SI 140A-02 Probability & Statistics for EECS, Fall 2024
Homework 10

Name: **Wenye Xiong**
Student ID: 2023533141

Due on Dec. 17, 2024, 11:59 UTC+8

Read all the instructions below carefully before you start working on the assignment, and before you make a submission.

- You are required to write down all the major steps towards making your conclusions; otherwise you may obtain limited points of the problem.
- Write your homework in English; otherwise you will get no points of this homework.
- Any form of plagiarism will lead to 0 point of this homework.

Problem 1

Let $U_i \sim \text{Unif}(0,1), i \geq 1$ be i.i.d. random variables. Define N as follows:

$$N = \max\left\{n : \prod_{i=1}^{n} U_i \geq e^{-1}\right\}$$

(a) Estimate $\text{E}(N)$ by generating 5000 samples of N and then use the sample mean.

(b) Estimate $\text{Var}(N)$.

(c) Estimate $P(N = i)$, for $i = 0, 1, 2, 3$.

(d) Can you find the exact distribution of N ?

Solution

(a), (b) and (c) can be solved by simulation. You can see the code in the appendix.
(d) We can find the exact distribution of N as follows:
Let's first transform the problem:

$$\begin{aligned} N &= \max\{n : \prod_{i=1}^{n} U_i \geq e^{-1}\} \\ &= \max\{n : \sum_{i=1}^{n} \ln(U_i) \geq -1\} \end{aligned}$$

Key observations:

- If $U \sim \text{Uniform}(0,1)$, then $-\ln(U) \sim \text{Exponential}(1)$
- Let $Y_i = -\ln(U_i)$, then $Y_i \sim \text{Exponential}(1)$
- Sum of exponential random variables follows Gamma distribution
- $S_n = \sum_{i=1}^{n} Y_i \sim \text{Gamma}(n, 1)$

Therefore:

$$\begin{aligned} P(N = k) &= P(S_k \leq 1 \text{ and } S_{k+1} > 1) \\ &= P(S_k \leq 1) - P(S_{k+1} \leq 1) \\ &= \frac{\gamma(1,k)}{\Gamma(k)} - \frac{\gamma(1,k+1)}{\Gamma(k+1)} \end{aligned}$$

where $\gamma(1,k)$ is the lower incomplete gamma function and $\Gamma(k)$ is the gamma function.
This gives us the exact probability mass function for N.

Problem 2

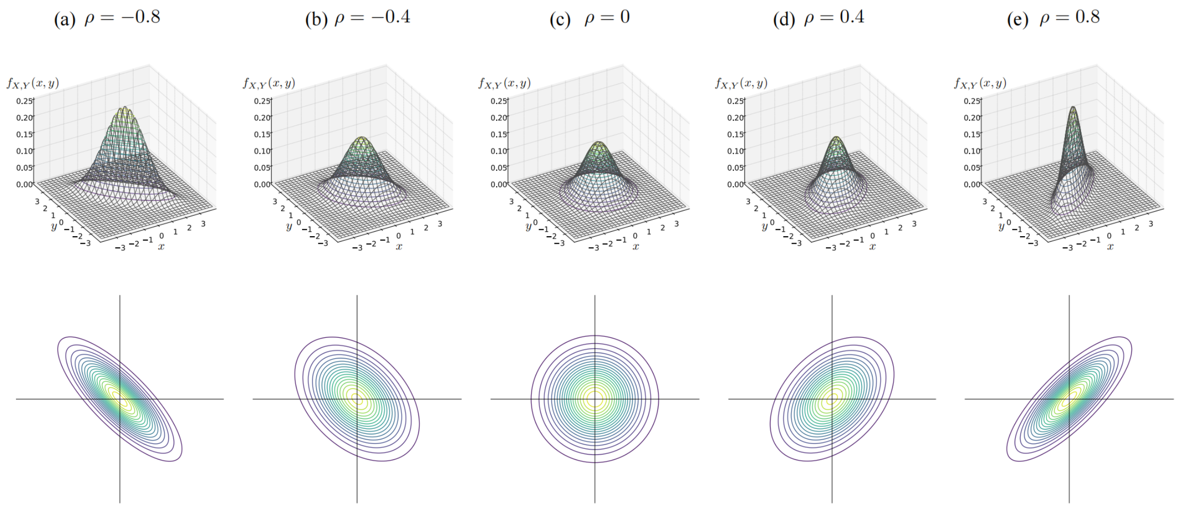
Visualization of Bivariate Normal Distribution with correlation coefficient ρ

(a) Use the following transformation to generate samples from bivariate Normal distribution with correlation coefficient ρ :

$$X = Z$$
$$Y = \rho Z + \sqrt{1-\rho^2}W$$

where $-1 < \rho < 1, Z$ and W are i.i.d. random variables satisfying $\mathcal{N}(0, 1)$.

(b) Plot the joint pdf function and the corresponding contour(or isocontour) as the following figure:



Solution

See the code in the appendix for the implementation of the above problem.

Problem 3

Let $X_1 \sim \text{Expo}(\lambda_1), X_2 \sim \text{Expo}(\lambda_2)$ and $X_3 \sim \text{Expo}(\lambda_3)$ be independent.

(a) Find $E(X_1 \mid X_1 > 2024)$

(b) Find $E(X_1 \mid X_1 < 1997)$

(c) Find $E(X_1 + X_2 + X_3 \mid X_1 > 1997, X_2 > 2014, X_3 > 2025)$ in terms of $\lambda_1, \lambda_2, \lambda_3$.

Solution

(a)
According to the memoryless property of exponential distribution, we have

$$E(X_1 \mid X_1 > 2024) = 2024 + E(X_1 - 2024 \mid X_1 > 2024) = 2024 + E(X_1) = 2024 + \frac{1}{\lambda_1}$$

(b)
Similarly, we have

$$E(X_1 \mid X_1 < 1997) = 1997 - E(1997 - X_1 \mid X_1 < 1997) = 1997 - E(X_1) = 1997 - \frac{1}{\lambda_1}$$

(c)
Since X_1, X_2, X_3 are independent, we have

$$\begin{gathered} E(X_1 + X_2 + X_3 \mid X_1 > 1997, X_2 > 2014, X_3 > 2025) = \\ E(X_1 \mid X_1 > 1997) + E(X_2 \mid X_2 > 2014) + E(X_3 \mid X_3 > 2025) \\ = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \frac{1}{\lambda_3} + 1997 + 2014 + 2025 = 6036 + \frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \frac{1}{\lambda_3} \end{gathered}$$

Problem 4

Let X and Y be two continuous random variables with joint PDF

$$f_{X,Y}(x,y) = \begin{cases} 6xy & \text{if } 0 \le x \le 1, 0 \le y \le \sqrt{x} \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the marginal distributions of X and Y. Are X and Y independent?

(b) Find $E[X \mid Y = y]$ and $\text{Var}[X \mid Y = y]$ for $0 \le y \le 1$.

(c) Find $E[X \mid Y]$ and $\text{Var}[X \mid Y]$.

Solution

(a) The supports of X and Y are both [0,1]. In this way, we have
and

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \\ &= \int_0^{\sqrt{x}} 6xydy \\ &= 3xy^2\Big|_{y=0}^{y=\sqrt{x}} \\ &= 3x^2, \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx \\ &= \int_{y^2}^{1} 6xydx \\ &= 3yx^2\Big|_{x=y^2}^{x=1} \\ &= 3y - 3y^5. \end{aligned}$$

$$f_X(x) = \begin{cases} 3x^2 & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$f_Y(y) = \begin{cases} 3y - 3y^5 & \text{if } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Since $f_{X,Y}(x,y) \neq f_X(x)f_Y(y), X$ and Y are not independent.
Therefore,

$$E[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx,$$

to calculate $E[X|Y=y]$, we need to first calculate $f_X|Y(x|y)$.
If $y^2 \le x \le 1$,

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{2x}{1-y^4}.$$

In this way,

$$f_{X|Y}(x|y) = \begin{cases} \frac{2x}{1-y^4} & \text{if} y^2 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

(b) Since

$$E[X|Y=y]=\int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx,$$

to calculate $E[X|Y=y]$, we need to first calculate $f_X|Y(x|y)$.
If $y^2 \le x \le 1$,

$$f_{X|Y}(x|y)=\frac{f_{X,Y}(x,y)}{f_Y(y)}=\frac{2x}{1-y^4}.$$

In this way,

$$f_{X|Y}(x|y)=\begin{cases} \frac{2x}{1-y^4} & \text{if} y^2 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned} E[X \mid Y=y] &= \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)dx \\ &= \int_{y^2}^{1} x \frac{2x}{1-y^4}dx \\ &= \frac{2}{3\left(1-y^4\right)} x^3 \Big|_{x=y^2}^{x=1} \\ &= \frac{2\left(1-y^6\right)}{3\left(1-y^4\right)} \\ &= \frac{2}{3} \cdot \frac{1+y^2+y^4}{1+y^2} \end{aligned}$$

Since

$$\operatorname{Var}[X \mid Y=y]=E\left[X^2 \mid Y=y\right]-(E[X \mid Y=y])^2$$

to calculate $\operatorname{Var}[X \mid Y=y]$, we need to first calculate $E\left[X^2 \mid Y=y\right]$. Since

$$\begin{aligned} E\left[X^2 \mid Y=y\right] &= \int_{-\infty}^{\infty} x^2 f_{X|Y}(x \mid y)dx \\ &= \int_{y^2}^{1} x^2 \frac{2x}{1-y^4}dx \\ &= \frac{1}{2\left(1-y^4\right)} x^4 \Big|_{x=y^2}^{x=1} \\ &= \frac{1-y^8}{2\left(1-y^4\right)} \\ &= \frac{1+y^4}{2} \end{aligned}$$

we have,

$$\begin{aligned} \operatorname{Var}[X \mid Y=y] &= E\left[X^2 \mid Y=y\right]-(E[X \mid Y=y])^2 \\ &= \frac{1+y^4}{2}-\left(\frac{2\left(1-y^6\right)}{3\left(1-y^4\right)}\right)^2 \\ &= \frac{1+y^4}{2}-\frac{4}{9} \cdot \frac{\left(1+y^2+y^4\right)^2}{\left(1+y^2\right)^2} \end{aligned}$$

(c) According to the result in question(b), we have

$$E[X \mid Y] = \frac{2}{3} \cdot \frac{1 + Y^2 + Y^4}{1 + Y^2}$$
$$\mathrm{Var}[X \mid Y] = \frac{1 + Y^4}{2} - \frac{4}{9} \cdot \frac{\left(1 + Y^2 + Y^4\right)^2}{\left(1 + Y^2\right)^2}$$

Problem 5

Let X be a discrete r.v. whose distinct possible values are $x_0, x_1, \ldots$, and let $p_k = P\left(X = x_k\right)$. The entropy of X is $H(X) = \sum_{k=0}^{\infty} p_k \log_2\left(1/p_k\right)$.

(a) Find $H(X)$ for $X \sim \text{Geom}(p)$.

(b) Let X and Y be i.i.d. discrete r.v.s. Show that $P(X = Y) \geq 2^{-H(X)}$.

Solution

(a)
The PMF of X is $P(X = k) = p(1-p)^k$ since $X \sim \text{Geom}(p)$. Thus we have

$$\begin{aligned}
H(X) &= -\sum_{k=0}^{\infty} p(1-p)^k \log_2\left(p(1-p)^k\right) \\
&= -p\sum_{k=0}^{\infty}(1-p)^k \log_2(1-p)^k - p\sum_{k=0}^{\infty}(1-p)^k \log_2 p \\
&= -p\log_2(1-p)\sum_{k=0}^{\infty} k(1-p)^k - p\log_2 p\sum_{k=0}^{\infty}(1-p)^k \\
&= -\log_2 p - \frac{1-p}{p}\log_2(1-p)
\end{aligned}$$

(b)
Since X and Y are i.i.d random variables, via LOTP, we have

$$\begin{aligned}
P(X = Y) &= \sum_{k=0}^{\infty} P(X = Y \mid Y = k) \cdot P(Y = k) \\
&= \sum_{k=0}^{\infty} P(X = k) \cdot P(Y = k) \\
&= \sum_{k=0}^{\infty} p_k^2
\end{aligned}$$

Denote Z as a new discrete random variable such that $P\left(Z = p_k\right) = p_k$, then we have:

$$E(Z) = \sum_{k=0}^{\infty} p_k \times p_k = P(X = Y)$$

Since $\log(\cdot)$ is a convex function, according to Jensen's inequality, we have $E(\log(Z)) \leq \log(E(Z))$, thus there is

$$\begin{aligned}
&\sum p_k \log_2 p_k \leq \log_2 \sum p_k^2 \\
\Leftrightarrow &-H(X) \leq \log_2 P(X = Y) \\
\Leftrightarrow &P(X = Y) \geq 2^{-H(X)}
\end{aligned}$$

Problem 6

Instead of predicting a single value for the parameter, we give an interval that is likely to contain the parameter: A $1-\delta$ confidence interval for a parameter p is an interval $[\hat{p}-\epsilon, \hat{p}+\epsilon]$ such that $Pr(p \in [\hat{p}-\epsilon, \hat{p}+\epsilon]) \geq 1-\delta$. Now we toss a coin with probability p landing heads and probability $1-p$ landing tails. The parameter p is unknown and we need to estimate its value from experiment results. We toss such coin N times. Let $X_i = 1$ if the i th result is head, otherwise 0 . We estimate p by using

$$\hat{p} = \frac{X_1 + \ldots + X_N}{N}$$

Find the $1-\delta$ confidence interval for p, then discuss the impacts of δ and N.

(a) Method 1: Adopt Chebyshev inequality to find the $1-\delta$ confidence interval for p, then discuss the impacts of δ and N.

(b) Method 2: Adopt Hoeffding bound to find the $1-\delta$ confidence interval for p, then discuss the impacts of δ and N.

(c) Discuss the pros and cons of the above two methods.

Solution

Since $X_i \sim \text{Bern}(p)$, where $X_i \in \{0,1\}$, we have $\mathbb{E}[X_i] = p$ and $\mathbb{V}[X_i] = p(1-p)$. Therefore, we have:

$$\mathbb{E}[\hat{p}] = p, \quad \mathbb{V}[\hat{p}] = \frac{p(1-p)}{N}$$

We need to find the interval $[\hat{p}-\epsilon, \hat{p}+\epsilon]$ such that:

$$P(p \in [\hat{p}-\epsilon, \hat{p}+\epsilon]) \geq 1-\delta \quad \Leftrightarrow \quad P(|\hat{p}-p| \geq \epsilon) \leq \delta$$

(a) Using Chebyshev's inequality on the random variable $\hat{p}$, we get:

$$P(|\hat{p}-p| \geq \epsilon) \leq \frac{p(1-p)}{N\epsilon^2}$$

Setting the right-hand side equal to δ, we have:

$$\delta = \frac{p(1-p)}{N\epsilon^2} \quad \Rightarrow \quad \epsilon = \sqrt{\frac{p(1-p)}{N\delta}}$$

Thus, δ negatively correlates with ϵ. Given a fixed number of samples N, there is a trade-off between accuracy and confidence. Specifically: 1. Fixing the confidence interval (parametrized by δ), reducing the estimation error ϵ requires increasing the number of samples N. 2. Fixing the estimation error ϵ, narrowing the confidence interval requires increasing the number of samples N.

(b) Using Hoeffding's inequality on the random variable $\hat{p}$, we get:

$$P(|\hat{p}-p| \geq \epsilon) \leq 2e^{-2N\epsilon^2}$$

Setting the right-hand side equal to δ, we have:

$$\delta = 2e^{-2N\epsilon^2} \quad \Rightarrow \quad \epsilon = \sqrt{\frac{\ln(2/\delta)}{2N}}$$

The effects of δ and N are similar to those discussed in (a).

(c) Comparison of the two methods:

Chebyshev's Inequality: - Pros: 1. Provides a general bound that is sharp and cannot be improved in general. 2. Can be improved with additional distributional information on polynomial moments. - Cons: 1. Requires the existence of moments up to the second order. 2. Has a quadratic convergence rate.

Hoeffding's Inequality: - Pros: 1. Provides an exponential convergence rate. 2. Does not require assumptions on moments. - Cons: 1. Works only for sub-Gaussian distributions. 2. Generally not sharp when the variance is small.

Problem 7

(**Optional Challenging Problem**) We consider the progressive Monty Hall problem. This time we assume there are n identical doors, where n is an integer satisfying $n \geq 3$. One door conceals a car, the other $n-1$ doors conceal goats. You choose one of the doors at random but do not open it. Monty then opens a door he knows to conceal a goat, always choosing randomly among the available doors. At this point he gives you the option either of sticking with your original door or switching to one of the remaining doors. You make your decision. Monty now eliminates another goat-concealing door (at random) and once more gives you the choice either of sticking or switching. This process continues until only two doors remain in play. What strategy should you follow to maximize your chances of winning? We consider three strategies: (1) Select a door at random and stick with it throughout. (2) Select a door at random, then switch doors at every opportunity, choosing your door randomly at each step. (3) Select a door at random, stick with your first choice until only two doors remain, and then switch. When $n = 10$ and $n = 1000$, **please run simulations to estimate the winning probability of each strategy. Check which strategy is best and provide the corresponding intuition.**