What Will This Chapter Tell Me?

Like many young boys in the UK my first career choice was to become a soccer star. My granddad (Harry) had been something of a local soccer hero in his day, and I wanted nothing more than to be like him. Harry had a huge influence on me: he had been a goalkeeper, and consequently I became a goalkeeper too. This decision, as it turned out, wasn't a great one because I was a bit short for my age, which meant that I never got picked to play in goal for my school. Instead, a taller boy was always chosen. I was technically a better goalkeeper than the other boy, but the trouble was that the opposition could just lob the ball over my head (so, technique aside, I was a worse goalkeeper). Instead, I typically got played at left back ('left back in the changing room' as the joke used to go) because, despite being right footed, I could kick with my left one too. The trouble was, having spent years trying to emulate my granddad's goalkeeping skills, I didn't really have a clue what a left back was supposed to do.[1] Consequently, I didn't exactly shine in the role, and for many years that put an end to my believing that I could play soccer. This example shows that a highly influential thing (like your granddad) can bias the conclusions you come to and that this can lead to quite dramatic consequences. The same thing happens in data analysis: sources of influence and bias lurk within the data, and unless we identify and correct for them we'll end up becoming goalkeepers despite being a short arse. Or something like that.

[1] In the 1970s at primary school, no one actually bothered to teach you anything about how to play soccer; they just shoved 11 boys onto a pitch and hoped for the best.

What is Bias?

You will all be familiar with the term 'bias'. For example, if you've ever watched a sports game you'll probably have accused a referee of being 'biased' at some point, or perhaps you've watched a TV show like *The X Factor* and felt that one of the judges was 'biased' towards the acts that they mentored. In these contexts, bias means that someone isn't evaluating the evidence (e.g., someone's singing) in an objective way: there are other things affecting their conclusions. Similarly, when we analyse data there can be things that lead us to the wrong conclusions.

A bit of revision. We saw in Chapter 2 that, having collected data, we usually fit a model that represents the hypothesis that we want to test. This model is usually a linear model, which takes the form of equation (2.4). To remind you, it looks like this:

$$\text{outcome}_i = \left(b_1 X_{1i} + b_2 X_{2i} + \cdots + b_n X_{ni}\right) + \text{error}_i$$

Therefore, we predict an outcome variable from some kind of model. That model is described by one or more predictor variables (the *X*s in the equation) and parameters (the *b*s in the equation) that tell us something about the relationship between the predictor and the outcome variable. Finally, the model will not predict the outcome perfectly, so for each observation there will be some error.

When we fit a model to the data, we estimate the parameters and we usually use the method of least squares (Section 2.4.3). We're not interested in our sample so much as a more general population to which we don't have access, so we use the sample data to estimate the value of the parameters in the population (that's why we call them estimates rather than values). When we estimate a parameter we also compute an estimate of how well it represents the population such as a standard error (Section 2.5.1) or confidence interval (Section 2.5.2). We also can test hypotheses about these parameters by computing test statistics and their associated probabilities (*p*-values, Section 2.6.1). Therefore, when we think about bias, we need to think about it within three contexts:

1 things that bias the parameter estimates (including effect sizes);

2 things that bias standard errors and confidence intervals;

3 things that bias test statistics and *p*-values.

These situations are related: first, if the standard error is biased then the confidence interval will be too because it is based on the standard error; second, test statistics are usually based on the standard error (or something related to it), so if the standard error is biased test statistics will be too; and third, if the test statistic is biased then so too will its *p*-value. It is important that we identify and eliminate anything that might affect the information that we use to draw conclusions about the world: if our test statistic is inaccurate (or biased) then our conclusions will be too.

Sources of bias come in the form of a two-headed, fire-breathing, green-scaled beast that jumps out from behind a mound of blood-soaked moss to try to eat us alive. One of its heads goes by the name of unusual scores, or 'outliers', whereas the other is called 'violations of assumptions'. These are probably names that led to it being teased at school, but, what the hell, it could breath fire from both heads so it could handle it. Onward into battle …