Central limit theorem

Formula

- The central limit theorem (CLT) suggests that when calculating sample means from some probability distribution, the sample means are approximately normally distributed.

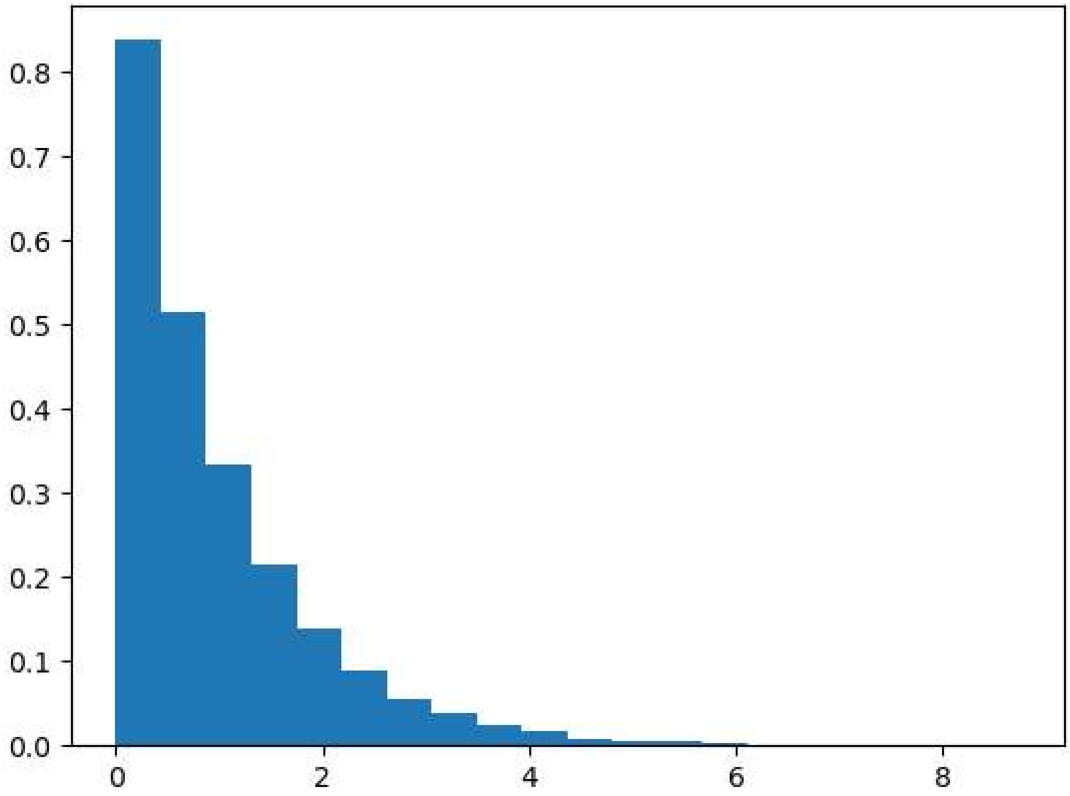
- Suppose we have a k amount of samples, represented by a sequence of independent and identically (iid) distributed random variables $X_1, X_2, \ldots, X_n$, each with mean μ and standard deviation σ.
- According to the CLT, as the sample size n and/or the amount of samples k increases, the distribution of sample means $\bar{X}$ approaches a normal distribution. Thus, for a large enough sample size $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

- In words, the sampling distribution $\bar{X}$ follows a normal distribution with mean μ and standard deviation $\frac{\sigma^2}{n}$, regardless of the underlying distribution of the X_i variables.

Simulation

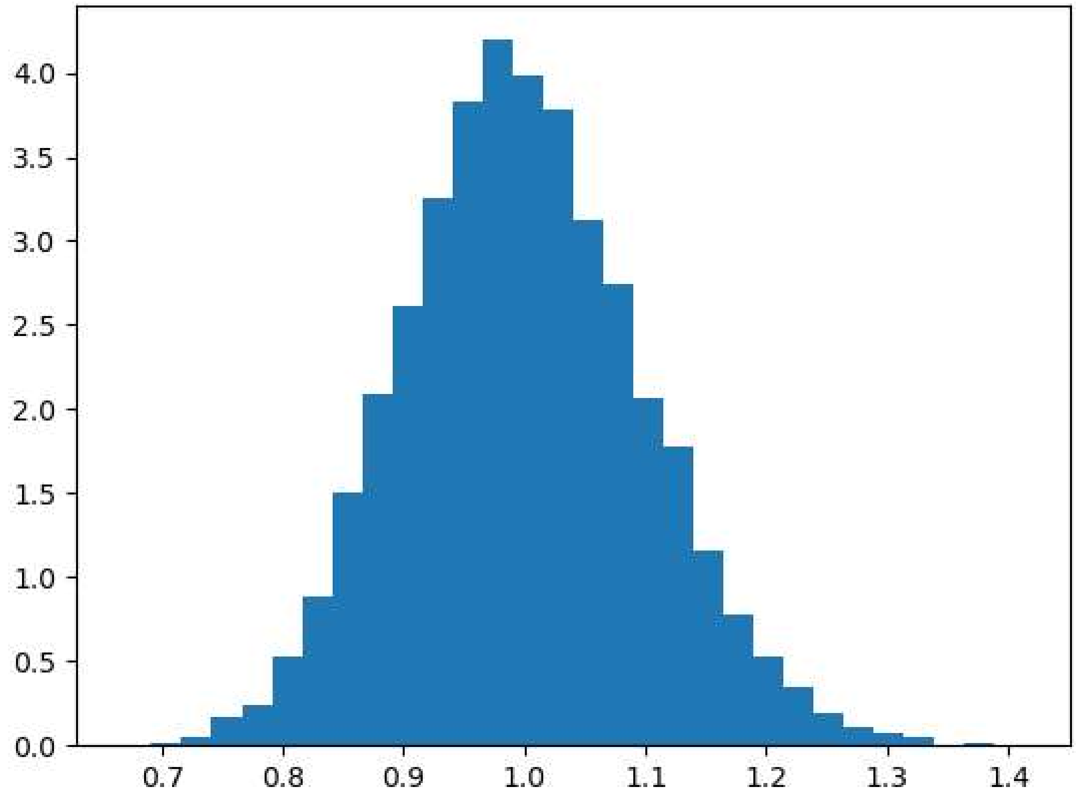
In [1]:
```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
```

In [2]:
```
# Let's draw a sample from the exponential function
numbers = np.random.exponential(1, 10000)

count, bins, ignored = plt.hist(numbers, 20, density = True)
plt.show()
```



In [3]:
```
# Now, let's repeatedly draw samples of size n, and calculate their means. The results are stored in a list.
sample_means = []
sample_size = 100

for i in range(10000):
    sample = np.random.exponential(1, sample_size)
    sample_mean = np.mean(sample)
    sample_means.append(sample_mean)
```

In [4]:
```
count, bins, ignored = plt.hist(sample_means, 30, density = True)
plt.show()
```



- The sample means are approximately normally distributed.

The mean μ and variance σ^2 of the exponential distribution vs the sampling distribution

In [5]:
```
# What we can expected based on the CLT
print("Mean of the original distribution:", round(np.mean(numbers),3))
print("Variance of the original distribution:", round(np.var(numbers)/sample_size,3))
```

```
Mean of the original distribution: 0.985
Variance of the original distribution: 0.01
```

In [6]:
```python
# What is actually obtained
print("Actual mean of the sampling distribution:", round(np.mean(sample_means), 3))
print("Actual standard deviation of the sampling distribution:", round(np.var(sample_means),3))
```

```
Actual mean of the sampling distribution: 0.999
Actual standard deviation of the sampling distribution: 0.01
```

- The mean and variance of the sampling distribution are very close to what the CLT suggests.