Homework 3

Instructions. This homework is due at the beginning of lab, Wednesday, September 20th. You can edit this file and copy/paste into it by opening it in Libreoffice on Linux, PDFfiller Google App, Acrobat, Microsoft Word and others. If you use the internet for information, please assess the verity of the source and document it as a source. Please turn in a pdf version of the written solutions, as well as a zip file of all code you use to solve the problems. Please name each piece of code by the question it is intended to answer or otherwise clearly indicate to me how to run the code to get each answer.

1. In the following, suppose you have read data from a Next Generation Sequencing (NGS) experiment sequencing your own DNA delivered in a Fasta file. Learn more about one of the technologies that can do generate such data under Large Genome Sequencing at the Illumina Whole Genome Sequencing website.
   (a) ___ (True/False) The proportion of nucleotide A in a read is a statistic.
   (b) ___ (True/False) The proportion of nucleotide A in your *FOXP2* gene is a statistic.
   (c) If you perform statistical inference using these data, what is the population you can extend your conclusions to? _____
   (d) One kind of inference that might be of interest is to determine whether you are a mutant. Suppose there are $n = 17$ reads covering a particular site in your genome that is known to be mutated from C to A in some individuals. If you observe $X = 4$ A nucleotides and $n - X = 13$ C at this site, can you assume you are mutant? Keep in mind that NGS data contain errors and you are diploid. (Also assume you are "mutant" as soon as you have at least one copy of the mutant nucleotide.) Please show all your work, including your choice of test statistic, your specific null hypothesis H_0, including any assumptions it makes, and the calculation of a p-value to support your conclusion.

2. In this question you will identify the nucleotide substitution that is occurring in the nonpermissive cells from the virus experiment. Suppose N_t is the target nucleotide and it is mutated to $\mathrm{N}_m \neq \mathrm{N}_t$, both in $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. Your goal is, therefore, to infer N_t and N_m.

   (a) For your **specific null hypothesis** H_0, you assume there is *no* mutation occurring, that therefore the virus genomes from permissive and nonpermissive cells are identically distributed, that the sequenced 500 bp fragments are independent and uniformly (every possible starting position is equally likely) drawn from the larger virus genome, and that the nucleotides in the virus genome (and by implication each fragment) are iid Multinoulli$(p_\mathrm{A}, p_\mathrm{C}, p_\mathrm{G}, p_\mathrm{T})$, where p_N is the probability of nucleotide N. Raise two biological criticisms of this model, identifying how it may fail to mimic real biological data, other than the interesting possibility that mutations may be *actually* happening.

   (b) Let $n_{n\mathrm{N}}$ be the total number of $\mathrm{N} \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ nucleotides in the fragments obtained from nonpermissive cells and $n_{p\mathrm{N}}$ the corresponding count for permissive cells. Let n_n and n_p be the total number of nucleotides in the reads from nonpermissive and permissive cells, respectively. The test statistic you will be using in this problem is

$$T = \sum_{\mathrm{N} \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}} \left(\frac{n_{n\mathrm{N}}}{n_n} - \frac{n_{p\mathrm{N}}}{n_p} \right)^2 .$$

      i. What numeric value will this test statistic be close to if H_0 is true? Why?

      ii. How will the value of the test statistic change if H_0 is not true?

      iii. Compute and report the observed value t_0 of the test statistic. What set of values of random variable T are as or more extreme than t_0? (Hint: Write a function to compute t_0 from sequence or count data for later reuse.)

(c) Propose estimates $\hat{p}_{\mathrm{A}}, \hat{p}_{\mathrm{C}}, \hat{p}_{\mathrm{G}}$, and $\hat{p}_{\mathrm{T}}$ of the model parameters $p_{\mathrm{A}}, p_{\mathrm{C}}, p_{\mathrm{G}}$, and p_{T} under the model assumptions stated in Part (a).

(d) Use Monte Carlo simulation to estimate the p-value.

3. In this question, you will identify whether there are any dinucleotide motifs (a dinucleotide is any pair of contiguous nucleotides, such as AC) that are specifically targeted by the mutation mechanism. For example, if C $\rightarrow$ A is targeted in the AC dinucleotide context, then mutations will occur specifically as AC $\rightarrow$ AA. All but one part is bonus because although many of you, especially those who have come through Stat 330 (or Stat 341), should know how to do the involved calculations, I cannot assume you can.

   (a) **Bonus.** Show mathematically that if nucleotide sites are independent and mutation $\mathrm{N}_t \rightarrow \mathrm{N}_m$ independently strikes N_t nucleotides with probability p, then the proportion of $\mathrm{N}_t\mathrm{N}_t$ dinucleotides will be altered in the mutated genomes relative to nonmutated genomes. Thus, you cannot check for changes in dinucleotide proportions to determine there is a dinucleotide motif.

   (b) **Bonus.** Show that if the dinucleotide AC is targeted for mutation to AA with probability p, then even if the nucleotides of the provirus were independent in the permissive cells, they will not be independent in the nonpermissive cells. So, nonindependence is a signal that could used to detect dinucleotide motifs.

   (c) **Bonus.** Discuss the appropriateness of the following test statistic for testing whether CC is a targeted motif (to AC). What value do you anticipate if there is no mutation? What value do you anticipate if there is C $\rightarrow$ A mutation, but no dinucleotide targeting? What value do you anticipate if CC is targeted to AC?

$$T = \frac{\frac{N_{n\mathrm{AC}} n_n^2}{(n_n - 1) N_{n\mathrm{A}} N_{n\mathrm{C}}}}{\frac{N_{p\mathrm{AC}} n_p^2}{(n_p - 1) N_{p\mathrm{A}} N_{p\mathrm{C}}}}, \tag{1}$$

   where $N_{n\mathrm{AC}}$ and $N_{p\mathrm{AC}}$ are the counts of overlapping occurrences of dinucleotide AC in the provirus genomes of nonpermissive and permissive cells, respectively.

(d) Using the specific null hypothesis H_0 from the previous question 2 (that assumes no mutation), simulate many T from Eq. 1 and plot a histogram of the values. (Yes, you can do this in Python!) Do you think there is evidence that CC is targeted? What about a more appropriate dinucleotide, motivated by your results from question 2?

(e) **Bonus.** Using your knowledge of the mutation from question 2, compute a p-value for the probability of obtaining a test statistic T as or more extreme than the observed one for your favorite motif.