Expt. No.

STATISTICAL METHODS IN A.I.
ASSIGNMENT-1

Question 1)

Answer: Given conditions on PMF are:

i) $P(X = x_i) = p_i \geq 0$, where 'X' = discrete random variable

ii) $\sum_{i=1}^{n} P(X = x_i) = \sum_{i=1}^{n} p_i = 1$

Example 1: $X \sim Bin(n,p)$ where $Bin(n,p)$ = Binomial distribution with parameters 'n' (= # of ~~trials~~ independent Bernoulli trials) & 'p' (= ~~out~~ success probability for each trial, as Binomial model the no. of successes in a sample of size 'n' drawn with replacement)

Here 'X' can take values '0', '1', '2', ..., 'n'

& $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, $\boxed{k = 0, 1, \ldots, n}$

prob that 'k' trials out of 'n' are ~~remain~~ successful

while the rest $(n-k)$ are not successful.

Checking conditions:

i) Since $n \in \mathbb{N}$ & $k \in [\mathbb{N} \cup \{0\}] \cap [0,n]$.

$\binom{n}{k} > 0$.

And $p \in [0,1] \Rightarrow p^k \in [0,1]$ & $(1-p)^{n-k} \in [0,1]$ too.

$\Rightarrow \boxed{P(X=k) \geq 0 \quad \forall\, k}$

ii) $\sum_{k=0}^{n} P(X=k) = \binom{n}{0}(1-p)^n + \binom{n}{1} p (1-p)^{n-1} + \cdots + \binom{n}{n} p^n$

$= [(1-p) + p]^n = 1^n = 1$

Nalanda

Teacher's Signature ....

Example 2: let $X \sim P(\lambda)$, where $P(\lambda)$ = poisson distribution with parameter λ (= time rate of occurence of given no. events). Here $P(\lambda)$ models no. of events occuring in a fixed interval or space if these events occur with a known constant mean rate & independently of the time since last event.

Condition check:

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0,1,2,\dots$$ (i.e. $k \in \mathbb{N} \cup \{0\}$) (where $\lambda > 0$)

Condition check:

i) Since $[\lambda > 0]$, $[e^{-x} > 0 \ \forall x \in \mathbb{R}]$, & $[k! > 0 \ \forall k]$, then $\boxed{P(X=k) \geq 0 \ \forall k}$

$\boxed{\text{or we can say } P(X=k) > 0 \ \forall k}$.

ii) $$\sum_{k=0}^{\infty} P(X=k) = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!}$$

$$= e^{-\lambda} + \lambda e^{-\lambda} + \frac{\lambda^2 e^{-\lambda}}{2!} + \dots$$

$$= e^{-\lambda}\left[1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots\right]$$

$$= e^{-\lambda} \cdot e^{\lambda}$$ ($\because e^x = 1 + x + \frac{x^2}{2!} + \dots$ from Maclaurin series expansion)

$$= 1$$

Expt. No.

Question 2.

Answer : Say, $U \sim$ Uniform(a,b) then

$$p_u' = \begin{cases} \frac{1}{b-a}, & \text{if } a \le x \le b \\ 0, & \text{otherewise} \end{cases}$$

Now, $Var(U) = E(U^2) - (EU)^2$ (using equation ⑤ as in notes)

$$= \int_{-\infty}^{\infty} t^2 \, p_u(t)\, dt - \left[\int_{-\infty}^{\infty} t \, p_u(t)\, dt\right]^2$$

$$= \int_{-\infty}^{a} t^2 \cdot 0 \cdot dt + \int_{a}^{b} \frac{t^2\, dt}{b-a} + \int_{b}^{\infty} t^2 \cdot 0 \cdot dt$$

$$- \left[\int_{-\infty}^{a} t \cdot 0 \cdot dt + \int_{a}^{b} \frac{t\, dt}{b-a} + \int_{b}^{\infty} t \cdot 0 \cdot dt\right]^2$$

$$= \left[\frac{t^3}{3}\right]_a^b \frac{1}{(b-a)} - \left[\left[\frac{t^2}{2}\right]_a^b \frac{1}{(b-a)}\right]^2$$

$$= \frac{b^3 - a^3}{3(b-a)} - \left[\frac{b^2 - a^2}{2(b-a)}\right]^2$$

$$= \frac{b^2 + ab + a^2}{3} - \frac{(b+a)^2}{4} \qquad [\because b^3 - a^3 = (b-a)(b^2+a^2+ab)]$$

$$= \frac{4b^2 + 4a^2 + 4ab - 3a^2 - 3b^2 - 6ab}{12}$$

$$= \frac{b^2 + a^2 - 2ab}{12} = \frac{(b-a)^2}{12}$$

Nalanda

Teacher's Signature ....

4.)

~~for discrete random variables~~

Say, $X \sim$ Discrete probability distribution

$f_X(x) =$ (Probability Mass function).

Now, $Var(X) = E[(X - E(X))^2]$

$$= \cancel{E\left(\left[X - \sum_{x \in \mathcal{X}} x f_X(x)\right]^2\right)}$$

boils down to.

$$= \sum_{i=1}^{n} (v_i - \mu)^2 f_X(v_i)$$, where $X = v_i$ & $\mu = \sum_{i=1}^{n} v_i f_X(v_i)$

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

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

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

$$= E(X^2) + \mu^2 - 2\mu^2 \quad \left(\because \sum_{i=1}^{n} v_i f_X(v_i) = \mu\right)$$

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

$$= E(X^2) - (EX)^2 \quad \left(\because E(X) = \mu = \sum_{i=1}^{n} v_i f_X(v_i)\right)$$

- Even if $n = \infty$ (i.e. for infinite possible values for X) the above derivatn will follow along the same line.
- Hence, for discrete distribution, we too have

$$Var(X) = E(X^2) - (EX)^2$$

Expt. No.

Question-5:

Answer: Objective: Say, $X \sim N(\mu, \sigma^2)$, Then we need to prove:

i) $E(X) = \mu$

ii) $Var(X) = \sigma^2$

Proof: Now, we know

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, x \in \mathbb{R}$$

i) $$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

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

$$= \frac{1}{\sqrt{2\pi}\cancel{\sigma}}\int_{-\infty}^{\infty} (\sigma.t+\mu)\, e^{-\frac{t^2}{2}}\, \cancel{\sigma}\, dt$$

$\boxed{\begin{array}{l} \text{take } t = \frac{x-\mu}{\sigma} \\ \Rightarrow \sigma\, dt = dx \end{array}}$

$$= \frac{\mu}{\sqrt{2\pi}}\underbrace{\int_{-\infty}^{\infty} e^{-t^2/2}\, dt}_{\sqrt{2\pi}} + \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty} t\, e^{-t^2/2}\, dt$$

$$= \left[\frac{\mu}{\cancel{\sqrt{2\pi}}} * \cancel{\sqrt{2\pi}}\right] + 0 \quad \left\{\begin{array}{l} \because t = \text{odd-func} \\ \& \ e^{-t^2/2} = \text{even-func} \end{array}\right.$$

($\because$ for standard normal,

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}\, dx = 1$$

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

Nalanda

Teacher's Signature

Hence, $E(X) = \mu + 0 = \mu$ — ▨

ii) $Var(X) = E(X^2) - (EX)^2 = E(X^2) - \mu^2$ ($\because$ we know $EX = \mu$ from above proof..)

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

let's compute $\to$ $E(X^2)$ only:

$$E(X^2) = \frac{1}{\sqrt{2\pi}\,\cancel{\sigma}}\int_{-\infty}^{\infty} (\sigma t+\mu)^2 e^{-t^2/2}\,\cancel{\sigma}\,dt$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} (\sigma^2 t^2 + \mu^2 + 2\mu\sigma t)\, e^{-t^2/2}\,dt$$

$$= \frac{2\sigma^2}{\sqrt{2\pi}}\int_0^{\infty} \underbrace{t^2 e^{-t^2/2}}_{\text{(as both are even)}}\,dt + \frac{\mu^2}{\cancel{\sqrt{2\pi}}}\cdot\cancel{\sqrt{2\pi}} + 2\mu\sigma\cdot 0$$

$$= \frac{2\sigma^2}{\cancel{\sqrt{2}}\sqrt{\pi}}\int_0^{\infty} \cancel{\sqrt{2}}\sqrt{z}\, e^{-z}\,dz + \mu^2$$ (taking $t^2/2 = z$)

$$= \frac{2\sigma^2}{\sqrt{\pi}}\int_0^{\infty} \sqrt{z}\, e^{-z}\,dz + \mu^2 = \frac{2\sigma^2}{\sqrt{\pi}}\Gamma(3/2) + \mu^2$$

($\because \Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x}\,dx$)

$$= \frac{\cancel{2}\sigma^2}{\cancel{\sqrt{\pi}}}\cdot\frac{\cancel{\sqrt{\pi}}}{\cancel{2}} + \mu^2$$

($\because \Gamma(n+1/2) = \frac{1\cdot 3\cdot 5\cdots(2n-1)}{2^n}\sqrt{\pi}$ where $n \in \mathbb{N}$)

$= \sigma^2$

$\Rightarrow Var(X) = (\sigma^2 + \mu^2) - \mu^2 = \sigma^2$ — ▨

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

Question 3

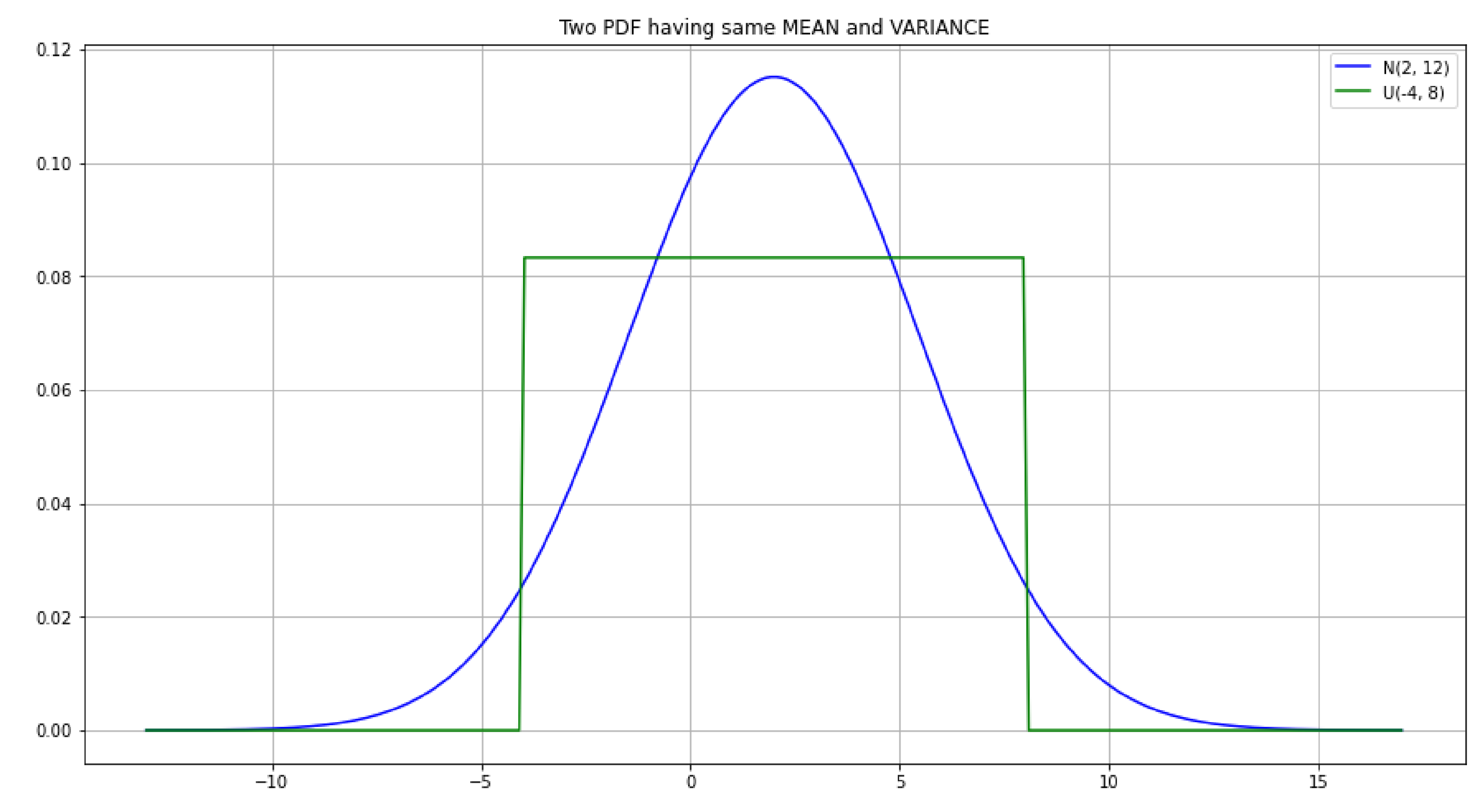
Show examples of two density functions (draw the function plots) that have the **same mean and variance**, but clearly **different distributions**. Plot both functions in the same graph with different colours.

Answer:

We take $X \sim N(2, 12)$, where $\mu = 2$, $\sigma^2 = 12$, and $Y \sim U[-4, 8]$, which comes out be $\mathbb{E}(Y) = \dfrac{8-4}{2} = 2$, and $Var(Y) = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 = \dfrac{(b-a)^2}{12} = 12$.

In [2]:
```python
# define normal distribution function.
def f_normal(x, mu, sigma2):
    """
    Input:
        x: a data point or an array of data points.
        mu: Mean parameter of the distribution function.
        sigma2: variance of the distribution function.
    Output:
        Corresponding probability value at x
    """
    return (1/np.sqrt(2*np.pi*sigma2))*np.exp(-(x-mu)**2/(2*sigma2))
# def Uniform distribution function
def f_uniform(x, a, b):
    """
    Input:
        x: a data point or an array of data points.
        a: Left limit of the distribution function.
        b: Right limit of the distribution function.
    Output:
        Corresponding probability value at x
    """
    if type(x) == np.ndarray: return np.array([1/(b-a) if a<=ele<=b else 0 for ele in np.squeeze(x)])
    else:
        if a<=x<=b: return 1/(b-a)
        else: return 0
```

In [3]:
```python
# plot the N(2, 4) and U(8, -4).
data = np.linspace(-13, 17, 240)
plt.figure(figsize = (15,8))
plt.plot(data,f_normal(data, 2, 12), color = "blue", label = "N(2, 12)")
plt.plot(data,f_uniform(data, -4, 8), color = "green", label = "U(-4, 8)")
plt.legend(loc = "best"); plt.title("Two PDF having same MEAN and VARIANCE"); plt.grid(); plt.show()
```



Question 6

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

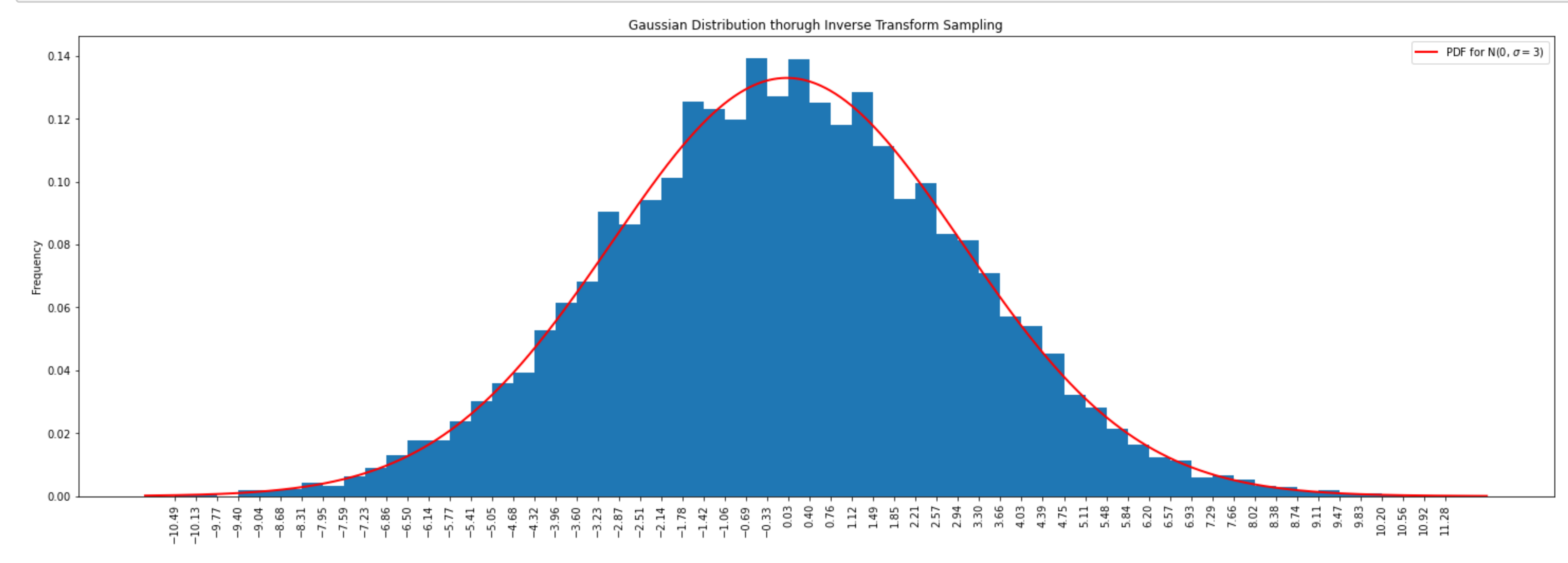
1. Normal density with $\mu = 0, \sigma = 3.0$.
2. Rayleigh density with $\sigma = 1.0$.
3. Exponential density with $\lambda = 1.5$.

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]`.

In [4]:
```python
# Generate 10,000 random numbers from U[0, 1].
data_uni = np.random.uniform(0, 1, 10000)
```

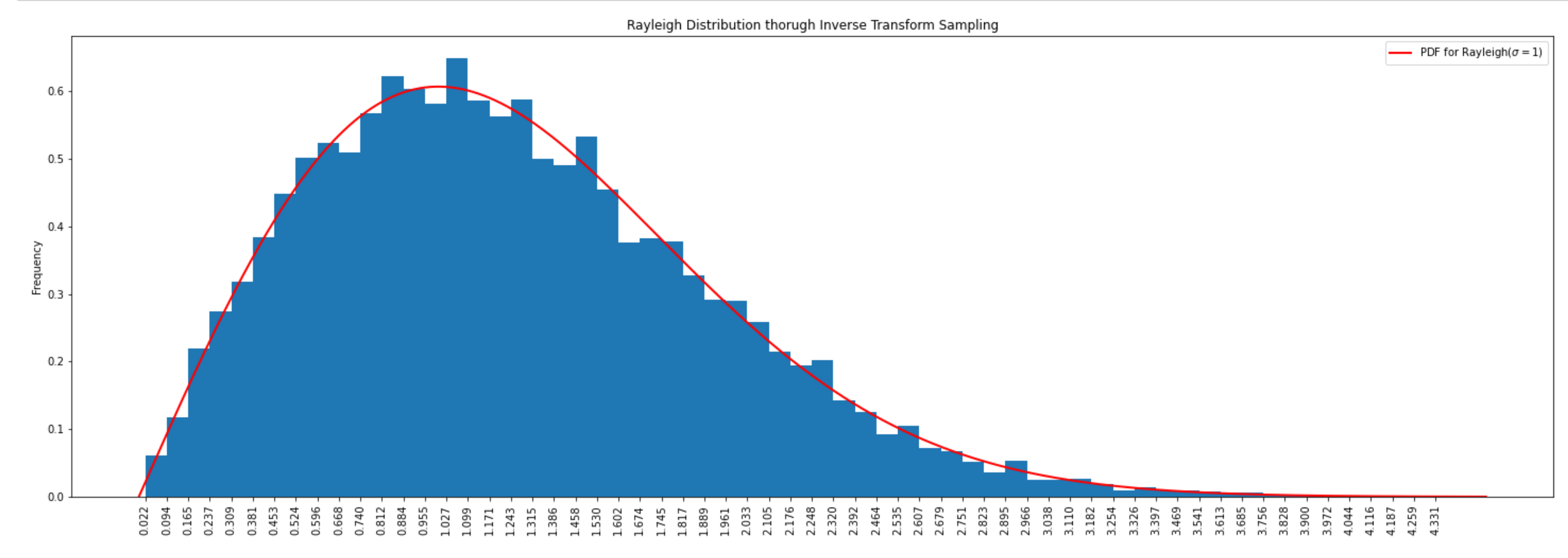
Normal density with $\mu = 0, \sigma = 3.0$.

In [11]:
```python
normal_cdf_inv = norm.ppf(data_uni, 0, 3)
#bins = [k for k in range(round(min(normal_cdf_inv)), round(max(normal_cdf_inv) + 1))]
plt.figure(figsize=(25, 8))
vals_at_x_norm, bins_norm, patches_norm = plt.hist(normal_cdf_inv, bins = 60, density = True)
pdf_x_points = np.linspace(-11,12,1000)
plt.plot(pdf_x_points, norm.pdf(pdf_x_points, 0, 3), color = "r", linewidth = 2, label = "PDF for N
(0, $\sigma = 3$)")
plt.xticks(bins_norm, rotation = 90); plt.ylabel("Frequency")
plt.title("Gaussian Distribution thorugh Inverse Transform Sampling"); plt.legend(loc = "best"); plt.
show()
```



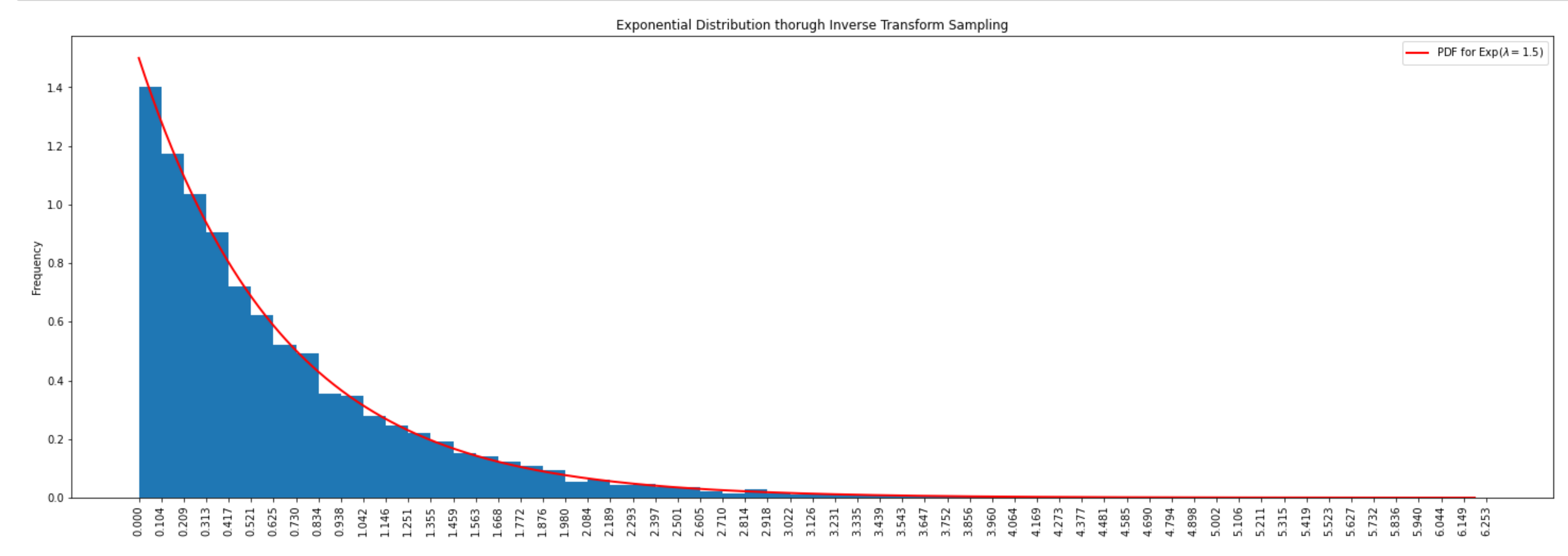
Rayleigh density with $\sigma = 1.0$.

In [18]:
```python
rayleigh_cdf_inv = rayleigh.ppf(data_uni, 0, 1)
#bins_ray = [k for k in range(round(min(rayleigh_cdf_inv)), round(max(rayleigh_cdf_inv) + 1))]
plt.figure(figsize=(25, 8))
_, bins_ray, patches_ray = plt.hist(rayleigh_cdf_inv, bins = 60, density = True)
pdf_x_points = np.linspace(0.0,4.5,1000)
plt.plot(pdf_x_points, rayleigh.pdf(pdf_x_points, 0, 1), color = "r", linewidth = 2, label = "PDF for
 Rayleigh($\sigma = 1$)")
plt.xticks(bins_ray, rotation = 90); plt.ylabel("Frequency")
plt.title("Rayleigh Distribution thorugh Inverse Transform Sampling"); plt.legend(loc = "best"); plt.
show()
```



Exponential density with $\lambda = 1.5$.

In [21]:
```python
exp_cdf_inv = expon.ppf(data_uni, 0, (1/1.5))
plt.figure(figsize=(25, 8))
_, bins_exp, patches_exp = plt.hist(exp_cdf_inv, bins = 60, density = True)
pdf_x_points = np.linspace(0,6.2,1000)
plt.plot(pdf_x_points, expon.pdf(pdf_x_points, 0, 1/1.5), color = "r", linewidth = 2, label = "PDF fo
r Exp($\lambda = 1.5$)")
plt.xticks(bins_exp, rotation = 90); plt.ylabel("Frequency")
plt.title("Exponential Distribution thorugh Inverse Transform Sampling"); plt.legend(loc = "best"); p
lt.show()
```



Inference from the plots —

1. Inverse of CDF (for a particular distribution) is a transformation of standard uniform random variable ($\sim U[0, 1]$) to such a random variable whose distribution is same as PDF for that particular random variable.
2. And, suppose X is a r.v. whose support (= domain) $= [a, b]$, where $a, b \in \mathbb{R}$ and $a < b$, and $F_X(x)$ is the corresponding CDF for X. Then the interval/ region where slope of CDF is maximum (since non-decreasing, it can't be negative, hence positive) i.e. strictly increasing, there we can observe most of the observations/data-points from $U[0, 1]$ to get mapped via inverse-transform method.

(Intuitively if we see, at places where slope rises the graph tends to cover a considerable part of the y-axis corresponding to a smaller region/interval on x-axis which in a sense maps a large chunk of y-axis to a small chunk in x-axis and this is why during inverse transform we get a hump or an accumulation of transformed data points.)

Question 7

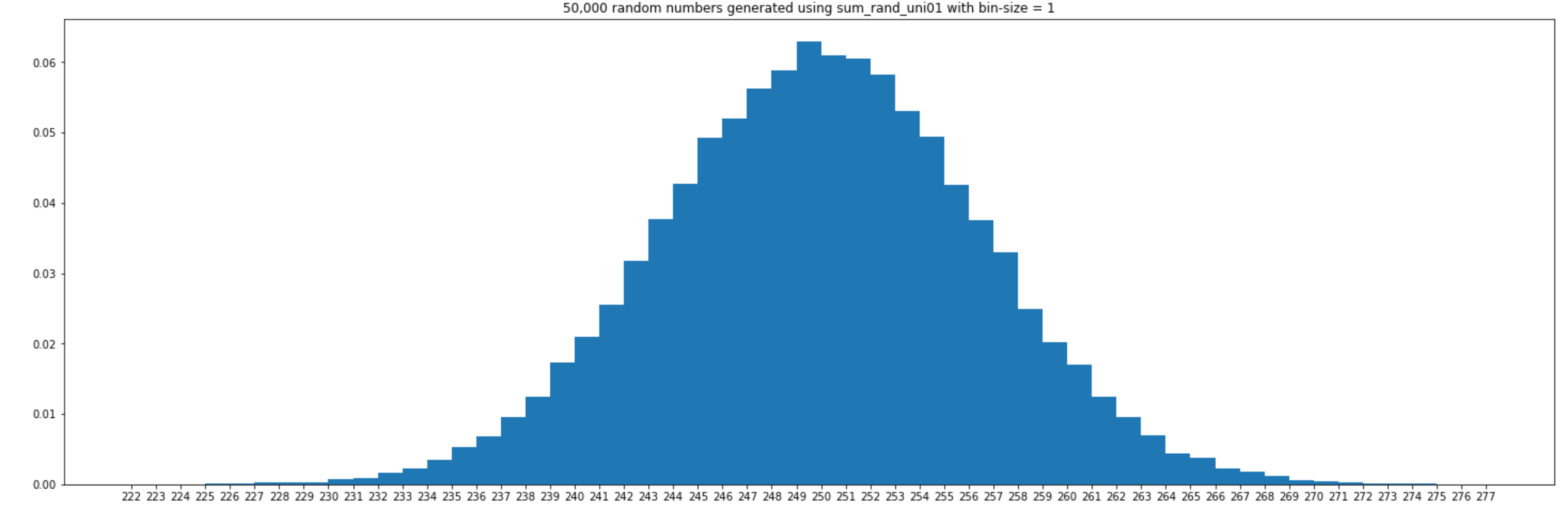
Write a function to generate a random number as follows: Every time the function is called, it generates 500 new random numbers from $U[0, 1]$ and outputs their sum.

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?**

In [20]:
```python
# random number generator
def sum_rand_uni01(seed):
    np.random.seed(seed)
    return np.sum(np.random.uniform(0, 1, 500))

# Generate 50,000 such numbers by calling the above function.
rand_lst = [sum_rand_uni01(i) for i in range(50000)]
bins = [j for j in range(np.floor(min(rand_lst)).astype(np.int32), np.ceil(max(rand_lst)).astype(np.i
nt32))]
```

In [21]:
```python
plt.figure(figsize=(25, 8))
# plot historam with these numbers
_, _, _ = plt.hist(rand_lst, bins = bins, density = True)
plt.xticks(bins); plt.title("50,000 random numbers generated using sum_rand_uni01 with bin-size = 1"
); plt.show()
```



Observations regarding above HISTOGRAM —

The shape of the resulting Histogram is mostly **BELL-Shaped curve or Gaussian curve** which can be credited to the **Central Limit Theorem** where we just rid of $\bar{X}$ by $N \times \bar{X}$.

Intuitive Statement — *suppose that a sample is obtained containing many observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic mean of the observed values is computed. If this procedure is performed many times, the* ***central limit theorem*** *says that the probability distribution of the average will closely approximate a normal distribution.*

In our case:\ Say, $X_1, X_2, X_3, \ldots, X_{500}$ are generated for $call1 =$ Call no. 1 and we sum them up and return those values. Denote: $\bar{X}_{500} = \frac{1}{500}\sum_{i=1}^{500} X_i$. Then $(500\bar{X}_{500}^{call\_1}, 500\bar{X}_{500}^{call\_2}, \ldots, 500\bar{X}_{500}^{call\_50{,}000})$ are generated for $50,000$ calls of function.

From **CLT**,\

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N(0, \sigma^2)$$

, where $X_1, X_2, X_3, \ldots, X_n$ are n random samples drawn from a population with overall mean μ and finite variance σ^2, and $\bar{X}_n$ is the sample mean $= \frac{1}{n}\sum_{i=1}^{n} X_i$

In our case: $n = 1, 2, \ldots, 500$

Now consider $n = 500$ to be large enough that:

$$\begin{aligned}
\left(\bar{X}_n - \mu\right) &\overset{approx}{\sim} N\left(\frac{0}{n}, \frac{\sigma^2}{n}\right) \\
\bar{X}_n &\sim N\left(\mu + \frac{0}{n}, \frac{\sigma^2}{n}\right) \\
500\bar{X}_{500} &\sim N\left(500\mu, \frac{500^2\sigma^2}{500}\right) \\
500\bar{X}_{500} &\sim N\left(250, \frac{500}{12}\right), \\
&\text{where } \mu = 1/2 \text{ as } X_i \text{ are generated from } U[0,1] \text{ and } \sigma^2 = (b-a)^2/12 = (1-0)^2/12 = 1/12.
\end{aligned}$$

That's why we get the peak of our histogram around 250 with variance $\dfrac{500}{12}$.