Sampling

2022-09-01

Contents

Reading:

Required Reading (everyone):

- Statistical Rethinking, Ch. 3 ("Sampling the Imaginary").

Reading Notes:

The concept of sampling is central to statistics. We use sampling in several ways in this book, and more generally in Bayesian statistics:

1. Central to the whole idea of statistics is that we want to draw inferences about a large population of individuals (people, animals or plants in nature, weather events, distribution of minerals in the ground, etc.) from a sample that's much smaller than the population.

   Some of the most familiar examples are around public opinion polling, where political scientists try to estimate the sentiments of millions, even hundreds of millions of people from the answers a few hundred to a few thousand answer questions on a survey. There are many other applications: Geologists will try to estimate the hazard of radon gas coming from the ground based on samples collected at a few dozen sites across a state. Biologists will estimate the population of a species in an area from samples of a few individuals observed in the wild. Engineers will estimate the failure rate of a manufacturing line in a factory from a few samples collected and tested. Medical researchers will estimate the prevalence of a diseae and the effectiveness of a medicine in treating it from relatively small samples of people tested or treated.

2. In Bayesian statistics, we often use sampling as an efficient way to approximate the values of difficult integrals that we can't solve using analytical mathematics.

3. In Bayesian statistics, we also use sampling to make predictions of the larger population from the inferences we drew from our sample of observed data. Sampling from the *posterior distribution* of probability allows us to not only make predictions, but to understand the uncertainty of our predictions in ways that we cannot when we use other approaches to statistical analysis.

4. Section 3.3 also discusses ways that we can use sampling from the posterior to check how well our model describes our sampled data and to make sure that our analysis software doesn't have bugs.

As you read the example of analyzing an imaginary test for detecting vampires at the beginning of the chapter, think about what McElreath writes about how easy it is to understand the analysis when you use Bayes's theorem in a formal mathematical sense versus presenting it in terms of *natural frequencies*. This relates to the distinction I drew in class when we discussed Chapter 2 about *parametric* versus *nonparametric* ways of applying Bayes's theorem. The formal mathematical approach is more like *parametric* statistics and the natural frequency approach is more like *nonparametric* statistics. This helps motivate why sampling methods (grid sampling for now, and Monte Carlo sampling later in the semester) can help us intuitively understand what we're doing when we do Bayesian data analysis. Complex mathematical formulas, especially ones with sums and integrals, can be intimidating and confusing to many people, and presenting frequencies and samples can be more intuitive and easier to understand.