Using the same rules as before, you will have almost two weeks to complete this homework.

These questions recreate an analysis done by a BCB colleague of yours. You will be modeling the abundance of pathogen genome variants collected from a massive, multi-year nationwide survey.

Question 1: Asymptotic results for multinomial probabilities.

(a) When deriving Maximum Likelihood Estimators (MLEs) with constraints, the Lagrange multiplier can be useful. In particular for multinomial data $\boldsymbol{X} \sim \text{Multinomial}(n, \boldsymbol{p})$, where $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ are occurrence counts of items falling in m categories and $\boldsymbol{p} = (p_1, p_2, \ldots, p_m)$ are the probabilities of each type[1], there is a constraint, $p_1 + p_2 + \cdots + p_m = 1$, that must be respected while maximizing the log likelihood $l(\boldsymbol{p} \mid \boldsymbol{x})$ after having observed $\boldsymbol{X} = \boldsymbol{x}$. If the constraint is of form $g(\boldsymbol{p}) = 0$, then the Lagrange multiplier λ is a non-zero constant used in formulating the Lagrange function

$$l(\boldsymbol{p} \mid \boldsymbol{x}) - \lambda g(\boldsymbol{p}),$$

which is then minimized instead of the log likelihood. Prove that the MLE $\hat{p}_i = \frac{x_i}{n}$.

(b) Obtain an estimate of the variance of MLE $\hat{p}_i$ by using the Central Limit Theorem[2], plugging in the MLE for the unknown p_i. Use these results to plot the MLEs from Part (a) and symmetric confidence intervals (using the asymptotic distribution provided by the CLT) for year 2013 and 2016 (or if you can, all years).

(c) The same variance you obtained in Part (b) can be obtained from inverting the Fisher Information and using asymptotic results for MLEs, but it takes considerabe more work. The benefit of using the Fisher Information, is that you also obtain the covariances $\text{Cov}(\hat{p}_i, \hat{p}_j) = -\frac{p_i p_j}{n}$. Use Wald test for $H_0 : p_A - p_B = 0$ to assess whether either of genotype A or B is significantly more abundant in either year 2013 or 2016.

Question 2: Bootstrap results for multinomial probabilities.

(a) Instead of relying on asymptotic theory, we can instead utilize the bootstrap to obtain confidence intervals. It will help to reshape the data (suppose `d` contains the data in the `count_genome.csv` file):

```
years <- 2009:2016
d.l <- reshape(d, varying = paste("X", years, sep=""),
        v.names = "cnt", timevar = "yr", time = years,
        direction = "long")
d.ll <- untable(d.l, d.l$cnt)
d.ll <- d.ll[, -c(3,4)]
```

Do you find any evidence that the confidence intervals from Question 1(b) are not actually symmetric?

(b) Formulate a nonparametric bootstrap test to replace the Wald test used in Question 1(c). Critique both tests by identifying their limitations and assess their trustworthiness in this context.

Question 3: Parametric bootstrap tests of a Poisson clustering model.

The researchers in this case were not satisfied with the pairwise comparison results. They wanted to know which genomes were significantly more abundant than other genomes in the same year. (If you think about it, using the pairwise results creates contradictions. For example, suppose you conclude that genome A is significantly more abundant than C, but not B. At the same time, you determine that genome B is not significantly more abundant than C. So is genome A significantly more abundant than all the others? No. Are genomes A and B significantly more abundant than all the others? No. But are there some genomes significantly more abundant? Yes. Right?)

[1] The multinomial is a generalization of the Binomial.

[2] Notice that $\hat{p}_i$ is the sample mean of iid Bernoulli random variables.

There is a relationship between the Multinomial model we have been using and independent Poisson distributions. In particular, if $X_1 \sim \text{Poisson}(\lambda_1), X_2 \sim \text{Poisson}(\lambda_2), \ldots, X_m \sim \text{Poisson}(\lambda_m)$ are independent, then the conditional distribution

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m \mid X_1 + X_2 + \cdots + X_m = n) = \frac{n!}{x_1! x_2! \cdots x_m!} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}$$

follows the Multinomial distribution, where $p_i = \frac{\lambda_i}{\sum_j \lambda_j}$ (see HW1 from BCB568, 2017). Using the independent Poissons model is appropriate if you wanted to generalize your conclusions to random experiments where the total sample size collected in each year was random. The Multinomial model conditions on the sample size and therefore only generalizes to samples of the same size. In a situation like this, where our goal is to perform inference on the past, it makes sense to condition on the sample size (there can be other reasons, see HW1, 2017). However, I want you to experience another model, so we will switch to the Poisson model.

(a) To answer the researchers' question, assume there are an unknown number K of genome groups, where all members of the same group k have the same true mean abundance λ_k. By maximizing over $\lambda_1, \lambda_2, \ldots, \lambda_K$, and the partition of genomes into K groups, we can give the researchers what they want by choose K through hypothesis testing. Specifically, the members of group $k_{\max} = \text{argmax}_{1 \le k \le K} \lambda_k$ are significantly more abundant than the others. (If they weren't, we would only have found evidence for $K = K - 1$ distinct groups.) Define $z_i \in \{1, 2, \ldots, K\}$ to be the discrete parameter indicating the group of genome $1 \le i \le m$. Write down a formula for the log likelihood $l_K(\lambda_1, \lambda_2, \ldots, \lambda_k, z_1, z_2, \ldots, z_m \mid x_1, x_2, \ldots, x_m)$ when K is given.

(b) Write code to maximize the log likelihood from Part (a). Given K and $z_1, z_2, \ldots, z_m$, one can find analytic formula for $\hat{\lambda}_k(z_1, z_2, \ldots, z_m)$, here written as functions of the unknown indicators. Plugging these formulae back into the log likelihood

$$l_K \left[\hat{\lambda}_1(z_1, z_2, \ldots, z_m), \hat{\lambda}_2(z_1, z_2, \ldots, z_m), \ldots, \hat{\lambda}_k(z_1, z_2, \ldots, z_m), z_1, z_2, \ldots, z_m \mid x_1, x_2, \ldots, x_m \right]$$

produces the profile log likelihood, now defined solely in terms of unknowns $z_1, z_2, \ldots, z_m$. Use the computer to maximize this log likelihood over all possible partitions. Demonstrate your code on years 2013 and 2016 for $K = 3$. (You may find function `combinations()` in Rlibrary `gtools` useful.)

(c) To compare $H_0 : K = k_0$ against $H_1 : K = k_0 + 1$, one can use the likelihood ratio Λ test (LRT) statistic, $-2\ln(\Lambda)$[3]. Assuming the necessary conditions apply, what is the asymptotic distribution of $-2\ln(\Lambda)$? Why don't the necessary conditions apply?

(d) Since the conditions for the asymptotic distribution of the LRT do not apply, perform a parametric bootstrap to test the null hypothesis $H_0 : K = k_0$ against $H_1 : K = k_0 + 1$. Start with $k_0 = 1$ and continue until you cannot reject H_0. Report the significantly more abundant genomes in years 2013 and 2016. How do they match up to your original plots?

[3]We use Λ to distinguish this ratio from the other λs of this question