STA248H1: Statistics for Computer Scientists

Course Instructor: Wei Q. Deng (wei.deng@mail.utoronto.ca)

July-August 2018

Course Description

A survey of statistical methodology with emphasis on data analysis and applications. The topics covered include descriptive statistics, data collection and the design of experiments, univariate and multivariate design, tests of significance and confidence intervals, concept of power, multiple regression. Students learn to use a statistical computer package as part of the course (Note: STA248H1 does not count as a distribution requirement course).

Programming languages

Both **R** and **Python** are acceptable computing languages in this course, we will see use of both throughout the course. Make sure you master at least one of them by end of this course. You will complete at least two of the specified DataCamp Programming modules as part of this course.

Course Website

Everything you need for the course is hosted here. Please check frequently for posting of lecture slides, practice problems and test information.

Textbook and learning resources

There is no official textbook for this course, but if you wish to have additional books/materials to look at, I suggest the following (all available through U of T library either as a physical copy or electronically).

- *Open Intro Statistics*, **3**rd edition available for download here
- Chapter 2-3 can be used as review for STA247
- topics in Chapter 1,4,5,7 and some of 8 are covered in this course
- *Probability and Statistics for Engineering and the Sciences* **9**th Edition by Jay L. Devore

Getting help

Please see website for schedule of office hours. Also, it is not a bad idea to google - and learning to use it effectively will become one of your most valuable skills in your future study/work.

Note that we will have additional office hours before midterm and final exam. If you wish to meet/discussion outside of office hours please schedule with me at least one day in advance.

IMPORTANT: please include in the subject head the course code "STA248" whenever you are emailing regarding this course.

Evaluation

- Completion of two R/Python programming modules 10% (5% each)
- Midterm 40%
- Final Exam 50%

Course content

Lecture 1: Welcome and introduction to statistics (July 4th, Wednesday)

- First half: course overview and connection with STA247 and STA302
- Second half: summarizing data (graphically and through summary statistics)
- Connection between statistics and parameters (ideas of a statistical model); population and sample.
- descriptive statistics (idea of a centre/location, spread/scale of a distribution)
- commonly used graphs (bar, histogram, boxplot, scatterplot, ecdf, qqplot)
- graphical representation and information from the graphs (outliers, model assumptions, dependence)
- data transformation

Lecture 2: Estimation - to think with data (July 9th, Monday)

- properties of estimators derived from sample
- unbiased, consistency
- method of moments
- maximum likelihood estimates
- sampling distribution, standard error

Lecture 3: Point estimation and interval estimation (July 11th, Wednesday)

- Second half: idea of inferential learning; bootstrap sampling
- confidence intervals (CI)
- parametric and non-parametric bootstrap
- CI for μ and σ^2 (normal, t-distribution, or binomial distributions), what about the median?
- bootstrap CIs.

Lecture 4: Hypothesis Tests and statistical significance (July 16th, Monday)

- formalize ideas of testing and inference
- tests of significance
- one-sided vs. two sided test
- confidence intervals
- power
- student's t-test (one sample, two sample, and paired)
- test of mean equality
- assumptions
- transformation

Lecture 5: Testing: Univariate and multivariate design -II (July 18th, Wednesday)

- analysis of variance (one-way)
- F-test in two samples
- K-samples
- know how to construct an one-way ANOVA table
- Levene's test (general idea for test of equality)
- generalize the idea of F-test as a test for goodness of fit.
- interactions (two-way - not tested)

Lecture 6: Midterm test (July 23rd, Monday)

Location to be announced the week before.

Lecture 7: Simple and multiple regression (July 25th, Wednesday)

- evidence for linear relationship
- correlation coefficients
- scatterplot and boxplot (when predictor variable is continuous or discrete)
- a linear model
- a simple linear model
- model assumptions
- idea of least squared method
- measure of fit

Lecture 8: Midterm review (July 30th, Monday)

Tianle subbing in

Lecture 9: Supervised learning with regression (August 1st, Wednesday)

Tianle subbing in

- linear regression as a supervised learning alogrithm

No lecture on August 6th

Lecture 10: More on linear regression (August 8th, Wednesday)

- extending to multiple variables
- collinearity
- feature selection
- measure of fit

Lecture 11: Makeup lecture/Final review (August 13th, Monday)