Quantitative Genomics and Genetics - Spring 2016
BTRY 4830/6830; PBSB 5201.01

Midterm - available online 11:59PM, Tues., March 15

For midterm exam, due before 11:59PM, Fri., March 18

PLEASE NOTE THE FOLLOWING INSTRUCTIONS:

1. You are to complete this exam alone. The exam is open book, so you are allowed to use any books or information available online, your own notes and your previously constructed code, etc. **HOWEVER YOU ARE NOT ALLOWED TO COMMUNICATE OR IN ANY WAY ASK ANYONE FOR ASSISTANCE WITH THIS EXAM IN ANY FORM** (the only exceptions are Mahya, Jin, and Dr. Mezey). As a non-exhaustive list this includes asking classmates or ANYONE else for advice or where to look for answers concerning problems, you are not allowed to ask anyone for access to their notes or to even look at their code whether constructed before the exam or not, etc. You are therefore only allowed to look at your own materials and materials you can access on your own. In short, work on your own! Please note that you will be violating Cornell's honor code if you act otherwise.

2. A complete answer to this exam will include two files: a SINGLE text file including all of your R code, and a SINGLE file including all of your typed answers and plots (where the latter may be a scan as long as we can read it). Please note that for your R code, to get full credit for all problems, we must be able to run your code and replicate all of your results (with ease!). We will attempt to run your code if you do not do this but we will deduct points accordingly (note that no code = no credit!).

3. Please pay attention to instructions and complete ALL requirements for ALL questions, e.g. some questions ask for R code, plots, AND written answers. We will give partial credit so it is to you advantage to attempt every part of every question.

4. The exam must be in Mahya's or Jin's (as appropriate) email inbox before 11:59PM Fri., March 18. It is your responsibility to make sure that it is in the appropriate email box before then and no excuses will be accepted (power outages, computer problems, Cornell's internet slowed to a crawl, etc.). Remember: you are welcome to hand this in early! We will deduct points for being late for exams received after this deadline (even if it is by minutes!!).

QUESTIONS (10 total, multiple parts per question) - make sure you answer all parts of all questions (!!):

Part 1: General Probability and Statistics Concepts

For Questions 1-3, consider a 'heights' system where you are interested in answering questions about the average height of all individuals in the United States, where you will be performing a 'measuring' experiment, such that you will measure heights of individuals selected from the United States. Assume that you will use $\mathbb{R}$ as your sample space Ω and the appropriate sigma algebra for this sample space $\mathcal{F}$. Assume a probability measure $Pr(\mathcal{F})$ on your sigma algebra and a random variable X that is a function on your sample space. Assume that the normal 'family' of probability models contains a probability model that is a correct description of your system and experiment. Assume that your observed sample of size n is iid.

1. Explain to your collaborator about the assumptions you are making with this framework by providing a one sentence description as to why each of the following assumptions is reasonable for this heights system / experiment AND at most two sentence description that explains specific aspects of these assumptions, which could be unrealistic or not exactly correct for this heights system / experiment (you are welcome to provide a hypothetical scenario) for EACH of the following: **(a)** Your choice of the sample space Ω, **(b)**. Your choice of the normal distribution to describe the possible 'family' of probability models. **(c)** Your choice of treating the sample as iid.

2. Recall that a statistic T is a function that maps each possible sample of size n, that could be generated under the correct model, to $\mathbb{R}$ and consider how we might construct various univariate statistics as estimators: **(a)** Could you define an estimator $\hat{\mu}$ for the parameter μ that is guaranteed to be incorrect no matter what the true value of μ? If your answer is 'yes', provide an example, if your answer is 'no' provide a one sentence explanation, **(b)** Could you define an estimator $\hat{\sigma}^2$ for the parameter σ^2 that is guaranteed to be incorrect no matter what the true value of σ^2? If your answer is 'yes', provide an example, if your answer is 'no' provide a one sentence explanation. **(c)** Using no more than one sentence, explain why any estimator you could define will have a probability distribution $Pr(\hat{\theta})$. **(d)** Using no more than two sentences, explain why it is not possible to construct an estimator $\hat{\mu}$ or $\hat{\sigma}^2$ that is guaranteed to be correct for your heights case. **(e)** Provide the definition of an unbiased estimator. Is $MLE(\hat{\mu})$ unbiased? Is $MLE(\hat{\sigma}^2)$ unbiased? If one or both of these is biased, provide a one sentence explanation as to why we might still use this estimator.

3. *Note that for parts 'a-d', you only need to provide the R code you used, the histograms, and the specific numbers you are asked to report, while part e requires a written answer!*

   Assume that *unknown to you* that the true parameter values of your system / experiment are $\mu = 1.5, \sigma^2 = 1.3$ and consider the null hypothesis $H_0 : \mu = 1$ and the alternative hypothesis $H_A : \mu \neq 1$: **(a)** Using R, simulate $M = 10,000$ samples of size $n = 20$ under the null hypothesis and $\sigma^2 = 1.3$. For each sample, calculate $-2ln\Lambda$ and plot a histogram of the values you obtain. Report the total number of $-2ln\Lambda$ values you obtained that are greater than the critical value $-2ln\Lambda_{\alpha=0.05} = 3.841459$. Use pchisq($-2ln\Lambda$, 1, lower.tail=FALSE) to calculate the p-value for each of your $M = 10,000$ statistics and plot a histogram of your p-values. **(b)** Repeat part 'a' but set $n = 100$. **(c)** Using R, simulate $M = 10,000$ samples of

size $n = 20$ under the true parameter values. For each, calculate $-2ln\Lambda$ and plot a histogram of the values you obtain. Report the number of $-2ln\Lambda$ values you obtained that are greater than the critical value $-2ln\Lambda_{\alpha=0.05} = 3.841459$. Use pchisq($-2ln\Lambda$, 1, lower.tail=FALSE) to calculate the p-value for each of your $M = 10,000$ statistics and plot a histogram of your p-values. **(d)** Repeat part 'c' but set $n = 100$. **(e)** Using no more than three sentences, explain why the histograms and reported numbers of parts 'a' and 'b' are similar AND why the histograms and reported numbers for part 'c' are different for those of part 'd'.

For Questions 4-5, consider a single locus in a human population that has two possible alleles 'A1' and 'A2'. Assume that you will perform a 'measuring' experiment where you will determine the genotype of a sample of individuals (without any errors).

4. **(a)** For the sample space $\Omega = \{A1A1, A1A2, A2A2\}$ write out the Sigma Algebra $\mathcal{F}$. **(b)** Assume that the true probability model has $Pr(A1) = 0.3, Pr(A2) = 0.7$ and where $Pr(A1 \cap A2) = Pr(A1)Pr(A2)$. Write out the probabilities of each set in $Pr(\mathcal{F})$. **(c)** Write out how a random variable, which you can call $X_{\#A1}$, maps each element of the sample space to an appropriate real number such that this random variable returns the number of A1 alleles in a given genotype. Describe how this random variable is related to X_a. **(d)** For $X_{\#A1}$, write down an appropriate 'family' of probability models and what the true parameter value(s) would be given the true probability model in part b.

5. **(a)** Consider a random variable X_{A1A1} which maps elements of the sample space in Question 4 to the reals as follows: $X_{A1A1}(A1A1) = 1$, $X_{A1A1}(A1A2) = 0$, $X_{A1A1}(A2A2) = 0$. Also consider the random variable X_a defined on the same sample space. Assuming the probability model in Question 4, show that these two random variables are not independent. **(b)** Assume that you have a sample of n iid observations of the random variable X_{A1A1}, i.e., a sample $\mathbf{x}_{A1A1} = [x_{1,A1A1}, ..., x_{n,A1A1}]$. Say that you are interested in estimating the true probability of individuals with the genotype A1A1. Provide a formula for the best estimator possible by defining a statistic on the random variable X_{A1A1}, i.e., this statistic will map your sample $\mathbf{x}_{A1A1}$ to a single number. **(c)** Say that you are interested in estimating the true probability of the 'A1' allele in the entire population (the Minor Allele Frequency or MAF) from this same sample. Provide a formula for best estimator possible by defining a statistic on the random variable X_{A1A1}. **(d)** For the estimator in parts 'b' and 'c', would these both be good estimators even if $Pr(A1 \cap A2) \neq Pr(A1)Pr(A2)$? Explain your answer for each using no more than two sentences (for each).

Part 2: Probability and Statistics as applied in Quantitative Genomics: Genome-Wide Association Studies (GWAS)

Your collaborator is interested in mapping genetic loci that can affect human Body Mass Index (BMI). They want to perform a GWAS experiment and they would like you to perform the analysis. They have collected data for a number of individuals sampled from a population and they have provided you relative measures of BMI in the file "QG16_phenotypes_midterm.txt" and SNP genotypes in the file "QG16_genotypes_midterm.txt". Note the following:

- There are two columns in the phenotype file, the first contains the name of each individual in the sample, the second contains the BMI of each individual in the sample. Each row of the phenotype file lists the name and BMI for an individual in the sample (n rows total).

- In the "genotypes" file, the first row contains the names of N genotypes, each of which has been measured for each of the n individuals in the sample. Each of the following rows contains all of the genotypes measured for a specific individual. For each of these rows, each consecutive PAIR OF COLUMNS represents a genotype of an individual (i.e., each genotype name refers to a genotype defined by a pair of columns), where there are a total of N genotypes for each individual (for the row corresponding to individual i, the 1st and 2nd column = 1st genotype of individual i, 3rd and 4th columns = 2nd genotype of individual i, ..., rows $2N-1$ and $2N$ $= N$ genotype for individual i). Each genotype of each individual is composed of a pair of alleles (e.g., the possible genotypes at the jth genotype for an individual i are 'A1A1', 'A1A2', 'A2A2'). In some cases, instead of an allele 'A1' or 'A2' there is a number '9' indicating that the measurement of the allele is missing.

6. **(a)** Plot a histogram of the phenotypes (provide your code!). **(b)** There is one individual with a phenotype that is clearly considerably larger than the others (an outlier). Report the name of this individual and remove this individual from the analysis by removing the phenotype AND the corresponding genotypes of this individual from the data set. **(c)** Re-plot the histogram of the phenotypes after removal of this individual. **(d)** The phenotypes should now look approximately normal. In no more than one sentence, explain why it is important that the phenotypes be well modeled by this distribution if we are going to use the genetic linear regression to model the relationships between genotypes and this phenotype.

   Once you have completed part 'a' your dataset now has $n-1$ individuals each with N genotypes. Analyze this filtered data set in Question 7.

7. **(a)** Write code to identify which individuals have a genotype with a missing allele. Report the names of individuals with at least one genotype with a missing allele and report the names of the genotypes for each of these individuals that have missing alleles. The entire list of these numbers represent genotypes for which at least one individual has missing data. **(b)** For your list in part 'a' remove these genotypes from the analysis, not just for the individual where they are missing but FOR EVERY individual in the data set. That is, remove the PAIRS OF COLUMNS corresponding to your missing genotype list.

   Once you have completed part 'd' your dataset now has $n-1$ individuals each with $N-$ the number of genotypes you removed. Analyze this filtered data set for Question 8-10.

8. **(a)** For EACH genotype, using the formulas provided in class, calculate the $MLE(\hat{\beta})$ for the three β parameters when using the linear regression model $y_i = \beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d + \epsilon_i$, with $\epsilon_i \sim N(0, \sigma^2_\epsilon)$ and plot a histogram for the estimates of each parameter = three histograms total (provide your code! And make sure you label your plots!). **(b)** For EACH genotype, calculate p-values for testing the null hypothesis $H_0 : \beta_a = 0 \cap \beta_d = 0$ versus the alternative hypothesis $H_A : \beta_a \neq 0 \cup \beta_d \neq 0$ using the formulas provided in class (i.e. the predicted value of the phenotype $\hat{y}_i$ for an individual i, the SSM, SSE, MSM, MSE, and the F-statistic), although you may use the function pf( ) with the option 'lower.tail=FALSE' to calculate the p-value from your F-statistic. **(c)** Plot a histogram of ALL p-values (not -log(p-values)! just the p-values!). If you ignore a few of the p-values that are extremely small (=highly significant), describe how the rest of the p-values are distributed (i.e., the probability distribution do they appear to resemble) and using no more than two sentences, explain what this indicates about the correct statistical model for the bulk of the genotype-phenotype relationships in the data.

(d) Plot a Manhattan plot using the -log(p-values) (provide your code!).

9. **(a)** For a Type 1 error of 0.05, how many genotypes are significant from your analysis in Question 8? **(b)** For a Type 1 error of 0.05 / $N-1$, how many genotypes are significant from your analysis in Question 3? **(c)** Using no more than two sentences, describe why the second (the lower) of these Type 1 errors is more appropriate for identifying the location of causal polymorphisms and justify your answer. **(d)** For the lower of the Type 1 errors, report the genotypes that you returned (i.e., a list of genotypes from 1 to $N-$ the number of genotypes you removed in Question 2). Using no more than three sentences, explain how many causal genotypes these two sets are likely indicating and provide a justification of your assertion.

10. **(a)** For the single most significant genotype in your analysis, produce two x-y plots: X_a vs Y and X_d vs Y (provide your code! And make sure you label your plots!). **(b)** Choose one of your other genotypes that has a p-value > 0.5 (i.e., choose any one of your genotypes with a p-value > 0.5) and produce two x-y plots: X_a vs Y and X_d vs Y. **(c)** Using no more than three sentences, describe how the plots in parts 'a' and 'b' differ, an explanation as to why you believe they differ, and why the plots make sense assuming that your belief is correct. **(d)** Using no more than two sentences, explain to your collaborator why the genotype you have plotted in part 'a' is not likely to be the causal polymorphism for BMI but why it indicates the position in the genome where a causal polymorphism may be located.