Programming Stats Homework Questions

1. Recreate the sampling distribution of the sample proportion following the Monte Carlo pseudo code from lecture. Be sure to use a bin size on the histogram of 0.05.

2. Two six sided dice are rolled, and the sum of the face values is six. What is the probability that at least one of the dice came up a three? Answer this by using Monte Carlo approximation. Hint find a function to sample the numbers 1 through 6.

3. A closet contains n pairs of shoes. If 2r shoes are chosen at random ($2r < n$), what is the probability that there will be no matching pair in the sample? A moderately difficult counting problem. Solution: $\binom{2n}{2r}$ ways to choose $2r$ shoes from $2n$ shoes. Ways to choose non-matching shoes:

   $$\begin{array}{cc} L_1 & R_1 \\ L_2 & R_2 \\ \vdots & \\ L_n & R_n \end{array}$$

   No matching paris if only 1 from each 'row' of shoes.
   $\binom{n}{2r}$ ways to select 1 from each 'row.'
   For each selection, there are 2 possible choices (R or L), giving 2^{2r} total ways to select left and right shoes.
   Therefore,

   $$P(\text{non-match}) = \frac{\binom{n}{2r}2^{2r}}{\binom{2n}{2r}}$$

   Rather than trying to figure that out you might resort to computation to enumerate things. Use the combination function in python to validate the above formula for the case where you have 5 pairs of shoes (10 total) and you are selecting 4 shoes.

4. Working group has 8 members. Three will be assigned to lead the group (one head leader, one assistant leader, and one workflow specialist). How many ways to assign the leaders? Use python to create all possible assignments and verify the answer from the notes.

5. Write python code to complete the pseudo code given for the problem:
   Ex: Working group has 8 members. Three will be assigned to lead the group (no difference in leadership positions). If there are 3 women and 5 men in the group, what is the probability of exactly 1 woman being asigned to the leadership (if assigned randomly)?

6. A manufacturer sends 100 parts, 10 of which are defective. The batch will be returned if there are more 5 defective units found in a sample of 20 of the parts. Let Y be the number of defective parts in the sample of size 20. Then Y follows a hypergeometric distribution. Use the hypergeometric PMF from the scipy package to find the probability the batch is returned.

7. A manufacturing plant uses a specific bulk product. The amount used (in tons) in one day has a gamma distribution with $\alpha = 3/2$ and $\lambda = 1/3$. Using the scipy package:

a. Find the probability that the plant will use between 3 and 4 tons on a given day.

b. How much of the bulk product should be stocked so that the plant's chance of running out of the product is only 0.05. Hint you'll need to use the ppf method.

8. The proportion of impurities per batch in a chemical product is a random variable that is well modeled by a *Beta* distribution with $\alpha = 3$ and $\beta = 2$. A batch with more than 40% impurities cannot be sold.

   a. Find the probability that a randomly selected batch cannot be sold because of excessive impurities.

   b. What is the mean proportion of impurities? the variance? (Hint: use an appropriate method)

9. Suppose Y = the amount spent on electricity (in dollars) per month for residents has a mean of $83 and a standard deviation of $11. A random sample of 20 resident's electricity bills is found. Approximate the probability the sample mean electricity bill is less than $80.

10. Suppose you have a Binomial random variable with $n = 20$ trials and $p = 0.4$. Use the binomial PMF/CDF and the normal approximation to the binomial to find $P(Y < n * 0.4)$ and $P(Y \leq n * 0.3)$. Repeat with $n = 50$.

11. Complete the power simulation corresponding to the 'Benford's Law' example from the notes.

12. The datasets illinois60, ..., illinois64 contain rainfall data from storms in the region over those years. The data can reasonably be assumed to come from a gamma distribution.
    This gives a random sample of $n = 227$ RVs from a $gamma(\alpha, \lambda)$ distribution. This implies that the likelihood is

$$L(\alpha, \lambda) = \frac{\lambda^{n\alpha}}{(\Gamma(\alpha))^n}\left(\prod_{i=1}^{n} y_i\right)^{\alpha-1} e^{-\lambda\sum_{i=1}^{n} y_i}$$

This gives a log likelihood of

$$l(\alpha, \lambda) = n\alpha ln(\lambda) - nln(\Gamma(\alpha)) + (\alpha - 1)\sum_{i=1}^{n} ln(y_i) - \lambda\sum_{i=1}^{n} y_i$$

The partial derivative are

$$\frac{\partial l(\alpha, \lambda)}{\partial \alpha} = nln(\lambda) - n\frac{\partial}{\partial \alpha}ln(\Gamma(\alpha)) + \sum_{i=1}^{n} ln(y_i)$$

$$\frac{\partial l(\alpha, \lambda)}{\partial \lambda} = n\alpha/\lambda - \sum_{i=1}^{n} y_i$$

There are no closed forms for the MLEs of λ or α. Use the observed data from the storms (you'll need to read in and combine the datasets) and gradient descent to find the approximate MLEs. The digamma function may be useful. Use alpha = 1 and lambda = 1 as initial values. Use a step size of 0.00001 and a tolerance of 0.000001.

13. The idea of a confidence level, say $(1-\alpha)*100\%$, is that in a large number of intervals created $(1-\alpha)$ of the intervals would contain the true value. Let's simulate this idea. Simulate 10,000 datasets each of size 15 from a normal distribution with mean 5 and standard deviation 1. For each data set calculate a 95% CI for μ:

$$\bar{y} \pm 1.96(1/\sqrt{15})$$

Determine the proportion of intervals that contain the true value of $\mu = 5$.

14. The CLT says that sample averages are observed in a normal distribution pattern if the sample size is 'large' enough. Let's simulate to check a particular scenario. Simulate 10,000 datasets each of size 30 from an exponential distribution with rate parameter equal to 1. Find the sample mean for each data set and plot all the sample means on a histogram. Use 25 bins and give the plot appropriate labels and a title. Overlay the corresponding normal distribution given by the CLT (you may need to look up the mean and variance of the exponential distribution). Repeat this for a sample size of 100. Is the rule of thumb $n > 30$ working here?