Genome 560
Homework 5
Due May 2, 2008

1. [5 pts] Complete the following ANOVA table from the values given. The sample size is N = 20.

Source of Variation	Sum of Squares	df	Mean Square	F
Group	56.7			
Error		14	13.5	
Total				

What is the probability of observing an F-statistic this large or larger under the null hypothesis?

2. [15 pts] Download the data file "Homework4.txt" from the course web-site. This file consists of a measure of genetic variation (column labeled Y) measured in 947 humans from geographically diverse populations. (FYI, the outcome variable in this case is the loadings from a Principal Component analysis of SNP genotypes for a polymorphism in the lactase gene). The additional variables in the file are Climate and Geographic Region that each individual was sampled from.

A. Make boxplots of the distribution of Y for both Climate and Geographic region. In doing our part to save the environment, put both plots on a single page.

B. We want to test the hypothesis that the mean response variable, Y, differs among Climate and Geographic Region. Perform a full two-way ANOVA. Submit the ANOVA table for this analysis and interpret the results.

C. After doing the analysis, you realize that perhaps it would be a good idea to check the assumptions of the ANOVA model. In particular, check for normality (by constructing a qq plot of the residuals from the model fit in part B above) and constant variance (by a Bartlett's test). Are you comfortable with the results obtained in part B given these analyses? Discuss approaches that may be used to make this data set more appropriate for ANOVA (i.e., what can be done to make the data meet the assumptions required for a valid ANOVA analysis)?

D. A non-parametric alternative to ANOVA is the Kruskal-Wallis rank sum test. Perform a Kruskal-Wallis test separately on both Climate and Geographic Region. Report the p-values and discuss whether this is consistent or inconsistent with the ANOVA results obtained in B.

5. [5 pts] The return of probability… Although ANOVA is a popular framework for hypothesis testing in genomics, unfortunately many data sets (such as microarray data) often violate the assumptions of ANOVA. One popular approach to analyze data when one is suspicious of potentially dubious assumptions is to use bootstrapping of the residuals of an ANOVA analysis to test hypotheses. The details of this approach need not concern us here. Rather, you're going to help me understand more fundamental probabilistic questions regarding the bootstrap. I sense your giddiness already, so without further adieu:

A. Recall, in constructing a bootstrap sample from a set of N data points, we randomly sample with replacement N observations. Thus, some observations may be sampled zero times, one time, etc. What is the probability that a particular data point is selected in a single bootstrap realization?

B. What is the probability that a particular data point is selected two times in a single bootstrap realization?

C. What is the probability that a particular data point is selected k times in a single bootstrap realization?