Approaches Toward Data Analysis

B. Gerstman, San Jose State University

Traditional, Test-Based Inference

The traditional form of statistical inference was developed in the early part of the 20th century to address the needs of agricultural and laboratory-based experiments. It started by specifying a single question and testable hypothesis (e.g., Does Fertilizer A result in a better crop yield than Fertilizer B?). Data were then collected, an underlying sampling distribution was assumed, and data were submitted to testing. Conclusions were categorical in the sense of rejecting or accepting a specific hypothesis (e.g., Fertilizer A increases crop yields). If the analysis beget a new question, the process was begun from scratch with primary data collection.

Today, this test-based approach toward does not serve all research questions equally well. For one, it does not uncover previously unrecognized patterns or associations. Second, it does not address non-sampling errors inherent in non-experimental data. And third, it fails to quantify the magnitude of any observed difference. Therefore, the traditional test-based method of inference is not appropriate for all research purposes; a more broad-based method of inference is needed.

John Tukey Suggests How Knowledge and Belief are Bought

The philosophy, development, and promotion of newer methods of statistical inference owe much to the work and words of statistician John W. Tukey. Tukey's reminds us that knowledge and belief have a price.[1] He asks "With what coin do we buy this belief?", and suggests the following "costs."

1. The care and insight with which data are collected and in which the study is planned and performed.
2 The effort involved in collecting enough data.
3. The formal error-rate that we are willing to accept as a basis for our conclusions.

In introductory statistics classes, we focus too often on coin 3 to the exclusion of coins 1 and 2.

Components of Current Analysis

The above discussion suggests that data analysis requires more than hypothesis testing. Specifically, it requires us to learn all we can about the data before it is submitted to the test. This is not a new idea, just one that needs to be continually "rediscovered" by each generation. We are, therefore, compelled to consider multiple elements of the data, including its genesis, quality, and use before submitting it to analysis. We are also compelled to pursue multiple forms of analysis for each set of data. Data exploration needs to brought to the fore, and measures of association need to consider both the direction and magnitude of observed differences. It is not enough to test the data for significance.

In suggesting a pragmatic but thorough course of analysis, the following steps are recommended:

1. Carefully consider the care and quality of the study design and data collection.
2. Explore the data with relevant summary statistics and graphs.
3. Calculate point estimates and confidence intervals for parameters of interest.
4. Follow leads with tests of significance, when necessary.
5. Always put results in plain English (or Spanish, or French, etc.).
6. Non-significant should not be discarded and should at minimum be submitted to power analysis.

Perhaps these steps are too rigid (good analysis is flexible), but they do serve to remind us that good data analysis is neither singular nor simple.

---

[1] Tukey, J. W. (1969.) Analyzing data: Sanctification or detective work? *American Psychologist,* 24, 83 - 91. Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, 6, 100 - 116.