Statistical Thinking for the 21st Century

Chapter 5

Fitting models to data

One of the fundamental activities in statistics is creating models that can summarize data using a small set of numbers, thus providing a compact description of the data. In this chapter we will discuss the concept of a statistical model and how it can be used to describe data.

5.1 What is a model?

In the physical world, "models" are generally simplifications of things in the real world that nonetheless convey the essence of the thing being modeled. A model of a building conveys the structure of the building while being small and light enough to pick up with one's hands; a model of a cell in biology is much larger than the actual thing, but again conveys the major parts of the cell and their relationships.

In statistics, a model is meant to provide a similarly condensed description, but for data rather than for a physical structure. Like physical models, a statistical model is generally much simpler than the data being described; it is meant to capture the structure of the data as simply as possible. In both cases, we realize that the model is a convenient fiction that necessarily glosses over some of the details of the actual thing being modeled. As the statistician George Box famously said: "All models are wrong but some are useful."

The basic structure of a statistical model is:

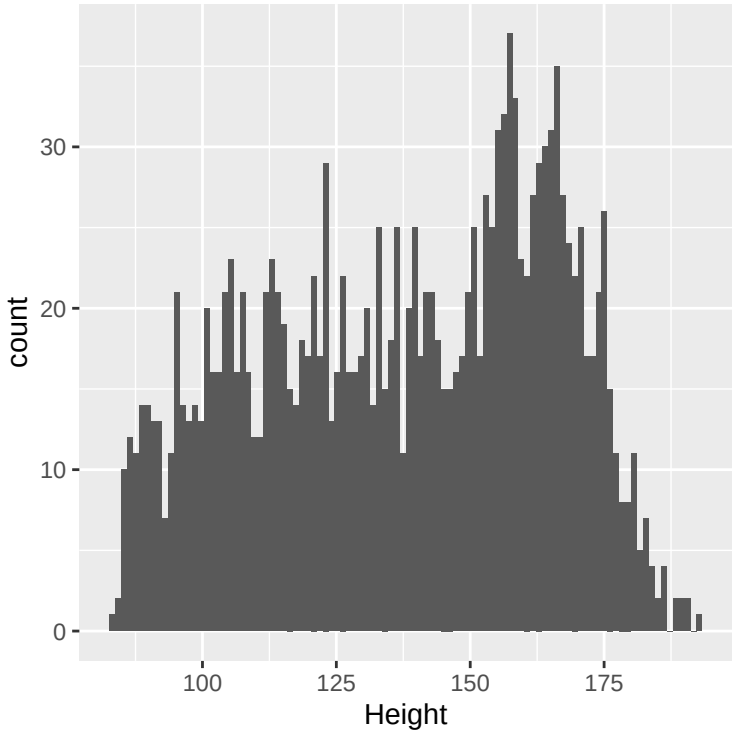

Figure 5.1: Histogram of height of children in NHANES.

$$data = model + error$$

This expresses the idea that the data can be described by a statistical model, which describes what we expect to occur in the data, along with the difference between the model and the data, which we refer to as the *error*.

5.2 Statistical modeling: An example

Let's look at an example of fitting a model to data, using the data from NHANES. In particular, we will try to build a model of the height of children in the NHANES sample (see Figure 5.1).

Remember that we want to describe the data as simply as possible while still capturing their important features. What is the simplest model we can imagine that might still capture the essence of the data?

How about the most common value in the dataset (which we call the *mode*)?

The mode describes the entire set of 1691 children in terms of a single number. If we wanted to predict the height of any new children, then our guess would be the same number: 166.5 centimeters.

$$\hat{height_i} = 166.5$$

We put the hat symbol over the name of the variable to show that this is our *predicted* value. The error for this individual would then be the difference between the predicted value ($\hat{height_i}$)and their actual height ($height_i$):

$$error_i = height_i - \hat{height_i}$$

How good of a model is this? In general we define the goodness of a model in terms of the error, which represents the difference between model and the data; all things being equal, the model that produces lower error is the better model.

What we find is that the average individual has a fairly large error of -28.8 centimeters. We would like to have a model where the average error is zero, and it turns out that if we use the arithmetic mean (commonly known as the *average*) as our model then this will be the case.

The mean (often denoted by a bar over the variable, such as $\bar{X}$) is the sum of all of the values, divided by the number of values. Mathematically, we express this as:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

We can prove mathematically that the sum of errors from the mean (and thus the average error) is zero. Given that the average error is zero, the mean seems like a better model than the mode.