John Tukey’s Introductory Paragraph

For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. ...All in all I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data

Why is it important?

- Introduced the term “data analysis” as a name for what applied statistician do
- Differentiating “data analysis” from formal statistical inference
- Embedded statistics in a larger entity

Four driving forces in the new science

1. The formal theories of statistics
2. Accelerating developments in computers and display devices
3. The challenge, in many fields, of more and ever larger bodies of data
4. The emphasis on quantification in an ever wider variety of disciplines

New Science, now what?

A field larger than what academic statistics could deliver

Action plan!

Cleveland’s 6 foci of activity

- Goal
  - expand the technical areas of statistics focuses on the data analyst
  - sets out six technical areas of work for a university department
  - advocates a specific allocation of resources devoted to research in each area and to courses in each area

Cleveland’s 6 foci of activity

- Multidisciplinary investigations (25%)
- Models and Methods for Data (20%)
- Computing with Data (15%)
- Pedagogy (15%)
- Tool Evaluation (5%)
- Theory (20%)

Breiman’s Two Cultures

Generative Modeling

- seeks to develop stochastic models which fit the data
- make inferences about the data-generating mechanism based on the structure of those models
- there is a true model generating the data, and often a truly “best” way to analyze the data

Breiman’s Two Cultures

Predictive modeling

- prioritizes prediction
- effectively silent about the underlying mechanism generating the data
- allows for many different predictive algorithms
- discuss only accuracy of prediction made by different algorithm on various datasets
- Example: Machine Learning

The Common Task Framework (CTF)

(a) A publicly available training dataset involving, for each observation, a list of (possibly many) feature measurements, and a class label for that observation.

(b) A set of enrolled competitors whose common task is to infer a class prediction rule from the training data.

(c) A scoring referee, to which competitors can submit their prediction rule. The referee runs the prediction rule against a testing dataset, which is sequestered behind a Chinese wall. The referee objectively and automatically reports the score (prediction accuracy) achieved by the submitted rule.