MECO 7312: Advanced Statistics and Probability

Dr. Khai Chiong

Fall, 2022

E-mail: khai.chiong@utdallas.edu
Office Hours: by appointment
Office: JSOM 13.324
Course modality: in-person

Web: www.khaichiong.com/teaching
Class Hours: Wednesday 1-3:45pm
Class Room: JSOM 13.501

Course Description

The goal of this PhD-level course is to develop a rigorous theoretical foundation necessary for research in applied econometrics and statistics. It covers the core topics of probability theory and statistical inference, including properties of random variables and probability distributions, frequentist and bayesian estimation, asymptotic theory, and various methods of hypothesis testing.

Texbooks

The main required textbook is *Statistical Inference* by Casella and Berger.

- Casella, George, and Roger L. Berger. *Statistical inference*. 2nd Edition
- Davidson, Russell, and James G. MacKinnon. *Econometric theory and methods*.
- Cameron, A. Colin, and Pravin K. Trivedi. *Microeconometrics: methods and applications.*

Grades

Course grade is based on the weighted average of four problem sets and one final exam. Please work together in a team of 4 to 5 for the problem sets – we often learn a great deal from our peers.

Programming and Statistical Computing

At the end of this course, students will be expected to be familiar with the following:

LaTeX

LaTeX is a program for typesetting mathematical notations. It is what I used to typeset all my lecture notes, research papers and journal publications. If your research is heavy on mathematical notations, it is a must over Microsoft Word.

There are many online resources for LaTeX. To install LaTeX:, see: https://guides.nyu.edu/LaTeX/installation. You will also need a front-end text editor, I recommend TeXstudio.

You will be expected to typeset all your assignments using LaTeX. Exams are exempted and can be handwritten. There is an initial start-up cost and a learning curve, but it will pay off in the longer run.

Another common way to use LaTeX is Overleaf https://www.overleaf.com/. It provides an online, web-based editor for LaTeX and it compiles LaTeX on the cloud. It is most often used in a group setting when several people are working on the same document. It tracks changes and handles versioning automatically. You also do not have to download anything to your computer. https://www.overleaf.com/learn/latex/Learn_LaTeX_in_30_minutes

R, Python

The Top-3 most used programming languages in data science are Python, R and SQL, according to a Kaggle Machine Learning & Data Science Survey.[1]

R is an open-source programming language for statistical computing. I use RStudio as the front-end graphical user interface for R: https://www.rstudio.com/

It has many advantages over Stata and SAS. In comparison to Python, R is a more specialized language used primarily in statistical and econometric analyses. For example, I use R for cleaning up data, exploratory descriptive statistics, regression analysis, and other econometric/statistical estimation procedures.

Python is more comparable to Matlab. It is a more general purpose programming language that is used beyond scientific and statistical computing. My typical workflow for research: SQL and R for cleaning data and initial analyses. If required, I use Python for more advanced computing such as numerical simulations, large-scale parallel computing using GPU, web scraping, machine learning tools analyzing unstructured data such as text, audio and video.

One of the easiest ways to install Python is through anaconda: https://www.anaconda.com/products/individual. It comes with Jupyter Notebook, which is my go-to front-end for writing and executing Python scripts.

[1] https://www.kaggle.com/sudhirnl7/data-science-survey-2018

Another easy way to run Python is through Google Colab. It is a free Jupyter Notebook environment that runs in the cloud and stores its notebooks on Google Drive. Since it online web-based, you do not have to install anything on your computer.

Although I feel you should learn both, for the purpose of this course, you can choose either R or Python.

Mathematica

Occasionally, we will use Mathematica. It is most useful for quickly visualizing functions and working through excessively tedious algebra.

It can be installed for free as a UTD student: https://oit.utdallas.edu/howto/mathematica/

Schedule

The schedule is tentative and subject to change.

Week 1: Basic probability theory (Casella-Berger, chapter 1): sample spaces, event spaces, probability spaces, random variables, probability density functions, cumulative density functions.

Week 2: Transformations and expectations of random variables (Casella-Berger, chapters 2): change of variables, probability integral transformation, moments, expectations.

Week 3: Multivariate random variables (random vectors) (Casella-Berger, chapter 4): joint and marginal distributions, conditional distributions, independence of random variables; covariance and correlation, bivariate transformations

Week 4: Common families of distributions (Casella-Berger, chapters 3): Multivariate Normal distribution, Gamma distribution, truncated random variables, probability inequalities.

Week 5: Properties of a random sample (Casella-Berger, chapter 5): sampling distributions, order statistics, unbiasedness and consistency, convergence concepts (convergence in probability, convergence almost surely, convergence in distribution).

Week 6: Point estimation (Casella-Berger, chapter 5): large sample theory, central limit theorem, delta method, big o notation, asymptotic variance

Week 7: Point estimation (Casella-Berger, chapter 7): Method of Moments estimator, Generalized Method of Moments, Maximum Likelihood Estimation

Week 8: (Point estimation (Casella-Berger, chapter 7): properties of the Maximum Likelihood estimator, loss functions, mean square error, Fisher's information, Probit models.

Week 9: (Casella-Berger, chapter 7. Cameron-Trivedi, Chapter 13): Bayesian methods versus Frequentist. Conjugate prior. Bayes Theorem. Bayesian estimation.

Week 10: (Casella-Berger, chapter 8): Hypothesis testing, Likelihood Ratio test, Wald's t-test, Type-1 and Type-2 errors.

Week 11: (Casella-Berger, chapter 8): Lagrange-multiplier test. Size and power of a test. p-values. Neyman-Pearson lemma. Asymptotic distribution of test statistics.

Week 12: (Casella-Berger, chapter 9): Confidence interval as inversion of test statistics. Coverage probabilities. Bayesian intervals.

Week 13: (Casella-Berger, appendix): Data-resampling and simulation techniques. Bootstrapping. Importance sampling. Monte Carlo sampling and integration. Non-parametric methods. Kernel estimators. k-nearest neighbors.

Week 14: (Davidson-Mackinnon, chapters 2-3): Statistical properties of linear regressions. Matrix notation. Bias and consistency. Frisch-Waugh-Lovell Theorem. The geometry of OLS estimation. Multicollinearity.

Week 15: (Davidson-Mackinnon, chapters 2-5): OLS covariance matrix. Hypothesis testing and inference involving OLS estimators. Heteroskedasticity consistent covariance matrix estimator. Serial correlation.

Optional class 1: Causal inference methods. Average treatment effect. Propensity score matching. Difference-in-difference methods.

Optional class 2: Discrete-choice models. Probit and Logit models. Binary and multinomial models. Time-series and panel data. Testing for serial correlations. Random and fixed effects.

Overlap with other courses

The first part of this course (the *probability* part of this course) overlaps with OPRE 7310 Probability and Stochastic Processes. The treatment of probability in OPRE 7310 is more rigorous and measure-theoretic than the treatment here.

COVID-19 Guidelines and Resources

The information contained in the following link lists the University?s COVID-19 resources for students and instructors of record.

Please see http://go.utdallas.edu/syllabus-policies.