What Use is Statistics for Massive Data?

BY DIANE LAMBERT

Bell Labs, Lucent Technologies

Statistics in the broad sense is about extracting information from data. The common view of statistics is much narrower, though. Often it is seen only as a set of cookbook methods that are designed for small sets of data that are obtained according to a known design or sampling plan. The massive dynamic sets of data with tens or hundreds of gigabytes or even terabytes of data that are increasingly common in business, manufacturing, environmental sciences, astronomy, data networking and many other areas are felt to be beyond the domain of statistics. Moreover, the most visible challenges for massive data involve computing, which can lead to the view that computer science is more appropriate for understanding massive data than statistics is. This paper, however, argues that the discipline of statistics and thoughtful practitioners and researchers are still essential for extracting information from really big sets of data.

1. What is Statistics? Most people are first exposed to statistics in a required undergraduate course that is filled with a hundred or more other students who are also required to be there. Partially in response to demands from other departments, the introductory course focuses on traditional methods for the kinds of small experiments, studies and surveys that are part of the curriculum for other departments. Typically, nearly all, if not all, the datasets considered have fewer than 100 observations with only a few variables on each, and the objective is to apply simple methods for finding means, variances, confidence intervals and p-values and for fitting linear regression models. There is no discussion of computing, probably both for lack of time and because sophisticated computing is not needed. There is some discussion of mathematical properties of techniques, but usually not in ways that are intuitive. Inference centers on confidence intervals and p-values, which are simple to compute but subtle to explain. What most people take away from the course is that statistics is neither particularly hard nor interesting (except for some peculiar use of language), it is only useful for simple data and simple questions (if at all), and any one can apply it (although it is best avoided), but hardly anyone can understand it.

But statistics is much different, and not just broader or deeper, from what is taught in undergraduate courses. Simply stated, statistics is about extracting information from data that are noisy or uncertain. The unstated position is that all data are noisy. Twenty measurements from a small experiment are noisy, and zillions of transaction records in a data warehouse are noisy. The twenty observations from an experiment are noisy because measurement errors are unavoidable. Transactions have noise because the individuals generating them vary, and transactions for even just one individual vary. In statisticians' terms, transactions vary across individuals and vary within any one individual over time and space. If the data arise by sampling, so not all individuals are included, then errors and variability are introduced