In [ ]:

Mehedi Azad
Data Analyst | Data Science | ML | Statistics | Mathematics | Power BI | Tableau | Pandas | Numpy | SQL | Python |
https://www.linkedin.com/in/mehediazad/

The central limit theorem

The central limit theorem states that if you take sufficiently large samples from a population, the samples' means will be normally distributed, even if the population isn't normally distributed.

In [ ]:

- What is the central limit theorem?

The central limit theorem relies on the concept of a `sampling distribution`, which is the `robability distribution` of a statistic for a large number of samples taken from a population.

Imagining an experiment may help you to understand sampling distributions:

- Suppose that you draw a random sample from a population and calculate a statistic for the sample, such as the mean.
- Now you draw another random sample of the same size, and again calculate the mean.
- You repeat this process many times, and end up with a large number of means, one for each sample.

The distribution of the sample means is an example of a sampling distribution.

In [ ]:

- The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough.
- Regardless of whether the population has a `normal`, `Poisson`, `binomial`, or `any other distribution`, the sampling distribution of the `mean` will be `normal`.
- A normal distribution is a `symmetrical`, `bell-shaped` distribution, with increasingly fewer observations the further from the `center of the distribution`.

`Central limit theorem` formula

Fortunately, you don't need to actually repeatedly sample a population to know the shape of the sampling distribution.

The parameters of the sampling distribution of the mean are determined by the parameters of the population:

- The `mean` of the `sampling distribution` is the `mean` of the `population`.

$$\mu_{\bar{x}} = \mu$$

- The `standard deviation` of the `sampling distribution` is the `standard deviation` of the `population` divided by the `square root` of the `sample size`.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- We can describe the sampling distribution of the mean using this notation:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Where:

- $\bar{X}$ is the sampling distribution of the sample means
- ~ means "follows the distribution"
- N is the normal distribution
- μ is the mean of the population
- σ is the standard deviation of the population
- n is the sample size

Sample size and the central limit theorem

- The `sample size (n)` is the `number of observations` drawn from the `population` for `each sample`.
- The sample size is the same for all samples.

The sample size affects the sampling distribution of the mean in two ways.

1. Sample size and normality

- The larger the sample size, the more closely the sampling distribution will follow a normal distribution.
- When the sample size is small, the sampling distribution of the mean is sometimes non-normal. That's because the central limit theorem only holds true when the sample size is "sufficiently large."
- By convention, we consider a sample size of 30 to be "sufficiently large."

- **When n < 30**, the central limit theorem doesn't apply.
  - The sampling distribution will follow a similar distribution to the population.
  - Therefore, the sampling distribution will only be normal if the population is normal.

- **When n ≥ 30,**, the central limit theorem applies.
  - The sampling distribution will approximately follow a normal distribution.

2. Sample size and standard deviations

- The sample size affects the standard deviation of the sampling distribution.
- Standard deviation is a measure of the variability or spread of the distribution (i.e., how wide or narrow it is).

- **When n is low**, the standard deviation is high.
  - There's a lot of spread in the samples' means because they aren't precise estimates of the population's mean.

- **When n is high**, the standard deviation is low.
  - There's not much spread in the samples' means because they're precise estimates of the population's mean.

Conditions of the central limit theorem

The central limit theorem states that the sampling distribution of the mean will always follow a normal distribution under the following conditions:

- The sample size is sufficiently large. This condition is usually met if the sample size is $n \geq 30$.
- The samples are independent and identically distributed (i.i.d.) random variables. This condition is usually met if the sampling is random.
- The population's distribution has finite variance. Central limit theorem doesn't apply to distributions with infinite variance, such as the Cauchy distribution. Most distributions have

finite variance.

Importance of the central limit theorem

The `central limit theorem` is one of the `most fundamental statistical theorems` . In fact, the “central” in “central limit theorem” refers to the importance of the theorem.

- Note:
  - **Parametric tests**, such as `t tests` , `ANOVAs` , and `linear regression` , have more statistical power than most non-parametric tests.
  - Their power comes from assumptions about populations distributions that are based on the `central limit theorem` .

In [ ]:

Let's Implement some python code to justify the `The Central LimiT Theorem`

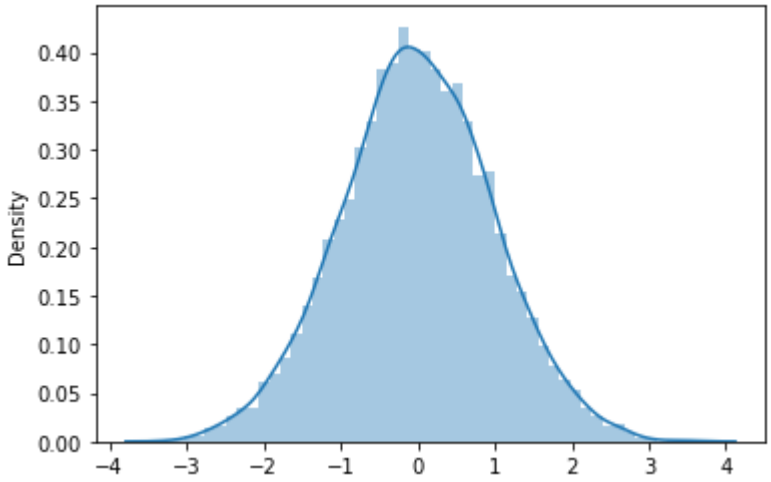
`Load Dependencies`

In [24]:
```
import  numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import statistics as stat
import warnings
warnings.simplefilter("ignore")
```

`Simulating a normally distributed population`

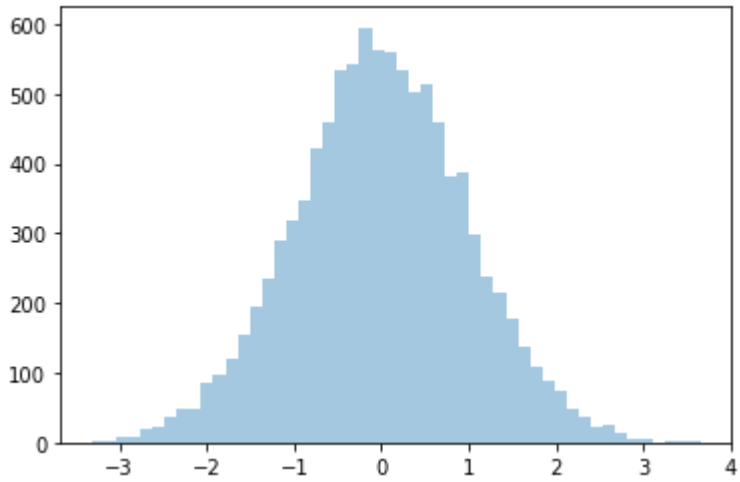
In [31]:
```
x = np.random.normal(size=10000)     #at default parameters, will be 'Standard' r
sns.distplot(x)
```

Out[31]:
```
<AxesSubplot:ylabel='Density'>
```



In [32]:
```
sns.distplot(x, kde=False)
```

Out[32]:
```
<AxesSubplot:>
```



Sampling from the normal distribution population

In [33]:
```
x_sample = np.random.choice(data, size=10, replace=False)
x_sample
```

Out[33]:
```
array([-0.62952575, -1.36314034, -0.67091729, -1.31520801,  0.41616666,
        0.27079118,  1.12566217,  0.26411123, -0.16663416, -0.30051029])
```

In [34]:
```
stat.mean(x_sample)
```

Out[34]:
```
-0.23692045980484663
```

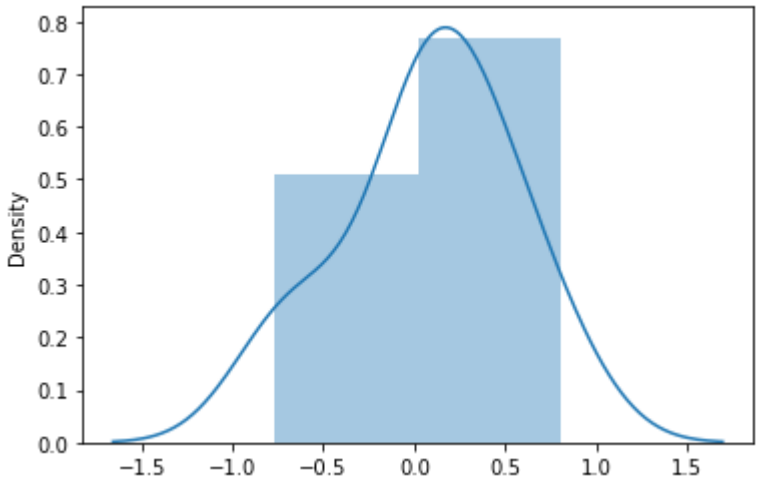
In [35]:
```
def sample_mean_calculator(population_array, sample_size, no_of_samples):
    sample_means = []
    for i in range(no_of_samples):
        sample = np.random.choice(population_array, size=sample_size, replace=Fal
        sample_mean = stat.mean(sample)
        sample_means.append(sample_mean)
    return sample_means
```

In [41]:
```
sample_mean_calculator(x, 10, 10)
```

Out[41]:
```
[-0.059940653732475604,
 0.4829323915612239,
 -0.11266108795811672,
 0.18044666834637957,
 0.15481304818970812,
 -0.38556611787583417,
 -0.5843782878616854,
 -0.18064006810312036,
 -0.33179871430561914,
 -0.12735910083704036]
```

In [42]:
```
sns.distplot(sample_mean_calculator(x, 10, 10))
```

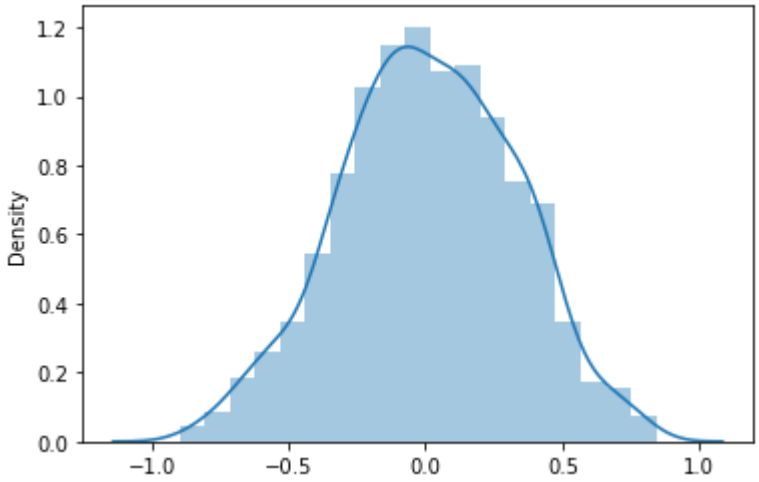
Out[42]:
```
<AxesSubplot:ylabel='Density'>
```



The more samples we take, the more likely the sampling distribution of the means will be normally distributed.

In [43]:
```
sns.distplot(sample_mean_calculator(x, 10, 1000))
```

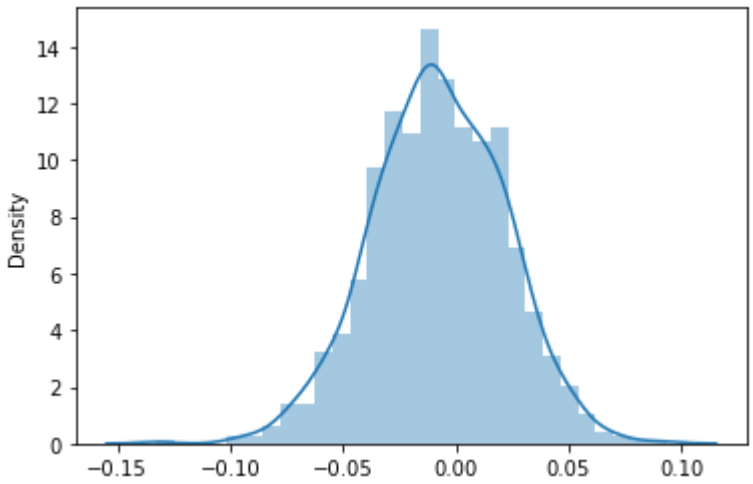
Out[43]: <AxesSubplot:ylabel='Density'>



The larger the sample, the tighter the sample means will tend to be around the population mean.

In [44]:
```
sns.distplot(sample_mean_calculator(x, 1000, 1000))
```

Out[44]: <AxesSubplot:ylabel='Density'>



In [ ]:

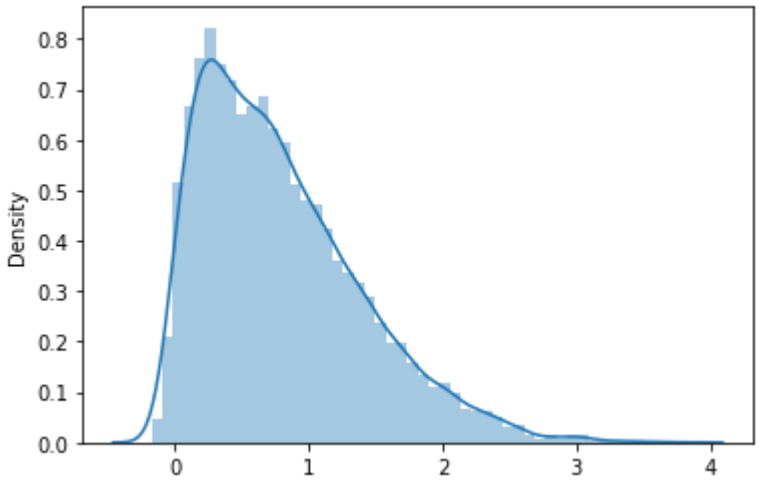
Sampling from a skewed population

In [46]:
```
from scipy.stats import skewnorm
```

In [47]:
```
s = skewnorm.rvs(12, size=10000)
```

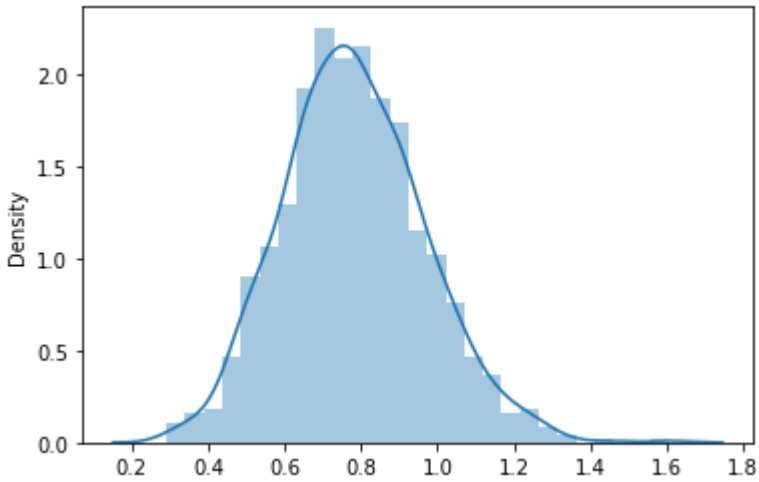
In [50]:
```
sns.distplot(s)
```

Out[50]: <AxesSubplot:ylabel='Density'>



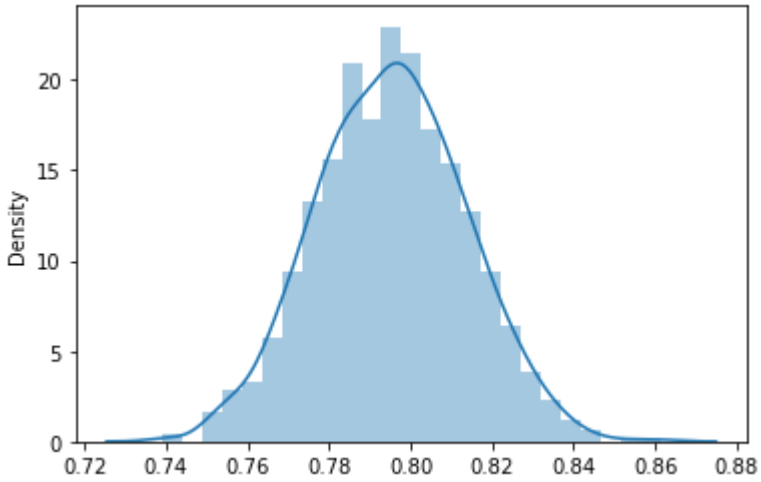
In [53]:
```
sns.distplot(sample_mean_calculator(s, 10, 1000))
```

Out[53]: <AxesSubplot:ylabel='Density'>



In [55]:
```
sns.distplot(sample_mean_calculator(s, 1000, 1000))
```

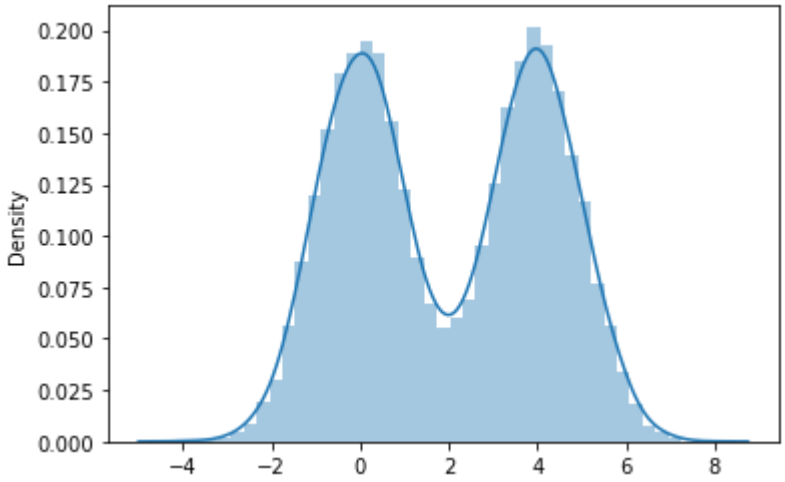
Out[55]: <AxesSubplot:ylabel='Density'>



Sampling from a multimodal distribution

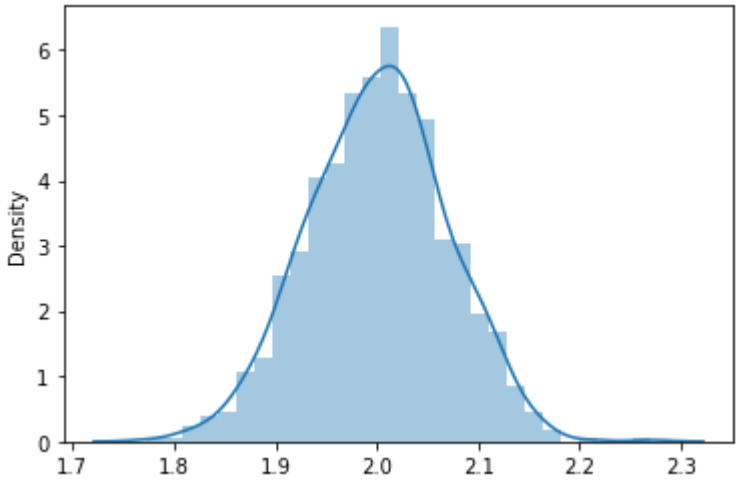
In [58]:
```
m = np.concatenate((np.random.normal(size=10000), np.random.normal(loc=4.0, size=
sns.distplot(m)
```

Out[58]: <AxesSubplot:ylabel='Density'>



In [59]:
```
sns.distplot(sample_mean_calculator(m, 1000, 1000))
```

Out[59]: <AxesSubplot:ylabel='Density'>

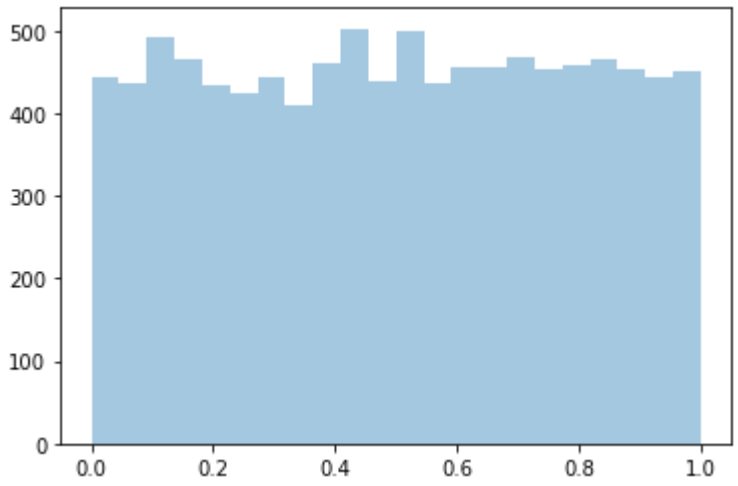


Sampling means from any distribution produces a normal sampling distribution

Even when sampling from a uniformm population

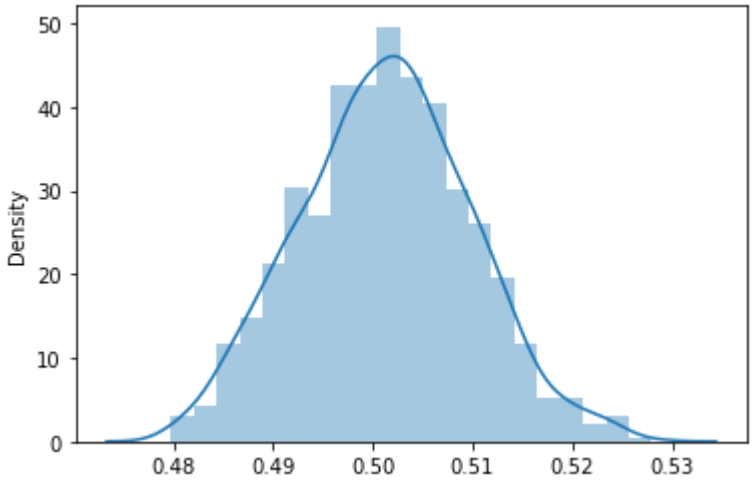
In [61]:
```
u = np.random.uniform(size=10000)
sns.distplot(u, kde=False)
```

Out[61]: <AxesSubplot:>



In [62]:
```
sns.distplot(sample_mean_calculator(u, 1000, 1000))
```

Out[62]:
```
<AxesSubplot:ylabel='Density'>
```



In [ ]:
```
```