SMAI Assignment 1

- Author: Vishal Reddy Mandadi
- Roll Number: 2019101119
- Date: 23rd September 2021

Question 1 solution:

Let us define PMF for a random variable x as

$$P(x) = \frac{1}{5} \; \forall x \in \{1, 2, 3, 4, 5\}$$

In the above function, we see that the range of x is finite and is given by $R = \{1, 2, 3, 4, 5\}$. Also, let's check if this PMF satisfies the required rules are not:

$$\sum_{x \in R} P(x) = P(1) + P(2) + P(3) + P(4) + P(5)$$
$$= 1/5 + 1/5 + 1/5 + 1/5 + 1/5 = 1$$

Thus, it satisfies the first rule, $\sum_{x \in R} P(x) = 1$. Also, you can see that $P(x) > 0 \;\; \forall x \in R$, this implies that the second rule is also satisfied, hence $P(x)$ is a valid PMF for finite range.

Now, let's show a valid PMF for infinite range. Let us define PMF of some random variable $X \in R$ where $R = \{1, 2, 3, ... \inf\}$ as:

$$P(x) = \frac{1}{2^x} \;\; \forall x \in R$$

The range of this variable is infinite. Now, let us check if $P(x)$ is a valid PMF or not.

$$\sum_{x \in R} P(x) = \sum_{x=1}^{\inf} P(x) = \sum_{x=1}^{\inf} \frac{1}{2^x} = \frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1 \; (Geometric \;\; series)$$

Thus, it satisfies the first rule, $\sum_{x \in R} P(x) = 1$. Also, you can see that $P(x) > 0 \;\; \forall x \in R$, this implies that the second rule is also satisfied, hence $P(x)$ is a valid PMF for infinite range. Thus, the required PMFs are:

1. For finite range

$$P(x) = \frac{1}{5} \; \forall x \in \{1, 2, 3, 4, 5\}$$

2. For infinite range

$$P(x) = \frac{1}{2^x} \;\; \forall x \in R$$

Question 2 solution

Let $U(a,b)$ be a uniform density function for a random variable x with range $R=[a,b]$, given by:

$$P(x) = U(a,b) = \{^{\frac{1}{b-a},\ if\ a \leqslant x \leqslant b}_{0,\ otherwise}$$

Now, mean of $U(a,b)$ is given by:

$$\mu(x,a,b) = E(x) = \int_a^b x * P(x)\, dx$$
$$E(x) = \int_a^b x * \frac{1}{b-a}\, dx$$
$$E(x) = \left[\frac{x^2}{2*(b-a)}\right]_a^b$$
$$E(x) = \frac{b^2-a^2}{2*(b-a)}$$
$$E(x) = \frac{b+a}{2}$$

Let us now calculate $E(x^2)$:

$$\mu(x^2,a,b) = E(x^2) = \int_a^b x^2 * P(x)\, dx$$
$$E(x^2) = \int_a^b x^2 * \frac{1}{b-a}\, dx$$
$$E(x^2) = \left[\frac{x^3}{3*(b-a)}\right]_a^b$$
$$E(x^2) = \frac{b^3-a^3}{3*(b-a)}$$
$$E(x^2) = \frac{a^2+ab+b^2}{3}$$

Now, the variance of x can be calculated by:

$$Var(x) = E(x^2) - E(x)^2$$
$$Var(x) = \frac{a^2+ab+b^2}{3} - \left(\frac{b+a}{2}\right)^2$$
$$Var(x) = \frac{a^2+ab+b^2}{3} - \frac{b^2+2ab+a^2}{4}$$
$$Var(x) = \frac{b^2-2ab+a^2}{12}$$
$$Var(x) = \frac{(b-a)^2}{12}$$

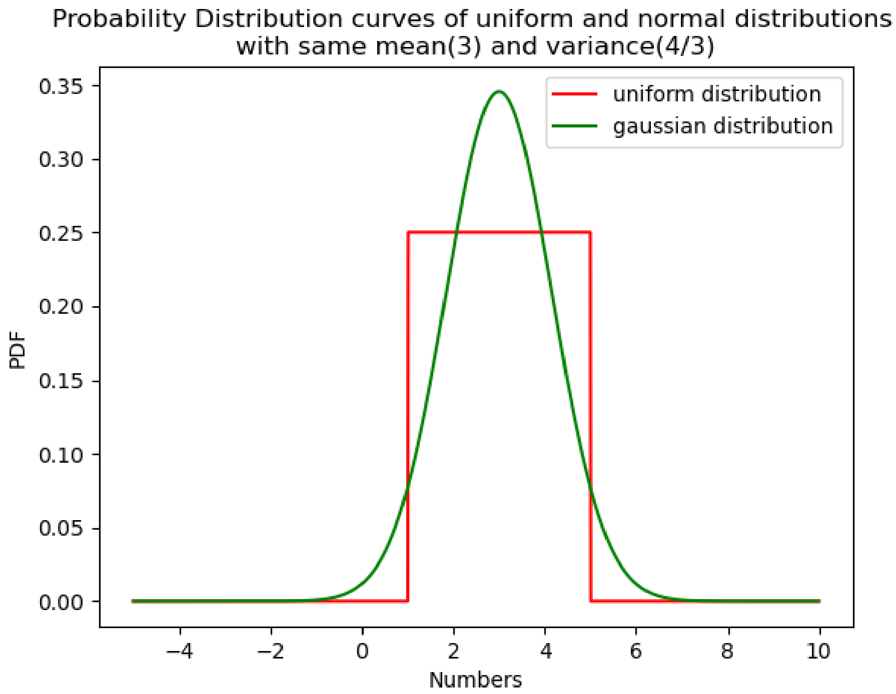
Hence, proved that $Var(x) = \frac{(b-a)^2}{12}$.

Question 3 solution (To be cont. (waiting for clarifications))

Let x be a random variable with uniform distribution $U(1,5)$, and let z be a random variable with normal distribution $N(\mu,\sigma^2)$ such that x and z have same mean and variance. This implies,

$$\mu = \frac{b+a}{2} = \frac{5+1}{2} = 3$$
$$\sigma^2 = \frac{(b-a)^2}{12} = \frac{(5-1)^2}{12} = \frac{(4)^2}{12} = \frac{4}{3}$$

This implies z has the distribution $N(3, \frac{4}{3})$. Now, let's plot the density functions of x and z to show that they have different density plots/functions despite having the same μ and σ^2. Following is the graph obtained on plotting these two probability distributions:



Please find the code used to generate the above graphs here:

```
# Created by Vishal Reddy Mandadi on 23-09-2021
# Q3 SMAI assignment 1
# Generating plots for normal and uniform distribution

import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm
import statistics

def normal_pdf(x_axis, mean, sd):
    return norm.pdf(x_axis, mean, sd)

def uniform_pdf(x=np.array([0], dtype=float), a=1, b=5):
    pdf = []
    for i in x:
        if i <= b and i >= a:
            pdf.append(1/(b-a))
        else:
            pdf.append(0)
    return np.array(pdf, dtype=float)

# Plot between -5 and 10 with .001 steps.
x_axis = np.arange(-5, 10, 0.01)
```

```
# Calculating mean and standard deviation
mean = 3 # statistics.mean(x_axis)
sd = np.sqrt(4/3) #statistics.stdev(x_axis)

y_uniform = uniform_pdf(x=x_axis, a=1, b=5)
y_gaussian = normal_pdf(x_axis, mean, sd)

fig, ax = plt.subplots()

ax.set(xlabel ='Numbers', ylabel ='PDF',
       title ='Probability Distribution curves of uniform and normal distributions \nwith same mean(3) and variance(4/3)')

ax.plot(x_axis, y_uniform, color='r', label='uniform distribution')
ax.plot(x_axis, y_gaussian, color='g', label='gaussian distribution')
plt.legend()

# plt.plot(x_axis, norm.pdf(x_axis, mean, sd))
plt.savefig(fname='pdfs_of_different_distributions.png')
plt.show()
```

Note: Since there were no restrictions mentioned on using library functions, I have used norm.pdf directly to generate normal pdf instead of using mathematical equations.

Question 4 solution

Let x be a discrete random variable with its probability mass function given by $P_X(x)$ defined on its discrete range R. Now, $E(x)$ and $E(x^2)$ can be defined as:

$$E(x) = \sum_{x \in R} x * P_X(x)$$
$$E(x^2) = \sum_{x^2 \in R} x * P_X(x)$$

The above definitions work on discrete random variables. Similarly, for discrete variables, $Var(x)$ or σ^2 is given by:

$$Var(x) = \sigma^2(x) = E((x - E(x))^2)$$
$$Var(x) = \sum_{x \in R} (x - E(x))^2 * P_X(x)$$
$$Var(x) = \sum_{x \in R} (x^2 + E(x)^2 - 2 * x * E(x)) * P_X(x)$$
$$Var(x) = \sum_{x \in R} x^2 * P_X(x) + \sum_{x \in R} E(x)^2 * P_X(x) - \sum_{x \in R} 2 * x * E(x) * P_X(x)$$
$$Var(x) = \sum_{x \in R} x^2 * P_X(x) + E(x)^2 * \sum_{x \in R} P_X(x) - 2 * E(x) * \left[\sum_{x \in R} x * P_X(x)\right]$$

Now here, we know that $E(x^2) = \sum_{x \in R} x * P_X(x)$, $\sum_{x \in R} P_X(x) = 1$ and $\sum_{x \in R} x * P_X(x) = E(x)$. Substituting these equations in the above equation, we get:

$$Var(x) = \sum_{x \in R} x^2 * P_X(x) + E(x)^2 * \sum_{x \in R} P_X(x) - 2 * E(x) * \left[\sum_{x \in R} x * P_X(x)\right]$$
$$Var(x) = E(x^2) + E(x)^2 - 2 * E(x) * (E(x))$$
$$Var(x) = E(x^2) + E(x)^2 - 2 * E(x)^2$$
$$Var(x) = E(x^2) - E(x)^2$$

Hence, proved that the equation $Var(x) = E(x^2) - E(x)^2$ is applicable to discrete random variables as well.

Question 5 solution

To prove:

1. $E_{x \sim N(\mu, \sigma^2)}(x) = \mu$
2. $Var_{x \sim N(\mu, \sigma^2)}(x) = \sigma^2$

Proof:

Let x be the normal random variable such that $x \sim N(\mu, \sigma^2)$ and range $R = (-\infty, \infty)$. Then PDF of x is given by:

$$P_X(x) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi} * \sigma} * e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

1. $E(x)$ is given by:

$$E(x) = \int_{x \in R} x * P_X(x) dx$$
$$E(x) = \int_{x \in R} (x - \mu + \mu) * P_X(x) dx$$
$$E(x) = \int_{x \in R} \mu * P_X(x) dx + \int_{x \in R} (x - \mu) * P_X(x) dx$$
$$E(x) = \mu * \int_{x \in R} P_X(x) dx + \int_{x \in R} (x - \mu) * P_X(x) dx$$
$$E(x) = \mu + \int_{x \in R} (x - \mu) * P_X(x) dx$$
$$E(x) = \mu + \int_{x \in R} (x - \mu) * \frac{1}{\sqrt{2\pi} * \sigma} * e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

   Now, substitute $(x - \mu) = t$ in the integral. On substitution, we get:

$$E(x) = \mu + \int_{t \in R} (t) * \frac{1}{\sqrt{2\pi} * \sigma} * e^{\frac{-t^2}{2\sigma^2}} dt$$

   Here, you can see that the function inside is an odd function passing through $t = 0$, $\implies$ the value of integral over the range of real numbers should be equal to zero i.e., $\int_{t \in R} (t) * \frac{1}{\sqrt{2\pi} * \sigma} * e^{\frac{-t^2}{2\sigma^2}} dt = 0$. This implies:

$$E(x) = \mu + \int_{t \in R} (t) * \frac{1}{\sqrt{2\pi} * \sigma} * e^{\frac{-t^2}{2\sigma^2}} dt$$
$$E(x) = \mu + 0$$
$$E(x) = \mu$$

   Hence, proved that mean of normal random variable $x \sim N(\mu, \sigma^2)$ is equal to μ.

2. $Var(x)$ is given by:

$$Var(x) = E((x - E(x))^2) = E((x - \mu)^2)$$
$$Var(x) = \int_{x \in R} (x - \mu)^2 * P_X(x) dx$$
$$Var(x) = \left[\int_{x \in R} (x - \mu)^2 * \frac{1}{\sqrt{2\pi} * \sigma} * e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \right]$$

Now, substitute $x - \mu = \sqrt{2} * \sigma * t$ in the integral $\implies dx = \sqrt{2}\sigma dt$. On substitution, we get:

$$Var(x) = \int_{t \in R} \frac{2\sigma^2}{\sqrt{\pi}} t^2 e^{-t^2} dt$$
$$Var(x) = \frac{2\sigma^2}{\sqrt{\pi}} \int_{t \in R} (t).(te^{-t^2}) dt$$
$$\text{Now apply integration by parts with } u = t \text{ and } v = t.e^{-t^2}$$
$$\text{Let us calculate integral of v first}$$
$$\int v = \int t.e^{-t^2} dt = \int \frac{-2t * e^{-t^2}}{(-2)} dt = \frac{-e^{-t^2}}{2} + c$$
$$\text{Using value of } \int v \text{ integral by parts we get:}$$
$$Var(x) = \frac{2\sigma^2}{\sqrt{\pi}} \left[\left[-\frac{te^{-t^2}}{2} \right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{-e^{-t^2}}{2} dt \right]$$
$$Var(x) = \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{e^{-t^2}}{2} dt$$

Now, let's find the value of integral separately using polar coordinates and then plug the value here. Let, $I = \int_{-\infty}^{\infty} \frac{e^{-t^2}}{2} dt$. Let us substitute $t = x$ so that it becomes $I = \int_{-\infty}^{\infty} \frac{e^{-x^2}}{2} dx$. Now, let us consider another equation $I_2 = \int_{-\infty}^{\infty} \frac{e^{-y^2}}{2} dy$. You can see that I and I_2 are equal as all we did was change the variables. On multiplying both the equations, we get:

$$I.I_2 = I.I = I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{e^{-x^2 - y^2}}{2} dx\, dy$$
$$\text{Now, let us convert the equation into polar coordinates}$$
$$\implies I^2 = \int_0^{\infty} \int_0^{2\pi} e^{-r^2} r d\theta\, dr = \int_0^{\infty} 2\pi e^{-r^2} r dr$$
$$I^2 = 2\pi \left[-\frac{e^{-r^2}}{2} \right]_0^{\infty} = 2\pi * \frac{1}{2} = \pi$$
$$\implies I = \sqrt{\pi}$$

Now, on plugging the value of I in $Var(x)$ equation, we get:

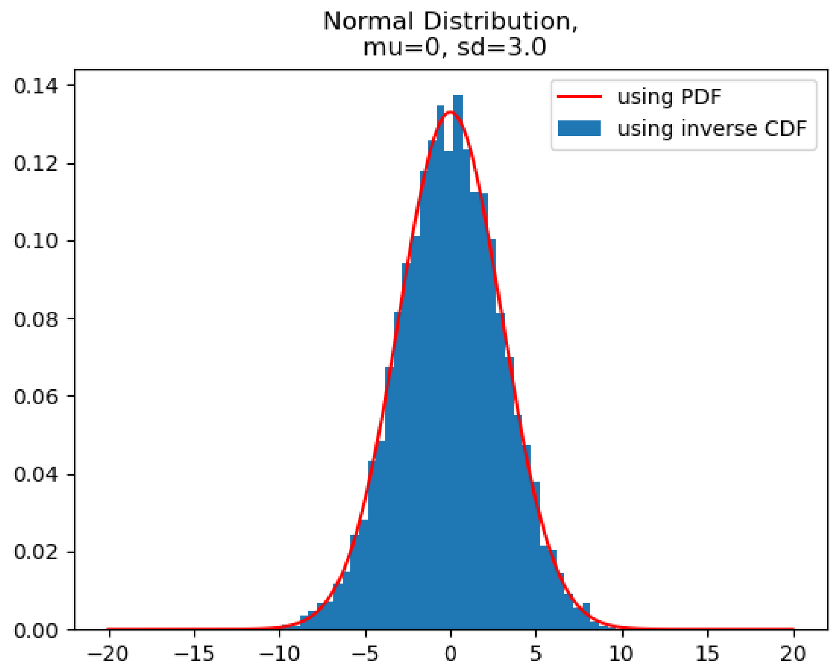
$$Var(x) = \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{e^{-t^2}}{2} dt = \frac{\sigma^2}{\sqrt{\pi}} I = \frac{\sigma^2}{\sqrt{\pi}} * \sqrt{\pi} = \sigma^2$$
$$Var(x) = \sigma^2$$

Hence, proved that $Var(x) = \sigma^2$. Therefore, the mean and the variance of normal distributions of form $N(\mu, \sigma^2)$ are μ and σ^2 respectively.

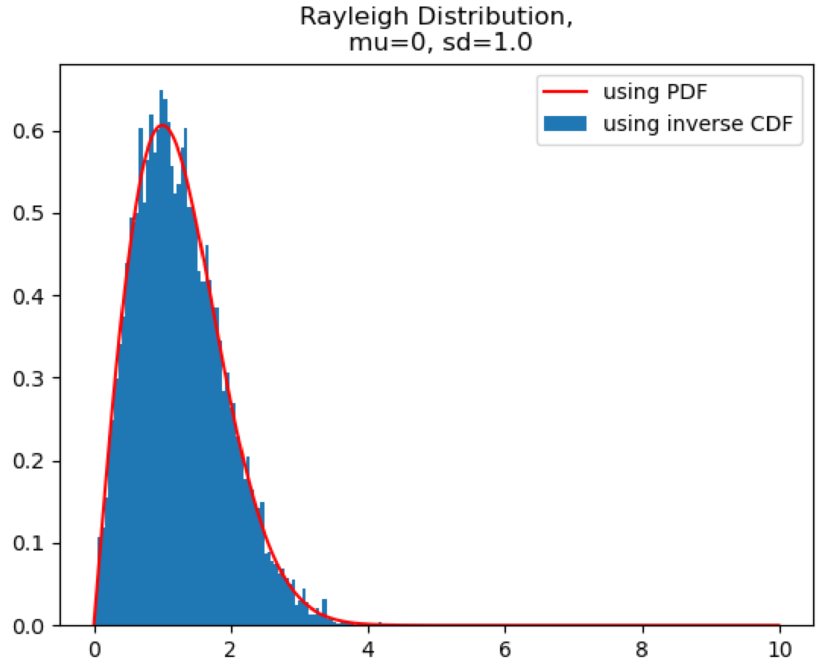
Question 6 solution

The following are the graphs generated using inverse CDF concept where a set of uniformly distributed random numbers are mapped to 3 distributions:

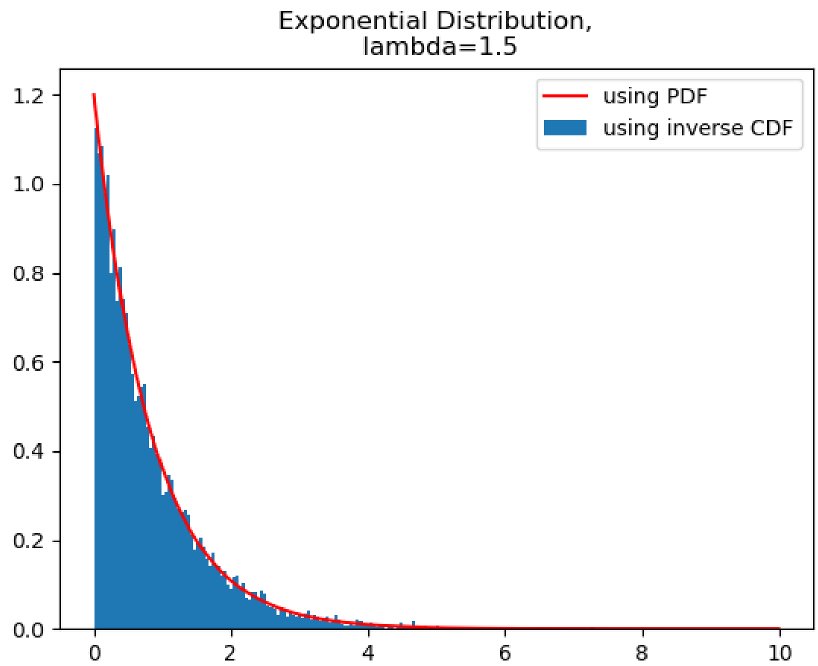
1. $U(0,1) \rightarrow N(\mu = 0, \sigma = 3)$



2. $U(0,1) \rightarrow Rayleigh(\sigma = 1.0)$



3. $U(0,1) \rightarrow Exponential(\lambda = 1.5)$



The code used to generate the above graphs is given below:

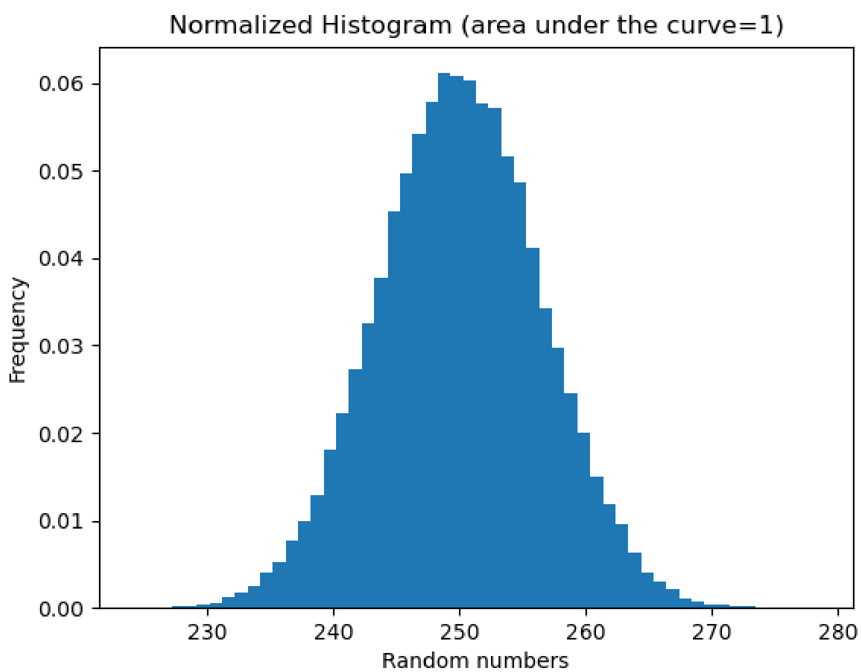
```
# Created by Vishal Reddy Mandadi on 23-09-2021
# Q6 SMAI Assignment 1

import numpy as np
import matplotlib.pyplot as plt
from numpy.core.fromnumeric import size
from scipy.stats import norm, rayleigh, expon
import statistics

def uniform_to_normal(x, mu=0.0, sd=3.0):
    # mu - mean, sd - standard_deviation
    return norm.ppf(x, loc=mu, scale=sd)

def uniform_to_rayleigh(x, mu=0.0, sd=1.0):
    return rayleigh.ppf(x, mu, sd)

def uniform_to_exponential(x, mu=2/3, sd=2/3):
    return expon.ppf(x, scale=sd)

if __name__=="__main__":
    # generate 10,000 rand nos
    rand_nos = np.random.uniform(low=0.0, high=1.0, size=(10000, 1))

    # Map then to Normal Distribution
    mean=0.0
    sd=3.0
    normal_nos = []
    normal_nos=uniform_to_normal(rand_nos, mu=mean, sd=sd)
    #print("Normal Nos: {}".format(normal_nos))
    bin_size=0.5
    no_of_bins = int((np.amax(normal_nos)-np.amin(normal_nos))/bin_size)
    fig, axes = plt.subplots(1, 1)
    axes.hist(normal_nos, bins=no_of_bins, density=True, label='using inverse CDF')
    axes.set_title("Normal Distribution,\n mu=0, sd=3.0")
    x_axis = np.arange(-20, 20, 0.01)
    axes.plot(x_axis, norm.pdf(x_axis, mean, sd), color='r', label='using PDF')
    axes.legend()
    plt.savefig("q6_normal.png")
    #plt.show()

    # Map them to Rayleigh Distribution
    mean=0.0
    sd=1.0
    rayleigh_nos = []
    rayleigh_nos=uniform_to_rayleigh(rand_nos, mu=mean, sd=sd)
    bin_size=0.05
    no_of_bins = int((np.amax(rayleigh_nos)-np.amin(rayleigh_nos))/bin_size)
    fig, axes = plt.subplots(1, 1)
    axes.hist(rayleigh_nos, bins=no_of_bins, density=True, label='using inverse CDF')
    axes.set_title("Rayleigh Distribution,\n mu=0, sd=1.0")
    x_axis = np.arange(-0, 10, 0.01)
    axes.plot(x_axis, rayleigh.pdf(x_axis, mean, sd), color='r', label='using PDF')
    axes.legend()
    plt.savefig("q6_rayleigh.png")
    #plt.show()

    # Map them to exponential distribution
    lmbda = 1.2 # 1.5
    mean = 1/lmbda
    sd = 1/lmbda
    exp_nos = []
    exp_nos=uniform_to_exponential(rand_nos, mu=mean, sd=sd)
    bin_size=0.045
    no_of_bins = int((np.amax(exp_nos)-np.amin(exp_nos))/bin_size)
    fig, axes = plt.subplots(1, 1)
    axes.hist(exp_nos, bins=no_of_bins, density=True, histtype='stepfilled', label='using inverse CDF')
    axes.set_title("Exponential Distribution,\n lambda=1.5")
    x_axis = np.arange(0, 10, 0.01)
    axes.plot(x_axis, expon.pdf(x_axis, scale=sd), color='r', label='using PDF')
```

```
    axes.legend()
    plt.savefig("q6_exp.png")
```

💡 **Inference:** From these graphs, we observe that the curves obtained via inverse CDF methods are very similar to the actual probability distribution functions. This helps us show that inverse CDFs methods work well in generating random numbers of different distributions with the help of uniform random distribution input. This method can sometimes be very complex to implement, in case distributions like Normal distribution ...etc., however, people were able to find intelligent approximations to bridge any such gaps (for example, R).

Question 7 solution

The following is the graph of the normalized histogram (normalized by area) which was asked in the question:



Note: Here, as asked in the question, we produced a normalized histogram i.e. Area under the curve = 1.

> *The **Central Limit Theorem** states that, In probability theory, the central limit theorem (CLT) establishes that, in many situations, when independent random variables are summed up, their properly normalized sum tends toward a normal distribution (informally a bell curve) even if the original variables themselves are not normally distributed. (Wikipedia)*

💡 **Inference:** The curve traced by the histogram resembles the shape of normal/Gaussian distribution. Our experiment, hence, provided experimental evidence of the Central Limit Theorem, proving its correctness, as the graphs generated from the sums of uniform random numbers match that of Gaussian, which is exactly what the Central Limit Theorem says.

The code which generates the above graph can be found below:

```
# Created by Vishal Reddy Mandadi on 23-09-2021
# Q7 of assignment1 SMAI
import numpy as np
import matplotlib.pyplot as plt

def gen_random_nos(n=500):
    random_nos=np.random.uniform(low=0.0, high=1.0, size=(500, 1))
    sum_of_nos=np.sum(random_nos, axis=0)
    return sum_of_nos

def plot_area_norm_hist(num_arr_np_ar, no_of_bins):
    fig, ax = plt.subplots()
    x, bins, p=ax.hist(num_arr_np_ar, bins=no_of_bins, density=True)# density=True is for normalization
    ax.set(xlabel ='Random numbers', ylabel ='Frequency',
       title ='Normalized Histogram (area under the curve=1)')
    plt.savefig(fname='q7_normalized_hist.png')

if __name__=="__main__":
    num_arr = []
    bin_size=1
    no_of_calls=50000
    for i in range(no_of_calls):
        num_arr.append(gen_random_nos())
        #print("Random nos sum: {}".format(gen_random_nos()))
    num_arr_np = np.array(num_arr, dtype=float)
    no_of_bins = int((np.amax(num_arr_np)-np.amin(num_arr_np))/bin_size)
    plot_area_norm_hist(num_arr_np, no_of_bins)
    plt.show()
```