11
Statistical Methods in Data Analytics

Spring 2018

Class Meetings: Friday 4:00 pm - 7:00 pm. @ HBS – 315

Course Instructor

Prof. Ioannis Pavlidis (correspond at ipavlidis@uh.edu)

Office Hour: 3-4 pm on Fridays @ HBS-306

Course TA

George Panagopoulos (correspond at gpanagopoulos@uh.edu)

Office Hour: 1-2 pm on Wednesdays, Thursdays @ HBS-306

Course Description

The course covers statistical methods in human and technology studies or experiments, from where the bulk of scientific and engineering data originate. The course starts with a contrast of hypothesis-driven research supported by statistical inference versus rigorous deduction based on first principles; this is in order to delineate the current from the past mode of science and engineering, motivating the subject. **Then, it proceeds in a step-wise manner building the student's background in the statistical tools of the trade, without which an MS thesis or PhD dissertation cannot be complete.** The course culminates with a series of sessions on experimental design, which is the cornerstone of a successful research project or industrial product.

Although the introduction and methodological sections of scientific papers differ from discipline to discipline (e.g., algorithms vs. assays), the results sections of papers should conform to a universal pattern, according to currently accepted best practices. The produced data should be derived according to appropriate study/experimental designs and should be subjected to relevant statistical tests. There is no such thing as statistics for computer scientists or statistics for biologists; statistics is the same for everybody. However, certain disciplines tend to use some tools more than others, and instruction needs

to be tailored according to the differing educational backgrounds. In computer science in particular, awakening to standard analysis of study/experimental results has been slow; most of this analysis used to be carried out heuristically. This has changed the last few years and several computer science disciplines have already adopted statistical methods as the standard in results analysis, while others are bound to follow sooner or later. Among the computer science communities that are at the forefront of this movement are the Human-Computer Interaction and Computer Vision communities. The Statistical Methods course aims to cover this need and is paced taking into account the typical background of graduate students in computer science. It is very practical in its orientation (no proofs), emphasizing the understanding of concepts and the ability to apply the right design or test to the right problem.

The main part of the course starts with the delineation between continuous and discrete variables and the enormous implication that this carries for the selection of tests. Then, it proceeds with the description of distributions, probabilities and error types that are fundamental to the construction of the *t*-tests, ANOVA tests, and non-parametric tests. In the second stage, the course visits regression in its various forms, completing the coverage of significance and association tests used in almost all scientific papers. **Emphasis is placed on multiple regression and linear modeling – a powerful and elegant method to examine the effect of multiple factors in a research problem; it is heavily used nowadays in MS and PhD research.** The treatment of symbolic data and the tools of last resort, that is, non-parametric methods complete the course's first part. In the course's second part, we visit the various experimental designs, including new methods, the so-called *Mixed Methods*. Before start analyzing data, one needs to know according to which principle to collect these data in order to address her/his hypothesis; for this, s/he needs to pick the right design. Even an impeccable testing will not save the day if the researcher picked the wrong study or experimental design (garbage in – garbage out). Hence, the student acquires towards the end of the course a 30,000 feet view of the scientific and engineering process, solidifying her/his ability to design, collect, and test.

The course has five homework assignments to reinforce the understanding of the concepts and methods. In the place of a final exam, the course has a semester long-project, where a problem is defined for the class, and then each group of students is required to come up with a study design, collect/quality control data, and perform tests, putting everything in the form of a term paper.

The students need to know R in order to process and plot the data. R is becoming one of the most useful tools for computer scientists in the data analytics business. We provide the students with online educational material and organize an R tutorial class.

Grade Plan

Module	Individual Weight	Total
Participation	10	10
Homework (x5)	10	50
Project	40	40
Total		100

Course Outline

Lesson 1: Data and Statistics 1/19/2018

Introduction; observations and variables; types of measurements for variables; distributions; numerical descriptive statistics; exploratory data analysis; bivariate data; data collection

Lesson 2: Probabilities and Sampling Distributions 1/26/2018

Probability; discrete probability distributions; continuous probability distributions; sampling distributions

Homework #1 Out

Lesson 3: Principles of Inference 2/2/2018

Hypothesis testing; estimation; sample size; assumptions

Assignment of Projects

Lesson 4: Inferences on a Single Population 2/9/2018

Inferences on the population mean; inferences on a proportion; inferences on the variance of one population; assumptions

Homework #1 Due

Homework #2 Out

Lesson 5: Inferences for Two Populations 2/16/2018

Inferences on the difference between means using independent samples; inferences on variances; inferences on means for dependent samples; inferences on proportions; assumptions

Lesson 6: Inferences for Two or More Means 2/23/2018

Analysis of variance; linear model; assumptions; specific comparisons; random models; unequal sample sizes; analysis of means

Homework #2 Due

Homework #3 Out

Lesson 7: Linear Regression 3/2/2018

The regression model; estimation of parameters; inferences for regression; correlation; regression diagnostics

Lesson 8: Multiple Regression 3/9/2018

The multiple regression model; estimation of coefficients; inferential procedures; correlations; special models; multicollinearity; variable selection; detection of outliers

Lesson 9: Other Linear Models 3/23/2018

The dummy variable model; unbalanced data; models with dummy and interval variables; weighted least squares; correlated errors

Homework #3 Due

Homework #4 Out

Lesson 10: Categorical Data 3/30/2018

Hypothesis test for a multinomial population; goodness of fit; contingency tables; loglinear model

Lesson 11: Nonparametric Methods 4/6/2018

One sample; two independent samples; more than two samples; rank correlation; the bootstrap

Homework #4 Due

Homework #5 Out

Lesson 12: Comparative and Single Factor Experiments 4/13/2018

Randomized designs; paired comparison designs; fixed effects model; random effects model

Lesson 13: Randomized Blocks, Latin Squares, and Related Designs 4/20/2018

Randomized complete block design; Latin square design; Greco-Latin square design; balanced incomplete block designs

Lesson 14: Factorial Designs 4/27/2018

Two-factor factorial design; general factorial design; fitting response curves and surfaces; blocking in factorial design

Homework #5 Due

Lesson 15: Project Presentations 5/4/2018

Project Reports Due

References

[1] Boddy, R. and Smith, G. *Statistical Methods in Practice for Scientists and Technologists.* Wiley, 2009.

[2] Freund, R. J., W. J. Wilson, and D. L. Mohr. *Statistical Methods.* 2010.

[3] Hinkelmann, Klaus, ed. *Design and Analysis of Experiments, Special Designs and Applications*. Vol. 3. John Wiley & Sons, 2011.

[4] Friedman, Lawrence M., Curt Furberg, and David L. DeMets. *Fundamentals of clinical trials*. Vol. 4. New York: Springer, 2010.