CUSP-GX-5004: APPLIED DATA SCIENCE

Foundation Session 1

Course Overview

- Big data is uniting researchers from multiple fields.
- Physical, computer, and social scientists.
- Each trying to bring the strengths of their respective disciplines.
- University-based multidisciplinary centers such as CUSP are the nexus.
- Ideally focused on practical questions to improve quality of life.
- This course aims to reflect this multidisciplinary view.

Course Goals

- Challenging to develop a class like this: different backgrounds and goals.
- Our goals are expose you to a wide variety of tools for data analytics.
- My piece is key ingredients of statistical theory and practice.
- Ultimately, we want you to be able to **intelligently** analyze real-world data, largely with an urban focus.

Course Topics

- Important elements from probability theory.
- Random variables, their moments, and key ideas on asymptotics.
- Introduction to machine learning. The bivariate linear model and hypothesis testing under ideal circumstances. Violation of ideal circumstances.
- The multivariate linear model, probability models, and generalized linear models.
- Time series analysis.
- Identification and treatment effects.
- Topics chosen by the class.

Suggested Readings/Textbooks

Angrist and Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 2009.

Hastie, et al., *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd Edition, Springer. (Online: http://statweb.stanford.edu/~tibs/ElemStatLearn/)

Kabacoff, *R in Action: Data analysis and graphics with R*, 1st Edition, Manning Publications.

McKinney, *Python for Data Analysis*, O'Reilly Media Inc., 2013.

Sheppard, *Introduction to Python for Econometrics, Statistics and Data Analysis*, August 2014. (Online: http://www.kevinsheppard.com/images/0/09/Python_introduction.pdf.)

Zumel and Mount, *Practical Data Science with R*, 1st Edition, Manning Publications Company, March 2014. (Select chapters available for free online: http://www.manning.com/zumel/)

An Introductory Lesson

- Economists have been doing “big data” for decades.
- Bruising endeavor with hard won lessons.

Measures of Economic Activity

- Prior to the 1930s, there were few government-sponsored measures of economic activity.
- As the US and Europe slid into the Great Depression, policy makers lacked basic information.
- In the US, the National Accounts were born in 1934 and greatly expanded during and after WWII.
- At the same time, Alfred Cowles established the Cowles Commission for Research in Economics.

The Cowles Commission

- Cowles approach was a probabilistic framework to estimate systems of simultaneous equations to model an economy.
- Ultimately would develop very large scale econometric models to examine a host of different economic variables.
  - Cowles was big data of the day (and of a sort).
- Main insight was a demonstrated bias of ordinary least squares estimates derived from such models.
  - Drove new statistical methods such as instrumental variables and full- and limited-information maximum likelihood.
- But such an approach was found to be inadequate for policy evaluation (“Goodhart’s law” and “Lucas critique”).

Goodhart and Lucas

- Goodhart "asserts that any economic relation tends to break down when used for policy purposes." (Wickens [2008].)
  - Proposed relationships, economic or otherwise, are not structural in nature (reduced or semi-reduced form).
  - Instead derived from fundamental behavioral relationships (structural).
- Lucas (1976) notes that individual decision rules affected by policy are driven by "deep structural parameters."
  - Decision rules and, therefore, decisions are contingent on the state of the system *as it is*.
  - Change the system through policy, change the decision rule.
  - Such changes may not be captured in non-structural models.

Structure and Experiment

- The Cowles approach yielded hard-won knowledge.
  - Substantial innovation in approaches to data analysis.
  - Highlighted fundamental limitations when applied to policy evaluation.
- Even with structural models, economists have long recognized that data are typically non-experimental ("found data").
  - May be the "digital exhaust" of human activity.
  - As a result, analysis is subject to potential selection bias.
  - Development of techniques to deal with such bias.
  - Moreover, direction of causation must be clearly understood ("umbrellas cause rain").

What Do You Want Out of This Course?

Optimization

Refer Notes for Foundation Session 1

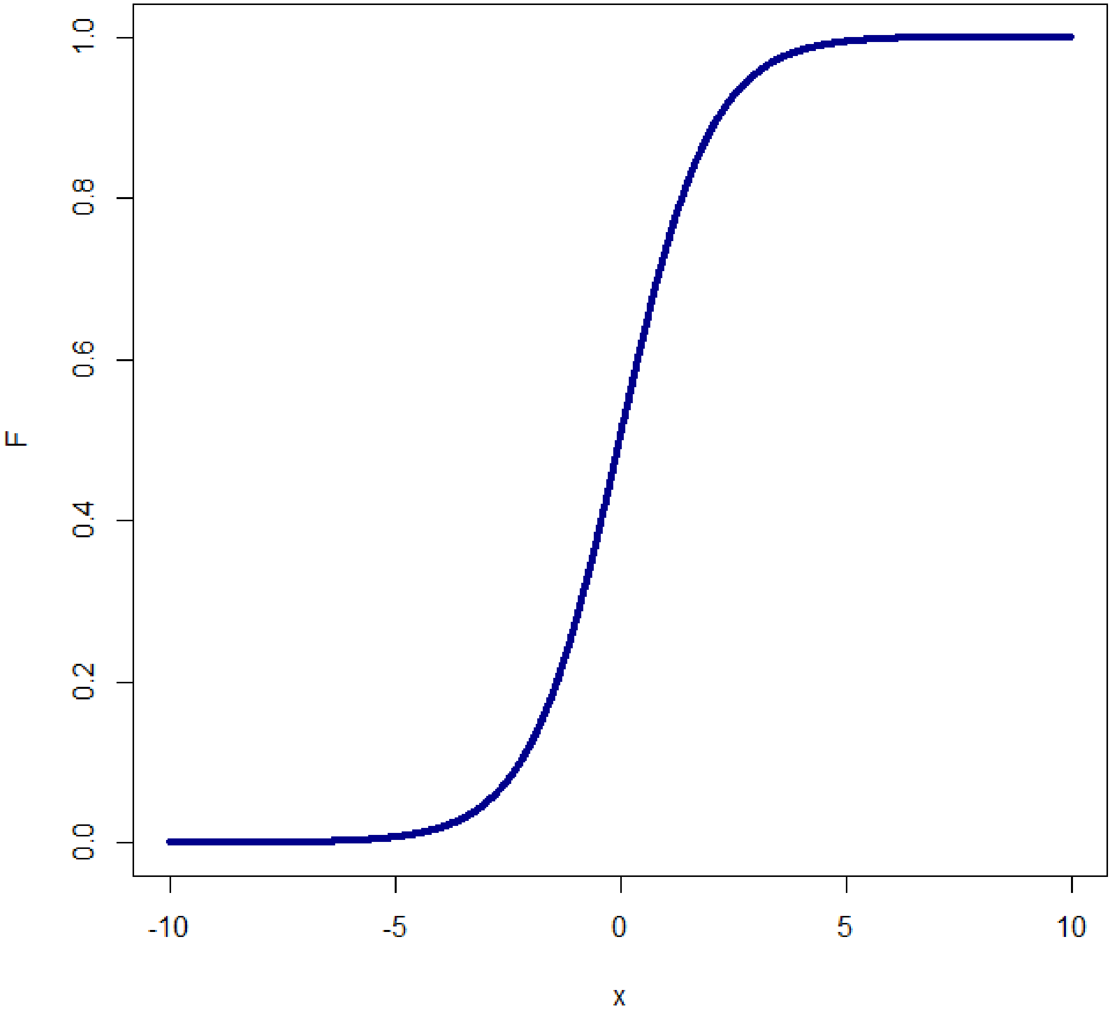
Important Points from Probability Theory

Refer Notes for Foundation Session 1

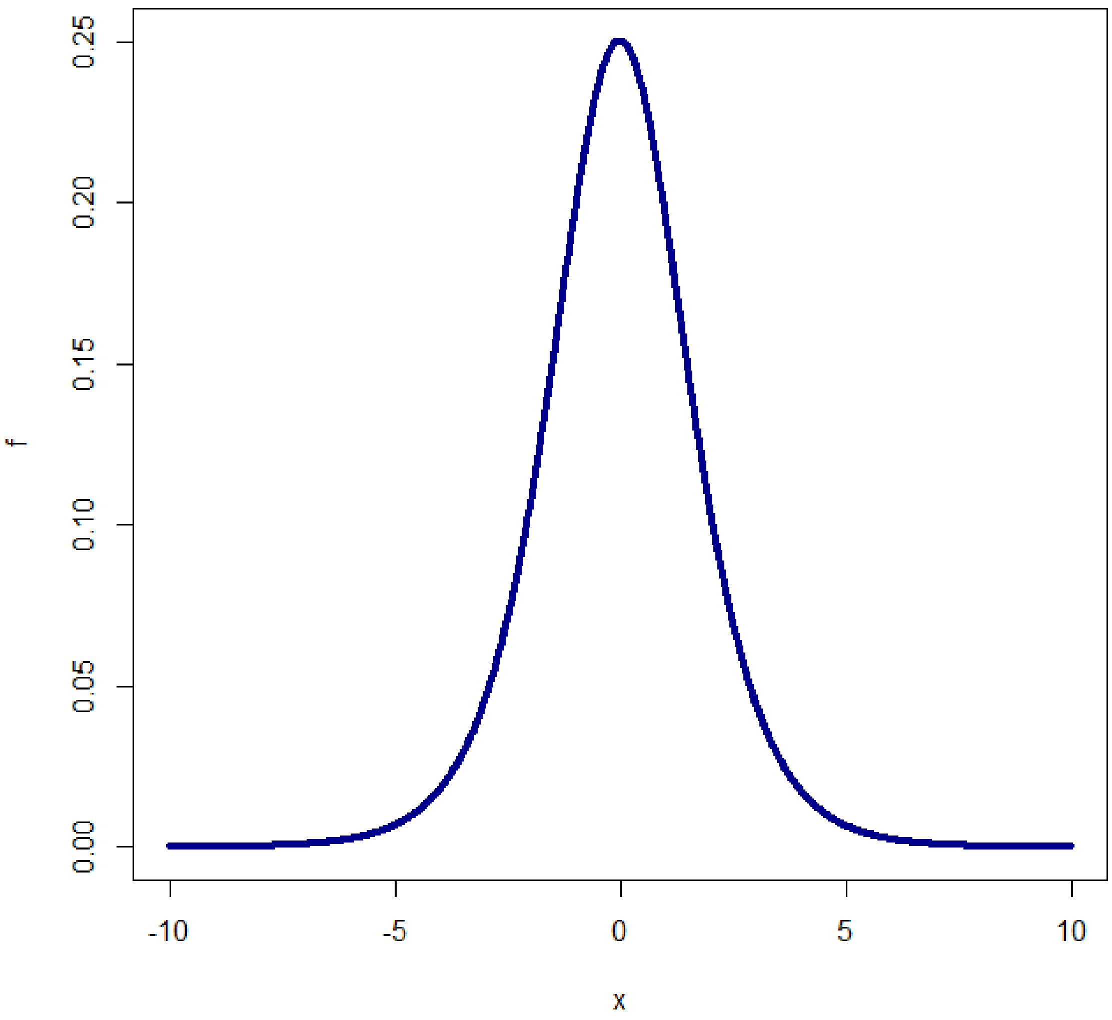
Random Variables and Frequentist Statistics

Refer Notes for Foundation Session 1

Logistic CDF



Logistic PDF



Common Discrete PDFs

- Bernoulli: coin flip (heads or tails)
- Binomial: multiple coin flips
- Multinomial: multiple outcomes (position A, B, C, or D)
- Discrete uniform: roll of a die or dice
- Poisson: integer valued, often countiing (number of visits to the doctor)

Common Continuous PDFs

- Normal: Nature, Law of Large Numbers, Central Limit Theorem (big data)
- t: small sample hypothesis testing
- Uniform: random number generator
- Chi-Squared: square of the normal
- Log normal: transformation of non-negative things like wages or stock returns
- Logistic: probability models