Assignment 1 – Random Variables

Laksh Nanwani – MS CSE - 2021701002 – laksh.nanwani@research.iiit.ac.in

---

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.

→ Probability Mass function = PMF = $P(v)$

The 2 conditions of PMF are

$$P(v_i) \geqslant 0 \text{ for all } i \quad - \text{ (1)}$$

$$\sum_{i=1}^{n} P(v_i) = 1 \quad - \text{ (2)}$$

For a finite range PMF, we can simply consider a coin toss or a dice roll. Considering a dice roll,

A roll of dice has 6 discrete outcomes $\{1,2,3,4,5,6\}$ each with an equal probability of $1/6$ ∴

we can say that $P(v) = 1/6 \;,\; v \in \{1,2,3,4,5,6\}$

As can be seen, $P(v) = 1/6 \geqslant 0$ → satisfies (1)

And $\sum_{i=1}^{6} P(v) = \sum_{i=1}^{6} 1/6 = 6 \times \frac{1}{6} = 1$ → satisfies (2)

∴ Roll of a dice has a PMF for its outcome and a finite range.

For an infinite range PMF, consider a function,

$$P(v) = \frac{1}{2^v}, \quad \text{where} \quad v \in N, \quad N = \text{natural numbers}$$

$\because N = \{1, 2, \ldots\} \Rightarrow N$ is discrete in nature with infinite range.

Now,

$$P(v) = \frac{1}{2^v} \Rightarrow \lim_{v \to \infty} P(v) = 0 \Rightarrow P(v) \geq 0 \rightarrow \text{satisfies } ①$$

$$\sum_{1}^{\infty} P(v) = \frac{1}{2} + \frac{1}{2^2} + \ldots \infty \quad \leftarrow \text{this is an infinite GP with } a = 1/2, \; r = 1/2$$

$$\therefore \sum P(v) = \frac{1/2}{1 - 1/2} = 1/2 \,/\, 1/2$$

$\because |r| \leq 1 \Rightarrow \sum = \frac{a}{1-r}$

$$\therefore \sum^{N} P(v) = 1 \rightarrow \text{satisfies } ②$$

$\therefore$ An event whose outcome is v with probability $1/2^v$, $v \in N$ is an example of PMF with infinite range

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.)

Equation 10 $\rightarrow$ $\sigma^2 = \dfrac{(b-a)^2}{12}$

Equation 5 $\rightarrow$ $\sigma^2 = E(x^2) - (E(x))^2$

$\Rightarrow$ Given, Uniform density function

$$U(a,b) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

$$E(x^2) = \int_a^b x^2 p(x)\,dx = \int_a^b x^2\, U(a,b)\,dx = \int_a^b x^2 \cdot \frac{1}{b-a}\,dx$$

$$= \frac{1}{b-a}\int_a^b x^2\,dx = \frac{1}{b-a}\,\frac{x^3}{3}\bigg|_a^b = \frac{1}{3(b-a)}\,(b^3-a^3)$$

$$\therefore E(x^2) = \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3} \Rightarrow \boxed{E(x^2) = \frac{b^2+ab+a^2}{3}}$$

$$E(x) = \int_a^b x\,U(a,b)\,dx = \int_a^b \frac{x}{b-a}\,dx = \frac{1}{b-a}\int_a^b x\,dx$$

$$= \frac{1}{2(b-a)}\,x^2\bigg|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2} \Rightarrow \boxed{E(x) = \frac{b+a}{2}}$$

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

$$\therefore \sigma^2 = \frac{4(b^2+ab+a^2) - 3(b^2+2ab+a^2)}{12} = \frac{4b^2+4ba+4a^2-3b^2-6ab-3a^2}{12}$$

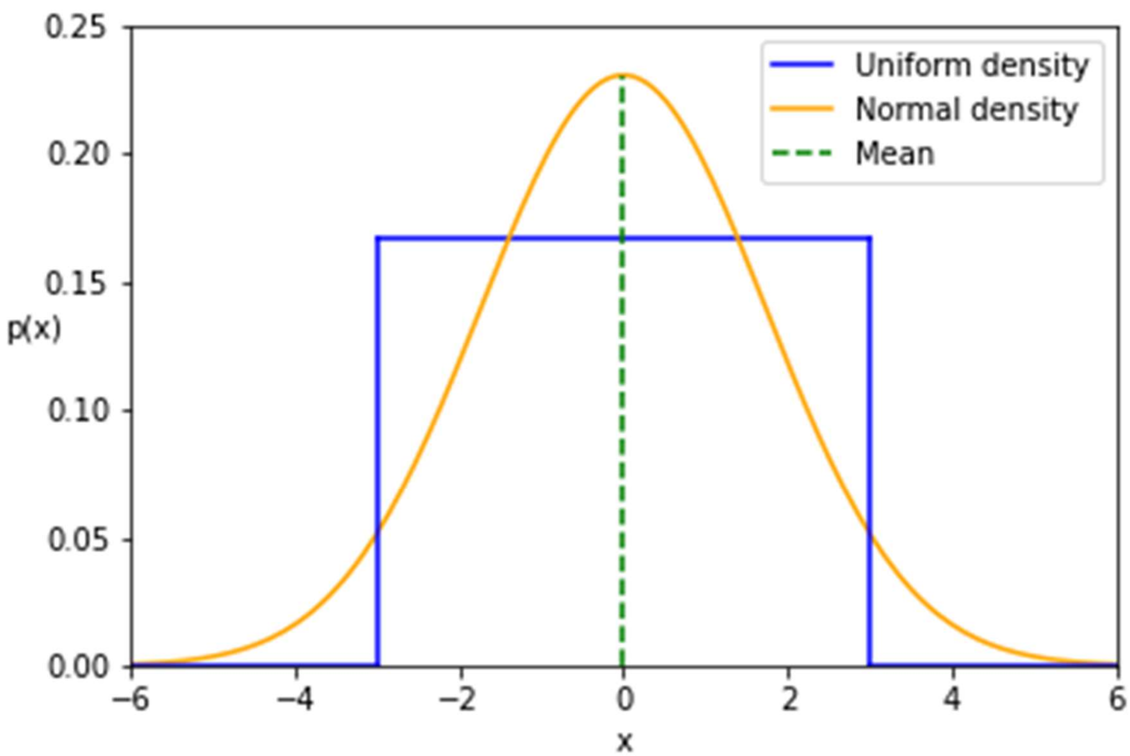
$$\therefore \sigma^2 = \frac{b^2-2ab+a^2}{12} = \frac{(b-a)^2}{12} \Rightarrow \boxed{\sigma^2 = \frac{(b-a)^2}{12}} \Rightarrow \text{Hence Proved.}$$

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 colors.

$\mu = 0 \; , \quad \sigma^2 = 3$

$$F_1 = U(a,b) = U(-3,3) = \begin{cases} 1/6 \; , & -3 \leq x \leq 3 \\ 0 \; , & \text{otherwise} \end{cases}$$

$$F_2 = N(\mu, \sigma^2) = N(0,3) = \frac{1}{\sqrt{6\pi}} e^{-\frac{x^2}{6}} \; , \quad -\infty < x < \infty$$



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

Equation 5 $\rightarrow$ $\sigma^2 = E(x^2) - (E(x))^2$

$\rightarrow$ For discrete random variables,

$$\sigma^2 = E((x-\mu)^2) = \sum_{i=1}^{n} (v_i - \mu)^2 P(v_i)$$

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

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

$$= E(x^2) - 2\mu \times \mu + \mu^2 \times 1$$

$$= E(x^2) - \mu^2 = E(x^2) - (E(x))^2$$

We already know,

$$\sum_{i=1}^{n} P(v_i) = 1 \text{ and}$$

$$E(x) = \mu = \sum_{i=1}^{n} v_i P(v_i)$$

$$\therefore \boxed{\sigma^2 = E(x^2) - (E(x))^2}$$

$\Rightarrow$ Hence proved.

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

$\rightarrow$ $$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$ $\leftarrow$ Normal density.

i) Mean $$= E(x) = \int_{-\infty}^{\infty} x N(\mu, \sigma^2)\, dx = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Let $$y = \frac{(x-\mu)}{\sqrt{2}\,\sigma} \Rightarrow \frac{dy}{dx} = \frac{1}{\sqrt{2}\,\sigma} \Rightarrow dy = \frac{dx}{\sqrt{2}\,\sigma}$$

$$\therefore E(x) = \int_{-\infty}^{\infty} \frac{\sqrt{2}\,\sigma y + \mu}{\sqrt{2\pi}\,\sigma} e^{-y^2} dy \times \sqrt{2}\,\sigma = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} y e^{-y^2} dy + \frac{\mu}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2} dy$$

$y e^{-y^2}$ is an odd function $$\Rightarrow \int_{-\infty}^{\infty} y e^{-y^2} dy = \int_{0}^{\infty} y e^{-y^2} dy - \int_{0}^{\infty} y e^{-y^2} dy = 0$$

$$\therefore E(x) = \frac{\mu}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2} dy$$

We know that $$\operatorname{erf}(\infty) = \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

$$\therefore E(x) = \frac{\mu}{\sqrt{\pi}} \times \sqrt{\pi} = \mu \Rightarrow \boxed{\text{Mean} = E(x) = \mu}$$

ii) Variance $$= E((x-\mu)^2) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Let $$y = \frac{x-\mu}{\sqrt{2}\,\sigma} \Rightarrow dy = \frac{dx}{\sqrt{2}\,\sigma}$$ and $$(x-\mu)^2 = 2\sigma^2 y^2$$

$$\therefore \quad Var = \int_{-\infty}^{\infty} 2\sigma^2 y^2 \cdot \frac{1}{\sqrt{2\pi}\,\sigma} e^{-y^2} \sqrt{2}\,\sigma \, dy = \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2} dy$$

Using Integration By Parts, $\int u\,dv = uv - \int v\,du$

Let $u = y \Rightarrow du = dy$

$dv = y e^{-y^2} dy \Rightarrow v = -\frac{e^{-y^2}}{2}$

$$\boxed{erf(\infty) = \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}}$$

$$\therefore \quad Var = \frac{2\sigma^2}{\sqrt{\pi}} \left(y \frac{e^{-y^2}}{2} \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -\frac{e^{-y^2}}{2} dy \right)$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \left(0 + \frac{1}{2}\sqrt{\pi} \right) = \frac{2\sigma^2}{\sqrt{\pi}} \times \frac{\sqrt{\pi}}{2} = \sigma^2$$

$$\therefore \quad \boxed{Var = \sigma^2}$$

Hence proved, that mean and variance of $N(\mu, \sigma^2)$ are infact the parameters μ and σ^2.

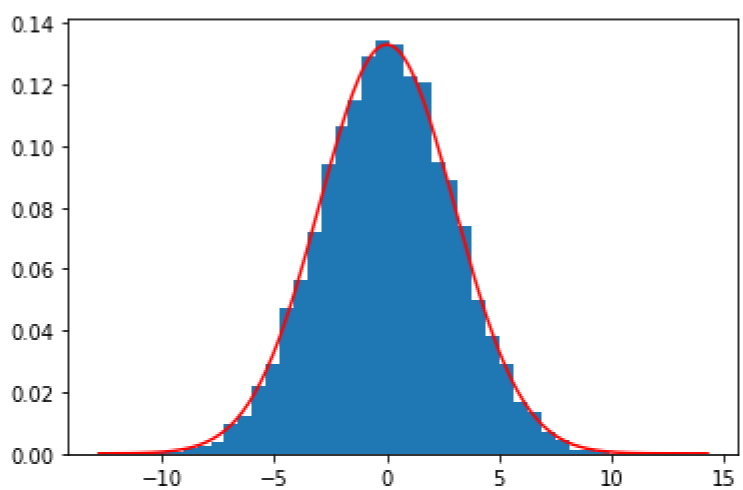
6. Using the inverse of CDFs, map a set of 10,000 random numbers from U[0, 1] to follow the following pdfs:
(a) Normal density with μ = 0, σ = 3.0.
(b) Rayleigh density with σ = 1.0.
(c) Exponential density with λ = 1.5.
Once the numbers are generated, plot the normalized histograms (the values in the 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, INT MAX].

In [1]:
```
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import norm, rayleigh, expon, chi
```

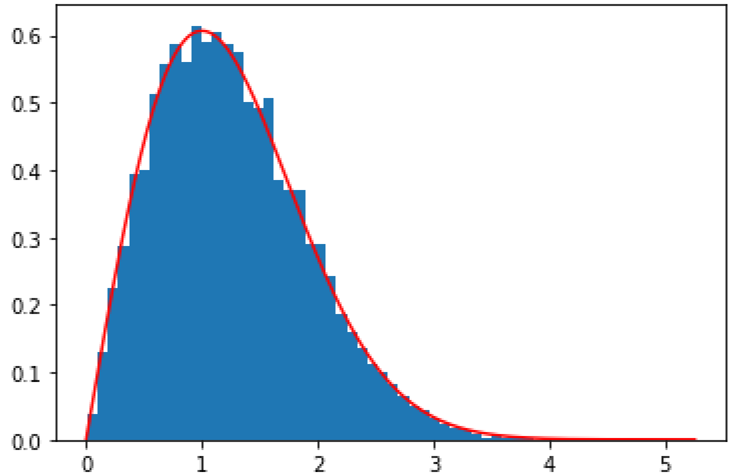
In [2]:
```
def generate_random_numbers(k):
    return np.random.rand(k)
```

In [3]:
```
random_numbers_10k = generate_random_numbers(10000)
normalize_hist = True
```

In [4]:
```
# a) Normal density with  μ  = 0,  σ  = 3.0

mean = 0
sd = 3
pdf_estimate = norm.ppf(random_numbers_10k, mean, sd)
pdf = np.linspace(norm.ppf(0.00001, mean, sd), norm.ppf(0.999999, mean, sd), 200)
bins = np.int((max(pdf_estimate) - min(pdf_estimate))/0.6)   # bin size = 0.6
plt.hist(pdf_estimate, bins, density = normalize_hist)
plt.plot(pdf, norm.pdf(pdf, mean, sd), color = 'red')
plt.show()
```

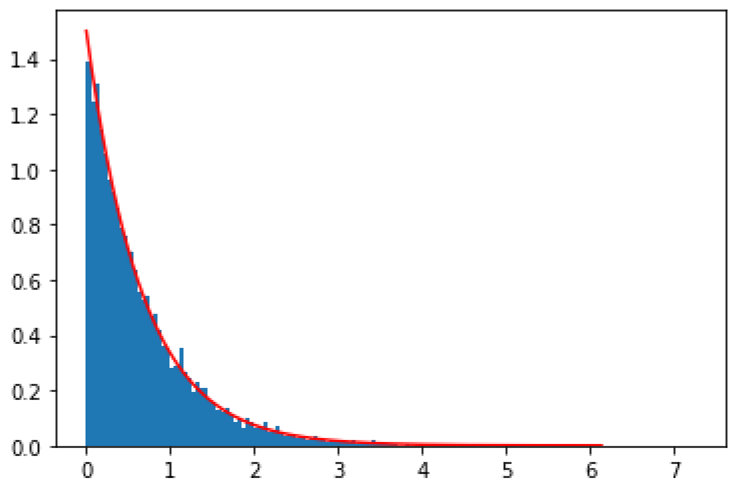


In [5]:
```
# b) Rayleigh density with  σ  = 1.0.

sigma = 1
pdf_estimate = rayleigh.ppf(random_numbers_10k, 0, sigma)
pdf = np.linspace(rayleigh.ppf(0, 0, sigma), rayleigh.ppf(0.999999, 0, sigma), 200)
bins = np.int((max(pdf_estimate) - min(pdf_estimate))/0.09) + 1   # bin size = 0.09
plt.hist(pdf_estimate, bins, density = normalize_hist)
plt.plot(pdf, rayleigh.pdf(pdf, 0, sigma), color = 'red')
plt.show()
```



In [6]:

```
# c) Exponential density with  λ  = 1.5
# this does verify with pdf = λ * exp(-λ*x)

lambda_ = 1.5  # implies f(0) = 1.5
pdf_estimate = expon.ppf(random_numbers_10k, 0, 1/lambda_)
pdf = np.linspace(0, expon.ppf(0.9999, 0, 1/lambda_), 200)
bins = np.int((max(pdf_estimate) - min(pdf_estimate))/0.05)   # bin size = 0.05
plt.hist(pdf_estimate, bins, density = normalize_hist)
plt.plot(pdf, expon.pdf(pdf, 0, 1/lambda_), color = 'red')
plt.show()
```



From the above plots, we can infact infer that given a random variable y that follows the probability distribution function(pdf) U[0, 1], the random variable x = $C^{-1}(y)$ will follow a probability distribution function(pdf) with corresponding cumulative distribution function(cdf) as C(), as mentioned in the shared pdf. The above plots show this for Normal, Rayleigh and Exponential density pdfs respectively.

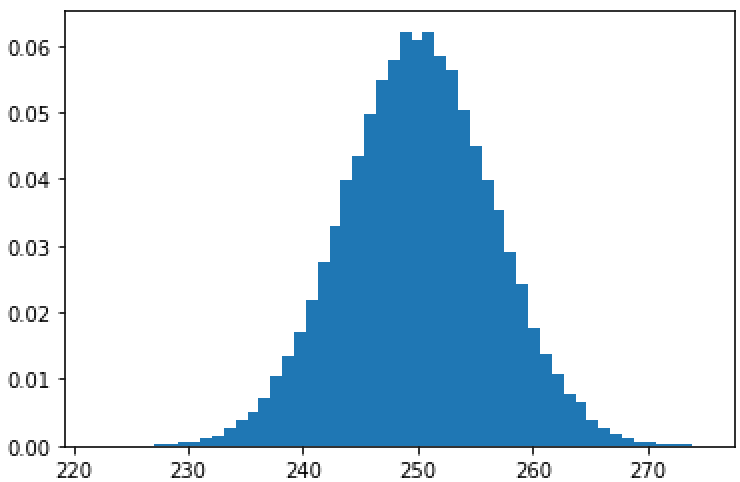
7. Write a function to generate a random number as follows: Every time the function is called, it generates 500 new random numbers from U[0, 1] and outputs their sum. Generate 50,000 random numbers by repeatedly calling the above function, and plot their normalized histogram (with bin-size = 1). What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

In [1]:
```
import matplotlib.pyplot as plt
import numpy as np
```

In [2]:
```
def generate_random(k):
    rand_array = np.random.rand(k)
    return sum(rand_array)

def generate_random_numbers(K, k):
    return np.array([generate_random(k) for _ in range(K)])
```

In [3]:
```
k = 500
K = 50000
random_numbers = generate_random_numbers(K, k)
```

In [4]:
```
normalized = True
bin_size = 1
bins = np.int((max(random_numbers) - min(random_numbers))/1)  #bin size = 1
plt.hist(random_numbers, bins, density = normalized)
plt.show()
```



The above histogram looks like a normal distribution curve or a bell curve with its mean(μ) at 250 which has been discretized or sampled at intervals of size=1.