CSC380 **Homework 3** **Fall 2024**

This homework is due on Friday, October 11^{th} at 11.59pm. Late submissions are not accepted.

Submission Guidelines

- You must make two submissions:
  - Your complete homework as a SINGLE PDF file by the stated deadline to the gradescope. **Include your code and output of the code as texts in the PDF.**
  - Your codes to a separate submission: a single notebook file including codes for questions 7 through 11.
- For your PDF submissions:
  - Select the page number for the answer to each question in the Gradescope.
  - You may submit typed or handwritten/scanned answers. If you decide to submit handwritten answers then please ensure that it is easily readable.
  - You can easily scan and upload your answers as a PDF using the Gradescope mobile app.

1. **[10pts]** Suppose that the height of men has mean 68 inches and standard deviation 4 inches. We draw 100 men at random. Find (approximately) the probability that the average height of men in our sample will be at least 68.5 inches.

2. **[10pts]** Suppose we have a book consisting of $n = 100$ pages. The number of misprints at each page is independent and has a Poisson distribution with mean 1. Find the probability that the total number of misprints is at least 80 and at most 90 using central limit theorem.

3. **[5pts]** Let $X_1, \cdots, X_n \sim$Poisson(λ) and let $\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Find the mean squared error of this estimator.

4. **[10pts]** We would like to build a simple model to predict the number of traffic accidents at a junction. The number of accidents is modeled as Poisson distributed. Recall that the Poisson is a discrete distribution over the number of arrivals (accidents) in a fixed time-frame. It has the probability function:

$$\text{Poisson}(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

The parameter λ is the *rate* parameter that represents the expected number of traffic accidents $E(x) = \lambda$ in a month. To fit the model we need to estimate the rate parameter using some data $X_1, \cdots, X_n$, representing the number of accidents in a sample of n months. For this purpose first write the logarithm of the joint probability distribution $\log p(X_1, \cdots, X_n; \lambda)$ using summations.

5. **[10pts]** Compute the maximum likelihood estimate of the rate parameter which maximizes the joint probability found in Question 4 by finding the zero-derivative solution.

6. **[5pts]** How many accidents are expected in the next month under this model, if in the last three months $X_1 = 2, X_2 = 5, X_3 = 3$ accidents were observed?

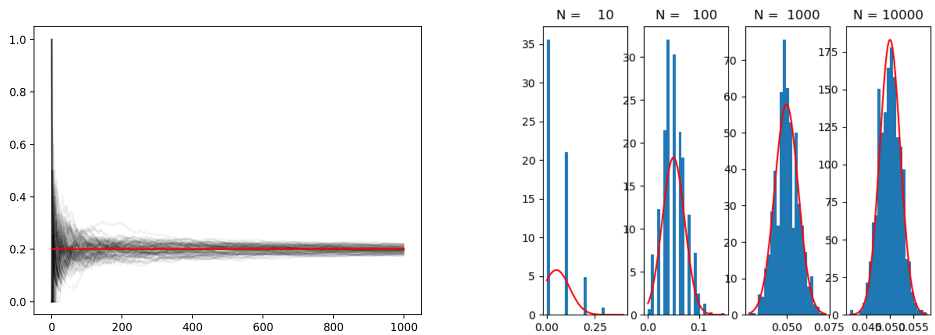

Figure 1: *Left:* Ideal plot for 100 sample mean trajectories of Question 7. *Right:* Ideal plot for Question 8.

7. **[12pts]** Let us numerically verify the law of large numbers. We will simulate $m = 100$ sample mean trajectories of $X_1, \ldots, X_N \sim \text{Bernoulli}(\mu = 0.2)$ and plot them altogether in one plot. Here, a sample mean trajectory means a sequence of $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_N$ where $\bar{X}_i$ is the sample mean using samples $X_1, \ldots, X_i$. We will plot $\bar{X}_n$ as a function of n, but do this multiple times. Take n from 1 to $N = 1000$. An ideal plot would look like Figure 1-Left. You must use the 'alpha' option to pyplot.plot() to give some transparency (you should obtain a similar look visualization as the one in figure). You may want to use the 'color' option to specify the color.

8. **[12pts]** Let us verify the central limit theorem (CLT) by simulation. For $N \in \{10, 100, 1000, 10000\}$, perform:

   - Take N samples from Bernoulli($\mu = 0.05$) and compute the sample mean. Repeat this 1000 times.
   - Plot those 1000 numbers as a histogram (pyplot.hist) with a proper number of bins. Use density=True.
   - With a red line, overlay the pdf of a Gaussian distribution with the parameters suggested by the CLT (figure this out!).

   An ideal answer would look like Figure 1-Right. To receive full credit, you must use pyplot.subplot to have four plots in one figure.

This question shows a way of estimating the correlation ρ of two random variables X, Y. For our chosen model, we will use a bivariate Gaussian distribution $(X, Y)^T$. Note that such a distribution denoted with $\mathcal{N}(\mu, \Sigma)$, has two parameters, where μ denotes the 2-dimensional mean vector consisting of μ_x, μ_y and Σ denotes the covariance matrix. The entry $(1, 1)$ of Σ is $Cov(X, X)$, the entry $(2, 2)$ is $Cov(Y, Y)$, and the entries $(1, 2), (2, 1)$ are $Cov(X, Y)$. For this example the means are $\mu_x = \mu_x = 0$, the standard deviations are $\sigma_x = \sigma_y = 1$, and the true (unknown) correlation is $\rho = 0.6$. Therefore the covariance matrix is

$$\begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$$

Using **numpy.random.seed** set your random number generator seed to 0 and answer the following:

9. **[10pts]** Create a dataset by drawing $N = 500$ samples from our model using **numpy.random** function **multivariate_normal**. Compute and report the plug-in estimator of correlation, given by:
$$\hat{\rho} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_j (Y_j - \bar{Y})^2}}$$
Where $\bar{X} = \frac{1}{N} \sum_i X_i$ is the sample mean (and similarly for $\bar{Y}$).

10. **[10pts]** Repeat the above process $m = 5,000$ times to generate $\hat{\rho}_1, \ldots, \hat{\rho}_m$, each one based on a fresh set of $N = 500$ samples. Display a histogram of your m estimates using **matplotlib.pyplot.hist** with 30 bins. Label the axes.

11. **[6pts]** Use m estimates obtained in the above question to estimate $\mathbb{E}\left[(\hat{\rho} - \rho)^2\right]$, the mean square error (MSE) of plug-in estimator $\hat{\rho}$. What is the value of your MSE estimate?