For all of these problems, if you need the prior probability of a genotype in the population, you can assume the following: $P(AA) = 0.95^2$, $P(BB) = 0.05^2$, and $P(AB)$ is the remainder.

Before generating random numbers, make sure to set a random seed so your results will be reproducible.

Problem 1

From biallelic models to triallelic models

In the "Error vs SNP" lecture, we limited ourselves to biallelic variation. Fortunately, most single nucleotide variation is biallelic (estimates of multiallelic variation are 3% of SNPs). Let's assume that at this particular position in the genome, we can have 3 possibilities: $\{G, A, C\}$. This isn't crazy - sometimes for unstable mutations alleles can drift and then create more stable variation (e.g. a G can turn into a C in some finite number of generations). We are going to restrict ourselves to the case where only two chromosomes exist.

(a) How many possible genotypes are there? (i.e. GG, GA, . . . , CC).

(b) If we assume an error can result in any base (e.g. a G can turn into a T), assume I observe a T in my data. How does that translate into the probabilistic model? Note: I'm not looking for a complicated answer. Simply describe how it might change the probability of observing in error is sufficient.

(c) Write the probability of observing read i for homozygous genotypes (i.e. GG, AA, CC). You can follow the example from class, but don't forget there are more possibilities than the example in class.

(d) Write the probability of observing read i given genotype GA. Remember that we have only two chromosomes but three possible alleles.

(e) Write down the probability of the remaining genotypes. If they reduce into functions of the others, feel free to be lazy and write them in terms of other probabilities.

Problem 2

Data analysis + the bootstrap

Consider the biallelic model again.

(a) Refer to the slides from `errors_and_snps.pdf`. If you assume $P(S = A) = P(S = B)$, does this reduce any of the likelihoods? For the remainder of this problem, please assume $P(S_i = A) = P(S_i = B)$ when there is than one allele in the truth.

(b) `reads.tsv` is some data in the following format:

observation	$P(E_i = 1)$	**indicator if actually an error**
A	0.02	FALSE
G	0.01	TRUE
...	...	...

This is a simulation and you don't need the final column; it is there for your personal enjoyment. You might be able to do something with it, but you don't need it to solve the problem. *Please don't use it to solve the problem.*

Write some code to estimate the posterior probability of the three possible genotypes given the data.

(c) Randomly sample 5 observations with replacement and re-estimate the posterior probability of each genotype. What are the results?

(d) Repeat (c) 1,000 times. That means you will have 1,000 estimates for each of your posterior probabilities, each using 5 observations. This procedure is a variation of the *bootstrap*. Make a histogram for each of the posterior probabilities. Please be mindful of the number of bins and the appearance of your histogram. No one likes an ugly histogram.

(e) Repeat (d), but this time instead of taking 5 observations, take 50. Again, make three histograms.

(f) How do the results from (d) and (e) compare? Feel free to take summary statistics like the mean and standard deviation from those resampled results.

(g) Implicitly, there are assumptions about the base caller, the prior probabilities, etc. What are these assumptions and how might they affect the results? An example, what if the base caller probability estimates were way off?

Problem 3

Generative models for sequencing

Consider the error distribution in `errors.tsv`. Each line is the probability of an error at position k in the read. In other words, if you are going to generate data from sequence `ACG`, the probability that you would observe an error in position 2 and observe some character other than `C` corresponds to the second line in `errors.tsv`. Assume that each read is independent.

Further, assume the error transition probability in `transitions.tsv`. The bases are in the following order: (A, C, G, T). For example, if the true base is C, then the second row third column tells you the probability of observing a G.

(a) Describe in words (or clear pseudocode) how you can use the error distribution and transition distribution to generate N "reads" that span a putative SNP. You need not worry about all of the sequence, only the base that spans the putative SNP position.

In other words, each “read” only contains one base, the base that spans the putative SNP position. You need to think about how reads are generated in DNA sequencing. In short, your output should look like the input from Problem 2. Tip: you will need to begin by choosing the true genotype of the data.

(b) Assume the biallelic model. Under this model we assumed that the only data you could observe is from allele `A` or from `B`. What happens if you observe a base that is not A or B? Describe what you plan to do with this observation and how it might affect your inference from the model and data.

(c) Assume the biallelic model (from Problem 2), but simulate 20 reads from the model you described in (a). Run this procedure 1,000 times and make histograms like you did in Problem 2. What is the probability of each of the 3 genotypes?

(d) Remove the dependency on `transitions.tsv` so that when you observe an error, there is only one possibility and that possibility is only genotype A or B. Repeat the analysis from (c). How does this assumption change your results? Describe.

Problem 4

Some calculations related to statistical power

Consider the biallelic model. Assume that errors are uniform, that is $P(E_i = 1) = e$.

(a) Assume I observed 100 reads and 30 of those observations were B. What is the probability of genotype AB given the data? That is, what is $P(AB \mid E_1, E_2, \ldots, E_{100})$?

(b) If the truth is AB, and I observe 100 more reads at random, how many do I expect to be 'B'? Note, your derivation should include e.

(c) Forget about the data in (a) and (b). Starting from no data, how many reads would I need to observe to have posterior probability 0.99 or greater of genotype AB? In other words, what does N need to be in order to to have $P(AB \mid E_1, \ldots, E_N) \geq 0.99$)? You can (deterministically) assume that half of your observations come from A and B. In other words, $N/2$ of your reads contain the observation A and $N/2$ contain B.

(d) Preface: this problem is a bit weird, but weird can be interesting. Assume I started with the data in (a) and for some strange reason I could only observe reads that turned out to be B. Nothing else about the data generation process has changed. For example, the error rate is still e and you are still equally likely to sample from each chromosome, but you will only observe reads that occur in a B (which may have arisen by an error or could be the correct base). How many observations would I need to observe to be 99% confident of genotype AB? Note: you might be able to analytically solve this, but I was too lazy to. It's probably easier to write some code to solve this.