Introduction to Statistical Inference: Assignment #1

Reyhaneh Ahani
reyhaneh.ahani@gmail.com

University of Tehran — November 22, 2023

Introduction

Initially, I address all the theoretical questions, followed by the programming section that includes output Figures and interpretations of the results.

1 Theoretical Section

1.1 Problem 1

Events Definition:

- A be the event that the alarm triggers.
- D be the event that conditions are dangerous.
- N be the event that conditions are normal.

so base on the question information we have:

- P(D) = 0.005 (probability of dangerous conditions occurring in a day).
- P(N) = 1 - P(D) = 1 - 0.005 = 0.995 (probability of normal conditions occurring in a day).
- P(A|D) = 0.95 (probability of triggering an alarm when dangerous conditions are present).
- P(A|N) = 0.005 (probability of triggering an alarm when conditions are normal).

1.1.1 The probability of a false alarm

We want to find P(N|A), the probability that conditions are normal given that the alarm triggers.

Firstly use Bayes' Theorem:

$$P(N|A) = \frac{P(A|N) * P(N)}{P(A)} \tag{1}$$

For P(A), we can calculate it using the law of total probability:

$$P(A) = P(A|D) * P(D) + P(A|N) * P(N) \tag{2}$$

Now we can calculate P(N|A):

$$P(N|A) = \frac{0.005 * 0.995}{0.95 * 0.005 + 0.005 * 0.995} = \frac{0.004975}{0.00475 + 0.004975} \approx \frac{0.004975}{0.009725} \approx 0.5111 \tag{3}$$

So, the probability of a false alarm is approximately 0.5111 or 51.11%.

1.1.2 The probability of an unidentified critical condition

The probability of an unidentified critical condition is the probability that conditions are dangerous when the system does not trigger.

We can calculate it with Bayes' Theorem as follows :

$$P(D|\neg A) = \frac{P(D) * P(\neg A|D)}{P(\neg A)} \tag{4}$$

Where:

- $\neg A$ is the event that the alarm does not trigger.
- $P(\neg A|D)$ is the probability of not triggering the alarm when dangerous conditions are present.

From the low of compliment we have: $P(\neg A|D) = 1 - P(A|D) = 1 - 0.95 = 0.05$

Now we can calculate $P(\neg A)$ using the law of total probability:

$$P(\neg A) = P(\neg A|D) * P(D) + P(\neg A|N) * P(N) \tag{5}$$

Given that $P(\neg A|N) = 1 - P(A|N) = 1 - 0.005 = 0.995$, we have:

$$P(\neg A) = 0.05 * 0.005 + 0.995 * 0.995 \approx 0.990275 \tag{6}$$

Now we can calculate $P(D|\neg A)$:

$$P(D|\neg A) = \frac{0.005 * 0.05}{0.990275} = \frac{0.00025}{0.990275} \approx 0.0002524 \tag{7}$$

So, the probability of an unidentified critical condition is approximately 0.0002524 or 0.02524%.

1.1.3 Expected number of false alarms and unidentified critical conditions over 10-years

For this part we can use the following formula:

$$\text{Expected Event within a Period} = \text{Probability of Event per Day} \times \text{Number of Days in the Period} \tag{8}$$

Given the following probabilities:

- Probability of False Alarms per day ($P_{\text{false alarms}}$) = 0.5111
- Probability of Unidentified Critical Conditions per day ($P_{\text{unidentified}}$) = 0.0002524

Now, let's calculate the expected values:

1. Expected False Alarms:

$$\text{Expected False Alarms} = 0.5111 \times (10 \times 365) = 0.5111 \times 3650 = 1868.15 \tag{9}$$

2. Expected Unidentified Critical Conditions:

$$\text{Expected Unidentified Critical Conditions} = 0.0002524 \times (10 \times 365) = 0.0002524 \times 3650 = 0.91966 \tag{10}$$

So, the expected number of false alarms over a 10-year period is approximately 1868.15, and the expected number of unidentified critical conditions is approximately 0.91966.

1.1.4 Effectiveness of the Alarming System

The effectiveness of the alarming system should be evaluated based on the trade-off between the number of false alarms and the number of unidentified critical conditions. In this case, there are a relatively high number of false alarms (1868.15) compared to a very low number of unidentified critical conditions (0.91966).
The system appears to have a high rate of false alarms, which could potentially lead to alert fatigue and decreased responsiveness. However, it also has a low rate of missing critical conditions, which is a positive aspect as critical conditions should not be missed.

1.2 Problem 2

Given two independent and identically distributed random variables, X_1 and X_2, with a common mean value μ and a common variance σ^2, calculate the followings

1.2.1 The mean value and variance of Y_1, where $Y_1 = X_1 + X_2$

To find the mean value and variance of Y_1, we can use the properties of expected values and variances for the sum of random variables:

Mean value of Y_1:

$$E(Y_1) = E(X_1 + X_2) = E(X_1) + E(X_2) = \mu + \mu = 2\mu \tag{11}$$

Variance of Y_1:

$$\text{Var}(Y_1) = \text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) = \sigma^2 + \sigma^2 = 2\sigma^2 \tag{12}$$

So, the mean value of Y_1 is 2μ, and the variance of Y_1 is $2\sigma^2$.

1.2.2 The mean value and variance of Y_2, where $Y_2 = 2X_1$)

To find the mean value and variance of Y_2, we can use the properties of expected values and variances for the scaled random variables:

Mean value of Y_2:

$$E(Y_2) = E(2X_1) = 2E(X_1) = 2\mu \tag{13}$$

Variance of Y_2:

$$\text{Var}(Y_2) = \text{Var}(2X_1) = 4\text{Var}(X_1) = 4\sigma^2 \tag{14}$$

So, the mean value of Y_2 is 2μ, and the variance of Y_2 is $4\sigma^2$.

1.2.3 Comparison of the mean value and variance of Y_1 and Y_2

No, the variances of Y_1 and Y_2 are not the same. The variance of Y_2 (which is $4\sigma^2$) is larger than the variance of Y_1 (which is $2\sigma^2$).

The difference in variances arises from the fact that Y_2 is a scaled version of X_1 (i.e., $Y_2 = 2X_1$), which amplifies the variability of the original random variable. When you scale a random variable by a constant factor, the variance is also scaled by the square of that factor. That's why the variance of Y_2 is four times the variance of X_1 ($4\sigma^2$), whereas Y_1 is the sum of two independent variables and its variance is just twice the variance of X_1 ($2\sigma^2$).

1.2.4 Covariance between Y_1 and Y_2

Covariance between Y_1 and Y_2:

$$\text{Cov}(Y_1, Y_2) = \text{Cov}(X_1 + X_2, 2X_1) = 2\text{Cov}(X_1, X_1) + \text{Cov}(X_2, 2X_1) \tag{15}$$

Since X_1 and X_2 are independent, $\text{Cov}(X_1, X_2) = 0$. Additionally, $\text{Cov}(X_1, X_1) = \text{Var}(X_1) = \sigma^2$. So, the covariance between Y_1 and Y_2 simplifies to:

$$\text{Cov}(Y_1, Y_2) = 2\sigma^2 + 0 = 2\sigma^2 \tag{16}$$

Therefore, the covariance between Y_1 and Y_2 is $2\sigma^2$.

1.3 Problem 3

To tackle this problem, firstly start by defining the probability function P for Ω and then proceed to find formulas for the probabilities of events A and B.

1.3.1 Probability Function P for Ω

The probability function P for the sample space Ω is defined as the probability of a specific sequence of birthdays for n individuals. Since each day of the year is equally likely for birthdays, we can express P(Ω) as the probability of any specific sequence of birthdays for n individuals. There are 365 possible birthdays for each individual, and since the birthdays are independent, the probability function P for Ω is:

$$P(\Omega) = \left(\frac{1}{365}\right)^n \tag{17}$$

1.3.2 Event A: "At least one person in the group shares your birthday."

Event A can be defined as the complement of the event that nobody in the group shares your birthday. Let p be the probability that nobody in the group shares your birthday, which is the probability of each of the n individuals having a birthday different from yours. The probability of event A, P(A), can be expressed as:

$$P(A) = 1 - p \tag{18}$$

To find p, we can calculate it for each individual and then multiply these probabilities together since the birthdays are independent:

$$p = \left(1 - \frac{1}{365}\right)^n \tag{19}$$

So, the probability of event A is:

$$P(A) = 1 - \left(1 - \frac{1}{365}\right)^n \tag{20}$$

To find the minimum value of n that makes P(A) exceed 0.5, we can solve the following inequality:

$$1 - \left(1 - \frac{1}{365}\right)^n > 0.5 \tag{21}$$

After solving the above equation, we find the minimum value of n that makes P(A) exceed 0.5 is 253.

1.3.3 Rationale explanation for why n must be larger than $\frac{365}{2}$

Let's try to reverse the problem. First, consider that we have n birthdays, and in the best-case scenario, let's assume that none of them share the same date. So, the probability of our birthdays falling within these n days is $\frac{n}{365}$, all based on the assumption that all n birthdays are unique. In order for $P(A)$ to exceed 0.5, n should be greater than $\frac{365}{2}$.

1.3.4 Python Simulation for Event B

The final result of the Python program is illustrated in Table 1. As shown, with a low number of iterations, the probability is more likely to exceed 0.9 sooner. However, with a large number of iterations, we observe a more certain range of values for n where $P(A) > 0.9$.

Table 1: Minimum n for $P(A) > 0.9$ for Different Iterations

Iteration	**Simulation**	**Minimum n**
10,000	1	41
	2	41
	3	41
	4	41
	5	42
	6	41
	7	41
	8	41
	9	41
	10	41
30	1	37
	2	34
	3	38
	4	36
	5	39
	6	34
	7	39
	8	32
	9	38
	10	33

1.3.5 Event B: "At least two people in the group share a birthday."

Event B can be defined as the complement of the event that everyone in the group has a unique birthday. Let q be the probability that everyone in the group has a unique birthday. The probability of event B, P(B), can be expressed as:

$$P(B) = 1 - q \tag{22}$$

To find q, we can calculate it for the first individual, then for the second individual, and so on, considering that the birthdays must be different for each individual:

$$q = 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \ldots \times \left(1 - \frac{n-1}{365}\right) \tag{23}$$

So, the probability of event B is:

$$P(B) = 1 - 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \ldots \times \left(1 - \frac{n-1}{365}\right) \tag{24}$$

You can calculate P(B) for any value of n using this formula.

1.3.6 Python Simulation for Event C

The minimum value of n required for $P(A) > 0.5$ was determined through 20 simulations. The results are as follows:

- Number of simulations with Minimum n = 87: 5 times
- Number of simulations with Minimum n = 88: 15 times Based on the observation, it is more likely that at least three people in the group share a birthday among 88 students.

1.4 Problem 4

Let's analyze each of the scenarios:

1. Mr. Jones is the parent of two children, and the older child is a girl. What is the likelihood that both children are girls? In this scenario, there are four possible combinations of children's genders: BB, BG, GB, and GG, where B represents a boy and G represents a girl. We know that the older child is a girl (GG, GB). So, the probability that both children are girls given that the older child is a girl is 1/2 or 50%.

2. Mr. Smith is the parent of two children, and at least one of them is a boy. What is the probability that both children are boys? In this scenario, there are three possible combinations of children's genders: BB, BG, and GB. We know that at least one of them is a boy (BB, BG, GB). To find the probability that both children are boys given that at least one of them is a boy, we exclude the case where both children are girls (GG). So, the probability that both children are boys given that at least one of them is a boy is 1/3 or approximately 33.33%.

Notice: The above probabilities are based on the assumption that each gender (boy or girl) is equally likely at birth and that the gender of one child does not affect the gender of the other child.

1.5 Problem 5

Ladies and gentlemen of the jury,

I stand before you to present a compelling case for reasonable doubt in the matter of my client, the driver of the blue taxi, who vehemently asserts his innocence in the hit-and-run incident.
The key evidence against him rests on a witness's claim of seeing a blue taxi leaving the scene. To challenge this, we conducted a scientific test to evaluate the witness's ability to differentiate between blue and green taxis under conditions similar to the night of the accident.
The collected data reveals that the witness correctly identifies blue taxis as blue 99% of the time, which

Table 2: Witness's Perceptions

	Actual Blue	Actual Green
Perceived as Blue	99%	2%
Misidentification	1%	98%

is commendable. However, it's crucial to note that the witness misidentifies green taxis as blue 2% of the time.
Considering there are 99 green taxis and only 1 blue taxi in the city, the probability of the witness mistakenly identifying a green taxi as blue is substantial.
Ladies and gentlemen, this raises a reasonable doubt about the accuracy of the witness's identification. It is possible that in the darkness of the night, the witness may have misjudged the color of the taxi involved in the incident.
I urge you to carefully consider this information and acknowledge the uncertainty that exists. Let us not rush to judgment, but rather, let reason and doubt guide your deliberations.

Thank you for your attention.

1.6 Problem 6

Given Information:
Inside a drawer, there are four dice: one with four sides (tetrahedron), one with six sides (cube), and two with eight sides (octahedra). Your friend discreetly selects one of these four dice at random. We'll denote the number of sides on the chosen dice as S.

1.6.1 The probability mass function (pmf) for S

The probability mass function (pmf) for S can be represented as:

$$P(S=s) = \begin{cases} \frac{1}{4} & \text{if } s=4 \text{ (tetrahedron)} \\ \frac{1}{4} & \text{if } s=6 \text{ (cube)} \\ \frac{1}{2} & \text{if } s=8 \text{ (octahedron)} \end{cases} \tag{25}$$

1.6.2 The selected dice without revealing the outcome

(A) Which dice is the most likely choice when R equals 3?

The calculation for obtain the conditional pmf of S given $R=3$ can be performed as follows:

$$P(S=i|R=3) = \frac{P(S=i)\cdot P(R=3|S=i)}{\sum_i P(S=i)\cdot P(R=3|S=i)} \tag{26}$$

Given that R is hidden from you, we can assume that each dice is equally likely to produce any outcome. Therefore:

$$P(R=3|S=i) = \begin{cases} \frac{1}{4} & \text{if } i=4 \\ \frac{1}{6} & \text{if } i=6 \\ \frac{1}{8} & \text{if } i=8 \end{cases} \tag{27}$$

(a) Substitute the values for $i=4$:

$$P(S=4|R=3) = \frac{\frac{1}{4}\cdot\frac{1}{4}}{\frac{1}{4}\cdot\frac{1}{4}+\frac{1}{4}\cdot\frac{1}{6}+\frac{1}{2}\cdot\frac{1}{8}} = 0.374999999 \tag{28}$$

(b) Substitute the values for $i=6$:

$$P(S=6|R=3) = \frac{\frac{1}{4}\cdot\frac{1}{6}}{\frac{1}{4}\cdot\frac{1}{4}+\frac{1}{4}\cdot\frac{1}{6}+\frac{1}{2}\cdot\frac{1}{8}} = 0.25 \tag{29}$$

(c) Substitute the values for $i=8$:

$$0.25P(S=8|R=3) = \frac{\frac{1}{2}\cdot\frac{1}{8}}{\frac{1}{4}\cdot\frac{1}{4}+\frac{1}{4}\cdot\frac{1}{6}+\frac{1}{2}\cdot\frac{1}{8}} = 0.374999999 \tag{30}$$

Based on the calculation, the tetrahedron and octahedron dice are equally the most probable choices when R is set to 3.

(B) Which dice is the most likely choice when R equals 6? The fraction of the Busyin rule is the same for both cube and octahedron dice, but the numerator for S is larger. Therefore, the octahedron die is considered the most probable choice when R is set to 6.

(C) Which dice is the most likely choice when R equals 7? The calculation is not required further since a result equal to 7 can only be achieved by an octahedron.

1.7 Problem 7

1.7.1 Standard Deviation

The standard deviation (σ) for a fair n-sided dice can be calculated using the formula:

$$\sigma = \sqrt{\frac{(n^2-1)}{12}} \tag{31}$$

For a 4-sided dice, $n=4$, and for a 6-sided dice, $n=6$. Therefore, the standard deviation for X and Y is:

$$\sigma_X = \sqrt{\frac{(4^2-1)}{12}} = \sqrt{\frac{5}{3}} \tag{32}$$

$$\sigma_Y = \sqrt{\frac{(6^2-1)}{12}} = \sqrt{\frac{35}{12}} \tag{33}$$

For Z, which is the average of X and Y:

$$\sigma_Z = \frac{\sqrt{\text{Var}(X) + \text{Var}(Y)}}{2} \tag{34}$$

$$\sigma_Z = \frac{\sqrt{\left(\sqrt{\frac{5}{3}}\right)^2 + \left(\sqrt{\frac{35}{12}}\right)^2}}{2} \tag{35}$$

1.7.2 Probability Mass Function (PMF) and Cumulative Distribution Function (CDF) for Z

The PMF for Z can be calculated by considering all possible combinations of outcomes from X and Y, taking the average, and determining the probabilities. The CDF for Z accumulates these probabilities for each possible value of Z.

PMF for Z:

$$P(Z = z) = \sum_{i=1}^{4} \sum_{j=1}^{6} P(X = i, Y = j) \times \delta\left(z = \frac{i+j}{2}\right) \tag{36}$$

where $\delta(x)$ is the indicator function.

CDF for Z:

$$F(z) = P(Z \leq z) = \sum_{k=-\infty}^{z} P(Z = k) \tag{37}$$

1.7.3 Expected Overall Gain or Loss

Let X and Y be two random variables representing the outcomes in a game. The rules of the game are as follows:

$$\text{If } X > Y, \text{ then the gain is } 2X \text{ dollars.} \tag{38}$$

$$\text{If } X \leq Y, \text{ then the loss is } 1 \text{ dollar.} \tag{39}$$

The expected gain or loss ($E(\text{gain})$) for each round can be expressed as:

$$E(\text{gain}) = P(X > Y) \cdot 2X - P(X \leq Y) \cdot 1 \tag{40}$$

Now, let's consider the overall gain or loss after engaging in this game for a total of 60 rounds. The expected overall gain or loss ($E(\text{overall})$) is given by:

$$E(\text{overall}) = 60 \cdot E(\text{gain}) \tag{41}$$

In this section, a Python code has been devised to compute $E(\text{overall})$ through 10,000 simulations. Based on the output, the following observations can be made:

$$E(\text{overall}) \approx 0.91 \tag{42}$$

1.8 Problem 8

1.8.1 Number of Raisins in the Cereal Box

The number of raisins in the cereal box can be found by integrating the density function $f(h) = 40h$ over the height range of the box. In this case, the height range is from 0 cm to 30 cm.

$$\text{Total number of raisins} = \int_0^{30} (40 - h)\, dh = 750 \tag{43}$$

1.8.2 Probability Density Function (pdf) for the Height of a Randomly Selected Raisin (H)

The probability density function for H can be derived from the density function $f(h)$ as follows:

$$g(h) = \frac{f(h)}{\int_0^{30} f(h)\, dh = 750} \tag{44}$$

Here, $g(h)$ is the pdf for the height of a randomly selected raisin.

1.8.3 Cumulative Distribution Function (cdf) for H

The cumulative distribution function $G(h)$ for H can be calculated by integrating the pdf $g(h)$ from 0 to h:

$$G(h) = \int_0^h g(h)\, dh \tag{45}$$

1.8.4 Probability of a Raisin in the Bottom Third of the Box

To find the probability that a randomly chosen raisin is located in the bottom third of the cereal box (i.e., H is less than or equal to 10 cm), you can use the cdf $G(h)$ as follows:

$$P(H \leq 10) = G(10) \tag{46}$$

1.9 Problem 9

1.9.1 Calculate the Joint Distribution, Covariance, and Correlation

The joint distribution of X and Y can be represented in a table as follows:

	$Y = 1$	$Y = -1$
$X = 1$	$P(X = 1, Y = 1)$	$P(X = 1, Y = -1)$
$X = -1$	$P(X = -1, Y = 1)$	$P(X = -1, Y = -1)$

(47)

Since $P(X = 1) = P(X = -1) = \frac{1}{2}$ and $P(Y = 1) = P(Y = -1) = \frac{1}{2}$, we have:

$$\begin{aligned} P(X = 1, Y = 1) &= c \\ P(X = 1, Y = -1) &= \frac{1}{2} - c \\ P(X = -1, Y = 1) &= \frac{1}{2} - c \\ P(X = -1, Y = -1) &= \frac{1}{2} + c \end{aligned} \tag{48}$$

The covariance $\text{Cov}(X, Y)$ can be calculated as:

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) \tag{49}$$

where E denotes the expected value.

Since X and Y are symmetric around 0, $E(X) = E(Y) = 0$. The expected value $E(XY)$ can be calculated as:

$$\begin{aligned} E(XY) &= (1 \cdot 1)P(X = 1, Y = 1) + (1 \cdot -1)P(X = 1, Y = -1) \\ &\quad + (-1 \cdot 1)P(X = -1, Y = 1) + (-1 \cdot -1)P(X = -1, Y = -1) \\ &= c - \frac{1}{2} + c - \frac{1}{2} \\ &= 2c - 1 \end{aligned} \tag{50}$$

The correlation $\text{Cor}(X, Y)$ is given by:

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \tag{51}$$

Since X and Y are symmetric, $\text{Var}(X) = \text{Var}(Y) = \frac{1}{2}$. Thus,

$$\begin{aligned} \text{Cor}(X, Y) &= \frac{2c - 1}{\sqrt{\frac{1}{2} \cdot \frac{1}{2}}} \\ &= 2(2c - 1) \\ &= 4c - 2 \end{aligned} \tag{52}$$

1.9.2 Determine the Values of c for Giving Conditions

- For independence, X and Y are independent if and only if $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all x, y. In our case, this implies:

$$\begin{aligned} c &= P(X = 1, Y = 1) \\ &= P(X = 1)P(Y = 1) \\ &= \frac{1}{2} \cdot \frac{1}{2} \\ &= \frac{1}{4} \end{aligned} \tag{53}$$

- For 100% correlation, $\text{Cor}(X, Y) = 1$. Thus,

$$\begin{aligned} 4c - 2 &= 1 \\ 4c &= 3 \\ c &= \frac{3}{4} \end{aligned} \tag{54}$$

  Therefore, X and Y are independent when $c = \frac{1}{4}$ and are 100% correlated when $c = \frac{3}{4}$.

1.10 Problem 11

1. Population mean: Not a random variable. It's a fixed value representing the average of a population.
2. Population size: Not a random variable. It's a fixed and known value representing the size of the entire population.
3. Sample size: Not a random variable. It's a fixed value representing the number of elements in a sample.
4. Sample mean: Yes, a random variable. The sample mean can vary depending on the elements included in a particular sample.
5. Variance of the sample mean: Not a random variable. It's a measure of the dispersion of sample means and is a fixed value based on the population variance and sample size.

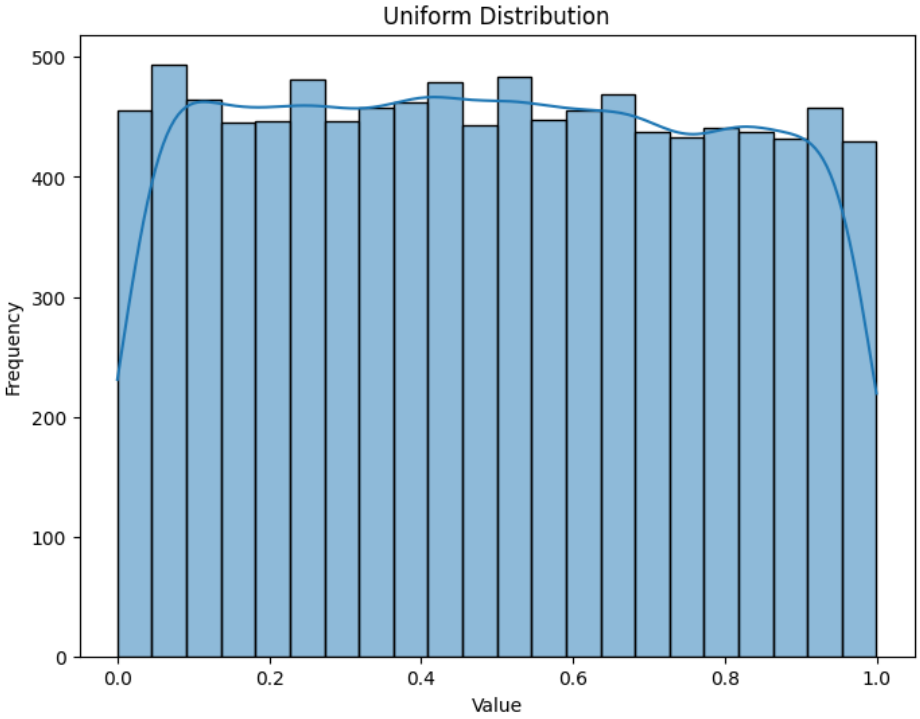


Figure 1: Uniform Distribution

6. The largest value in the sample: Not a random variable. It's a specific, deterministic value based on the sample.
7. Population variance: Not a random variable. It's a fixed value representing the variability of the entire population.

   So, the random variable among the listed options is the Sample mean.

2 Implementation

2.1 Problem 10 (Random Distributions)

2.1.1 Random Distributions Plots

By utilizing the Numpy random package to generate random variables, they can be observed through Figures 1 to 5.

2.1.2 Sample Means Distribution

The sample means for a normal distribution over 1000 iterations can be observed in Figure 6.

Base on the Central Limit Theorem (CLT), regardless of the original distribution of the data, the distribution of the sample means tends to follow a normal distribution as the sample size increases. Therefore, for a sufficiently large sample size, the distribution of sample means should be approximately normal. This is reflected in the final plot, where the distribution of means appears to be normal.

2.2 Problem 10 (Prob10 Data-set)

2.2.1 Data Cleaning Process

1. The `convert_to_numeric(df)` function is used for converting numeric values in the dataframe `df` into non-numeric values. This ensures that all values in the dataframe are converted to a non-numeric format, which is useful for further data processing or analysis. The data types of the columns in the

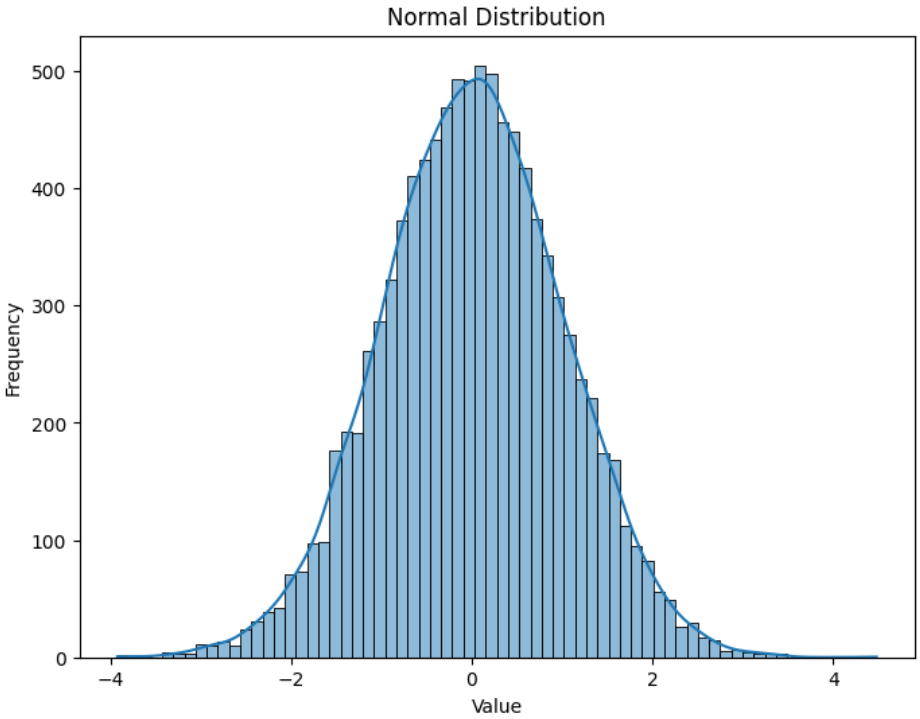


Figure 2: Normal Distribution

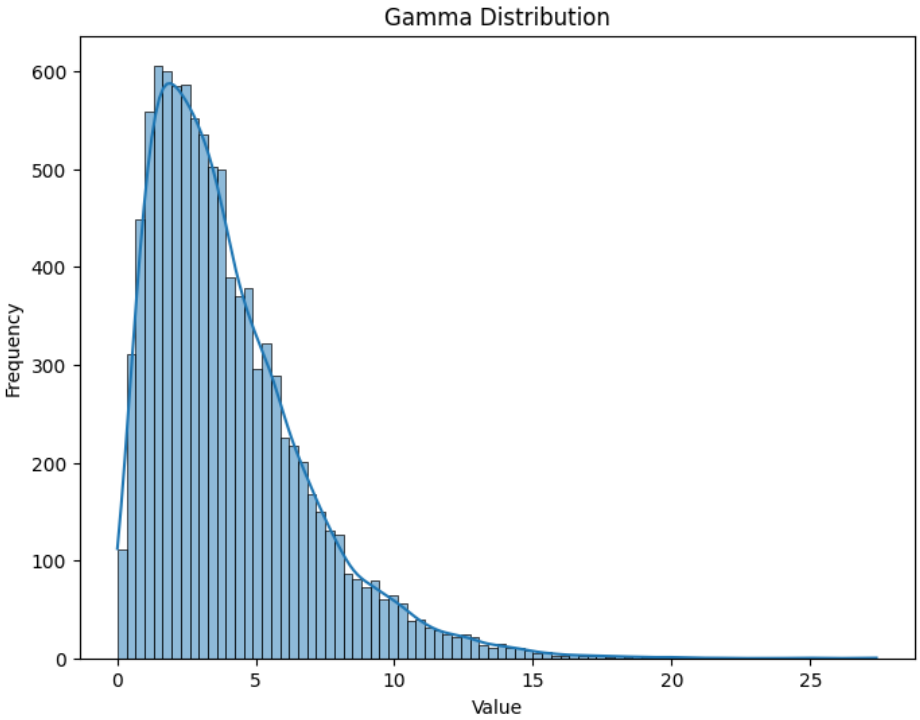


Figure 3: Gamma Distribution

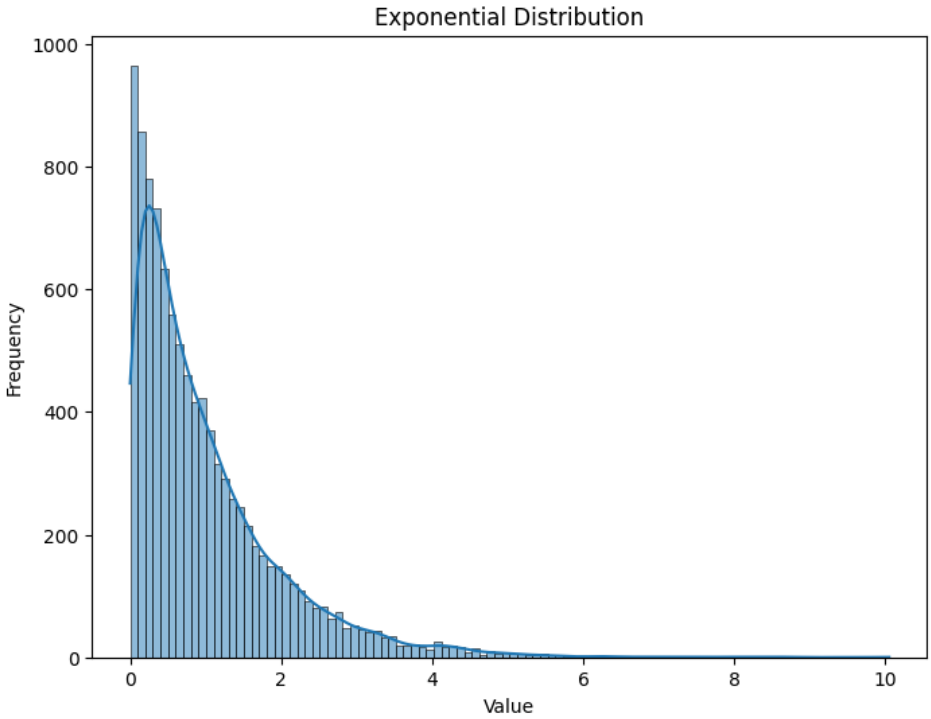


Figure 4: Exponential Distribution

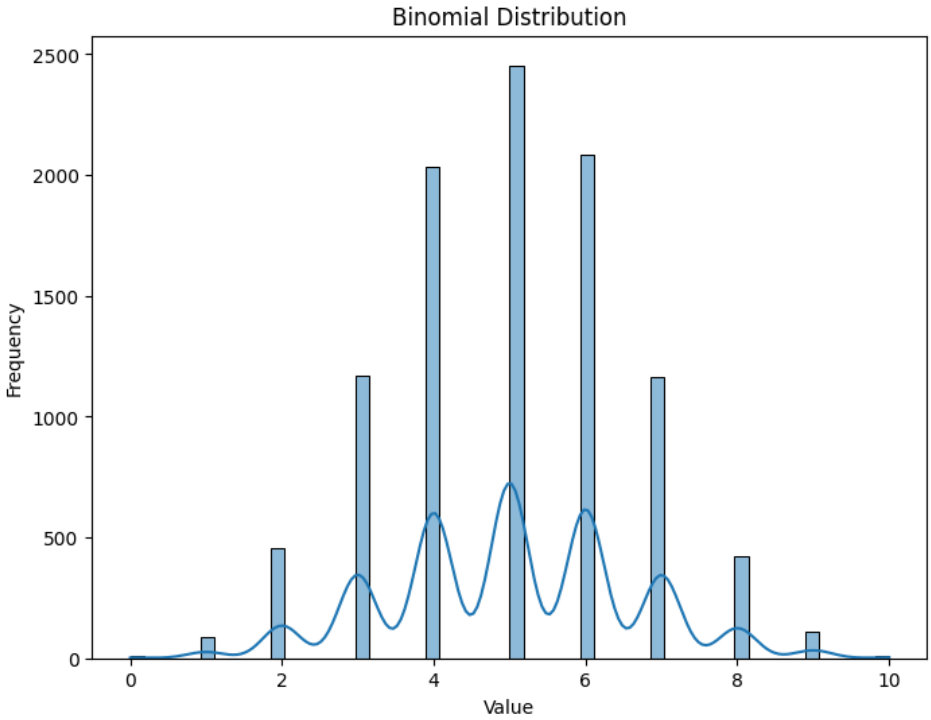


Figure 5: Binomial Distribution

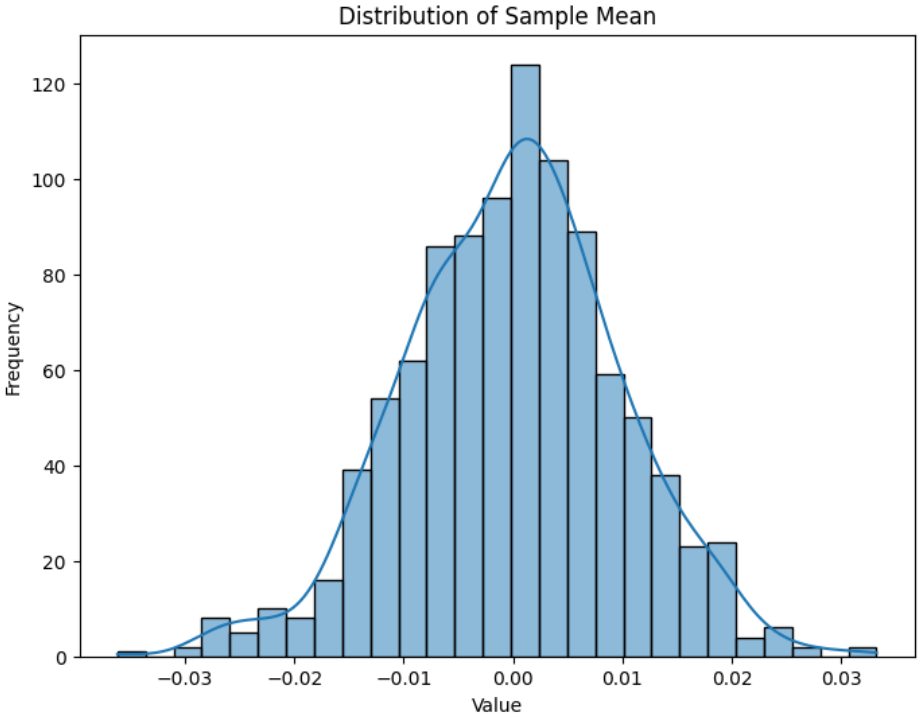


Figure 6: Sample Means Distribution

dataframe `df`, along with the label 'SeconderydTypes', are printed using the `printtxt(df.dtypes, 'Secondery_dTypes')` function. This helps in understanding the types of data that are present in each column.

2. Each column in the dataframe `df` is iterated over by the `for column in df.columns` loop, allowing for operations to be performed on each individual column. The condition `df[column].dtype == "object"` checks if the current column's data type (`df[column]`) is classified as an "object" type. This is typically used for identifying categorical columns in the dataframe.

3. The statement `df[column].fillna(df[column].mode()[0], inplace=True)` is used to fill any missing values (`NaN`) in the current column with the mode (most frequent value) of that column for object-type columns. This aids in handling missing categorical data.

4. The statement `df[column].fillna(df[column].mean(), inplace=True)` is used to fill any missing values (`NaN`) in the current column with the mean value of that column for numerical-type columns. This aids in handling missing numerical data.

5. Finally, the dataframe `df` is displayed by using the `df.head()` method after the missing values have been replaced. This allows for a quick inspection of the updated dataframe and ensures that the missing value replacements have been applied correctly.

2.2.2 Describing Columns of the Data-set

The Prob10 data-set comprises car details for 205 entries within the Range-Index, spanning from 0 to 204. A total of 26 columns, encompassing features such as 'symboling,' 'normalized-losses,' and 'make,' have been observed. Information on data availability is provided by non-null counts, and diverse data types, including float64, int64, and object, are represented. The memory usage is 41.8+ KB.
Concerning nonsensical data, missing values, represented by "?", are observed in certain columns. These missing values should be attended to during the process of data cleaning.

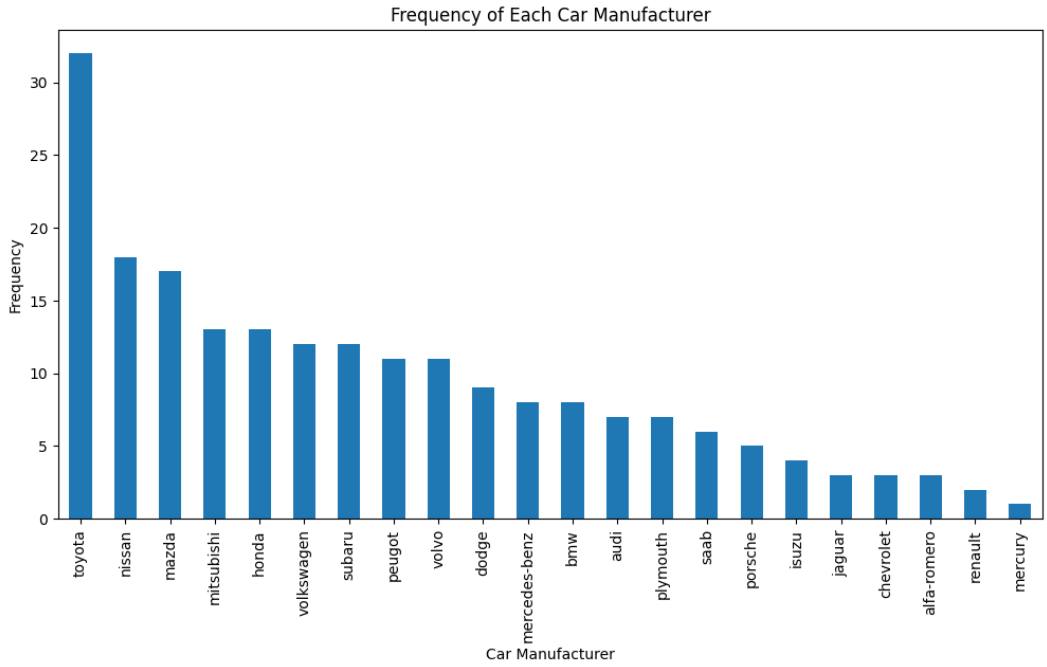


Figure 7: Car Manufacturer

2.2.3 Most frequent car manufacturer

After learning more about the dataset and completing some preprocessing, the number of cars produced by each company is counted using the `value_counts()` function. The results are presented in Figure 7

2.2.4 Dispersion of the Data-set

The analysis provided indicates that a longer tail on the right side of the distribution is suggested by the positive skewness values (above 0) observed in columns such as ”normalized-losses,” ”bore,” ”wheel-base,” ”width,” ”curb-weight,” ”engine-size,” ”compression-ratio,” ”horsepower,” ”city-mpg,” ”highway-mpg,” and ”price.” Conversely, a longer tail on the left side of the distribution is suggested by the negative skewness value observed in the ”stroke” column.

Table 3: Skewness and Kurtosis Values

Variable	Skewness	Kurtosis
`symboling`	0.209525	-0.689015
`normalized-losses`	0.848535	1.341492
`wheel-base`	1.042514	0.963276
`length`	0.154810	-0.110013
`width`	0.897375	0.656614
`height`	0.062660	-0.462188
`curb-weight`	0.676402	-0.070942
`engine-size`	1.933375	5.148030
`bore`	0.020063	-0.795149
`stroke`	-0.684727	2.092670
`compression-ratio`	2.591720	5.077161
`horsepower`	1.387515	2.584179
`peak-rpm`	0.073052	0.055542
`city-mpg`	0.658838	0.535505
`highway-mpg`	0.536038	0.400284
`price`	1.813926	3.243836

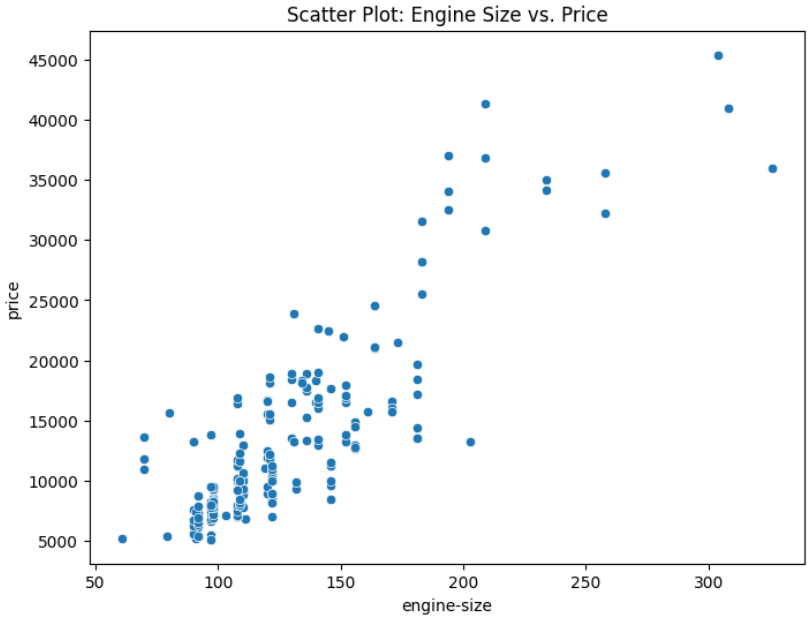


Figure 8: Scatter-Plot between Engine Size vs Price

2.2.5 The Scatter Plot between Engine-Size and Price Value

A positive association is observed. This association is illustrated by Figure 8.

2.2.6 Pair-plot for Multivariate Analysis

A pair plot matrix, as depicted in Figure 9, has been generated utilizing the columns 'length,' 'wheel-base,' 'curb-weight,' 'horsepower,' 'city-mpg,' and 'price.'

2.2.7 Correlation Heat-map for All Numerical Records

The Figure 10 is generated by utilizing Pandas' `corr()` function and Seaborn's `heatmap()`.

2.2.8 Box-Plot for a Few Categorical Columns

To examine the influence of specific categorical columns on the price of the car, several box plots were created. This is illustrated in Figures 11 and 12.

2.3 Problem 12

- Fi The Monte Carlo method involves the random generation of points within a given region, and the value of pi is estimated by using the proportion of points falling within a specific shape, such as a circle.
- Secondly, an inference can be drawn from the code that, as the number of random points (N) is increased, the accuracy of the estimated pi values is enhanced. This improvement is reflected by the decrease observed in the differences between the estimated pi values and the actual pi value as N is increased. The final result can be observed in Figure **??**.
- Overall, the usefulness and effectiveness of the Monte Carlo method for estimating mathematical constants, particularly pi, are showcased by this code. The gradual increase in the number of random points leads to the approach of the estimation towards the actual pi value, thereby validating the value and reliability of the Monte Carlo method within this specific context.

probability

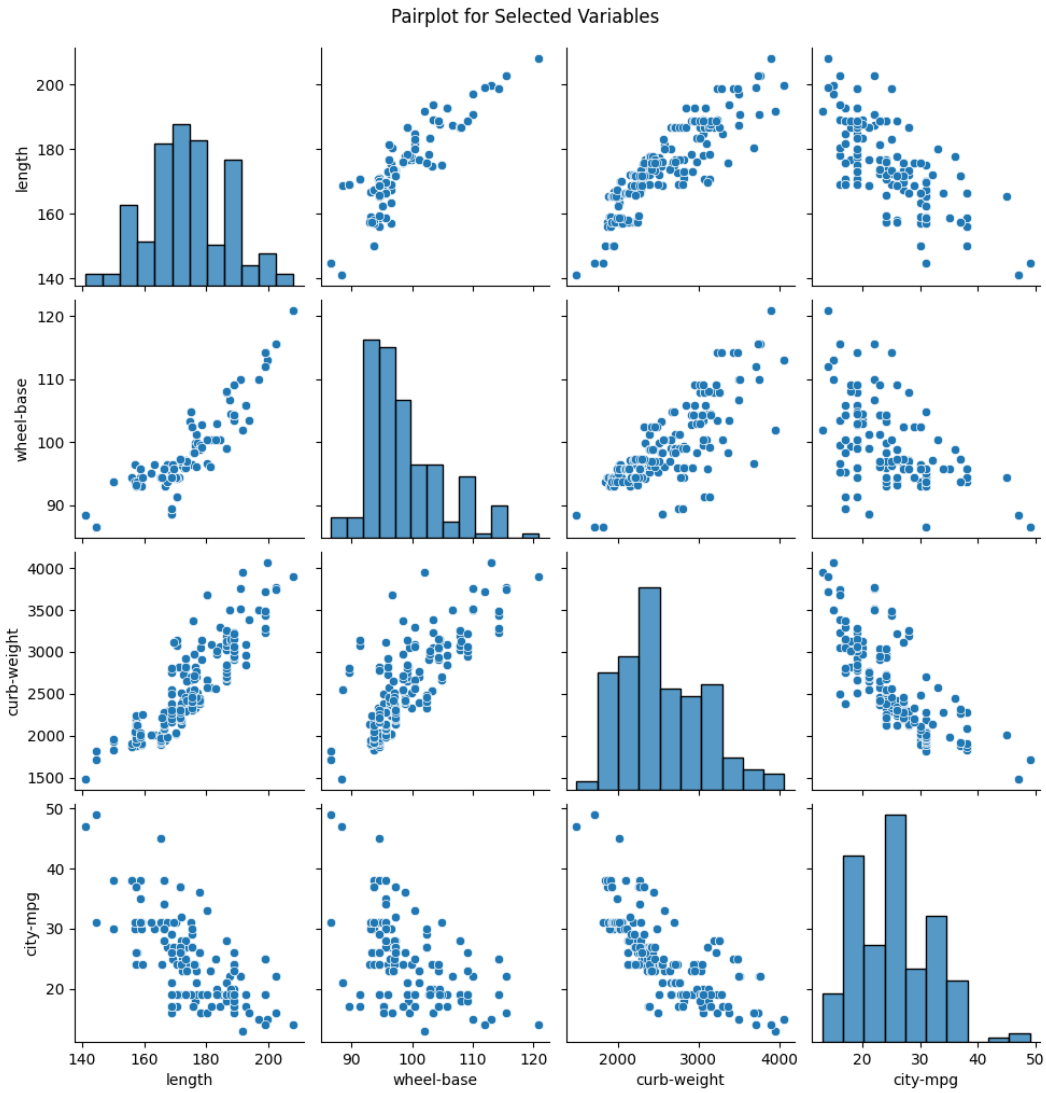


Figure 9: Pair-Plot for a list of numerical columns

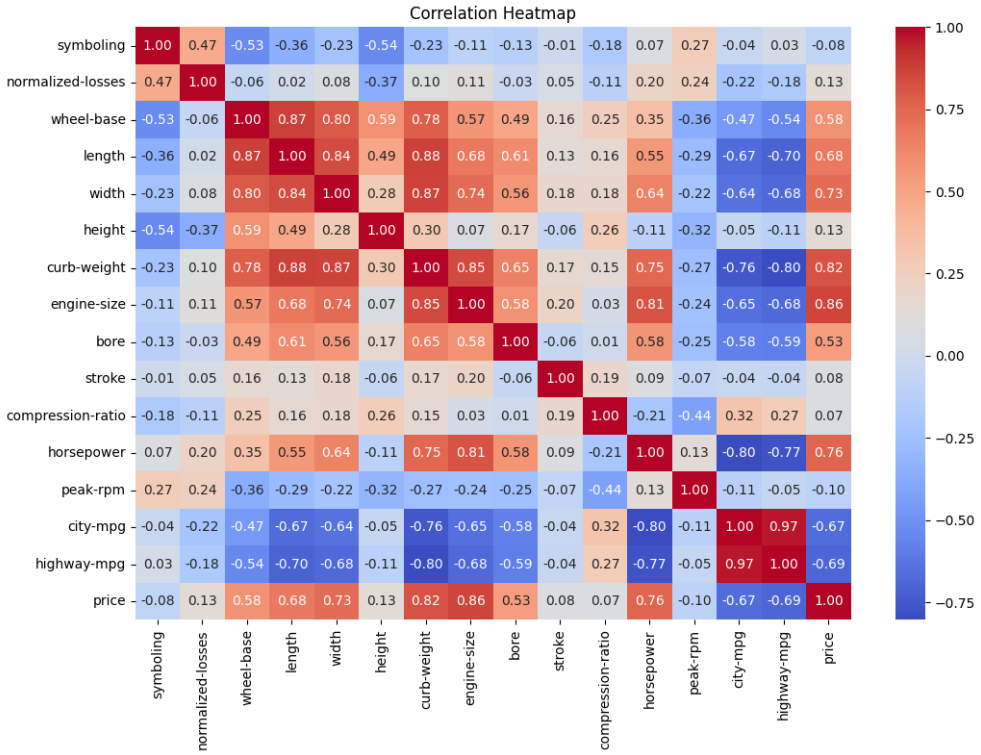
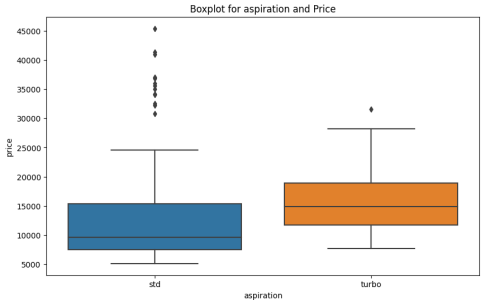
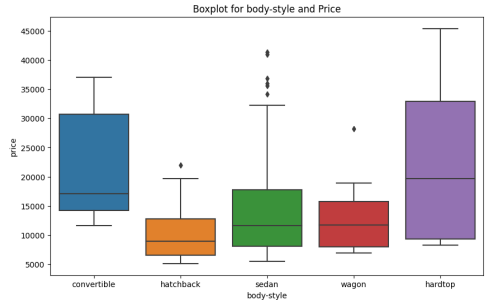


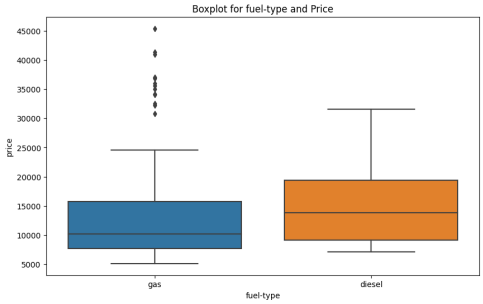
Figure 10: Correlation heatmap over all entire numerical records using Pandas `.corr()`



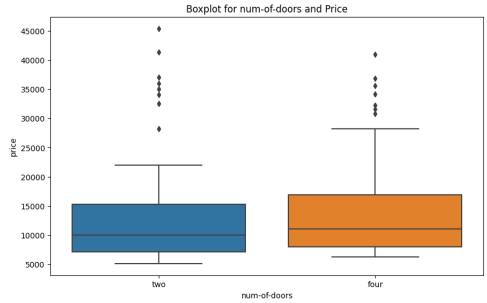
(a) BoxPlot of Price over Aspiration



(b) BoxPlot of Price over BodyStyle



(c) BoxPlot of Price over Fuel



(d) BoxPlot of Price over NumOfDoor

Figure 11: Box-Plot for Various Categorical Columns

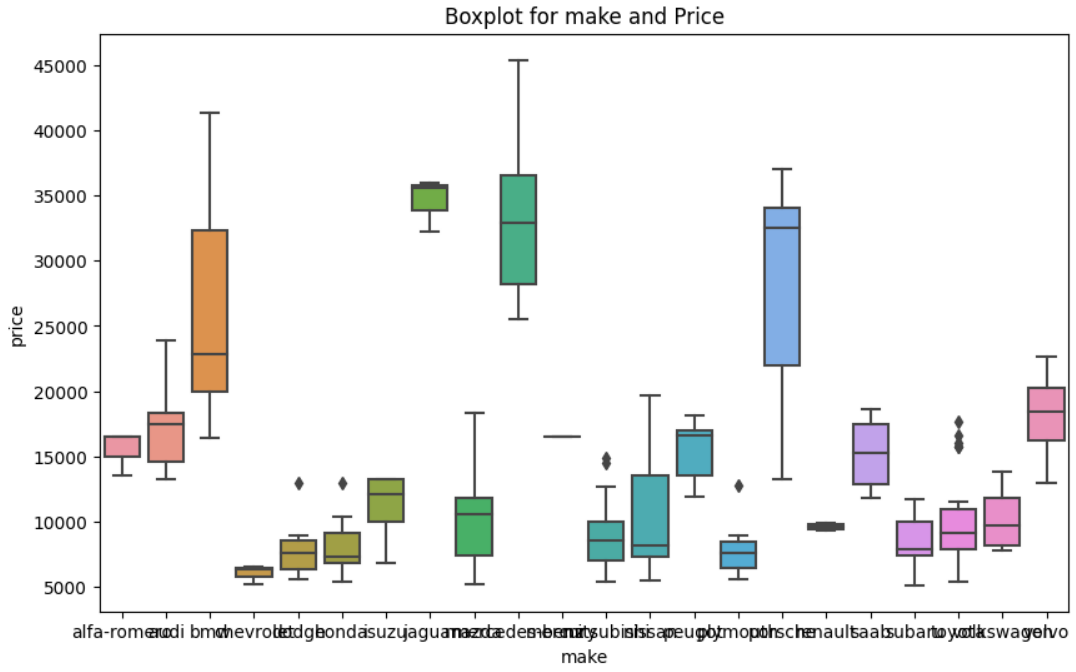


Figure 12: Box-Plot of Price over Manufacturer

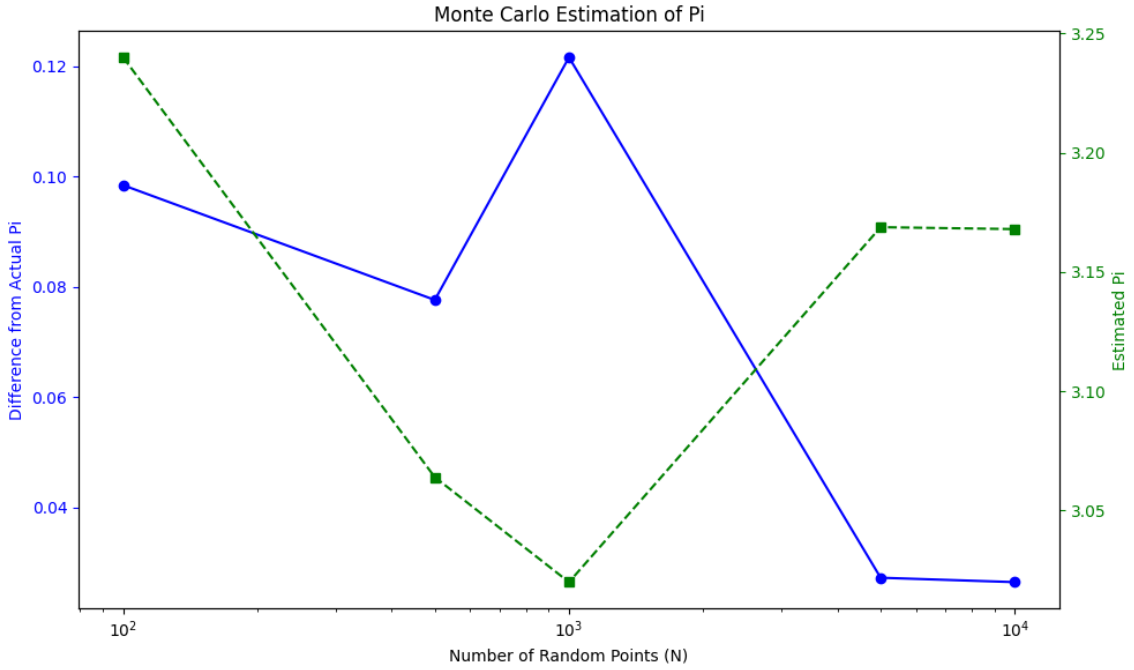


Figure 13: Monte Carlo Estimation of PI

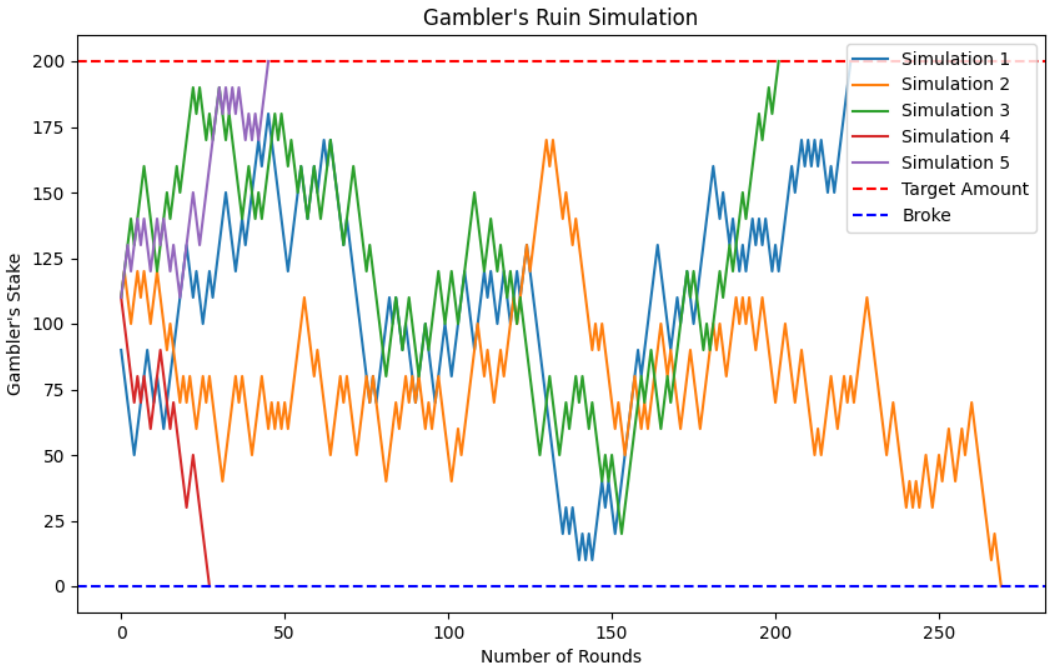


Figure 14: Simulate the Gambling Scenario for 1000 Rounds

2.4 Problem 13

- The plot of the simulation of this scenario is shown in Figure 14. The target amount is represented by the red dashed line in the plot, indicating the objective that the gambler aims to reach. Observations can be made regarding the success of any of the simulations in reaching the target before going broke.
- The point where the gambler goes broke (stake ≤ 0) is represented by the blue dashed line in the plot. Observations can be made to determine if this point is reached by any simulations before the target amount is achieved.
- The gambler's stake after each round for a single simulation is represented by each line in the plot. By analyzing the trajectories, an understanding can be gained of how the gambler's stake fluctuates over time.
- Figure 15 illustrates the overall probability of reaching the target amount before going broke. As expected, it hovers around 0.5, which was the initial win probability for the gambling. Since the distance between the target amount and the broke line from the initial stake is the same, the mean value of the probability of winning distribution should be the same as the win probability for each gambling round. The output, 0.48800, indicates this statement.

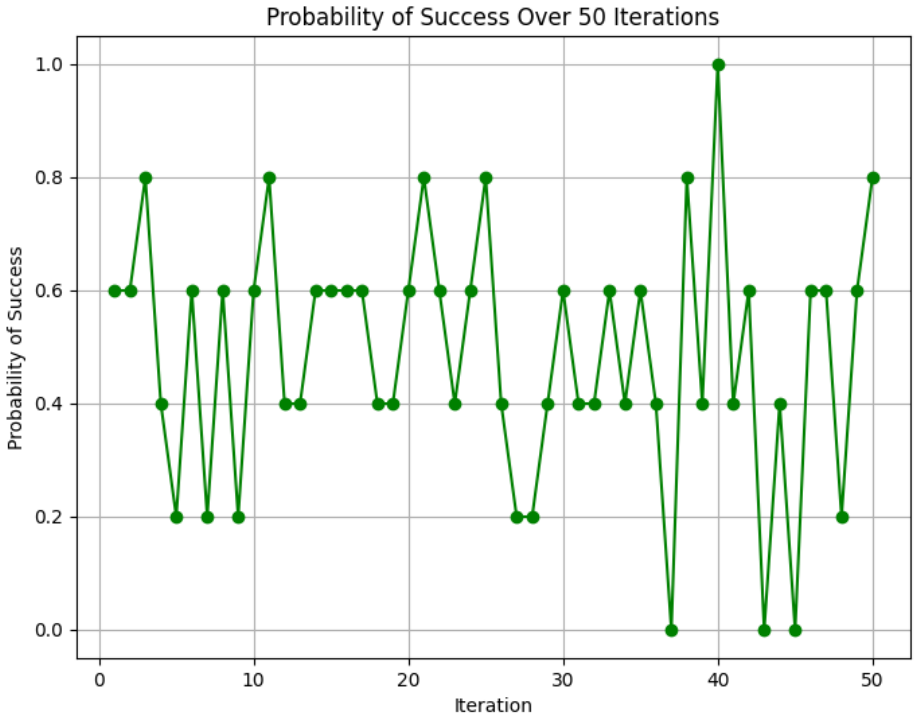


Figure 15: Probability of Reaching the Target Amount over 50 Iteration