Central Limit Theorem

In this lesson, concepts of the central limit theorem are explored.

We'll cover the following

In inferential statistics, operations are performed on a sample of data, and then predictions are made about an entire population. The central limit theorem is considered an important concept in inferential statistics due to its widely used applications.

What is it?

The **central limit theorem** or **CLT** states that:

> ***The mean of a random sample will closely resemble the mean of the whole population as the sample size increases, regardless of the shape of the data distribution.***

The above statement means that whatever the data distribution is, whether it be **uniform**, **binomial**, **normal**, etc., if the sample size or the size of the subset of the data keeps increasing, the average value of the entire population of data can be inferred by taking the average of that sample data. Let's test this theorem with an example to verify its credibility.

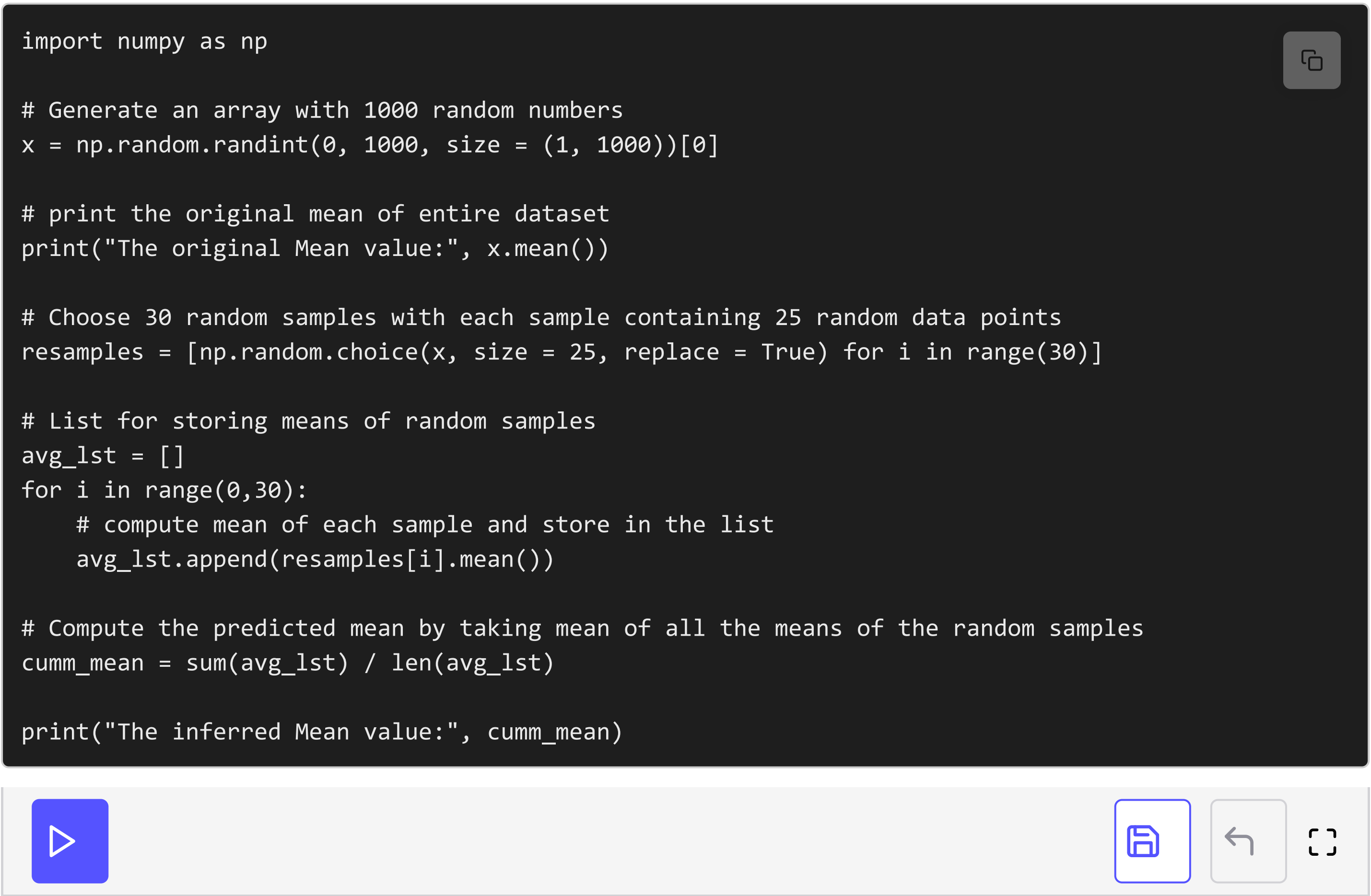
Example

The following steps will be performed in this example:

- An array with **one-thousand** random values will be created.
- Then, the average of this dataset is computed, which is our original **mean**

value. Later it will be compared with the inferred one.

- Then, **thirty** random samples are gathered with each sample containing **twenty-five** data points.
- The average value of each random sample is computed and stored in a list.
- The inferred mean value is calculated by taking an average of the values in the list.



```
import numpy as np

# Generate an array with 1000 random numbers
x = np.random.randint(0, 1000, size = (1, 1000))[0]

# print the original mean of entire dataset
print("The original Mean value:", x.mean())

# Choose 30 random samples with each sample containing 25 random data points
resamples = [np.random.choice(x, size = 25, replace = True) for i in range(30)]

# List for storing means of random samples
avg_lst = []
for i in range(0,30):
    # compute mean of each sample and store in the list
    avg_lst.append(resamples[i].mean())

# Compute the predicted mean by taking mean of all the means of the random samples
cumm_mean = sum(avg_lst) / len(avg_lst)

print("The inferred Mean value:", cumm_mean)
```

After running the above example, we can observe that the difference between the original and inferred value is not very large. This confirms the above **central limit theorem**. If the sample size keeps increasing the two mean values will come even closer.

Importance

Now imagine that we have a dataset with **ten million**+ data points. Performing analysis on this large dataset would consume a lot of resources. However, using the central limit theorem, a small random sample can be selected to do the required analysis and results can then be generalized for the entire data population. This technique is very famous in the financial sector, and we’ll also be using this in one of our projects later in the course.

This marks the end of the statistics for data analysis chapter. The next chapter discusses different data manipulation techniques using the `NumPy` and `pandas` packages.