Statistical Methods in AI: Assignment 1

Due on September 25, 2021 at 12:00 am

Professor Anoop Namboodiri

Tushar Choudhary
2019111019

Problem 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.

Solution

Finite Range

For finite range, we can take the example of a Bernoulli distribution.
For a real number p such that $0 < p < 1$, consider the PMF:

$$P_X(x) = \begin{cases} 1-p, & \text{if } x = 0 \\ p, & \text{if } x = 1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Here the range of the PMF is $R_x = \{0, 1\}$ which is finite.
Next, we check whether the conditions of PMF are satisfied by this example

$$\begin{aligned} \sum_{x \in R_x} P_X(x) &= P_X(0) + P_X(1) \\ &= (1-p) + (p) \\ &= 1 \end{aligned} \tag{2}$$

Hence, this is a valid PMF with a finite range.

Infinite Range

For infinite range, we can take the example of a random variable X which represents the number of coin tosses we do till we get the first tails.
Consider the PMF:

$$P_X(x) = \begin{cases} \frac{1}{2^x}, & \text{if } x \in \mathbb{N} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Here the range of the PMF is $R_x = N$ which is infinite.
Next, we check whether the conditions of PMF are satisfied by this example

$$\begin{aligned} \sum_{x \in R_x} P_X(x) &= \sum_{x \in \mathbb{N}} P_X(x) \\ &= \sum_{x \in \mathbb{N}} \frac{1}{2^x} \\ &= \frac{1/2}{1 - 1/2} \\ &= 1 \end{aligned} \tag{4}$$

Hence, this is a valid PMF with an infinite range.

Problem 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.)

Solution

If we have the PDF of a uniform distribution as $\mathrm{f}_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a <= x <= b \\ 0, & \text{otherwise} \end{cases}$

Then required result is $\sigma^2 = \frac{(b-a)^2}{12}$.

We know that for a random variable X, we have σ^2 =E[X^2] $- (E[X])^2$.
Hence, if we know the values for $E[X^2]$ and $E[X]^2$ we can calculate the value of σ^2.

Calculating $E[X]$:

$$\begin{aligned} E[x] &= \int_a^b \frac{x}{b-a}\,dx \\ &= \frac{1}{b-a}\int_a^b x\,dx \\ &= \frac{1}{2(b-a)}\,(b^2-a^2) \\ &= \frac{a+b}{2} \end{aligned} \tag{5}$$

Calculating $E[X^2]$:

$$\begin{aligned} E[x^2] &= \int_a^b \frac{x^2}{b-a}\,dx \\ &= \frac{1}{b-a}\int_a^b x^2\,dx \\ &= \frac{1}{3(b-a)}\,(b^3-a^3) \\ &= \frac{a^2+ab+b^2}{3} \end{aligned} \tag{6}$$

Calculating σ^2:

$$\begin{aligned} \sigma^2 &= E[x^2] - (E[x])^2 \\ &= \frac{a^2+ab+b^2}{3} - (\frac{a+b}{2})^2 \\ &= \frac{a^2+ab+b^2}{3} - (\frac{a^2+2ab+b^2}{4}) \\ &= \frac{a^2+2ab+b^2}{12} \\ &= \frac{(b-a)^2}{12} \end{aligned} \tag{7}$$

Hence, proved.

Problem 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.

Solution

Consider an exponential distribution with parameter $\lambda = 1$. The mean of this distribution will be $1/\lambda = 1/1 = 1$ and variance will be $1/\lambda^2 = 1/1 = 1$.

Consider a normal distribution with parameters $\mu = 1$ and $\sigma = 1$. The mean of this distribution will be equal $\mu = 1$ and variance will be equal to $\sigma^2 = 1$.

Hence, the mean and variance for both these distributions is 1.

I have written the following code in python3 to generate the graph:

```
import matplotlib.pyplot as plt
import numpy as np

x = np.linspace(0, 5, 500, endpoint=True) # 500 points equally distributed from 0 to 4
y1 = np.exp(-x) # Exponential distribution with lambda = 1
y2 = np.exp(-((x-1)**2)/2) / np.sqrt(2*np.pi) # Normal distribution with mu = 1 and sigma = 1
plt.plot(x, y1, label="Exponential Distribution")
plt.plot(x, y2, label="Gaussian Distribution")
plt.legend()
plt.title('Exponential Distribution v/s Gaussian Distribution')
plt.show()
```

The graph obtained is:

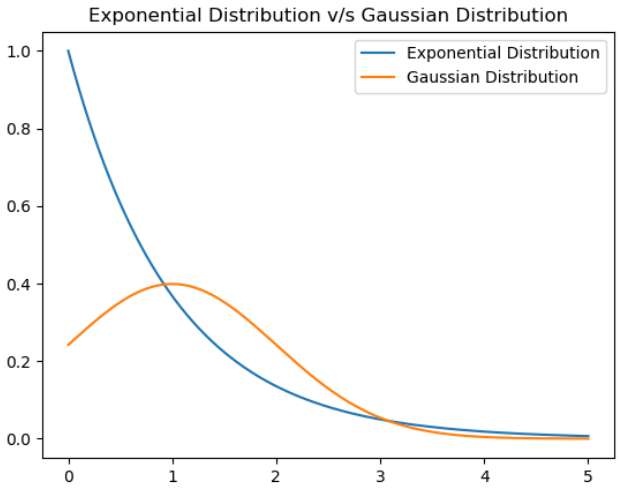


Figure 1: Exponential Distribution v/s Gaussian Distribution

Problem 4

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

Solution

If we have the PMF of a discrete random variable X given by $P_X(x)$ over range R_X, then required result is σ^2 =E[X^2] $- E[X]^2$.

We know that for a random variable X, we have σ^2 =E[(X - E[X])^2].

Calculating σ^2:

$$\begin{aligned}
\sigma^2 &= E[(X - E[X])^2] \\
&= \sum_{x \in R_x} P_X(x) * (x - E[X])^2 \\
&= \sum_{x \in R_x} P_X(x) * (x^2 - 2xE[X] + E[X]^2) \\
&= \sum_{x \in R_x} P_X(x)x^2 - 2E[X] \sum_{x \in R_x} P_X(x)x + E[X]^2 \sum_{x \in R_x} P_X(x) \\
&= E[X^2] - 2E[X] * E[X] + E[X]^2 * 1 \\
&= E[X^2] - E[X]^2
\end{aligned} \tag{8}$$

Hence, proved.

Problem 5

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

Solution

We know that PDF of a normal random variable $X \sim N(\mu, \sigma^2)$ is given by $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. For this distribution, the mean E[X] can be calculated by $E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx$, using which we can get the variance Var(X) = $E[(X - E[X])^2]$.

Calculating E[X]:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_{-\infty}^{\infty} \frac{x}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \end{aligned} \tag{9}$$

Now in the above equation we will substitute $x = x' + \mu$ to simplify the term, $(x - \mu)^2$.
Also, d(x) = $d(x' + \mu)$ = dx' (since the mean μ is constant). The limits shall still stay $-\infty$ to ∞.
So now we have,

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} \frac{x' + \mu}{\sigma\sqrt{2\pi}} e^{-\frac{x'^2}{2\sigma^2}}\, dx' \\ &= \int_{-\infty}^{\infty} \frac{x'}{\sigma\sqrt{2\pi}} e^{-\frac{x'^2}{2\sigma^2}}\, dx' + \mu \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x'^2}{2\sigma^2}}\, dx' \end{aligned} \tag{10}$$

Now on observing carefully, we can see that the function in the first integral is an odd function and hence its integral from $-\infty$ to ∞ will be 0. So we're left with just the second term.

$$\begin{aligned} E[X] &= \mu \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x'^2}{2\sigma^2}}\, dx' \\ &= \mu \int_{-\infty}^{\infty} f_X(x') dx' \\ &= \mu * 1 \\ &= \mu \end{aligned} \tag{11}$$

(Using $\int_{-\infty}^{\infty} f_X(x') dx' = 1$, which is property of probability density function.)

Calculating Var(X):

$$\begin{aligned} Var(X) &= E[(X - E[X])^2] \\ &= \int_{-\infty}^{\infty} \frac{(x - E[X])^2}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \\ &= \int_{-\infty}^{\infty} \frac{x'^2}{\sigma\sqrt{2\pi}} e^{-\frac{x'^2}{2\sigma^2}}\, dx' \end{aligned} \tag{12}$$

Since the function inside the above integral is an even function, the value of the integral from $-\infty$ to ∞ will be twice the value of value of integral from 0 to ∞.

Problem 5 continued on next page...

So we have:

$$Var(X) = 2 * \int_0^\infty \frac{x'^2}{\sigma\sqrt{2\pi}} e^{-\frac{x'^2}{2\sigma^2}} dx' \tag{13}$$

Now in the above function, we substitute $k = \frac{x'^2}{2\sigma^2}$.
We have $x'^2 = 2\sigma^2 k$ and from this we can get $dx' = \frac{\sigma dk}{\sqrt{2k}}$.

Next,

$$\begin{aligned} Var(X) &= 2 * \int_0^\infty \frac{2\sigma^2 k}{\sigma\sqrt{2\pi}} \ e^{-k} \ \frac{\sigma dk}{\sqrt{2k}} \\ &= \frac{2\sigma^2}{\sqrt{\pi}} * \int_0^\infty k^{1/2} e^{-k} dk \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \Gamma\left(\frac{3}{2}\right) \end{aligned} \tag{14}$$

In the above equation, Γ is the Gamma function.
We know $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ and $\Gamma(t+1) = t * \Gamma(t)$. This gives us $\Gamma\left(\frac{3}{2}\right) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{\pi}}{2}$.

This gives us,

$$Var(X) = \frac{2\sigma^2}{\sqrt{\pi}} * \frac{\sqrt{\pi}}{2} = \sigma^2 \tag{15}$$

Problem 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,INT MAX]

Solution

I have written the following code in python3 to generate the required graphs:

```
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import norm, rayleigh, expon

# Generating the random numbers
x = np.sort(np.random.rand(10000))

# For part A - Normal density
y = norm.ppf(x,scale=3)
pdf = norm.pdf(y,scale=3)
plt.figure(1)
plt.hist(y,density=True,bins=500,label='Histogram')
plt.plot(y,pdf,lw=2,label='PDF')
plt.title('Histogram and PDF for Gaussian Distribution')
plt.legend()

# For part B - Rayleigh density
y = rayleigh.ppf(x)
pdf = rayleigh.pdf(y)
plt.figure(2)
plt.hist(y,density=True,bins=500,label='Histogram')
plt.plot(y,pdf,lw=2,label='PDF')
plt.title('Histogram and PDF for Rayleigh Distribution')
plt.legend()

# For part C - Exponential density
y = expon.ppf(x,scale=1/1.5)
pdf = expon.pdf(y,scale=1/1.5)
plt.figure(3)
plt.hist(y,density=True,bins=500,label='Histogram')
plt.plot(y,pdf,lw=2,label='PDF')
plt.title('Histogram and PDF for Exponential Distribution')
plt.legend()

plt.show()
```

Part A

Normal density with $\mu = 0$, $\sigma = 3$. For a random variable X with Normal distribution $N(\mu, \sigma^2)$ we have:
* PDF = $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
* CDF = $F_X(x) = \phi\left(\frac{x-\mu}{\sigma}\right)$
* Inverse of CDF = $F_X^{-1}(x) = \mu + \sigma\phi^{-1}(x)$
(ϕ here represents the CDF of a Standard Normal Distribution)

The graph obtained is:

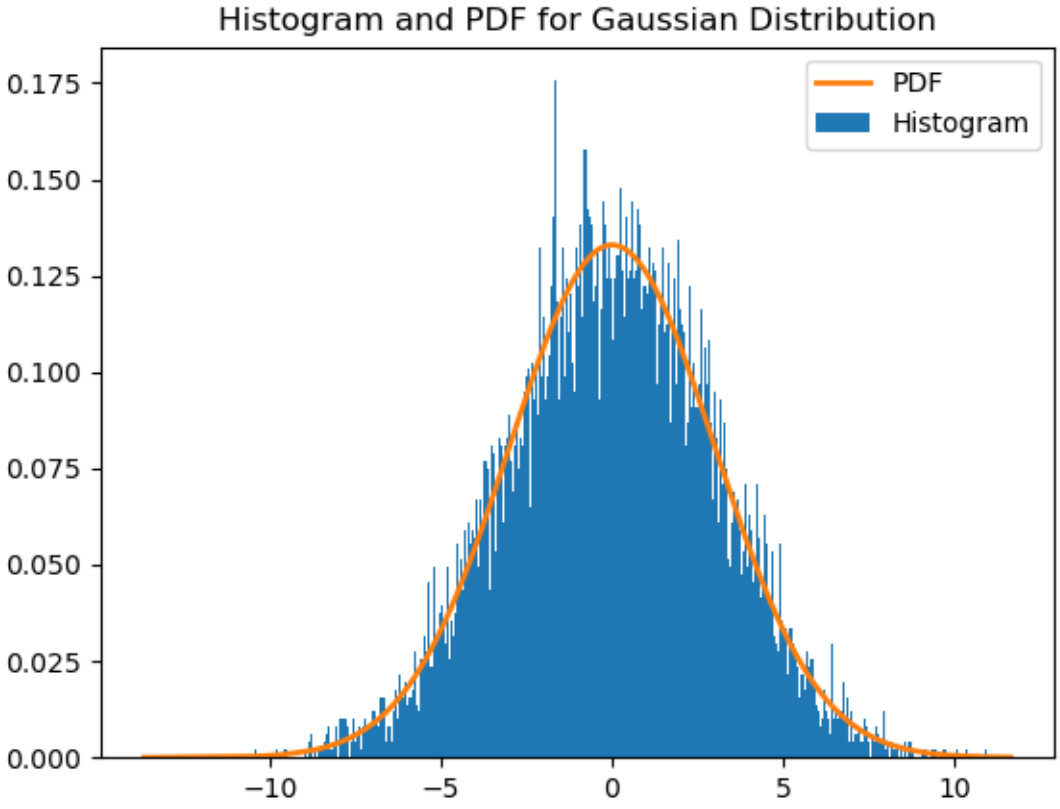


Figure 2: Histogram and PDF for Gaussian Distribution

Inference- - The values obtained using the CDF inverse have a Gaussian distribution, and we can see the histogram of the distribution matches the PDF.

Problem 6 continued on next page...

Part B

For Rayleigh density with $\sigma = 1$ with:
* PDF = $f_X(x) = \frac{x}{\sigma^2} e^{-\frac{x}{2\sigma^2}}$
* CDF = $F_X(x) = 1 - e^{-\frac{x}{2\sigma^2}}$
* Inverse of CDF = $F_X^{-1}(x) = \sigma\sqrt{-2ln(1-x)}$

The graph obtained is:

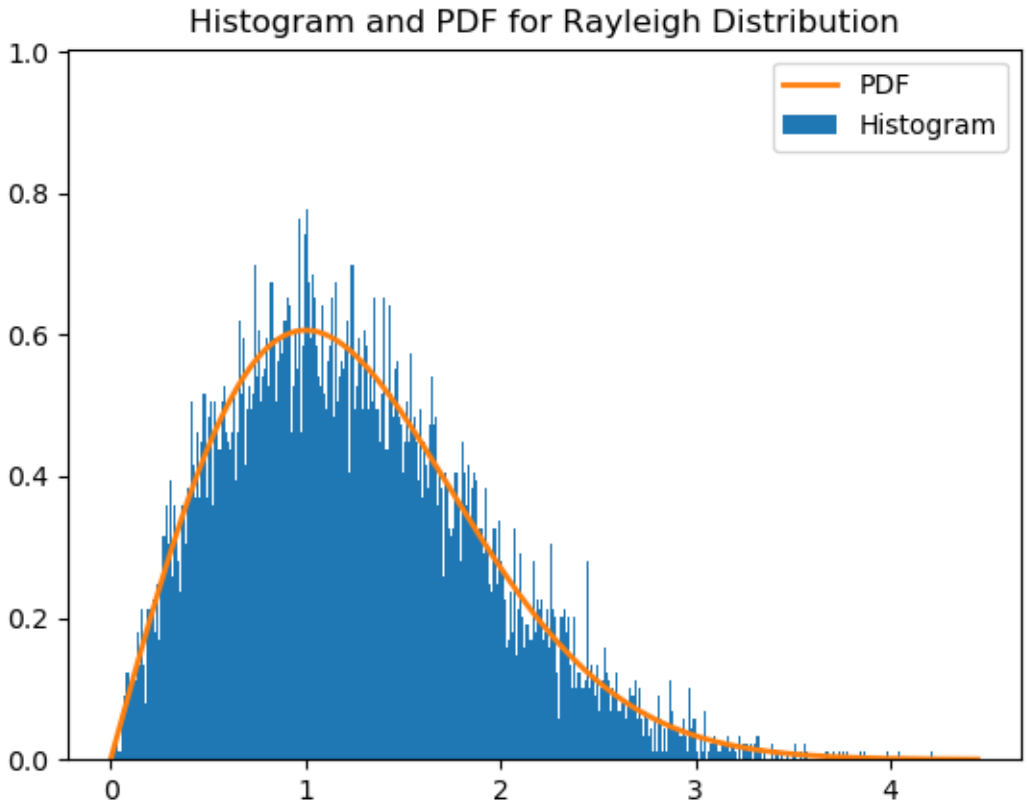


Figure 3: Histogram and PDF for Rayleigh Distribution

Inference- - The values obtained using the CDF inverse have a Rayleigh distribution, and we can see the histogram of the distribution matches the PDF.

Problem 6 continued on next page...

Part C

For Exponential density with λ=1.5, we have:
* PDF = $f_X(x) = \lambda e^{-\lambda x}$
* CDF = $F_X(x) = 1 - e^{-\lambda x}$
* Inverse of CDF = $F_X^{-1}(x) = \frac{-ln(1-x)}{\lambda}$
(The standard deviation $\sigma = \frac{1}{\lambda}$).

The graph obtained is:

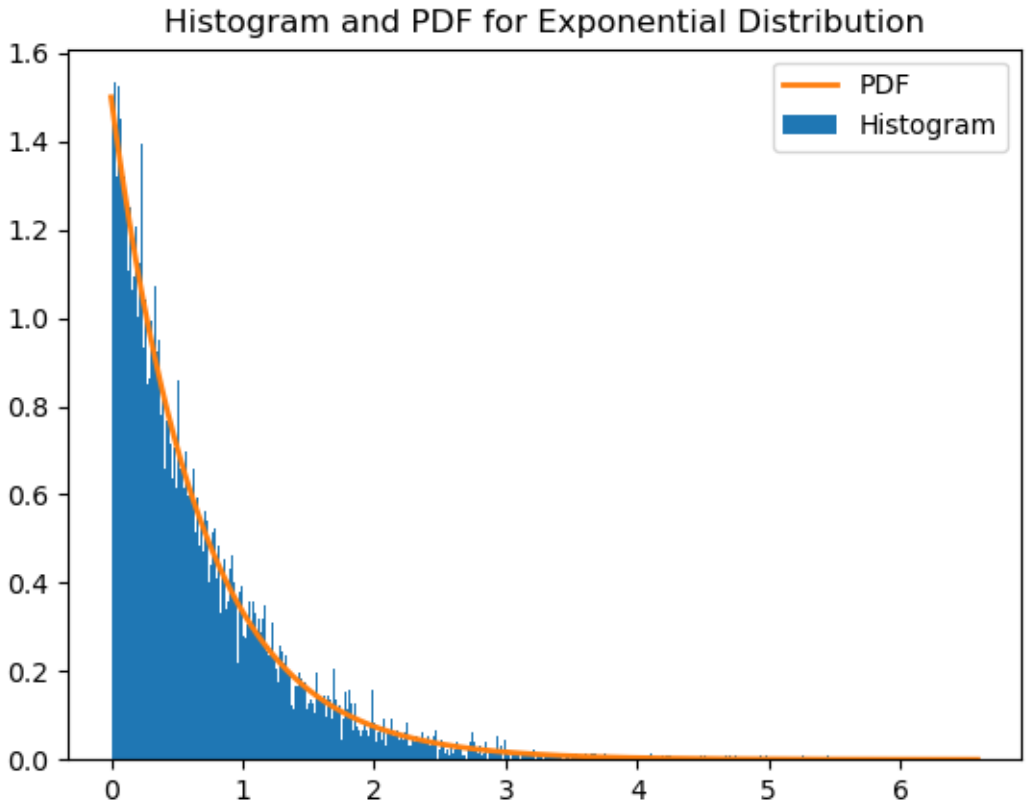


Figure 4: Histogram and PDF for Exponential Distribution

Inference- - The values obtained using the CDF inverse have an Exponential distribution, and we can see the histogram of the distribution matches the PDF..

Problem 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?

Solution

I have written the following code in python3 to generate the required graph:

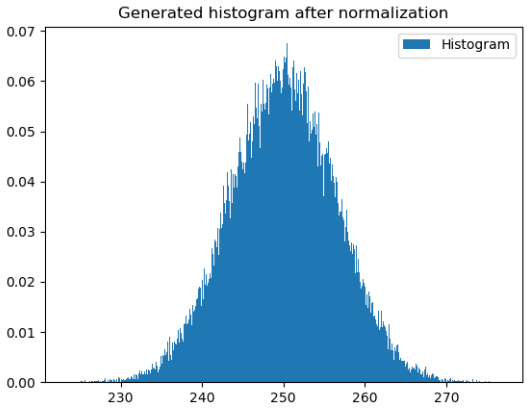
```
import matplotlib.pyplot as plt
import numpy as np

def sum500():
  return np.random.random(500).sum()

arr = []
for i in range(50000):
    arr.append(sum500())

plt.hist(arr,density=True,bins=500,label='Histogram')
plt.title('Generated histogram after normalization')
plt.legend()
plt.show()
```

The graph obtained is:



Observation:
We can see the distribution in the histogram corresponds to a Gaussian distribution with the peak at 250. This result can be explained using Central Limit Theorem, which states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.