Statistical Inferences

Statistical analysis

Statistical analysis has two main focuses:

1. Descriptive statistics
2. Inferential statistics.

Descriptive statistics summarize data using graphs and summary values such as the mean and interquartile range.

- Descriptive statistics can help us identify relationships and patterns.
- Descriptive statistics do not draw conclusions

Inferential statistical analysis does allow us to make conclusions beyond the data we have to the population from which it was drawn.

A definition of inference is: The process of drawing conclusions about population parameters based on a sample taken from the population

Inference

- A sample is likely to be a good representation of the population
- There is an element of uncertainty as to how well the sampler presents the population
- The way the sample is taken matters

What is Statistical Inference?

Statistical inference is the process of using a sample to infer the properties of a population. Statistical procedures use sample data to estimate the characteristics of the whole population from which the sample was drawn.

Image of a scientist who wants to make a statistical inference. Scientists typically want to learn about a population. When studying a phenomenon, such as the effects of a new medication or public opinion, understanding the results at a population level is much more valuable than understanding only the comparatively few participants in a study.

Unfortunately, populations are usually too large to measure fully. Consequently, researchers must use a manageable subset of that population to learn about it.

By using procedures that can make statistical inferences, you can estimate the properties and processes of a population. More specifically, sample statistics can estimate population parameters. Learn more about the differences between sample statistics and population parameters.

How to Make Statistical Inferences

In its simplest form, the process of making a statistical inference requires you to do the following:

1.Draw a sample that adequately represents the population.
2.Measure your variables of interest.
3.Use appropriate statistical methodology to generalize your sample results to the population while accounting for sampling error.

Statistical inference requires using specialized sampling methods that tend to produce representative samples. If the sample does not look like the larger population you're studying, you can't trust any inferences from the sample. Consequently, using an appropriate method to obtain your sample is crucial. The best sampling methods tend to produce samples that look like the target population.

Inferences in statistics can help you make predictions and conclusions about the populations you are looking at by interpreting the results of random samples from that population. The two main applications of inferential statistics that help us to draw these conclusions are **hypothesis testing** and **confidence intervals** of the data.

Statistical inferences are dependent on three main components:

•The **size of samples;**
•**Variability in the samples;** and
•The **size of the observed differences**.

The **population** refers to a group of units (persons, objects, or other items) enumerated in a census or from which a sample is drawn.

A **sample** is defined as a subset of a population selected for measurement, observation, or questioning, to provide statistical information about the population.

To conduct statistical inference, the following conditions must be met:

1.The data for the experiment should be obtained through **random samples or randomized experiments**
2.The distribution of the sample means must be approximately normal
3.Individual observations must be independent

Sampling Process

Sampling is the process of selecting a subset of a population to study. The goal of sampling is to collect data from a representative sample that accurately reflects the characteristics of the population. There are several types of sampling methods, including random sampling, stratified sampling, and cluster sampling.

Why is sampling needed?

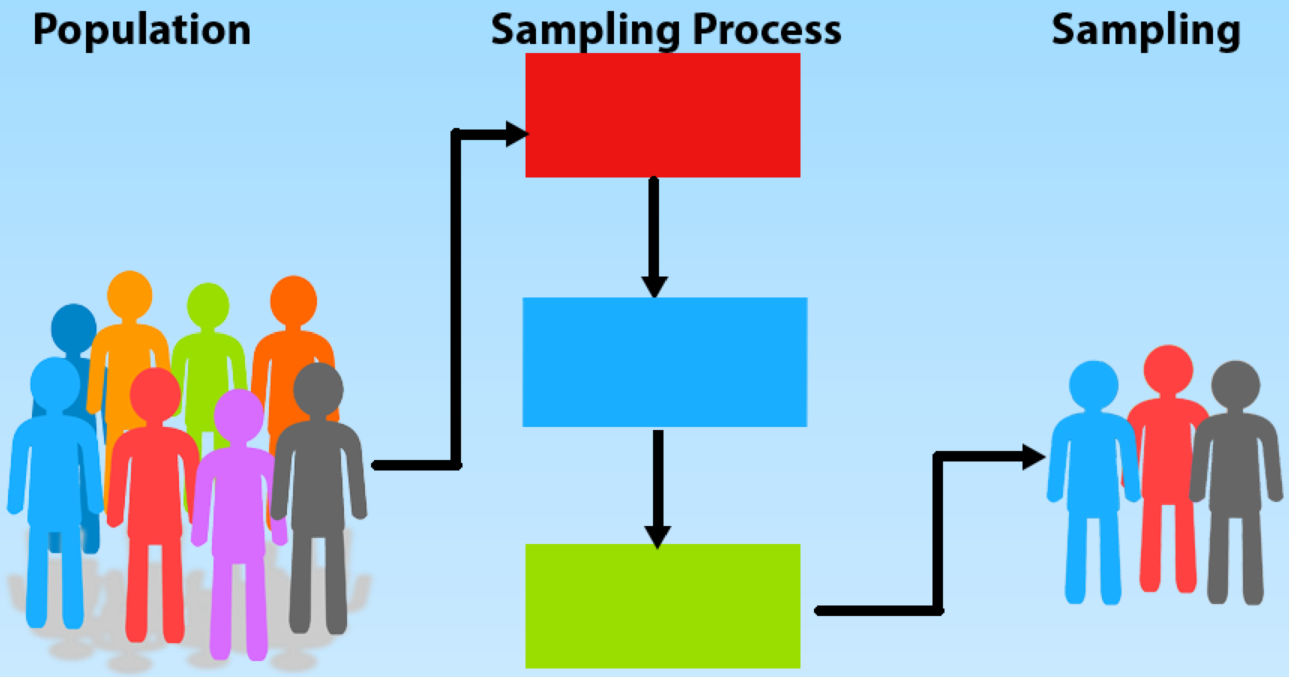
The sampling process is used to collect the sample data which helps us make inferences about the population data. It allows us to draw conclusions about the population from sample data.

There are various reasons why sampling is needed, such as:

•It is a cost-efficient method as we do not have to use data from the total population to build a machine learning model.

•It is not feasible to study the total population, therefore the sampling process makes our work easier.

•The size of the sample data is always smaller, therefore we can clean and process our data with more efficacy.



Population

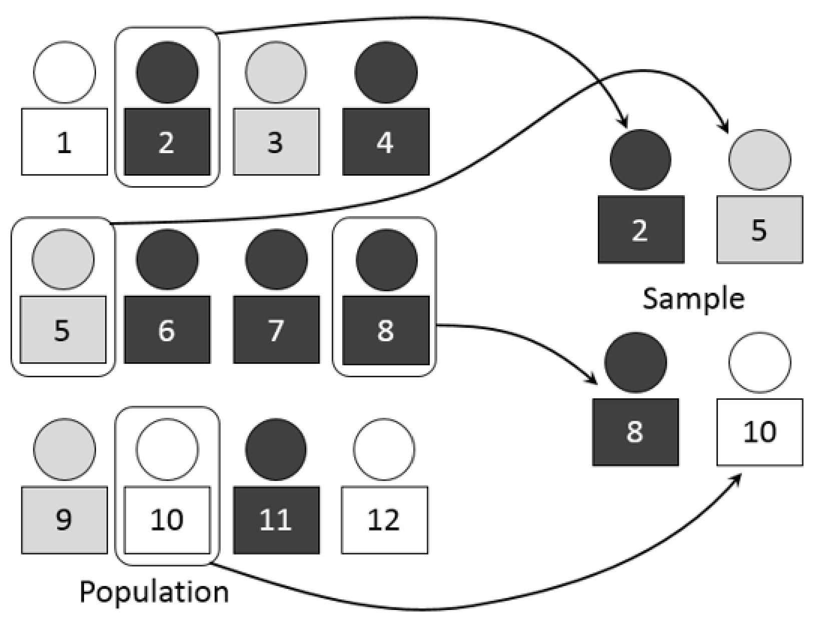
Sampling Process

Sampling

Simple Random Sampling (SRS)

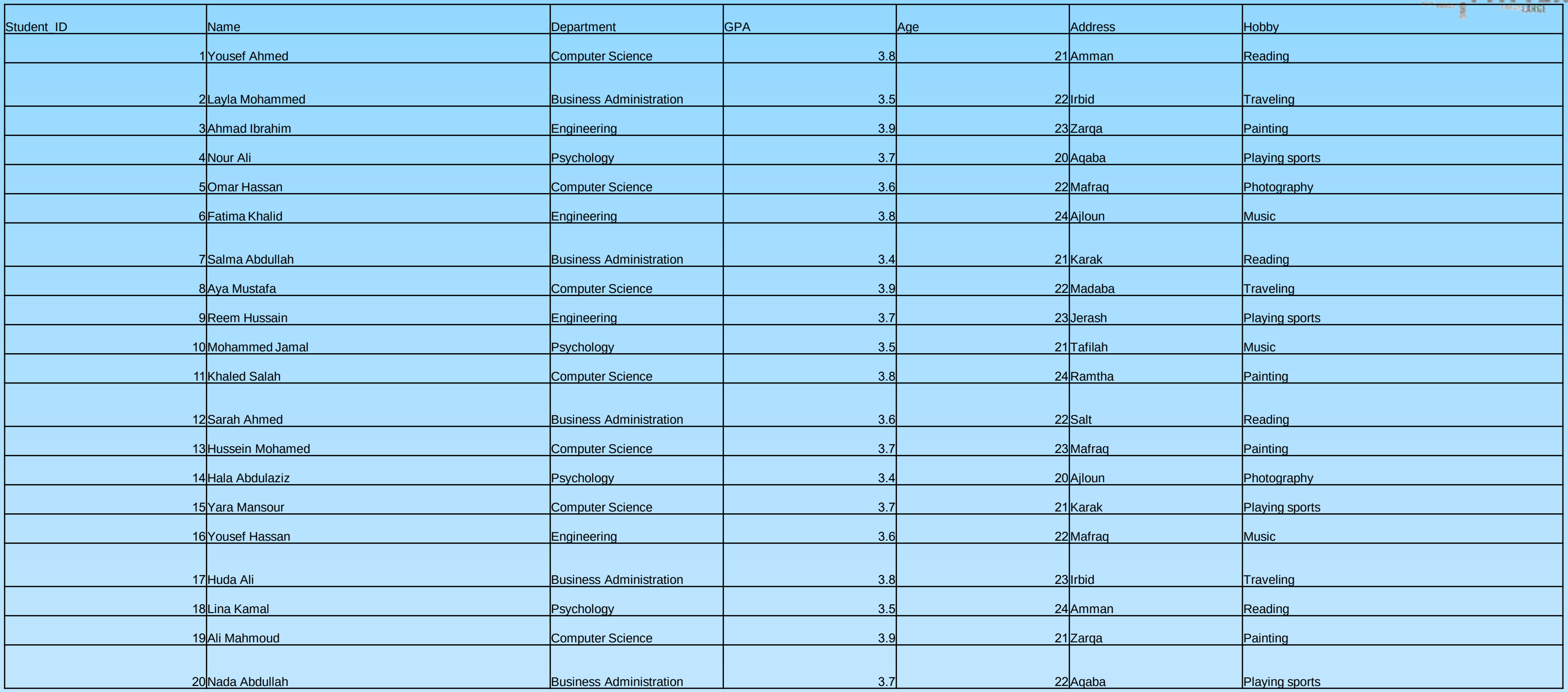
In Simple Random Sampling (**SRS**), **everyone in the population has an equal chance of being selected** for the sample.

To prepare a sample using **SRS**, we **randomly select the desired number of members from the population**.



الجامعة الألمانية الأردنية
German Jordanian University

Original Data



Student_ID	Name	Department	GPA	Age	Address	Hobby
1	Yousef Ahmed	Computer Science	3.8	21	Amman	Reading
2	Layla Mohammed	Business Administration	3.5	22	Irbid	Traveling
3	Ahmad Ibrahim	Engineering	3.9	23	Zarqa	Painting
4	Nour Ali	Psychology	3.7	20	Aqaba	Playing sports
5	Omar Hassan	Computer Science	3.6	22	Mafraq	Photography
6	Fatima Khalid	Engineering	3.8	24	Ajloun	Music
7	Salma Abdullah	Business Administration	3.4	21	Karak	Reading
8	Aya Mustafa	Computer Science	3.9	22	Madaba	Traveling
9	Reem Hussain	Engineering	3.7	23	Jerash	Playing sports
10	Mohammed Jamal	Psychology	3.5	21	Tafilah	Music
11	Khaled Salah	Computer Science	3.8	24	Ramtha	Painting
12	Sarah Ahmed	Business Administration	3.6	22	Salt	Reading
13	Hussein Mohamed	Computer Science	3.7	23	Mafraq	Painting
14	Hala Abdulaziz	Psychology	3.4	20	Ajloun	Photography
15	Yara Mansour	Computer Science	3.7	21	Karak	Playing sports
16	Yousef Hassan	Engineering	3.6	22	Mafraq	Music
17	Huda Ali	Business Administration	3.8	23	Irbid	Traveling
18	Lina Kamal	Psychology	3.5	24	Amman	Reading
19	Ali Mahmoud	Computer Science	3.9	21	Zarqa	Painting
20	Nada Abdullah	Business Administration	3.7	22	Aqaba	Playing sports

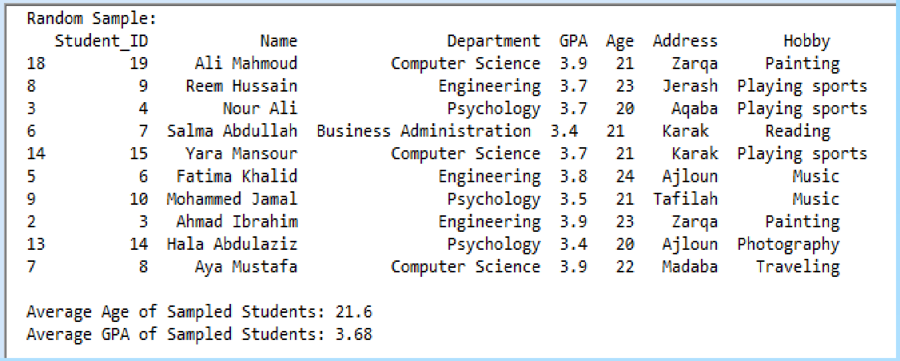
```
import pandas as pd
original_data = pd.read_csv('StudentsForSample.csv')
original_department_counts = original_data['Department'].value_counts()
# Display the counts for each department
print("Number of students for each department in the original dataset:")
print(original_department_counts)
```

Number of students for each department in the original dataset:

```
Computer Science 7
Engineering 5
Business Administration 4
 Psychology 4
```

Random Sampling and Average Statistics

```
import pandas as pd

# Read the dataset
original_data = pd.read_csv('StudentsForSample.csv')

# Perform random sampling
random_sample = original_data.sample(n=10, random_state=42)

avg_age = random_sample['Age'].mean()
avg_gpa = random_sample['GPA'].mean()

# Display the random sample and average statistics
print("Random Sample:")
print(random_sample)
print("\nAverage Age of the Random Sample:", avg_age)
print("Average GPA of the Random Sample:", avg_gpa)
```



```
Random Sample:
    Student_ID             Name               Department  GPA  Age  Address        Hobby
18          19      Ali Mahmoud         Computer Science  3.9   21    Zarqa     Painting
8            9     Reem Hussain              Engineering  3.7   23   Jerash  Playing sports
3            4         Nour Ali               Psychology  3.7   20    Aqaba  Playing sports
6            7   Salma Abdullah  Business Administration  3.4   21    Karak      Reading
14          15     Yara Mansour         Computer Science  3.7   21    Karak  Playing sports
5            6    Fatima Khalid              Engineering  3.8   24   Ajloun        Music
9           10   Mohammed Jamal               Psychology  3.5   21  Tafilah        Music
2            3    Ahmad Ibrahim              Engineering  3.9   23    Zarqa     Painting
13          14   Hala Abdulaziz               Psychology  3.4   20   Ajloun  Photography
7            8      Aya Mustafa         Computer Science  3.9   22   Madaba    Traveling

Average Age of Sampled Students: 21.6
Average GPA of Sampled Students: 3.68
```

Number of students for each department in the random sample:

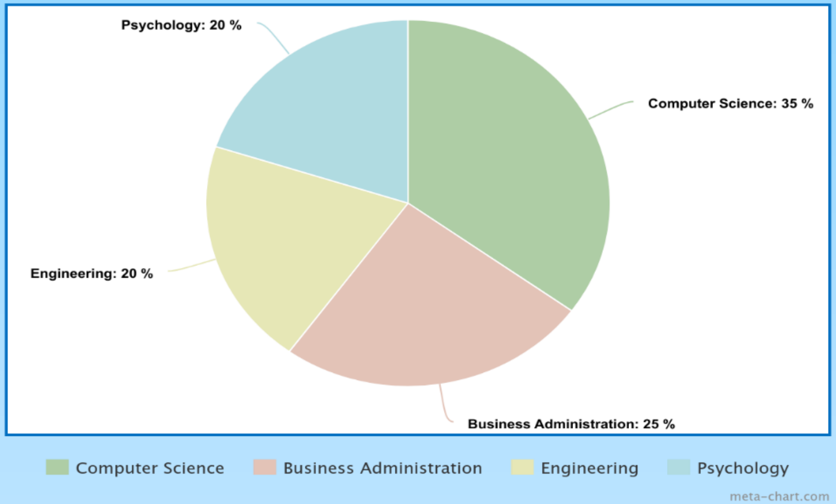
Computer Science 3

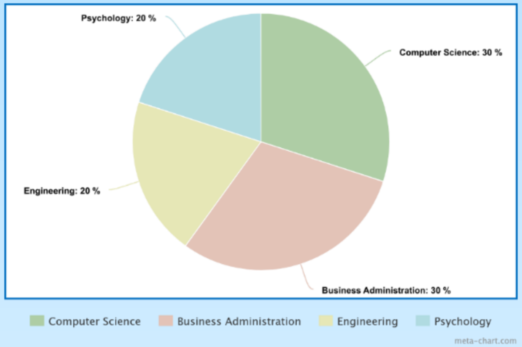
Engineering 3

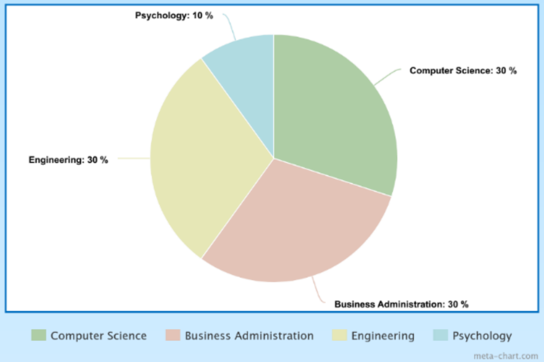
Business Administration 2

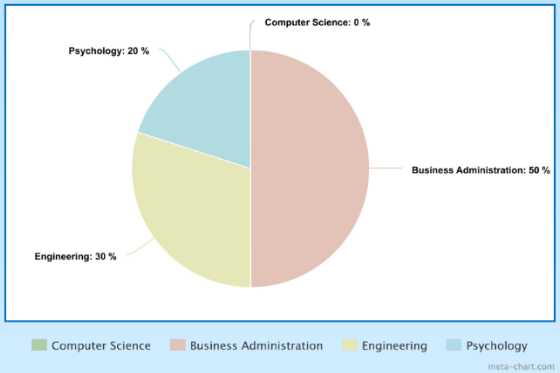
Psychology 2

Random Samples



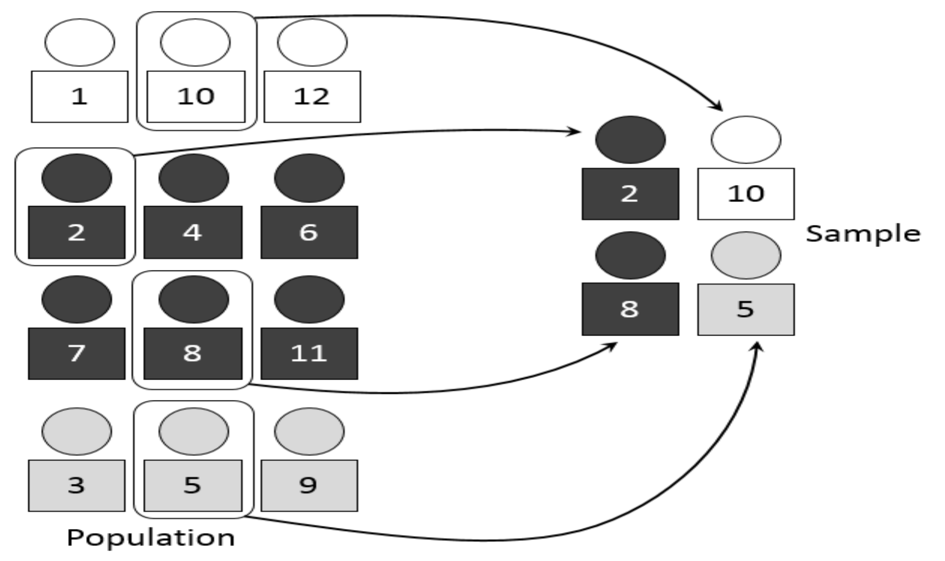




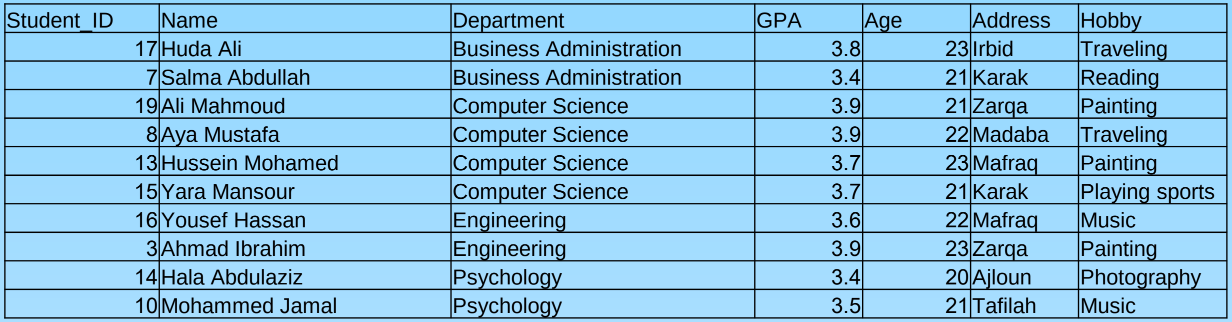


Stratified sampling

Stratified sampling is a method of obtaining a representative sample from a population that researchers have divided into relatively similar subpopulations (strata). Researchers use stratified sampling to ensure specific subgroups are present in their sample. It also helps them obtain precise estimates of each group's characteristics.

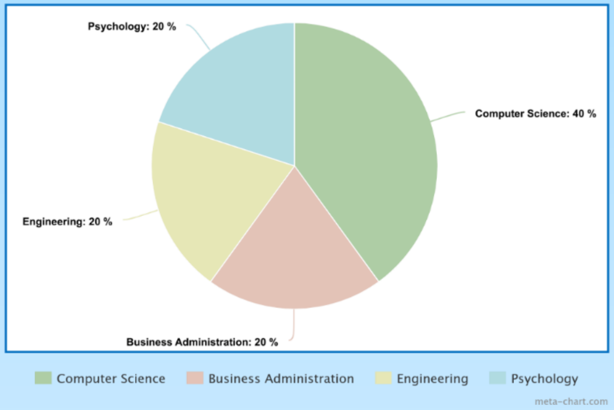


Stratified sampling



Student_ID	Name	Department	GPA	Age	Address	Hobby
17	Huda Ali	Business Administration	3.8	23	Irbid	Traveling
7	Salma Abdullah	Business Administration	3.4	21	Karak	Reading
19	Ali Mahmoud	Computer Science	3.9	21	Zarqa	Painting
8	Aya Mustafa	Computer Science	3.9	22	Madaba	Traveling
13	Hussein Mohamed	Computer Science	3.7	23	Mafraq	Painting
15	Yara Mansour	Computer Science	3.7	21	Karak	Playing sports
16	Yousef Hassan	Engineering	3.6	22	Mafraq	Music
3	Ahmad Ibrahim	Engineering	3.9	23	Zarqa	Painting
14	Hala Abdulaziz	Psychology	3.4	20	Ajloun	Photography
10	Mohammed Jamal	Psychology	3.5	21	Tafilah	Music

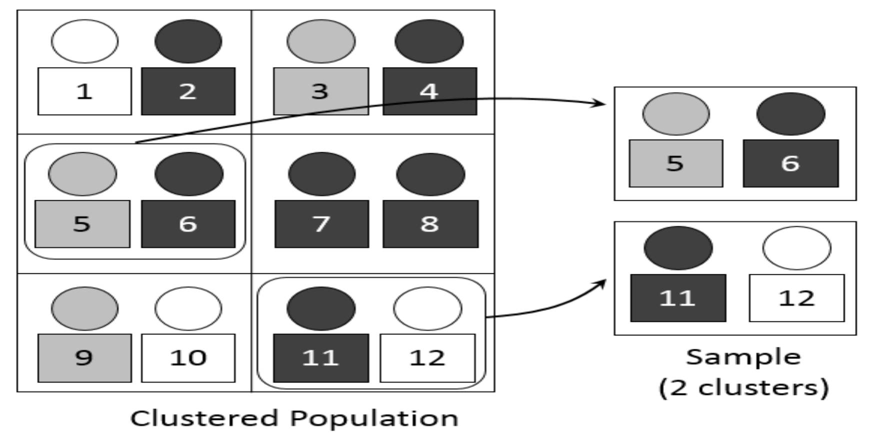
```
# Read the dataset
original_data = pd.read_csv('StudentsForSample.csv')

# Perform random sampling
stratified_sample = original_data.groupby('Department').apply(
    lambda x: x.sample(frac=0.50)
)
# Calculate the average age and average GPA of the random sample
avg_age = stratified_sample['Age'].mean()
avg_gpa = stratified_sample['GPA'].mean()

# Display the random sample and average statistics
print(stratified_sample)
print("\nAverage Age of the Stratified Sample:", avg_age)
print("Average GPA of the Stratified Sample:", avg_gpa)
```



```
Average Age of the Random Sample: 21.7
Average GPA of the Random Sample: 3.679
```

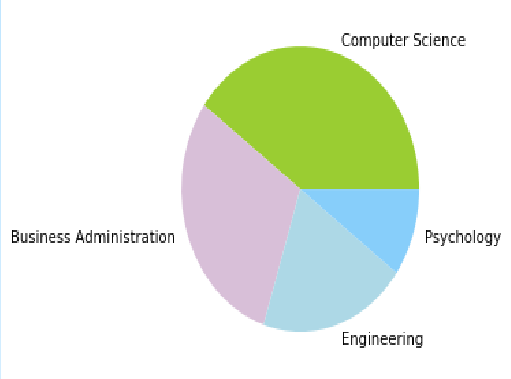
Cluster Sampling

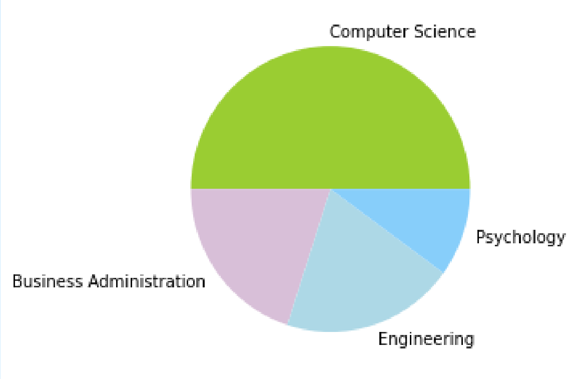
Cluster sampling is a sampling method in which the population is divided into clusters (groups of units) and a random sample of clusters is selected. It is a useful method when it is impractical or expensive to sample the entire population and the clusters are representative of the population.

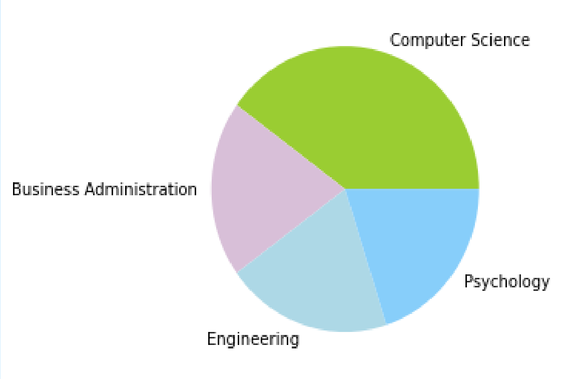


Cluster Sampling

```
original_data = original_data.sample(frac=1)

clusters = np.array_split(original_data, 4)
chosen_clusters = np.random.choice(len(clusters), 2, replace=False)

# Get the individuals in the chosen clusters
sample = np.concatenate([clusters[i] for i in chosen_clusters])
df = pd.DataFrame(sample, columns=['Student_ID', 'Name', 'Department','GPA', 'Age',
'Address', 'Hobby'])
```


Computer Science
Business Administration
Psychology
Engineering


Computer Science
Business Administration
Psychology
Engineering


Computer Science
Business Administration
Psychology
Engineering