1. Compare cluster sampling and stratified sampling by giving detailed definitions, examples, similarities, and differences. Explain when to use which.

- Cluster Sampling,

Definition,

Cluster sampling is defined as a sampling method where the researcher creates multiple clusters of people from a population where they are indicative of homogeneous characteristics and have an equal chance of being a part of the sample.

In cluster sampling, we keep the number of samplings in each group equal and another important point is that each sample can just be in one group, but all samples should be separated. After separating samples into different groups, the researcher selects one group to continue.

Types,

There are two ways to classify this sampling technique. The first way is based on the number of stages followed to obtain the cluster sample, and the second way is the representation of the groups in the entire cluster. Normally, cluster sampling is done in several stages, and in each stage, the researcher can choose different sampling methods.

Advantages,

- Consumes less time and cost,
- Convenient access,
- Data accuracy,
- Ease of implementation

Example,

An organization intends to survey to analyze the performance of smartphones across Germany. They can divide the entire country's population into cities (clusters) and select cities with the highest population and also filter those using mobile devices.

- Stratified Sampling

Definition,

Stratified random sampling is a type of probability sampling using which a research organization can branch off the entire population into multiple non-overlapping, homogeneous groups (strata) and randomly choose final members from the various strata for research which reduces cost and improves efficiency. Members in each of these groups should be distinct so that every member of all groups gets an equal opportunity to be selected using simple probability.

Advantages,

- Better accuracy in results in comparison to other probability sampling methods such as cluster sampling, simple random sampling, systematic sampling, or non-probability methods such as convenience sampling.
- Convenient to train a team to stratify a sample due to the exactness of the nature of this sampling technique.
- Due to the statistical accuracy of this method, smaller sample sizes can also retrieve highly useful results for a researcher.
- This sampling technique covers the maximum population as the researchers have complete charge over the strata division.

- Cluster sampling vs stratified sampling

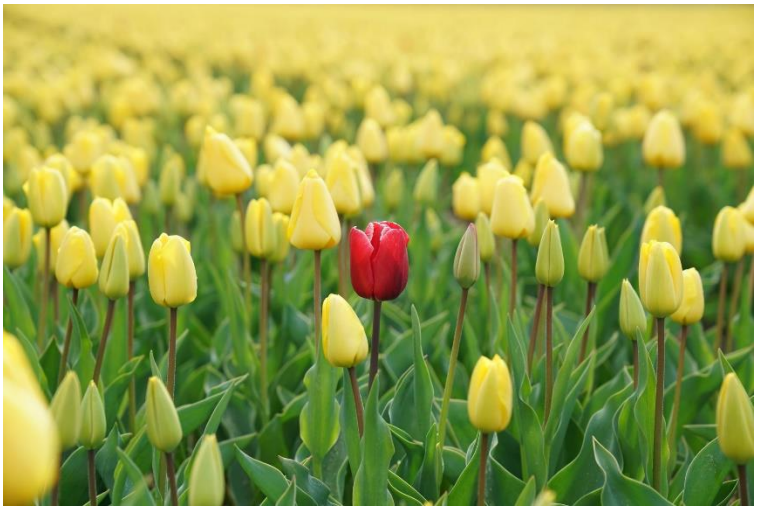
Cluster sampling	Stratified sampling
Elements of a population are randomly selected to be a part of groups (clusters).	The researcher divides the entire population into even segments (strata).
Members from randomly selected clusters are a part of this sample.	Researchers consider individual components of the strata randomly to be a part of sampling units.
Researchers maintain homogeneity between clusters.	Researchers maintain homogeneity within the strata.

Reference,

https://www.questionpro.com/blog/cluster-sampling/

https://www.questionpro.com/blog/stratified-random-sampling/

2. What is an outlier? Which one is more affected by the presence of outliers, the mean or the median? Explain.



An outlier is an observation that lies an abnormal distance from other values in a random sample of a population. Some outliers represent true values from natural variation in the population. Other outliers may result from incorrect data entry, equipment malfunctions, or other measurement errors.

- Four ways of calculating outliers

Sorting Methods,

It is a simple way of finding outliers, sorting values and watching them, and finding outliers

Using visualizations,

use different graphs to find the range and distribution of data, the famous graphs to use are, scatter plot, box-and-whisker plot, histogram, and stem-and-leaf graph

Statistical outlier detection,

You can use z scores that are found from standard deviations to select the outliers

Using the interquartile range,

The famous way to find outliers, you need to find the first quartile and third quartile, then the difference (Q3 - Q1) is called the interquartile range or IQ.

lower inner fence: Q1 - 1.5*IQ
upper inner fence: Q3 + 1.5*IQ
lower outer fence: Q1 - 3*IQ
upper outer fence: Q3 + 3*IQ

The outliers can affect the mean more than the median value, please see the below sample of data,

30, 171, 184, 201, 212, 250, 265, 270, 272, 289, 305, 306, 322, 322, 336, 346, 351, 370, 390, 404, 409, 411, 436, 437, 439, 441, 444, 448, 451, 453, 470, 480, 482, 487, 494, 495, 499, 503, 514, 521, 522, 527, 548, 550, 559, 560, 570, 572, 574, 578, 585, 592, 592, 607, 616, 618, 621, 629, 637, 638, 640, 656, 668, 707, 709, 719, 737, 739, 752, 758, 766, 792, 792, 794, 802, 818, 830, 832, 843, 858, 860, 869, 918, 925, 953, 991, 1000, 1005, 1068, 1441

Mean = 576.078

Median = 559.5

According to the interquartile range method, the number 1441 is an outlier, so we remove it from the data set and calculate again,

Mean = 566.360

Median = 559

The reason is, the outliers are few and can't affect the number of samples a lot but their values are very larger or very smaller than the normal number in the range, so

- Median relates to place not much change
- Mean relates to the value of the numbers change

Reference,

https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

https://www.scribbr.com/statistics/outliers/