Section 1
Chapter 1

# Basic concepts

# Why data never speak for themselves

> Science teaches us to doubt, and in ignorance, to refrain.
>
> Claude Bernard (Silverman, 1998; p. 1)

The beginning of wisdom is to recognize our own ignorance. We mental health clinicians need to start by acknowledging that we are ignorant; we do not know what to do; if we did, we would not need to read anything, much less this book – we could then just treat our patients with the infallible knowledge that we already possess. Although there are dogmatists (and many of them) of this variety – who think that they can be good mental health professionals by simply applying the truths of, say, Freud (or Prozac) to all – this book is addressed to those who know that they do not know, or who at least want to know more.

When faced with persons with mental illnesses, we clinicians need to first determine what their problems are, and then what kinds of treatments to give them. In both cases, in particular the matter of treatment, we need to turn somewhere for guidance: how should we treat patients?

We no longer live in the era of Galen: pointing to the opinions of a wise man is insufficient (though many still do this). Many have accepted that we should turn to science; some kind of empirical research should guide us.

If we accept this view – that science is our guide – then the first question is how are we to understand science?

## Science is not simple

This book would be unnecessary if science was simple. I would like to disabuse the reader of any simple notion of science, specifically "positivism": the view that science consists of positive facts, piled on each other one after another, each of which represents an absolute truth, or an independent reality, our business being simply to discover those truths or realities.

This is simply not the case. Science is much more complex.

For the past century scientists and philosophers have debated this matter, and it comes down to this: facts cannot be separated from theories; science involves deduction, and not just induction. In this way, no facts are observed without a preceding hypothesis. Sometimes, the hypothesis is not even fully formulated or even conscious; I may have a number of assumptions that direct me to look at certain facts. It is in this sense that philosophers say that facts are "theory-laden"; between fact and theory no sharp line can be drawn.

## How statistics came to be

A broad outline of how statistics came to be is as follows (Salsburg, 2001): Statistics were developed in the eighteenth century because scientists and mathematicians began to recognize the inherent role of uncertainty in all scientific work. In physics and astronomy, for

instance, Pierre Laplace realized that certain error was inherent in all calculations. Instead of ignoring the error, he chose to quantify it, and the field of statistics was born. He even showed that there was a mathematical distribution to the likelihood of errors observed in given experiments. Statistical notions were first explicitly applied to human beings by the nineteenth-century Belgian Lambert Adolphe Quetelet, who applied it to the normal population, and the nineteenth-century French physician Pierre Louis, who applied it to sick persons. In the late nineteenth-century, Francis Galton, a founder of genetics and a mathematical leader, applied it to human psychology (studies of intelligence) and worked out the probabilistic nature of statistical inference more fully. His student, Karl Pearson, then took Laplace one step further and showed that not only is there a probability to the likelihood of error, but even our own measurements are probabilities: "Looking at the data accumulated in biology, Pearson conceived the measurements themselves, rather than errors in the measurement, as having a probability distribution." (Salsburg, 2001; p. 16.) Pearson called our observed measurements "parameters" (Greek for "almost measurements"), and he developed staple notions like the mean and standard deviation. Pearson's revolutionary work laid the basis for modern statistics. But if he was the Marx of statistics (he actually was a socialist), the Lenin of statistics would be the early twentieth-century geneticist Ronald Fisher, who introduced randomization and p-values, followed by A. Bradford Hill in the mid twentieth-century, who applied these concepts to medical illnesses and founded clinical epidemiology. (The reader will see some of these names repeatedly in the rest of this book; the ideas of these thinkers form the basis of understanding statistics.)

It was Fisher who first coined the term "statistic" (Louis had called it the "numerical method"), by which he meant the observed measurements in an experiment, seen as a reflection of all possible measurements. It is "a number that is derived from the observed measurements and that estimates a parameter of the distribution." (Salsburg, 2001; p. 89.) He saw the observed measurement as a random number among the possible measurements that could have been made, and thus "since a statistic is random, it makes no sense to talk about how accurate a single value of it is . . . What is needed is a criterion that depends on the probability distribution of the statistic . . ." (Salsburg, 2001; p. 66). How probably valid is the observed measurement, asked Fisher? Statistical tests are all about establishing these probabilities, and statistical concepts are about how we can use mathematical probability to know whether our observations are more or less likely to be correct.

## A scientific revolution

This process was really a revolution; it was a major change in our thinking about science. Prior to these developments, even the most enlightened thinkers (such as the French Encyclopedists of the eighteenth century, and Auguste Comte in the nineteenth century) saw science as the process of developing absolutely certain knowledge through refinements of sense-observation. Statistics rests on the concept that scientific knowledge, derived from observation using our five senses aided by technologies, is not absolute. Hence, "the basic idea behind the statistical revolution is that the real things of science are distributions of number, which can then be described by parameters. It is mathematically convenient to embed that concept into probability theory and deal with probability distributions." (Salsburg, 2001; pp. 307–8.)

It is thus not an option to avoid statistics, if one cares about science. And if one understands science correctly, not as a matter of absolute positive knowledge but as a much

more complex probabilistic endeavor (see Chapter 11), then statistics are part and parcel of science.

Some doctors hate statistics; but they claim to support science. They cannot have it both ways.

## A benefit to humankind

Statistics thus developed outside of medicine, in other sciences in which researchers realized that uncertainty and error were in the nature of science. Once the wish for absolute truth was jettisoned, statistics would become an essential aspect of all science. And if physics involves uncertainty, how much more uncertainty is there in medicine? Human beings are much more uncertain than atoms and electrons.

The practical results of statistics in medicine are undeniable. If nothing else had been achieved but two things – in the nineteenth century, the end of bleeding, purging, and leeching as a result of Louis' studies (Louis, 1835); and in the twentieth century the proof of cigarette smoking related lung cancer as a result of Hill's studies (Hill, 1971) – we would have to admit that medical statistics have delivered humanity from two powerful scourges.

## Numbers do not stand alone

The history of science shows us that scientific knowledge is not absolute, and that all science involves uncertainty. These truths lead us to a need for statistics. Thus, in learning about statistics, the reader should not expect pure facts; the result of statistical analyses is not unadorned and irrefutable fact; all statistics is an act of interpretation, and the result of statistics is more interpretation. This is, in reality, the nature of all science: it is all interpretation of facts, not simply facts by themselves.

This statistical reality – the fact that data do not speak for themselves and that therefore positivistic reliance on facts is wrong – is called *confounding bias*. As discussed in Chapter 2, observation is fallible: we sometimes think we see what is *not* in fact there. This is especially the case in research on human beings. Consider: caffeine causes cancer; numerous studies have shown this; the observation has been made over and over again: among those with cancer, coffee use is high compared to those without cancer. Those are the unadorned facts – and they are wrong. Why? Because coffee drinkers also smoke cigarettes more than non-coffee drinkers. Cigarettes are a confounding factor in this observation, and our lives are chock full of such confounding factors. Meaning: we cannot believe our eyes. Observation is not enough for science; one must try to observe *accurately*, by removing confounding factors. How? In two ways: 1. Experiment, by which we control all other factors in the environment except one, thus knowing that any changes are due to the impact of that one factor. This can be done with animals in a laboratory, but human beings cannot be controlled in this way (ethically). Enter the randomized clinical trial (RCT). These are how we experiment with humans to be able to observe accurately. 2. Statistics: certain methods (such as regression modeling, see Chapter 6) have been devised to mathematically correct for the impact of measured confounding factors.

We thus need statistics, either through the design of RCTs or through special analyses, so that we can make our observations accurate, and so that we can correctly (and not spuriously) accept or reject our hypotheses.

Science is about hypotheses and hypothesis-testing, about confirmation and refutation, about confounding bias and experiment, about RCTs and statistical analysis: in a word, it is

not just about facts. Facts always need to be interpreted. And that is the job of statistics: not to tell us the truth, but to help us get closer to the truth by understanding how to interpret the facts.

## Knowing less, doing more

That is the goal of this book. If you are a researcher, perhaps this book will explain why you do some of the things you do in your analyses and studies, and how you might improve them. If you are a clinician, hopefully it will put you in a place where you can begin to make independent judgments about studies, and not simply be at the mercy of the interpretations of others. It may help you realize that the facts are much more complex than they seem; you may end up "knowing" less than you do now, in the sense that you will realize that much that passes for knowledge is only one among other interpretations, but at the same time I hope this statistical wisdom proves liberating: you will be less at the mercy of numbers and more in charge of knowing how to interpret numbers. You will know less, but at the same time, what you do know will be more valid and more solid, and thus you will become a better clinician: applying accurate knowledge rather than speculation, and being more clearly aware of where the region of our knowledge ends and where the realm of our ignorance begins.

# Chapter 2 Why you cannot believe your eyes: the Three C's

> Believe nothing you hear, and only one half that you see.
>
> Edgar Allan Poe (Poe, 1845)

A core concept in this book is that the validity of any study involves the sequential assessment of Confounding bias, followed by Chance, followed by Causation (what has been called the Three C's) (Abramson and Abramson, 2001).

Any study needs to pass these three hurdles before you should consider accepting its results. Once we accept that no fact or study result is accepted at face value (because no facts can be observed purely, but rather all are interpreted), then we can turn to statistics to see what kinds of methods we should use to analyze those facts. These three steps are widely accepted and form the core of statistics and epidemiology.

## The first C: bias (confounding)

The first step is bias, by which we mean *systematic* error (as opposed to the random error of chance). Systematic error means that one makes the same mistake over and over again because of some inherent problem with the observations being made. There are subtypes of bias (selection, confounding, measurement), and they are all important, but I will emphasize here what is perhaps the most common and insufficiently appreciated kind of bias: confounding. Confounding has to do with factors, of which we are unaware, that influence our observed results. The concept is best visualized in Figure 2.1.

### Hormone replacement therapy

As seen in Figure 2.1, the confounding factor is associated with the exposure (or what we think is the cause) and leads to the result. The *real* cause is the confounding factor; the *apparent* cause, which we observe, is just along for the ride. The example of caffeine, cigarettes, and cancer was given in Chapter 1. Another key example is the case of hormone replacement therapy (HRT). For decades, with much observational experience and large observational studies, most physicians were convinced that HRT had beneficial medical effects in women, especially postmenopausally. Those women who used HRT did better than those who did not use HRT. When finally put to the test in a huge randomized clinical trial (RCT), HRT was found to lead to actually worse cardiovascular and cancer outcomes than placebo. Why had the observational results been wrong? Because of confounding bias: those women who had used HRT also had better diets and exercised more than women who did not use HRT. Diet and exercise were the confounding factors: they led to better medical outcomes directly, and they were associated with HRT. When the RCT equalized all women who received HRT versus placebo on diet and exercise (as well as all other factors), the direct effect of HRT could

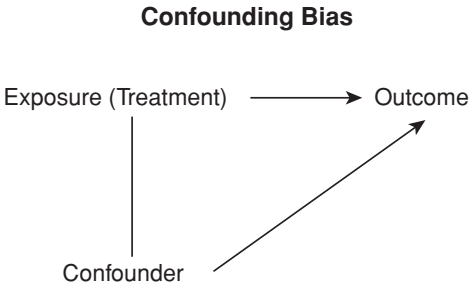


**Figure 2.1** Confounding bias.

finally be observed accurately; and it was harmful to boot (Prentice *et al.*, 2006). (This example is discussed more in Chapter 9.)

## The eternal triangle

As one author puts it: "Confounding is the epidemiologist's eternal triangle. Any time a risk factor, patient characteristic, or intervention appears to be causing a disease, side effect, or outcome, the relationship needs to be challenged. Are we seeing cause and effect, or is a confounding factor exerting its unappreciated influence? . . . Confounding factors are always lurking, ready to cast doubt on the interpretation of studies." (Gehlbach, 2006; pp. 227–8.)

This is the lesson of confounding bias: *we cannot believe our eyes.* Or perhaps more accurately, we cannot be sure when our observations are right, and when they are wrong. Sometimes they are one way or the other, but, more often than not, observation is wrong rather than right due to the high prevalence of confounding factors in the world of medical care.

The kind of confounding bias that led to the HRT debacle had to do with intrinsic characteristics of the population. The doctors had nothing to do with the patients' diets and exercise; the patients themselves controlled those factors. It could turn out that completely independent features, such as hair color or age or gender, are confounding factors in any particular study. These are not controlled by patients or doctors; they are just there in the population and they can affect the results. Two other types of confounding factors exist which are the result of the behavior of patients and doctors: confounding by indication, and measurement bias.

## Confounding by indication

The major confounding factor that results from the behavior of doctors is *confounding by indication* (also called selection bias). This is a classic and extremely poorly appreciated source of confusion in medical research:

As a clinician, you are trained to be a non-randomized treater. What this means is that you are taught, through years of supervision and more years of clinical experience, to tailor your treatment decisions to each individual patient. You do not treat patients randomly. You do not say to patient A, take drug X; and to patient B, take drug Y; and to patient C, take drug X; and to patient D, take drug Y – you do not do this without thinking any further about the matter, about why each patient should receive the one drug and not the other. You do not practice randomly; if you did, you should be appropriately sued. However, by practicing non-randomly, you automatically bias all your experience. You think your patients are doing well

because of your treatments, whereas they should be doing well because you are tailoring your treatments *to those who would do well with them*. In other words, it often is not the treatment effects that you are observing, but the treatment effects in specially chosen populations. If you then generalize from those specific patients to the wider population of patients, you will be mistaken.

## Measurement bias: blinding

I have focused on the first C as confounding bias. The larger topic here is bias, or systematic error, and besides confounding bias, there is one other major source of bias: measurement bias (sometimes also called information bias). Here the issue is not that the outcomes are due to unanalyzed confounding factors, but rather that the outcomes themselves may be inaccurate. The way the outcomes are measured, or the information on which the outcomes are based, is false. Often this can be related to the impact of either the patients' wishes or the doctors' beliefs; thus double-blinding is the usual means of handling measurement bias.

Randomization is the best means of addressing confounding bias, and blinding the means for measurement bias. While blinding is important, it is not as important as randomization. Confounding bias is much more prominent and multivaried than measurement bias. Clinicians often focus on blinding as the means of handling bias; this only addresses the minor part of bias. Unless randomization occurs, or regression modeling or other statistical analyses are conducted, the problem of confounding bias will render study results invalid.

## The second C: chance

If a study is randomized and blinded successfully, or if observational data are appropriately analyzed with regression or other methods, and there still seems to be a relationship between a treatment and an outcome, we can then turn to the question of chance. We can then say that this relationship does not seem to be systematically erroneous due to some hidden bias in our observations; now the question is whether it just happened by chance, whether it represents *random* error.

I will discuss the nature of the hypothesis-testing approach in statistics in more detail in Chapter 8; suffice it to say here that the convention is that a relationship is viewed as being unlikely erroneous due to chance if, using mathematical equations designed to measure chance occurrence of associations, it is likely to have occurred 5% of the time, or less frequently, due to chance. This is the famous p-value, which I will discuss more in Chapter 7.

The application of those mathematical equations is a simple matter, and thus the assessment of chance is not complex at all. It is much simpler than assessing bias, but it is correspondingly less important. Usually, it is no big deal to assess chance; bias is the tough part. Yet again many clinicians equate statistics with p-values and assessing chance. This is one of the least important parts of statistics.

Often what happens is that the first C is ignored, bias is insufficiently examined, and the second C is exaggerated: not just 1, or 2, but 20 or 50 p-values are thrust upon the reader in the course of an article. The p-value is abused until it becomes useless, or, worse, misleading (see Chapter 7).

The problem with chance, usually, is that we focus too much on it, and we misinterpret our statistics. The problem with bias, usually, is we focus too little on it, and we don't even bother with statistics to assess it.

## The third C: causation

Should a study pass the first two hurdles, bias and chance, it still should not be seen as valid unless we assess it in terms of causation. This is an even more complex topic, and a part of statistics where clinicians cannot simply look for a number or a p-value to give them an answer. We actually have to use our minds here, and think in terms of ideas, and not simply numbers.

The problem of causation is this: if X is associated with Y, and there is no bias or chance error, still we need to then show that X causes Y. Not just that Prozac *is associated with* less depression, but that Prozac *causes* less depression. How can we do this? A p-value will not do it for us.

This is a problem that has been central to the field of clinical epidemiology for decades. The classic handling of it has been ascribed to the work of the great medical epidemiologist A. Bradford Hill, who was central to the research on tobacco and lung cancer. A major problem with that research was that randomized studies could not be done: you smoke, you don't, and see me in 40 years to see who has cancer. This could not practically or ethically be done. This research was observational and liable to bias; Hill and others devised methods to assess bias, but they always had the problem of never being able to remove doubt completely. The cigarette companies, of course, constantly exploited this matter to magnify this doubt and delay the inevitable day when they would be forced to back off on their dangerous business.

With all this observational research, they would argue to Hill and his colleagues, you still cannot prove that cigarettes *cause* lung cancer. And they were right. So Hill set about trying to clarify how one might prove that something causes anything in medical research with human beings.

I will discuss this topic in more detail in Chapter 10. Hill basically pointed out that causation cannot be derived from any one source, but that it could be inferred by an accumulation of evidence from multiple sources (see Table 10.1).

It is not enough to say a study is valid; one also wants to know if these results are replicated by multiple studies, if they are supported by biological studies in animals on mechanisms of effect, if they follow certain patterns consistent with causation (like a dose–response relationship) and so on.

For our purposes, we might at least insist on replication. No single study should stand on its own, no matter how well done. Even after crossing the barriers of bias and chance, we should ask of a study that it be replicated and confirmed in other samples and other settings.

## Summary

Confounding bias, chance, and causation – these are the three basic notions that underlie statistics and epidemiology. If clinicians understand these three concepts, then they will be able to believe their eyes more validly.

# Chapter 3 Levels of evidence

> With a somewhat ready assumption of cause and effect and, equally, a neglect of the laws of chance, the literature becomes filled with conflicting cries and claims, assertions and counterassertions.
>
> Austin Bradford Hill (Hill, 1962; p. 4)

The term *evidence* has become about as controversial as the word "unconscious" had been in the Freudian heyday, or as the term "proletariat" was in another arena. It means many things to many people, and for some, it elicits reverent awe – or reflexive aversion. This is because, like the other terms, it is linked to a movement – in this case evidence-based medicine (EBM) – which is currently quite influential and, with this influence, has attracted both supporters and critics.

This book is not about EBM per se, nor is it simply an application of EBM, although it is, in my view, consistent with EBM, rightly understood. I will expand on that topic further in Chapter 12, but for now, I would like to emphasize at the very start what I take to be the most important feature of EBM: the concept of *levels of evidence*.

## Origins of EBM

It may be worthwhile to note that the originators of the EBM movement in Canada (such as David Sackett) toyed with different names for what they wanted to do; they initially thought about the phrase "science-based medicine" but opted for the term evidence instead. This is perhaps unfortunate since science tends to engender respect, while evidence seems a more vague concept. Hence we often see proponents of EBM (mistakenly, in my view) saying things like: "That opinion is not evidence-based" or "Those articles are not evidence-based." The folly of this kind of language is evident if we use the term "science" instead: "That opinion is not science-based" or "Those articles are not science-based." Once we use the term science, it becomes clear that such statements beg the question of what science means. Most of us would be open to such a discussion (which I touched on in the introduction). Yet (ironically perhaps due to the success of the EBM movement) many use the term "evidence" without pausing to think what it means. If some study is not "evidence-based," then what is it? "Non-evidence" based? "Opinion" based? But is there such a thing as "non-evidence"? Is there no opinion in evidence? Stated otherwise, do the facts speak for themselves? We have seen that they do not, which tells us that those who say such things as "That study is not evidence-based" are basically revealing their positivism: they could just as well say "That study is not science-based" because they have a very specific meaning in mind for science, which is in fact positivism. Since positivism is false, this extreme and confused notion of evidence is also false.

**Table 3.1** Levels of evidence

| |
|---|
| *Level I: Double-blind randomized trials* |
| Ia: Placebo-controlled monotherapy |
| Ib: Non placebo-controlled comparison trials, or placebo-controlled add-on therapy trials |
| *Level II: Open randomized trials* |
| *Level III: Observational studies* |
| IIIa: Nonrandomized, controlled studies |
| IIIb: Large nonrandomized, uncontrolled studies (n > 100) |
| IIIc: Medium-sized nonrandomized, uncontrolled studies (100 > n > 50) |
| *Level IV: Small observational studies (nonrandomized, uncontrolled,* 50 > n > 10) |
| *Level V: Case series* (n < 10), *Case report* (n = 1), *Expert opinion* |

From Soldani *et al.* (2005), with permission from Blackwell Publishing.

There is no inherent opposition between evidence and opinion, because "evidence" if meant to be "facts" always involves interpretation (which involves opinions or subjective assessments) as we discussed earlier.

In other words, all opinions are types of evidence; any perspective at all is based on some kind of evidence: there is no such thing as non-evidence.

In my reading of EBM, the basic idea is that we need to understand what kinds of evidence we use, and we need to use the best kinds we can: this is the concept of *levels* of evidence. Evidence-based medicine is *not* about an opposition between having evidence or not having evidence; it is about ranking different kinds of evidence (since we always have some kind of evidence or another).

## Specific levels of evidence

The EBM literature has various definitions of specific levels of evidence. The main EBM text uses letters (A through D). I prefer numbers (1 through 5), and I think the specific content of the levels should vary depending on the field of study. The basic constant idea is that randomized studies are higher levels of evidence than non-randomized studies, and that the lowest level of evidence consists of case reports, expert opinion, or the consensus of the opinion of clinicians or investigators.

Levels of evidence provide clinicians and researchers with a road map that allows consistent and justified comparison of different studies so as to adequately compare and contrast their findings. Various disciplines have applied the concept of levels of evidence in slightly different ways, and in psychiatry, no consensus definition exists. In my view, in mental health, the following five levels of evidence best apply (Table 3.1), ranked from level I as highest and level V as lowest.

The key feature of levels of evidence to keep in mind is that each level has its own strengths and weaknesses, and, as a result, no single level is completely useful or useless. All other things being equal, however, as one moves from level V to level I, increasing rigor and probable scientific accuracy occurs.

Level V means a case report or a case series (a few case reports strung together), or an expert's opinion, or the consensus of experts or clinicians or investigators' opinions (such as