biasSampling

November 15, 2024

1. Definition

- Bias Sampling refers to a situation where the method of selecting a sample leads to a sample that is not representative of the population. In other words, some members of the population are systematically more likely to be included than others, which leads to biased results.
- There are different types of biases that can occur in sampling, including but not limited to:
  - Selection Bias: When certain individuals or groups are more likely to be selected for the sample than others.
  - Non-response Bias: When certain individuals chosen for the sample do not respond or participate.
  - Undercoverage Bias: When some members of the population are not represented in the sampling frame.
  - Overcoverage Bias: When members outside the target population are included in the sample.

2. Theory with Important Formulas

- Bias = Bias = Sample Estimate - True Value
- Selection Bias: Occurs when the method of selecting a sample systematically favors certain members of the population over others.
- Non-response Bias: Arises when those who do not respond to a survey or study differ in important ways from those who do respond.
- Undercoverage Bias: Happens when certain subgroups are not adequately represented in the sample.

3. Examples

- Example 1:
  - A market research company wants to estimate the average income of people living in a city. They decide to conduct a survey by calling people listed in a phone directory. However, people who do not have landline phones are excluded, which could lead to selection bias.
- Example 2:
  - An election poll surveys 1,000 people about their voting preferences but only surveys those who can be reached through online surveys. If the sample is not diverse (e.g., older people are less likely to respond to online surveys), this leads to non-response bias.

4. Practical Usages

- Bias sampling can severely affect the reliability of results in various fields:

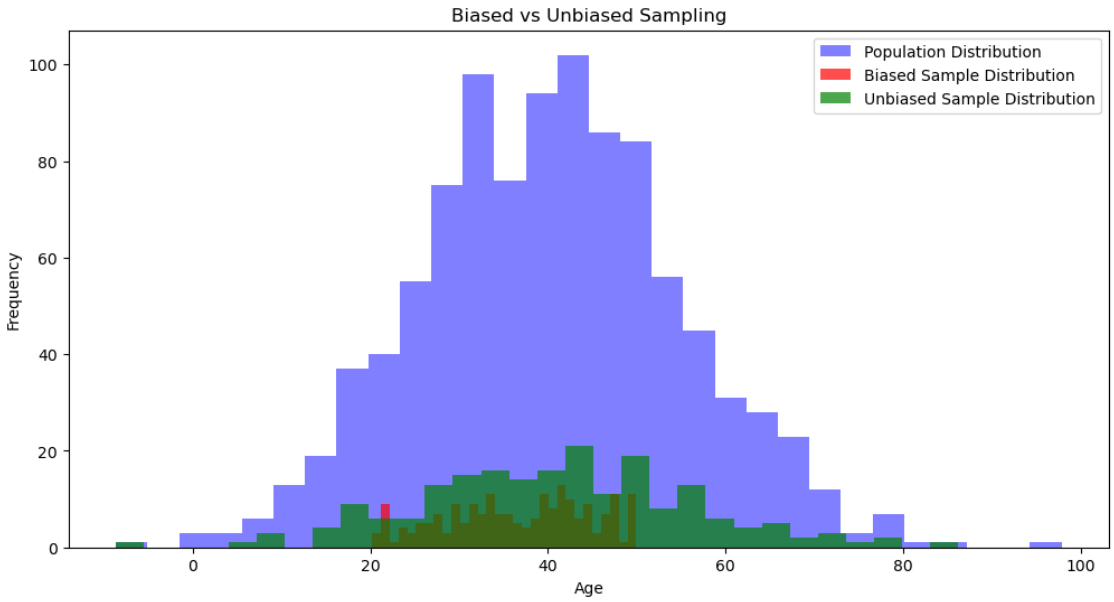
- Healthcare Studies: Bias can occur when certain groups (e.g., elderly or disabled) are underrepresented in clinical trials, which can result in inaccurate conclusions about the efficacy of treatments.
- Market Research: If a company only surveys a certain demographic (e.g., only high-income individuals), the results won't reflect the entire target population.
- Political Polls: If a poll excludes certain groups (e.g., individuals without internet access), the poll results might not be an accurate representation of the broader population.

5. Python Code for Explanation and Visualization

- In this Python code, we simulate a biased sampling process and compare it with unbiased sampling.

[2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Simulating a population with a biased distribution
np.random.seed(42)
population_size = 1000
age_population = np.random.normal(40, 15, population_size)  # Mean = 40, SD = 15

# Simulate biased sampling where we only sample from ages between 20 and 50␣
 ↪(biased sample)
biased_sample = np.random.choice(age_population[(age_population >= 20) &␣
 ↪(age_population <= 50)], size=200, replace=False)

# Simulate unbiased sampling (random sample)
unbiased_sample = np.random.choice(age_population, size=200, replace=False)

# Plotting the results
plt.figure(figsize=(12, 6))

# Plotting the population distribution
plt.hist(age_population, bins=30, alpha=0.5, label='Population Distribution',␣
 ↪color='blue')

# Plotting the biased sample distribution
plt.hist(biased_sample, bins=30, alpha=0.7, label='Biased Sample Distribution',␣
 ↪color='red')

# Plotting the unbiased sample distribution
plt.hist(unbiased_sample, bins=30, alpha=0.7, label='Unbiased Sample␣
 ↪Distribution', color='green')

plt.title('Biased vs Unbiased Sampling')
plt.xlabel('Age')
plt.ylabel('Frequency')
```

```
plt.legend()
plt.show()
```



6. Diagram/Graph/Plot Used

- Histogram: The histogram will compare the population distribution, biased sample distribution, and unbiased sample distribution. The biased sample will show a restricted range of ages, while the unbiased sample will closely mirror the population.

7. Additional Important Information

- Minimizing Bias:
  - Random Sampling: Ensures every individual has an equal chance of being selected.
  - Stratified Sampling: When certain groups are likely to be underrepresented, stratified sampling ensures all subgroups are - properly represented.
  - Ensuring Non-Response: Offering multiple response methods (e.g., email, phone) and following up with non-respondents can reduce non-response bias.
- Common Types of Bias:
  - Response Bias: When respondents provide inaccurate answers (e.g., social desirability bias).
  - Measurement Bias: When the method of data collection leads to systematic errors (e.g., a miscalibrated instrument).

8. Scenario

- A research institute wants to understand the average income of people living in a city. The sampling method used involves sending surveys via email to a list of people who have previously participated in other studies. The institute does not send the survey to anyone who hasn't been in their previous surveys.

9. Problem Statement

- The institute's sampling method could lead to selection bias. By only surveying those who have previously participated in studies, they might exclude certain demographic groups, such as younger or less affluent individuals, who are less likely to respond to surveys. This could result in an inaccurate estimate of the city's average income.

10. Solution

- To minimize bias, the institute should use random sampling from the entire population. This way, every individual in the population has an equal chance of being selected. Additionally, they can offer multiple ways to respond to the survey (e.g., email, phone, in-person) to reduce non-response bias.
- Why Bias sampling? This topic is crucial because bias in sampling leads to unreliable data, and using methods like random sampling can help ensure that the data represents the broader population more accurately.

11. Alternate Solutions

- Stratified Sampling: If the population is diverse and certain groups are underrepresented, stratified sampling ensures that all groups are represented.
- Cluster Sampling: If the population is large and geographically spread out, cluster sampling can be more practical, although it may still introduce biases if clusters are not homogeneous.