Statistical Inference HW#1

Student Name:
Pouya Haji Mohammadi Gohari

SID:810102113

Date of deadline
Wednesday 8th November, 2023

Dept. of Computer Engineering

University of Tehran

Contents

1 Question 1

1.1 Normal when alarm triggers

According to the question we have a conditional probability which is the sensor has probability of 0.95 of triggering an alarm when dangerous conditions are present in a given day and probability of triggering an alarm when conditions are normal during the day is 0.005.
Eventually we have probability of 0.005 for dangerous conditions occurring in a day.

$$
\begin{gathered}
P(triggering|dangerous) = 0.95 \\
P(triggering|normal) = 0.005 \\
P(dangerous) = 0.005
\end{gathered}
\tag{1}
$$

So from equation 1 we have another probabilities which are shown in the equation 2.

$$
\begin{gathered}
P(\overline{triggering}|dangerous) = 0.05 \\
P(\overline{triggering}|normal) = 0.995 \\
P(normal) = 0.995
\end{gathered}
\tag{2}
$$

Let's go to the first question, it wants us to find a conditional probability which is $P(normal|triggering)$. From Bayes' theorem we have:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

And also if we consider Law of Total Probability we reach the equation below:

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

So if we apply this to our question, questioned probability would be 0.511568123 due to 3.

$$
\begin{gathered}
P(normal|triggering) = \frac{P(triggering|normal)P(normal)}{P(triggering|normal)P(normal) + P(triggering|dangerous)P(dangerous)} \\
\\
P(normal|triggering) = \frac{0.005 * 0.995}{0.005 * 0.995 + 0.95 * 0.005} \\
\\
P(normal|triggering) = 0.511568123
\end{gathered}
\tag{3}
$$

1.2 Dangerous when alarm not triggers

In this part we must find conditional probability of $P(dangerous|\overline{triggering})$ which we can use the Bayes' theorem and law of total probability in here as well that can be seen in the equation 4.

$$P(dangerous|\overline{triggering}) = \frac{P(\overline{triggering}|dangerous)P(dangerous)}{P(\overline{triggering}|P(dangerous)P(dangerous) + P(\overline{triggering}|normal)P(noraml)}$$

$$P(dangerous|\overline{triggering}) = \frac{0.05 * 0.005}{0.05 * 0.005 + 0.995 * 0.995} = 0.000252455 \tag{4}$$

1.3 Expectation of false alarms and critical conditions

In order to solving this question, we assume each day as Bernoulli distribution and we defined x as follows. If false alarm then x is equal to 1 otherwise is 0.The p parameter is equal to answer of question 1, which was 0.511568123. If X be the number of false alarms in whole $n = 3650$ days then we have a Binomial distribution.

$$X \sim Bin(n = 3650,\ p = 0.511568123)$$

$$f_X(x \mid n, p) = \begin{cases} \left(\begin{array}{c} n \\ x \end{array}\right) p^x(1-p)^{n-x}, & \text{if } x = 0, 1, \ldots, n \\ 0, & \text{otherwise.} \end{cases}$$

In order to find the mean for X we can assume 3650 f Bernoulli distributions which are i.i.d and you can see in equation 5.

$$E[X] = \sum_{i=1}^{3650} E[X_i] \tag{5}$$

Cause X_i are i.i.d and the mean for each distribution is p then we can rewrite equation 5 as 6.

$$E[X] = 3650 * E[X_1] = 3650 * p = 3650 * 0.511568123 = 1867.22364895 \tag{6}$$

For unidentified critical conditions we can do the same thing and we assume Y as Binomial distribution with parameter $n = 3650, p = 0.000252455$ and for mean we have:

$$E[Y] = 3650 * E[Y_1] = 3650 * p = 3650 * 0.000252455 = 0.92146075 \tag{7}$$

It is effective because expected of unidentified critical section(in equation 6) is approximately 1 day out of 3650 days and it says mean of times that conditions are dangerous and alarm doesn't trigger is approximately 1 so it is so effective.

2 Question 2

2.1 Mean and variance of Y_1

Consider X_1 and X_2 are distributed random variables with a common mean of m and common variance of σ^2. Let Y be $aX_1 + bX_2$ then we can find mean and variance of Y as equations 8 and 9:

$$\begin{aligned} E[Y] &= E[aX_1 + bX_2] \\ E[Y] &= aE[X_1] + bE[X_2] \end{aligned} \tag{8}$$

$$\begin{aligned} Var[Y] &= Var[aX_1 + bX_2] \\ Var[Y] &= a^2Var[X_1] + b^2Var[X_2] + 2abCov(X_1, X_2) \end{aligned} \tag{9}$$

More generally we can extend these to equations 10 and 11 which Y is $\sum_{i=1}^{n} a_iX_i$.

$$\begin{aligned} E[Y] &= E[\sum_{i=1}^{n} a_iX_i] \\ E[Y] &= \sum_{i=1}^{n} a_iE[X_i] \end{aligned} \tag{10}$$

$$\begin{aligned} Var[Y] &= Var[\sum_{i=1}^{n} a_iX_i] \\ E[Y] &= \sum_{i=1}^{n} a_i^2Var[X_i] + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Cov(X_i, X_j) \end{aligned} \tag{11}$$

But in the question was declared that X_1 and X_2 are i.i.d so the $Cov(X_1, X_2)$ would be 0 and for $Y_1 = X_1 + X_2$ we have $\{a = 1,\ b = 1\}$.

$$\begin{aligned} E[Y_1] &= E[X_1] + E[X_2] = 2m \\ Var[Y_1] &= 1^2Var[X_1] + 1^2Var[X_2] = 2\sigma^2 \end{aligned} \tag{12}$$

As you can see in equation 12 mean and variance of Y_1 are $\{2m,\ 2\sigma^2\}$.

2.2 Mean and variance of Y_2

For this part of question, we have $Y_2 = 2X_1$ and our parameters are $\{a = 2,\ b = 0\}$. We can use equations 8 and 9 again to solve this part.

$$\begin{aligned} E[Y_2] &= 2E[X_1] = 2m \\ Var[Y_2] &= 4Var[X_1] = 4\sigma^2 \end{aligned} \tag{13}$$

2.3 Are they same

No they are not.Intuitively speaking, coefficient of variance will be quadratic while sum of variance will be added linearly.

2.4 Covariance between Y_1 & Y_2

We can find the covariance between Y_1 and Y_2 with definition of covariance which is in equation 14

$$Cov(Y_1, Y_2) = E[Y_1 Y_2] - E[Y_1]E[Y_2] \tag{14}$$

From sections b and c we have:

$$\begin{gathered} Y_1 = X_1 + X_2 \\ Y_2 = 2X_1 \end{gathered} \tag{15}$$

With equations 14 and 15 we can find the covariance between Y_1 and Y_2 easily:

$$\begin{aligned} Cov(Y_1, Y_2) &= E[(X_1 + X_2) * (2X_1)] - E[(X_1 + X_2)]E[2X_1] \\ &= E[2X_1^2 + 2X_1X_2] - 2(E[X_1] + E[X_2])E[X_1] \\ &= 2E[X_1^2] + 2E[X_1X_2] - 2E[X_1]^2 - 2E[X_1]E[X_2] \\ &= 2(E[X_1^2] - E[X_1]^2) + 2(E[X_1X_2] - E[X_1]E[X_2]) \\ &= 2Var(X_1) + 2Cov(X_1, X_2) \\ &= 2\sigma^2 + 0 \\ &= 2\sigma^2 \end{aligned} \tag{16}$$

3 Question 3

3.1 Probability for Ω

Each person in the sample space Ω can be born in one of 365 days so if we wrap these things up, the probability function would be:

$$P(X = \{x_1,\ x_2,\ \cdots, x_n\}) = \left(\frac{1}{365}\right)^n$$

Attention: $for\ i \in \{1, 2, \cdots, n\}\ : x_i \in \{1,\ 2,\ 3, \cdots,\ 365\}$.

3.2 $P(A)$

In order to find the probability, we can use complement of the event:

$$P(A) = 1 - P(\overline{A}) \tag{17}$$

We defined $\overline{A}$ an event that nobody matches with my birthday so each person's birthdays in the sample space can be in other 364 days of the year and probability of complement of A would be:

$$P(\overline{A}) = \left(\frac{364}{365}\right)^n \tag{18}$$

With equation 17 and 18 we can easily find the probability of A:

$$P(A) = 1 - \left(\frac{364}{365}\right)^n \tag{19}$$

At the next step we will find such n that satisfy the inequality in 20.

$$\begin{aligned} P(A) &> 0.5 \\ 1 - \left(\frac{364}{365}\right)^n &> 0.5 \\ 0.5 &> \left(\frac{364}{365}\right)^n \\ \log(0.5) &< nlog\left(\frac{364}{365}\right) \\ 252.651988844 &< n \end{aligned} \tag{20}$$

According to equation 20, the minimum value satisfying this inequality is 253.

3.3 Rational reason

Whenever n gets bigger then the chance of someone's birthday matching with mine, will be bigger. Till n is equal or grater than 365 days then the possibility of matching birthdays will be more close to 1.If i want to be more precise about it, see the equation 21.

$$\lim_{n\to\infty} P(A) = 1 \tag{21}$$

So this equation shows us that $P(A)$ is increasing due to value of n and if we want such a n that exceeds half then n must bigger than half of year which is $\frac{365}{2}$.

3.4 Simulation $P(B)$

We assume that n is less than 366 because if n is greater or equal to 366 then we are 100% sure that 2 person match their birthdays .

- $B = np.zeros(shape = (1, 366), dtype = float)$: This line initializes a numpy array B with dimensions (1, 366) to store the results of each simulation. This array will help us calculate the probability wanted in this question(from 1 to 365).
- for k in range(10): this will runs our simulations 10 times in order to see the variation of our results.
- for i in range(1, 365): This for loop show that in each iteration we are calculating probability for i individual persons.
- for j in range(10000): we have 10000 trials for each i.
- $trials = np.random.randint(low = 1, high = 366, size = (1, i))$ for each person we have random birth day and this line generate random day for each persons' birthday.
- Next part we use unique function which is pre-defined in numpy library. In order to see weather we have 2 person matching their birthdays or not.
- If 2 persons' birthdays matched then we add 1 to $B[0][i]$
- After ending 10000 trials we divide $B[0][i]$ by 10000 and check if it is bigger than 0.9 or not.If it is bigger than 0.9 then break, otherwise continue the loop.

The left one is number of n and the right one is probability:

$$n,\ P(B) = \begin{cases} 42, & 0.9163 \\ 42, & 0.9133916300000001 \\ 41, & 0.901889988944 \\ 41, & 0.90599018899889945 \\ 41, & 0.9022905990189 \\ 42, & 0.9157913391629999 \\ 41, & 0.907389839022906 \\ 42, & 0.9085915791339163 \\ 41, & 0.9073899790738984 \\ 41, & 0.9000907389979075 \end{cases} \tag{22}$$

In order to validating by 30 trials we do the same thing but in 'j for loop' we have 30 trials not 10000 times. And our numpy array is B_validating.
For the B array we can see approximately the answer is 41 for n that the probability exceeding 0.9 and it does not vary too much.

But for validating with 30 trials we have so many variation which the range of variation is $(34, 42)$.
The left one is number of n and the right one is probability:

$$n,\ validate\_P(B) = \begin{cases} 40, & 1.0 \\ 37, & 0.9288888888888889 \\ 35, & 0.9287407407407408 \\ 34, & 0.9287345679012345 \\ 41, & 0.9666666666666667 \\ 31, & 0.9240293676268861 \\ 33, & 0.9564773991769546 \\ 37, & 0.9621432098765432 \\ 37, & 0.9320714403292182 \\ 37, & 0.9977357146776406 \end{cases} \tag{23}$$

3.5 Determination of $P(B)$ using calculus

In this part we are going to determine the precise formula for our $P(B)$.Again we using complement of B and we defined $\overline{B}$ as one's birthday does not match to another person.

$$P(\overline{B}) = \frac{\prod_{i=1}^{n}(365 - i)}{365^n} \tag{24}$$

Attention: if $n \geq 366$ then we the probability of B is equal to 1 and $\overline{B}$ is to 0(cause of pigeonhole principle).Eventually probability of B is:

$$\begin{aligned} P(B) &= 1 - P(\overline{B}) \\ &= 1 - \frac{\prod_{i=1}^{n}(365 - i)}{365^n} \end{aligned} \tag{25}$$

3.6 Simulation of $P(C)$

- $C = np.zeros(shape = (1, 366), dtype = float)$: This line initializes a numpy array C with dimensions (1, 366) to store the results of each simulation. This array will help us calculate the probability wanted in this question(from 1 to 365).
- for l in range(10): this will runs our simulations 10 times in order to see the variation of our results.
- for i in range(1, 365): This for loop show that in each iteration we are calculating probability for i individual persons.
- for j in range(10000): we have 10000 trials for each i.
- $trials = np.random.randint(low = 1, high = 366, size = (1, i))$ for each person we have random birth day and this line generate random day for each persons' birthday.
- Next part we use sort function which is pre-defined in numpy library.

- If we loop over the sorted array and check 3 continuous index that match together then if we find any 3 continuous index matching together then we add 1 to $C[0][i]$ and break this loop.
- After ending 10000 trials we divide $C[0][i]$ by 10000 and check if it is bigger than 0.5 or not.If it is bigger than 0.9 then break, otherwise continue the loop.

If we repeat the simulation 10 times then we can easily find out that we have 2 different answers. It is whether 87 or 88 for n exceeding probability of C from 0.5.

$$n,\ P(C) = \begin{cases} 87, & 0.5031 \\ 87, & 0.50235031 \\ 88, & 0.5075 \\ 87, & 0.5046498050235031 \\ 87, & 0.5015504649805024 \\ 88, & 0.51115075 \\ 87, & 0.5007499050155046 \\ 87, & 0.5063500749905016 \\ 88, & 0.5044511150750001 \\ 87, & 0.5068495150635008 \end{cases} \tag{26}$$

3.7 Plausible

In the real life, the likelihood of shared birthdays is caused by this fact that some birth dates occur more frequently.Considering actual probabilities of different birthdays the chance of sharing birthdays is slightly more higher than discussed model. And some other features like holidays or christmas holidays or some cultural events can effect our problem too.

4 Question 4

4.1 Likelihood of both children are girls

Let's define our sample space which you can see that in equation 27.

$$\Omega = (\{girl, girl\}\ , \ \{girl, boy\}\ , \{boy, girl\}\ , \ \{boy, boy\}) \tag{27}$$

The question said we have already older girl so the sample space is restricted to the equation 28.

$$\Omega_{restricted} = (\{girl, girl\}\ , \ \{girl, boy\}) \tag{28}$$

So out of two choices in the $\Omega_{restricted}$, we have just one option so the probability of both children being girls given that the older child is a girl is 0.5.
We can solve this with another method explained in equation 29.

$$\begin{aligned} P(Girl_2|Girl_1) &= \frac{P(Girl_1 \cap Girl_2)}{P(Girl1)} \\ &= \frac{\frac{1}{4}}{\frac{1}{2}} \\ &= 0.5 \end{aligned} \tag{29}$$

Attention: The sequence is important here because we have a older girl. To be more precise: $\{girl, boy\}$ and $\{boy, girl\}$ are different cause in first one we have older girl and younger boy while in second one we have older boy and younger girl.

4.2 Likelihood of both children are boys

We don't know about boy being older or younger, so the sequence is not important we have same sample space discussed in equation 27.We restrict sample space that have at least one of them boy.

$$\Omega_{restricted} = (\{boy, girl\}\ , \ \{girl, boy\}\ , \ \{boy, boy\}) \tag{30}$$

Probability of that both children are boys is one out of three cause to restricted sample space.
Another approach for this is using the Bayes' theorem in equation 31:

$$\begin{aligned} P(\text{both are boy}|\text{one is boy}) &= \frac{P(\text{both are boy})}{P(\text{one is boy})} \\ &= \frac{\frac{1}{4}}{\frac{3}{4}} \\ &= \frac{1}{3} \end{aligned} \tag{31}$$

5 Question 5

We rewrite whole probabilities that are given in the question in equation 32.

$$\begin{aligned} P(blue) &= 0.01 \\ P(green) &= 0.99 \\ P(detect\_blue|blue) &= 0.99 \\ P(detect\_green|blue) &= 0.01 \\ P(detect\_blue|green) &= 0.02 \\ P(detect\_green|green) &= 0.98 \end{aligned} \tag{32}$$

we will calculate 4 conditional probabilities as follow:

$$\begin{aligned} P(blue|detect\_blue) &= \frac{P(detect\_blue|blue)P(blue)}{P(detect\_blue|green)P(green) + P(detect\_blue|blue)P(blue)} \\ &= \frac{0.99 * 0.01}{0.02 * 0.99 + 0.99 + 0.01} \\ &= \frac{0.0099}{0.0297} \\ &= 0.333333333 \end{aligned} \tag{33}$$

$$\begin{aligned} P(green|detect\_blue) &= \frac{P(detect\_blue|green)P(green)}{P(detect\_blue|green)P(green) + P(detect\_blue|blue)P(blue)} \\ &= \frac{0.02 * 0.99}{0.02 * 0.99 + 0.99 + 0.01} \\ &= \frac{0.0198}{0.0297} \\ &= 0.666666667 \end{aligned} \tag{34}$$

$$\begin{aligned} P(green|detect\_green) &= \frac{P(detect\_green|green)P(green)}{P(detect\_green|green)P(green) + P(detect\_green|blue)P(blue)} \\ &= \frac{0.98 * 0.99}{0.98 * 0.99 + 0.01 * 0.01} \\ &= \frac{0.9702}{0.9703} \\ &= 0.999896939 \end{aligned} \tag{35}$$

$$\begin{aligned} P(blue|detect\_green) &= \frac{P(detect\_green|blue)P(blue)}{P(detect\_green|green)P(green) + P(detect\_green|blue)P(blue)} \\ &= \frac{0.01 * 0.01}{0.98 * 0.99 + 0.01 * 0.01} \\ &= \frac{0.001}{0.9703} \\ &= 0.000103061 \end{aligned} \tag{36}$$

We define our table as you can see in figure 1:

	$detect\_green$	$detect\_blue$
$green$	$P(green \mid detect\_green)$	$P(green \mid detect\_blue)$
$blue$	$P(blue \mid detect\_green)$	$P(blue \mid detect\_blue)$

Figure 1: Conditional probability

And if we will the table then table would be like in figure 2.

	$detect\_green$	$detect\_blue$
$green$	0.999896939	0.666666667
$blue$	0.000103061	0.333333333

Figure 2: Conditional probability filled

6 Question 6

6.1 Probability mass function for S

we have 4 dices with $\{4,\ 6,\ 8,\ 8\}$ sides. As question said us the S is a discrete random variable which denoted to the number of sides on the chosen die.
$P_S(s)$ can be calculated as in the equation 37.

$$P_S(s) = \begin{cases} \frac{1}{4} & if\ s = 4 \\ \frac{1}{4} & if\ s = 6 \\ \frac{1}{2} & if\ s = 8 \end{cases} \tag{37}$$

6.2 $S_{given} R = 3$

We must calculate $P(S = 4|R = 3),\ P(S = 6|R = 3),\ P(S = 8|R = 3)$ with Bayes' rule:

$$\begin{aligned} P(S = 4|R = 3) &= \frac{P(S = 4 \cap R = 3)}{P(R = 3)} \\ &= \frac{P(R = 3|S = 4)P(S = 4)}{P(R = 3|S = 4)P(S = 4) + P(R = 3|S = 6)P(S = 6) + P(R = 3|S = 8)P(S = 8)} \\ &= \frac{\left(\frac{1}{4} * \frac{1}{4}\right)}{\left(\frac{1}{4} * \frac{1}{4}\right) + \left(\frac{1}{6} * \frac{1}{4}\right) + \left(\frac{1}{8} * \frac{1}{2}\right)} \\ &= \frac{\frac{1}{16}}{\frac{1}{16} + \frac{1}{24} + \frac{1}{16}} \\ &= \frac{\frac{1}{16}}{\frac{1}{6}} \\ &= 0.375 \end{aligned} \tag{38}$$

$$\begin{aligned} P(S = 6|R = 3) &= \frac{P(S = 6 \cap R = 3)}{P(R = 3)} \\ &= \frac{P(R = 3|S = 6)P(S = 6)}{P(R = 3|S = 4)P(S = 4) + P(R = 3|S = 6)P(S = 6) + P(R = 3|S = 8)P(S = 8)} \\ &= \frac{\left(\frac{1}{6} * \frac{1}{4}\right)}{\left(\frac{1}{4} * \frac{1}{4}\right) + \left(\frac{1}{6} * \frac{1}{4}\right) + \left(\frac{1}{8} * \frac{1}{2}\right)} \\ &= \frac{\frac{1}{24}}{\frac{1}{16} + \frac{1}{24} + \frac{1}{16}} \\ &= \frac{\frac{1}{24}}{\frac{1}{6}} \\ &= 0.25 \end{aligned} \tag{39}$$

$$
\begin{aligned}
P(S=8|R=3) &= \frac{P(S=8 \cap R=3)}{P(R=3)} \\
&= \frac{P(R=3|S=8)P(S=8)}{P(R=3|S=4)P(S=4)+P(R=3|S=6)P(S=6)+P(R=3|S=8)P(S=8)} \\
&= \frac{\left(\frac{1}{8}*\frac{1}{2}\right)}{\left(\frac{1}{4}*\frac{1}{4}\right)+\left(\frac{1}{6}*\frac{1}{4}\right)+\left(\frac{1}{8}*\frac{1}{2}\right)} \\
&= \frac{\frac{1}{16}}{\frac{1}{16}+\frac{1}{24}+\frac{1}{16}} \\
&= \frac{\frac{1}{16}}{\frac{1}{6}} \\
&= 0.375
\end{aligned}
\tag{40}
$$

So if we check the results of our conditional probabilities in equations 38, 39 and 40 then $P(S=4|R=3)$, $P(S=8|R=3)$ have same and highest probabilities among others.

6.3 $S_{given}R=6$

In this section we already know that $P(S=4|R=6)=0$ cause $S=4$ has only 4 sides. So calculating $P(S=6|R=8),\ P(S=8|R=6)$ would be enough.

$$
\begin{aligned}
P(S=6|R=6) &= \frac{P(S=6 \cap R=6)}{P(R=6)} \\
&= \frac{P(R=6|S=6)P(S=6)}{P(R=6|S=4)P(S=4)+P(R=6|S=6)P(S=6)+P(R=6|S=8)P(S=8)} \\
&= \frac{\left(\frac{1}{4}*\frac{1}{6}\right)}{0+\left(\frac{1}{6}*\frac{1}{4}\right)+\left(\frac{1}{8}*\frac{1}{2}\right)} \\
&= \frac{\frac{1}{24}}{\frac{1}{24}+\frac{1}{16}} \\
&= \frac{\frac{1}{24}}{\frac{5}{48}} \\
&= 0.4
\end{aligned}
\tag{41}
$$

$$
\begin{aligned}
P(S=8|R=6) &= \frac{P(S=8 \cap R=6)}{P(R=6)} \\
&= \frac{P(R=6|S=8)P(S=8)}{P(R=6|S=4)P(S=4)+P(R=6|S=6)P(S=6)+P(R=6|S=8)P(S=8)} \\
&= \frac{\left(\frac{1}{8}*\frac{1}{2}\right)}{0+\left(\frac{1}{6}*\frac{1}{4}\right)+\left(\frac{1}{8}*\frac{1}{2}\right)} \\
&= \frac{\frac{1}{16}}{\frac{1}{24}+\frac{1}{16}} \\
&= \frac{\frac{1}{16}}{\frac{5}{48}} \\
&= 0.6
\end{aligned} \tag{42}
$$

According to equations 41 and 42 the dice with 8 sides have most likely choice.

6.4 $S_{given}R = 8$

In this case we can easily tell that 8 sided dice has most likely choice cause if we chose 4 sided dice it doesn't have any side bigger than 4. Similar to 4 sided dice, if we chose 6 sided then it doesn't have any side more than 6 so both probabilities for $S = 4,\ S = 6$ would be 0.

7 Question 7

7.1 STD of $X,\ Y$ and Z

As question said:

$$Z = \frac{X+Y}{2} \tag{43}$$

At first we are going to determine the probability mass functions for each random variables $X,\ Y$.

$$P_X(x) = \begin{cases} \frac{1}{4} & x=1 \\ \frac{1}{4} & x=2 \\ \frac{1}{4} & x=3 \\ \frac{1}{4} & x=4 \end{cases} \tag{44}$$

$$P_Y(y) = \begin{cases} \frac{1}{6} & y=1 \\ \frac{1}{6} & y=2 \\ \frac{1}{6} & y=3 \\ \frac{1}{6} & y=4 \\ \frac{1}{6} & y=5 \\ \frac{1}{6} & y=6 \end{cases} \tag{45}$$

Next step is find the mean and variance for each random variable:

$$\begin{aligned} E[X] &= \sum_{x=1}^{x=4} x P_X(x) \\ &= \left(\frac{1}{4} * (1+2+3+4)\right) \\ &= \left(\frac{1}{4} * \frac{(4*5)}{2}\right) \\ &= 2.5 \end{aligned} \tag{46}$$

$$\begin{aligned} E[Y] &= \sum_{y=1}^{y=6} y P_Y(y) \\ &= \left(\frac{1}{6} * (1+2+3+4+5+6)\right) \\ &= \left(\frac{1}{6} * \frac{(6*7)}{2}\right) \\ &= 3.5 \end{aligned} \tag{47}$$

To calculating $Var[X],\ Var[Y]$, we need to calculate $E[X^2],\ E[Y^2]$ at first.

$$\begin{aligned} E[X^2] &= \sum_{x=1}^{x=4} x^2 P_X(x) \\ &= \left(\frac{1}{4} * (1+4+9+16)\right) \\ &= \left(\frac{1}{4} * \frac{(4*5*9)}{6}\right) \\ &= 7.5 \end{aligned} \tag{48}$$

$$\begin{aligned} E[Y^2] &= \sum_{y=1}^{y=6} y^2 P_Y(y) \\ &= \left(\frac{1}{6} * (1+4+9+16+25+36)\right) \\ &= \left(\frac{1}{6} * \frac{(6*7*13)}{6}\right) \\ &= \frac{91}{6} \end{aligned} \tag{49}$$

$$\begin{aligned} Var[X] &= E[X^2] - E[X]^2 \\ &= 7.5 - 6.25 \\ &= 1.25 \\ \sigma_X &= \frac{\sqrt{5}}{2} \end{aligned} \tag{50}$$

$$\begin{aligned} Var[Y] &= E[Y^2] - E[Y]^2 \\ &= \frac{91}{6} - \frac{49}{4} \\ &= \frac{35}{12} \\ \sigma_Y &= \frac{\sqrt{35}}{2\sqrt{3}} \end{aligned} \tag{51}$$

The σ_X and σ_Y calculated in equations 50 and 51. Instead of calculating standard deviation of Z, we calculate variance of Z.

$$\begin{aligned}
Var[Z] &= Var\left[\left(\frac{X+Y}{2}\right)\right] \\
&= \frac{Var\left[X+Y\right]}{4} \\
&= \frac{Var[X]+Var[Y]+2Cov(X,Y)}{4} \\
&= \frac{\frac{5}{4}+\frac{35}{12}+0}{4} \\
&= \frac{\frac{50}{12}}{4} \\
&= \frac{25}{24}
\end{aligned} \tag{52}$$

Attention: $Cov(X,Y)$ is zero because X and Y are independent.Let's prove it this way, if any event occurred in X then it won't give us any information about Y so they are independent.(It is like tossing same dice over and over again if you see the number of first toss then it won't give you any information about number of second toss)

7.2 PMF and CDF of Z

At first we define our sample space for Z:

$$Z = \{2,3,4,5,6,7,8,9,10\} \tag{53}$$

we have 4 choice from X and 6 choice from Y then we have 24 choices for (x,y) pairs. Next step we are finding pairs which sum of them are equal to $Z=2$ and we do that again for $Z=3$ and so on.

$$pairs(x,y)\ for\ each\ z \begin{cases}
\{(1,1)\} & z=2 \\
\{(1,2),(2,1)\} & z=3 \\
\{(1,3),(2,2),(2,2)\} & z=4 \\
\{(1,4),(2,3),(3,2),(4,1)\} & z=5 \\
\{(1,5),(2,4),(3,3),(4,2)\} & z=6 \\
\{(1,6),(2,5),(3,4),(4,3)\} & z=7 \\
\{(2,6),(3,5),(4,4)\} & z=8 \\
\{(3,6),(4,5)\} & z=9 \\
\{(4,6)\} & z=10
\end{cases} \tag{54}$$

So the probability mass function for Z:

$$P_Z(z) = \begin{cases} \frac{1}{24} & z = 2 \\ \frac{2}{24} & z = 3 \\ \frac{3}{24} & z = 4 \\ \frac{4}{24} & z = 5 \\ \frac{4}{24} & z = 6 \\ \frac{4}{24} & z = 7 \\ \frac{3}{24} & z = 8 \\ \frac{2}{24} & z = 9 \\ \frac{1}{24} & z = 10 \end{cases} \tag{55}$$

For example if $z = 2$ then we have one pair out twenty four pairs that this event comes true.And Also for cumulative distribution function for Z:

$$F_Z(z) = \begin{cases} 0 & z < 2 \\ \frac{1}{24} & z < 3 \\ \frac{3}{24} & z < 4 \\ \frac{6}{24} & z < 5 \\ \frac{10}{24} & z < 6 \\ \frac{14}{24} & z < 7 \\ \frac{18}{24} & z < 8 \\ \frac{21}{24} & z < 9 \\ \frac{23}{24} & z < 10 \\ 1 & z \geq 10 \end{cases} \tag{56}$$

7.3 Overall gain

At first we are going to calculate what is the probability if $X > Y$:

$$Pair(x, y) = \{(2,1), (3,1), (3,2), (4,1), (4,2), (4,3)\} \text{ where } x > y \tag{57}$$

According to the 57 we have 6 pairs which $x > y$, so x and y of 16 pairs would satisfy $x \leq y$. Then the probability of win a round is:

$$P(\text{wining a round}) = P(X > Y) = \frac{6}{24} = 0.25 \tag{58}$$

Assume that for each round we have a Bernoulli distribution which if $x = 1$ then we win $2X$ dollars and if $x = 0$ then we loss 1 dolor.The mean of gain for each round will be:

$$\begin{aligned} E[\text{gain a round}] &= E[2X] * P(\text{wining a round}) * 1 + 0 : x > y \\ &= 2 * (\frac{(2+3+3+4+4+4)}{6}) * 0.25 \\ &= \frac{5}{3} \end{aligned} \tag{59}$$

So the mean of gain for each round is 1.25. To calculating mean of gain for 60 rounds, assume having a Binomial distribution with this parameters:

$$M \sim Bin(n = 60, p = P(\text{wining a round}))$$

Binomial distribution will be considered as n i.i.d Bernoulli distributions (each X_i) and the mean of M can easily calculated in equation 60.

$$\begin{aligned} E[M] &= \sum_{i=1}^{60} E[X_i] \\ &= 60 * E[X_1] \\ &= 60 * \frac{5}{3} \\ &= 100 \end{aligned} \tag{60}$$

Similar to mean of gain part, we can define X as Bernoulli distribution which if $x = 1$ we lose the game else we win the game so probability of losing game in a round will be:

$$P(\text{losing a round}) = P(X \leq Y) = \frac{18}{24} = 0.75 \tag{61}$$

Mean of loss for each round will be:

$$\begin{aligned} E[\text{lose a round}] &= P(\text{losing a round}) * 1 + 0 \\ &= 0.75 \end{aligned} \tag{62}$$

Again for 60 rounds we have a Binomial distribution (L):

$$L \sim Bin(n = 60, p = P(\text{losing a round}))$$

$$\begin{aligned} E[M] &= \sum_{i=1}^{60} E[X_i] \\ &= 60 * E[X_1] \\ &= 60 * 0.75 \\ &= 45 \end{aligned} \tag{63}$$

Overall expected value would be:

$$\begin{aligned} E[Overall] &= E[M] - E[L] \\ &= 100 - 45 \\ &= 55 \end{aligned} \tag{64}$$

8 Question 8

8.1 Raisins can be found in one box

We are given a function which it's for how many raisins are per height and also given that range of h is $(0, 30)$. To calculating how many raisins are in a cereal box we must use integral:

$$\begin{aligned}
\text{raisins in cereal box} &= \int_{h=0}^{h=30} f(h)dh \\
&= \int_{0}^{30} (40-h)dh \\
&= \left[40h - \frac{h^2}{2}\right]_0^{30} \\
&= 40*30 - \frac{900}{2} \\
&= 1200 - 450 \\
&= 750
\end{aligned} \tag{65}$$

750 raisins can be found in one cereal box.

8.2 PDF for H

Maximum of raisins in the box is 750 and we can have $f(h)$ raisins for each height so the $g(h)$ will be:

$$g(h) = \begin{cases} \frac{40-h}{750} & 0 \leq h \leq 30 \\ 0 & o.w. \end{cases} \tag{66}$$

8.3 CDF for H

To find the $G(h)$ we must calculate integral of $g(h)$, in other words:

$$G(h) = \begin{cases} 0 & h \leq 0 \\ \frac{40h - \frac{h^2}{2}}{750} & 0 \leq h \leq 30 \\ 1 & 30 \leq h \end{cases} \tag{67}$$

8.4 Probability of bottom third

we must find the probability of heights are in the range of $(0, 10)$:

$$\begin{aligned}
P(0 \leq h \leq 10) &= G(10) - G(0) \\
&= \frac{350}{750} \\
&= \frac{7}{15}
\end{aligned} \tag{68}$$

9 Question 9

9.1 Joint distribution and Covariance and Correlation of $X,\ Y$

Both of random variable are :

$$P_X(x) = \begin{cases} \frac{1}{2} & x = 1 \\ \frac{1}{2} & x = -1 \\ 0 & o.w. \end{cases} \tag{69}$$

$$P_Y(y) = \begin{cases} \frac{1}{2} & y = 1 \\ \frac{1}{2} & y = -1 \\ 0 & o.w. \end{cases} \tag{70}$$

If c denote as the probability of $(x = 1, y = 1)$ then for joint distribution of X, Y we have:

	$x = -1$	$x = 1$
$y = -1$	$p(x = -1, y = -1)$	$p(x = 1, y = -1)$
$y = 1$	$p(x = -1, y = 1)$	$p(x = 1, y = 1)$

Figure 3: Joint distribution of X, Y

we know that sum of each row or column must be $\frac{1}{2}$.(for example in first row $y = -1$ the sum of probability must be $\frac{1}{2}$ due to probability of $p(y = -1) = 1/2$): we can use equation 71 to calculate the $Cov(X, Y)$.

	$x = -1$	$x = 1$
$y = -1$	c	$\frac{1}{2} - c$
$y = 1$	$\frac{1}{2} - c$	c

Figure 4: Fill with c

$$\begin{aligned} Cov(X, Y) &= E[XY] - E[X]E[Y] \\ &= \sum_x \sum_y xyP_{X,Y}(x, y) - \sum_x xP_X(x) * \sum_y yP_Y(y) \\ &= \left(c + c - \frac{1}{2} + c - \frac{1}{2} + c\right) - \left(\frac{1}{2} - \frac{1}{2}\right) * \left(\frac{1}{2} - \frac{1}{2}\right) \\ &= 4c - 1 \end{aligned} \tag{71}$$

Attention: mean of both variable are zero.
For correlation we have :

$$\begin{aligned}
Cor(X,Y) &= \frac{Cov(X,Y)}{\sigma_x \sigma_y} \\
&= \frac{4c-1}{\sqrt{(E[X^2]-E[X]^2)}\sqrt{(E[Y^2]-E[Y]^2)}} \\
&= \frac{4c-1}{\sqrt{E[X^2]}\sqrt{E[Y^2]}} \\
&= \frac{4c-1}{\sqrt{\sum_x x^2 P_X(x)}\sqrt{\sum_y y^2 P_Y(y)}} \\
&= \frac{4c-1}{\sqrt{(\frac{1}{2}*2)}\sqrt{(\frac{1}{2}*2)}} \\
&= 4c-1
\end{aligned} \tag{72}$$

9.2 Determine value c

If X, Y are independent then $f_{X,Y}(x,y)$ must be equal to $f_X(x)f_Y(y)$ Then the values in table 4 must equal to table 5.

	$x=-1$	$x=1$
$y=-1$	0.25	0.25
$y=1$	0.25	0.25

Figure 5: Joint distribution for X, Y if they are independent

Attention:Due to independent between X, Y we can fill table 5 as follow:
$f_{X,Y}(x=1, y=1) = f_X(x=1)f_Y(y=1) = \frac{1}{2} * \frac{1}{2} = 0.25$
So value of c would be $\frac{1}{4}$ in order to these tables match their values.
For having 100% correlation between X, Y then $Cor(X,Y)$ must be equal to weather 1 or -1.From equation 72 we can calculate possible values for c.

$$Cor(X,Y) = \begin{cases} 1 & \rightarrow 4c-1=1 \rightarrow c=0.5 \\ -1 & \rightarrow 4c-1=-1 \rightarrow c=0 \end{cases} \tag{73}$$

If value of c is equal to 0.5 or 0, then correlation will be 100%.

10 Question 10

10.1 Generate random variables

we generated random numbers for 5 different distribution and plot via histogram in python:

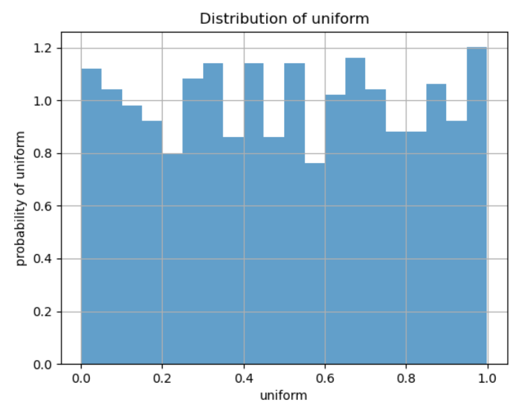

Figure 6: Uniform

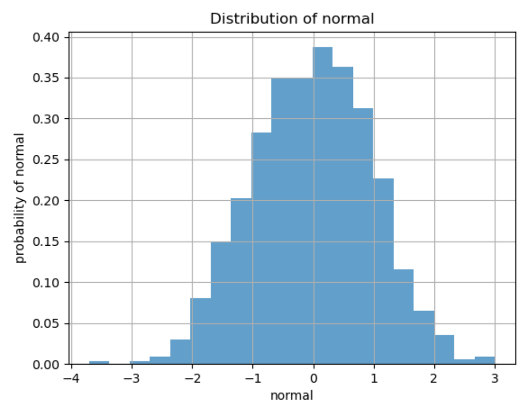

Figure 7: Normal

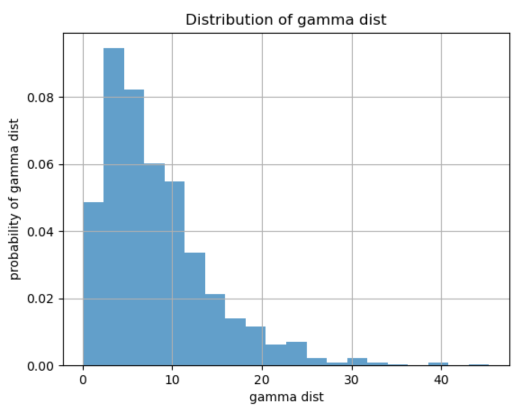

Figure 8: Gamma with parameter $(2, 4)$

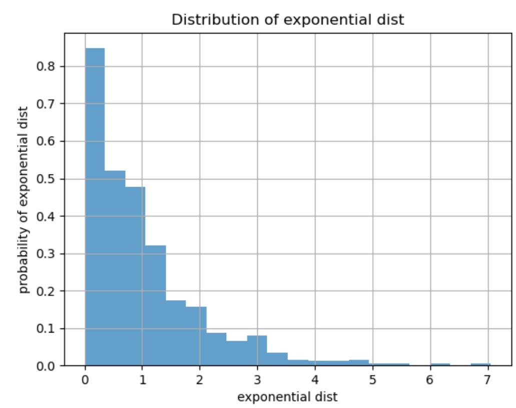

Figure 9: Exponential

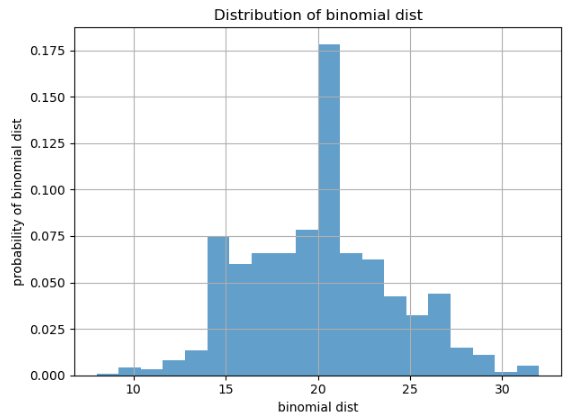

Figure 10: Binomial with parameter $(n = 100, p = 0.2)$

As you can see we have generated random variables which question asked us to.

10.2 Mean values

Mean of each are shown in figures 11 ,12,13,14,15.

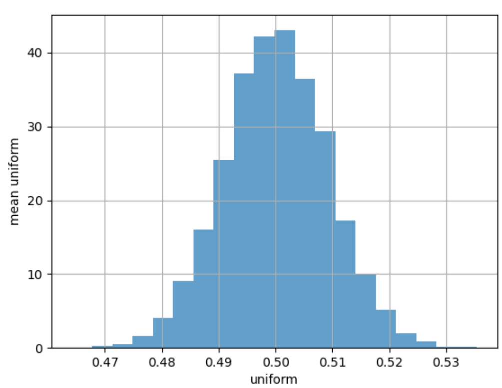

Figure 11: mean of Uniform

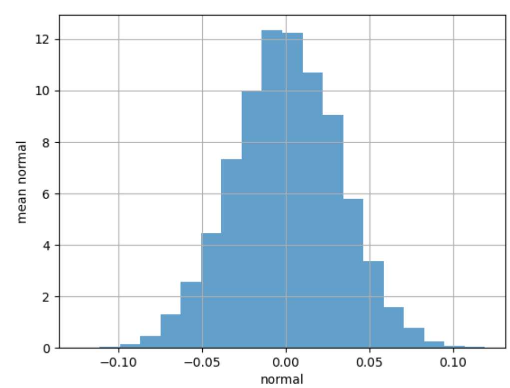

Figure 12: mean of Normal

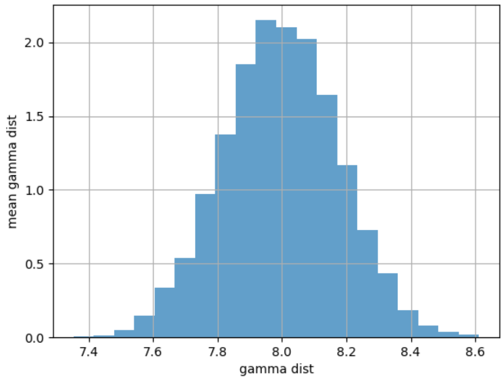

Figure 13: mean of Gamma with parameter $(2, 4)$

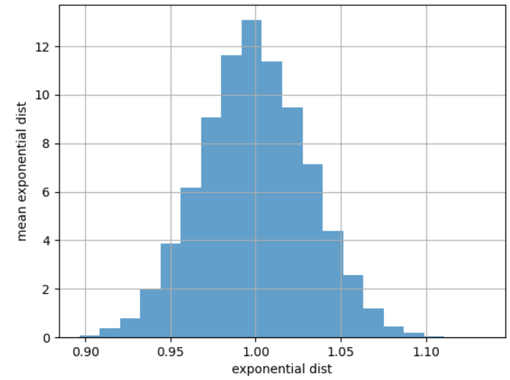

Figure 14: mean of Exponential

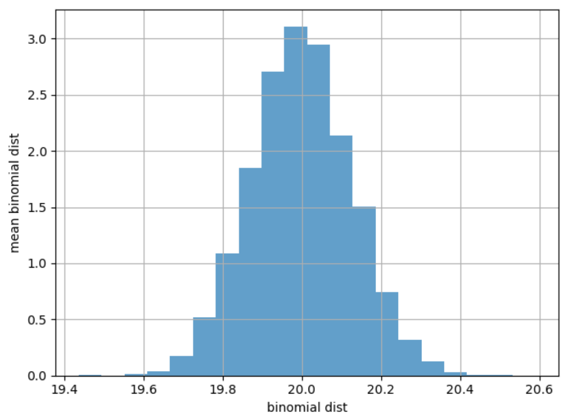

Figure 15: mean of Binomial with parameter $(n = 100, p = 0.2)$

From Central limit theorem (CLT) we know if we have $X_1, X_2, \cdots, X_n, \cdots$ random variables which are sample from a single distribution and are independent then the $\overline{X_n}$ for $n \to \infty$ will converge to Normal distribution. We are getting samples each iterations so every sample are independent to another and they are identical so from central limit theorem mean of them converges to normal distribution which you can see figures above.

10.3 Cleaning data and non-sense data

Load dataset with pandas library and the shape of data is $(205, 26)$.
Given data has missing values which are filled by question mark.We are deleting every row if data was invalid(filled by question marked).
In the next step we are going to find the duplicate rows and delete them from our data.
After doing this we have a cleaned data with the shape $(159, 26)$ and non-sense data will be deleted.

10.4 Car's Frequency

For bar plotting car's frequency, we must find how many unique cars are in our dataset so we use value_counts function to do such a thing, after that we can plot bar which can be seen in figure 16.

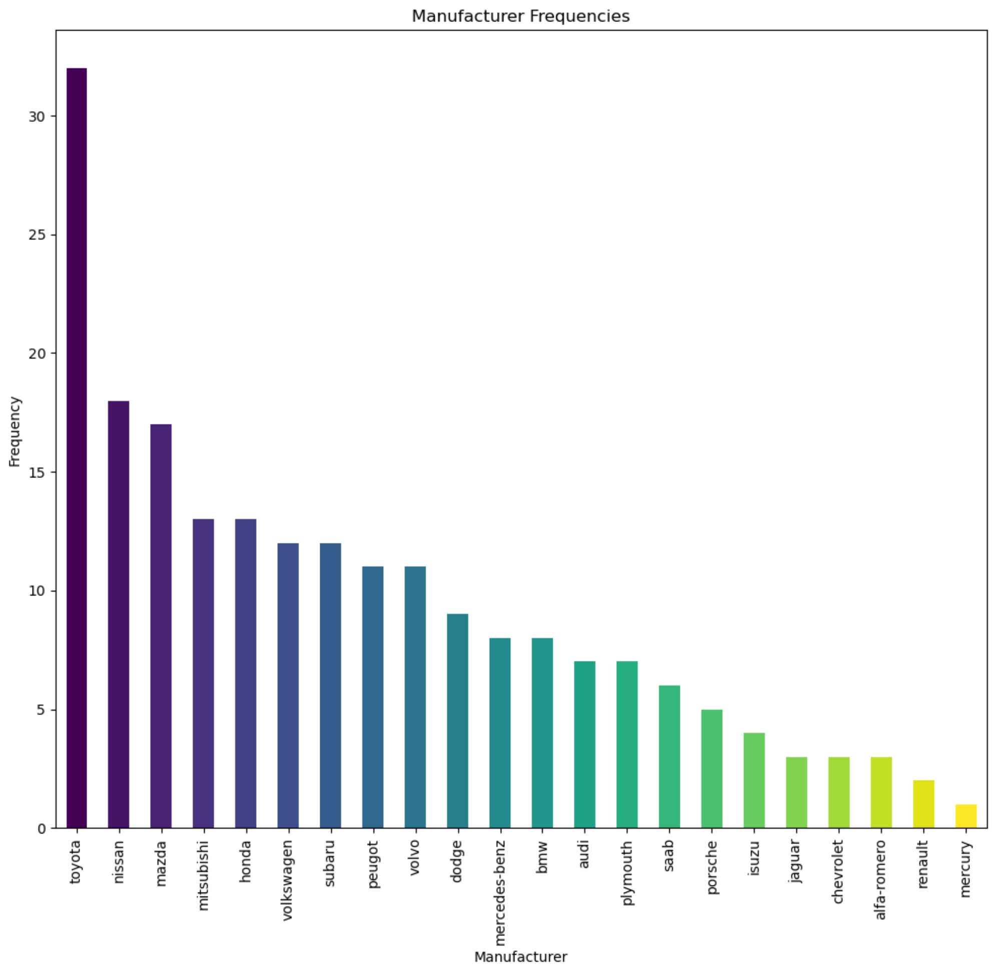

Figure 16: Car's frequency

As you can see in figure 16 Toyota has most cars.

10.5 Skewness and kurtosis

In order to calculate skewness and kurtosis for each column, we must find numeric columns(columns with numbers). After finding numeric columns, from scipy.stats we import skew and kurtosis functions:
skewness and kurtosis of symboling are: $(0.09405221860667115, -0.5492671501283724)$
skewness and kurtosis of wheel-base are: $(0.9060976267244266, 0.5794342110215762)$
skewness and kurtosis of length are: $(-0.06535155385138845, -0.2387109601067401)$
skewness and kurtosis of width are: $(0.9081735680297252, 0.7873934040163175)$
skewness and kurtosis of height are: $(0.16680954927243458, -0.30849161683950355)$
skewness and kurtosis of curb-weight are: $(0.77463831626637, 0.11112944373537648)$

skewness and kurtosis of engine-size are: $(1.4765102751920993, 2.820810722328159)$
skewness and kurtosis of compression-ratio are: $(2.68460608233467, 5.509811851769003)$
skewness and kurtosis of city-mpg are: $(0.7267259620086014, 1.0759516945468226)$
skewness and kurtosis of highway-mpg are: $(0.595366801752236, 0.7657863653351233)$
Concept of skewness and kurtosis are very important concepts which shown in figure 17.

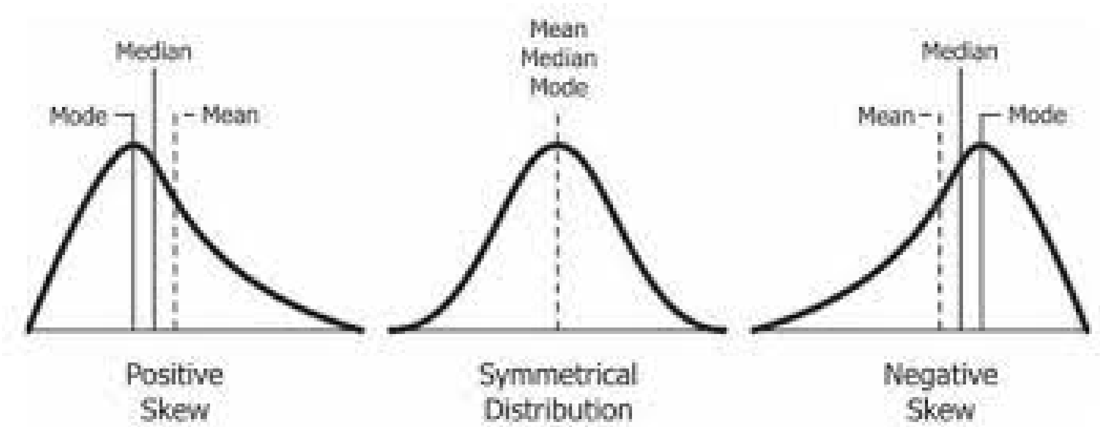

Figure 17: Skewness concept

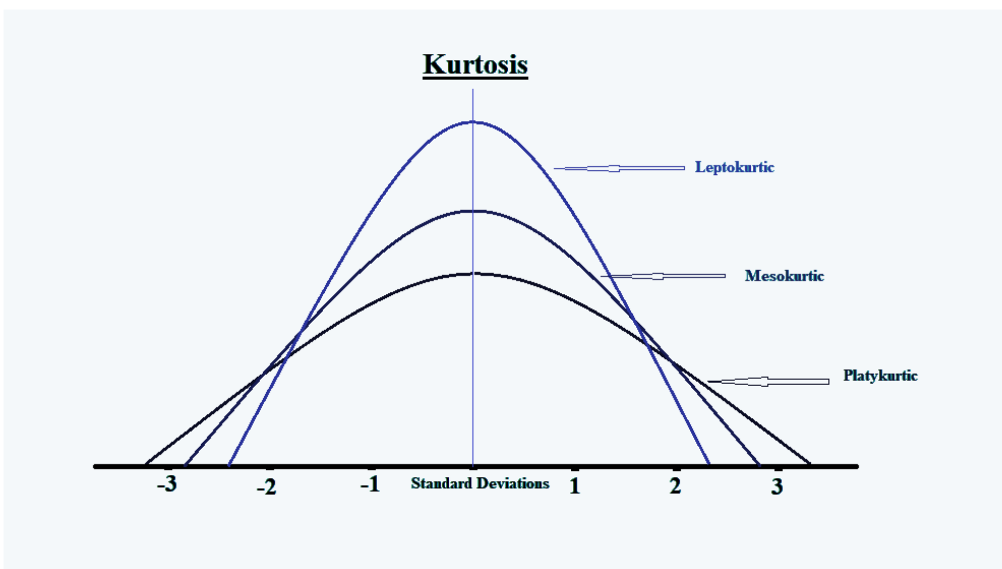

Figure 18: Kurtosis concept

More precisely: skewness measures asymmetry of our numeric data.It's concept help us determine whether data is skewed to the right(positive skewed) or to the left(negative skewed) while kurtosis describes the extent to which data is heavy-tailed (has outliers) or light-tailed (data clusters near the mean).If positive kurtosis we call it Leptokurtic and if kurtosis be zero then it called Mesokurtic otherwise we call it Platykurtic.

10.6 Scatter engine-size and price

we use scatter plot from matplotlib.pyplot for do this for us.At first we get engine-size and price from our cleaned data and pass it to our function which can be shown in figure 19.

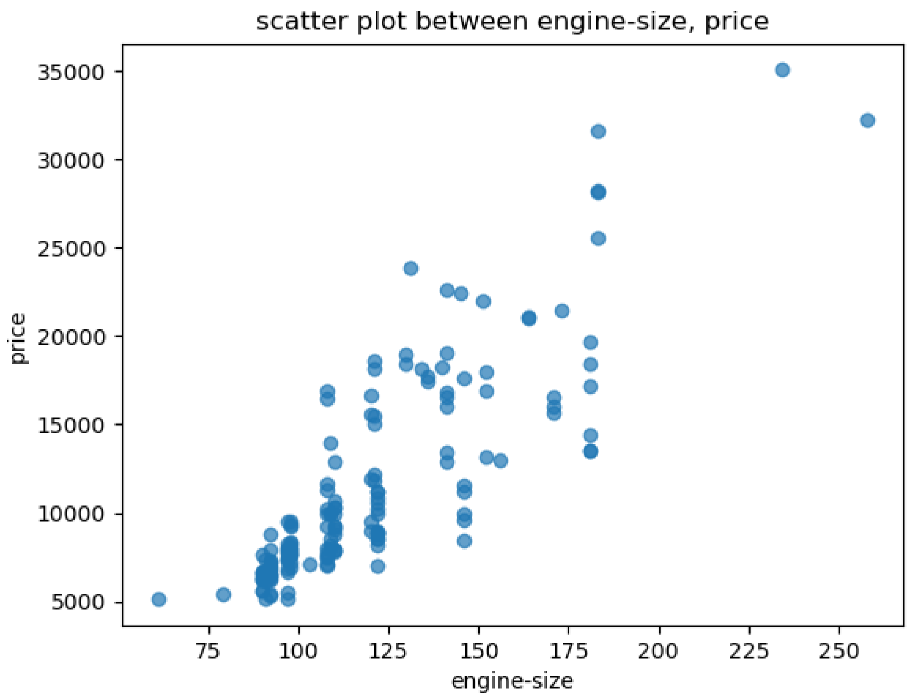

Figure 19: scatter between engine-size and price

As you can see in figure 19 engine-size and price are correlated and they are positive associated.
In other word increasing one effect on another one so they are absolutely correlated and positive associated.

10.7 Pair plot

Figure 20: Kurtosis concept

As you can see in the 20 we have pair plot for multivariate analysis and it will show us scatter plot for multiple variants in our data.Also we can see if any data are correlated to other or not.

10.8 Correlation matrix

We know that correlation of two random variables are a number in range of $[-1, 1]$. So if correlation of two different random variable be closer to 1 or -1 then they are more correlated.
See the figure 21 correlation between wheel-base and length are close to 1 so they are more correlated or correlation between city-mpg and highway-mpg are close to 1 so they are more correlated.

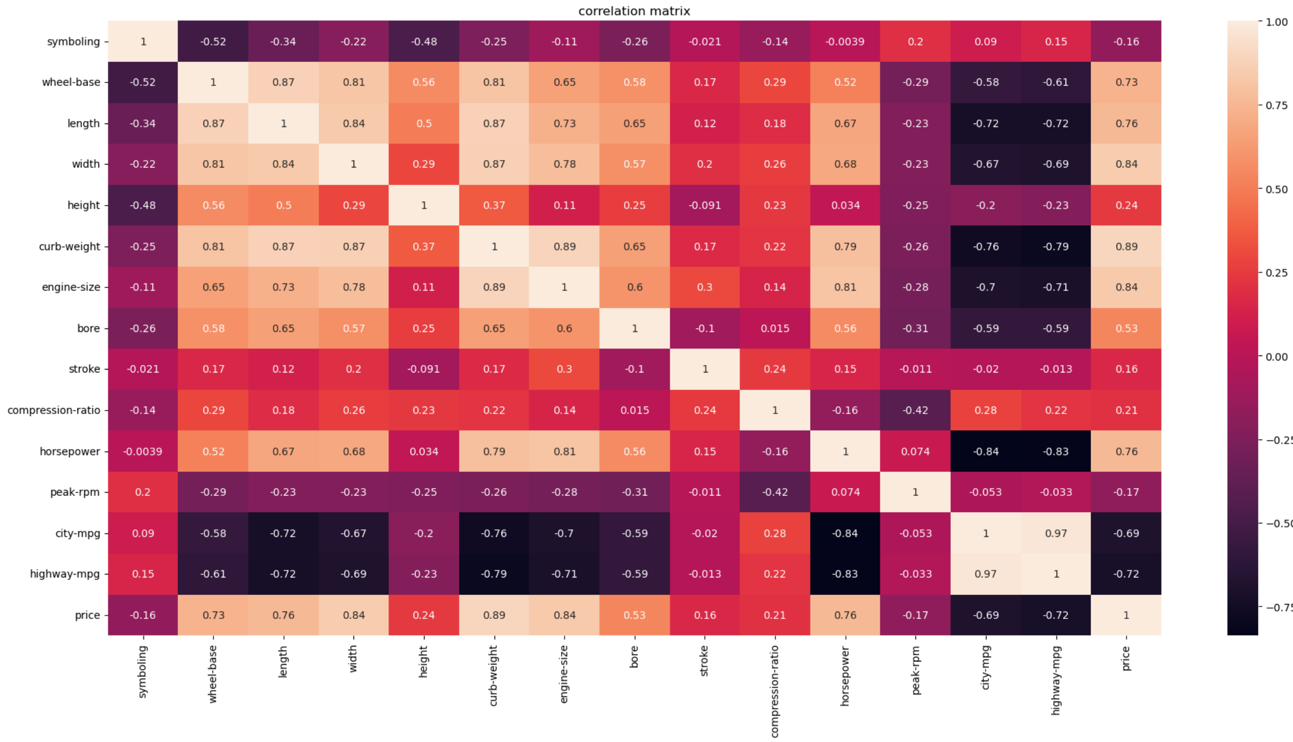

Figure 21: Kurtosis concept

10.9 Numerical columns box plots

We are going to box plot every numeric columns:

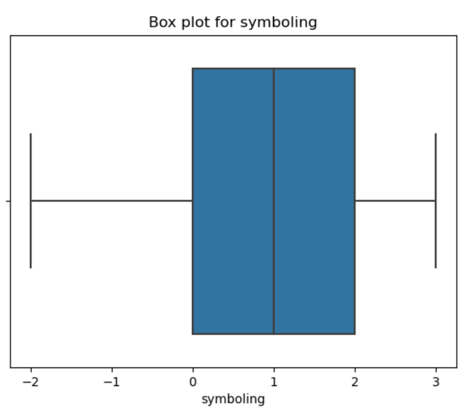

Figure 22: symboling, percentile $= (0.000, 2.00)$, $IQR = 2$, whiskers $= (-2, 3)$

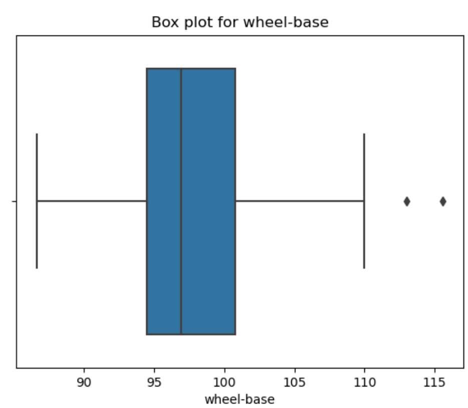

Figure 23: wheel-base, percentile $= (94.500, 100.80)$, $IQR = 6.3$, whiskers $= (86.6, 110.0)$

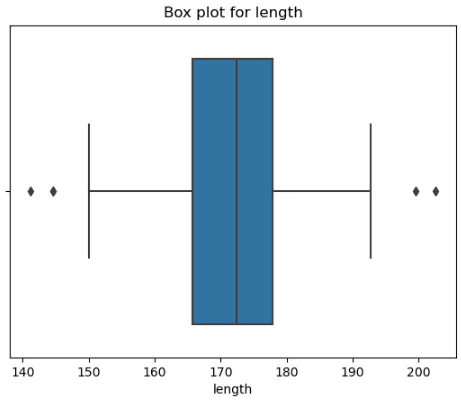

Figure 24: length, percentile $= (165.650, 177.80)$, $IQR = 12.150$, whiskers $= (150.0, 192.7)$

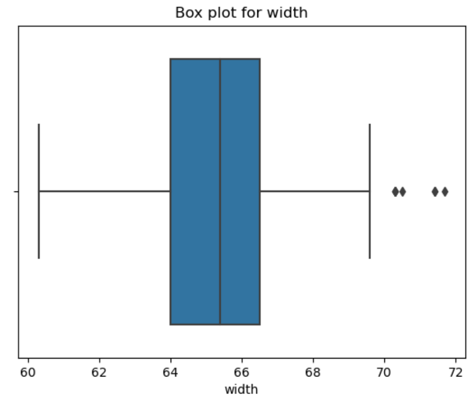

Figure 25: width, percentile $= (64.000, 66.50)$, $IQR = 2.5$, whiskers $= (60.3, 69.6)$

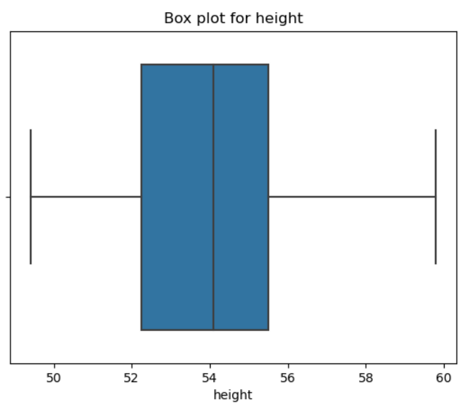

Figure 26: height, percentile $= (52.250, 55.50)$, $IQR = 3.25$, whiskers $= (49.4, 59.8)$

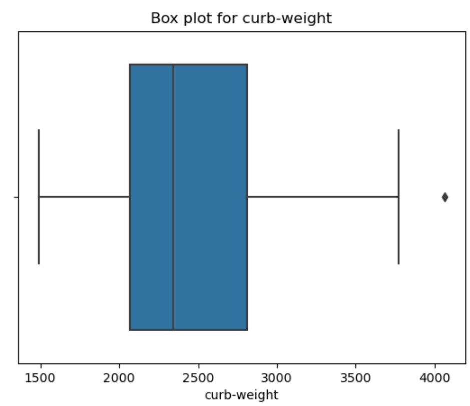

Figure 27: curb-weight, percentile $= (2065.500, 2809.50)$, $IQR = 744.00$, whiskers $= (1488, 3770)$

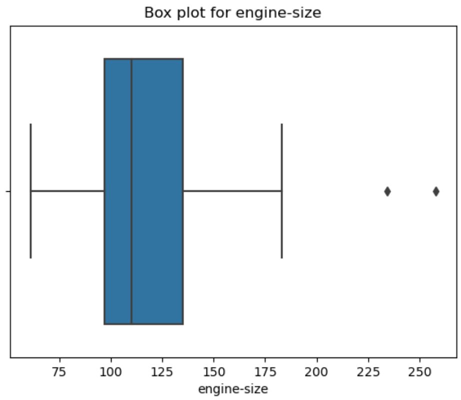

Figure 28: engine-size, percentile = $(97.000, 135.00)$, $IQR = 38.00$, whiskers = $(61, 183)$

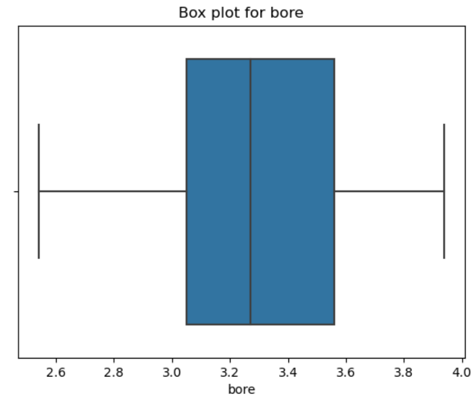

Figure 29: bore, percentile = $(3.050, 3.56)$, $IQR = 0.51$, whiskers = $(2.54, 3.94)$

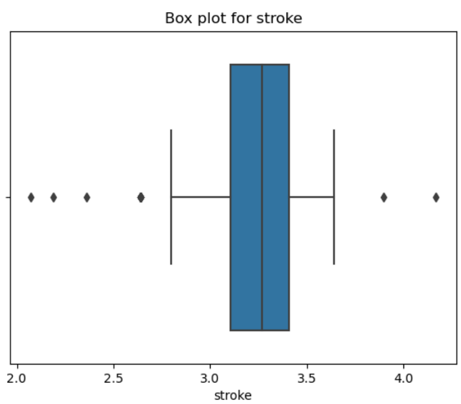

Figure 30: stroke, percentile = $(3.105, 3.41)$, $IQR = 0.305$, whiskers = $(2.8, 3.64)$

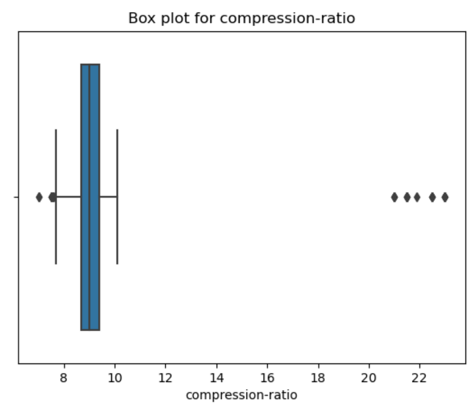

Figure 31: compression-ratio, percentile = $(8.700, 9.40)$, $IQR = 0.7$, whiskers = $(7.7, 10.1)$

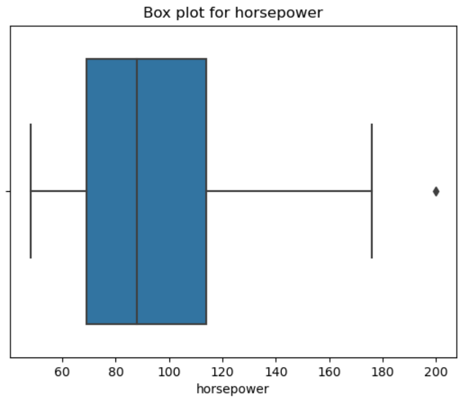

Figure 32: horsepower, percentile = $(69.000, 114.00)$, $IQR = 45$, whiskers = $(48, 176)$

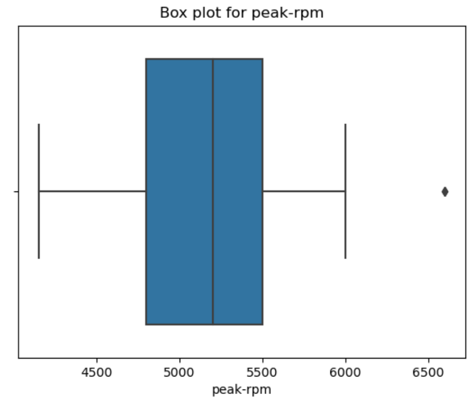

Figure 33: peak-rpm, percentile = $(4800.000, 5500.00)$, $IQR = 700$, whiskers = $(4150, 6000)$

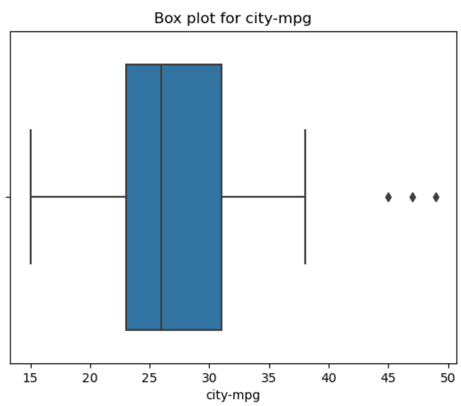

Figure 34: city-mpg, percentile = $(23.000, 31.00)$, $IQR = 8$, whiskers = $(15, 38)$

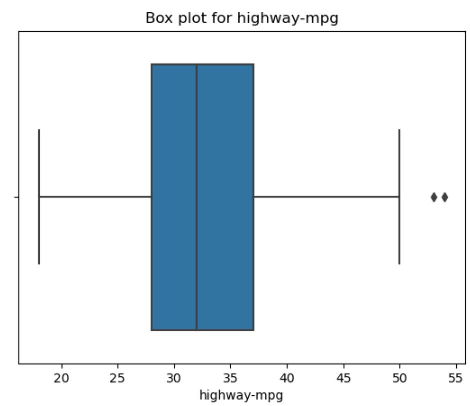

Figure 35: highway-mpg, percentile = $(28.000, 37.00)$, $IQR = 9$, whiskers = $(18, 50)$

11 Question 11

From the list given just two of them can be random variable which they are: Sample mean and The largest value in the sample.
Because sample mean is calculating from sample random variable and it can vary from sample to sample and the largest value in the sample depends on the values in the sample therefor it is a random variable.
Population mean is a fixed number.
Population size is a fixed value as well.
Variance of the sample mean it is not a random variable because variance of a random variable is fixed.
Population variance is a fixed number as well.

12 Question 12

At first we defined a function named estimate pi which gets a trials size and number of figure iterator in order to plot the figure for each iteration you want.
Then We are going use 2 random variable which are uniform distribution with the domain of $D_f = [0, 1]$. Afterward we must check how many pair points are in the circle,to be more precise about this we are checking that $x[i]^2 + y[i]^2 <= 1$.Further, if that i^{th} point satisfy this then we will add it to our circle points.
Eventually we divide number of circle's points from given trials size and cause we are checking a quarter of a circle then we must multiply that number by 4.
As you can see in the figures 36, 37, 38 and 39.

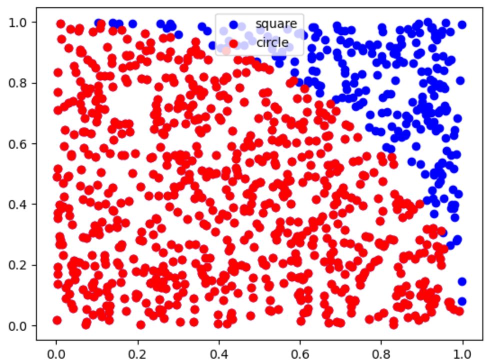

Figure 36: trials size = 1000, estimated pi = 3.176, real pi = 3.141592653589793

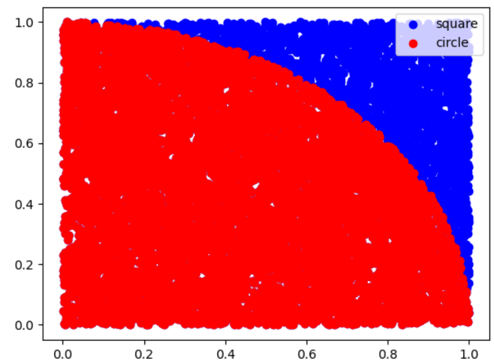

Figure 37: trials size = 10000, estimated pi = 3.1224, real pi = 3.141592653589793

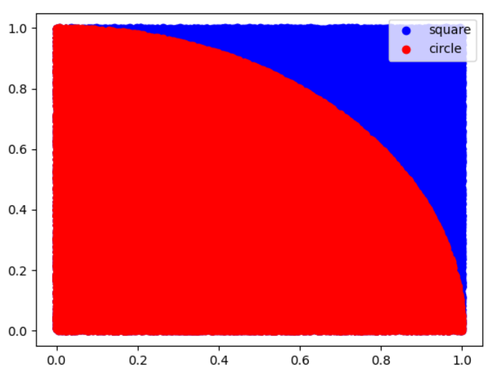

Figure 38: trials size = 100000, estimated pi = 3.1436, real pi = 3.141592653589793

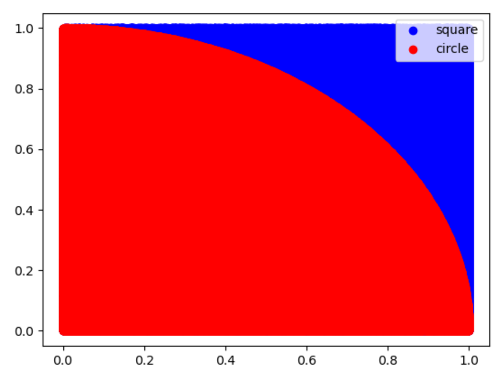

Figure 39: trials size = 1000000, estimated pi = 3.143132, real pi = 3.141592653589793

From Monte Carlo method we now if our trials size gets bigger then estimated pi would be more close to real pi.

13 Question 13

13.1 Create program simulating gambler's ruin scenario

For simulate the Gambler's Ruin scenario , we defined a function named gambler scenario which gets initial stake, bet amount, target amount and probability.
In the function we getting a random number between 0 and 1 then we check this if it is lower than our probability, if was lower then we lost the round otherwise we won the round.
After that we have conditions which are defined in equation 74.

$$\begin{aligned} \text{if win} = 1: &\begin{cases} \text{initial stake} + \text{bet mount} & initial\ stake + bet\ amount < target\ amount \\ \text{target amount} & o.w. \end{cases} \\ \text{else:} &\begin{cases} \text{initial stake} - \text{bet amount} & initial\ stake - bet\ amount > 0 \\ 0 & o.w. \end{cases} \end{aligned} \tag{74}$$

13.2 Simulate till broke or reach to target amount

We simulate infinite rounds for gambler's ruin and track the gambler's stake in order to reach the state that we are whether broke or target amount which can be seen in figure 40.

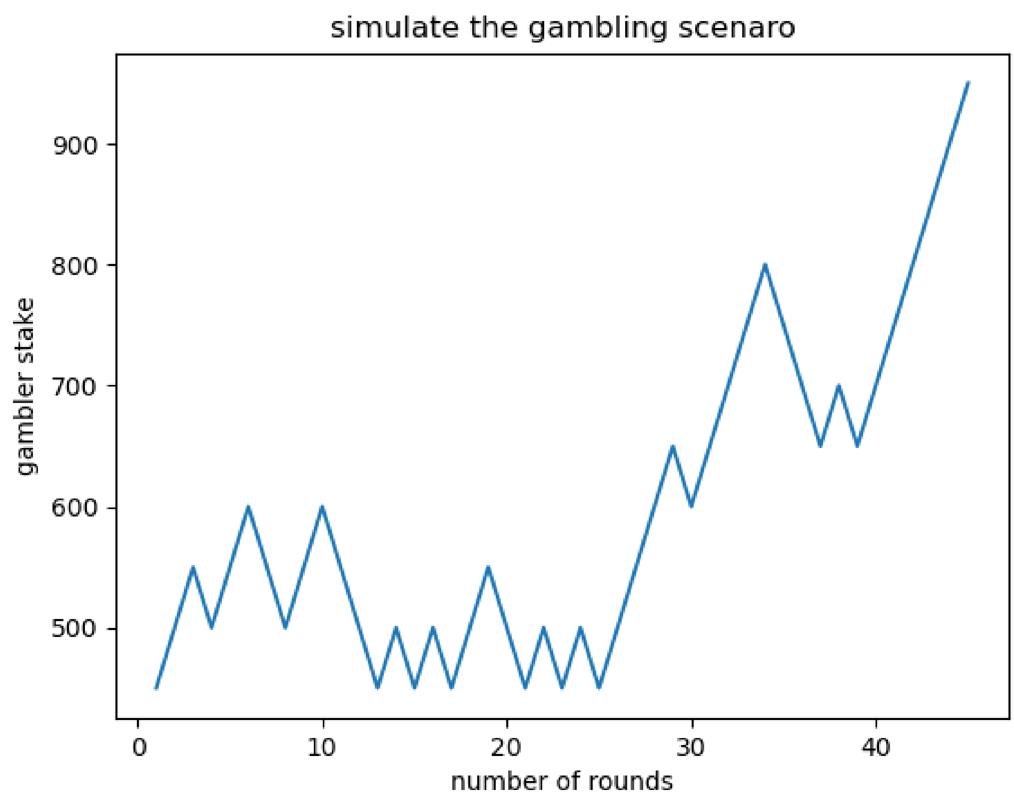

Figure 40: simulate scenario with initial stake = 500, target amount = 1000, prob = 0.5, bet = 50

As you can see in figure 40, approximately it takes 50 rounds that gambler went broke.

13.3 Probability of reaching target amount

For 10000 times we simulate for gambler's ruin in order to see how many times in 10000 times we reach the target amount and the probability of reach that state was:

$$P(reach\ target\ amount) = 0.505 \tag{75}$$

13.4 Mean value

If we add a duration that each iterate reaching target amount and append it to an array then we calculate the mean of that array the answer of this part will be:

$$E[reach\ target\ amount] = 100.78336633663366 \tag{76}$$

we calculate each mean for iterations and plot it in histogram which can be seen in figure 41

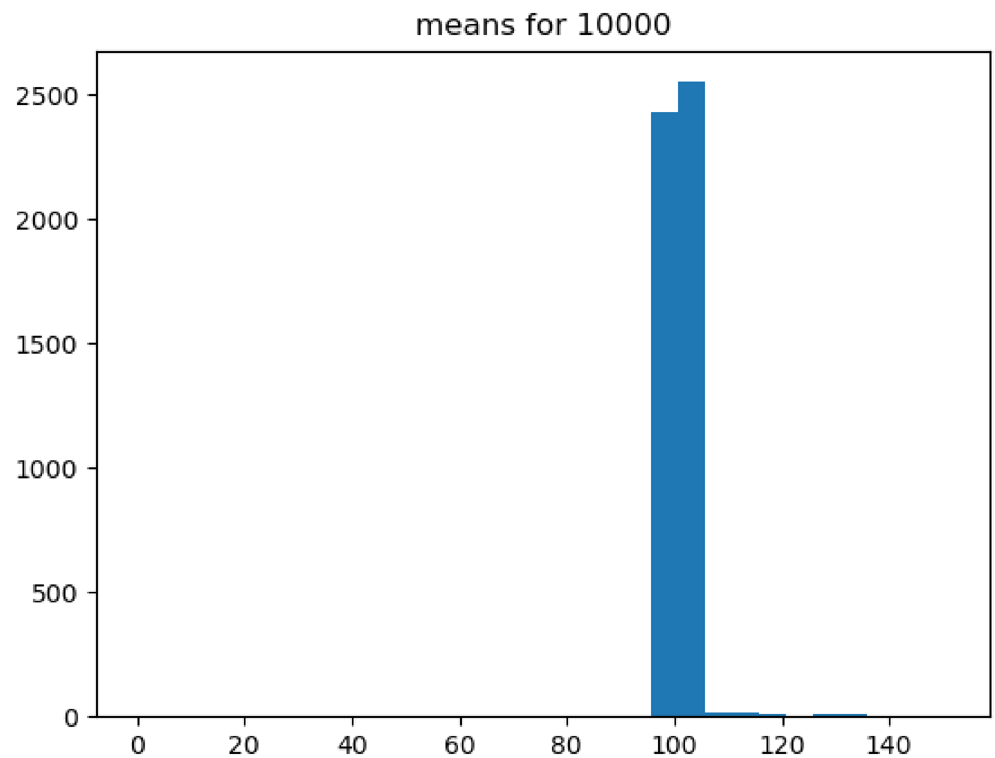

Figure 41: means initial stake = 500, target amount = 1000, prob = 0.5, bet = 50

And we are going to plot more about the means of our iterations in order to see the convergence of mean in figure ??.

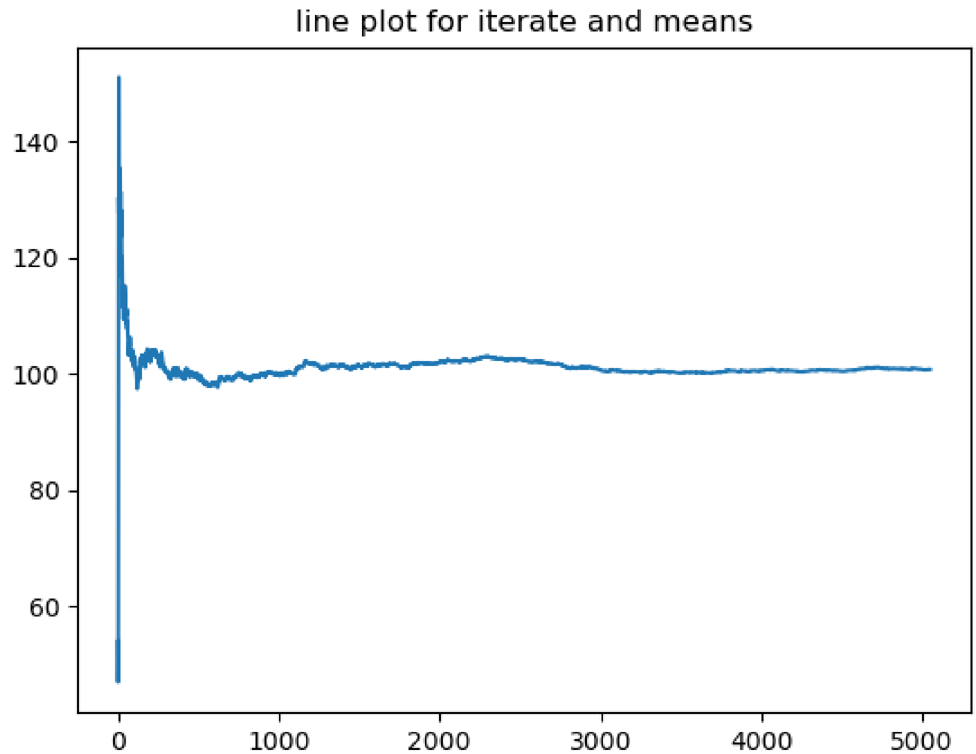

Figure 42: convergence of means stake = 500, target amount = 1000, prob = 0.5, bet = 50

13.5 Visualize another gambling scenario

Assume that we initialize gambling parameters as follows:

$$\{initial\ stake,\ bet,\ target\ amount,\ prob\} = \{1000,\ 50,\ 10000,\ 0.6\} \tag{77}$$

then the line plot showing gambler's stake over time in figure 43.

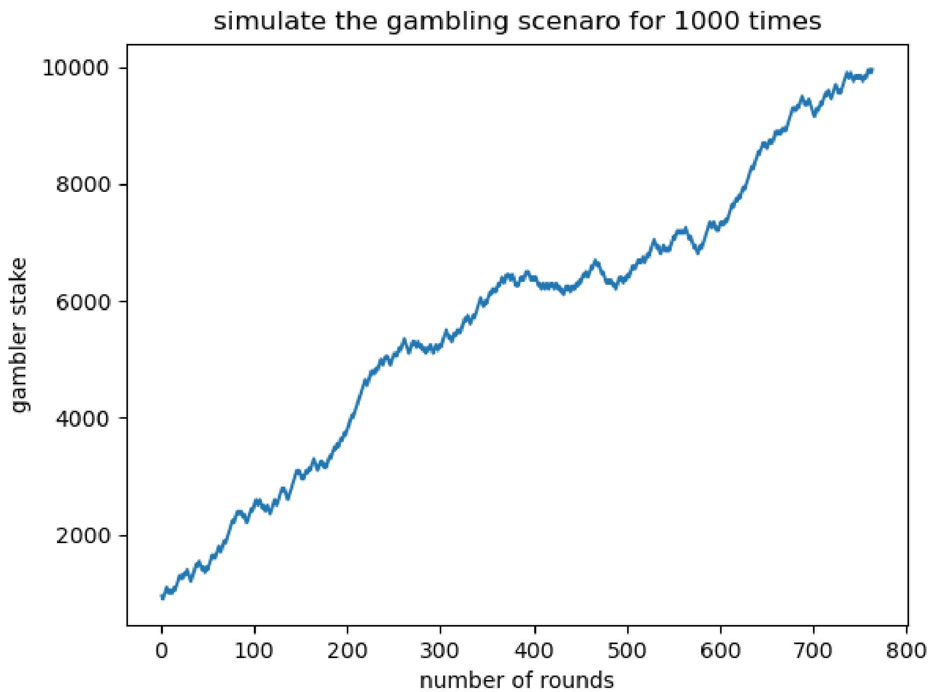

Figure 43: Gambler scenario for new parameters: initial stake = 1000, target amount = 10000, prob = 0.6, bet = 50

As you can see in figure43 after 800 rounds we getting our target amount.
Probability of reaching target amount:

$$P(reach\ target\ amount) = 0.9996 \tag{78}$$

mean of reaching target amount state:

$$E[reach\ target\ amount] = 899.0700280112045 \tag{79}$$

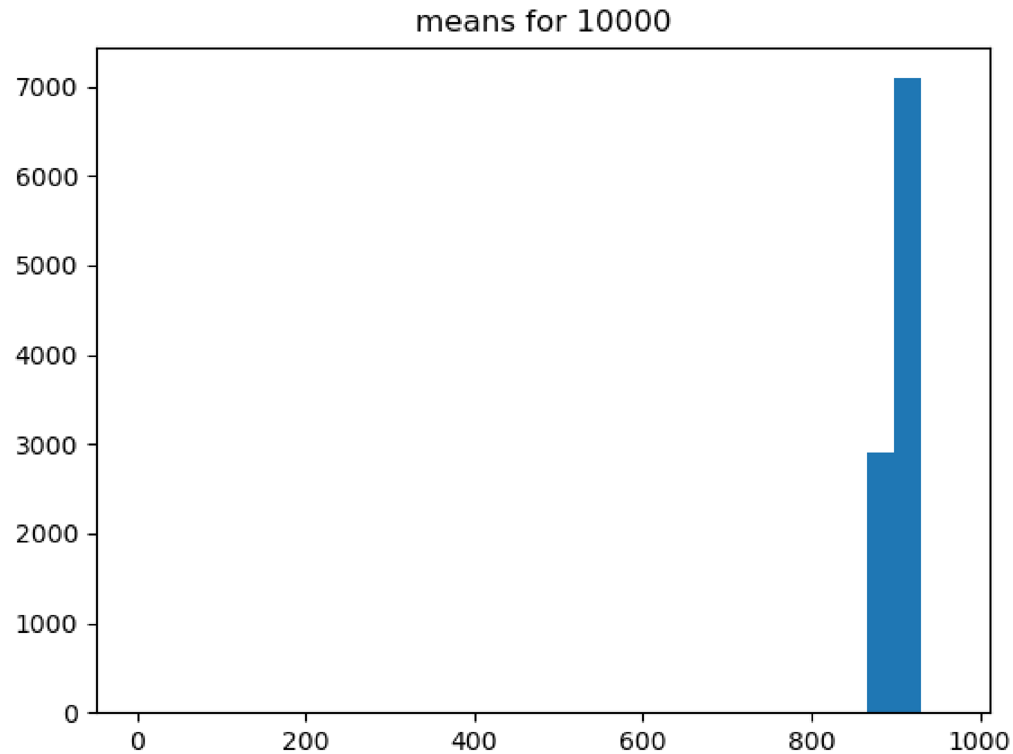

Figure 44: means with initial stake = 1000, target amount = 10000, prob = 0.6, bet = 50 parameters

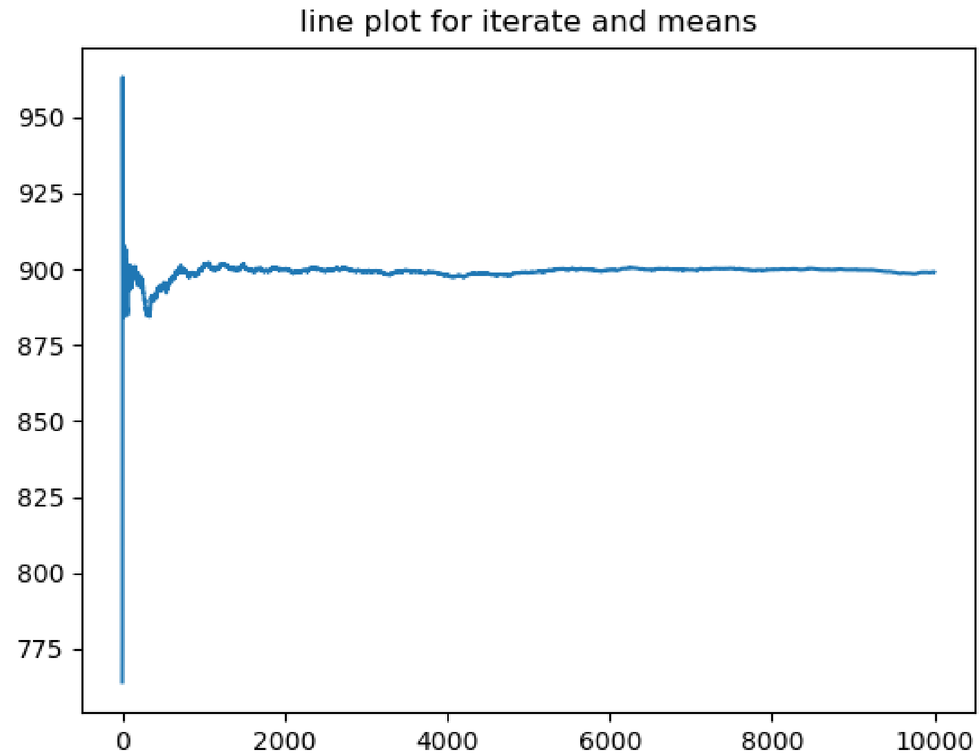

Figure 45: convergence of means with stake = 1000, target amount = 10000, prob = 0.6, bet = 50 parameters