1.4: The Big Ideas of Statistics

There are a number of very basic ideas that cut through nearly all aspects of statistical thinking. Several of these are outlined by Stigler (2016) in his outstanding book "The Seven Pillars of Statistical Wisdom", which I have augmented here.

1.4.1 Learning from data

One way to think of statistics is as a set of tools that enable us to learn from data. In any situation, we start with a set of ideas or *hypotheses* about what might be the case. In the PURE study, the researchers may have started out with the expectation that eating more fat would lead to higher death rates, given the prevailing negative dogma about saturated fats. Later in the course we will introduce the idea of *prior knowledge*, which is meant to reflect the knowledge that we bring to a situation. This prior knowledge can vary in its strength, often based on our amount of experience; if I visit a restaurant for the first time I am likely to have a weak expectation of how good it will be, but if I visit a restaurant where I have eaten ten times before, my expectations will be much stronger. Similarly, if I look at a restaurant review site and see that a restaurant's average rating of four stars is only based on three reviews, I will have a weaker expectation than I would if it was based on 300 reviews.

Statistics provides us with a way to describe how new data can be best used to update our beliefs, and in this way there are deep links between statistics and psychology. In fact, many theories of human and animal learning from psychology are closely aligned with ideas from the new field of *machine learning*. Machine learning is a field at the interface of statistics and computer science that focuses on how to build computer algorithms that can learn from experience. While statistics and machine learning often try to solve the same problems, researchers from these fields often take very different approaches; the famous statistician Leo Breiman once referred to them as "The Two Cultures" to reflect how different their approaches can be (Breiman 2001). In this book I will try to blend the two cultures together because both approaches provide useful tools for thinking about data.

1.4.2 Aggregation

Another way to think of statistics is "the science of throwing away data". In the example of the PURE study above, we took more than 100,000 numbers and condensed them into ten. It is this kind of *aggregation* that is one of the most important concepts in statistics. When it was first advanced, this was revolutionary: If we throw out all of the details about every one of the participants, then how can we be sure that we aren't missing something important?

As we will see, statistics provides us ways to characterize the structure of aggregates of data, and with theoretical foundations that explain why this usually works well. However, it's also important to keep in mind that aggregation can go too far, and later we will encounter cases where a summary can provide a misleading picture of the data being summarized.

1.4.3 Uncertainty

The world is an uncertain place. We now know that cigarette smoking causes lung cancer, but this causation is probabilistic: A 68-year-old man who smoked two packs a day for the past 50 years and continues to smoke has a 15% (1 out of 7) risk of getting lung cancer, which is much higher than the chance of lung cancer in a nonsmoker. However, it also means that there will be many people who smoke their entire lives and never get lung cancer. Statistics provides us with the tools to characterize uncertainty, to make decisions under uncertainty, and to make predictions whose uncertainty we can quantify.

One often sees journalists write that scientific researchers have "proven" some hypothesis. But statistical analysis can never "prove" a hypothesis, in the sense of demonstrating that it must be true (as one would in a logical or mathematical proof). Statistics can provide us with evidence, but it's always tentative and subject to the uncertainty that is always present in the real world.

1.4.4 Sampling

The concept of aggregation implies that we can make useful insights by collapsing across data – but how much data do we need? The idea of *sampling* says that we can summarize an entire population based on just a small number of samples from the population, as long as those samples are obtained in the right way. For example, the PURE study enrolled a sample of about 135,000 people, but its goal was to provide insights about the billions of humans who make up the population from which those people were sampled. As we already discussed above, the way that the study sample is obtained is critical, as it determines how broadly we can generalize the results. Another fundamental insight about sampling is that while larger samples are always better (in terms of their ability to accurately represent the entire population), there are diminishing returns as the sample gets larger. In fact,

the rate at which the benefit of larger samples decreases follows a simple mathematical rule, growing as the square root of the sample size, such that in order to double the quality of our data we need to quadruple the size of our sample.
