# Introduction to Statistics (S1201D)

Wenda Zhou

May 22, 2017

## 1 Description

This course is designed for students who desire a strong grounding in statistical concepts with some mathematical rigor. We will be covering basics of probability followed by a selection of topics in statistics covering prediction, estimation and testing. We will also cover how to make use of these statistical methods through software – specifically the R programming language. No previous programming experience is expected or required.

## 2 Books

Lecture notes will be provided covering the topics discussed in class. We may also refer to the following books:

- *An Introduction to Statistical Learning*, James, Witten, Hastie and Tibshirani. Available freely online at `http://www-bcf.usc.edu/~gareth/ISL/`
- *OpenIntro Statitics*, Diez, Barr and Cetinkaya-Rundel. Available freely online at `https://www.openintro.org/stat/textbook.php`.

## 3 Software

We will be using the R programming language (`https://cran.r-project.org/`). It is recommended to also use RStudio Desktop (`https://www.rstudio.com/products/rstudio/`, freely available on Windows, Mac and Linux) as your development environment.

## 4 Topics

The topics below are listed in no particular order and do not necessarily reflect the intended course structure (only its contents).

**Introduction to data and visualization** Overview of descriptive statistics – expectation, variance, covariance and correlations. Statistical visualizations and graphics.

**Introduction to probability** Axioms of probability. Discrete and continuous random variables – common distributions. Conditional probability and expectation. Joint and multivariate distributions.

**Estimation** Maximum likelihood estimation. Method of moments. Confidence intervals. Bias, variance and mean-squared error of estimators. Unbiased estimators. Linear regression and ordinary least squares. Introduction to generalized linear models.

**Hypothesis testing** Definition of a statistical test. P-value and significance. Testing of a binomial proportion. Chi-square test of contingency table. Testing in linear models. ANOVA.

**Prediction** Prediction risk. Training and testing risk. Cross-validation. Introduction to machine learning.

**Advanced statistical topics** Overview and discussion of statistical challenges that arise in practice. Spatial and temporal data. Causality. Survey sampling. Missing data.

# 5 Courseworks and grading

The grade for the course will consist of four parts with the following weights.

**15%** : Homework assignments

**15%** : Project

**30%** : Midterm exam

**40%** : Final exam

## 5.1 Homework assignments

There will be about 10 homework assignments (2 per week), due on Tuesday and Thursday. The homeworks due Tuesday will be mostly theoretical and use paper submission. The homeworks due Thursdays will be mostly practical and software based and will use online submission.

Collaboration is allowed for homework but the final submission must be your own work.

## 5.2 Project

The project will be a somewhat larger homework assignment where you will be asked to analyze one of a few given datasets with specific directions.

## 5.3 Midterm exam

The midterm exam will be held in class on June 5th.

## 5.4 Final exam

The final exam will be held in class on June 29th.

# 6 Student conduct and academic integrity

All participants of the class are expected to abide by the columbia code of conduct which can be found at http://studentconduct.columbia.edu/. I would like to draw particular attention to the acadamic integrity section.