Probability and Random Variables with Python Simulations

Applied Stochastic Processes: HW 2

Due Date: October 1, 2024

Objectives

This assignment is designed to help students develop a deep understanding of probability and random variables. Students will enhance their problem-solving skills and gain practical experience through Python-based simulations.

Policy

- You can discuss HW problems with other students, but the work you submit must be written in your own words and not copied from anywhere else. This includes codes.
- However, do write down (at the top of the first page of your HW solutions) the names of all the people with whom you discussed this HW assignment.
- You may decide to write out your solution with pen and paper and use a scanning app to turn in a PDF submission or may choose to type out your solution with LATEX. We strongly encourage the latter.

Expected Topics Covered

1. Discrete and Continuous Random Variables
2. Expectation and Variance
3. Two Random Variables and Conditional Probability
4. Random Vectors

Submission Guidelines

- Submit your code solutions as a Jupyter notebook file (.ipynb) or as Python scripts (.py). You can also convert them to pdf and attach them to your final submission document. Be sure to indicate which code belongs to what question.
- For those using **LaTeX**, you can paste your code by importing the python environment `pythonhiglights`.

```
\usepackage{pythonhiglights}

\begin{python}
# import necessary libraries
import math
import random

# example arithmetic evaluation
a, b = 2, 3
c = a + b
print(f"The sum of the numbers {a} and {b} is {c}")
\end{python}
```

- Ensure all code is well-documented and includes comments explaining each step.
- Provide a brief report summarizing your findings and the results of your simulations.

Q1: Discrete Random Variables and Real Life Applications (16 Points)

Reading: Discrete random variables are fundamental components in the field of probability and statistics. They are defined as variables that can take on a countable number of distinct values, often associated with counting processes or categorical outcomes. For instance, the number of heads obtained when flipping a coin multiple times, or the number of customers arriving at a store within an hour, are both examples of discrete random variables.

These variables are typically described by a probability mass function (PMF), which assigns the probability to each possible value that the random variable can take. The sum of all probabilities in the PMF must equal one, satisfying the fundamental property of probability.

Discrete random variables are often classified further into types, such as binomial or Poisson random variables, each characterized by their unique properties and applications. The binomial random variable arises in contexts where there are fixed numbers of independent trials, each with two possible outcomes (success or failure). Meanwhile, the Poisson random variable is useful for modeling the number of events occurring within a fixed interval of time or space, given a known average rate of occurrence.

1. A factory produces electronic components, with each component either passing a quality check or being rejected. Let X represent the number of components that pass out of 5 tested components in a day, where the probability of each component passing the quality check is $p = 0.8$.

   (a) **(2 points):** Define the probability mass function (PMF) for X as a binomial distribution. Show that the total probability is 1 by summing the PMF over all possible values of X.

   (b) **(3 points):** Suppose the factory wants to predict the likelihood of a specific number of components passing the quality check. Find the expected value and variance of X, and explain how the factory can use this information to estimate daily production quality.

   (c) **(3 points):** Calculate the probability that exactly 3 out of 5 components pass the quality check using the PMF. Based on this, discuss how rare or frequent it is for the factory to encounter this scenario, and explain the implications for managing production quality.

2. A warehouse tracks the number of defective products returned by customers daily. Let X represent the number of defective products returned on a given day. The company has noticed that on average, 4 defective products are returned daily. Based on historical data, the number of defective products returned follows a Poisson distribution with parameter $\lambda = 4$, which represents the average number of returns per day.

   (a) **(2 points):** Define the PMF for X, where X follows a Poisson distribution with parameter $\lambda = 4$. Verify that the sum of all probabilities over the range of X (from 0 to infinity) equals 1, proving that the PMF is valid.

   (b) **(3 points):** Find the expected value $E(X)$ and the variance $\text{Var}(X)$ of the number of defective products returned. Explain the significance of these values for the warehouse's operations and how they can be used to set performance targets or identify unusual trends in returns.

   (c) **(3 points):** Calculate the probability that the warehouse will receive fewer than 3 defective products on a given day, i.e., $P(X < 3)$. Discuss what this probability means in terms of real-life decision-making for warehouse management, particularly in planning inventory and customer service responses.

Q2: Asymptotic Relationship Between Binomial and Poisson Distributions (18 Points)

Discrete random variables are essential in probability and statistics, representing countable outcomes. Two key types are the Binomial and Poisson distributions, each with unique applications.
The **binomial distribution** is used for a fixed number of independent trials with two outcomes: success or failure. For instance, flipping a coin 10 times results in a binomial distribution for the number of heads. The probability mass function (PMF) for a binomial random variable X with parameters n (trials) and p (success probability) is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

This distribution is ideal for discrete events with a set number of trials.
In contrast, the **Poisson distribution** models the number of events in a fixed time or space interval, assuming a constant average rate and independence. For example, it can represent the number of phone calls received by a call center in an hour. The Poisson random variable Y with parameter λ (average rate) is given by:

$$P(Y = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

This distribution is particularly useful for rare events over many trials. The relationship between binomial and Poisson distributions emerges when n is large and p is small, keeping $\lambda = np$ constant. Here, the binomial distribution $B(n, p)$ can be approximated by a Poisson distribution with parameter λ, facilitating easier calculations in large datasets.
In summary, while both distributions model discrete random variables, they fit different scenarios: the binomial for fixed trials with a constant success probability, and the Poisson for event counts over intervals. Understanding their relationship aids in accurately modeling discrete events in real-world situations.

A small factory produces electronic components where each component has a probability $p = 0.01$ of being defective. The factory inspects batches of 100 components each day.

Proof of Concept (6 points)

(a) **(2 points)** Let X be the number of defective components in a batch of $n = 100$ components, where each component has a defect probability $p = 0.01$. Show that as n becomes large and p becomes small such that $\lambda = np$ remains constant, the binomial distribution $X \sim B(n, p)$ approaches a Poisson distribution with parameter $\lambda = np$.

(b) **(2 points)** Prove that:

$$\lim_{n \to \infty,\, p \to 0} P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where $\lambda = np$ is fixed.

(c) **(2 points)** Provide a detailed explanation of why the Poisson distribution is a good approximation for the binomial distribution in this context.

Poisson Approximation for Defect Probability (4 points)

(a) **(1 point)** Using the Poisson approximation, calculate the probability of finding exactly 3 defective components in a batch of 100.

(b) **(1 point)** Compute the probability of finding at most 5 defective components using the Poisson distribution.

(c) **(2 points)** Explain the implications of these probabilities for quality control in the factory, and discuss how this approximation simplifies the analysis.

Simulation Exercise: Empirical Comparison (8 points)

In the above question, we calculated probabilities related to defective components in a batch using the Poisson approximation. This theoretical framework helps us understand the defect probabilities in a large batch of components where the average number of defects is relatively small.
To deepen our understanding and validate the accuracy of the Poisson approximation, we will now perform a simulation exercise. This exercise will help us empirically compare the binomial distribution (which is more accurate for small defect probabilities) with the Poisson distribution (which is used as an approximation).

(a) Simulate 10,000 batches of 100 components each using the binomial distribution $B(100, 0.01)$.

(b) Simulate 10,000 batches of 100 components each using the Poisson distribution with parameter $\lambda = 1$.

(c) Plot histograms comparing the two distributions. Calculate and compare the means and variances of both simulated distributions.

(d) Analyze the results to determine how well the Poisson distribution approximates the binomial distribution in practice. Discuss any discrepancies and their possible causes.

Q3: Poisson Distribution in Real-World Applications (35 points)

Reading: In many real-world situations, we encounter events that occur independently within a fixed time frame. For instance, the number of phone calls made in a minute can be modeled using the Poisson distribution. This distribution is often applied in scenarios where events occur randomly but with a known average rate. The Poisson distribution is characterized by a single parameter, λ, which represents the expected number of events (in this case, calls) within a given time period.
In telecommunication systems, the Poisson distribution helps predict the number of incoming or outgoing calls, assisting companies like Airtel Rwanda in optimizing their network resources and setting pricing strategies. By analyzing how call patterns change over time, companies can adjust their tariffs to maximize returns, while maintaining affordability for customers.
In this exercise, you will use the Poisson distribution to model the number of calls made per minute by students and calculate the expected revenue before and after a tariff reduction. Simulations will help to analyze how these changes impact Airtel Rwanda's revenue.

Airtel Rwanda provides communication services to high school students in Rwanda at a subsidized cost. To maximize their market returns, their Data Analyst collects data and reports that students make an average of 5 calls per minute with a 7:3 ratio for intra-network to inter-network calls. Following her recommendation to reduce call tariffs, her subsequent monthly report reveals that the average number of calls increases by 2 calls per minute.

Given the following tariff information:

- **Before reduction:**
  - Intra-network calls: 70 RwF per minute
  - Inter-network calls: 90 RwF per minute
- **After reduction:**
  - Intra-network calls: 60 RwF per minute
  - Inter-network calls: 80 RwF per minute

Assume the number of calls follows a Poisson distribution and that a cap of 10 calls are made per minute.

Questions:

1. **(1 point)** What is the probability that exactly 7 calls are made in a given minute before the tariff reduction? Use the Poisson distribution with parameter $\lambda = 5$.
2. **(1 point)** What is the probability that fewer than 3 calls are made in a given minute after the tariff reduction? Use the Poisson distribution with parameter $\lambda = 7$.
3. **(1 point)** Before the tariff reduction, what is the expected number of intra-network and inter-network calls per minute based on the 7:3 ratio?
4. **(2 points)** After the tariff reduction, what is the expected number of intra-network and inter-network calls per minute, assuming the same ratio holds?
5. **(2 points)** Calculate the expected revenue per minute before and after the tariff reduction based on the expected number of calls and the given tariffs.
6. **(8 points)** Simulate the probability that the revenue in a given minute exceeds 500 RwF before and after the tariff reduction for 100,000 repetitions. (Hint: For each case, calculate the Poisson parameter for total calls $\lambda = 5$ before and $\lambda = 7$ after the reduction, simulate the number of calls per minute using the Poisson distribution, and compute the probability that the revenue exceeds 500 RwF.)

Simulation Exercise: Revenue Analysis Before and After Tariff Reduction (20 points)

Tasks:

1. Simulate the number of calls per minute before and after the tariff reduction using the Poisson distribution.
2. Calculate the revenue for each minute based on the simulated number of calls and the given tariff rates.
3. Plot the revenue distribution using histograms to visualize the revenue per minute before and after the tariff reduction.
4. Use a line plot instead of histograms to visualize the trend in revenue over time (minutes in a day).
5. Calculate the total revenue in a day (assume 1440 minutes) before and after the tariff reduction.
6. Compute the percentage increase in revenue after the tariff reduction compared to before.
7. Interpret the histograms and the line plot to understand the distribution and central tendencies of revenue before and after the tariff reduction. Discuss the implications of the revenue distribution for Airtel Rwanda's strategy in maintaining an affordable service while maximizing revenue.

Q4: Expectation, Variance, and Moments of Discrete Random Variables (13 points)

Reading: In probability theory and statistics, expectation and variance are key to understanding random variables. The **expectation** (or mean) of a random variable quantifies its central tendency, reflecting the average value expected over many repetitions of an experiment. For a discrete random variable X with probability mass function $P(X = x_i)$, the expectation is calculated as:

$$E[X] = \sum_i x_i \cdot P(X = x_i)$$

Variance, in contrast, indicates the spread or variability of a random variable around its mean. It is defined as the expected squared deviation from the mean, quantifying how outcomes differ from the expected value:

$$\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

Expectation and variance are examples of **moments**, specific mathematical quantities that capture different characteristics of a random variable's distribution. The first moment is the expectation, while the second moment relates to the variance. Higher moments, such as the third and fourth, describe skewness and kurtosis, measuring asymmetry and "tailedness" of the distribution, respectively thus, aiding in the analysis of distribution shapes.
In practical applications, expectation and variance offer insights into real-world situations. For example, in finance, the expected value can represent the average return on an investment, while variance indicates its risk or volatility. Understanding these concepts is essential for informed decision-making across fields like economics, engineering, and decision theory.

1. A programming problem set has three tasks to solve for an interview of interns. For each task solved, a dollar is awarded to the prospective intern. Let X represent the amount of money earned by an intern, which depends on how many tasks they solve. The distribution of solving the problems is such that it is twice as likely to solve one task as it is to solve no task. It is four times and three times as likely to solve two tasks and three tasks, respectively, as it is to solve no task. If a candidate does not solve any task in 2 out of 5 attempts, answer the following:

   (a) **(2 points)** What is the average amount of money a candidate can be rewarded with? (**Hint:** Use the probabilities for solving 0, 1, 2, and 3 tasks to compute the expectation.)

   (b) **(3 points)** How would the amount of money earned vary with respect to the number of tasks solved? (**Hint:** Calculate the variance of X to assess the variability in rewards.)

2. In a series of independent trials with probability 0.3 of success, we want to calculate the probability that the first success occurs on the 4th trial. Let X_1 be a random variable representing the number of trials until the first success.

   (a) **(1 point)** What is the probability that the first success occurs on the 4th trial? (**Hint:** Use the PMF of the geometric distribution.)

   $$P(X_1 = k) = (1-p)^{k-1}p$$

   (b) During a fishing competition, each angler has a 10% chance of catching a fish with each cast. Assuming independence in each cast, use X_r to represent the number of casts needed to secure r successful catches.

      (i) **(1 point)** Calculate $P(X_1 = x_1)$, the probability of securing the first catch on the x_1-th cast.

      (ii) **(2 points)** Determine the expected number of casts to catch the first fish, $E[X_1]$.

      (iii) **(2 points)** Find $P(X_4 = x_4)$, the probability that it takes exactly x_4 casts to catch four fish.

      (iv) **(2 points)** Find $E[X_6]$, the expected number of casts to catch 6 fish. **(Hint:** Use the fact that the sum of independent geometric random variables follows a negative binomial distribution.)

Q5: Expectation, Variance, and Moments of Continuous Random Variables (36 points)

Reading: In probability theory, continuous random variables can take any value in a continuous range, and their behavior is described using a probability density function (PDF). A key concept when working with continuous distributions is **expectation (mean)**, which represents the average value of the random variable over a large number of experiments. The **variance** measures how much the values of the random variable deviate from the mean.
Mathematically, for a continuous random variable X with PDF $f_X(x)$, the expectation $\mathbb{E}[X]$ and variance $\text{Var}(X)$ are defined as:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \int_{-\infty}^{\infty} x^2 f_X(x)\, dx - \left(\int_{-\infty}^{\infty} x f_X(x)\, dx \right)^2$$

Higher-order **moments** like skewness and kurtosis provide additional insights into the distribution's shape,

Exponential Distribution and Reliability (12 points)

A machine part fails according to an **exponential distribution** with a mean time between failures of 10 hours. Let X represent the time (in hours) until the next failure.

1. **(2 points)**: Write the PDF for the exponential distribution and calculate the expectation $\mathbb{E}[X]$.
2. **(3 points)**: Derive the variance of X. Interpret what this variance means in terms of the variability of failure times.
3. **(7 points)**: Suppose the machine part is replaced after each failure, and you track 1000 failure times. Simulate the total downtime over these 1000 failures, assuming each failure follows the exponential distribution with a mean of 10 hours. Plot the histogram of failure times and discuss the practical implications for managing the machine's operational efficiency.

Normal Distribution and Production Quality (12 points)

In a factory, the weight of a certain product follows a **normal distribution** with a mean of 500 grams and a standard deviation of 10 grams.

1. **(2 points)** Write the PDF of the normal distribution for this product weight. Identify the expectation $\mathbb{E}[X]$ and variance $\text{Var}(X)$.
2. **(3 points)** Calculate the probability that a randomly selected product weighs between 490 and 510 grams. How would this result influence quality control decisions?
3. **(4 points)** If the factory wants to ensure that 95% of the products weigh between 495 and 505 grams, what should the standard deviation be? What does this say about how much the manufacturing process needs to be improved?
4. **(3 points)** Simulate 10,000 product weights using the normal distribution with the given parameters. Plot the distribution and calculate the percentage of products that fall within the range of 495 to 505 grams. Compare this with the theoretical value and discuss any discrepancies.

Lognormal Distribution and Stock Returns (12 points)

The daily returns of a stock are modeled using a **lognormal distribution** with parameters $\mu = 0.001$ and $\sigma = 0.02$.

1. **(2 points)** Write the PDF for the lognormal distribution and calculate the expectation and variance of the stock returns.
2. **(3 points)** Calculate the probability that the stock returns more than 5% in a day. Discuss the implications of this for investors considering the stock as a high-risk, high-reward asset.
3. **(4 points)** Simulate 1,000 days of stock returns using the lognormal distribution. Plot the histogram of daily returns and calculate the proportion of days where the return is positive. How does this relate to typical stock market behavior?
4. **(3 points)** Calculate the 95th percentile of the daily returns and explain its significance in a financial risk management context.

Q6: Joint Probability Distributions and Covariance (32 points)

Reading: In probability theory, when we deal with two or more random variables, we often need to analyze how these variables relate to one another. The **joint probability distribution** describes the probability that two or more random variables take on certain values simultaneously. If X and Y are random variables, the joint probability distribution is represented by $P(X = x, Y = y)$ for discrete variables or a joint probability density function (PDF) $f_{X,Y}(x, y)$ for continuous variables.
For two continuous random variables X and Y, the joint PDF must satisfy the condition:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx\, dy = 1$$

From the joint distribution, we can compute **marginal distributions**, which describe the behavior of individual random variables by integrating or summing over the other variables:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx$$

Conditional probability distributions describe the distribution of one variable given that the other has taken on a specific value, computed as:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Finally, **covariance** and **correlation** help us quantify how much two variables change together. For two random variables X and Y, the covariance is defined as

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Correlation is a normalized form of covariance:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

This provides a measure of the linear relationship between the variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation).

1. In a certain suburb, each household reported the number of cars and the number of television sets that they own. Let X represent the number of cars, and Y represent the number of television sets owned by a randomly selected household. The table below gives the joint probability distribution for X and Y:

X / Y	**1**	**2**	**3**	**4**
1	0.1	0	0.1	0
2	0.3	0	0.1	0.2
3	0	0.2	0	0

Table 1: Joint Distribution of X and Y

(a) **(2 points)** Find the probability that a randomly selected household owns **at most two cars** and **at most two television sets**. That is, compute:

$$P(X \leq 2 \text{ and } Y \leq 2)$$

Provide an explanation of how the probabilities from the table contribute to this total probability.

(b) **(3 points)** Using the joint probability distribution, determine the **marginal distribution** for the number of cars X owned by a household. In other words, compute the probabilities $P(X = x)$ for each possible value of X. Specifically:

$$P(X = 1), P(X = 2), P(X = 3)$$

Summarize these values in a clear format and provide an interpretation of what the marginal distribution of cars tells us about car ownership in the suburb.

(c) **(3 points)** Similarly, determine the **marginal distribution** for the number of television sets Y owned by a household. Compute the probabilities $P(Y = y)$ for each possible value of Y. Specifically:

$$P(Y = 1), P(Y = 2), P(Y = 3), P(Y = 4)$$

Present the results in a clear format and discuss what the marginal distribution reveals about television set ownership in the suburb.

(d) **(3 points)** Are the number of cars X and the number of television sets Y owned by a household **independent**? Recall that X and Y are independent if and only if:

$$P(X = x, Y = y) = P(X = x)P(Y = y) \quad \text{for all } x, y$$

For this part, verify whether this condition holds for any pair of values of X and Y. Show your calculations and state whether X and Y are independent or not.

(e) **(2 points)** Find the **conditional probability** that a randomly selected household owns **exactly two television sets** given that they own **exactly two cars**. That is, calculate:

$$P(Y = 2 \,|\, X = 2)$$

Explain the result and provide insight into how owning two cars may affect the likelihood of owning two television sets.

(f) **(3 points)** Using the marginal distributions, compute the **expected number** of cars $\mathbb{E}[X]$ and the **expected number** of television sets $\mathbb{E}[Y]$ a randomly selected household owns. Use the formulas for the expected value of a discrete random variable:

$$\mathbb{E}[X] = \sum_x x \cdot P(X = x), \quad \mathbb{E}[Y] = \sum_y y \cdot P(Y = y)$$

Interpret these results in the context of the suburb's car and television set ownership patterns.

2. The daily returns of two correlated stocks, X and Y, follow a joint lognormal distribution with the following parameters:

$$\mu_X = 0.001, \quad \mu_Y = 0.002, \quad \sigma_X = 0.02, \quad \sigma_Y = 0.03, \quad \rho_{X,Y} = 0.8$$

   (a) **(2 points)** Write the joint PDF for the lognormal distribution of X and Y.

   (b) **(3 points)** Calculate the marginal distributions of X and Y.

   (c) **(4 points)** Simulate 1,000 days of returns for both stocks using the joint lognormal distribution. Plot the scatter plot and calculate the empirical correlation.

   (d) **(2 points)** Using the simulated data, calculate the percentage of days where both stocks have positive returns. Compare this to the theoretical correlation.

3. **(5 points)** Let X be the size of a surgical claim and let Y denote the size of the associate hospital claim. A risk analyst uses a model in which $E(X) = 5$, $E(X^2) = 27.4$, $E(Y) = 7$, $E(Y^2) = 51.4$ and $\text{Var}(X+Y) = 8$. Let $C_1 = X+Y$ denote the size of the combined claims before the application of a 20% surcharge on the hospital claim and C_2 denote the size of the combined claim after the application of the surcharge. Calculate $\text{Cov}(C_1, C_2)$.