# STA613/CBB540 HOMEWORK 2

DUE THURSDAY, JAN $31^{ST}$

(1) *Hypothesis testing.* Say you are given the following table, representing the set of counts, for a particular threshold, of a statistical test applied to some number of features. Let *significant* and *non-significant* represent the results of a particular test with the given threshold, and *positive* represent the number of truly alternative features, and *negative* represent the number of truly null features.

| | non-significant | significant |
|---|---|---|
| negative | 1945 | 54 |
| positive | 188 | 192 |

(a) In this example, how many false positives, false negatives, true positives, and true negatives are there?
(b) What is the false positive rate for this threshold? What is the false discovery rate?
(c) If you are going to give the set of significant features to your collaborator, what are you going to tell that collaborator about the make-up of that set in terms of the false positive rate and the false discovery rate?
(d) What are the sensitivity and specificity in terms of these results?

(2) *Exponential family.*
(a) Write out the exponential distribution in the exponential family form. What is the sufficient statistic for this family ($T(x)$)? What is the response function? What is the natural parameter?
(b) Now set $\eta$, the natural parameter, to $\theta^T X$, and write out the conditional distribution $Y|\theta, X$ in terms of the exponential distribution generalized linear model.
(c) Write out the log likelihood of this model for a set of data $D = \{(x_1, y_1), ..., (x_n, y_n)\}$.
(d) How would you estimate $\theta$? (Just in words).

(3) *False positives and family-wise error rate.* You are trying to identify eQTLs among 8,724 genes with 52,827 SNPs. How many false positives do you expect for p-value threshold:
(a) $t(x) = 0.05$?
(b) $t(x) = 5 \times 10^{-8}$?
(c) If you choose to use the Bonferroni correction (FWER) for p-value=0.05, what will your p-value threshold be?
(d) How many false positives would you expect under this Bonferroni correction scenario (conservatively)?

(4) *Finding eQTLs.* Download the set of simulated genotype and gene expression data from last week's homework.

(a) Compute and sort the set of p-values from modeling the data $Y|X, \beta$ as a normal distribution (i.e., linear regression); don't show them here.
(b) How many tests were performed? What is the Bonferroni corrected threshold for p-value=0.05?
(c) At this threshold, how many associations are significant?
(d) Permute the labels on the genotypes (same permutation for all SNPs), and compute the p-values from the same model. Plot a quantile-quantile plot of the sorted p-values from the permuted tests versus the p-values from the actual tests (show this plot). What does this plot tell you about the data?
(e) For a false discovery rate of $\leq 10\%$, what is the p-value threshold (compute this using both the permuted p-values and the actual p-values)?
(f) Try permuting the SNPs again, and compute the p-value threshold for an FDR $\leq 10\%$ again. How robust is your estimate of the threshold to different permutations? How might you improve this estimate?

(5) *Case-control study.* Using the same genotypes from last week, download the case control data as part of this week's assignment. There are two columns, corresponding to two simulated binary phenotypes.
(a) For each of the five SNPs and each phenotype, compute the odds ratio using logistic regression (see: glm); for which SNP/phenotype pairs do you see possible associations? Plot the two associations with the odds ratios farthest from 1, including the case control status and associated genotypes, and the logistic regression function (for values of $x$ between 0 and 2).
(b) What is the variance of these odds ratios according to the model?
(c) In words, what do the variance estimates say about your measurement of the odds ratio?
(d) For each of the three genotypes $\{0, 1, 2\}$, what would your predictions be about the probability of a case or a control given the model you estimated here?