Advanced Statistics for Genomics
BINF 6310/8310 – 6310L/8310L
Spring 2015

<table>
<thead>
<tr>
<th>Instructor: Anthony Fodor</th>
<th>Contact Information</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><a href="mailto:afodor@uncc.edu">afodor@uncc.edu</a></td>
</tr>
</tbody>
</table>

Office Hours by appointment
Monday/Wed 9:30-10:45
Wed: 2:00-3:15

Textbooks Needed:
Zuur et al. Mixed Effect Models and Extensions in Ecology With R.

A.) COURSE DESCRIPTION

In this course, you will learn to apply statistical models to the large, complex datasets of post-genomic biology.
B.) PRE/CO-REQUISITES
You should be able to program in some high-level language: Java, C#, Perl, Python, Ruby, C++, C+, Fortran, whatever. By “program”, I mean be proficient enough to parse large files and implement simple algorithms. You should have a basic knowledge of molecular biology. You should know the meaning of the terms “DNA”, “RNA”, gene, protein, etc. You should have had at least one semester of statistics.

C.) LEARNING OBJECTIVES

i.) To achieve a thorough understanding of canonical and the Bayesian statistical models, what they can and can’t do and what their assumptions are.

ii.) To learn to apply these statistical models to the enormous datasets of post-genomic biology.

iii.) To learn to understand the way the statistical models are used (and misused) in the bioinformatics literature.

iv.) To learn to program in R.

v.) (Depending on time): To extend canonical linear models to mixed linear models.

vi.) (Depending on time): To extend Bayesian models to Hidden Markov Models.

vii.) (Depending on time): To extend canonical linear models to machine learning on high dimensional genomics datasets.
D.) INSTRUCTIONAL METHODS
Lecture: Monday - Wed 9:30-11:00
Lab: Wed 2:00-3:30

E.) GRADING

Problem sets and assignments will be worth 50% of your grade. Problem sets will include assigned problems (to be solved in R). Ph.D. students (8000 level) will have a problem set in which they re-analyze data from a paper published in their field of interest.

There will be a mid-term (25%) and a final (25%). The mid-term and final are on the materials in the lectures. All exams are cumulative.
F.) TENTATIVE SCHEDULE

I. Introduction – What is a probability distribution – The Uniform Distribution – The law of large numbers.

II. Probability estimated by relative frequency of sequence counts. Sequence entropy.

III. Canonical and Bayesian approaches to using the binomial distribution for inference.

IV. Canonical linear models and hypothesis testing.
   a. The normal distribution and the central limit theorem.
   b. The t-distribution.
   c. T-tests and their non-parametric equivalent.
   d. ANOVA.
   e. Simple linear regression.
   f. Multiple linear regression.
   g. PCA clustering in genomics experiments.

V. Multiple hypothesis testing and the power problem in genomic studies.

VI. Applying inference to sequence counts – the Poisson and negative binomial distributions, over-dispersion, dseq (or other rna-seq analysis packages).

VII. (If time) Mixed linear models.

VIII. (If time) Probabilistic models in sequence analysis
   a. Dynamic programming and pairwise alignment.
   c. Markov chains with Gaussian emission probabilities.
IX. (If time) Linear models for high-dimensional machine learning
   a. Over fitting vs. under-fitting; bias vs. variance trade-offs.
   b. Ridge and Lasso (penalized) regression models.

G.) POLICIES AND PROCEDURES
   a. ACADEMIC INTEGRITY
      All students are required to read and abide by the Code of Student Academic Integrity. Violations of the Code of Student Academic Integrity, including plagiarism, will result in disciplinary action as provided in the Code. Definitions and examples of plagiarism are set forth in the Code. The Code is available from the Dean of Students Office or online at: http://www.legal.uncc.edu/policies/ps-105.html. A set of links to various resources on plagiarism and how to avoid it is available at the UNCC Library website: http://library.uncc.edu/display/?dept=instruction&format=open&page=920.

   b. ATTENDANCE
      Attendance at both laboratory and lecture is mandatory.