Statistical Techniques

Modern data analysis involves techniques from several fields, including statistics and machine learning.

Statistics is the body of knowledge about drawing conclusions from incomplete information. Put another way, statistics allows us to reason about data in the presence of uncertainty.

Among the statistical tools that we will use are hypothesis testing and tools for prediction of new data.

Because we are doing data science, we will take a rather different approach from standard statistical teaching. Data science puts coding at the heart of data analysis. Coding gives us a much more powerful set of tools than have been available in the past. This has great benefits for analysis, but also for teaching.

The benefits for analysis are that we can get, clean, and analyze a much wider range of data. It becomes natural to extend our methods of analysis to techniques based in computation, like machine learning.

The benefits for teaching are two-fold. The first benefit is our ability to analyze real data. The greater power and range of our tools allow us to analyze real-world, messy data instead of cleaned-up toy datasets, so you are better prepared to analyze the real data you will soon have to deal with in your education and work. The second benefit is that an emphasis on computation allows us to use richer, simpler and more powerful techniques, based in resampling, that are easier to explain, and have a deeper relationship to the models that we are using. We will rely much less on mathematics, and that gives us time to explain the ideas in another way, and in more depth.

The discipline of statistics has long addressed the same fundamental challenge as data science: how to draw robust conclusions about the world using incomplete information. One of the most important contributions of statistics is a consistent and precise vocabulary for describing the relationship between observations and conclusions. This text continues in the same tradition, focusing on a set of core inferential problems from statistics: testing hypotheses, estimating confidence, and predicting unknown quantities.

Data science extends the field of statistics by taking full advantage of computing, data visualization, machine learning, optimization, and access to information. The combination of fast computers and the Internet gives anyone the ability to access and analyze vast datasets: millions of news articles, full encyclopedias, databases for any domain, and massive repositories of music, photos, and video.

Applications to real data sets motivate the statistical techniques that we describe throughout the text. Real data often do not follow regular patterns or match standard equations. The interesting variation in real data can be lost by focusing too much attention on simplistic summaries such as average values. Computers enable a family of methods based on resampling that apply to a wide range of different inference problems, take into account all available information, and require few assumptions or conditions. Although these techniques have often been reserved for advanced courses in statistics, their flexibility and simplicity are a natural fit for data science applications.