Probability and Statistics for data analysis

Ioannis Vrontos, Associate Professor, Department of Statistics, AUEB,

Office: Spetson and Troias 2, 4th floor, office 412, Tel: +30-210-8203927, Email: vrontos@aueb.gr

Overview

This course will provide the fundamental theory in probability and statistics, which is necessary in basic research and in data analysis tasks. This graduate level course will focus in generalizing the basic undergraduate knowledge of probability and statistics, providing the underline theory and implementation using the statistical package R.

Key Outcomes

By completing the course the students will be able to:

- Know the basic concepts in probability.
- Learn the fundamentals in statistical inferenceallowing them to understand which type of analysis is necessary and how it can be correctly implemented.
- Learn about the theory and the accurate practice of regression analysis.
- Learn about the Bayesian approach, know how to apply it in practice, and understand thedifferences from the respective classical (frequentist) methods.

Requirements and Prerequisites

The students should have a basic quantitative background. Specifically, basic knowledge in the fields of calculus, probability and statistics will be necessary for this course.

Bibliography

1. **Probability and Statistics:**
    - “Statistical Inference” by G. Casella and R.L. Berger, 2nd edition, Duxbury Press, 2001.
2. **Linear Regression**
    - “Applied Linear Regression”, by S. Weisberg, 3rd edition Wiley, 2005.
    - An R Companion to Applied Regression, by J. Fox and S. Weisberg, 2nd edition,SAGE Publications Inc, 2011.
3. **Bayesian Statistics:**
    - “Bayesian Methods for Data Analysis”, by B.P. Carlin and T.A. Louis, 3rd edition,Chapman and Hall/CRC, 2008.
    - “Introduction to Bayesian Statistics”, by P. Dellaportas and P. Tsiamyrtzs, lecture notes, AUEBpublishing.

Software/Computing requirements

The computational aspects of this course will be implemented exclusively in R, a free software environment for statistical computing and graphics. R can be downloaded at https://www.r-project.org and installed on all types of environments (Windows, Mac, UNIX). We will make use of the data sets available in the "datasets" library of R. Other data sets that will be used in class and/or assignments they will be available in the eclass page of this course.

Grading

There will be a total of 3 homework assignments (given approximately at the units 3, 6 and 10) that will contribute 30% in the final grade. The remaining 70% will be determined by the in class final exam. Please note that one needs to write at least 5 (out of 10) in the final exam (independently of the grades in the homework assignments) not to fail the course.

Course Syllabus

The course comprises of ten units of three hours each.

Unit 1: Introduction and Probability

The role of statistics in big data will be given and certain case studies will illustrate its necessity. Basic concepts in probability will be presented: axiomatic definition and interpretations, calculus of probabilities, counting, conditional probabilities, independence, law of total probability and Bayes theorem.

Unit 2: Univariate Random Variables

We will cover the basic theory in probability mass/density functions, cumulative distribution functions and moments.Various cases of discrete random variables (e.g.Binomial, Poisson, Hypergeometric etc.) along with severalcontinuous random variables (e.g. Gamma, Normal, Beta, etc.) will be given with their functional forms and properties,along with a series of examples and applications.

Unit 3: Multivariate Random Variables

Moving from single to many recorded random variables, we will generalize the definitions from the univariate to the multivariate case. Furthermore, we will study the topics of joint, marginal and conditional distributions, independence,covariance/correlation and hierarchical modeling.

Unit 4: Properties of Random Sample

We will provide the basic concepts for a random sample and its order statistics. Convergence concepts will be given with emphasis on the law of large numbers and central limit theorem.

Unit 5: Statistical Inference

The Sufficiency principle in statistics will be explored and its importance in big data will be utilized.Methods of obtaining point estimates,with emphasis to maximum likelihood estimation, will be given along with the basic principles of evaluating them. Point estimation in mean, variance, proportion etc. will illustrate the proposed theory.

Unit 6: Statistical Inference

Thebasic theory and interpretation of confidence interval (CI) and hypothesis testing (HT)will be studied. CI and HT for the mean, variance, proportion etc. will illustrate the respective theory.

Unit 7: Statistical Inference

In this unit we will generalize inference moving from the single to two populations and also present the analysis of variance (ANOVA) methodology. Multiple comparison issues will be treated and the basic principles in statistical modeling (Regression/GLM) will be outlined.

Unit 8 – 10: Regression

We will provide the basic theory of regression modeling insimple/multiple linear regression and give a brief introduction to generalized linear models. We will also emphasize the accurate implementation of the theory, in examining assumptions, checking the diagnostics and correctly interpreting the results. Issues like multicollinearity, dummy variables and model selection will be examined while methods like Lasso will be presented.

Participation

In-class contribution is a significant part of your grade and an important part of our shared learning experience. Your active participation helps us to evaluate your overall performance. You can excel in this area if you come to class on time and contribute to the course by:

- Providing strong evidence of having thought through the material.
- Advancing the discussion by contributing insightful comments and questions.
- Listening attentively in class.
- Demonstrating interest in your peers' comments, questions, and presentations.
- Giving constructive feedback to your peers when appropriate.

*Please*arrive to class on time and stay to the end of the class period. Chronically arriving late or leaving class early is unprofessional and disruptive to the entire class. Repeated tardiness will have an impact on your grade.

Turn off all electronic devices prior to the start of class. Cell phones tablets and other electronic devices are a distraction to everyone.

Assignments

Late assignments will either not be accepted or will incur a grade penalty unless due to documented serious illness or family emergency. Exceptions to this policy for reasons of civic obligations will only be made available when the assignment cannot reasonably be completed prior to the due date, you make suitable arrangements, and give notice for late submission in advance.

Attendance Requirements

Class attendance is essential to succeed in this course and is part of your grade. An excused absence can only be granted in cases of serious illness or grave family emergencies and must be documented. Job interviews and incompatible travel plans are considered unexcused absences. Where possible, please notify the instructor in advance of an excused absence.

Students are responsible for keeping up with the course material, including lectures, from the first day of this class, forward. It is the student's obligation to bring oneself up to date on any missed coursework.

Code of Ethics

Students may not work together on individual graded assignments unless the instructor gives express permission.

Exercise integrity in all aspects of one's academic work including, but not limited to, the preparation and completion of all other course requirements by not engaging in any method or means that provides an unfair advantage. In any case of doubt, students must be able to prove that they are the sole authors of their work by demonstrating their knowledge to the instructor.

Clearly acknowledge the work and efforts of others when submitting written work as one’s own. Ideas, data, direct quotations (which should be designated with quotation marks), paraphrasing, creative expression, or any other incorporation of the work of others should be fully referenced. No plagiarism of any sort will be tolerated. This includes any material found on the internet. Reuse of material found in question and answer forums, code repositories, other lecture sites, etc., is unacceptable. You may use online material to deepen your understanding of a concept, not for finding answers.

Please report observed violations of this policy. Any violations will incur a fail grade at the course and reporting to the senate for further disciplinary action.