Statistics 431: Statistical Inference

Syllabus, Summer 2012

Classes:	Mon/Tues/Wed/Thur 10:45a.m.–12:15 p.m., in F45 JMHH
Instructor:	Hyunseung Kang
Email:	khyuns@wharton.upenn.edu
Office:	434 JMHH
Office hours:	Mon/Tues/Wed/Thur 12:15 p.m. – 1:30pm
Course website:	http://stat.wharton.upenn.edu/∼khyuns/stat431/

Course overview

The course is aimed to equip students with the tools needed to analyze real-world data and to justify their use through theory. Together, we will study basic concepts related to statistical inference and examine commonly used methods, with an emphasis on understanding when and how to apply them. Students wil also learn how use these methods on the statistical software `R`.

Prerequisites

The official prerequisite of the course is STAT 430. The effective prerequisite is *fluency* with basic probabilistic reasoning and analysis (e.g., probability distributions and densities; joint distributions; conditional probability, independence, correlation, and covariance; moment generating functions; law of large numbers; central limit theorem; etc.). For a refresher/overview of these topics, please refer to *A First Course in Probability* by Sheldon Ross.

It would be helpful to have previous exposure to linear algebra, but it is *not required*. Previous exposure to the statistical computing software `R` is also *not required*.

Textbook

There is no required textbook for this course. All course material will largely consist of taking the best parts of each textbook listed below and presented through lecture and lecture notes. However, if you wish to purchase a textbook, Devore is available at the Bookstore.

1. *(Recommended) Probability and Statistics for Engineering and the Sciences,* 8th *Ed.*, J. Devore, Brooks/Cole - Cengage Learning, 2011.

2. *(Recommended) Regression Analysis by Example,* 4th *Ed.*, S. Chatterjee and A. Hadi, Wiley-Interscience, 2006.

3. *(Recommended) Statistics and Data Analysis: from Elementary to Intermediate*, A. C. Tamhane and D. D. Dunlop, Prentice Hall, 2000.

All the textbooks are on reserve at the Lippincott Library in Van Pelt.

Statistical computing software

The statistical computing software `R` (latest version) will be used in the course. It is free, and can be downloaded at the `R`-project website:

`http://www.r-project.org/`.

The above website also contains a list of manuals for using the software. Basic usage of `R` will be illustrated in class and through sample codes posted on the course website. Again, *no previous exposure to the software is required.*

Grading policy

- Assignments (25%), *due every Monday before class begins!*
- Weekly quizzes (35%) , *every Monday at the beginning of class*
- Final project (40%), *due Thursday, August 9th*

Assignments will be handed out every Monday and will be due the following Monday *before class begins*. Weekly quizzes will be be given every Monday at the beginning of class. They will be 15 minutes long and will be based on the previous week's lectures and assignment.

Final Project (Due Thur, August 9th)

In the final project, students will analyze a real-world data set of their choosing using the tools learned from the class. The final project should focus on what statistical tools were used, whether the tools were appropriate in the setting, and why the tools were important in the analysis. Students may also develop new tools for analysis, as long as it is justified by theory.

Students may work in groups up to three people. Each group will submit a one-page, single-spaced, 12-point type, 1-inch margin, executive summary providing an overview of the project. Also, the group will submit a technical report containing the details of the group's analysis. Both documents must be in a single PDF file (no .txt, .doc, .docx, .tex, etc.). In the technical report, students are expected to provide some mathematical justification of their analysis and include relevant numerical analysis (e.g. p-values, t-tests, F-tests, etc.) of the data set.

If students are interested and if the quality of the analysis is exceptional, your instructor will help you get the final project published in an academic journal.

Lecture Schedule

1. Population and Sample ($\approx$ one lecture): Summarizing data, estimation of mean and variance, unbiased estimators, and risk
2. Sampling Distribution ($\approx$ three lectures):
   (a) Derivation of distributions for the mean and the variance
   (b) Introduction to the chi-square distribution, F-distribution, and the t distribution.

3. One-Sample and Two-Sample Hypothesis Testing ($\approx$ four lectures):
   (a) Hypothesis testing for $H_0 : \mu = 0$ vs $H_a : \mu \neq 0$ for one-sample, two-sample, and paired samples.
   (b) Introduction to confidence intervals, derivation of confidence intervals, and illustration of statistical power.
   (c) 2-by-2 Factorial design and count data: Chi-square test for independence
4. Regression ($\approx$ 12+ lectures):
   (a) Simple linear regression: $y_i = \beta_0 + \beta_1 x_i$
   (b) Multiple linear regression: ANOVA, MANOVA, ANCOVA, polynomial regression, weighted least squares regression
   (c) Generalized linear models (GLMs): Logistic regression, logit regression, probit regression, and Poission regression
   (d) Time series models: AR and ARMA models
   (e) Model diagnostics and model selection procedures: Forward, Backward, Stepwise, Lasso, Ridge, AIC, BIC, and Mallow's Cp
5. Additional Topics ($\approx$ four lectures):
   (a) Nonparametric regresssion: Kernel methods, moving average processes, and b-Splines
   (b) Nonparametric inference: Permutation Test, Welch's Test, Signed-Rank Test, Kolmogorov-Smirnov Test
   (c) Bootstrapping, Bayesian inference, and other computational procedures for inference
   (d) Likelihood-based inference: Maximum likelihood estimators (MLEs) and inference for MLEs
   (e) Multivariate methods (if time permits): PCA,CCA, and SVD