Connect, Message, Like,Follow&Share, 100% Free Counselling➔ThankYou

Follow me on Linkedin

Statistical Inference

1. Sampling Methods

To reduce the maximum permissible error, confidence level, and population variance/ standard deviation. maximum permissible error is defined as the difference between actual output and predicted output.

Confidence Level is defined as the probability that the value of a parameter falls within a specified range of values.

Population Variance is defined as the value of variance that is calculated from population data.

The determination of sample size depends on three major factors, such as:

1. confidence level

2. maximum permissible error

3.population variance/standard deviation

The formula given below can be used to calculate sample size:

$$n = \left[\frac{Z_{critical} * \sigma}{E}\right]^2$$

Where,

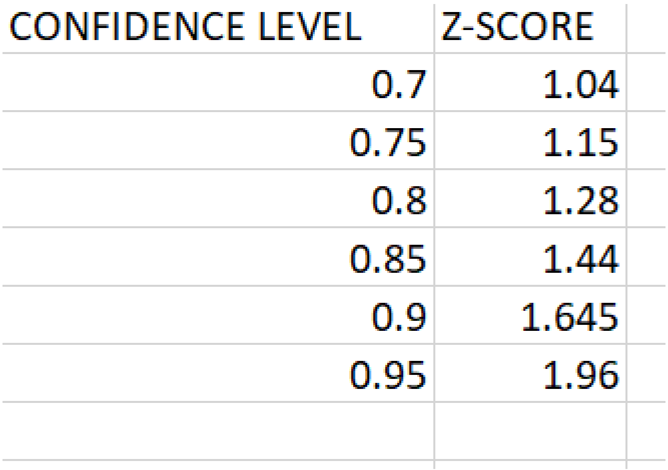
n = sample size

critical = critical Z statistic for the specified confidence level

σ = population standard deviation

E = maximum permissible error

The critical Z value for a specified confidence level is found in the Z table. Z table for various confidence levels is given below.



CONFIDENCE LEVEL	Z-SCORE
0.7	1.04
0.75	1.15
0.8	1.28
0.85	1.44
0.9	1.645
0.95	1.96

There are two types of sampling methods

1.1 Random Sampling

1.2. Stratified Sampling

Let's take an example, Suppose 1000 students are present in a class and select 10 students with different characteristics(like marks, etc..) from them.

population=1000

sample =10

1.1. Random Sampling

If we use random sampling, It randomly selects the 10 students(sample) from 1000 students(population).

Drawback

By using Random Sampling, there is a chance to select the same type of students as a sample. It may give a biased result and such a sample is called a Biased Sample.

```
a=[]
for i in range(1,1001):
    a.append(i)
#Importing the NumPy library
import numpy as np
#Choosing 8 Random states from ‘states’ without repetition
np.random.choice(a, size=10, replace=False)
```

output:

```
array([715, 864,  18, 911, 309, 115, 598, 294, 651, 578])
```

To overcome this we can use Stratified Sampling.

1.2. Stratified Sampling

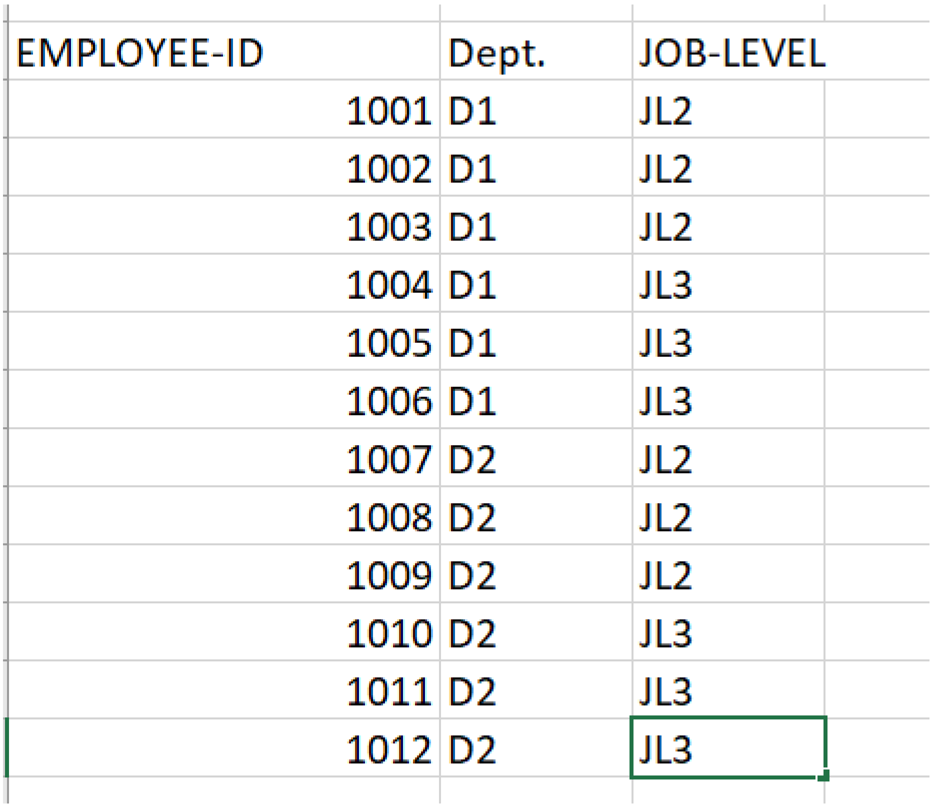
If we use Stratified Sampling, the population is divided into groups based on characteristics. these groups are called Strata.

The sample is chosen randomly from each of these groups.

Suppose you have a list of 12 employees along with their department and job level information

You can sample the data by grouping it based on department and job level.

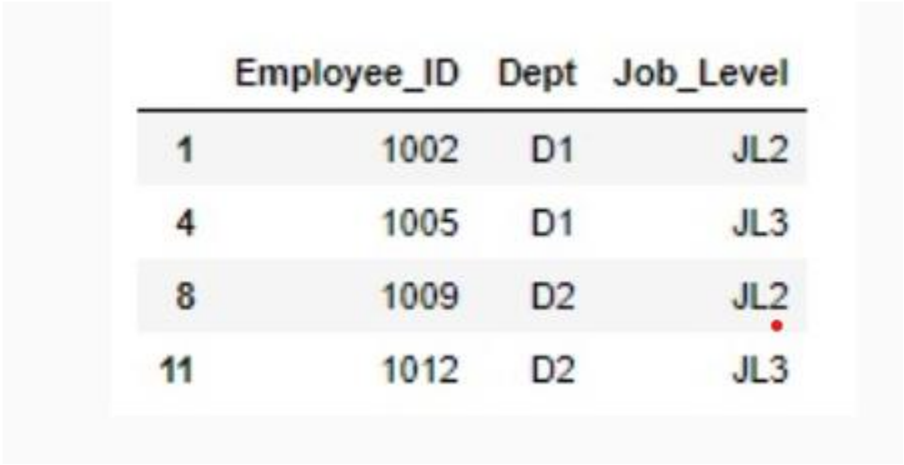
There are two departments (D1 and D2) and two job levels (2 and 3).



EMPLOYEE-ID	Dept.	JOB-LEVEL
1001	D1	JL2
1002	D1	JL2
1003	D1	JL2
1004	D1	JL3
1005	D1	JL3
1006	D1	JL3
1007	D2	JL2
1008	D2	JL2
1009	D2	JL2
1010	D2	JL3
1011	D2	JL3
1012	D2	JL3

```
#Taking a random sample from the population using the groupby function based on Department and Job Level
data.groupby(['Dept','Job_Level'], group_keys=False).apply(lambda x: x.sample(1))
```

output:



	Employee_ID	Dept	Job_Level
1	1002	D1	JL2
4	1005	D1	JL3
8	1009	D2	JL2
11	1012	D2	JL3

You can observe that the output sample has all combinations of ‘Dept’ and ‘Job_Level’.

The above output is the stratified sample.

2. Hypothesis Testing

Why hypothesis testing?

Let’s consider an example, The data scientist has successfully estimated the population means, population variance, and standard deviation using various point estimation techniques.

According to the data scientist, the confidence interval of the population mean is (408 to 417). The client wants to verify this claim.

Therefore, the client randomly chooses 100 students from the country and conducts the surprise test (consisting of the same questions) for them. The mean marks scored by those 100 students are found to be 403, which does not lie in the range specified by the data scientist.

Now the question arises, whether this observation is sufficient for the client to conclude that the estimation by the data scientist is not valid.

In inferential statistics, a Hypothesis Test is conducted to find answers to such questions.

What is hypothesis testing?

In statistics, the hypothesis is a statement about the population and it deals with collecting enough evidence about the hypothesis. Then, based on the evidence collected, the test either accepts or rejects the hypothesis about the population.

Hypothesis testing needs to be performed to find evidence in support of this hypothesis. Based on the evidence found, this hypothesis can be accepted or rejected.

In each hypothesis testing, there are two parameters Null hypothesis and Alternate hypothesis.

Null Hypothesis (H0): It is a statement that rejects the observation based on which the hypothesis is made. You can start the hypothesis testing considering the null hypothesis to be true. It cannot be rejected until there is evidence that suggests otherwise.

Alternate Hypothesis (Ha): It is a statement that is contradictory to the null hypothesis. If you find enough evidence to reject the null hypothesis, then the alternative hypothesis is accepted.

if the probability of occurrence of the given data is less than the level of significance (0.05) you can reject the null hypothesis.

if the probability of occurrence of the given data is greater than or equal to the level of significance (0.05) you cannot reject the null hypothesis.

steps to calculate the Hypothesis:-

Step 1: Let assume the null hypothesis, alternate hypothesis, and the level of significance.

Step 2: Calculate the P-value.

Step 3: Conclude whether to reject the null hypothesis or not based on the P-value i.e.

- If P-value < significance level, then reject the null hypothesis
- If P-value >= significance level, the null hypothesis cannot be rejected

Step 4: State the conclusion.

For the above example,

Step-1

Null hypothesis(H0): The estimate given by the data scientist is correct.

Alternate Hypothesis(Ha): The estimate given by the data scientist is incorrect.

Step-2

calculate P-Value:

```
#Importing the norm function from the scipy.stats module
from scipy.stats import norm
#Finding the probability of getting a value that is 0.56 standard deviation from mean using norm.cdf() function
print(norm.cdf(0.56))
```

output:- 0.712

Step-3

P-Value > level of significance.

So, it is failed to reject the null hypothesis.

Step-4

The estimate given by the data scientist is correct.