every pair, the randomization may be biased by the way in which the coin was tossed from one matched pair to the next. This introduces additional 'noise' in the system and is easily avoided by using the computer to generate the random numbers needed.

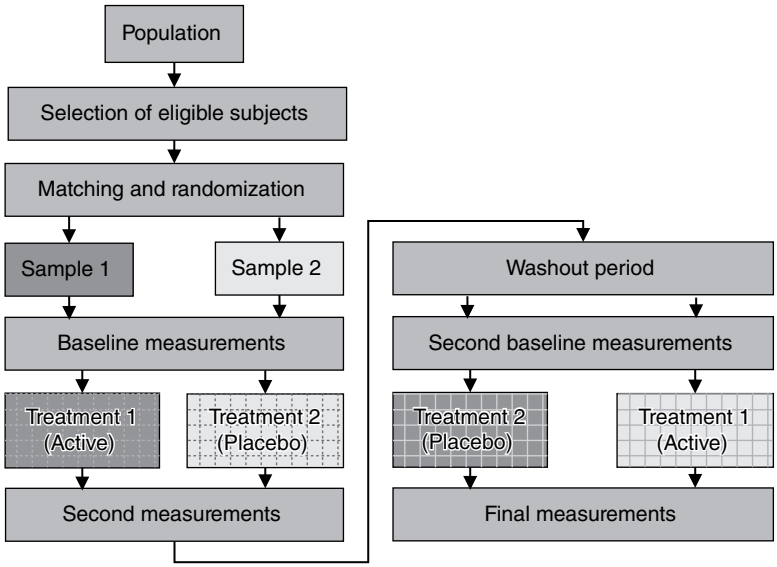

FIGURE 1.5 Randomized placebo-controlled cross-over trial.

the population as a whole. To address these circumstances, alternate designs and analytical approaches have been developed in the last decade or more that aim to take complex, real-world circumstances into account [8]. The existing guidance is due to be updated in 2019.

1.6.4 Epidemiological Studies

Epidemiological studies examine relationships between exposures and health-related outcomes in populations. In the context of nutritional epidemiology, exposures might include individual diet, community health intervention programmes, supplementation, food advertising, dietary advice, or other nutrition-related variables. Outcomes can include changes in nutrition-related blood biochemistry (e.g. cholesterol levels, haemoglobin), clinical outcomes (e.g. xerophthalmia, obesity), or morbidity or mortality statistics relating to nutrition.

Epidemiological studies fall into three categories.

Descriptive Studies

Descriptive studies in epidemiology include ecological studies, cross-sectional studies, and time trend analysis. They are useful for generating hypotheses. Measurements can be made in individuals at a given point in time (cross-sectional studies) or accumulated over time in groups of people (ecological studies). They are used to relate measures of exposure and outcome in groups of people that share common characteristics (e.g. vegetarians versus omnivores) or to compare regions or countries. For example, they might compare diet and disease patterns between countries (are heart disease rates lower in countries where people eat lots of oily fish?) or between subgroups (do vegetarians have lower risk of heart disease compared to non-vegetarians?).

There are two main problems with this type of study. First, there may be other factors that could explain an observed association or changes in the population over time. For example, populations with higher oily fish consumption may be more active or less obese. Second, not everyone in the population or subgroup is exposed at the same level: some individuals in the population may eat lots of oily fish, while others may eat very little. Are the people with low oily fish consumption the ones that have higher rates of heart disease?

Analytical Studies

These include cohort and case-control studies. Their primary characteristic is that they relate exposures in individuals (factors that are likely to influence the occurrence of disease or mortality within the population) to outcomes (disease or mortality rates). Analytical studies are usually based on observations relating to large numbers of people in the population (hundreds or thousands). They provide much stronger evidence of diet–disease relationships than descriptive studies. In terms of the Bradford Hill model of causality (Box 1.4), they provide evidence of temporal association and dose-response. If blood or urine samples are collected, they may also provide evidence of a plausible physiological mechanism for causation. Detailed descriptions of these designs and their statistical analysis are given in epidemiological texts such as Rothman [9] or Margetts and Nelson [10].

Cohort studies are prospective in nature; they look *forward* in time, following a group of people who are free from the disease of interest at baseline. Relevant exposures are measured at baseline or time zero. Over time (usually years), the appearance of disease (morbidity), or relevant changes in blood or urine biochemistry, or nutrition-related mortality, is monitored (Figure 1.6).

The risk of developing disease (or related outcomes) in the exposed group is compared with the risk of developing disease in the unexposed group. This is known as the Relative Risk.

In nutritional epidemiology, the meanings of 'Exposed' and 'Unexposed' are different from the meanings in infectious disease or occupational epidemiology. In infectious disease epidemiology, for example, a subject either is or is not exposed to a particular bacterium that can cause illness. In occupational epidemiology, a subject may be classified either as exposed or not exposed to a potential risk factor (e.g. asbestos dust). In nutritional epidemiology, in contrast, variables are usually continuous – no one has a diet that does not contain energy or iron, for example. Subjects are therefore classified into bands of exposure. 'Exposed' might mean someone in the top half of the distribution of intake, and 'Unexposed' might mean someone in the bottom half. Of course, degrees of exposure may also be ascertained in other spheres of epidemiology, but in nutritional epidemiology it is the norm and forms the basis for most analysis of risk.

In cohort studies, the time sequence of cause and effect is not in question. There may be other factors, however, which over time, explain the apparent associations between the measures at baseline and the outcome measures collected years later (e.g. some individuals may gain more weight than others, or their socio-economic status at follow-up may not be the same as it was at baseline). The strength of cohort studies is that if these factors are measured at baseline and again at follow-up, they can be taken into account in the analysis. The main disadvantage of cohort studies is the length of time it takes to accumulate relevant outcome measures in

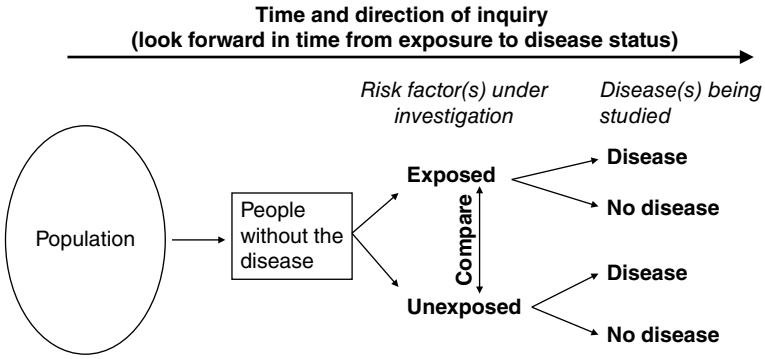

FIGURE 1.6 Cohort study design.

sufficient numbers of individuals, and the cost of collecting the measurements in hundreds or thousands of individuals over long periods of time.

Case-control studies, in contrast, start by identifying a group of subjects with a given disease or condition. Controls are subjects who do not have the disease or outcome being investigated. They are matched individually with the cases for age and sex, and often for other variables such as smoking or BMI or income or occupation. The study then looks *backward* in time to measure exposures that are potentially causal (Figure 1.7).

Measures of exposure in the past are usually determined by asking cases and controls about their habits in the past using a questionnaire (although sometimes there are records of exposure, for example the type of work someone did, or hospital records of birth weight). The past exposures of the cases are then compared with the past exposures of the controls. The relevant statistical outcome is the Odds Ratio: what are the chances of being exposed and in the disease group compared with being exposed and in the non-disease group. This is the best estimate of the Relative Risk, which cannot be measured directly in case-control studies [9, 10].

Case-control studies are much cheaper to carry out than cohort studies, but accurate measurements of past exposure are often difficult to collect. For example, asking someone what their diet was like 10 or 15 years ago is likely to be heavily influenced by the type of diet they consume now.

Experimental Studies

In epidemiology, experimental studies take two forms: clinical trials and community trials. The aim is to compare the impact of the active treatment on the relevant outcomes with the impact of a placebo.

Clinical trials in epidemiology usually take the form of randomized controlled trials. The aim is to see if an intervention to alter exposure results, over time, in a change in outcomes. In epidemiology, clinical trials can involve thousands of subjects followed over many years, whereas studies relating to drug treatments or food interventions relating to changes in blood or urine biochemistry may involve only tens of subjects. The purpose in having such large studies in epidemiology is to be able to generalize to the population with confidence, and to take into account the many confounders (see below) that may operate in the real world.

Researchers usually strive to achieve 'double blind' status in their study designs, but in nutritional interventions this may not always be possible. If the intervention involves changing diet, for example, or providing nutritional advice to increase fruit and vegetable consumption, the subject may be aware of the changes and is therefore no longer 'blind'. Similarly, the researcher involved in administering the changes may be aware of which subjects are in the treatment group and which are in the placebo group. If this is the case, it is important to ensure that the person undertaking the statistical analysis is blinded

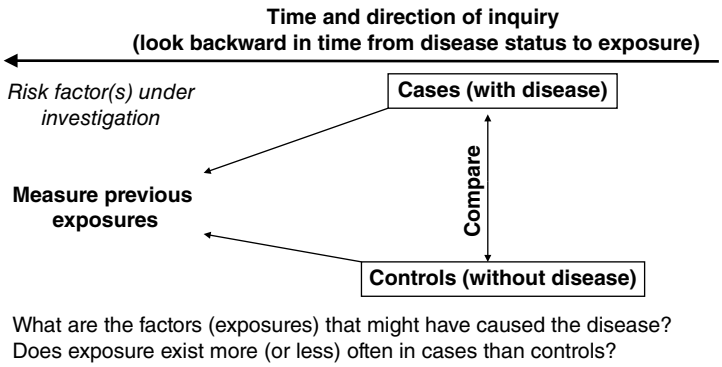

FIGURE 1.7 Case-control study design.

to the identity of the groups being compared. This can be done through the coding of results for computer analysis so that the comparison is simply between group A and group B. Only at the end of the analysis is the identity of the group revealed. Even here, it may not always be possible to blind the analyst. In that case, the procedures for carrying out the statistical analyses should be described in advance so as to avoid a 'fishing expedition', hunting for statistically significant findings.

Community trials are intervention studies carried out in whole communities to see if some change in exposure is associated with a subsequent change in disease or mortality rates. Again, the study will involve a treatment community and a placebo community. The communities are matched (e.g. age and sex structure of the population, percentage of the population not in work). There may be more than one community in each group.

Community trials are pragmatic in nature. The aim is to see if community-based interventions have sufficient penetration and impact to bring about changes in nutrition-related outcomes (e.g. the percentage of adults over the age of 45 who are overweight or obese). The identity of the individuals in the community who make the desired changes and their individual outcomes may not be known. A questionnaire can be used to find out the proportion of the population who were aware of the intervention (for example, advice on healthy eating provided in a GP surgery by the community dietitian), and height and weight data could be taken from the routine 'healthy man' or 'healthy woman' screening programmes being run in the same groups of GP surgeries.

Community trials are much cheaper to run than clinical trials. They do not, however, provide the same wealth of detail about individual behaviours and outcomes and, as a consequence, provide less insight into causal mechanisms.

Confounding and Bias

A key feature of epidemiological studies is the attention paid to *confounding factors* and *bias*. Confounding factors are associated with both the exposure and the outcome. Suppose that we were comparing a group of cases who had recently had their first heart attack with a group of controls who had never had a heart attack. Say that we were interested in knowing if higher oily fish consumption was associated with lower rates of heart attack. We could measure oily fish consumption using a food frequency questionnaire that asked about usual diet. Suppose we found that it was lower among the cases. Is this good evidence that higher oily fish consumption is protective against having a heart attack? What if the cases, on average, were 10 years older than the controls, and that younger people tended to have higher levels of oily fish in their diet. This could explain the apparent association of higher oily fish consumption with decreased risk of heart attack. In this case, *age* would be referred to as a confounding factor. Confounding factors need to be associated with both the exposure and the outcome that we are interested in. We could match our cases with controls of the same age. Alternatively, we could use a statistical approach that took age into account in the analysis. The most common confounding factors – things like age, gender, social class, and education – need to be controlled for when comparing one group with another. Other factors such as smoking, disease status (e.g. diabetes), or body mass index (BMI) may also be taken into account, but these may be explanatory factors or factors in the causal pathway rather than true confounders.

Bias is a problem associated with measuring instruments or interviewers. *Systematic bias* occurs when everyone is measured with an instrument that always gives an answer that is too high or too low (like an inaccurate weighing machine). Bias can be *constant* (every measurement is inaccurate by the same amount) and or *proportional* (the size of the error is proportional to the size of the measurement, e.g. the more you weigh the greater the inaccuracy in the measurement). Bias is a factor that can affect any study and should be carefully controlled.

Some types of bias may simply reduce our ability to detect associations between exposure and outcome. This is 'noise in the system'. It means that

there may be an association between exposure and outcome, but our data are too 'noisy' for us to be able to detect it. For example, we know that there is day-to-day and week-to-week variation in food and drink consumption. We need to try and collect sufficient information to be able to classify subjects according to their 'usual' consumption.

Other types of bias mean that the information that we obtain is influenced by the respondent's ability to give us accurate information. Subjects who are overweight or obese, for example, or who have higher levels of dietary restraint, tend to under-report their overall food consumption, especially things like confectionery or foods perceived as 'fatty'. Subjects who are more health-conscious may over-report their fruit and vegetable consumption because they regard these foods as 'healthy' and want to make a good impression on the interviewer. In these instances, making comparisons between groups becomes problematic because the amount of bias is related to the type of individual which may in turn be related to their disease risk.

Dealing with issues such as confounding, residual confounding, factors in the causal pathway, and different types of bias are fully addressed in epidemiological textbooks [9, 10].

1.7 DATA, RESULTS, AND PRESENTATION

First of all, a few definitions are needed:

- Statistic – a numerical observation
- Statistics – numerical facts systematically collected (also the science of the analysis of data)
- Data – what you collect (the word 'data' is plural – the singular is 'datum' – so we say 'the data are...' not 'the data is...')
- Results – a summary of your data

1.7.1 Data Are What You Collect, Results Are What You Report

No one else is as interested in your data as you are. You must love your data, look after them carefully (think of the cuddly statistician), and cherish each observation. You must make sure that every observation collected is accurate, and that when the data are entered into a spreadsheet, they do not contain any errors. When you have entered all your data, you need to 'clean' your data, making sure that there are no rogue values, and that the mean and the distribution of values is roughly what you were expecting. Trapping the errors at this stage is essential. There is nothing worse than spending days or weeks undertaking detailed statistical analysis and preparing tables and figures for a report, only to discover that there are errors in your data set, meaning that you have to go back and do everything all over again.

TIP

Allow adequate time in your project to clean the data properly. This means

- Check for values that are outside the range of permitted values.
- Look at the distributions of variables and check for extreme values. Some extreme values may be genuine. Others may be a result of 'fat finger' syndrome (like typing an extra zero and ending up with 100 rather than 10 as a data point).
- Understand how to use 'missing values' in SPSS. These help you to identify gaps in the data and how to handle them (for example, the difference between 'I don't know', Not Applicable, or missing measurement).

If you find an unusual observation, check it with your supervisor or research colleagues. They may want to inspect your data to see that your observations are correct. Don't try and hide an unusual observation (or worse still, ignore it, or leave it out of the data set without telling anyone). Always be frank and open about letting others inspect your data, *especially* if you or they think there may be something wrong. We all make mistakes. It is no great shame if there are some errors in the data that we missed and that someone else helpfully spots for us. Be thick-skinned about this. The real embarrassment comes

if we do lots of analysis and the errors in the data only come to light when we make a presentation of our results.

1.7.2 Never Present Endless Detailed Tables Containing Raw Data

It is your job as a scientist to summarize data in a coherent form (in tables, graphs, and figures), tell an interesting story about the relationships between the variables you have measured, and interpret the results intelligently for your reader, using appropriate statistical analyses.

Of course, you need to keep accurate records of observations, and make sure that your data set (spreadsheet) is stored securely and that you have backup copies of everything. Bulking up a report with tables of raw data is bad practice, however. No one will read them.

Chapter 15 provides lots of examples about how to summarize data to make the presentation of results interesting. It also shows how to present results according to the type of audience you are talking to. If I am presenting new results about the impact of school food on attainment to scientific colleagues, I will include lots of information about the methods that I used to identify my samples, make observations, and analyze the data, as well as details about the findings themselves. My scientific colleagues will need enough information to be confident that my data are unbiased, that I have used the right analytical approaches, and that the results are statistically significant. This is the same basic approach that I will take when I am writing a paper for submission to a peer-reviewed journal. In contrast, if I am presenting results on the same topic to a group of teachers, I will use lots of graphs and charts to summarize the results in a way that tells a good story. The underlying data will be the same, of course. But the teachers are likely to be bored by too much detail about methods – they probably just want to know the headline about whether better school food has a positive impact on attainment, and how big that impact is likely to be.

1.7.3 Significant Digits and Rounding

It always surprises me, when teaching undergraduate and postgraduate students, that they often don't know how to round numbers properly. So when asked to present a result to two decimal places, for example, either they provide a string of numbers after the decimal place (far more than two) in the mistaken hope that somehow that is 'better' or 'more accurate', statistically speaking. Alternatively, it becomes evident that the concept of 'rounding' is not familiar to them, and their answers vary like leaves blowing in the wind.

The underlying principle is that when undertaking calculations, it is useful to retain as many digits as possible during the course of the calculation. This is what happens when you use your calculator, Excel, or SPSS [11]. This will produce the most mathematically precise result. However, when the final value is presented, it should contain no more significant digits than the numbers from which it is derived.

For example, calculate the average height of a group of five students, each of whom had been measured to the nearest centimetre (Table 1.1).

The average (the arithmetic mean) of these five values is 164.4 cm. However, presenting the result to the nearest millimetre (tenth of a centimetre) would be misleading, as it would imply that your starting observations had a similar level of precision. You should, instead, round the value for the result to the number of significant digits with which you started. In this case, round 164.4 cm (four significant digits) to 164 cm (three significant digits, to the nearest whole centimetre). This is the value that you should report for your result, as it reflects the level of precision of your original observations.[12]

TABLE 1.1 Height of Five Students (cm)

163
152
176
166
165

[12] On the other hand, keep all the significant digits when doing your calculations. We will see how this works in later chapters.

TABLE 1.2 Rules for Rounding

Rule	Original Value	Rounded Value
If the final digit is less than 5, round to the value of the preceding digit	164.4	164
If the final digit is greater than 5, round to the value which is one higher than the preceding digit	164.6	165
If the final digit is equal to 5, and the preceding digit is odd, round up to the next even number	163.5	164
If the final digit is equal to 5, and the preceding digit is even, round down to the preceding number	164.5	164

The standard conventions for rounding are as shown in Table 1.2.

These conventions may differ from those which you have been taught, but they are the conventions followed by all calculators and all computers when undertaking computations.

Some people have been taught always to round up if the last digit is equal to 5, but the calculations in Table 1.3 illustrate the error which is introduced if that rule is followed.

Although this error seems small, it can be substantial if we are dealing with small numbers. For example, if the original values had been 3.5, 4.5, 5.5, and 6.5, and we were rounding to whole numbers, the average would be equal to 5 for the original values and correctly rounded values (4, 4, 6, 6), but the average for the incorrectly rounded values would be 5.5 – an error of 10%!

Curiously, Excel and some statistical packages (e.g. Minitab) display numbers which are *always rounded up* when the last digit is 5. However, underlying calculations are based on the correct rules for rounding shown in Table 1.2. To ensure that your calculations are correct, always follow these rules.

Finally, be careful when reporting values from Excel or Minitab – even though the calculations will be right, the final displayed value may not accord with the rules above.

TIP

Use common sense to report final values according to the correct rules of rounding – don't just report what the computer tells you. This principle – using your common sense to interpret computer output – will come up again and again throughout this book.

TABLE 1.3 Impact of Rounding Styles on Mean Value

Original Value	Correctly Rounded to Nearest Even Digit	Incorrectly Rounded (Always Upward)
163.5	164	164
164.5	164	165
165.5	166	166
166.5	166	167
Average = 165	Average = 165	Average = 165.5 (rounded = 166)

1.8 READING

The concepts which underlie research design and statistics set out in this book may be very different from any that you have had to think about. Persistence will bring rewards, however, even if the topics seem difficult at first.

Very often, one author's treatment of a topic will seem totally incomprehensible, while another's will seem crystal clear. Below, therefore, is a list of books to which you might like to refer – try at least two on the same topic if you are having difficulty grasping a particular idea. Don't worry about the publication dates – the approaches to basic statistics have not changed for the last few decades. The volumes cited use examples from medicine rather than nutrition and dietetics, but the statistical principles remain the same.

The books marked with an '*' are particularly highly recommended. Note that a number of these

TABLE 1.4 Impact of Screening for Breast Cancer on 5-year Mortality in 62 000 Women

		Cause of Death			
		Breast Cancer		All Other Causes	
	n	*n*	Rate/1 000	*n*	Rate/1 000
Screening group					
Examined	20 200	23	1.1	428	21
Refused	10 800	16	1.5	409	38
Total	31 000	39	1.3	837	27
Control group	31 000	63	2.0	879	28

texts are now available as e-books, usually at a substantially reduced price compared with the hard copy.

* Armitage P, Berry G, Mathews JNS. *Statistical Methods in Medical Research*. 4th edition. Blackwell Science. Oxford. 2001. Very thorough and wide-ranging, with excellent explanations of the mathematics behind the statistics. Highly recommended.

* Bland M. *An Introduction to Medical Statistics*. 4th edition. Oxford University Press. 2015. Good, basic introduction, wide ranging, with lots of useful examples including SPSS. Might be a bit difficult for beginners, as he tends to jump straight in without too much underlying explanation.

* Bowling Ann. *Research Methods in Health*. 4th edition. Open University Press. Buckingham. 2014. This text provides an excellent overview of issues in both quantitative and qualitative research. Together with Ann Bowling's other text ('Measuring Health', 4th edition, McGraw Hill Education, 2017), it provides a practical and straightforward guide to research design for the health sciences.

Campbell MJ, Machin D, Walters SJ. *Medical Statistics: A Textbook for the Health Sciences (Medical Statistics)*. 4th edition. Wiley-Blackwell. Chichester. 2007. A comprehensive text, the book has lots of 'common sense' tips and useful comments sprinkled liberally throughout.

* Campbell MJ. *Statistics at Square One*. 11th edition. BMJ Books. 2009. Useful little 'cookbook', good for quick reference to remind you of the underlying statistical formulae, but not very good about explaining the principles underlying the tests or why they work as they do.

Campbell MJ. *Statistics at Square Two*. 2nd edition. BMJ Books. 2006. This is a useful follow-on to Statistics at Square One. It explores more complex statistical analyses involving multiple variables and complex study designs.

Freedman D, Pisani R, Purves R. Statistics. 4th edition. Norton. London. 2014. Basic and very readable text. May at times seem long-winded, but lots of detailed examples and useful exercises (with answers).

Glantz SA. *Primer of Biostatistics*. 7th edition. McGraw-Hill. London. 2012. First rate text, very clear descriptions of tests with medical examples and problems for solving.

Corder GW, Forman DI. *Nonparametric Statistics: A Step-by-Step Approach*. 2nd edition. Wiley. 2014. Very good at explaining the fundamentals of nonparametric statistics.

* Juster, Norton. *The Phantom Tollbooth*. Harper Collins. New York. 2008. Illustrations by Jules Feiffer. A wonderful allegory of scientific thinking concerning two kidnapped Princesses: Rhyme and Reason. The illustrations by Jules Feiffer perfectly complement Milo's struggle to make sense of the world. The ideal fantasy for when you're fed up with statistics.

Mead R, Curnow RN and Hasted A (editor). *Statistical Methods in Agriculture and Experimental Biology*. 3rd edition. Chapman and Hall CRC Press. London. 2002. The authors worked in the Applied Statistics Department of the Food Research Institute at Reading, so the examples are geared towards food science.

Moser CA and Kalton G. *Survey Methods in Social Investigation*. 2nd edition. Heinemann Educational. London. 1985. Detailed and practical

advice on survey techniques and design. Nonmathematical. A classic.

* Norman GR and Streiner DL. *Biostatistics: the Bare Essentials*. 4th edition. People's Medical Publishing House. USA. 2014. Clear, funny, and irreverent. Goes into the lower reaches of the upper echelons of statistics (i.e. covers the basics plus some of the more advanced stuff).

* Riegelman R. *Studying a Study and Testing a Test: How to Read the Medical Evidence*. 5th edition. Lippincott Williams and Wilkins. 2004. Good for interpreting the medical literature.

1.9 EXERCISES

Answers to these exercises can be found in Part 4, at the end of the chapters.

1.9.1 Rounding and significant digits

a. Round the following numbers to two decimal places. Use the rules plus common sense.

12.2345
144.5673
73.665
13.6652
99.4545

b. Round the same numbers to three significant digits.

1.9.2 Interpreting data: does screening save lives?

62 000 women in a health insurance scheme were randomly allocated to either a screening programme for breast cancer, or no screening. After 5 years, the following results were observed (Table 1.4):

a. Does screening save lives? Which numbers tell you so?
b. Among the women in the screening group, the death rate from all other causes in the 'Refused' group was almost twice that in the 'Examined' group. Did screening cut the death rate in half? Explain briefly.
c. Was the study blind?

1.9.3 Hypothesis and null hypothesis

In a study designed to assess whether undernutrition is a cause of short stature in poor inner-city children:

a. State the hypothesis (H_1)
b. State the null hypothesis (H_0)
c. Consider:
   i. confounding factors
   ii. sources of systematic bias
c. Consider ways to minimize the effects of confounding and systematic bias

REFERENCES

1. Popper Karl R. *Conjectures and Refutations. The Growth of Scientific Knowledge*. 5th edition. Routledge. London. 2002.
2. Fisher RA. *Design of Experiments*. 7th edition. Oliver and Boyd. London. 1960.
3. Juster Norton. *The Phantom Tollbooth*. Random House. New York. 2000.
4. Bradford Hill A. The environment and disease: association or causation? *Proc R Soc Med.* 1965 May; 58(5): 295–300.
5. Bradford Hill A. *A Short Textbook of Medical Statistics*. 11th edition. Hodder and Stoughton. London. 1985.
6. Fedak KM, Bernal A, Capshaw ZA, Gross S. Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerg Themes Epidemiol.* 2015; 12: 14. doi: https://doi.org/10.1186/s12982-015-0037-4 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4589117/
7. Shein-Chung Chow, Jen-Pei Liu. *Design and Analysis of Clinical Trials: Concepts and Methodologies*. 3rd edition. Wiley.London. ISBN: 978-0-470-88765-3. 892. January 2014

8. Medical Research Council. *Developing and Evaluating Complex Interventions: New Guidance*. Medical Research Council. London. 2008. https://mrc.ukri.org/documents/pdf/developing-and-evaluating-complex-interventions/
9. Rothman K. *Epidemiology: An Introduction*. 2nd edition. Oxford University Press. 2012.
10. Margetts B and Nelson M. *Design Concepts in Nutritional Epidemiology*. 2nd edition. Oxford University Press. 1997.
11. IBM SPSS Statistics. Version 24. © Copyright IBM Corporation and others, 1989, 2016.