Introduction to Statistical Genomics

GENOME 560, Spring 2012

Course Overview

The data intensive nature of the 21st century biology made it very important for scientists to have a basic proficiency in statistics, as technological advances now allow complex and high dimensional datasets to be routinely collected. Whether it is thousands of gene expression levels that have been measured by microarrays, millions of polymorphisms that have been genotyped for a case control study of a disease phenotype, or more general questions of how to properly design an experiment, you will constantly be confronted with how to collect, analyze and interpret data through your research careers.

This course provides the key statistical concepts and methods necessary for extracting biological insights from these types of datasets. As this is only a five-week course, we will not be able to cover every specific topic that might arise in the course of your research. Thus, we will focus on **rigorous understanding of fundamental concepts** that will provide you with the tools necessary to address routine statistical analyses and the **foundation to understand and learn more specialized topics**.

Throughout this course, we will often make use of the freely available statistical software R (available at http://www.r-project.org/). R has become one of the most widely used platforms for statistical analysis in genomics, because it is powerful, easy to share code, and makes publications quality graphics. Problem sets will require the use of statistical software, and while you are free to use whatever you feel comfortable with (such as MatLab, SAS, STATA, or perhaps even Microsoft Excel), I highly encourage you to use R.

Learning Goals

The primary objective of this course is to provide a strong foundation into fundamental statistical concepts, particularly as they relate to genomics, and thus better prepare you for a successful scientific career.

Required Texts and Websites

There is no required textbook for this course. Each lecture will be accompanied by a handout that covers all of the in class material. There are also several excellent introductory statistics resources available on-line. The following websites contain on-line textbooks that cover all of the material presented in this class.

1. http://www.math.wm.edu/~trosset/Courses/351/book.pdf
2. http://www.statsoft.com/textbook/stathome.html
3. http://www.stat.berkeley.edu/~stark/SticiGui/Text/toc.htm

For students who would like to have general reference book, I recommend:

1. Probability and Statistics for Engineering and the Scientists 6th Ed. Jay L. Devore (2004). Duxbury press, Thompson-Brooks/Cole.

2. Statistical Inference. Casella, G. and Berger, R. L. (1990). Wadsworth, Belmont, CA.
3. Probabilistic Graphical Models: Principles and Techniques. Koller, D. and Friedman, N. (2009). MIT Press.

The first textbook is fairly extensive in its coverage, but does not focus on equations. The second text book covers much of the same material, but in a quantitatively more rigorous, mathematical, and theoretical manor. The third textbook primarily focuses on learning and inference in probabilistic graphical models.

Grading

Grades will be based on five problem sets that will each comprise 20% of your total grade. Problem sets will be distributed on Thursday and due the following Thursday at the start of class. Late assignments will not be accepted in general. It will be only accepted in exceptional circumstances or if prior arrangements have been made with the instructor.

Class Meetings

Class meets twice a week – Tue/Thu 9:00-10:20am in Foege S110.

Each class will last for 80 minutes and be primarily lecture based, but will include other forms of learning and interactions. In particular, we will often interrupt lectures to work on problems in small groups as well as work through statistical analyses using the freely available software R.

Instructor

Professor: Su-In Lee
Office: Paul Allen Center 536
Office hours: Wed 9-10am
Phone: (206) 685-1418
Email: suinlee@cs.washington.edu

Course Schedule (Tentative)

1. May 1: **Descriptive statistics**
   Key concepts: averages, standard deviation, percentiles

2. May 3: **Random variables and probability theories**
   Key concepts: random variables, expectations, joint probability, conditional probability

3. May 8: **Distributions**
   Key concepts: probability density functions, discrete distributions, sampling distributions

4. May 10: **Parameter estimation**
   Key concepts: maximum likelihood estimation, point estimation, Bayesian estimation, confidence intervals

5. May 15: **Regression methods**
   Key concepts: univariate and multivariate linear regression

6. May 17: **Hypothesis testing I –** t-test
   Key concepts: confidence intervals, t-test, p-values

7. May 22: **Hypothesis testing II –** ANOVA
   Key concepts: single factor ANOVA

8. May 24: **Hypothesis testing III – categorical data**
   Key concepts: contingency tables, Chi-square tests, Fisher exact tests

9. May 29: **Bootstrapping, cross validation and permutation tests**

10. May 31: **Assessing significance in high dimensional experiments**
    Key concepts: multiple hypothesis testing, false discovery rates, q-values

Special topics that may be discussed in class include Bayesian networks, Expectation Maximization (EM) algorithm, principal component analysis (PCA), and singular value decomposition (SVD).