Setup

In [ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[20 Points] Task 01 - Functions of Gaussian Random Variable

Gaussian/Normal random variables are the most common type of random variables. They are characterized by a mean μ and a standard deviation σ. The probability density function of a Gaussian/Normal random variable is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The standard normal distribution is the distribution of the Normal random variable with mean 0 and standard deviation 1 which is denoted by $X \sim \mathcal{N}(0, 1)$. In this task, you are required to comment on the PDF shape of different functions of a Gaussian random variables.

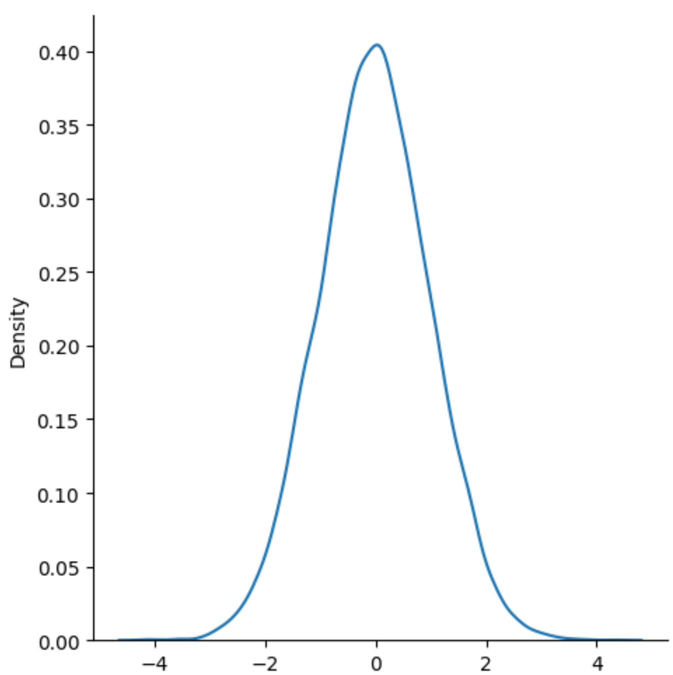
The following code generates 10,000 samples of X random variable and plots an approximate PDF of X (based on its 10,000 samples) using the displot() function of Seaborn library.

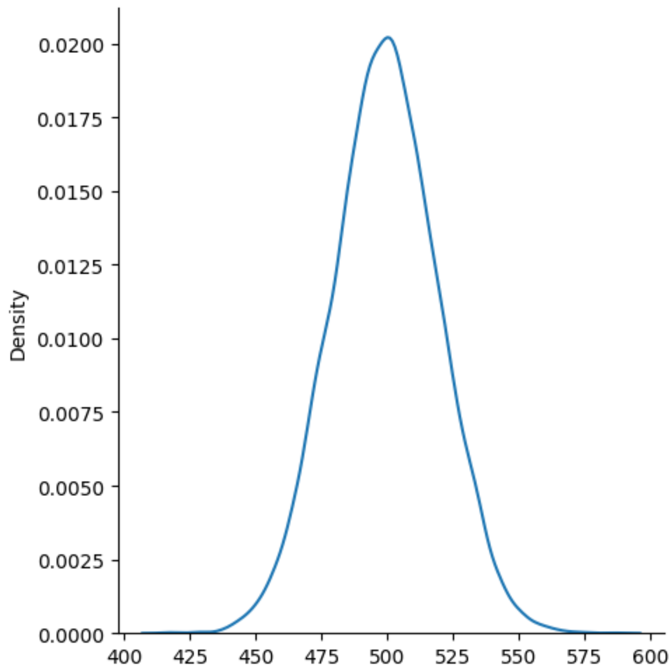
Modify the code to also plot the approximate PDF of $Y = 20X + 500$ and $Z = X^2 + X + 500$.

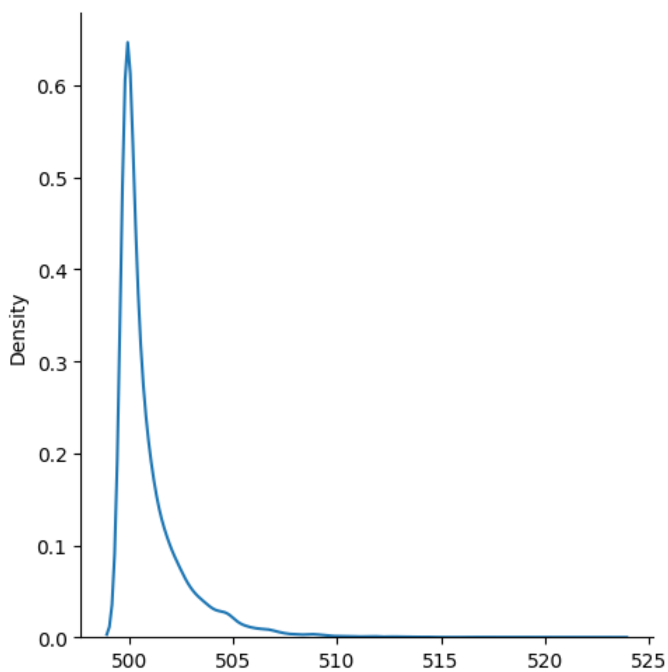
Comment on the PDF shapes of Y and Z, compared to X.

In [ ]:
```
# Ali Muhammad Asad aa07190

sample_size = 10000

x_sample = np.random.normal(0, 1, sample_size)

sns.displot(data=x_sample,kind="kde")

## For Y = 20X + 500 ##
Y = 20*x_sample + 500
sns.displot(data = Y, kind = "kde")

## For Z = X^2 + X + 500 ##
Z = x_sample**2 + x_sample + 500
sns.displot(data = Z, kind = "kde")
```

Out[ ]: <seaborn.axisgrid.FacetGrid at 0x12b682ce7d0>







Ali Muhammad Asad aa07190 Comments

- The PDF of Y is a Normal Distribution of Mean of 500 and Variance of 400 times of X. Since Y is a linear function of X, the mean of Y can be calculated using Y = 20(E[X]) + 500. Since E[X] = 0, the Mean of Y is 500. Similarly the Variance can be calculated by the formula Var(Y) = 20^2 Var(X). Since Var(X) = 1, Var(Y) = 400. Therefore, the mean should be shifted to the right at 500 where we should observe the peak of the PDF of Y, and since we have a higher variance, we should have more dispersed data than X thus giving a smaller height as there is greater variability in the data. This can be observed in the PDF of Y; we have a lesser height but more a greater range so data is more spread out indicating a higher variance. The peak also occurs at 500 indicating the mean value.
- The PDF of Z although doesn't seem like a Normal Distribution at first glance, but the PDF of an Exponential Random Variable, however, it is still a Normal Distribution only positively skewed a lot. This shape occurs due to Z taking a non-linear input of X; the quadratic X^2 term due to which it follows a non-linear pattern. The skewed pattern of the PDF also shows that more values are found towards the right side of the mean indicating that there is a greater likelihood of observing larger values of Z as compared to X and lower values of Z. This again is a result of the quadratic term X^2 in Z due to which there should be more values that are positive and towards the right side of the mean as $X^2 > X$ for more values than $X^2 < X$. This also results in the non-symmetrical distribution than a regular Normal Distribution and results in a skewed like shape.

[30 points] Task 02 - Central Limit Theorem

Suppose $X_1, X_2, \cdots, X_n$ are independent random variables with the same underlying distribution. In this case, we say that the X_i are independent and identically distributed or i.i.d. In particular, the X_i, all have the same mean μ and standard deviation σ.

Let S_n be the sum of n i.i.d random variables:

$$S_n = X_1 + X_2 + \cdots + X_n$$

[10 points] Part A: Uniform Random Variable

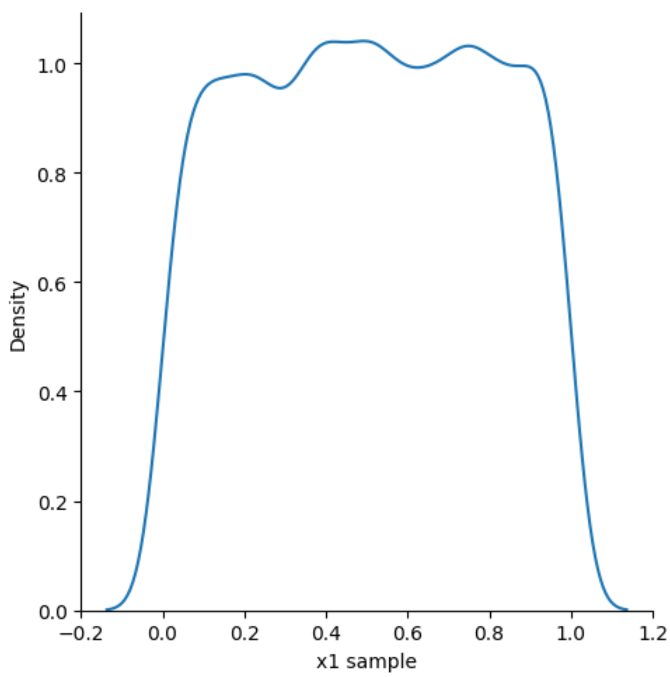
In this part, we consider X_is to be continuous uniform random variables. The following code generates 10,000 samples of X_1 random variable and plots an approximate PDF of X_1 (based on its 10,000 samples) using the displot() function of Seaborn library.

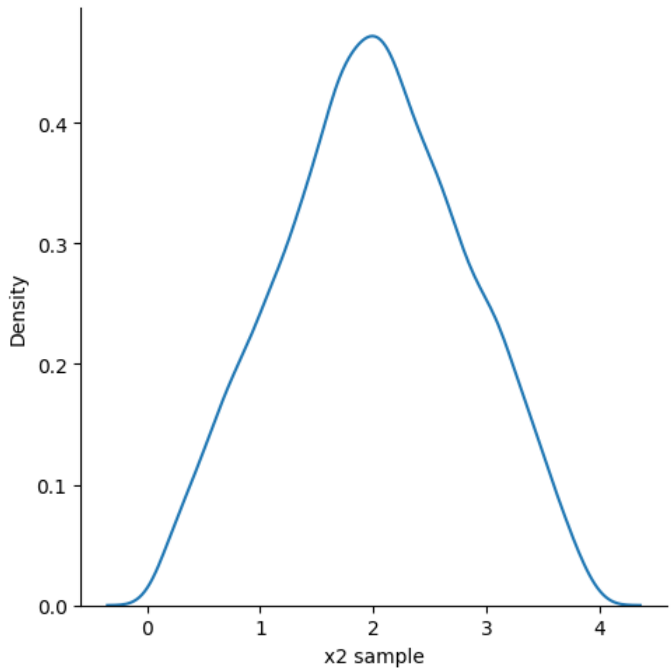
Modify the given code to generate 10,000 samples of $X_2, X_3, \cdots, X_n$, and plot the approximate PDF of S_n using the displot() function of Seaborn library for the following values of n: 1,2,3,5,10,50,100.

Comment on the evolution of PDF shape of S_n as n increases.

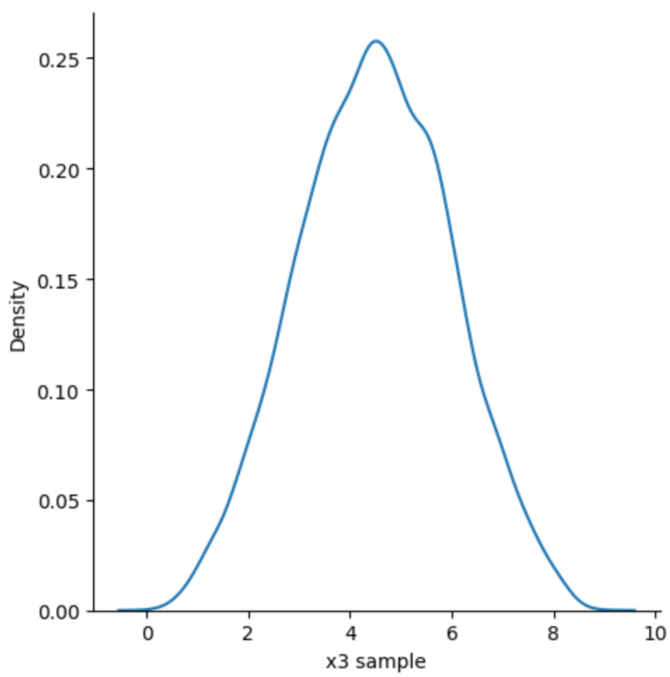
In [ ]:

```
# Ali Muhammad Asad aa07190

sample_size = 10000

df = pd.DataFrame()

n = [1, 2, 3, 5, 10, 50, 100]
for i in n:  # iterating over each element in n
    x_sample = 0 # setting the x_sample to 0 initially
    for x in range(1, i + 1): # iterating from one, all the way to i value to get X1, X2, X3, ..., Xn values
        x_sample += np.random.uniform(0, i, sample_size) # Summing up all the values of X1, X2, X3, ..., Xn
        col = f'x{i} sample'
        df[col] = x_sample
    sns.displot(data=df[col], kind="kde") # Plotting the pdf of Sn where Sn = X1 + X2 + X3 + ... + Xn
```

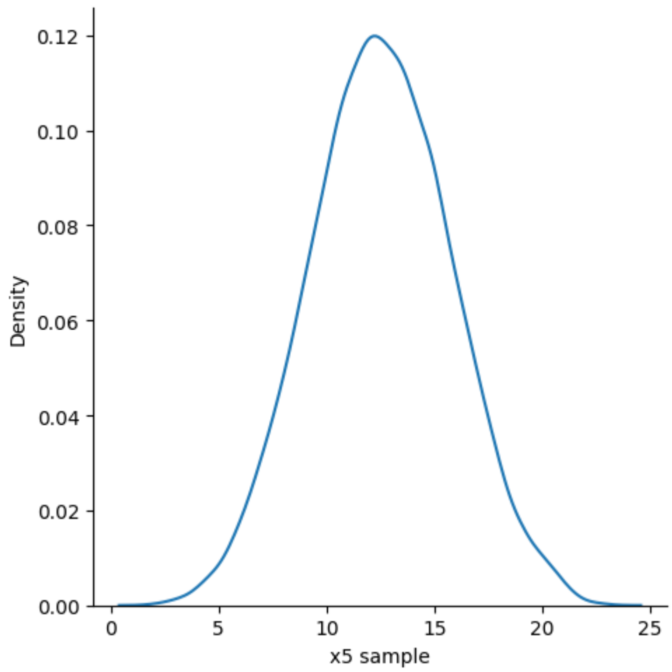




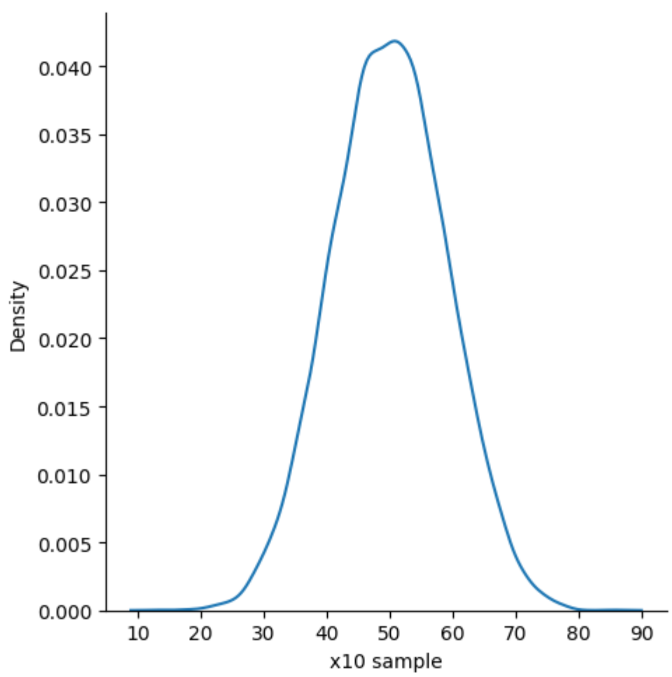
Density
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
x2 sample



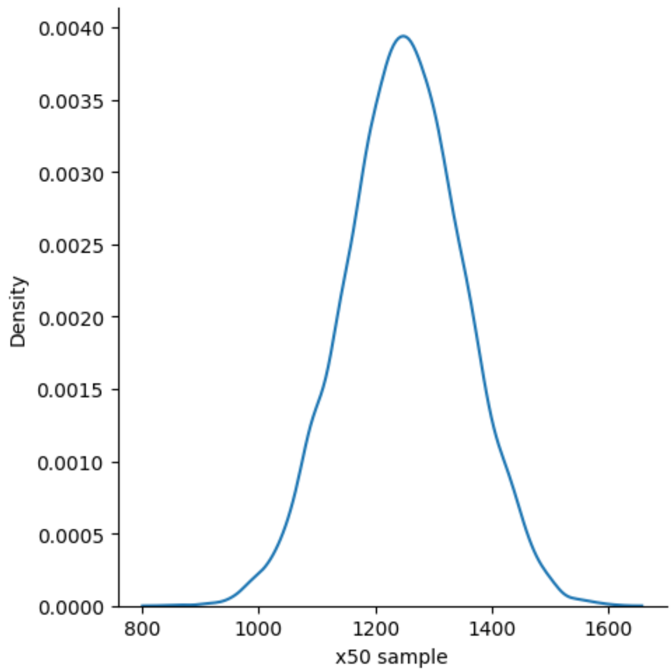
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
x3 sample

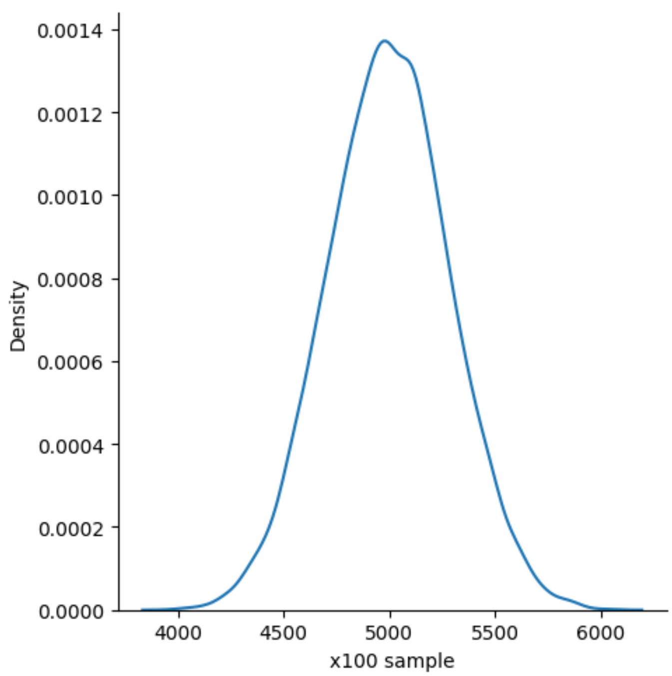


0.12
0.10
0.08
0.06
0.04
0.02
0.00
Density
0
5
10
15
20
25
x5 sample



0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
10
20
30
40
50
60
70
80
90
x10 sample





Ali Muhammad Asad aa07190

Comments

At first when $n = 1$, we have a curve that is sort of flat on top which shows the uniform random variable as S_n only had X_1 first. But as we increase the values of n, we see the the curve shape gradually turns to the bell curve shape or the shape of the PDF of a Normal Random Variable - this is clearly visible in the $x100$ sample. This is in accordance with the **Central Limit Theorem** which states that the sum of a large number of independent, identically distributed random variables approaches a normal distribution regardless of the individual random variables. Further, as n increases, the mean shifts towards the right, which is as expected since we have larger number of random variables, however, the height decreases and the variance increases so as to keep the area under the curve to be equal to 1 so that the PDF remains valid.

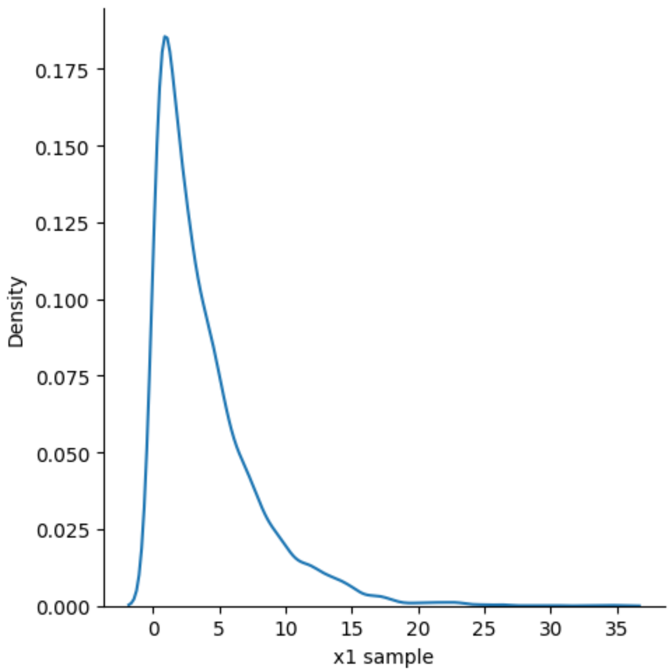
[10 points] Part B: Exponential Random Variable

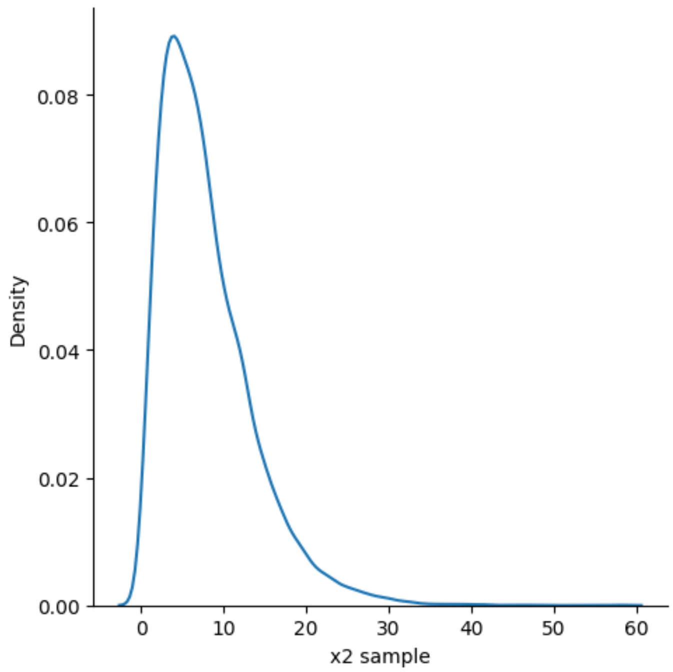
In this part, we consider X_is to be exponential random variables. The following code generates 10,000 samples of X_1 random variable and plots an approximate PDF of X_1 (based on its 10,000 samples) using the displot() function of Seaborn library.

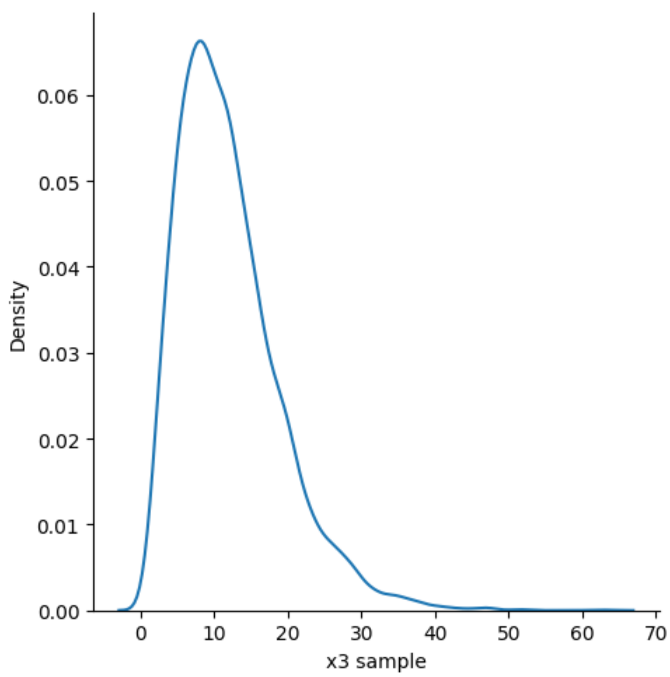
Modify the given code to generate 10,000 samples of $X_2, X_3, \cdots, X_n$, and plot the approximate PDF of S_n using the displot() function of Seaborn library for the following values of n: 1,2,3,5,10,50,100.

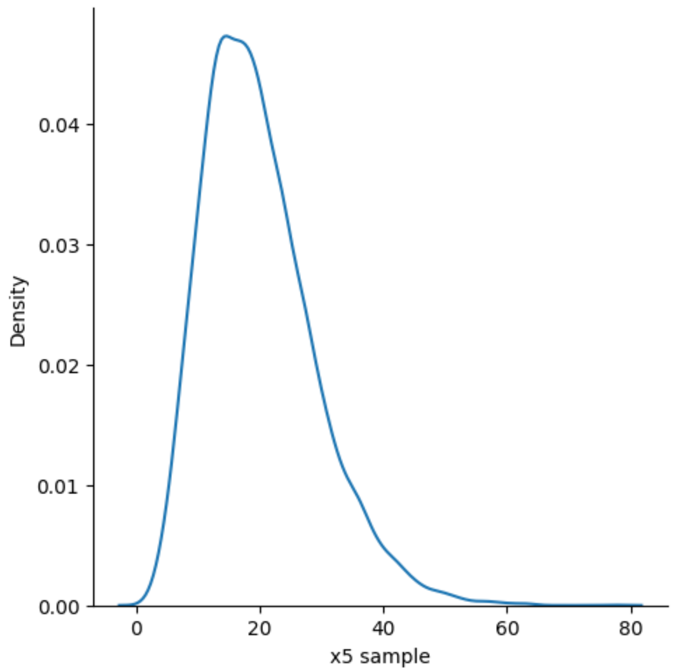
Comment on the evolution of PDF shape of S_n as n increases.

In [ ]:
```
# Ali Muhammad Asad aa07190

sample_size = 10000

rate = 0.25 #Rate of exponential random variables

df = pd.DataFrame()

n = [1, 2, 3, 5, 10, 50, 100]

for i in n:
    x_sample = 0
    for x in range(1, i + 1):
        x_sample += np.random.exponential((1/rate),sample_size)
        col = f'x{i} sample'
        df[col] = x_sample
    sns.displot(data=df[col], kind="kde")
```

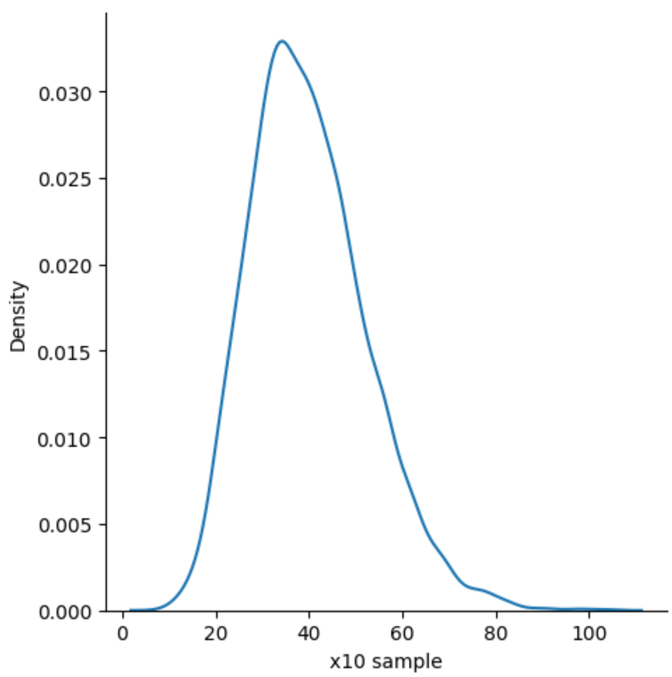



0.08
0.06
0.04
0.02
0.00
Density
0
10
20
30
40
50
60
x2 sample

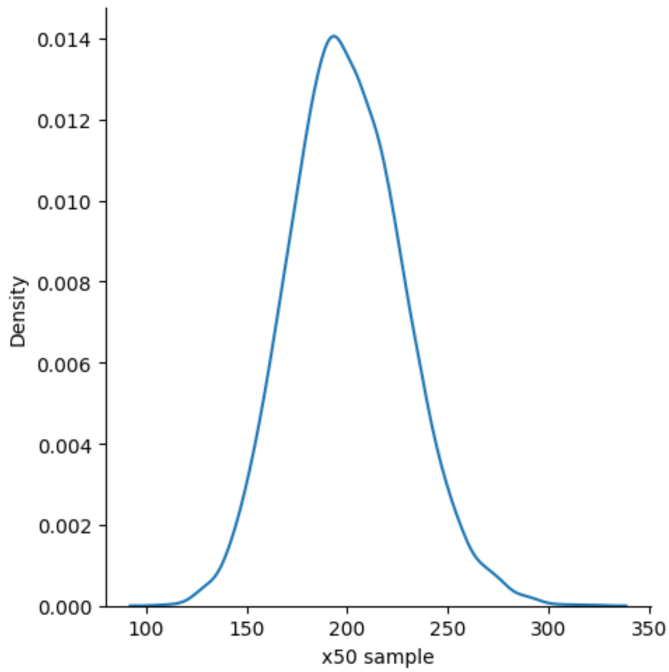

0.06
0.05
0.04
0.03
0.02
0.01
0.00
Density
0
10
20
30
40
50
60
70
x3 sample

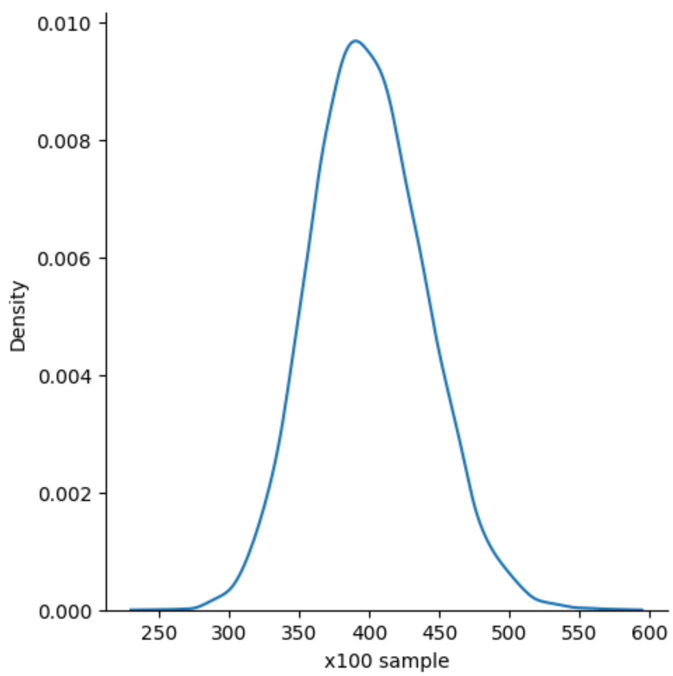


Density
0.04
0.03
0.02
0.01
0.00
0
20
40
60
80
x5 sample



Density
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
x10 sample





Ali Muhammad Asad aa07190

Comments

At first when $n = 1$, the shape of the PDF S_n is that which is expected of an Exponential Random Variable to have. However, as we increase the value of n, we observe that the PDF S_n gradually becomes similar to the bell shaped curve - the Normal Distribution. Moreover, as n increases, the mean shifts towards the right gradually, the height decreases and the variance increases so as to keep the area under the curve equal to 1 so that the PDF remains valid. Again this is in accordance with the **Central Limit Theorem** which states that the sum of a large number of independent, identically distributed random variables approaches a normal distribution regardless of the individual random variables.

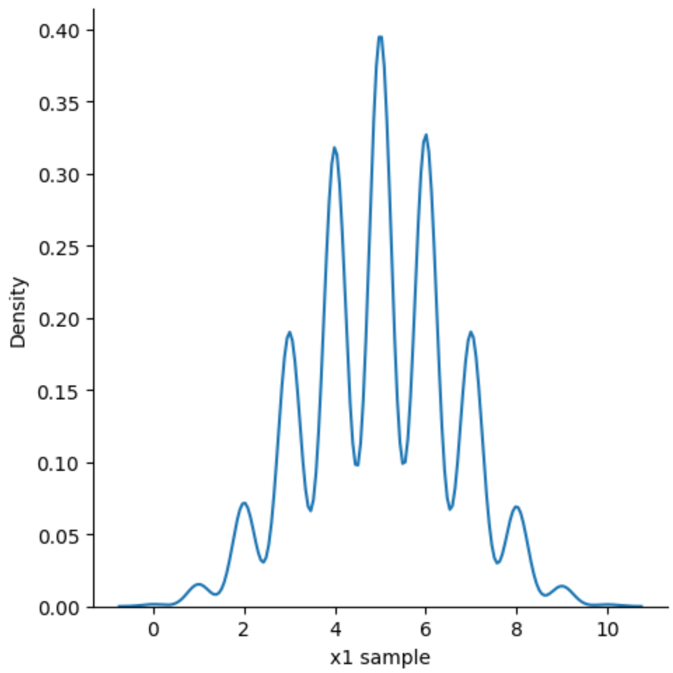
[10 points] Part C: Binomial Random Variable

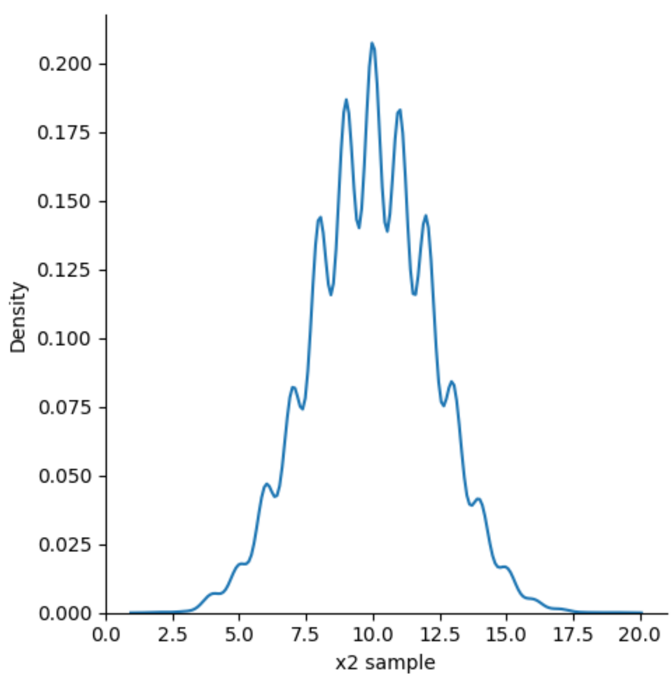
In this part, we consider X_is to be binomial random variables. The following code generates 10,000 samples of X_1 random variable and plots an approximate PDF of X_1 (based on its 10,000 samples) using the displot() function of Seaborn library.

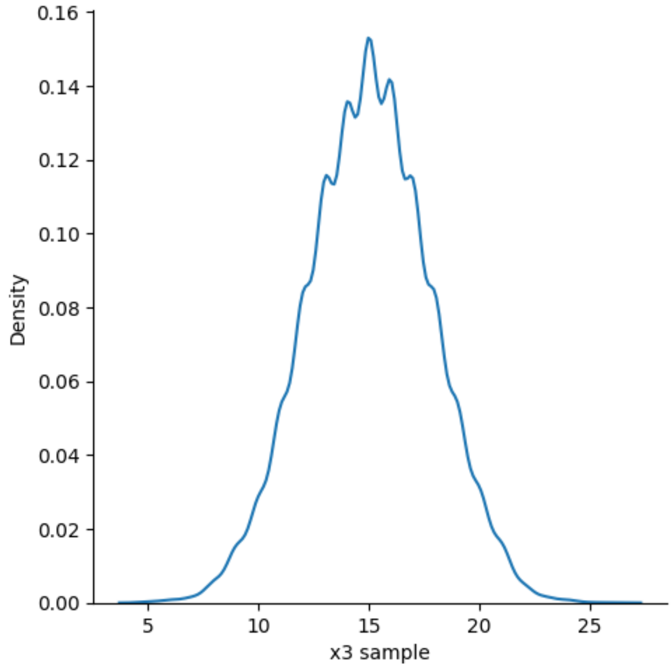
Modify the given code to generate 10,000 samples of $X_2, X_3, \cdots, X_n$, and plot the approximate PDF of S_n using the displot() function of Seaborn library for the following values of n: 1,2,3,5,10,50,100.

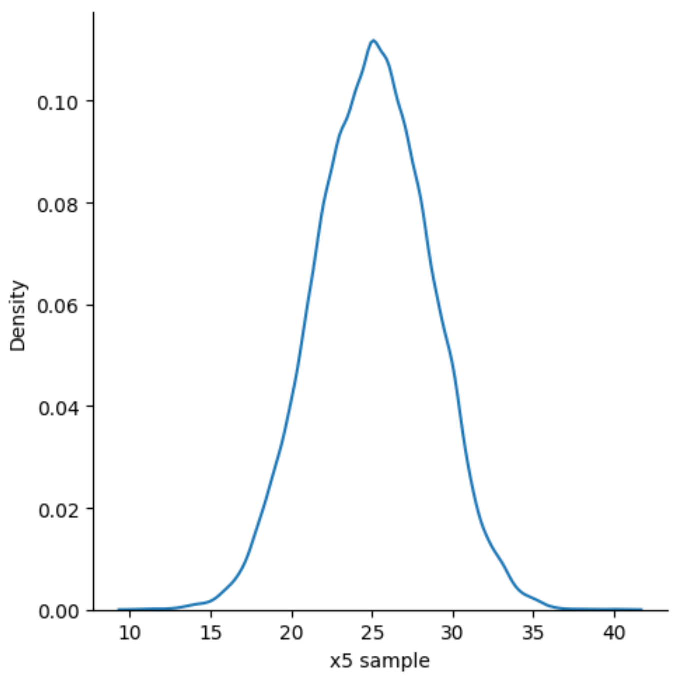
Comment on the evolution of PDF shape of S_n as n increases.

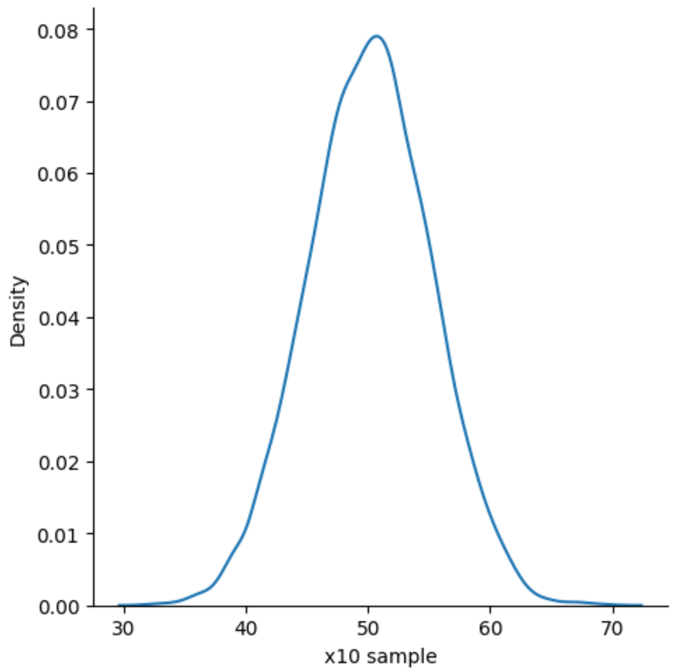
In [ ]:
```
# Ali Muhammad Asad aa07190

sample_size = 10000

df = pd.DataFrame()

n = [1, 2, 3, 5, 10, 50, 100]
for i in n:
    x_sample = 0
    for x in range(1, i + 1):
        x_sample += np.random.binomial(10,0.5,size=sample_size)
        col = f'x{i} sample'
        df[col] = x_sample
    sns.displot(data=df[col], kind="kde")
```

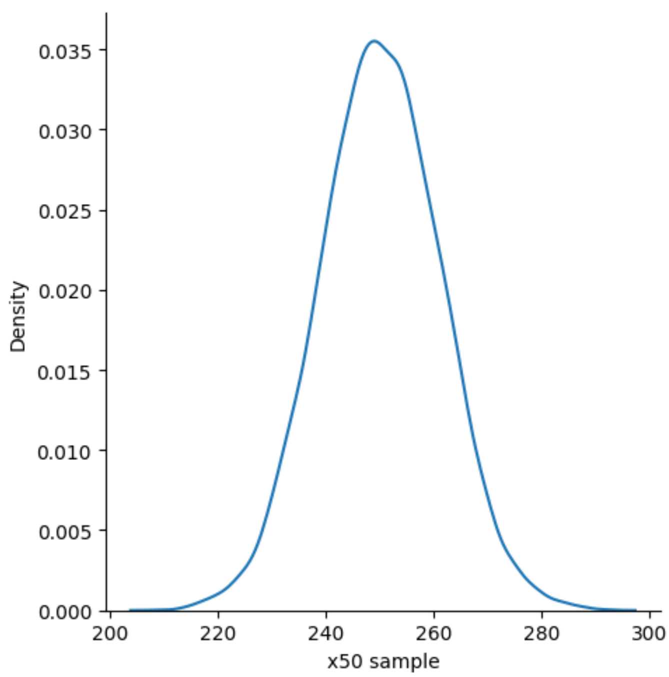


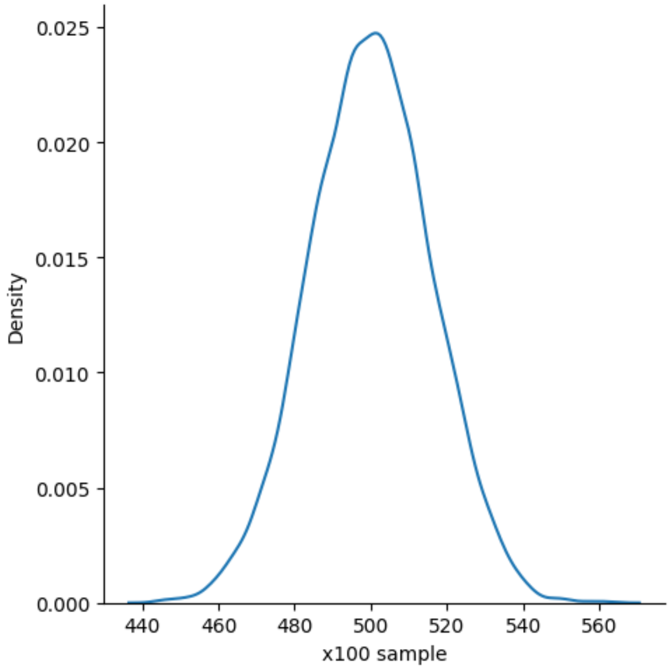



0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Density
5
10
15
20
25
x3 sample


0.10
0.08
0.06
0.04
0.02
0.00
Density
10
15
20
25
30
35
40
x5 sample


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Density
30
40
50
60
70
x10 sample


0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
200
220
240
260
280
300
x50 sample



Ali Muhammad Asad aa07190

Comments

At first when $n = 1$, we have fewer peaks, and the PDF S_n is as which should be of a binomial distribution with x number of states. However, as we increase n, we observe that the number of peaks increase, but also move closer to each other, and at large values of n, the PDF S_n resembles the bell shaped curve - the Normal Distribution. Moreover, the mean shifts towards the right, the peaks converge to a single peak, the height decreases and the variance increases so as to keep the area under the curve to be 1 so that the PDF remains valid. Again this is in accordance with the **Central Limit Theorem** which states that the sum of a large number of independent, identically distributed random variables approaches a normal distribution regardless of the individual random variables.