SMAI Assignment 1

Anirudh Kaushik(2020111015)

Question 1

Give an example each of probability mass functions with finite and infinite ranges. Show that the conditions on PMF are satisfied by your example.

A1. (a) Infinite range: Consider the probability mass function following the infinite geometric progression with common ration $\frac{1}{2}$ and first term $\frac{1}{2}$

$$P(X = x_k) = \frac{1}{2^k} \tag{1}$$

As per the property of a PMF

$$\sum_{i=1}^{\infty} p_i = 1 \tag{2}$$

Thus

$$\sum_{i=1}^{\infty} p_i = \sum_{k=1}^{\infty} P(X = x_k) = \sum_{k=1}^{\infty} \frac{1}{2^k} \tag{3}$$

Now, sum of a geometric progression with first term a and common ration r is $\frac{a}{1-r}$, substituting $a = \frac{1}{2}$ and $r = \frac{1}{2}$ we get

$$\frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1 \tag{4}$$

Hence, the given function is a PMF with infinite range.

(b) Finite Range: Consider a fair coin being tossed twice. Let X be the the number of heads observed.The possible outcomes for the coin toss are $\{HH, TH, HT, TT\}$. The range of X is given by $R_x = 0, 1, 2$ corresponding to 0 heads (TT), 1 head (HT and TH) and 2 heads (HH). The corresponding PMF is given by:

$$\begin{aligned} P(0) &= \frac{1}{4} \\ P(1) = \frac{2}{4} &= \frac{1}{2} \\ P(2) &= \frac{1}{4} \end{aligned} \tag{5}$$

Thus the PMF is:

$$P(X = x_k) = \{\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\} \tag{6}$$

Verifying the condition of a PMF:

$$\sum_{k=1}^{n} p_i = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1 \tag{7}$$

Hence, the given function is a PMF with finite range.

Question 2

Show with complete steps that the variance of uniform density is given by equation 10. (Hint: use the expression for variance in equation 5.)

A2. Uniform density:

$$U(a,b) = \begin{cases} \frac{1}{(b-a)}, \ x \ \in \ (a,b) \\ 0, \ otherwise \end{cases}$$

We have

$$\sigma^2 = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 \tag{8}$$

Now, $\mathbb{E}[x] = \mu$, and hence $(\mathbb{E}[x])^2 = \mu^2 = (\frac{b+a}{2})^2$

$$\begin{aligned} \mathbb{E}[x^2] &= \int_a^b x^2 \frac{1}{(b-a)} dx \\ &= \frac{1}{(b-a)} \int_a^b x^2 dx \\ &= \frac{1}{(b-a)} \left[\frac{x^3}{3} \right]_a^b \\ &= \frac{1}{(b-a)} \left[\frac{b^3 - a^3}{3} \right] \\ &= \frac{1}{(b-a)} \left[\frac{(b-a)(b^2 + a^2 + ab)}{3} \right] \\ &= \frac{(b^2 + a^2 + ab)}{3} \end{aligned}$$

Substituting this and the value of μ in equation 8 we get:

$$\begin{aligned}
&\frac{(b^2+a^2+ab)}{3}-\left(\frac{b+a}{2}\right)^2\\
&=\frac{(4b^2+4a^2+4ab)}{3*4}-\frac{3(b^2+a^2+2ab)}{4*3}\\
&=\frac{(4b^2+4a^2+4ab-3b^2-3a^2-6ab)}{12}\\
&=\frac{(b^2+a^2-2ab)}{12}\\
&=\frac{(b-a)^2}{12}
\end{aligned}$$

Hence proved that $\sigma^2 = \frac{(b-a)^2}{12}$

Question 3

Show examples of two density functions (draw the function plots) that have the same mean and variance, but clearly different distributions. Plot both functions in the same graph with different colours.

A3. Consider Normal distribution with $\mu = 1$ and $\sigma^2 = 1$, i.e. $N(1, 1)$. Here, the mean of the distribution is $\mu = 1$ and the variance is $\sigma^2 = 1$.
Now consider exponential distribution with $\lambda = 1$. The mean in this case is $\frac{1}{\lambda} = 1$ and the variance is $\frac{1}{\lambda^2} = 1$.
Thus, both these distributions have mean and variance equal to one.
The plots for both these functions however are extremely different as shown.

Imports

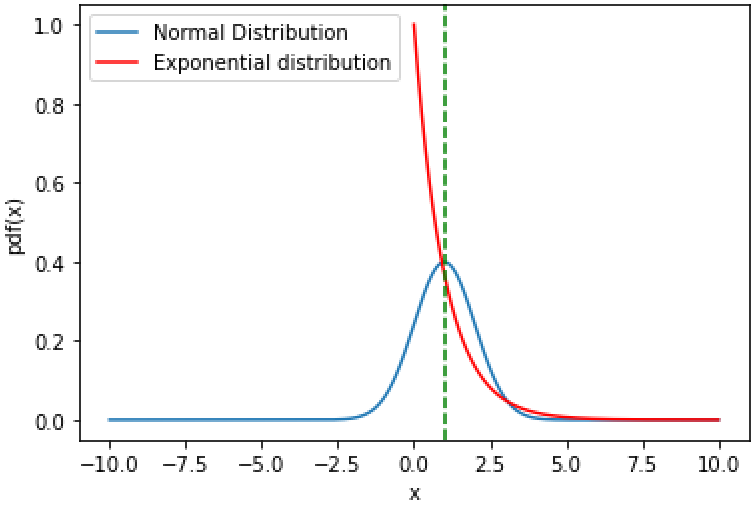
```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm
from scipy.stats import expon
%matplotlib inline
```

PDF of Normal distribution and exponential distribution

```
num_pts = 50000
norm_dist = norm.pdf(np.linspace(-10,10,num_pts), loc=1.0, scale=1.0)
expon_dist = expon.pdf(np.linspace(0,10,num_pts), scale=1.0)
```

Plot PDFs

```
plt.plot(np.linspace(-10,10,num_pts),norm_dist,label="Normal Distribution")
plt.plot(np.linspace(0,10,num_pts),expon_dist, color='r', label="Exponential distribution")
plt.axvline(x=1, color='g',Linestyle='--')
plt.xlabel("x")
plt.ylabel("pdf(x)")
plt.legend(loc="upper left")
plt.show()
```



Question 4

Show that the alternate expression for variance given in equation 5 holds for discrete random variables as well.

A4. By definition of σ for a discrete random variable X with range $\{x_1, x_2, ..., x_N\}$ we have,

$$\sigma^2 = \sum_{i=1}^{N}(x_i - \mu)^2 P(x_i)$$

$$= \sum_{i=1}^{N}(x_i^2 + \mu^2 - 2x_i\mu)P(x_i)$$

$$= \sum_{i=1}^{N}(x_i)^2 P(x_i) + \mu^2 - 2\sum_{i=1}^{N} x_i\mu P(x_i)$$

$$= \sum_{i=1}^{N}(x_i)^2 P(x_i) + \mu^2 - 2\mu\sum_{i=1}^{N} x_i P(x_i)$$

where PMF is the Probability Mass Function of X.
Now, $\mu = \mathbb{E}[x] = \sum_{i=1}^{N} x_i P(x_i)$. Substituting this in the above expression we get:

$$\sum_{i=1}^{N}(x_i)^2 P(x_i) + (\mathbb{E}[x])^2 - 2\mathbb{E}[x]\mathbb{E}[x]$$

$$= \mathbb{E}[x^2] - (\mathbb{E}[x])^2$$

Thus, this expression is the same as for the variance of a continuous random variable. Hence proved that $\sigma^2 = \mathbb{E}[x^2] - (\mathbb{E}[x])^2$

Question 5

Prove that the mean and variance of a normal density, $N(\mu, \sigma^2)$ are indeed its parameters, μ and σ^2 .

A5. Before proceeding, we assume the following:

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

Proof for Mean:
Now $N(\mu, \sigma) = \frac{1}{\sqrt{2\pi}} e^{-(\frac{x-\mu}{\sqrt{2}\sigma})^2}$ and $\mu = \int_{-\infty}^{\infty} xp(x)dx$

substituting pdf into the above equation

$$\mu = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} xe^{-(\frac{x-\mu}{\sqrt{2}\sigma})^2}dx$$
$$let\ y = x - \mu$$
$$dy = dx$$
$$x = y + \mu$$

substituting y in the integral:

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}(y+\mu)e^{-(\frac{y}{\sqrt{2}\sigma})^2}dy$$
$$= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} ye^{-(\frac{y}{\sqrt{2}\sigma})^2}dy + \frac{1}{\sqrt{2\pi}\sigma}\mu\int_{-\infty}^{\infty} e^{-(\frac{y}{\sqrt{2}\sigma})^2}dy - (a)$$

Now, Let I $= \frac{1}{\sqrt{2\pi}\sigma}\mu\int_{-\infty}^{\infty} e^{-(\frac{y}{\sqrt{2}\sigma})^2}dy = \frac{1}{\sqrt{2\pi}\sigma}\mu\int_{-\infty}^{\infty} e^{-(\frac{x}{\sqrt{2}\sigma})^2}dx$

Thus, $I^2 = (\frac{1}{\sqrt{2\pi}\sigma}\mu)^2\int_{-\infty}^{\infty} e^{-(\frac{y}{\sqrt{2}\sigma})^2}dy * \int_{-\infty}^{\infty} e^{-(\frac{x}{\sqrt{2}\sigma})^2}dx$

$$= (\frac{1}{\sqrt{2\pi}\sigma}\mu)^2\int_{-\infty}^{\infty} e^{-((\frac{y}{\sqrt{2}\sigma})^2+(\frac{x}{\sqrt{2}\sigma})^2)}dydx$$

Let $x^{'} = \frac{x}{\sqrt{2}\sigma}$ $and\ y^{'} = \frac{y}{\sqrt{2}\sigma}$

substituting $r^2 = x^{'2} + y^{'2}$and using substituting the Jacobian correction

$$= (\frac{1}{\sqrt{2\pi}\sigma}\mu)^2\int_{-\infty}^{\infty} e^{-r^2}rdrd\theta$$

The above is simply the integral of the function and the term inside the integral will evaluate to $2\pi\sigma^2$

$$= (\frac{1}{\sqrt{2\pi}\sigma}\mu)^2 * \pi * 2\sigma^2 = \mu^2$$
$$\implies I^2 = \mu^2$$
$$\implies I = \mu$$

Substituing the value of I in (a)

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} ye^{-(\frac{y}{\sqrt{2}\sigma})^2}dy + \mu$$

$$
\begin{aligned}
&let\ t^2 = \frac{y^2}{2\sigma^2}\\
&\quad t = \frac{y}{\sqrt{2}\sigma}\\
&\quad dy = \sqrt{2}\sigma dt\\
\implies &\frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} te^{-t^2}dt + \mu\\
&= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{0} te^{-t^2}dt + \frac{1}{\sqrt{\pi}}\int_{0}^{\infty} te^{-t^2}dt + \mu\\
&\quad \text{Substitute } \lambda = t^2 \implies d\lambda = 2tdt\\
&= \frac{1}{2\sqrt{\pi}}\int_{-\infty}^{0} e^{-\lambda}dt + \frac{1}{2\sqrt{\pi}}\int_{0}^{\infty} e^{-\lambda}dt + \mu\\
&= \frac{1}{2\sqrt{\pi}}\int_{0}^{\infty} e^{-\lambda}dt - \frac{1}{2\sqrt{\pi}}\int_{0}^{\infty} e^{-\lambda}dt + \mu\\
&= frac12\sqrt{\pi}(\frac{1}{2} - \frac{1}{2}) = 0 + \mu = \mu
\end{aligned}
$$

Hence Proved that the mean comes out to be μ

Proof for Variance:

Variance $= \sigma^2 = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 = \mathbb{E}([x-\mu])^2$

$$
\begin{aligned}
\implies &\sigma^2 = \int_{-\infty}^{\infty}(x-\mu)^2 e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx\\
&\quad substitute\ y = x - \mu \implies dy = dx\\
&= \int_{-\infty}^{\infty}(y)^2 e^{\frac{-(y)^2}{2\sigma^2}}dy
\end{aligned}
$$

Using Integration By Parts

$$
\begin{aligned}
&\int udv = uv - \int vdu\\
&substituting\ u = y \implies du = dy\\
&and\ dv = ye^{\frac{-y^2}{2\sigma^2}}dy \implies v = -\sigma^2 e^{\frac{-y^2}{2\sigma^2}}\\
&= \frac{1}{\sqrt{2\pi}\sigma}\left[-\sigma^2 ye^{\frac{-y^2}{2\sigma^2}}\right] + \sigma^2\int_{-\infty}^{\infty} e^{\frac{-y^2}{2\sigma^2}}dy\\
&= 0 + \sigma^2 * 1\\
\implies &variance(x) = \sigma^2
\end{aligned}
$$

Hence Proved that variance of Normal distribution $= \sigma^2$ and standard deviation $= \sqrt{\sigma^2} = \sigma$

Question 6

Using the inverse of CDFs, map a set of 10,000 random numbers from $U[0,1]$ to follow the following pdfs:

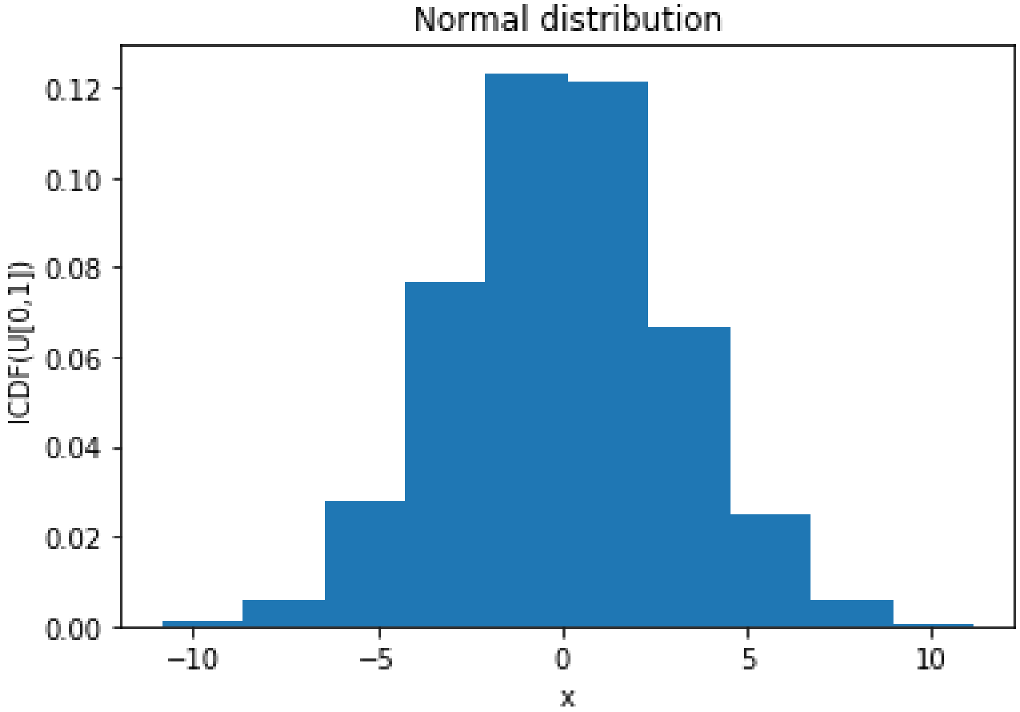
(A) Normal density with $\mu = 0$, $sigma = 3.0$.

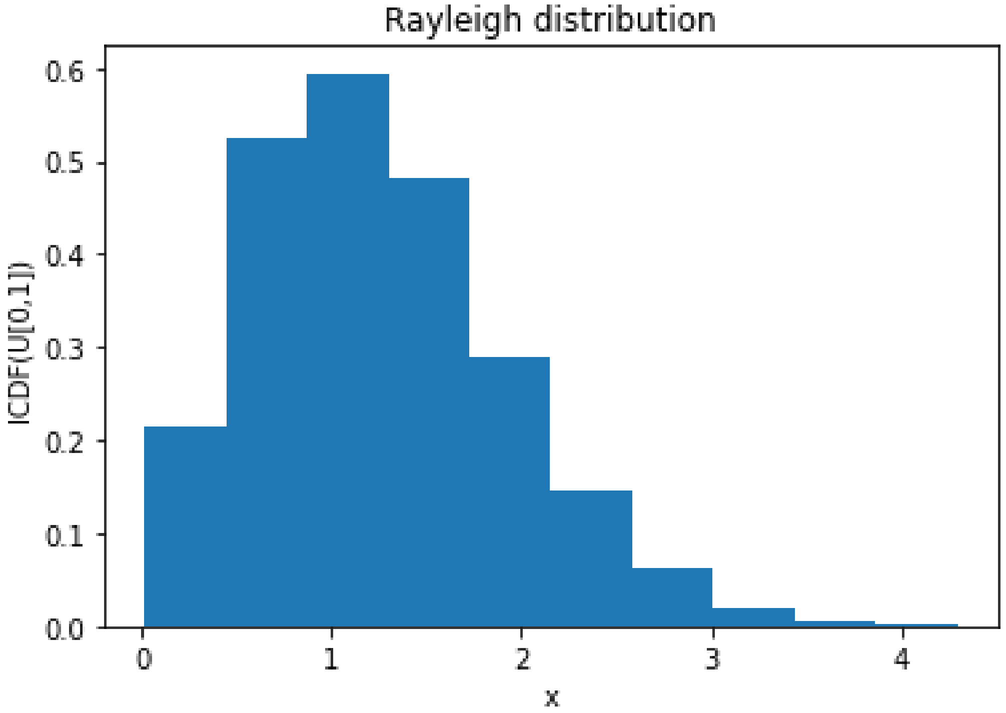
(B) Rayleigh density with $\sigma = 1.0$.

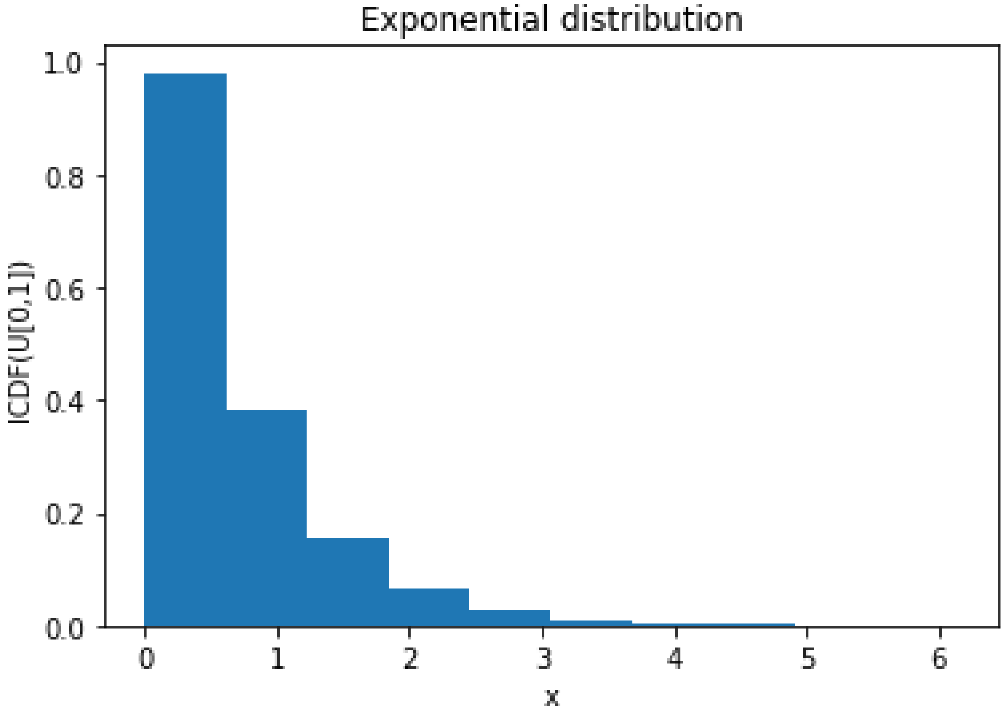
(C) Exponential density with $\lambda = 1.5$.

Once the numbers are generated, plot the normalized histograms (the values in the bins should add up to 1) of the new random numbers with appropriate bin sizes in each case; along with their pdfs. What do you infer from the plots? Note: see $rand()$ function in C for $U[0, INTMAX]$.

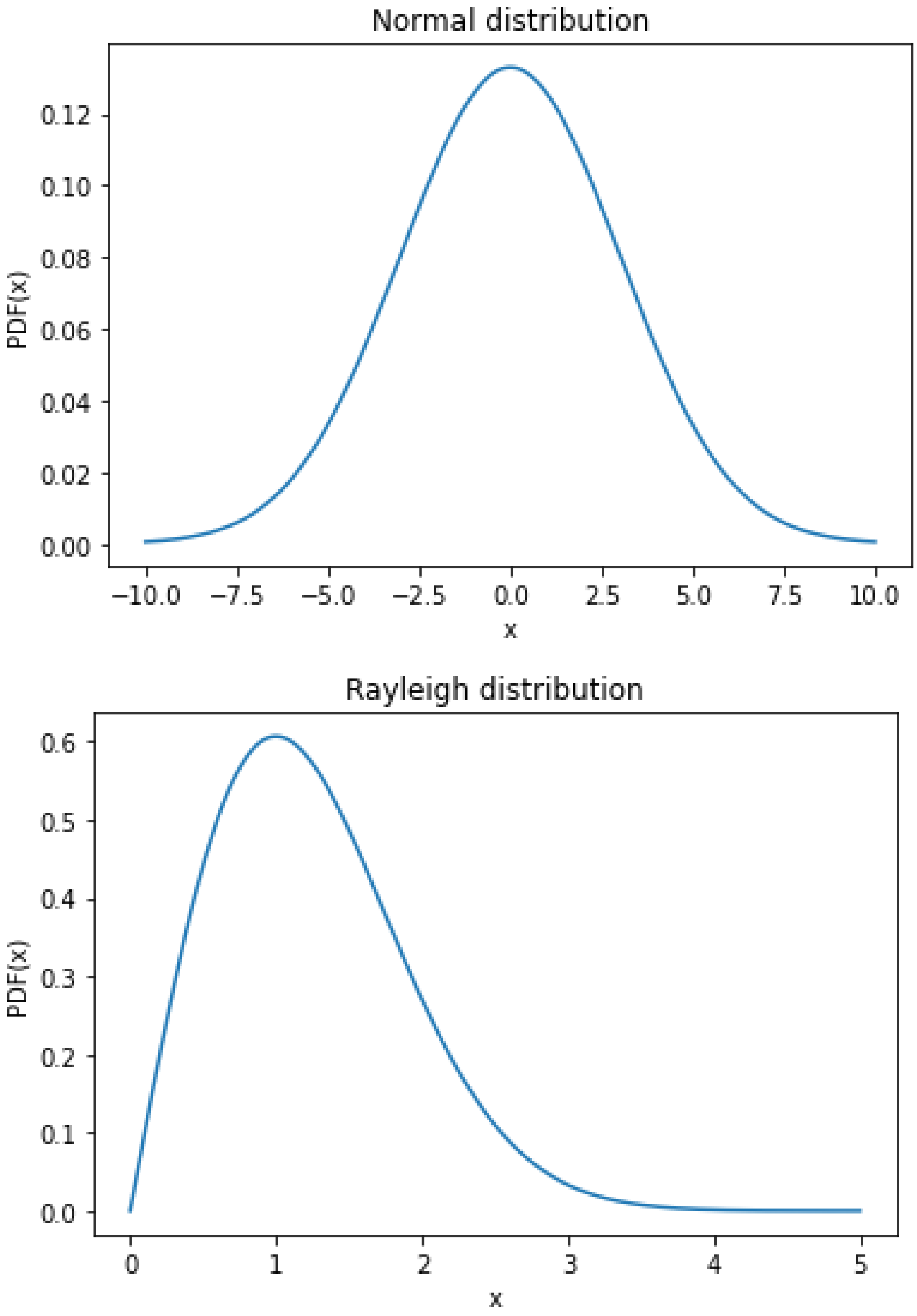
A6. The following are the plots obtained by using the inverse cdf method:





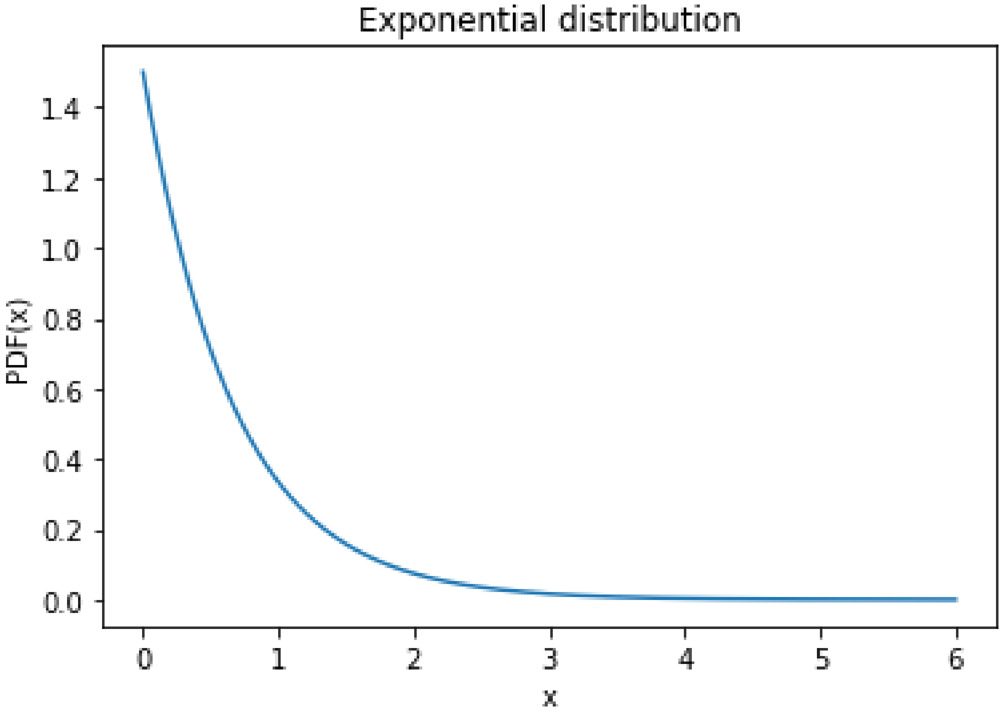


The following are the plots of the pdfs of the 3 distributions with corresponding mean and variance plotted over the same range of values:



Normal distribution

Rayleigh distribution


Exponential distribution
PDF(x)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
x

Imports

```
from scipy.stats import norm
from scipy.stats import rayleigh
from scipy.stats import expon
import random
import scipy
import numpy as np
import matplotlib.pyplot as plt
from prettytable import PrettyTable
%matplotlib inline
```

Generate random numbers from uniform distribution

```
N = 10000
rand_val = np.random.uniform(low = 0.0, high = 1.0, size = N)
```

Inverse CDF functions

```
def norm_inv_cdf(variance, mean, values):
  # erf^-1(2y-1)*root 2 * sigma + mu
  return (scipy.special.erfinv(2*values - 1)*variance*(np.sqrt(2))) + mean

def rayleigh_inv_cdf(sigma, values):
  # root 2/sigma * root(-ln(1-y))
  return np.sqrt(2)*sigma*np.sqrt(-np.log(1-values))

def expon_inv_cdf(lamda, values):
  # -1/lambda * ln(1-y)
  return -1/lamda * np.log(1-values)
```

PDF functions

```
def norm_pdf(variance, mean, values):
  return(1/(variance*np.sqrt(2*np.pi)))*np.exp((-0.5)*(((values-mean)/variance)**2))
def rayleigh_pdf(sigma, values):
  return (values/(sigma**2) * np.exp(-(values**2)/(2*(sigma**2))))
def expon_pdf(lamda, values):
  return lamda*np.exp(-(lamda*values))
```

Generate Distributions

```
# Normal Distribution
sigma = 3.0
mu = 0.0
norm_dist = norm_inv_cdf(sigma,mu, rand_val)

# Rayleigh Distribution
sigma = 1.0
rayleigh_dist = rayleigh_inv_cdf(sigma, rand_val)


# Exponential Distribution
lamda = 1.5
expon_dist = expon_inv_cdf(lamda,rand_val)
```

Plot Normalized Histograms

```
x, bins, p = plt.hist(norm_dist, density=True)
plt.show()
x, bins, p = plt.hist(rayleigh_dist, density=True)
plt.show()
plt.show()
x, bins, p = plt.hist(expon_dist, density=True)
plt.show()
```

Generate PDFs

```
norm_pdf_dist = norm_pdf(3.0, 0.0, np.linspace(-10,10,N))
rayleigh_pdf_dist = rayleigh_pdf(1.0, np.linspace(0,5,N))
expon_pdf_dist = expon_pdf(1.5, np.linspace(0,6,N))
```

Plot PDFs

```
plt.plot(np.linspace(-10,10,N),norm_pdf_dist)
plt.show()
plt.plot(np.linspace(0,5,N),rayleigh_pdf_dist)
plt.show()
plt.plot(np.linspace(0,6,N),expon_pdf_dist)
plt.show()
```

```
l = [["Normal",np.mean(norm_dist)],["Rayleigh", np.mean(rayleigh_dist)],["Exponential", np.me
table = PrettyTable(['Distribution', 'Mean'])

for rec in l:
    table.add_row(rec)

print(table)
```

```
+--------------+----------------------+
| Distribution |         Mean         |
+--------------+----------------------+
|    Normal    | 0.026529182842217842 |
|   Rayleigh   |  1.259900431024384   |
| Exponential  |  0.6743978237253746  |
+--------------+----------------------+
```

Observations

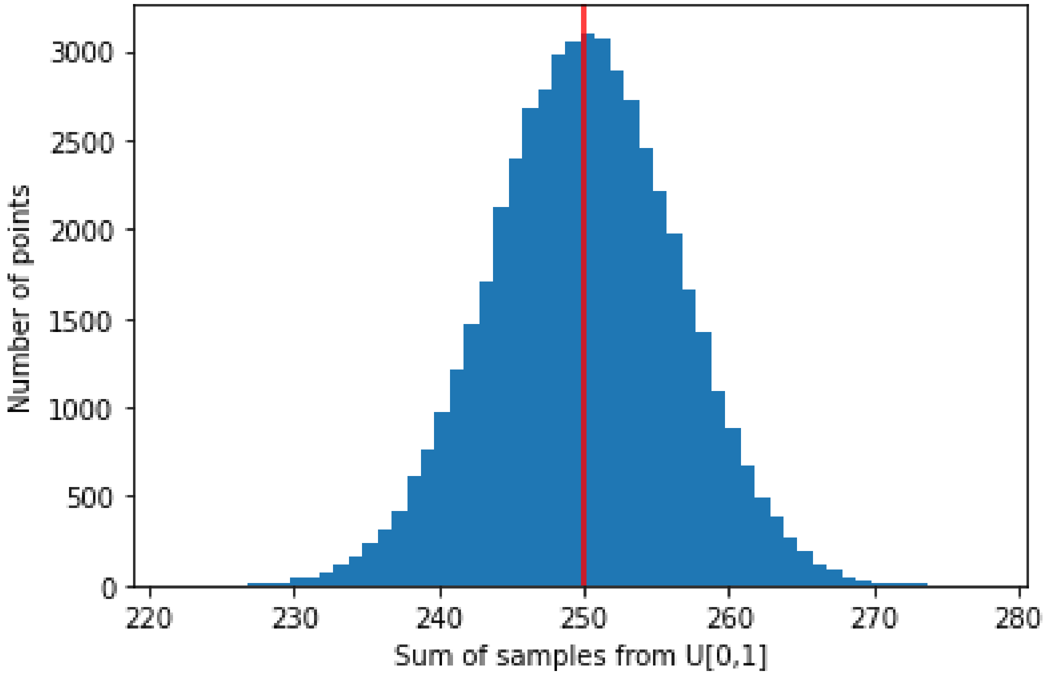
- Upon sampling 10000 random numbers from uniform distribution and using them in the inverse cdf functions of various distributions, we obtain distribution curves resembling the pdf of the corresponding distributions.
- In other words, by using random numbers from uniform distribution $U[0,1]$ and puggling these values into the inverse cdf of a given distribution (Normal, rayleigh and exponential), we obtain random numbers from the corresponding distributions.
- Ex: U[0,1] -> inv_cdf_normal(U[0,1]) -> number from Normal distribution

✓ 0s completed at 8:50 PM

Question 7

Write a function to generate a random number as follows: Every time the function is called, it generates 500 new random numbers from $U[0,1]$ and outputs their sum. Generate 50, 000 random numbers by repeatedly calling the above function, and plot their normalized histogram (with bin-size = 1). What do you find about the shape of the resulting histogram?

A7. The histogram obtained with bin size 1 is as follows:



Imports

```
import numpy as np
import random
import matplotlib.pyplot as plt
from prettytable import PrettyTable
```

Random number generator

```
def gen_rand():
  # Gen random number
  # return np.sum(np.random.unform(size=500))# vectorized implementation
  x = 0
  for i in range(0,500):
    x += random.uniform(0,1)
  return x

def gen_randn(N):
  # generate N random numbers and return a list
  values = []
  for i in range(N):
    values.append(gen_rand())
  values = np.array(values)
  return values
```

Plot Histogram

```
values = gen_randn(50000)
mean = np.mean(values)
e_x2 = np.mean(values**2)#E[x^2]
variance = e_x2 - mean**2
x, bins, p = plt.hist(values,bins=np.arange(min(values), max(values) + 1, 1))
plt.axvline(x=mean, color='r')
plt.show()
```

Tabulate Results

```
l = [[mean, 1, 50000, variance]]
table = PrettyTable(['Mean', 'bin-size', 'Number of points', 'Variance^2'])

for rec in l:
    table.add_row(rec)
```

```
print(table)
```

```
+--------------------+----------+------------------+--------------------+
|        Mean        | bin-size | Number of points |     Variance^2     |
+--------------------+----------+------------------+--------------------+
| 249.99339593340133 |    1     |      50000       | 41.58612367406022 |
+--------------------+----------+------------------+--------------------+
```

Observations

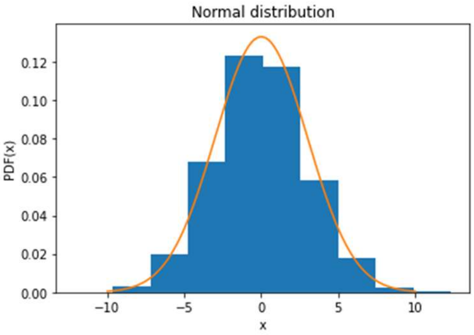
- We observe that the graph follows a bell-shaped symmetrical curve centered around the mean (250, since the range of the values is from 0 to 500 i.e 500*[0,1]).
- This is characteristic of the normal distribution.
- Thus, if we add a large number of samples from uniform distribution and do this to generate multiple samples we will obtain a good approximation of the normal distribution known as the Irwin-hall distribution.
- This also follows from central limit theorem.
- The pdf for the function obtained using this method is as follows $\frac{1}{(n-1)!}\sum_{k=0}^{n}(-1)^k\binom{n}{k}(x-k)^n sgn(x-k)$
- Here, sgn() is the sign function.

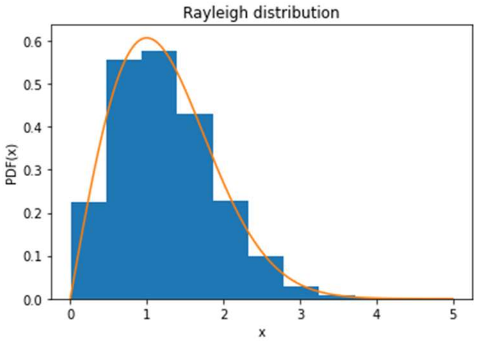
Source for formula of pdf: https://en.wikipedia.org/wiki/Irwin%E2%80%93Hall_distribution

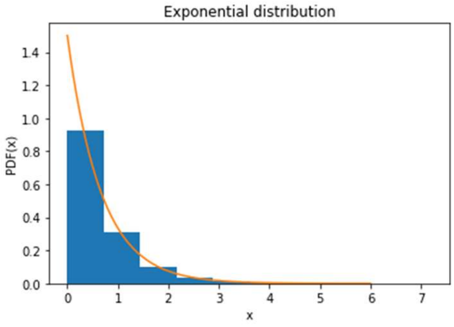
✓ 0s completed at 8:56 PM

Additional Observations

Q6) The pdf of the distributions obtained by using the inverse cdf of the corresponding distribution on random samples from U[0,1] given an approximate distribution. The pdf becomes more accurate i.e. error in the samples reduce as we increase the number of samples. For 10,000 samples we get a very good approximation of the three distributions, Normal, Rayleigh and Exponential.







Here, in the case of exponential distribution, the max value has been scaled down a bit by the hist function of matplotlib.pyploy with density=True to normalise it. However, the distribution is still followed perfectly and the result is the same since the same formula was used for inverse cdf.

Q7) Note that the Normal distribution obtained is approximate and is known as Irwin-Hall distribution. We can get better results by taking a higher number of samples of the sum of uniform distribution samples. Increasing the sum (i.e summing more numbers, let's say 1000 instead of 500) shifts the mean and also provides a more accurate distribution. We can approximate a normal distribution by drawing 12 samples from the uniform distribution U[0,1].