Genome 560: Introduction to Statistical Genomics

Course overview and curriculum content

> *"There are very few things which we know, which are not capable of being reduc'd to a Mathematical Reasoning, ... and where a Mathematical Reasoning can be had, it's as great folly to make use of any other, as to grope for a thing in the dark when you have a Candle standing by you"*
>
> John Arburhnot (1692)

Statistical concepts and principles permeate all aspects of science ranging from designing experiments, analyzing data, testing hypotheses, reporting results, and interpreting the research literature. Although this is true for all scientific disciplines, the need for a basic proficiency in statistics has become even more important in genetics and genomics 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 throughout your research careers.

This course covers 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 a 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. We begin by considering what data is, how to describe it, and good practices for visualizing data and presenting it to others (as well as common pitfalls in visualizing data). Particular emphasis will be placed on critical aspects underlying good experimental design, and how easy erroneous inferences can be made when experiments are designed poorly. Next we will explore fundamental concepts in probability, including what random variables are, why they are the foundation of statistical inference, and how they are distributed. The next few weeks will focus on what many of you may consider the backbone of statistics: estimating parameters and testing hypothesis. Although the key statistical tests often used in research will be covered (t-tests, chi-square tests, ANOVA), we will not focus on cataloging every statistical method that has ever been developed. Not only is this impossible to do in a short course, but more importantly, I do not want to leave you with the impression that statistics is a cookbook with "recipes" to follow. Rather, parameter estimation and hypothesis testing is a dynamic branch of statistics in and of itself, with new methods constantly being developed. However, they are all based on a handful of concepts (such as likelihood), which we will focus on. A thorough understanding of these basic principles will enable you to confidently and correctly analyze your data, and know when to either learn more or seek additional assistance. Finally, we will discuss the multiple testing problem and solutions for dealing with it, which has become an acute issue in genetics and genomics.

Throughout this course, we will often make use of the freely available statistical software R (available at http://cran.r-project.org/). R has become one of the most widely used platforms for statistical analysis in genetics and genomics, because it is powerful, easy to share code, and makes publication quality graphics (the free part helps too). Problem sets will require the use of statistical software, and while you are free to use whatever you feel comfortable with (such as SAS, STATA, or perhaps even Microsoft Excel) I highly encourage you to use R. During the first lecture, I will provide a tutorial on how to install and get started with R.

Learning Goals and Objectives

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

Specific goals are:

- To appreciate the stochastic nature of biological phenomena encourage students to "***think probabilistically***."
- To identify elements of ***good experimental design***.
- To ***summarize and visualize*** data.
- ***Choose*** statistical methods that are appropriate to the scientific question of interest and to the specific features of the available data.
- To ***estimate parameters and quantify uncertainty***, such as estimating the mean and confidence interval of allele frequencies.
- To ***assess statistical significance*** when multiple hypothesis tests are performed, such as in the analysis of differential gene expression measured by mircroarrays.
- To ***critically*** evaluate and interpret statistical methods used in primary research articles.
- ***Implement*** a variety of analytical tasks on realistic datasets, using freely available software and widely available hardware.

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, as well as more in-depth material that is beyond the scope of this course, but you may find useful nonetheless. In addition, there are 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 a general reference book, I recommend:

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

2. Statistical Inference. Casella, G. and Berger, R. L. (1990). Wadsworth, Belmont, CA.
The first textbook is fairly extensive in its coverage, but does not focus on equations. The second textbook is covers much of the same material, but in a quantitatively more rigorous, mathematical, and theoretical manor.

Grading
The best way to learn statistics is to ***do*** statistics. Therefore, grades will be based on five problem sets that will each comprise 20% of your total grade. Problem sets will be distributed on Thursdays and due the following Tuesday at the start of class. Late assignments will not be accepted, accepted in exceptional circumstances or if prior arrangements have been made with the instructor.

Expectations for Success
I intend for this course to be a challenging introduction to statistical methods and concepts that scientists in Genetics and Genomics frequently encounter. However, the key word here is "challenging" not impossible. Although statistics in particular, and quantitative courses in general, sometimes invoke fear, the elements for success are simple: attend lectures, always ask questions when you have them, and complete all assigned problem sets.

Characteristics of Class Meetings
Class meets twice a week. Each class will last for 80 minutes and be primarily lecture based, but will include other forms of learning and interaction. In particular, the course has been organized to stimulate student discussion and interaction, and 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.

Course Schedule:

April 1: **Descriptive statistics and visualizing data**
Key concepts: histograms, scatter plots, averages, standard deviation, percentiles

April 3: **Collecting Data: Experiments, Sampling, and Experimental design**
Key concepts: randomization, confounding

April 8: **Randomness and probability**
Key concepts: random variables, expectation, permutations and combinations

April 10: **Distributions**
Key concepts: discrete distributions, continuous distributions, sampling distributions

April 15: **Estimating Parameters**
Key concepts: point estimates, confidence intervals, likelihood

April 17: **Hypothesis testing 1: Inferences based on one or two samples**
Key concepts: proportion test, t-test, nonparametric analogs

April 22: **Hypothesis testing 2: ANOVA**
Key concepts: single factor ANOVA, nonparametric analogs

April 24: **Linear regression**
Key concepts: Q-Q plots, simple and multiple regression

April 29: **Analysis of categorical data**
Key concepts: Chi-square tests, Fisher Exact Tests

May 1: **Assessing significance in high dimensional experiments**
Key concepts: Multiple Testing, False Discovery Rates, q-value