



Department of Computer Science and Engineering (Data Science)

Lab Manual

Subject: Foundations of Data Analysis Laboratory (DJ19DSL303)

Semester: III

Experiment 9

(Sampling Methods)

NAME: Dev Patel

SAP ID: 60009200016

BATCH: K/K1

DATE: 25/01/2022

Aim: Perform different sampling methods for creating samples on a given dataset.

Theory:

Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population. Different sampling methods are widely used by researchers in market research so that they do not need to research the entire population to collect actionable insights.

It is also a time-convenient and a cost-effective method and hence forms the basis of any research design. Sampling techniques can be used in a research survey software for optimum derivation.

For example, if a drug manufacturer would like to research the adverse side effects of a drug on the country's population, it is almost impossible to conduct a research study that involves everyone. In this case, the researcher decides a sample of people from each demographic and then researches them, giving him/her indicative feedback on the drug's behaviour.

Types of sampling: sampling methods

Sampling in market research is of two types – probability sampling and non-probability sampling. Let's take a closer look at these two methods of sampling.

1. **Probability sampling:** Probability sampling is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly. All the



Department of Computer Science and Engineering (Data Science)

members have an equal opportunity to be a part of the sample with this selection parameter.

2. Non-probability sampling: In non-probability sampling, the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.

In this blog, we discuss the various probability and non-probability sampling methods that you can implement in any market research study.

Types of probability sampling with examples:

Probability sampling is a sampling technique in which researchers choose samples from a larger population using a method based on the theory of probability. This sampling method considers every member of the population and forms samples based on a fixed process.

For example, in a population of 1000 members, every member will have a $1/1000$ chance of being selected to be a part of a sample. Probability sampling eliminates bias in the population and gives all members a fair chance to be included in the sample.

- Simple random sampling: One of the best probability sampling techniques that helps in saving time and resources, is the Simple Random Sampling method. It is a reliable method of obtaining information where every single member of a population is chosen randomly, merely by chance. Each individual has the same probability of being chosen to be a part of a sample.

For example, in an organization of 500 employees, if the HR team decides on conducting team building activities, it is highly likely that they would prefer picking chits out of a bowl. In this case, each of the 500 employees has an equal opportunity of being selected.

- Cluster sampling: Cluster sampling is a method where the researchers divide the entire population into sections or clusters that represent a population. Clusters are identified and included in a sample based on demographic parameters like age, sex, location, etc. This makes it very simple for a survey creator to derive effective inference from the feedback.

For example, if the United States government wishes to evaluate the number of immigrants living in the Mainland US, they can divide it into clusters based on states such as California, Texas, Florida, Massachusetts, Colorado, Hawaii, etc. This way of



Department of Computer Science and Engineering (Data Science)

conducting a survey will be more effective as the results will be organized into states and provide insightful immigration data.

- **Systematic sampling:** Researchers use the systematic sampling method to choose the sample members of a population at regular intervals. It requires the selection of a starting point for the sample and sample size that can be repeated at regular intervals. This type of sampling method has a predefined range, and hence this sampling technique is the least time-consuming.
For example, a researcher intends to collect a systematic sample of 500 people in a population of 5000. He/she numbers each element of the population from 1-5000 and will choose every 10th individual to be a part of the sample (Total population/ Sample Size = $5000/500 = 10$).
- **Stratified random sampling:** Stratified random sampling is a method in which the researcher divides the population into smaller groups that don't overlap but represent the entire population. While sampling, these groups can be organized and then draw a sample from each group separately.
For example, a researcher looking to analyse the characteristics of people belonging to different annual income divisions will create strata (groups) according to the annual family income. E.g. – less than \$20,000, \$21,000 – \$30,000, \$31,000 to \$40,000, \$41,000 to \$50,000, etc. By doing this, the researcher concludes the characteristics of people belonging to different income groups. Marketers can analyse which income groups to target and which ones to eliminate to create a roadmap that would bear fruitful results.

Lab Assignments to complete in this session

1. Consider a randomly generated dataset which has id, group number (1-10) and Values as three attributes. Implement systematic sampling, stratified sampling (3 stratas) and Cluster sampling.
2. Consider titanic dataset and implement strata sampling and cluster sampling.

Lab Work:

QUESTION 1

```
import numpy as np
import pandas as pd
```

```
# Define total number of groups
number_of_groups = 10
# Create data dictionary
data = {'Id': np.arange(1, number_of_groups+1).tolist(),
        'Group Number': np.arange(1, number_of_groups+1).tolist(),
        'Value': [19, 71, 58, 62, 12, 91, 60, 75, 38, 51]}
# Transform dictionary into a data frame
df1 = pd.DataFrame(data)

display(df1)
```

	Id	Group Number	Value
0	1	1	19
1	2	2	71
2	3	3	58
3	4	4	62
4	5	5	12
5	6	6	91
6	7	7	60
7	8	8	75
8	9	9	38
9	10	10	51

```
# Define systematic sampling function
def systematic_sampling(df, step):

    indexes = np.arange(0, len(df), step=step)
    systematic_sample = df.iloc[indexes]
    return systematic_sample
```

```
# Obtain a systematic sample and save it in a new variable
systematic_sample = systematic_sampling(df1, 3)
# View sampled data frame
display(systematic_sample)
```

	Id	Group	Number	Value
0	1		1	19
3	4		4	62
6	7		7	60
9	10		10	51



```
# Create a dictionary of students
students = {
    'Name': ['Lisa', 'Kate', 'Ben', 'Kim', 'Josh',
            'Alex', 'Evan', 'Greg', 'Sam', 'Ella'],
    'ID': ['001', '002', '003', '004', '005', '006',
          '007', '008', '009', '010'],
    'Grade': ['A', 'A', 'C', 'B', 'B', 'B', 'C',
             'A', 'A', 'A'],
    'Category': [2, 3, 1, 3, 2, 3, 3, 1, 2, 1]
}

# Create dataframe from students dictionary
df = pd.DataFrame(students)

# View the dataframe
df
```

	Name	ID	Grade	Category
0	Lisa	001	A	2
1	Kate	002	A	3
2	Ben	003	C	1
3	Kim	004	B	3
4	Josh	005	B	2
5	Alex	006	B	3
6	Evan	007	C	3
7	Greg	008	A	1
8	Sam	009	A	2
9	Ella	010	A	1



```
df.groupby('Grade', group_keys=False).apply(lambda x: x.sample(2))
```

	Name	ID	Grade	Category
9	Ella	010	A	1
1	Kate	002	A	3
2	Lisa	001	A	2

```
df.groupby('Grade', group_keys=False).apply(lambda x: x.sample(frac=0.6))
```

	Name	ID	Grade	Category
0	Lisa	001	A	2
1	Kate	002	A	3
7	Greg	008	A	1
5	Alex	006	B	3
4	Josh	005	B	2
6	Evan	007	C	3

```
#Make this example reproducible
np.random.seed(0)
```

```
#Create DataFrame
```

```
df = pd.DataFrame({'tour': np.repeat(np.arange(1,11), 20),
                    'experience': np.random.normal(loc=7, scale=1, size=200)})
df
```


	tour	experience
0	1	8.764052
1	1	7.400157
2	1	7.978738
3	1	9.240893
4	1	8.867558
...
195	10	6.828454
196	10	7.771791
197	10	7.823504
198	10	9.163236
199	10	8.336528

200 rows × 2 columns

```
#Randomly choose 4 tour groups out of the 10
clusters = np.random.choice(np.arange(1,11), size=4, replace=False)
```

```
#Define sample as all members who belong to one of the 4 tour groups
cluster_sample = df[df['tour'].isin(clusters)]

#View first six rows of sample
cluster_sample
```

	tour	experience	
0	1	8.764052	
1	1	7.400157	
2	1	7.978738	
3	1	9.240893	
4	1	8.867558	
...	
195	10	6.828454	
196	10	7.771791	
197	10	7.823504	
198	10	9.163236	
199	10	8.336528	

80 rows × 2 columns

```
#Find how many observations came from each tour group
cluster_sample['tour'].value_counts()
```

```
10    20
8      20
6      20
1      20
Name: tour, dtype: int64
```

▼ QUESTION 2

```
import pandas as pd
import numpy as np
```

```
df = pd.read_csv('/Titanic_fda.csv')
df
```

	Name	PClass	Age	Sex	Survived
0	Allen, Miss Elisabeth Walton	1st	29.00	female	1
1	Allison, Miss Helen Loraine	1st	2.00	female	0
2	Allison, Mr Hudson Joshua Creighton	1st	30.00	male	0
3	Allison, Mrs Hudson JC (Bessie Waldo Daniels)	1st	25.00	female	0
4	Allison, Master Hudson Trevor	1st	0.92	male	1
...
1308	Zakarian, Mr Artun	3rd	27.00	male	0
1309	Zakarian, Mr Maprieder	3rd	26.00	male	0
1310	Zenni, Mr Philip	3rd	22.00	male	0
1311	Lievens, Mr Rene	3rd	24.00	male	0



```
df.isnull().sum()
```

```
Name      0
PClass    1
Age      557
Sex        0
Survived   0
dtype: int64
```

```
df['Age'].round(decimals = 0)
```

```
0      29.0
1       2.0
2      30.0
3      25.0
4       1.0
...
1308    27.0
1309    26.0
1310    22.0
1311    24.0
1312    29.0
Name: Age, Length: 1313, dtype: float64
```

```
df['Age'].fillna(df.Age.mean(), inplace=True)
df
```


	Name	PClass	Age	Sex	Survived
0	Allen, Miss Elisabeth Walton	1st	29.00	female	1
1	Allison, Miss Helen Loraine	1st	2.00	female	0
2	Allison, Mr Hudson Joshua Creighton	1st	30.00	male	0
3	Allison, Mrs Hudson JC (Bessie Waldo Daniels)	1st	25.00	female	0
4	Allison, Master Hudson Trevor	1st	0.92	male	1
...



```
df.dropna() #Drop NaN values
```

	Name	PClass	Age	Sex	Survived
0	Allen, Miss Elisabeth Walton	1st	29.00	female	1
1	Allison, Miss Helen Loraine	1st	2.00	female	0
2	Allison, Mr Hudson Joshua Creighton	1st	30.00	male	0
3	Allison, Mrs Hudson JC (Bessie Waldo Daniels)	1st	25.00	female	0
4	Allison, Master Hudson Trevor	1st	0.92	male	1
...
1308	Zakarian, Mr Artun	3rd	27.00	male	0
1309	Zakarian, Mr Maprieder	3rd	26.00	male	0
1310	Zenni, Mr Philip	3rd	22.00	male	0
1311	Lievens, Mr Rene	3rd	24.00	male	0
1312	Zimmerman, Leo	3rd	29.00	male	0



1312 rows × 5 columns

```
df.duplicated().sum() #Check for duplicate values
```

0

```
# A discontinuous class column 'Age_group' is made
```

```
age_group = np.arange(0,100,10)
age_group_labels = [f"{i} - {i+10}" for i in range(0,90,10)]
df['Age_group'] = pd.cut(df['Age'], bins=age_group, labels=age_group_labels)
```

▼ Stratified Sampling On PClass

```
df['PClass'].value_counts()
```

```

3rd    711
1st    322
2nd    279
Name: PClass, dtype: int64

```

'''There are 54.2% 3rd class passengers, 24.54% 1st class passengers, and 21.3% 1st class passengers. Create a sample of 787 passengers disproportionately (equal number of passengers from each PClass stratum)

Dispropotionate Sampling: Using pandas groupby, seperate the passengers into groups based on their grade i.e. A, B and C and randomly sample 262 passengers from each class group using the sample function

```

...
df1 = df.groupby('PClass', group_keys=False).apply(lambda x: x.sample(262))
df1

```

	Name	PClass	Age	Sex	Survived
290	Bazzani, Ms Albi	1st	30.397989	female	1
233	Smart, Mr John Montgomery	1st	56.000000	male	0
49	Carter, Miss Lucile Polk	1st	14.000000	female	1
275	White, Mr Richard Frasar	1st	21.000000	male	0
70	Cornell, Mrs Robert Clifford (Malvi Helen Lamson)	1st	55.000000	female	1
...
1133	Petroff, Mr Pentcho	3rd	30.397989	male	0
812	Franklin, Mr Charles	3rd	30.397989	male	0
886	Johanson, Mr Jakob Alfred	3rd	34.000000	male	0
1309	Zakarian, Mr Maprieder	3rd	26.000000	male	0
1029	Moran, Mr Daniel J	3rd	30.397989	male	0

786 rows × 6 columns

```
df1['PClass'].value_counts()
```

```

1st    262
3rd    262
2nd    262
Name: PClass, dtype: int64

```

```
...
```

Sample out 60% of passengers proportionately (create population samples from each stratum based on its propotion in the sample)

Proportionate Sampling: Using pandas groupby, seperate the passengers in groups based on t i.e. 1st, 2nd, 3rd, and random sample from each group based on population proportion. The total sample size is 60% (0.6) of the population.

```
'''
df2 = df.groupby('PClass', group_keys=False).apply(lambda x: x.sample(frac=0.6))
df2
```

	Name	PClass	Age	Sex	Survived	Age_group
240	Spedden, Mr Frederick Oakley	1st	45.000000	male	1	40 - 50
155	Kenyon, Mrs Frederick R (Marion)	1st	30.397989	female	1	30 - 40
256	Taussig, Mr Emil	1st	52.000000	male	0	50 - 60
275	White, Mr Richard Frasar	1st	21.000000	male	0	20 - 30
199	Pears, Mr Thomas	1st	30.397989	male	0	30 - 40
...
691	Brocklebank, Mr William Alfred	3rd	35.000000	male	0	30 - 40
685	Bradley, Miss Bridget Delia	3rd	18.000000	female	1	10 - 20
982	Madigan, Miss Margaret	3rd	30.397989	female	1	30 - 40
834	Guest, Mr Robert	3rd	30.397989	male	0	30 - 40
936	Klasen, Miss Gertrud Emilia	3rd	1.500000	female	0	0 - 10

787 rows × 6 columns

```
df2['PClass'].value_counts()
```

```
3rd    427
1st    193
2nd    167
Name: PClass, dtype: int64
```

```
'''
Notice that even in the sample, there are 54.2% 3rd class passengers,
24.54% 1st class passengers, and 21.3% 1st class passengers.
'''
```

```
'\nNotice that even in the sample, there are 54.2% 3rd class passengers,\n24.54% 1st
and 21.3% 1st class passengers \n'
```

▼ Cluster Sampling on Age

```
df.isnull().sum()
```

```
Name          0
PClass        1
Age           0
Sex           0
Survived      0
Age_group     0
dtype: int64
```

```
df.dropna()
```

	Name	PClass	Age	Sex	Survived	Age_g
0	Allen, Miss Elisabeth Walton	1st	29.00	female	1	20
1	Allison, Miss Helen Loraine	1st	2.00	female	0	0
2	Allison, Mr Hudson Joshua Creighton	1st	30.00	male	0	20
3	Allison, Mrs Hudson JC (Bessie Waldo Daniels)	1st	25.00	female	0	20
4	Allison, Master Hudson Trevor	1st	0.92	male	1	0
...
1308	Zakarian, Mr Artun	3rd	27.00	male	0	20
1309	Zakarian, Mr Maprieder	3rd	26.00	male	0	20
1310	Zenni, Mr Philip	3rd	22.00	male	0	20
1311	Lievens, Mr Rene	3rd	24.00	male	0	20
1312	Zimmerman, Leo	3rd	29.00	male	0	20

1312 rows × 6 columns

```
df['Age_group'].value_counts()
```

```

30 - 40    707
20 - 30    260
10 - 20    117
40 - 50    104
0 - 10      55
50 - 60     48
60 - 70     19
70 - 80      3
80 - 90      0

```

Name: Age_group, dtype: int64

```

# Randomly choose 525 age groups out of 1312
clusters = np.random.choice(list(df.Age_group.unique()), size=4, replace=False)

# Define sample as all passenegrs belong to one of the 21 age groups
cluster_sample = df[df['Age_group'].isin(clusters)]

# View first 6 rows of sample
cluster_sample

```

	Name	PClass	Age	Sex	Survived
1	Allison, Miss Helen Loraine	1st	2.000000	female	0
4	Allison, Master Hudson Trevor	1st	0.920000	male	1
6	Andrews, Miss Kornelia Theodosia	1st	63.000000	female	1
7	Andrews, Mr Thomas, jr	1st	39.000000	male	0
11	Astor, Mrs John Jacob (Madeleine Talmadge Force)	1st	19.000000	female	1
...
1302	Yalsevac, Mr Ivan	3rd	30.397989	male	1
1304	Yasbeck, Mrs Antoni	3rd	15.000000	female	1
1305	Yousef, Mr Gassan	3rd	20.000000	male	0

```
# find how many observations came from each age group
cluster_sample['Age_group'].value_counts()
```

```
30 - 40    707
10 - 20    117
0 - 10      55
60 - 70     19
80 - 90      0
70 - 80      0
50 - 60      0
40 - 50      0
20 - 30      0
Name: Age_group, dtype: int64
```