stratifiedSample

November 15, 2024

1. Definition

- A Stratified Sample is a method of sampling where the population is divided into subgroups (strata) based on a characteristic, and then a random sample is taken from each subgroup. This ensures that every subgroup is represented proportionally in the sample.

2. Theory with Important Formulas

- Strata: The population is divided into distinct strata (groups). Each stratum should be homogeneous with respect to the characteristic of interest but different from other strata.
- Sampling Process:
  - Divide the population into L strata.
  - Randomly sample from each stratum (either proportionally or equally).
- Proportional Stratified Sampling: If the sample size from each stratum is proportional to the stratum's size in the population:
  - Sample Size from Stratum = (Size of Stratum / Size of Population) x Total Sample Size
  - Formula:
    - n_i = (N_i / N) * n
    - Where:
      - n_i = sample size from stratum i
      - N_i = total population size of stratum i
      - N = total population size (sum of all strata)
      - n = total sample size
- Equal Allocation Stratified Sampling: Each stratum is given the same sample size, regardless of its size in the population.
- Estimating the Population Mean (for Proportional Sampling):

3. Examples

- Scenario: A company has employees across 5 different departments. To understand the average salary of all employees, the company decides to use stratified sampling.

1. Population: 500 employees across 5 departments.
2. Strata: 5 departments (e.g., Marketing, Sales, HR, Finance, Engineering).
3. Sampling: Take 20% of employees from each department as a sample.

- In this case, we ensure that each department (stratum) is properly represented in the sample, avoiding bias.

4. Practical Usages

- Healthcare Research: Ensuring that each age group, gender, or ethnicity is represented in the sample for a health study.
- Education: Selecting a sample of students from different grade levels or school types to ensure the study is representative.
- Market Research: Sampling customers from different income brackets to understand buying preferences across economic classes.

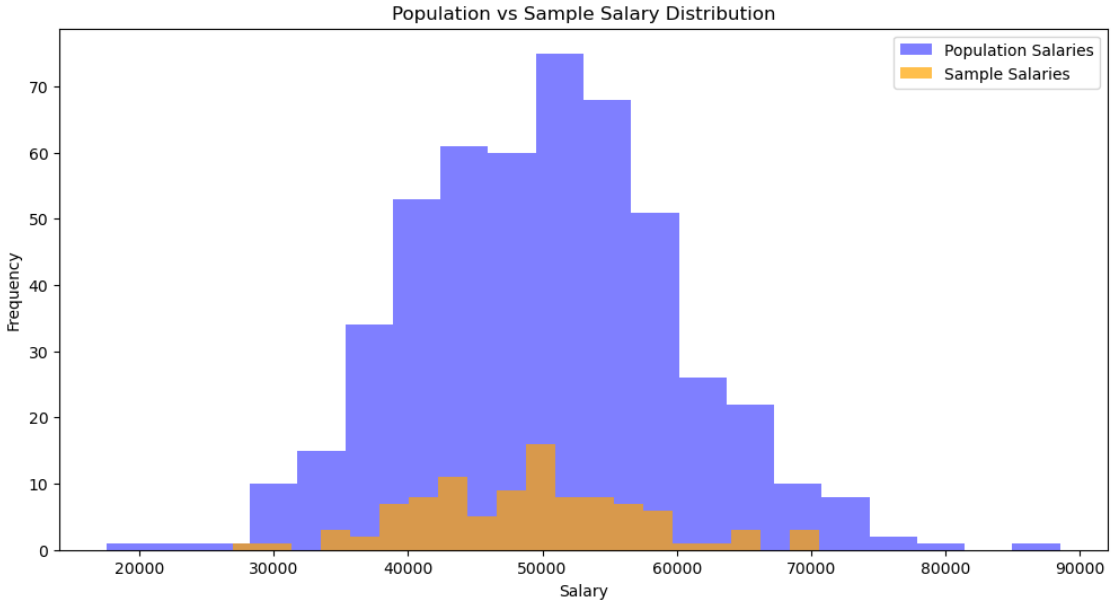
5. Python Code for Explanation and Visualization

[3]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Example: Stratified sampling for employee salary data
# Step 1: Simulating a population of employees from 5 departments
np.random.seed(42)
departments = ['Marketing', 'Sales', 'HR', 'Finance', 'Engineering']
department_sizes = [100, 150, 50, 80, 120]  # Number of employees in each␣
 ↪department
salaries = np.random.normal(50000, 10000, sum(department_sizes))  # Random␣
 ↪salaries with mean=50000, std=10000
departments_list = np.array([department for department, size in␣
 ↪zip(departments, department_sizes) for _ in range(size)])

# Create a DataFrame with employee salary data
population = pd.DataFrame({'Department': departments_list, 'Salary': salaries})

# Step 2: Stratified Sampling (take 20% from each department)
sample = population.groupby('Department').apply(lambda x: x.sample(frac=0.2,␣
 ↪random_state=42)).reset_index(drop=True)

# Print sample summary
print(sample.describe())

# Step 3: Visualizing the population and sample salary distributions
plt.figure(figsize=(12, 6))

# Plot the population salary distribution
plt.hist(population['Salary'], bins=20, alpha=0.5, label='Population Salaries',␣
 ↪color='blue')

# Plot the sample salary distribution
plt.hist(sample['Salary'], bins=20, alpha=0.7, label='Sample Salaries',␣
 ↪color='orange')

plt.title('Population vs Sample Salary Distribution')
plt.xlabel('Salary')
```

```
plt.ylabel('Frequency')
plt.legend()
plt.show()
```

```
C:\Users\rohit\AppData\Local\Temp\ipykernel_12384\2813813989.py:17:
DeprecationWarning: DataFrameGroupBy.apply operated on the grouping columns.
This behavior is deprecated, and in a future version of pandas the grouping
columns will be excluded from the operation. Either pass `include_groups=False`
to exclude the groupings or explicitly select the grouping columns after groupby
to silence this warning.
  sample = population.groupby('Department').apply(lambda x: x.sample(frac=0.2,
random_state=42)).reset_index(drop=True)
```

	Salary
count	100.000000
mean	48951.279462
std	8466.485592
min	26980.788353
25%	42716.828026
50%	49314.374935
75%	54695.573986
max	70607.479249



6. Diagram/Graph/Plot Used

- Histogram: Visualizes the salary distribution of both the population and the stratified sample.
- Box Plot: Can be used to highlight the central tendency and variability within the sample and population.

7. Additional Important Information

- Advantages:
  - More accurate estimates than simple random sampling when population is heterogeneous.
  - Ensures proportional representation from each subgroup.
- Limitations:
  - Requires knowledge of the population structure to divide it into strata.
  - Can be complex to implement if strata are not well-defined.

8. Scenario

- A university wants to understand students' satisfaction across different departments and ensure that each department is properly represented in the survey sample.

9. Problem Statement

- If the university just selects students randomly, it might over-represent or under-represent certain departments. For instance, the university might have more students in the Engineering department than in others, so random sampling could lead to biased results.

10. Solution

- By using Stratified Sampling, the university can ensure that students from all departments are properly represented, regardless of the department's size. The university can take a sample proportionate to the size of each department, ensuring that each department's student population is accurately reflected in the survey.
- Why stratified sampling? Stratified sampling guarantees that every subgroup (department) is fairly represented, leading to more reliable insights about the overall student satisfaction.

11. Alternate Solutions

- Simple Random Sampling: In cases where the departments are similar in size and characteristics, a simple random sample might still provide accurate results.
- Cluster Sampling: If it's not possible to stratify by department, the university might randomly select a few departments (clusters) and survey all students within those departments.