GENOME 560 Problem Set #1

(Due April 12th 8:59 am)

1. [**40 points**] **Computing Descriptive Statistics and Visualization**

   In this question, you will do some analysis on a DMS data set using R. Download a data file containing many DMSes from the course website `http://faculty.washington.edu/dfowler/teaching/2016/GNOM560/560_dataset_reduced.txt`, and do the following using R. Be sure to submit all R code and output.

   (a) [**5 points**] First, you'll need to open the downloaded file and isolate a DMS dataset from all the ones in the file. You can use the WW domain dataset we discussed in class or, if you're interested, you can pick a different one. Hint: use the which() function along with the [] index operator. The dms_id column of the dataframe may be of use. Show your code.

   (b) [**5 points**] Next, use the "range", "quantile", "mean", "var" and "sd" functions to obtain these descriptive statistics for the reported functional scores (e.g. the 'reported_scores' column). Submit the results. For quantiles, let's try 2.5%, 5%, 25%, 50%, 75%, 95% and 97.5% points.

   The lower 25% and the upper 25% point in a Normal distribution will be -0.67448 and 0.67488 standard deviations from the mean. Is this approximately true for the reported functional scores in your dataset?

   (c) [**5 points**] Use the "hist" command with 50 bins to examine the distribution of the functional scores. Include the histogram and the commands to generate it in your RMarkdown document.

   (d) [**5 points**] In class, we calculated the mean for the reported effects of a few different amino acids over all positions. Repeat these calculations for all 20 amino acids.

   (Hint: There are many ways to do this including manually or with a simple for loop, iterating over each amino acid.)

   Submit the commands that you used to obtain the desired means.

   (e) [**5 points**] Make a barplot consisting of the mean reported effect scores for every amino acid. Show the commands you used and the figure.

   (f) [**10 points**] Make boxplots of the score distributions for stop codons ('X'), alanine and tryptophan. Modify the histograms' x-label, y-label, and title to be more descriptive than the default values. Show the commands and figure.

(g) [**5 points**] Repeat (f) with "permuted" scores (e.g. randomize reported effects between alanine and trypotphan). Create permuted labels using sample(), for example:

```
> all_scores = c(trp_scores, ala_scores)

 >permuted_trp_scores = sample(all_scores, length(trp_scores))
```

Submit the commands you used and the figures. Based on the comparison with the results from part (f), what is your conclusion?

2. [**40 points**] **Probability Theory and Distributions**

You are interested in a particular haploid organism, *R. Waterstonii.* This organism has a locus with two known alleles, A and B, in the population. The organism is hard to find and cannot be cultured, though, so you only have four of them to study. You genotype this locus to determine the allele carried by each of your four isolates.

(a) [**10 points**] What is the sample space for this experiment? How many items does the sample space contain? Can you think of a general rule for finding the size of the sample space if you collect n isolates? Use this rule to find the sample space for 20 isolates.

(b) [**10 points**] Assume that the two alleles have equal frequency in the population. If you sampled four isolates at random, what is the probability that you will fail to observe A in any of your isolates?

(c) [**10 points**] Now calculate the probability that you will observe A once, twice, three or four times. Plot the results to define the probability distribution governing the possible outcomes.

(d) [**10 points**] Define a random variable to express the outcomes A and B for an isolate numerically. Calculate the expected value of random variable given that the alleles have equal frequency.

3. [**20 points**] **Measures of spread**

Develop a measure of spread for numerical data that is distinct from variance. If you are totally stumped, you can Google it but please give it 10m of quality thought before doing so!

(a) [**10 points**] Write down the equation for your statistic.

(b) [**10 points**] In a few sentences, compare and contrast your statistic with variance.

4. [**30 points**] **Inter-species Cross**

   Let's say that we are interested in functionally characterizing offsprings between inter-species cross. The Saccharomyces Genome Resequencing Project (SGRP) has completely resequenced 37 strains of the closely related species *S.cerevisiae* (25 natural isolates, 7 laboratory strains, and 5 clinical isolates) and 27 strains of the closely related species *S.paradoxus* (22 natural isolates and 5 laboratory strains).

   (a) If we want to perform all pair-wise crosses between the SGRP *S.cerevisiae* and *S.paradoxus* strains, how many matings should we do?

   (b) Let's assume that none of the crosses performed above worked so instead, we decide to perform all pair-wise matings within each species. How many matings should we now do for *S.cerevisiae* and *S.paradocus*?

   (c) More bad news. In setting up the cross in part (b), we somehow managed to switch the labels on the plates for the *S.cerevisiae* strains and we have no idea which one is which. What we're really interested in are cross between the natural and clinical isolates. Rather than doing the obvious and reordering the strains, we decided to gamble and randomly pick two *S.cerevisiae* strains to cross. What is the probability that we guess correctly and mate a natural and clinical isolate?

5. [**10 points**] **Hypergeometric Distribution**

   Say that in a gene expression study, we identified 100 differentially expressed genes. We performed a Gene Ontology analysis and find that 15 out of the 100 are annotated as "sensory perception of sound". Our microarray contains 10,000 genes in total, of which 1,000 are annotated as "sensory perception of sound". What is the probability that we would observe 15 or more genes out of 100 annotated with this term by chance?

6. [**30 points**] **Expression QTLs**

   There has been considerable recent interest in mapping loci that influence inter-individual variation in gene expression levels. In these experiments, gene expression levels are treated as a quantitative trait and linkage analysis is performed to find positions in the genome that contribute to variation in transcript abundance.

   Let's say that you have performed a linkage analysis experiment to map gene expression quantitative trait loci (eQTL). In total, you have detected significant linkage for 1,013 expression traits. Next, you want to test for the presence of linkage "hotspots", which are regions in the genome showing linkage to multiple gene expression traits.

To detect “hotspots”, you divide the genome into 579 bins of equal size. Define the random variable X_i to be the number of linkages observed in the i^{th} bin.

(a) What distribution does X follow? Briefly explain your choice.

(b) Assuming that eQTLs are randomly distributed, what is the probability that a bin contains no eQTLs.

(c) What is the probability that a bin contains 40 or more eQTLs?

(d) Find the number of eQTL in a bin x, such that $P(X \geq x) = 0.05$.

7. [**30 points**] **Parameter Estimation**

Imagine you have collected data on the number of *de novo* mutations per generation in *R. Waterstonii.* In 12 isolates, you observe the following number of *de novo* mutations:

Isolate	# of mutations	Isolate	# of mutations
1	12	7	18
2	15	8	12
3	10	9	7
4	4	10	11
5	6	11	5
6	15	12	14

(a) First, assume that the number if *de novo* mutations has a Poisson distribution with the unknown parameter λ. Come up with a common-sense estimator for λ and use it to find a point estimate using your observed data.

(b) The Poisson distribution has the interesting property that the mean and variance are equal. Find the variance of your sample. Compare it to your estimate of λ. Do they seem ”close enough” to you, or are you worried that the Poisson distribution might not be a good assumption?

(c) Let’s consider another way of answering part **b**. Given your point estimate of λ, generate 1,000 random samples of size 12 from a Poisson distribution. Find the sample variance for each of the 1,000 samples. How many of these simulated data sets have a sample variance greater than your observed data? Do you still agree with your response in part **b**?