BioE 145 Assignment 3 Due Tuesday 2/11 11:59 PM

Note: Turn in your submission to this assignment in Gradescope by Tuesday February 11th, 11:59 PM. Attach a PDF printout of the IPython sections of the homework as an appendix, as well as any code used to find your answers to the written questions.

Underlying Distributions - Genome Expression

1. **MLE of Poisson Distribution.** We have n independent data points x_i sampled from discrete random variable X that has a Poisson distribution.

   (a) Given the definition for likelihood and the probability mass function for the Poisson distribution below, write out the likelihood for λ given the n x_i samples.

$$\mathcal{L}(\lambda; x) = P_\lambda(X = x) \tag{1}$$

$$P_\lambda(X = x_i) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \tag{2}$$

   (b) Starting with the likelihood from your answer to Q1.a, show that the MLE of the n samples given an underlying Poisson distribution is the average of the samples.

$$\hat{\lambda}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3}$$

   Hints:

   i. A good place to start would be to look back at the use of log-likelihoods in solving for the MLE.

   ii. A reminder from calculus: taking the derivative and setting it equal to 0 usually helps when solving for a maximum or a minimum.

2. **Poisson Counts.** Here we'll carry out some statistical analysis to determine the significance that sets of counts for gene Z in two conditions (the data points) arise from different distributions. The two sets, x and y, are of size n each.

   The null and alternate hypotheses for the scenario will be:

   H_0 : The counts for both conditions come from the same Poisson distribution.

   H_a : The two sets of data points are sampled from different Poisson distributions.

   (a) Find expressions for the Poisson parameters for each of the two hypotheses: λ_0 for H_0 and λ_x, λ_y for H_a

   (b) Compute the lambdas by implementing your answer from Q2.a using the provided dataset. [*Jupyter Notebook*]

   (c) Find expressions for $\mathcal{L}_0$ and $\mathcal{L}_a$, the likelihoods for the two hypotheses.

   (d) Compute the likelihoods by implementing your answer from Q2.c using the provided dataset. [*Jupyter Notebook*]

   (e) Using the likelihood ratio test for a p-value of 0.05 (using a χ^2 table for 1 degree of freedom), can you safely reject the null hypothesis?

3. **Comparing Distributions.** The two provided datasets are *each* expression counts of 1000 genes across several replicates, with the expression compared in two conditions (for all genes and each replicate). One is simulated using a Poisson distribution, while the other is via a negative binomial (Pascal) distribution.

   (a) Find mean and variance for each gene, under each condition, for both sets. I.e. across replicates. Keep the two sets separate. Take a look at Q3.b for ideas on the form of the data structure to use. [*Jupyter Notebook*]

   (b) Create two separate plots for the 2 datasets, plotting the previously found means against variances. Determine which distribution is the basis for each (just based on visual interpretation). [*Jupyter Notebook*]

   (c) **Analysis of Individual Genes** As stated before, each dataset contains counts for genes expressed in two conditions. Many genes will be expressed similarly across both conditions, though not all.

      i. For each gene (and for both datasets), determine if there is a statistical difference in expression counts across the two conditions using the likelihood ratio test as in Q2, assuming a Poisson distribution for both. [*Jupyter Notebook*]

      ii. Compare between datasets the number of significant differences in gene expression across the two conditions i.e. number of genes expressed differently in dataset 1 vs. dataset 2. Using this comparison, determine the underlying distribution for each dataset. *Briefly* explain why (<10 words).