Stratified Random Sampling

Group Members:
Disha Savant, Shubham Kelkar

CSD-505 Math. Foundations for Data Science (pract)

0

Definition

Stratified sampling is a method in which a population is divided into distinct groups, known as strata, that share a specific attribute. Random samples are subsequently drawn from each group (stratum) to ensure that every group is adequately represented in the overall sample.

Explanation

1. Population Division: Begin by dividing the entire population into smaller, homogeneous groups known as strata. Each group shares a common characteristic.

2. Sampling Within group: Draw a random sample from each group such that the proportion of individuals sampled from each group reflects its relative size within the entire population.

3. Sample Aggregation: The last step involves combining the individual samples from each group to form the complete sample. This process guarantees that every group of the population is represented in the sample.

Examples

Studying travel preferences by a travel company.
Features: age, gender, transport modes, budget group.

Developing a model to predict whether patients have a certain disease.
Features: gender, age group, symptoms, prior medical conditions.

Identify the types of books that are most frequently borrowed from the library.
Features: fiction, non-fiction, children's books, reference materials.

Frequently used in machine learning to ensure that the training and test data sets are representative of the overall data.

Implementation

Teachers from GBS have started a combined class for students in the MSc DS, MSc AI, and MCA programs. They want to gather feedback from 20 students. Given that there are 20 MSc DS students, 20 MSc AI students, and 60 MCA students, how many students should be selected from each course to ensure fair representation?
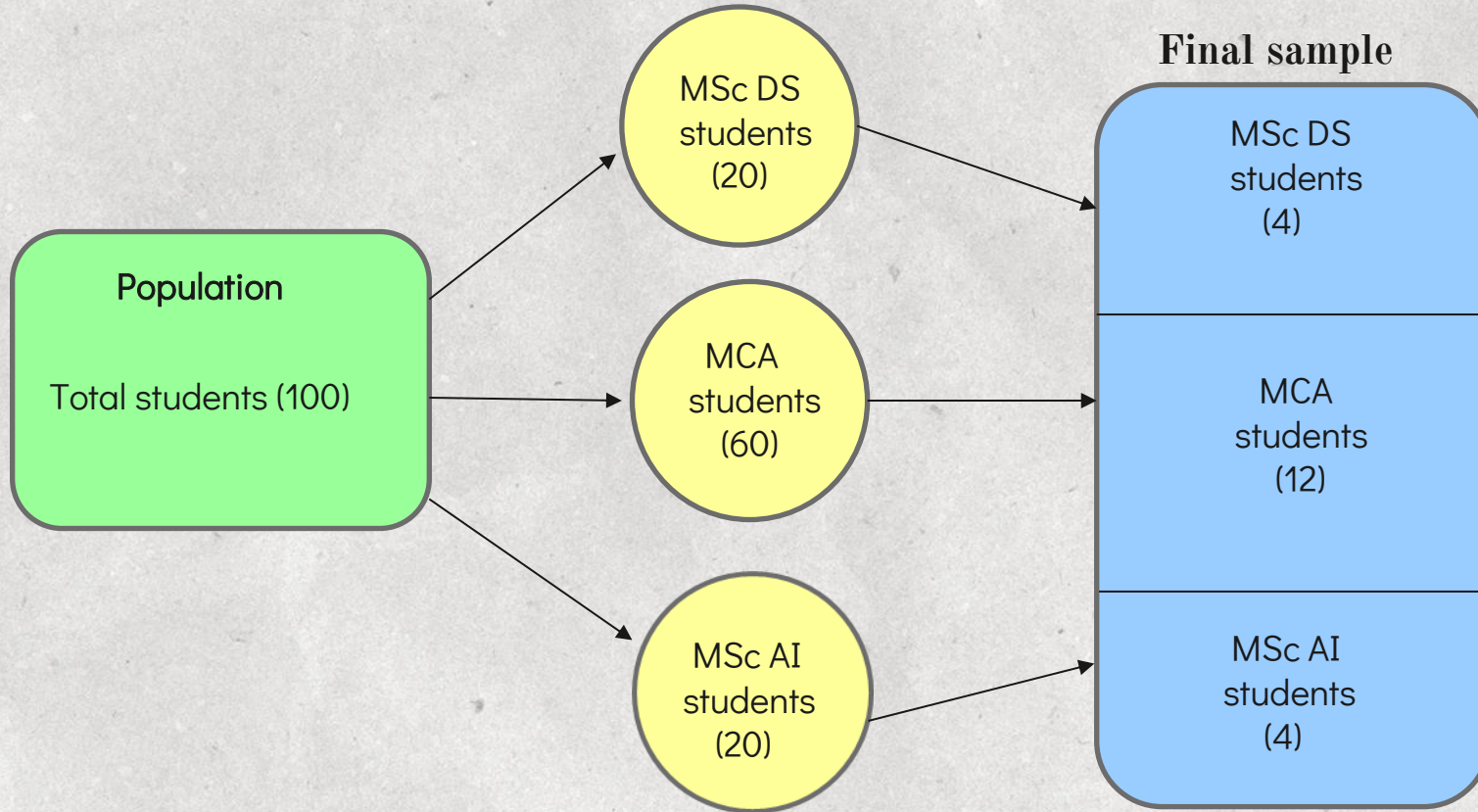
Population size: 100

- MSc DS: 20 --- 20%
- MSc AI: 20 --- 20%
- MCA: 60 --- 60%

Proportionate representation:
Sample size: 20

- MSc DS: 20/100 x 20 = 4
- MSc AI: 20/100 x 20 = 4
- MCA: 60/100 x 20 = 12

Implementation

Real-life usage

When to use ?

1. When the population is heterogeneous and can be divided into distinct groups that are internally homogeneous.

2. Definable groups, when you can clearly define and access information about the groups before sampling.

3. When you have to analyse the different groups. This enables meaningful comparisons.

Real-life usage

When not to use ?

1. When the population is homogeneous.

2. When the population size is small.

3. If resources (time, money, personnel) are limited, simpler methods might be preferable.

4. When the groups are not well defined due to lack of information.

Case study

Stratified Sampling of Neighborhood Sections for Population Estimation: A Case Study of Bo City, Sierra Leone (2015)

A study in Bo City, Sierra Leone (West Africa), used stratified random sampling to improve population estimates in rapidly growing areas where census data is hard to get. By dividing neighborhoods based on residential structures and occupancy rates, researchers gained accurate insights into population density and distribution. This method reduces estimation uncertainty and helps allocate resources better for urban planning and public health in fast-growing regions.

References

1. <u>Stratified Sampling: Definition, Formula, Examples, Types (surveysparrow.com)</u>

2. <u>How Stratified Random Sampling Works, With Examples (investopedia.com)</u>

3. ChatGpt.

4. <u>Stratified Sampling of Neighborhood Sections for Population Estimation: A Case Study of Bo City, Sierra Leone (ncbi.nlm.nih.gov)</u>

Thank You !