

Introduction to Advanced Data Analysis

Professor: Hammou El Barmi
Columbia University

Course Description: This course reviews and expands upon core topics in statistics through the study and practice of data analysis. It will be computationally intensive and we will use the R and or SAS languages and environments for statistical computing and graphics. I will assume that you had regression. The material that will be covered include: **basic statistical procedures (including nonparametric procedures), linear regression, design of experiments, categorical data, survival analysis and multivariate analysis.** Because of the nature of the course, I will use several books and they include

- The Statistical Sleuth: A Course in Methods of Data Analysis: Ramsey and Schafer, Duxbury
- Applied Linear Statistical Models: Kutner, Nachtsheim, Neter and Li (5th Edition), McGraw Hill
- Methods of Multivariate Analysis, 2nd Edition, Alvin C. Rencher
- Categorical Data Analysis, 2nd Edition, Alain Agresti, Wiley
- The Statistical Analysis of Failure Time Data, 2nd Edition, J. D. Kalbfleisch and Ross L. Prentice, Wiley

Grading: Your final grade will be based on an in class exam, take home assignments (HWs) and a project. The homework will be assigned approximately every week and will be collected a week later. Late assignments will not be accepted and they will result automatically in a zero. HWs submitted by email will NOT be graded and the lowest score on the homework assignments will be dropped. The exam will count for 40% of your final score and the HW will count for 30%. The project will count for the remaining 30%.

- 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 \Rightarrow 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.

控制实验可以用于探究因果关系：控制变量，增加随机性消除其他因素的干扰，实现双盲减少主管的干扰 》》》 增加试验组的可比性和实验结果的可靠性

观测学习可以用于探究相关关系

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 interest factors 就是后面需要被treat的。treat的方式一般就是通过分组。
 - nuisance factors nuisance factors 就是后面需要被blocked variables
- Decide which variables to measure and control:
 - treatment variables 想要研究对观测值变动是否影响的 那个变量
 - potential large sources of variation in the units (blocking variables) 除了想要研究的之外需要通过随机化或者双盲从而实现对试验影响尽可能小的其他变量
- Decide on the experimental procedure and how treatments are to be randomly assigned.