STA613/CBB540 HOMEWORK 1

DUE THURSDAY, JAN 24^{TH}

For the questions that require a written answer: be brief. Only include the figures in your answers when they are specifically requested.

(1) *Poisson and the negative binomial.* In a program like R, let's look at some examples of draws from a distribution. For this question, include the code that you used to answer these questions, and your answers, but not the plots themselves.
  (a) Sample 1000 IID draws from a Poisson distribution with λ parameter equal to your age (you can be approximate if you are concerned about privacy issues) (see: rpois). Plot a histogram of these data (see: hist). How many of those 1000 draws are above 50 and how many are below 10 (don't use the histogram to answer this; instead see: sum)? What is the variance of the sample (see: var)?
  (b) Sample 1000 IID draws from a negative binomial distribution with size parameter equal to the same age as above and $p = 0.5$ (see: rnbinom, in the stats library, see: library). Plot a histogram of these data. How many of those 1000 draws are above 50 and how many are below 10? What is the variance of the sample? How does this compare with the Poisson?
  (c) Sort and plot the draws from the negative binomial and the Poisson (see: sort, plot). What can you interpret from this plot?

(2) *Maximum likelihood estimates.* Assume you have a set of data $D = \{x_1, x_2, ..., x_n\}$, where each x_i is an integer between zero and 73. You decide to model these data as drawn from a Poisson distribution.
  (a) Derive the maximum likelihood estimates for the Poisson parameter λ. Start by writing out the log likelihood of the data.
  (b) Does this estimate make intuitive sense to you? Why or why not?
  (c) Is the λ parameter sufficient to fully estimate the model? In other words, is λ the full set of *sufficient statistics* for the Poisson distribution? Why or why not?

(3) *Gene expression data.* Download the matrix of gene expression data (simulated) from the course website. You will find a matrix with $n = 500$ individuals and $p = 10$ genes. Read in these data to a program like R (type: "?read.table" for how to do that in R). For this question, include the code that you used to answer these questions, and your answers, but not the plots themselves.
  (a) Plot a histogram of each gene (see: hist). Which ones look the farthest from normally distributed?
  (b) Plot a quantile-quantile plot for each gene, comparing samples drawn from a normal distribution with the same mean and standard deviation as the gene in question with the actual samples (see: qqplot, rnorm, mean, sd,

abline(0,1)). Which genes look the farthest from normally distributed? How do these plots compare with examining the histograms?

(c) Standardize one gene, and plot the same quantile-quantile plot with the standardized data. Does standardizing make the data look more or less similar to what is expected from a normal distribution? Are outliers controlled?

(d) Project one gene to the quantiles of a standard normal, and plot the same quantile-quantile plot with the quantile normalized data (see: qqnorm). How does quantile normalization impact the quantile-quantile plots? Are outliers controlled?

(e) Plot the covariance of the genes (see: image, cov). Which genes appear to covary?

(f) Plot the correlation of the genes (see: cor). Which genes appear to be correlated?

(4) *Finding eQTLs.* Download a corresponding set of (simulated) genotypes from the course website.

(a) Which gene-SNP pairs appear to be associated (see: lm)?

(b) Choose one associated gene-SNP pair, and one pair that does not appear associated. For both pairs, plot the SNP (x-axis) versus gene (y-axis) (see: plot). Overlay on top of this the best fit linear regression model parameters β (see: abline or lines). Include this figure in your solutions.

(c) What are the RSS for both of the pairs you have chosen? What is the r^2 value?

(d) Write out the likelihood of the gene expression data given the genotype data. Compute the likelihood of the gene expression data given the genotypes for your two chosen gene-SNP pairs (assume the residual variance $\sigma^2 = 1$).