MICHIGAN STATE UNIVERSITY
Department of Statistics and Probability

COLLOQUIUM

Marianne Huebner
Michigan State University

A Contemporary Framework for Initial Data Analysis

Tuesday, September 6, 2016
10:20 a.m. - 11:10 am
Refreshments 10:00 am
C405 Wells Hall

Abstract

Scientists perform initial data analyses (IDA) as part of their research studies, often informally and unstructured. For some researchers this may mean data cleaning, or basic data summaries, or others perform explorative data analyses which are then formalized in statistical models. This may have a large and non-transparent impact on results and conclusions presented in publications. With standard software packages many researchers may rush to perform sophisticated analyses, without systematically checking for errors in the data and without a clear understanding about the underlying features of the data. To improve this situation, IDA has to be established as a necessary and genuine step in the mind of all researchers. Systematically checking for errors or getting an understanding of the underlying features of the data are necessary steps. Although there have been articles listing and discussing important elements of IDA, we lack a general, conceptual framework. How to perform IDA in a structured and strategic way needs to be discussed and reporting guidelines need to be developed. We try to clarify the scope and the aim of IDA, focusing on the question "Why do we do what"? We divide IDA into five major steps:
1.Data cleaning
2.Data screening that consists of understanding the properties of the data
3.Initial data reporting that informs all potential collaborators working with the data in future about insights
4.Refining and updating the analysis plan that translates the relevant findings into adaptations to the analysis plan
5.Reporting of IDA in research papers that document steps that impact the interpretation of results.

Nontransparent changes or data driven hypotheses are to be avoided in IDA. Modern challenges for IDA are size and complexity of datasets, including data from different sources, or data that were collected for administrative instead of research purposes. This work is being developed as part of the STRATOS Initiative (http://stratos-initiative.org/). The STRATOS initiative is a large collaboration of statisticians (82 statisticians from 15 countries). The objective of STRATOS is to provide accessible and accurate guidance in the design and analysis of observational studies.

To request an interpreter or other accommodations for people with disabilities, please call the Department of Statistics and Probability at 517-355-9589.