What will this chapter tell me?

Although I had learnt a lot about golf clubs randomly appearing out of nowhere and hitting you around the face, I still felt that there was much about the world that I didn't understand. For one thing, could I learn to predict the presence of these golf clubs that seemed inexplicably drawn towards my apparently magnetic head? A child's survival depends upon being able to predict reliably what will happen in certain situations; consequently they develop a model of the world based on the data they have (previous experience) and they then test this model by collecting new data/experiences. Based on how well the new experiences fit with their original model, a child might revise their model of the world.

According to my parents (conveniently I have no memory of this at all), while at nursery school the model of the world that I was most enthusiastic to try out was 'If I get my penis out, it will be really funny'. To my considerable disappointment, this model turned out to be a poor predictor of positive outcomes. Thankfully for all concerned, I soon revised this model of the world to be 'If I get my penis out at nursery school the teachers and mummy and daddy will be quite annoyed'. This revised model was a better 'fit' of the observed data. Fitting models that accurately reflect the observed data is important to establish whether a theory is true.

You'll be relieved to know that this chapter is not about my penis but is about fitting statistical models. We edge sneakily away from the frying pan of research methods and trip accidentally into the fires of statistics hell. We will start to see how we can use the properties of data to go beyond our observations and to draw inferences about the world at large. This chapter lays the foundation for the whole of the rest of the book.

Building statistical models

We saw in the previous chapter that scientists are interested in discovering something about a phenomenon that we assume actually exists (a 'real-world' phenomenon). These real-world phenomena can be anything from the behaviour of interest rates in the economic market to the behaviour of undergraduates at the end-of-exam party. Whatever the phenomenon we desire to explain, we collect data from the real world to test our hypotheses about that phenomenon. Testing these hypotheses involves building statistical models of the phenomenon of interest.

Let's begin with an analogy. Imagine an engineer wishes to build a bridge across a river. That engineer would be pretty daft if she just built any old bridge, because it might fall down. Instead, the engineer collects data from the real world: she looks at existing bridges and sees from what materials they are made, their structure, size and so on (she might even collect data about whether these bridges are still standing). She uses this information to construct an idea of what her new bridge will be (this is a 'model'). It's expensive and impractical for her to build a full size version of her bridge, so she builds a scaled-down version. The model may differ from reality in several ways – it will be smaller for a start – but the engineer will try to build a model that best fits the situation of interest based on the data available. Once the model has been built, it can be used to predict things about the real world: for example, the engineer might test whether the bridge can withstand strong winds by placing the model in a wind tunnel. It is important that the model is an accurate representation of the real world or her conclusions based on the model can't be extrapolated to the real-world bridge.

Scientists do much the same: they build (statistical) models of real-world processes in an attempt to predict how these processes operate under certain conditions (see Jane Superbrain Box 2.1). Unlike engineers, we don't have access to the real-world situation and so we can only ever *infer* things about psychological, societal, biological or economic processes based upon the models we build. However, just like the engineer, we want our model to be as accurate as possible so that we can be confident that the predictions we make about the real world are also accurate; the statistical model we build must represent the data collected (the *observed data*) as closely as possible. The degree to which a statistical model represents the data collected is known as the **fit** of the model.

Figure 2.2 illustrates three models that an engineer might build to represent the real-world bridge that she wants to create. The first model is an excellent representation of the real-world situation and is said to be a *good fit*. If the engineer uses this model to make predictions about the real world then, because it so closely resembles reality, she can be confident that these predictions will be accurate. So, if the model collapses in a strong wind, then there is a good chance that the real bridge would collapse also. The second model has some similarities to the real world: the model includes some of the basic structural features, but there are some big differences too (e.g., the absence of one of the supporting towers). We might consider this model to have a *moderate fit* (i.e., there are some similarities to reality but also some important differences). If the engineer uses this model to make predictions about the real world then these predictions may be inaccurate or even catastrophic (e.g., the model predicts that the bridge will collapse in a strong wind, causing the real bridge to be closed down, creating 100-mile tailbacks with everyone stranded in the snow, all of which was unnecessary because the real bridge was perfectly safe – the model was a bad representation of reality). We can have some confidence, but not complete confidence, in

predictions from this model. The final model is completely different to the real-world situation; it bears no structural similarities to the real bridge and is a *poor fit*. Any predictions based on this model are likely to be completely inaccurate. Extending this analogy to science, if our model is a poor fit of the observed data then the predictions we make from it will be equally poor.