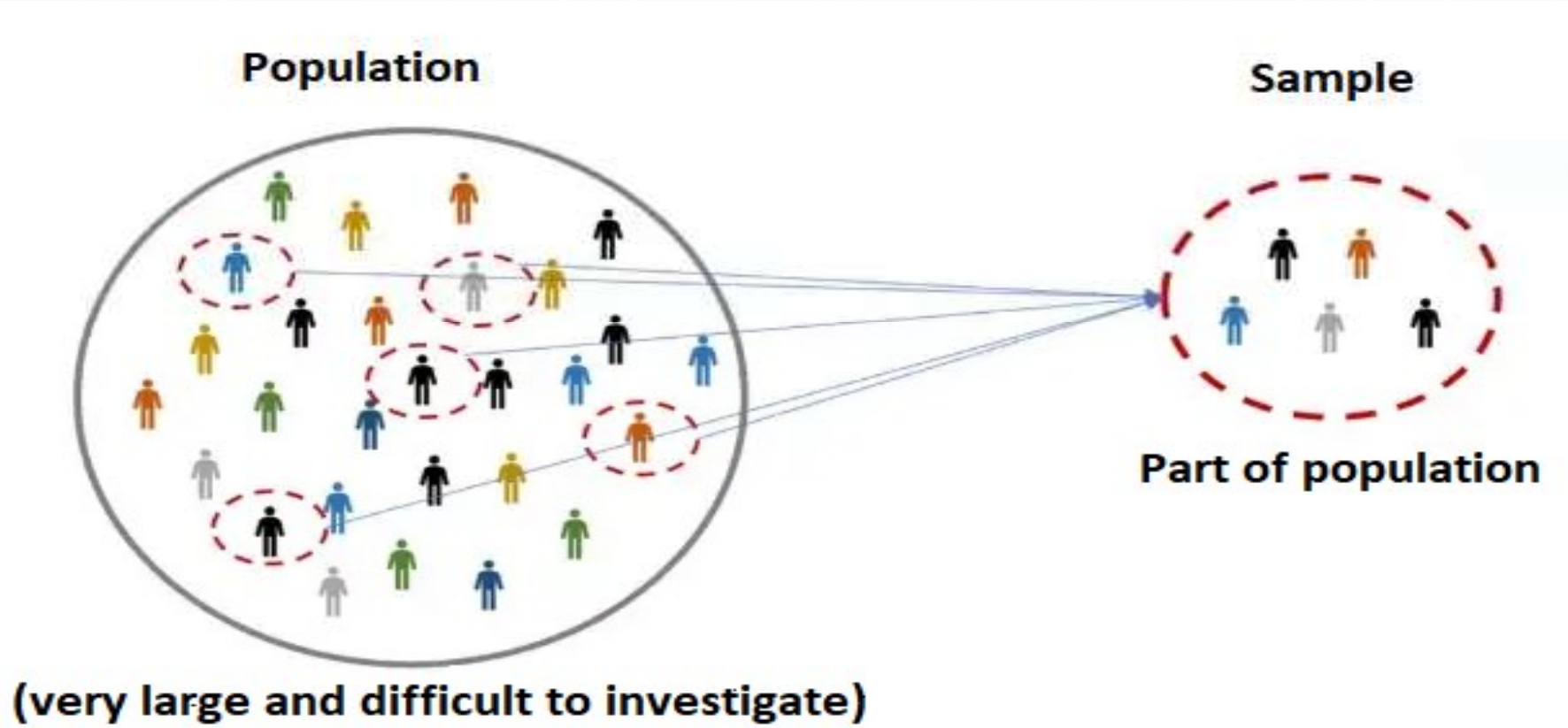
Sampling

By
Dr. Anjula Mehto
Assistant Professor
Thapar Institute of Engineering and Technology,
Patiala, Punjab
Email-anjula.mehto@thapar.edu

Sampling


Population
Sample
Part of population
(very large and difficult to investigate)

Sampling

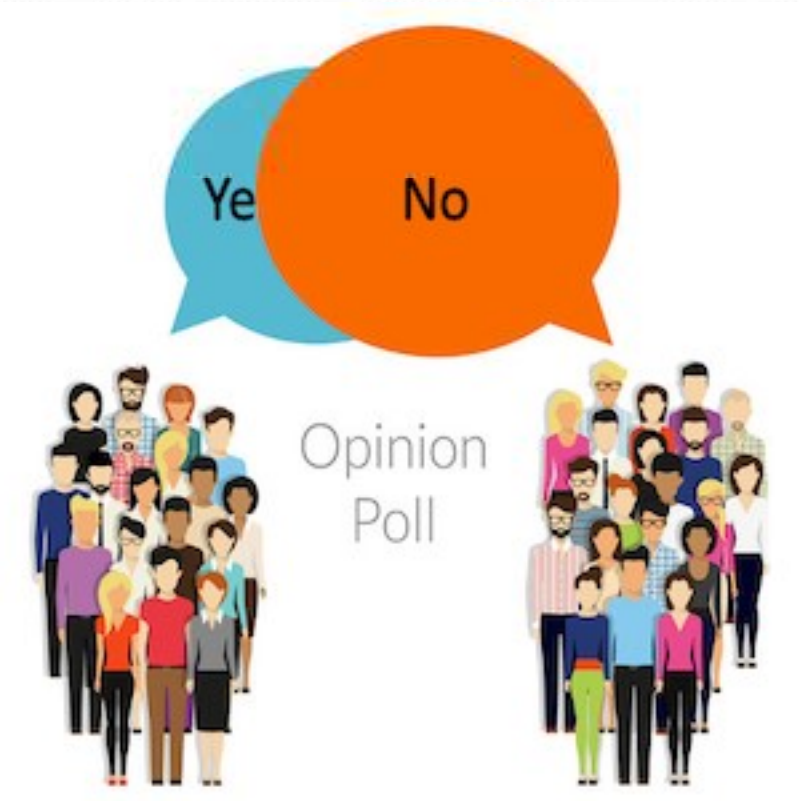
- A process used in statistical analysis.
- Process of selecting a portion or subset (sample) of the population to represent the entire population.
- Sample- A representative subset of a population.
- Representative subset accurately reflects the characteristics of the population as a whole.

Examples

1. Find the average height of all adults in this class.
2. Find the average height of all adult females in Patiala.

Examples

Surveys and Opinion Polls:



Examples

Medical Studies:

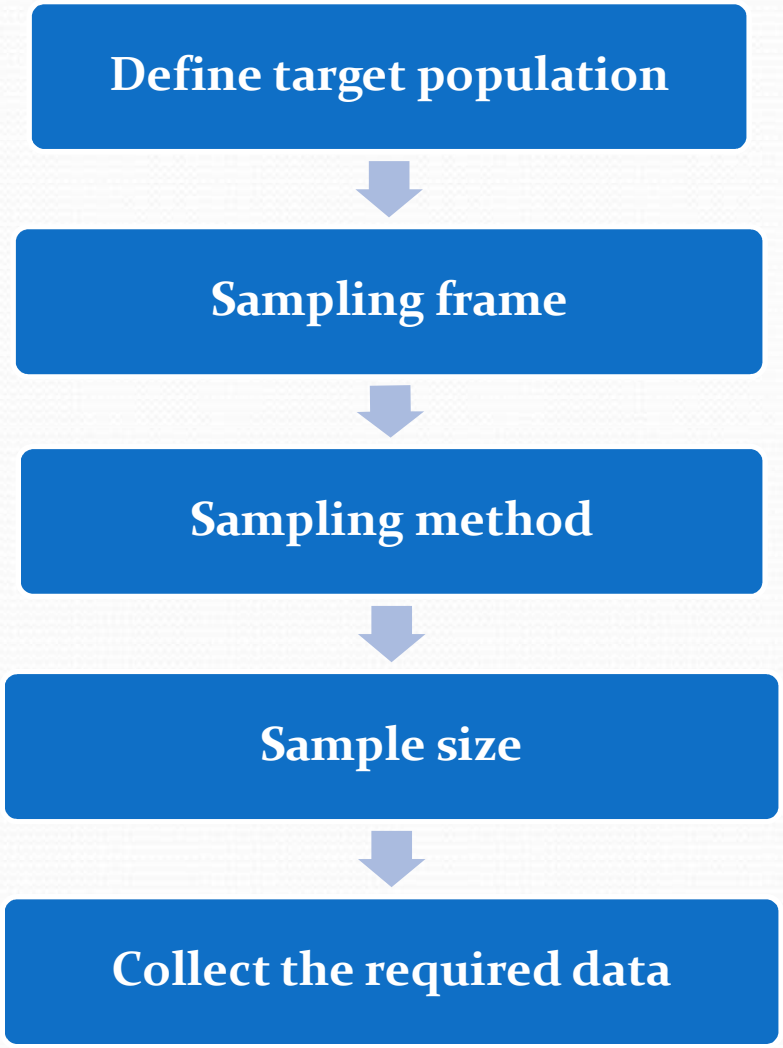
- Sampling a group of patients to study the effectiveness of a new treatment.

Examples

Agriculture:

- Sampling crops to determine their yield and quality.

Steps Involved in Sampling



Steps Involved in Sampling

1. Define Target Population:
   - Based on the objective of the study, clearly target the population.
   - Example- For regional election, the target population would be all people who are domiciled in the region and eligible to vote.

Steps Involved in Sampling

2. Define Sampling Frame:
   - A subset of the population that defines the pool from which a sample to be selected.
   - Example- The sampling frame would consist of all the people from the population who are in the state and can participate in the study.
3. Select Sampling Technique:
   - Select an appropriate sampling technique.

Steps Involved in Sampling

4. Determine Sample Size:
   - Refers to the number of items or elements selected from sample frame.
   - If the sample size is too small, the sample may not accurately represent the population.
   - If the sample size is too large, it can be inefficient and time-consuming to collect and analyze the data.

Steps Involved in Sampling

- The appropriate sample size depends on factors such as:
  - The size of the population
  - The confidence level and
  - The margin of error

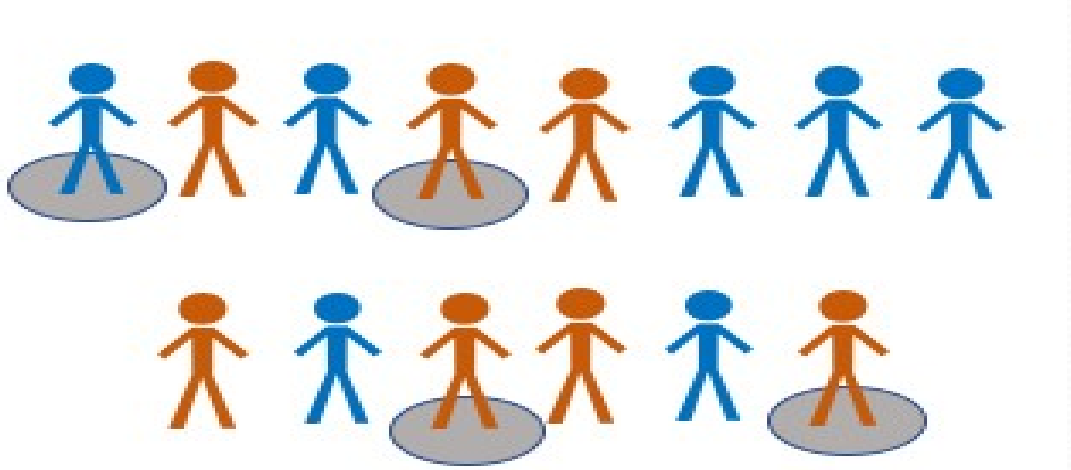
5. Data collection:
   - Once the sample has been selected, the next step is to collect data from the sample.

Types of Sampling

- Simple Random Sampling
- Systematic Sampling
- Stratified Sampling
- Cluster Sampling

Simple Random Sampling

- Pick the sample, at random.
- Applicable when population is small, and homogeneous.
- Example- I want to ask a question in this class, then I can randomly pick any student



Simple Random Sampling

Advantages:

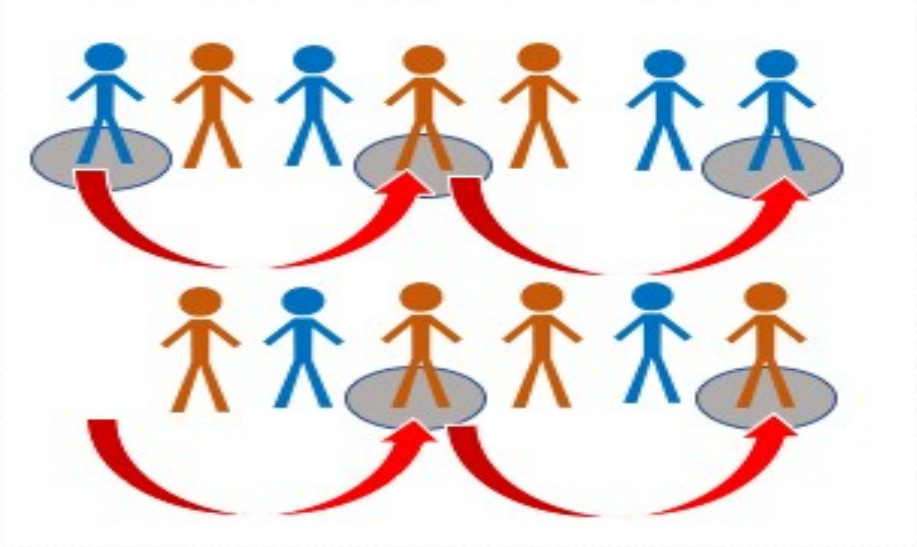
- Easy calculation.
- Every participant has an equal probability of being selected

Disadvantages:

- If sampling frame is large, then it is impracticable.
- Minority subgroups of interest in population may not be present in sample.

Systematic Sampling

- The samples are chosen at regular intervals.
- A random start and then proceeds with the selection of every ‘kth’ element.
- Example- Select every 10th name from the telephone directory.



Systematic Sampling

Advantages:

- Samples are easy to select.
- Sample evenly spread over entire population.

Disadvantages:

- Sample may be biased if hidden periodicity in population coincides with that of selection.

Stratified Sampling

- The population is divided into subgroups or strata based on a certain characteristic or criterion (age groups, ethnic origin, and gender).
- Each stratum is then sampled as an independent sub-population.
- Individual elements from a sub-population can be randomly selected.

Stratified Sampling

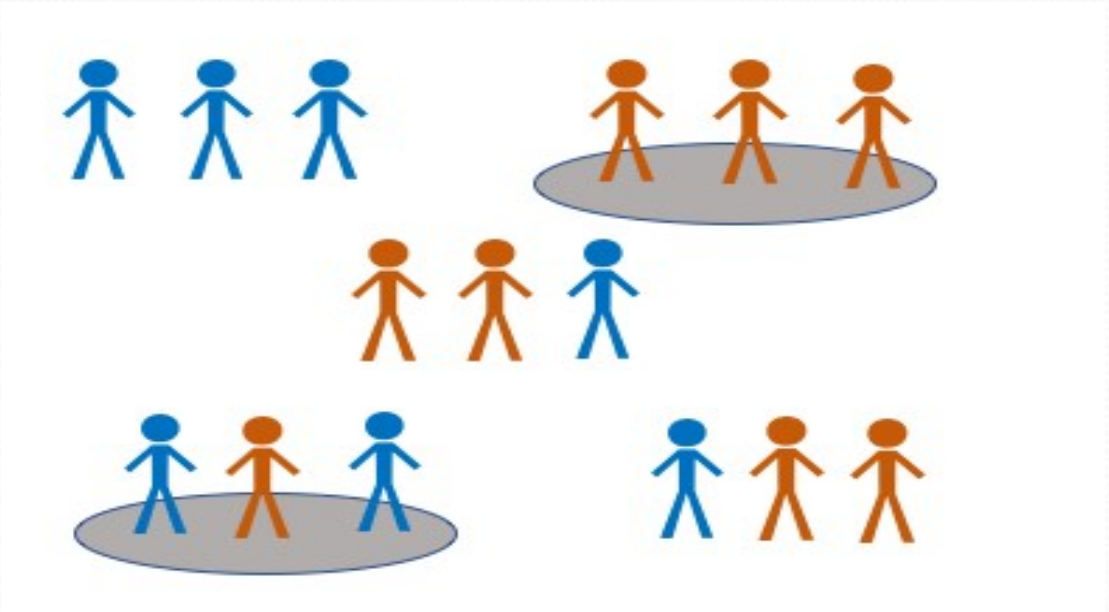
- Example-
- If a company has 500 male employees and 100 female employees.
- To ensure that the sample reflects the gender as well use stratified sampling

Stratified Sampling

- **Advantages:**
  - Improves the representativeness of the sample
  - Increases efficiency
- **Disadvantages:**
  - Requires prior knowledge
  - Can be complex
  - Can be time-consuming

Cluster Sampling

- The entire population is divided into smaller groups or clusters, and then a random sample of these clusters is selected.
- The sample size is then determined based on the number of clusters selected.



Example

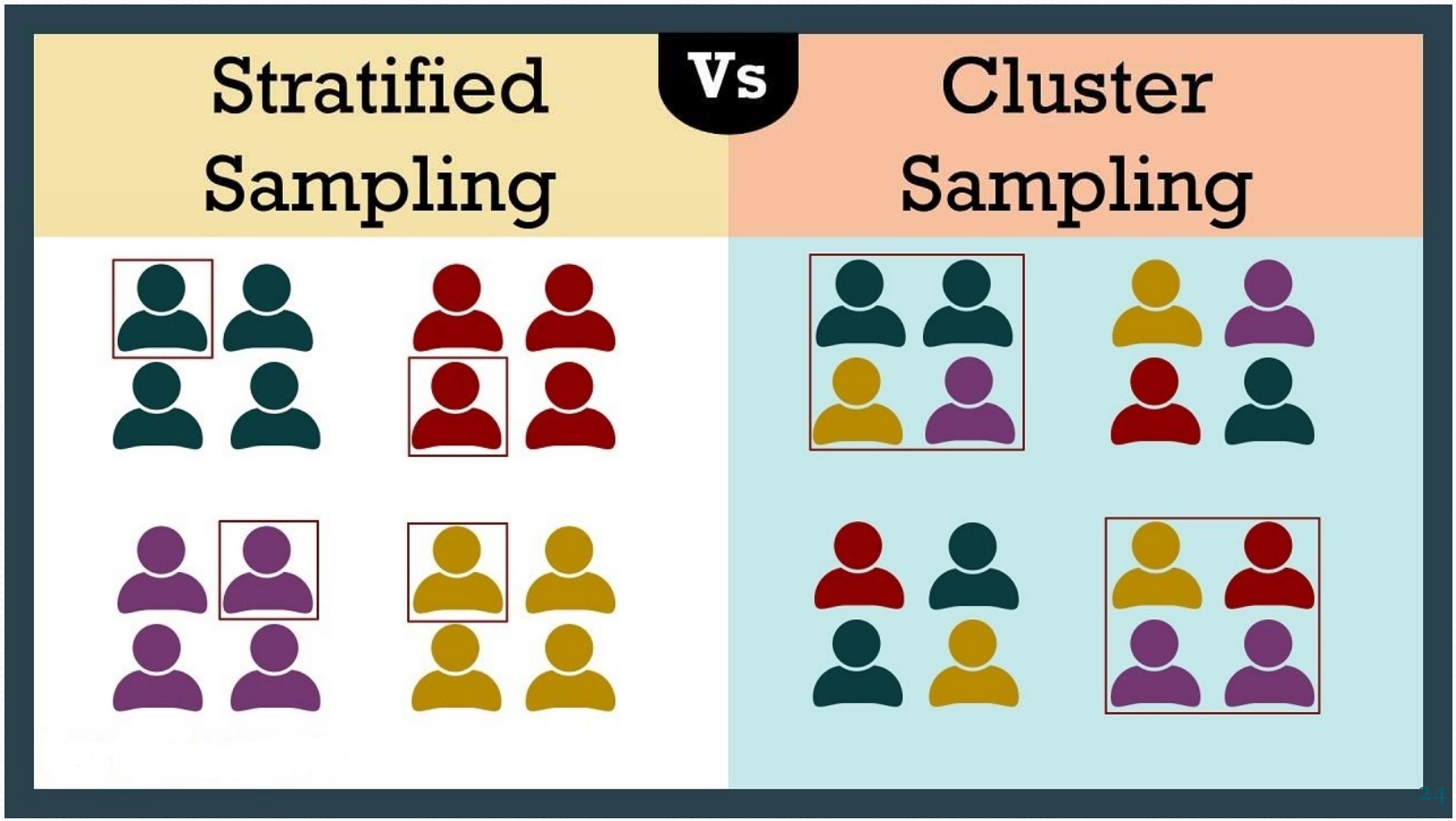
- Sample a population of 10,000 schools to estimate the average number of students per school.
- Divide the schools into smaller groups or clusters based on geographical locations, for example, schools in each state.
- Now, instead of visiting all 10,000 schools, select a random sample of 5 states, and visit all the schools within those 5 states.
- The data collected from the schools within the 5 selected states will then be used to estimate the average number of students per school for the entire population of 10,000 schools.

Cluster Sampling

- **Advantages:**
  - Cuts down on the cost of preparing a sampling frame.
- **Disadvantages:**
  - Can introduce bias: If the clusters are not representative of the population, leading to inaccurate results.

Cluster Vs Stratified Sampling

- The main difference between these two methods is how the sample is selected.



Which Sampling Technique to Use ?

- Size of the data set:
  - Small data sets- Simple random sampling or Systematic Sampling.
  - Large data sets- May use a more complex technique like Stratified sampling or Cluster sampling.

Which Sampling Technique to Use ?

- Nature of the data:
  - For well-defined structured data, Stratified Sampling may be a good choice.
  - While for more complex data sets, Cluster or Systematic Sampling may be better.

Which Sampling Technique to Use ?

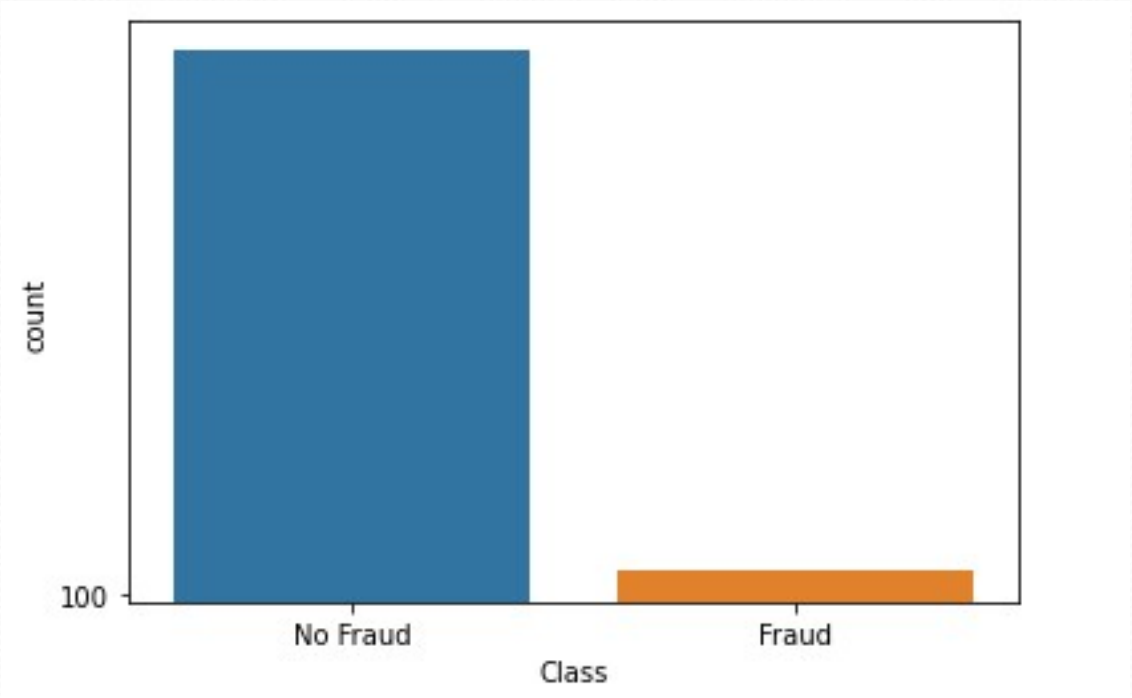
- Cost and feasibility:
- Simple random sampling may be the most straightforward and cost-effective.
- While more complex techniques like Cluster or Stratified sampling may be more expensive and require more resources.

Why need Sampling?

- Reduces computational cost.
- Reduces processing time of the entire dataset.
- Provides Balance class distribution or treats imbalanced classes.
- Avoid Overfitting.

Imbalanced classes

- Example: To detect fraudulent credit card transactions.
- Suppose your model gives 94% accuracy.



Imbalanced classes

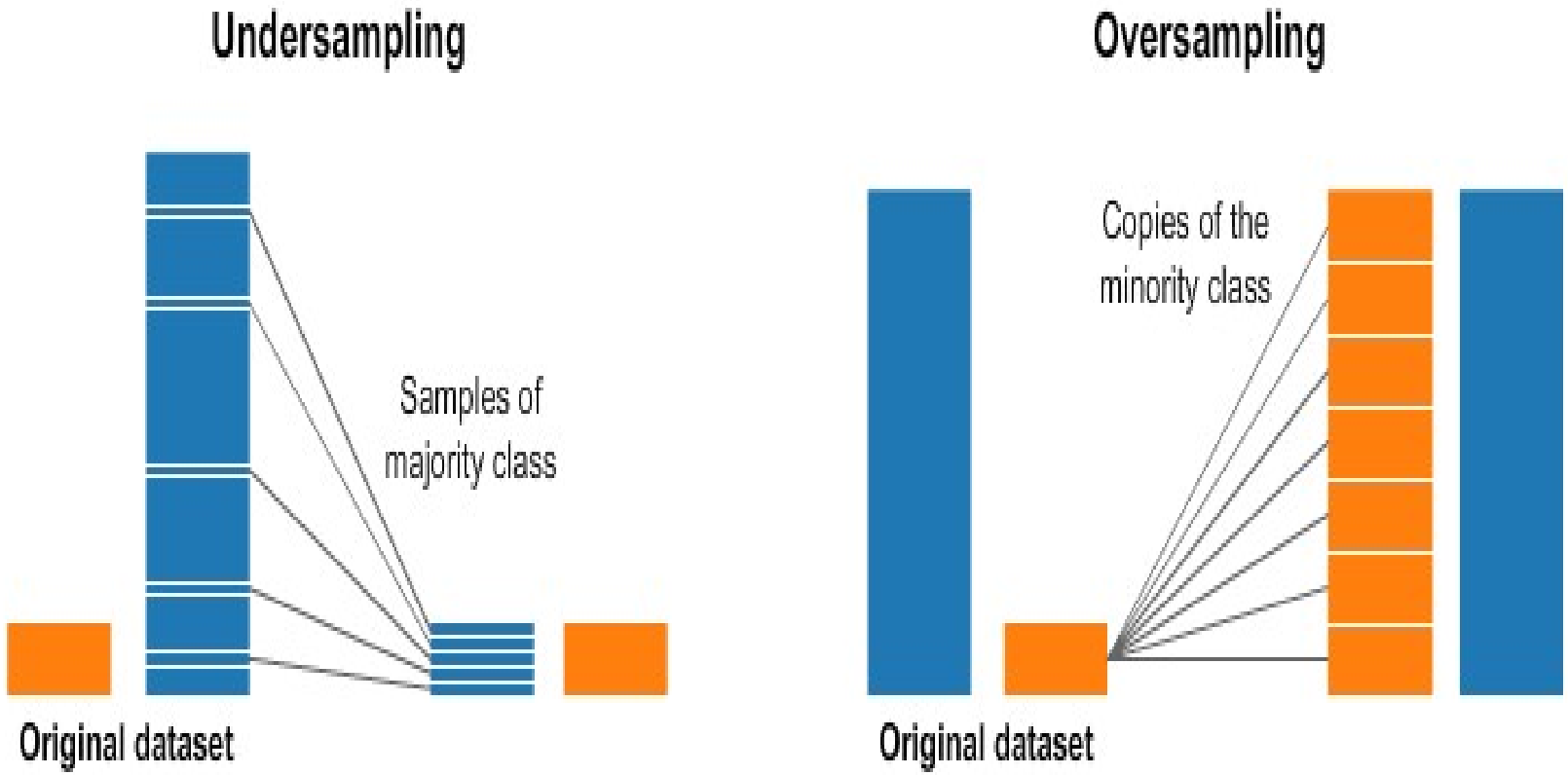
- Unfortunately, that accuracy (94%) is misleading:
  - All those **non-fraudulent** transactions, would have "100%" accuracy.
  - Those transactions which are **fraudulent**, would have "0%" accuracy.
  - Your overall **accuracy would be high** simply because the most transaction is not fraudulent.

How to Balance Class Distribution?

- Use techniques such as oversampling the minority class or
- Use undersampling the majority class.

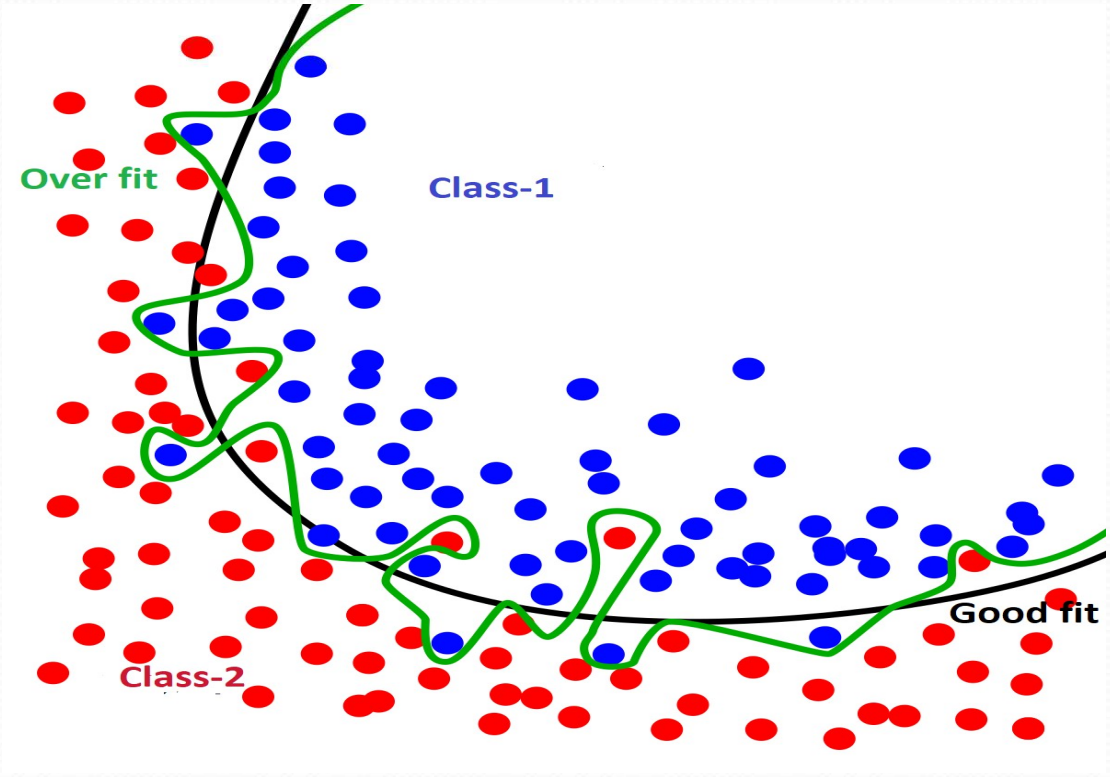
How to Balance Class Distribution?

Undersampling and Oversampling

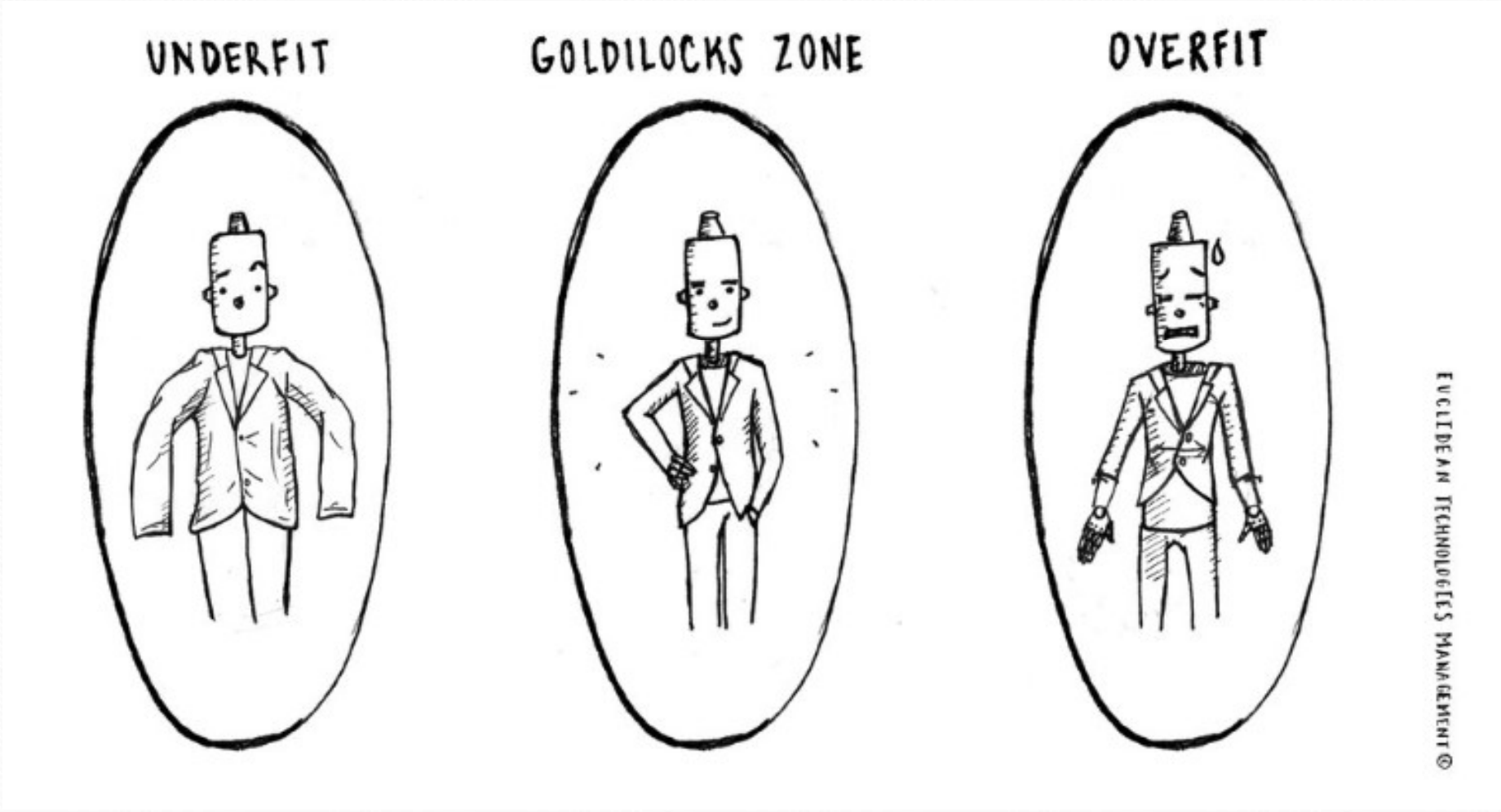


Overfitting

- When a model is too closely aligned to a limited set of data points, but fails to generalize to new data.

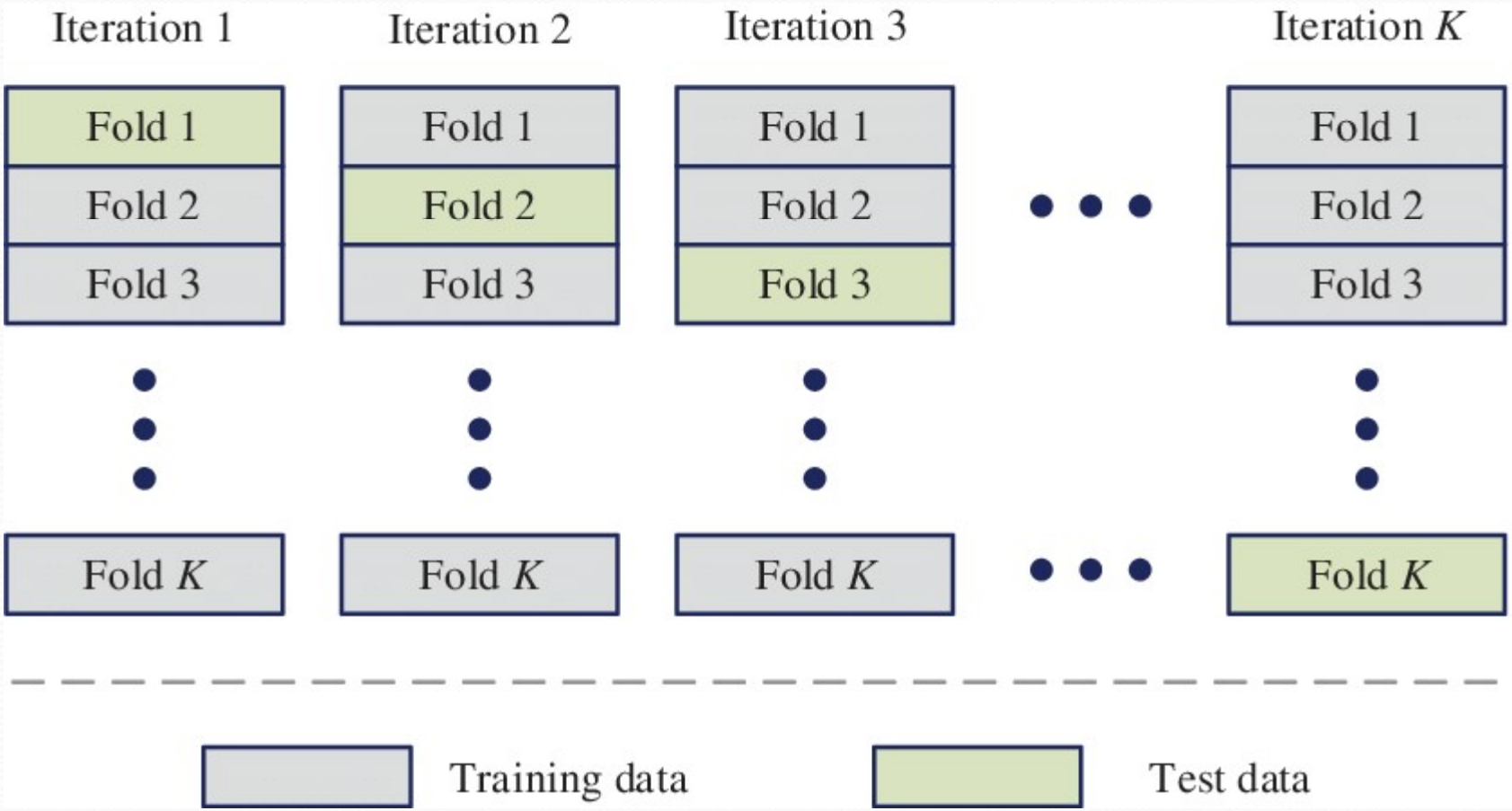


Overfitting vs Underfitting



How to Reduce Overfitting

K-fold cross-validation



K-fold Cross-Validation

- The original dataset is divided into ‘k’ subsets or folds of roughly equal size.
- The model is trained on ‘k-1’ folds.
- The model is evaluated on the remaining fold.
- This process is repeated ‘k’ times so that each fold is used for evaluation once.
- The average performance across all ‘k’ iterations is used to evaluate the model's overall performance.

How to Check Goodness of a Sample?

Central tendency measures:

- Describe the center or middle of the sample data.
- Common measures of central tendency include the mean, median, and mode.

How to Check Goodness of a Sample?

Central tendency measures:

- Example: Dataset- {1, 1, 1, 2, 2}
- The mean: (1 + 1 + 1 + 2 + 2) / 5 = 7 / 5 = 1.4
- The mode: 1
- The median: 1
- Mean, Median, and Mode are similar, it indicates that the sample is representative of the population

How to Check Goodness of a Sample?

Dispersion measures:

- Describe how spread out the sample data is.
- Common measures of dispersion include range, variance, and standard deviation.

How to Check Goodness of a Sample?

Dispersion measures:

- Example: Dataset- {1, 1, 1, 2, 2}
- The range- 1
- The variance: 1.6
- The standard deviation: 0.6
- The range, variance, and standard deviation are small, it indicates that the sample is tightly grouped and has a good dispersion

How to Check Goodness of a Sample?

Normality tests:

- Checks whether the sample data follows a normal distribution.
- Common normality tests include the Shapiro-Wilk test, and the Anderson-Darling test.
- Both the tests can be implemented using the scipy.stats library in Python.

How to Check Goodness of a Sample?

On the basis of type of dataset

- Numerical data:
  - Can use measures of central tendency, dispersion, and normality.
- Categorical data:
  - Can use measures such as relative frequencies or proportions.

How to Check Goodness of a Sample?

On the basis of type of dataset:

- Time series data:
  - Can use measures such as autocorrelation and partial autocorrelation.
  - **statsmodels** library in Python provides functions to calculate these measures.

How to Check Goodness of a Sample?

On the basis of type of dataset:

- Image or multimedia data:
  - Can use measures such as image quality metrics, such as mean squared error or structural similarity index.
  - **scikit-image** library in Python provides functions to calculate these metrics.

How to Check Goodness of a Sample?

1. Suppose there is a dataset

[3, 1, 6, 2, 4, 4, 2, 0, 1, 6, 4, 6, 4, 5, 9, 6, 8, 9, 7, 7] with four samples: [6, 2, 2, 4, 9], [7, 2, 8, 2, 9], [7, 8, 9, 5, 4], [1, 2, 0, 1, 3], Which one is a good sample?)

How to Check Goodness of a Sample?

- Range- Difference between the largest and smallest values.
- Variance = $(1 / (n - 1)) * \sum(x_i - \bar{x})^2$

  where

  x_i represents each individual data point,

  $\bar{x}$ represents the sample mean,

  n data points in the sample
- Standard Deviation- The square root of the variance

How to Check Goodness of a Sample?

Samples 1: [6, 2, 2, 4, 9]

Sample Mean: 4.6

Sample Median: 4.0

Sample Mode: 2

Sample Range: 7

Sample Variance: 7.0

Sample Standard Deviation: 2.7

How to Check Goodness of a Sample?

Samples 2: [7, 2, 8, 2, 9]

Sample Mean: 5.6

Sample Median: 7.0

Sample Mode: 2

Sample Range: 7

Sample Variance: 9.0

Sample Standard Deviation: 3.0

How to Check Goodness of a Sample?

Samples 3: [7, 8, 9, 5, 4]

Sample Mean: 6.6

Sample Median: 7.0

Sample Mode: 4

Sample Range: 5

Sample Variance: 3.44

Sample Standard Deviation: 1.8

How to Check Goodness of a Sample?

Samples 4: [1,2,0,1,3]

Sample Mean: 1.4

Sample Median: 1.0

Sample Mode: 1

Sample Range: 3

Sample Variance: 1.04

Sample Standard Deviation: 1.0

How to Check Goodness of a Sample?

- **Answer- Sample 4 is good.**

How to Check Goodness of a Sample?

2. Data set [80, 41, 95, 53, 82, 33, 84, 32, 7, 27, 30, 23, 40, 20, 44, 16, 19, 28, 21, 97] with 4 samples [16, 28, 82, 80, 97], [30, 53, 84, 33, 97], [19, 7, 41, 21, 28], [7, 19, 80, 53, 27]. Which one is good?

How to Check Goodness of a Sample?

Sample 1

Sample Mean: 60.6, Sample Median: 80.0 Sample Mode: 16, Sample Range: 81, Sample Variance: 1042.24, Sample Standard Deviation: 32.2

Sample-2

Sample Mean: 59.4, Sample Median: 53.0, Sample Mode: 30, Sample Range: 67, Sample Variance: 724.24, Sample Standard Deviation: 26.91

How to Check Goodness of a Sample?

Sample 3

Sample Mean: 23.2, Sample Median: 21.0, Sample Mode: 7, Sample Range: 34, Sample Variance: 124.96, Sample Standard Deviation: 11.17

Sample-4

Sample Mean: 37.2, Sample Median: 27.0, Sample Mode: 7, Sample Range: 73, Sample Variance: 685.76, Sample Standard Deviation: 26.18

Calculate Size of a Sample

- **Simple Random Sampling:**

$$n = (Z^2 * p * (1-p)) / E^2$$

where

- n - sample size
- Z - Z-score corresponding to the desired level of confidence (e.g. 1.96 for 95% confidence)
- p - estimated proportion of the population with a certain characteristic (often assumed to be 0.5)
- E - desired margin of error

Calculate Size of a Sample

- **Stratified Sampling:**

$$n = (Z^2 * p * (1-p)) / (E/S)^2$$

where

- n - sample size,
- Z - Z-score corresponding to the desired level of confidence (e.g. 1.96 for 95% confidence)
- p - estimated proportion of the population with a certain characteristic (often assumed to be 0.5),
- E - desired margin of error
- S - number of strata.

Calculate Size of a Sample

- **Cluster Sampling:**

$$n = (Z^2 * p * (1-p)) / (E/C)^2$$

where

- n - sample size,
- Z - Z-score corresponding to the desired level of confidence (e.g. 1.96 for 95% confidence)
- p - estimated proportion of the population with a certain characteristic (often assumed to be 0.5),
- E - desired margin of error
- C - average size of the clusters.

Calculate Size of a Sample

- Consider a population of 10,000 individuals, and a desired confidence level of 95% and margin of error of 5%. Determine the sample size.
- **Solution- Simple random sampling:**

  n = (Z^2 * p * (1-p)) / E^2

  n = (1.96^2 * 0.5 * (1 - 0.5)) / (0.05^2) = 384.16

  n=384.16

  n=384