simpleRandomSample

November 15, 2024

1. Definition

- A Simple Random Sample (SRS) is a subset of a population in which every individual has an equal chance of being chosen. It is a foundational technique in statistics for ensuring unbiased representation of the population.

2. Theory with Important Formulas

- Key Properties:
  - Every individual in the population has an equal probability of selection.
  - Ensures that the sample is unbiased if implemented correctly.
- Sampling Process:
  - If a population consists of N individuals, and the sample size is n, the probability of any specific subset being chosen is:
    - P(Subset) = 1 / (N choose n)
    - Where:
    - N = Total population size
    - n = Sample size
    - (N choose n) = Number of ways to choose n items from N
- Other Formulas:
  - Sample Mean ($\bar{x}$):
    - $\bar{x} = (\Sigma x) / n$
    - Where:
      - Σx = Sum of all sampled values
      - n = Sample size
  - Sample Variance (s^2):
    - $s^2 = \Sigma(xi - \bar{x})^2 / (n - 1)$
    - Where:
      - xi = Individual sample value
      - $\bar{x}$ = Sample mean
      - n = Sample size
  - Sample Standard Deviation (s):
    - $s = \sqrt{s^2}$
    - Where:
      - s^2 = Sample variance
  - Population Proportion Estimate ($\hat{p}$):
    - $\hat{p} = x / n$
    - Where:
      - x = Number of items with a specific characteristic in the sample
      - n = Sample size

- Margin of Error (E) for a Confidence Interval:
  - $E = Z * (s / \sqrt{n})$
  - Where:
    - Z = Z-value (from the standard normal distribution for a given confidence level)
    - s = Sample standard deviation
    - n = Sample size
- Confidence Interval for the Population Mean ( ):
  - $CI = \bar{x} \pm E$

3. Examples

- Population: A school has 500 students.
- Sample: Select 50 students randomly for a survey about school facilities.
- Method: Assign each student a number and use a random number generator to select 50 students.

4. Practical Usages

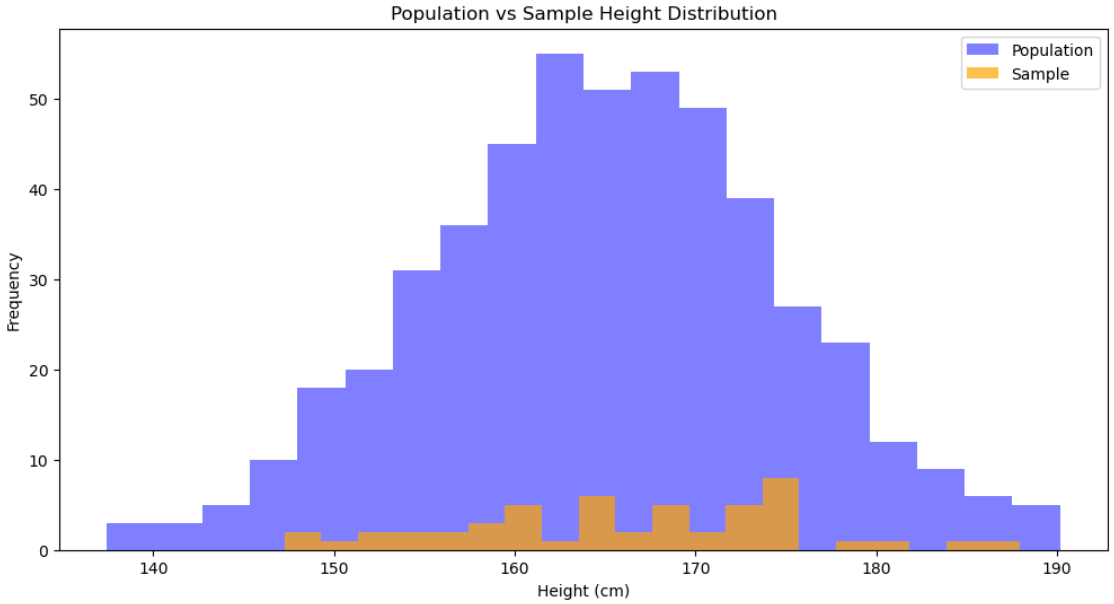
- Market Research: Collecting unbiased opinions from customers.
- Healthcare: Testing a new drug on a random subset of patients.
- Quality Control: Inspecting a random sample of products from a manufacturing line.
- Polling: Understanding political preferences by sampling voters randomly.

5. Python Code for Explanation and Visualization

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

# Example: Selecting a simple random sample from a population
# Step 1: Create a population of 500 individuals
population = pd.DataFrame({
    'ID': range(1, 501),
    'Age': np.random.randint(18, 60, 500),
    'Height': np.random.normal(165, 10, 500)  # Mean = 165 cm, Std Dev = 10 cm
})

# Step 2: Take a simple random sample of size 50
sample_size = 50
random_sample = population.sample(n=sample_size, random_state=42)

# Print sample summary
print("Sample Summary:")
print(random_sample.describe())

# Step 3: Visualize population vs sample distribution
plt.figure(figsize=(12, 6))
plt.hist(population['Height'], bins=20, alpha=0.5, label='Population',␣
 ↪color='blue')
```

```
plt.hist(random_sample['Height'], bins=20, alpha=0.7, label='Sample',␣
 ↪color='orange')
plt.title('Population vs Sample Height Distribution')
plt.xlabel('Height (cm)')
plt.ylabel('Frequency')
plt.legend()
plt.show()
```

```
Sample Summary:
                ID         Age      Height
count    50.00000   50.000000   50.000000
mean    254.86000   36.940000  166.143609
std     171.07726   10.788145    9.412016
min       1.00000   19.000000  147.250057
25%      79.75000   29.000000  159.557968
50%     299.00000   36.500000  166.991778
75%     404.00000   43.750000  173.284875
max     498.00000   58.000000  187.963500
```



6. Diagram/Graph/Plot Used

- Histogram: Shows the comparison between the population and the sample distribution.
- Box Plot: Highlights the spread and central tendency of the sampled data compared to the population.

7. Additional Important Information

- Advantages:
  - Easy to implement and understand.

  - Provides an unbiased representation of the population if sample size is adequate.
- Limitations:
  - Requires a complete list of the population.
  - May not account for variability within subgroups (solution: stratified sampling).

8. Scenario

- A university wants to understand the average height of students across all departments.

9. Problem Statement

- It is impractical to measure the height of all 20,000 students due to time and resource constraints. The university needs a reliable method to estimate the average height.

10. Solution

- By using a Simple Random Sample, the university can select 500 students randomly and calculate the average height from the sample. This approach ensures that every student has an equal chance of being chosen, providing an unbiased estimate of the population average.
- Why this topic?
  - SRS is easy to implement, unbiased, and ensures that the sample represents the population fairly.

11. Alternate Solutions

- Stratified Sampling: If departments have uneven student numbers, divide the population by department and sample proportionally.
- Systematic Sampling: Select every k-th student from a sorted list.
- Cluster Sampling: Divide students into clusters (e.g., dormitories) and randomly select clusters to measure.