

Introduction to Advanced Data Analysis

Professor: Hammou El Barmi
Columbia University

- Experiments are performed to generate data in order to make decisions.
- Much of the scientific knowledge about processes and systems is based on induction: reasoning from the specific to the general.
- Example(survey): Do NYC residents favor allowing guns in schools?
 - Specific cases: 100 NYC adults are called for a telephone survey
 - Inferential goal: get information about whether or not NYC residents favor allowing guns in schools.
- Example: Does hormone replacement improve health status in post-menopausal women?
 - Specific cases: the health status monitored in 1000 women over a 3-year period. Some took hormones, others did not.
 - Inferential goal: Determine if hormones improve the health of women not in the study.

- In general we are interested in how do the inputs of a process affect an output.
- The input variables can be divided into three categories:
 - controllable factors: measured and determined by scientist.
 - uncontrollable factors: measured but not determined by scientist.
 - noise factors: unmeasured, uncontrolled factors, often called experimental variability or error.

- For any interesting process, there are inputs such that:

variability in input \implies variability in output

- If variability in x leads to variability y , we say x is a source of variation.
- Good design and analysis of experiments can identify sources of variation.

Information on how inputs affect the output can be gained from:

- **Observational studies:**
 - Input and output variables are observed from a pre-existing population.
 - It may be hard to say what is input and what is output.
- **Controlled experiments:** One or more input variables are controlled and manipulated by the experimenter to determine their effect on the output.

- Randomized, controlled, double-blind experiments
 - Randomization guards against selection bias.
 - Can eliminate correlation between x and y due to a different cause. aka a confounder.
No causation without randomization
 - Ensure that the groups are comparable
 - Double-blind: minimizes bias in the response and in the evaluations of the experimental outcome
- Observational studies:
 - Assignment of experimental units to study groups is not done by the researcher
 - May lack advantages of controlled trials
 - May help establish association but not causation.
 - Can suggest good experiments to run, but can't definitively show causation.

When conducting a data analysis you need to ask the following questions:

- What is the objective of the analysis and/or the original experiment?
- What was the design of the study?
 - Randomized controlled or observational?
 - If a controlled trial, how were subjects assigned to the different groups?
 - Was the assignment process controlled by the investigator?
 - If an observational study: Are the groups comparable? What factors are confounded with treatment?
- What procedure would be appropriate for the data?
 - Exploratory data analysis techniques?
 - Inferential statistical techniques?
 - Model building?
- Implementation of analysis plan?
- Interpretation of Results?

- Preliminary look at data:
 - Evaluating data quality
 - Missing values
 - Outliers/Influential points
- Checking assumptions: Distributions, relationships, etc.
- Compute measures of location & dispersion

To do so, compute

- Numerical descriptive statistics (Measure of location, dispersion, skewness, kurtosis)
- Graphical descriptive statistics (histograms, box plots, stem and leaf plots, QQ plots)

- Identify research hypotheses to be tested.
- Choose a set of experimental units, which are the units to which treatments will be randomized.
- Choose a response/output variable.
- Determine potential sources of variation in response:
 - factors of interest
 - nuisance factors
- Decide which variables to measure and control:
 - treatment variables
 - potential large sources of variation in the units (blocking variables)
- Decide on the experimental procedure and how treatments are to be randomly assigned.