**NUIN 408 – Homework #1**

Due 4/13/2020 at 11:59 PM

This homework covers material from Lectures 1 – 4.

**Problem 1:**

A RNAseq experiment focuses on changes in the level of expression of three genes. The experiment is repeated 20 times. The resulting values of fold change for these three genes are:

Gene A: **5, 4, 4, 3, 4, 5, 4, 3, 5, 3, 2, 2, 6, 5, 4, 4, 4, 4, 2, 5**

Gene B: **6, 6, 7, 5, 6, 5, 7, 6, 5, 7, 8, 8, 7, 6, 5, 6, 5, 6, 4, 9**

Gene C: **5, 3, 2, 4, 6, 4, 3, 4, 5, 4, 1, 2, 2, 2, 3, 2, 6, 5, 3, 5**

For each one of these three genes, think of its fold change as a random variable with a probability distribution that is unknown to us. The listed values are 20 samples of the underlying probability distribution for each gene. For each of these three genes:

a)  List the values that this variable takes in the provided sample and create a vector with these values ordered from lowest to highest.

b)  Estimate the probability of each value taken by the random variable and create a vector with these probabilities, ordered as in (a).

c)  Show that the estimated probabilities add up to one

d)  Graph a bar chart that shows the sample probability distribution of the fold values.

e)  Compute the estimated mean (sample mean) of the random variable.

f)  Compute the estimated variance and standard deviation (sample variance and standard deviation) of the random variable.

**Problem 2:**

Consider two fair dice. The outcome of throwing each of them is a random variable X that can take six different values xi, all with the same probability pi=P(X=xi):

Consider a new random variable Y=X1+X2, where X1 is the random variable associated with the outcome of throwing the first dice, and X2 is the random variable associated with the outcome of throwing the second dice. The random variable Y describes the sum of these two outcomes.

Let’s characterize Y as a random variable:

a)  List the values that this random variable takes and their respective probabilities; create an appropriately ordered vector for each.

b)  Show that the true probabilities add up to one

c)  Graph a bar chart that shows the true probability distribution of this random variable.

d)  Compute the true mean of this random variable.

e)  Compute the true variance and the standard deviation of this random variable.

f) Compute the true variance and standard deviation of the variable X. How do these values compare to those of Y? What key principle does this demonstrate?

g)  Compute the Fano Factor and the Coefficient of Variation of the distribution of Y

h)  What is the name of the distribution of variable Y?

**Problem 3:**

Now instead of knowing the true probabilities, we are going to do a digital “experiment” to estimate the probability distribution of Y.

a) Run 100 random trials, rolling the two dice and collect the resulting distribution of total values.

b) What is the estimated probability of each of the possible values? Plot a histogram.

c) Re-run the experiment, 1000, 10,000, and 100,000 times. Use the subplot command, to plot all 4 histograms on the same figure. What happens when you run more trials?

d) Compute the variance of each of your 4 sample distributions. What is the trend as you increase the number of trials?

e) In a separate graph, plot the cumulative density function for each of these 4 experiments, this time on the same axes. Use ‘legend’ to label the lines.

**Problem 4:**

Consider a random variable K described by Poisson statistics:

𝑃(𝐾=𝑘)= 𝜆𝑘𝑒−𝜆 𝑘!

For this question, we will be characterizing the Poisson distribution for the following values of λ: 2, 4, 10, and 50.

a)  Plot each Poisson distribution on a single graph with an x-axis from 0 to 100. For smooth curves, use a step-size less than one.

b)  What are the mean and variance of each Poisson distribution? Do you know them without calculating?

c)  Intuitively, how do you expect the skewness to change as lambda increases? Calculate the skewness of each Poisson distribution. Look up the equation for the Skewness of a Poisson on Wikipedia.

d)  For each value of λ, plot the Poisson and Gaussian distribution with same mean and variance on one plot. Plot all four plots in a single figure using the subplot function.

e)  By eye, which value of λ gives the closest approximation to a Gaussian?

f)  Now plot a Q-Q plot for each value of λ versus the Gaussian fit. Plot all four Q-Q plots on a single figure using subplot. Which value of λ is the most Gaussian? How does the Q-Q plot demonstrate this?

g)  Describe the how a Poisson distribution relates to a Gaussian distribution as the value of λ increases.