Statistics Essentials for Data Science

Sampling and Sampling Techniques

Learning Objectives

By the end of this lesson, you will be able to:

- Understand the concept of sampling along with its advantages and disadvantages
- Identify various sampling techniques
- Discover the process employed in non-probability sampling methods
- Differentiate between the situations when probability and non-probability sampling are applied
- Clarify the role of probability distribution in sampling

Business Scenario

ABC is a government organization that stores and maintains an extensive dataset on a country's population. The organization is struggling to segment the data, as it is scattered and does not provide any valuable knowledge.

To address this issue, ABC plans to categorize the data by city and analyze it to predict population trends in the coming years.

In this endeavor, the organization will investigate various methods including:

- Exploring data sampling
- Implementing probability sampling
- Utilizing systematic sampling

This approach aims to enhance their analytical work and extract meaningful insights from the dataset.

Introduction to Sampling and Sampling Errors

Discussion

Discussion: Data Sampling

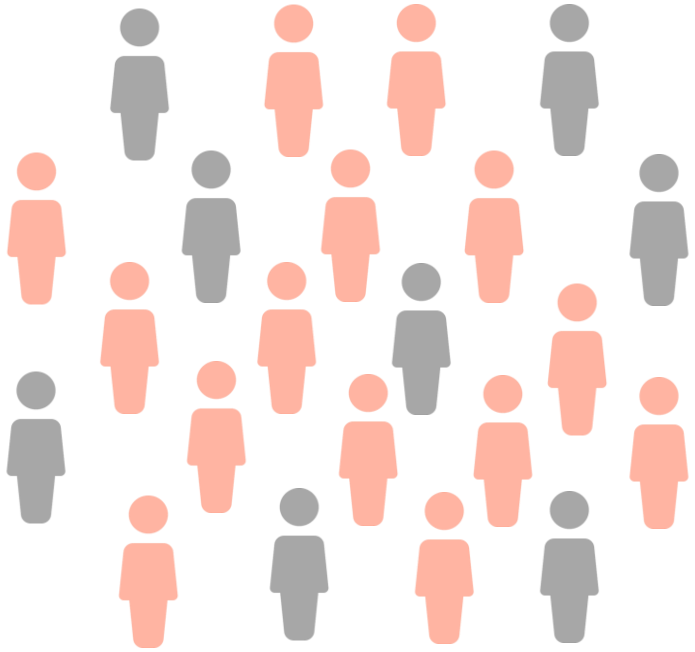
Duration: 15 minutes

How does sampling of data help in decision-making?

- What does data sampling mean?
- What are the advantages of sampling?

Sampling

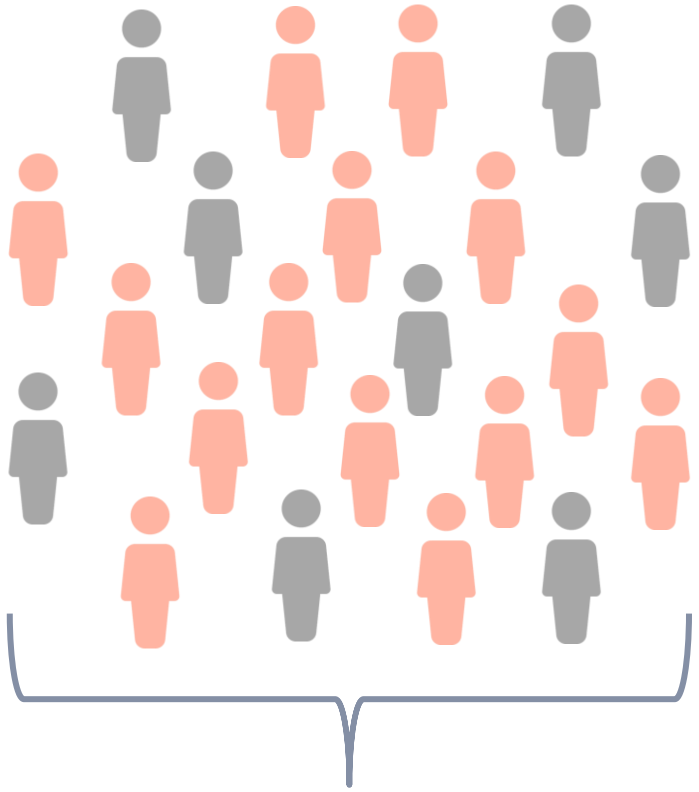
Sampling is the process of selecting a subset of a population to represent the entire population.



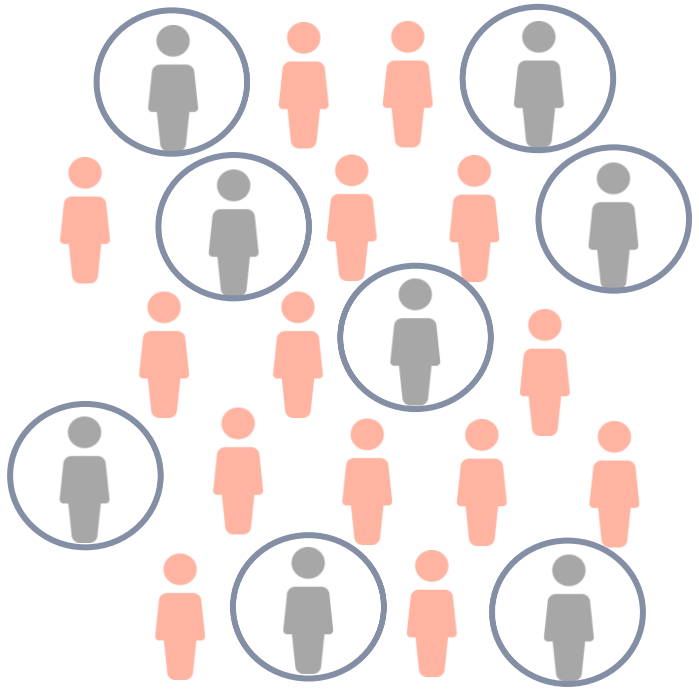
In statistics, sampling allows for the testing of a hypothesis about a population's characteristics.

Sampling

Consider a sample that is adequate to make certain conclusions about the entire group



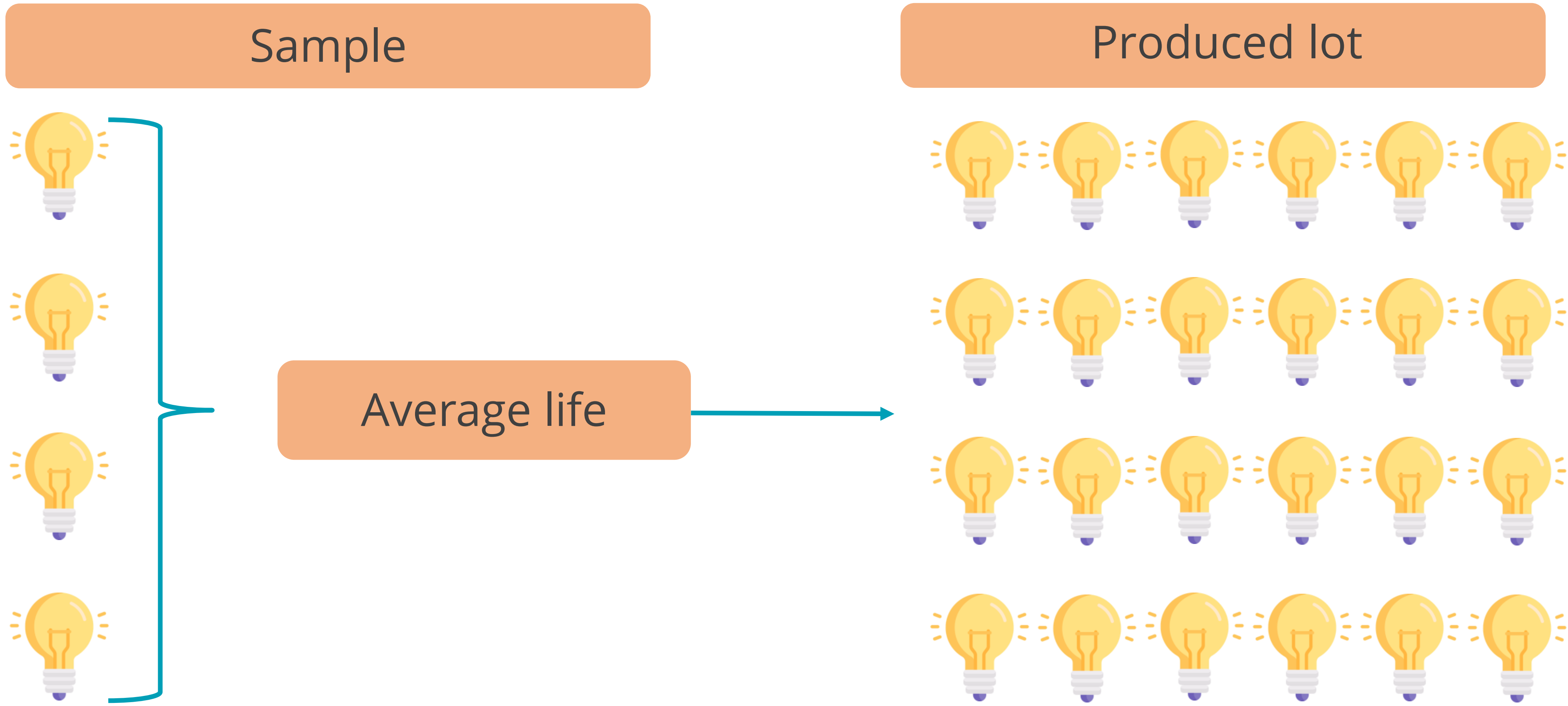
Representative sample of the population



Different samples can provide varied numerical results

Sampling

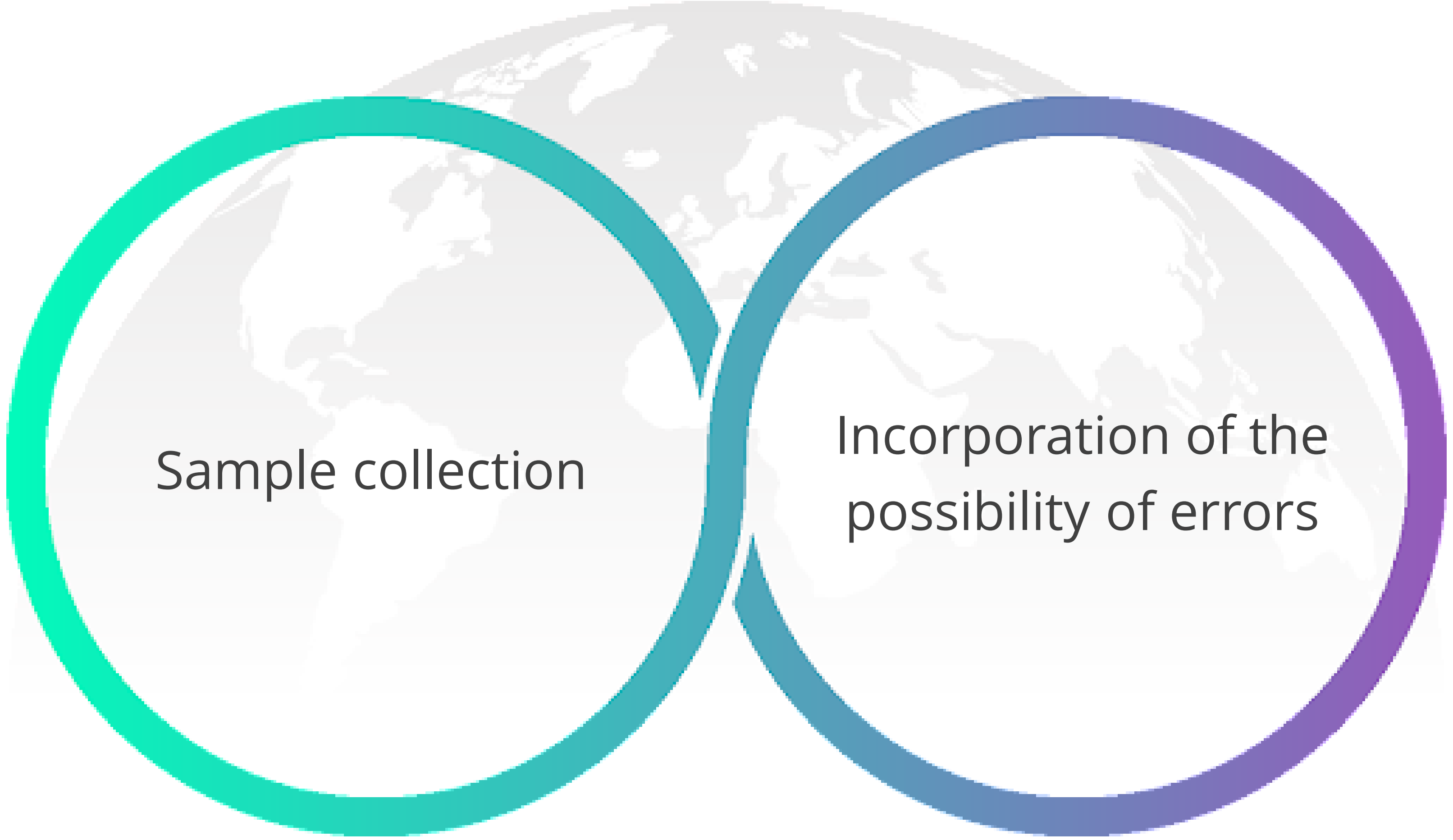
The process of making an inductive inference about an entire population based on a sample is called statistical inference.



Example: The average life of the sample of bulbs tested is taken as an estimate of the average life of the entire produced lot.

Sampling

While making conclusions, the following terms are important to understand:

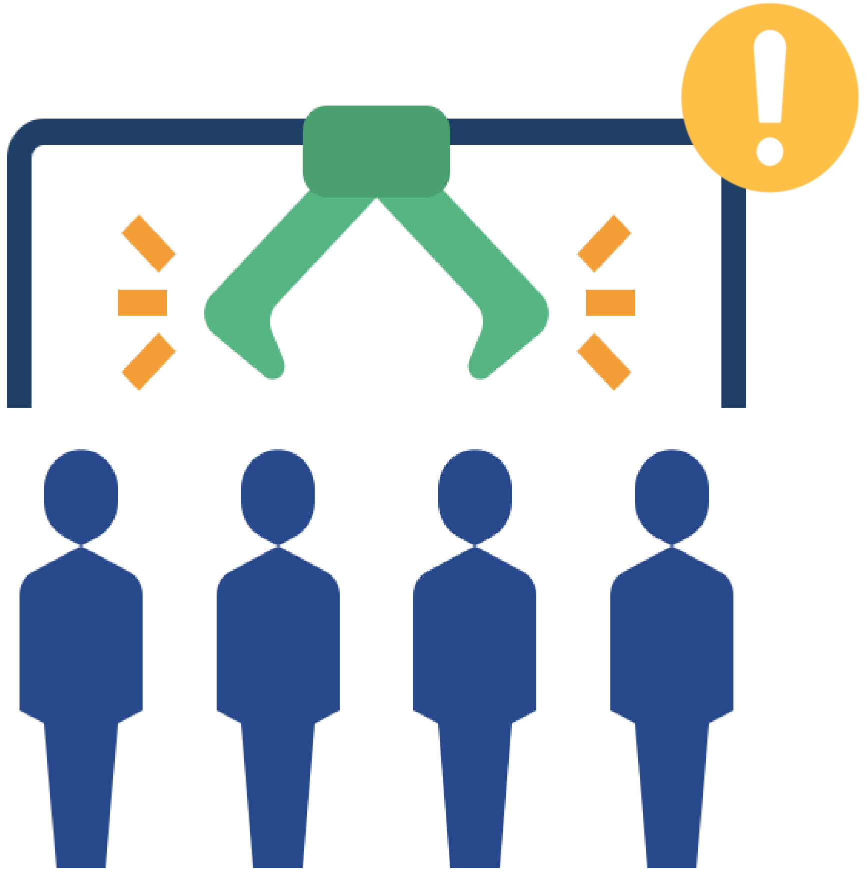


Sample collection

Incorporation of the possibility of errors

Sampling Error

The error in the inductive inference from a sample to a population is known as sampling error.

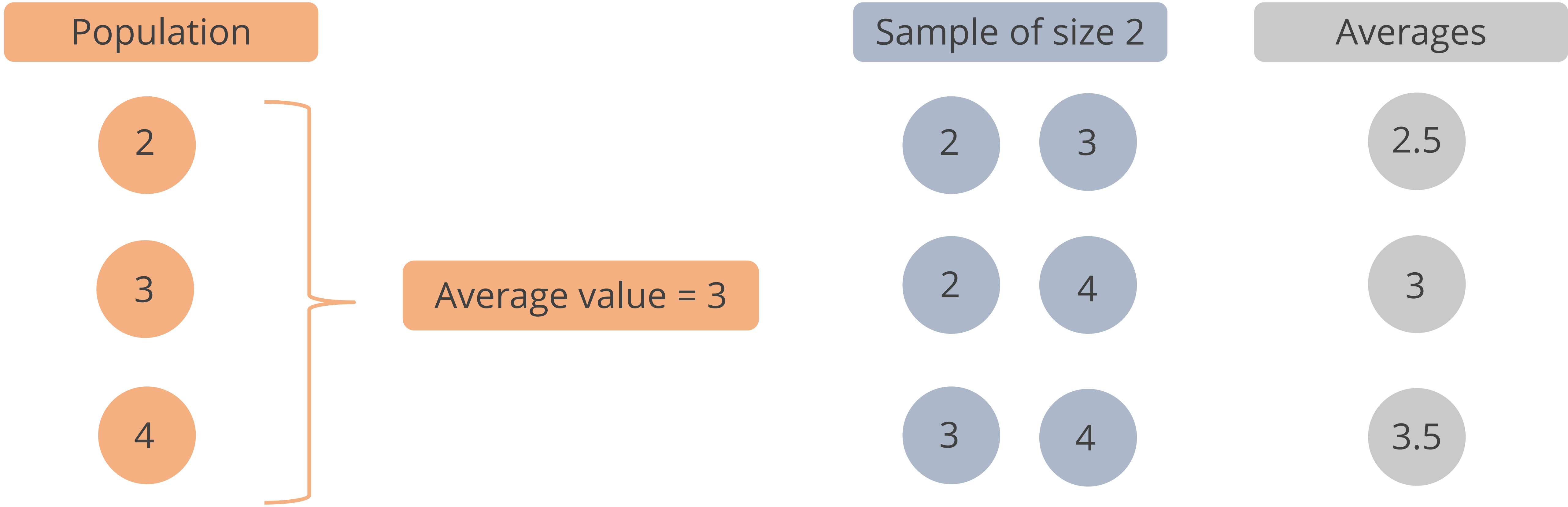


Non-sampling errors encompass all errors not classified as sampling errors. Issues in data collection or biases in responses may induce sampling errors.

Sampling Error

Consider a population that comprises three units with associated numerical values of 2, 3, and 4 and with 3 as the average value.

Consider various samples of size 2. These are (2, 3), (2, 4), and (3,4). Their averages are 2.5, 3, and 3.5.



Population

2

3

4

Average value = 3

Sample of size 2

2 3

2 4

3 4

Averages

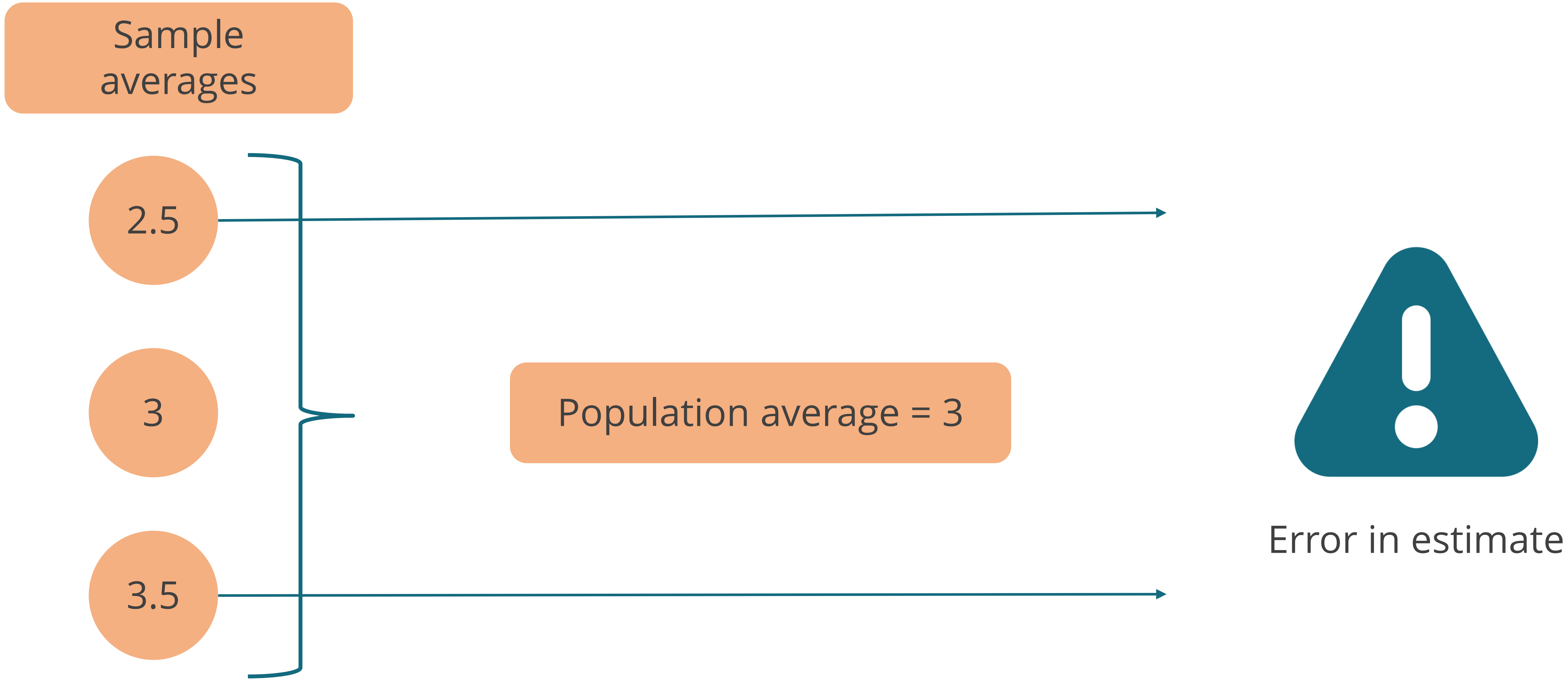
2.5

3

3.5

Sampling Error

When the sample average is taken as an estimate of the population average, there is an error in two of the three possible samples.

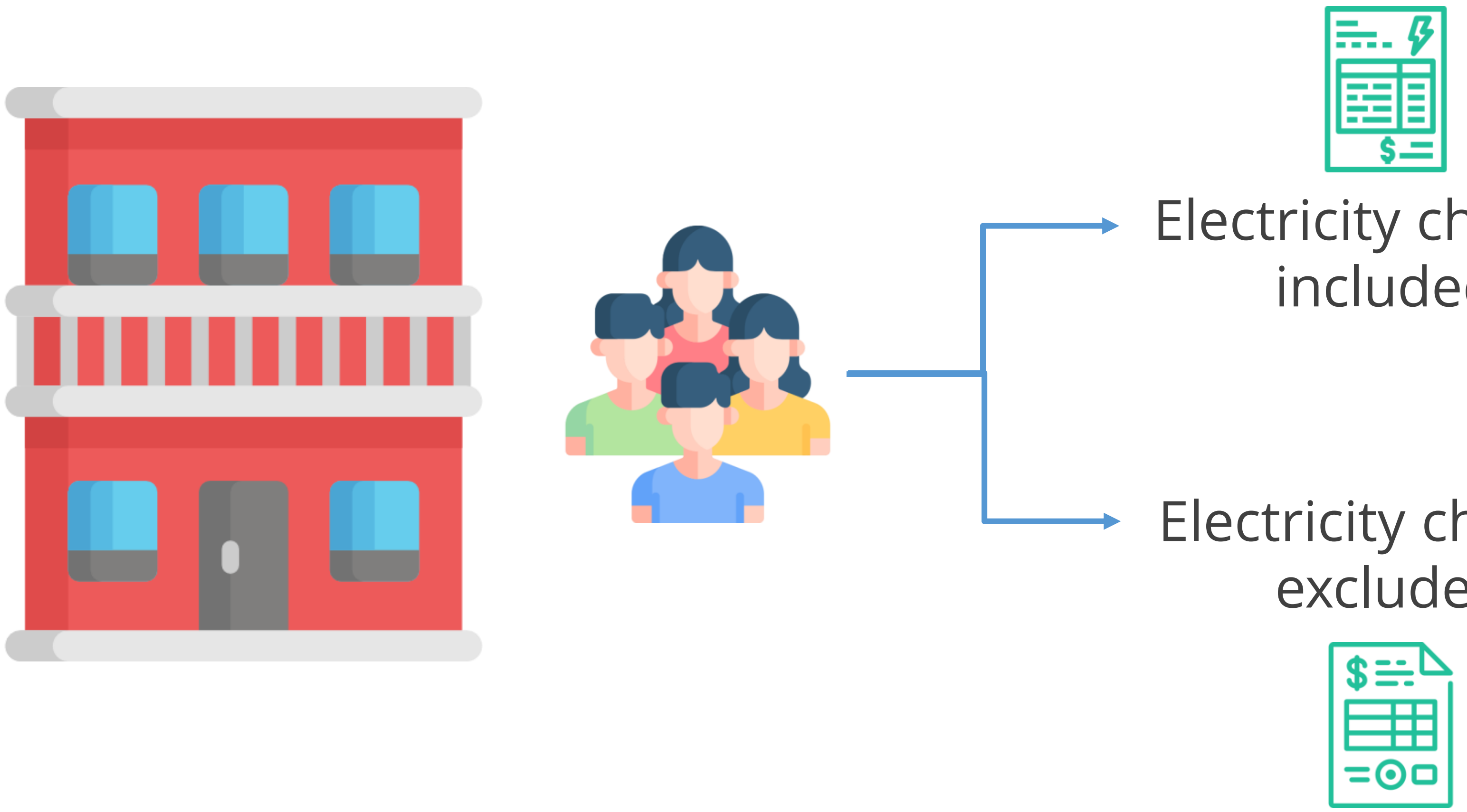


Such an error is referred to as a sampling error.

Data Collection Errors

Example: Data collected on rentals for paying guest accommodation

The population may consist of tenants who are charged separately for electricity and those who are not charged.



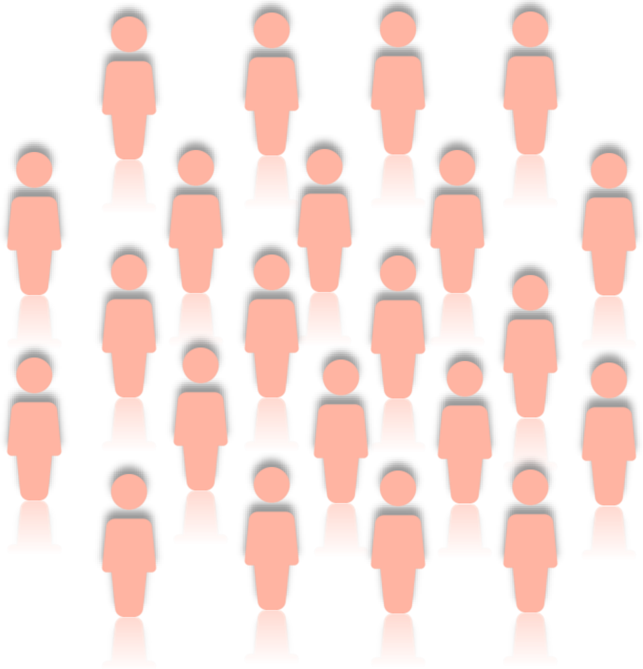
Electricity charges included

Electricity charges excluded

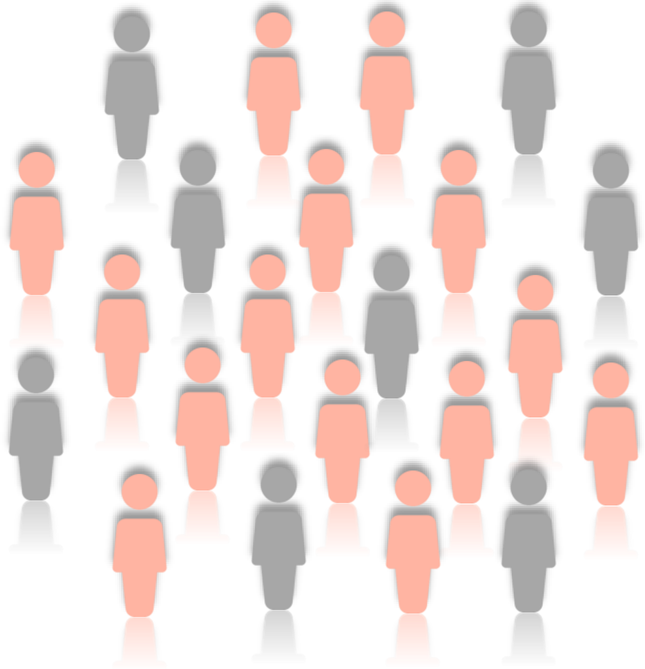
When such details are ignored, the results can be misleading.

Non-Sampling Errors

It is important to note that non-sampling errors occur both in complete enumeration and in sampling studies.



Complete enumeration



Sampling studies

It is important to select a sample that represents the population.

Advantages and Disadvantages of Sampling

Advantages of Sampling

Sampling saves time, because it reduces the volume of data and minimizes the need to go over each item.

The sampling process avoids monotony and eliminates repeated inquiries for each dataset.

Sampling makes surveys feasible when time and resources are limited.

Limitations of Sampling

A few of the major limitations of sampling are:

Chances of bias

Data excluded from sampling if it is not homogeneous

Difficulties in selecting a sample

This has an impact on the accuracy of the result.

Discussion: Data Sampling

Duration: 15 minutes

How does sampling of data help in decision-making?

- What does data sampling mean?

Answer: Sampling is the process of selecting a subset of a population to represent the entire population.

- What are the advantages of sampling?

Answer: It is the only option when destructive testing is involved. It becomes necessary to restrict sampling to small sizes. The impact is particularly significant when a large mass of data is involved.

Probability Sampling Methods

Discussion

Discussion: Probability Sampling

Duration: 15 minutes

Assume you have been given a large data set about a country's population, and you are tasked with dividing the data by city. Your objective is to analyze this data to predict the country's population in the next five years.

To accomplish this, you should:

- Comprehend probability sampling and its various methods

Probability Sampling

Probability sampling is the selection of a sample from a population based on the randomization principle.

It ensures that the sample is representative of the population, allows researchers to estimate the level of uncertainty in their findings, and permits the results to be generalized across the whole population.

Approaches to Probability Sampling

There are four approaches to probability sampling:

Simple random sampling

Systematic sampling

Stratified random sampling

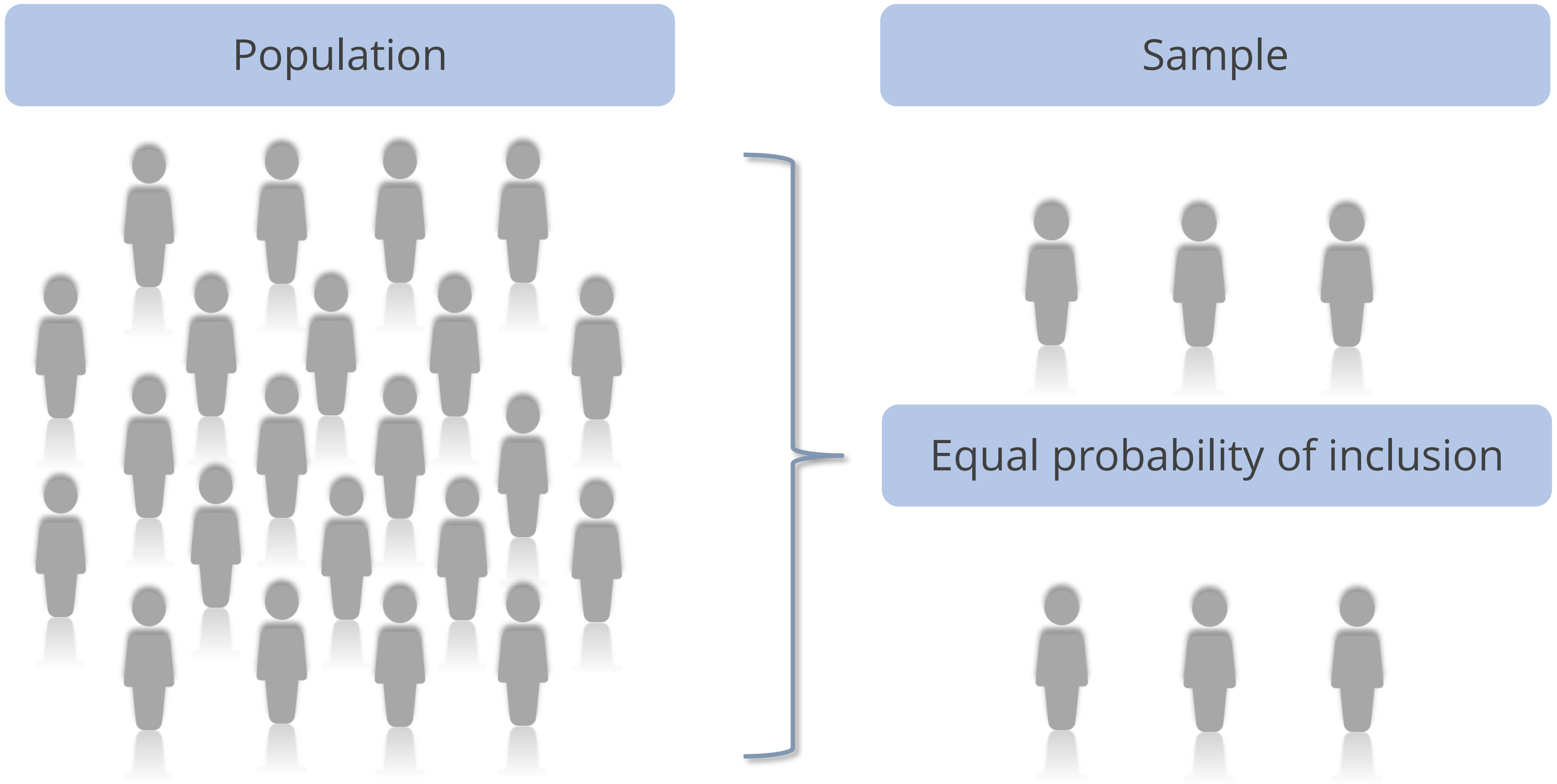
Cluster sampling

Simple Random Sampling

Simple random sampling is a technique of sample selection in which every sample has an equal probability of selection.

Simple Random Sampling

Every unit in the sample also has an equal probability of inclusion in the sample.



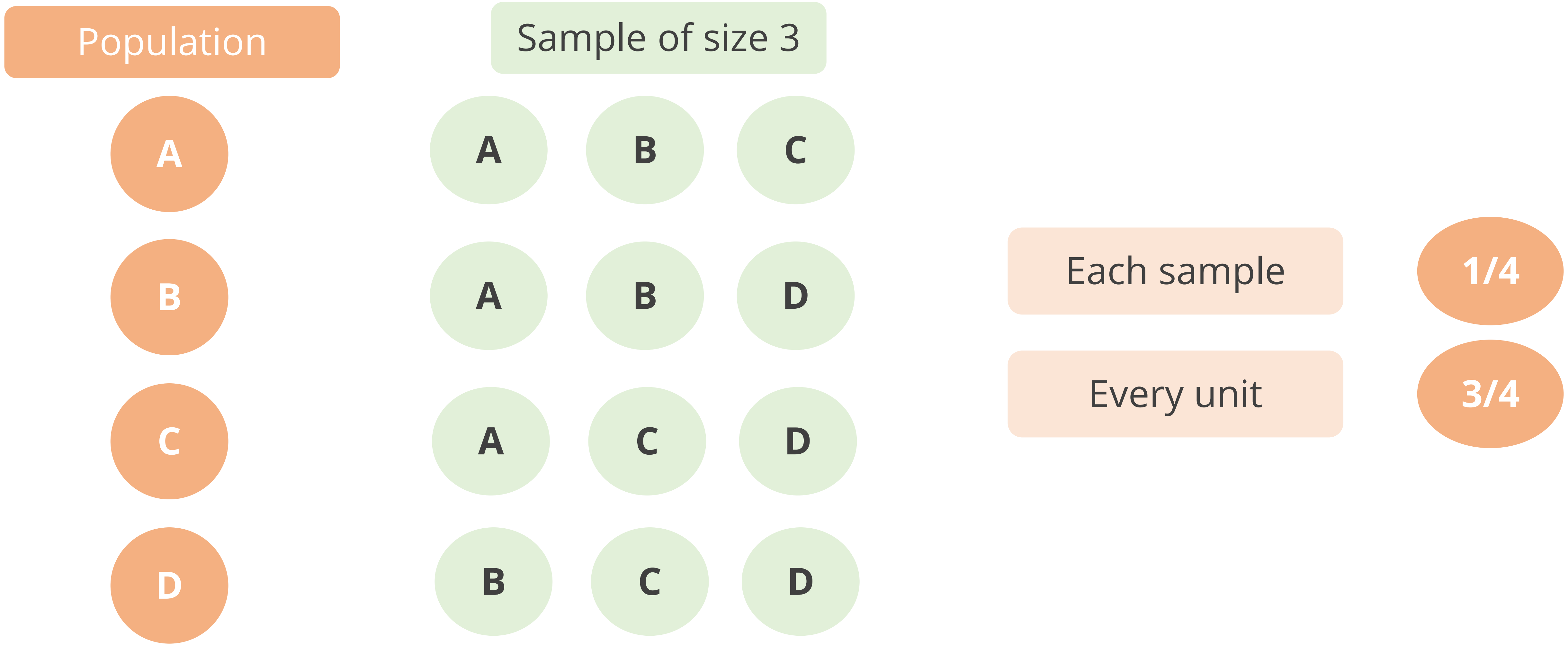
Population

Sample

Equal probability of inclusion

Simple Random Sampling: Example

Consider a population consisting of four units: A, B, C, and D.



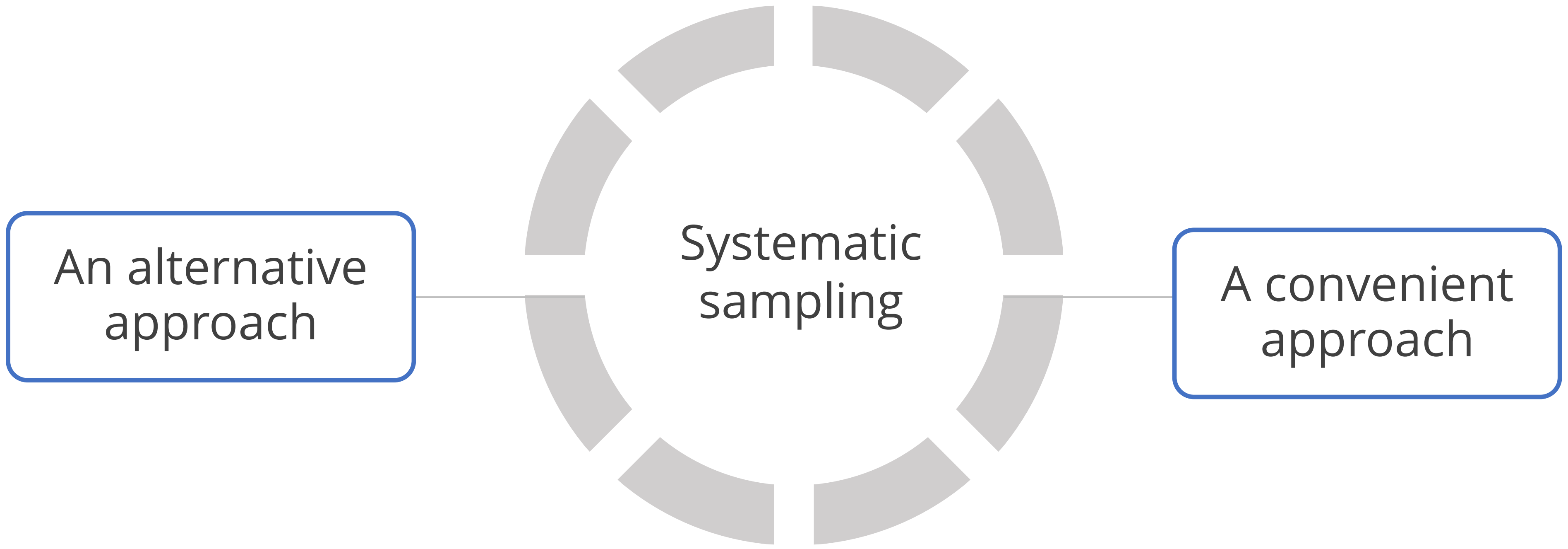
Population	Sample of size 3		
A	A	B	C
B	A	B	D
C	A	C	D
D	B	C	D

Each sample	1/4
Every unit	3/4

Each sample has a ¼ probability of being selected. Also, every unit has a ¾ probability of being selected.

Systematic Sampling

Systematic sampling is a probability sampling technique in which researchers randomly select members of a population at regular intervals.

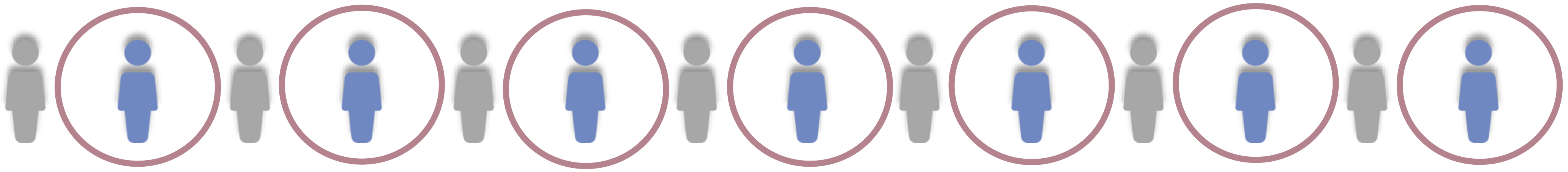


An alternative approach

Systematic sampling

A convenient approach

The process involves systematically selecting elements from a subset that is arranged in sequential order.



Systematic Sampling: Example

Suppose a population consists of $N = 500$ units and a sample of $n = 50$ units

Population $N = 500$

Sample will be every k^{th} unit

Sample $n = 50$

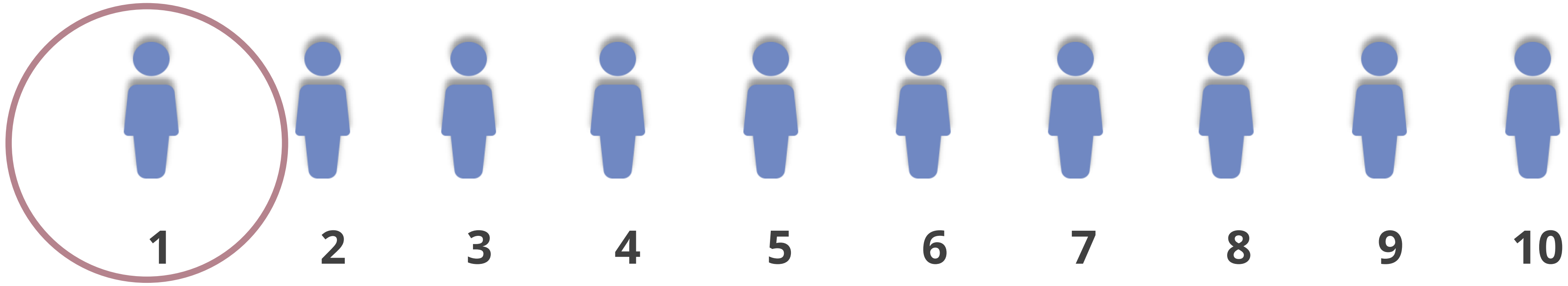
$k = N/n = 500/50 = 10$

Then, every k^{th} unit, where $k = N/n = 500/50 = 10$ in the population is chosen.

Systematic Sampling: Example

The first unit is randomly selected from units 1 to k = 10.

1 to k = 10



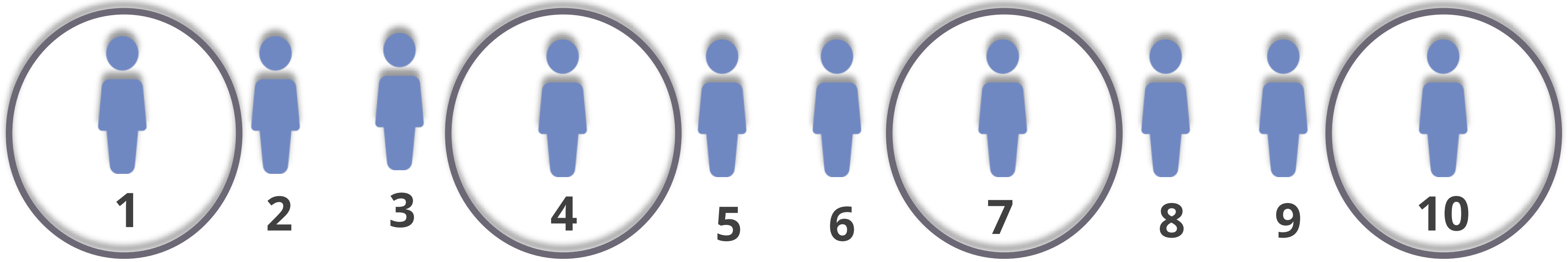
k is the sampling interval

Every unit in the population has the probability of 1/k = 0.1 of being chosen.

Systematic Sampling: Example

In a sequential arrangement of units, systematic sampling maintains both efficiency and the original order of the units.

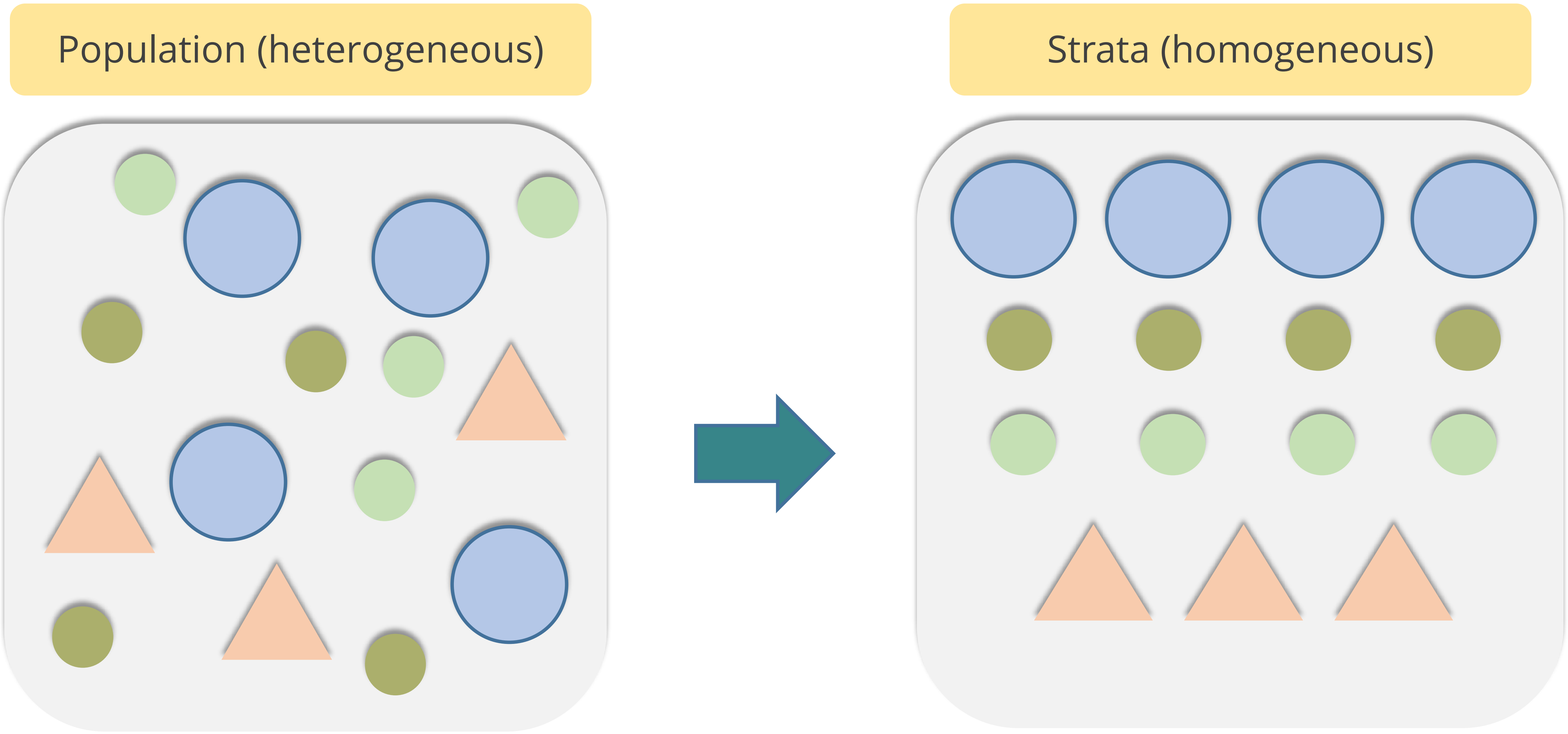
Samples are equidistant.



The distance is being chosen logically.

Stratified Random Sampling

Stratified random sampling is a sampling method that divides a population into smaller subgroups known as strata.

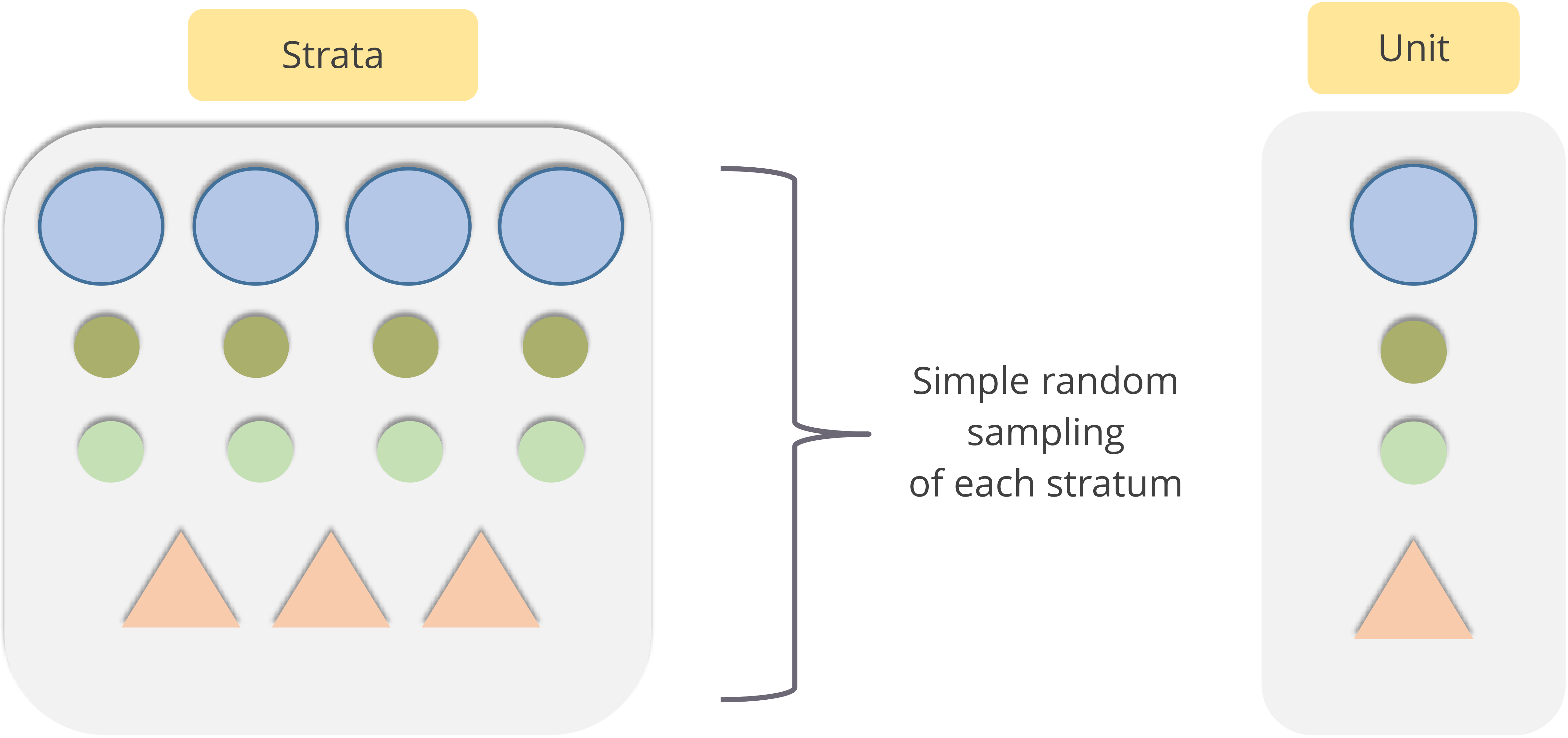


Population (heterogeneous)

Strata (homogeneous)

Stratified Random Sampling

A random sample is drawn from each stratum in a number proportional to the size of the stratum in comparison to the population.



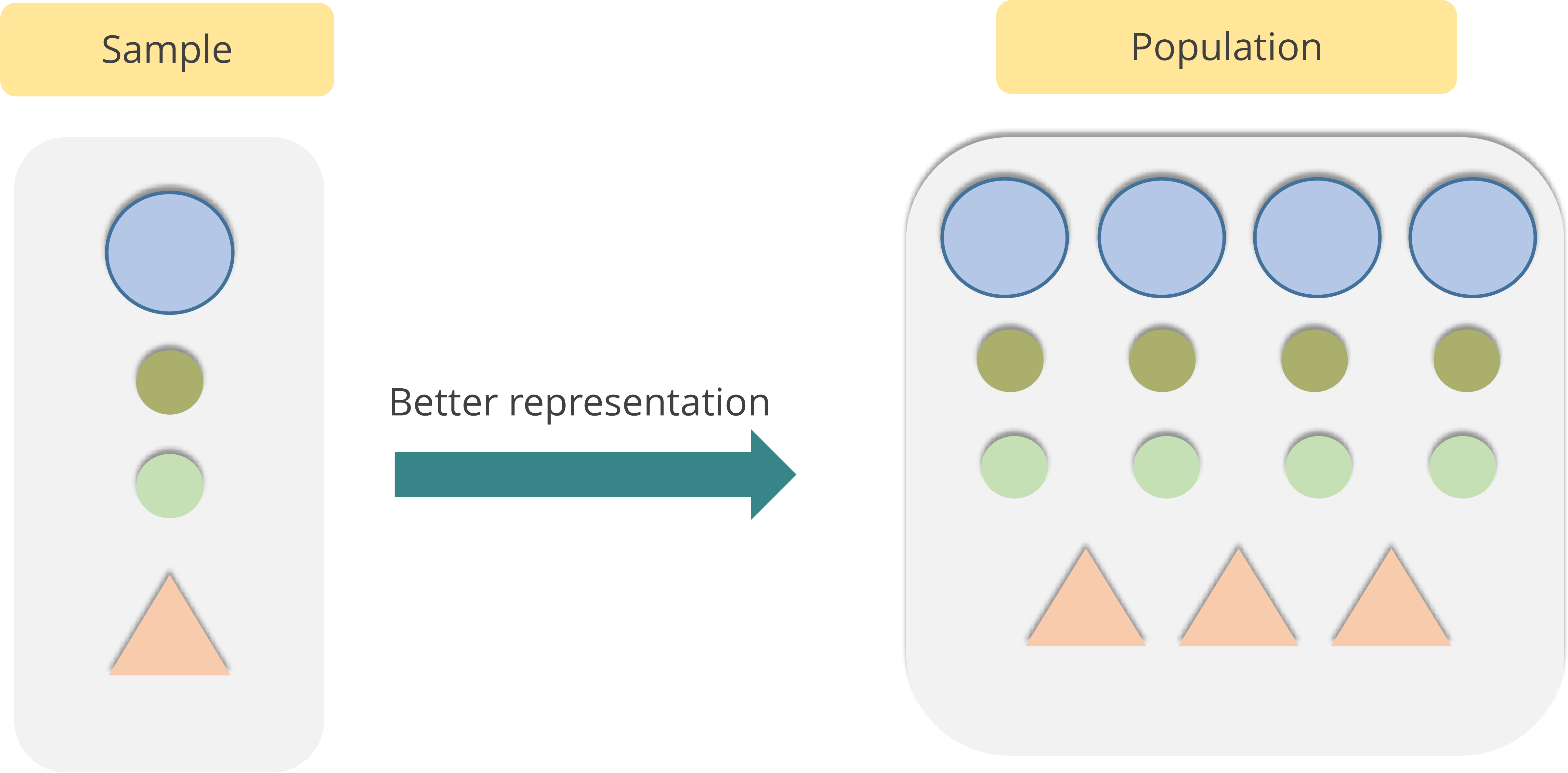
Strata

Unit

Simple random sampling of each stratum

Stratified Random Sampling

In comparison to a sample drawn at random, the chosen sample includes members of all strata, making it a more accurate representation of the population.



Sample

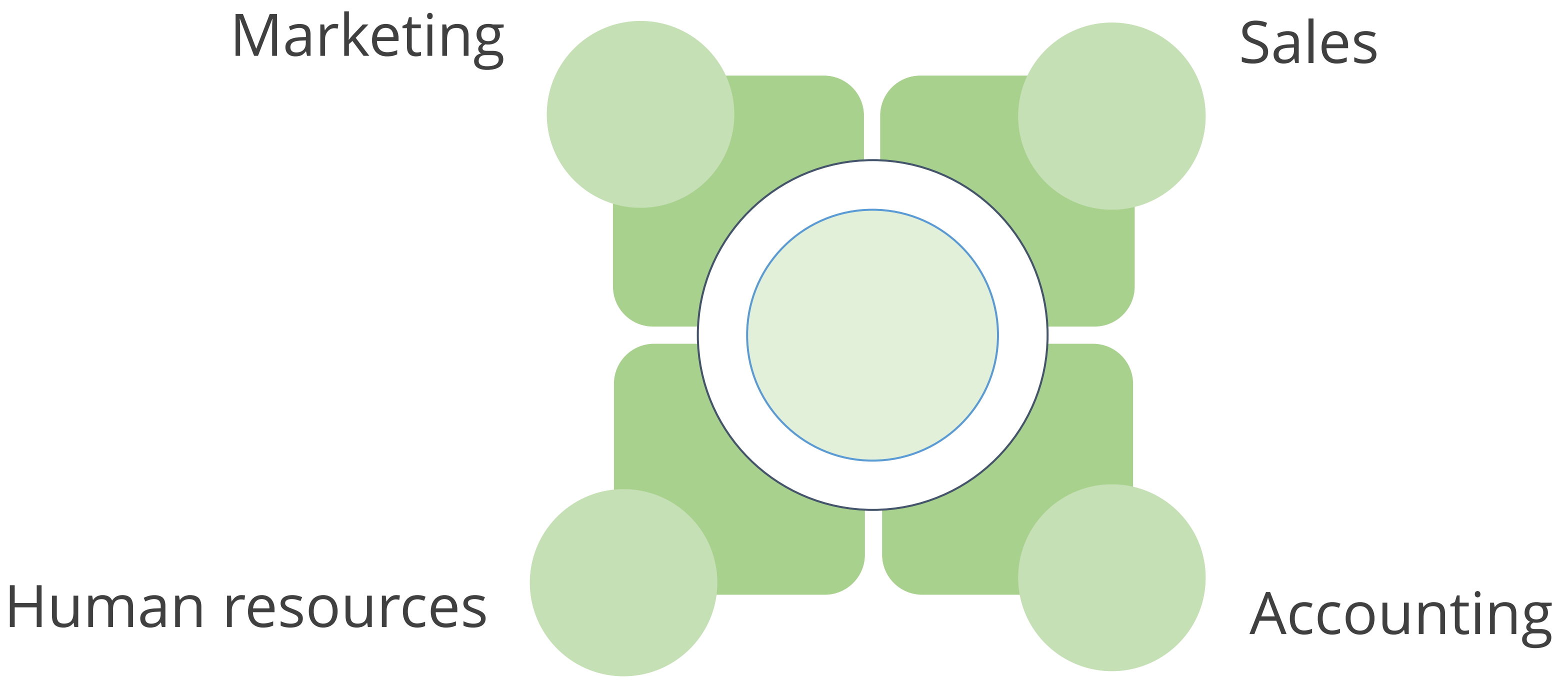
Better representation

Population

Stratified Random Sampling: Example

An employee satisfaction survey is being conducted in a large company.

The employees’ opinions could vary at different levels in the hierarchy and in different functions, like:



Cluster Sampling

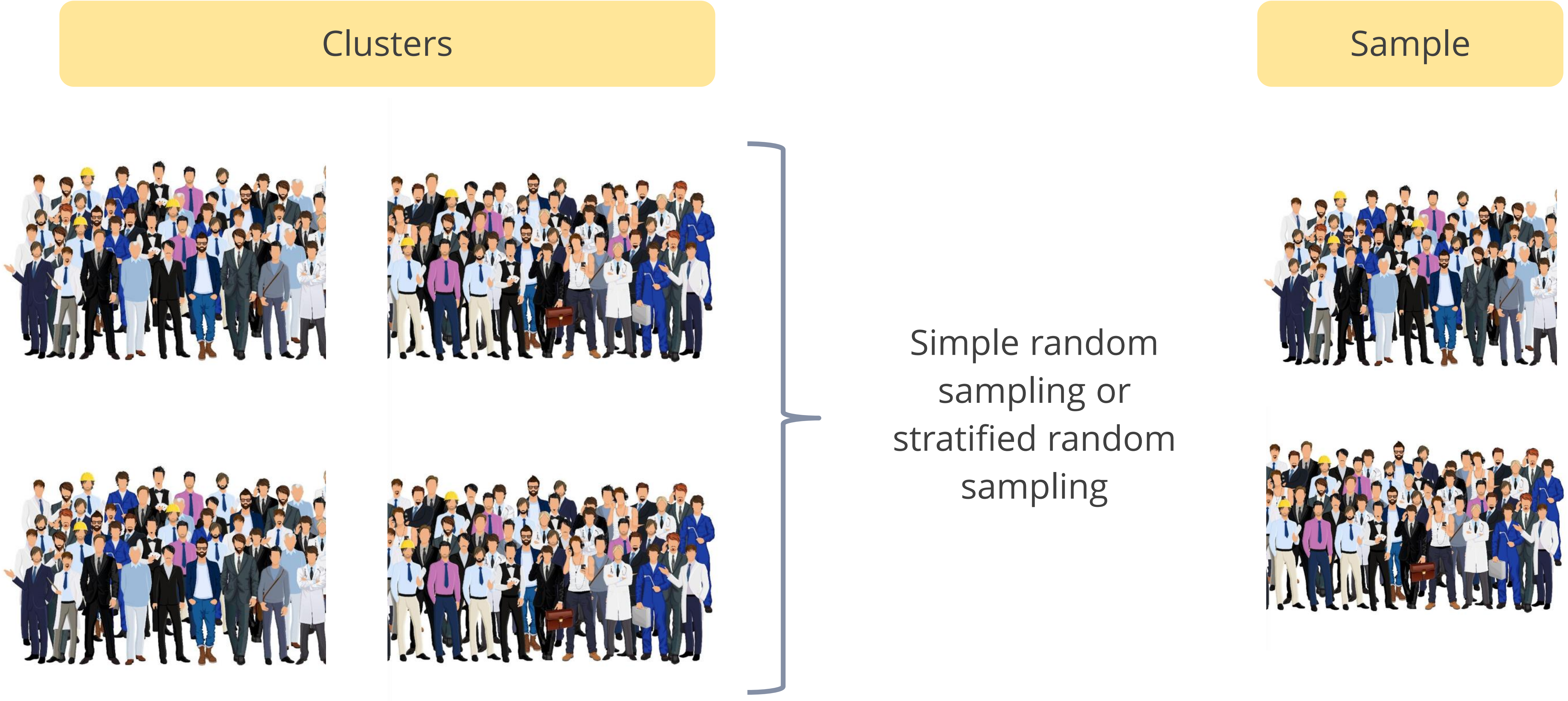
Cluster sampling is a method in which the population is divided into several clusters. All clusters are almost equally heterogeneous.

Population

Clusters

Cluster Sampling

A sample of clusters is selected using simple or stratified random sampling.



Clusters

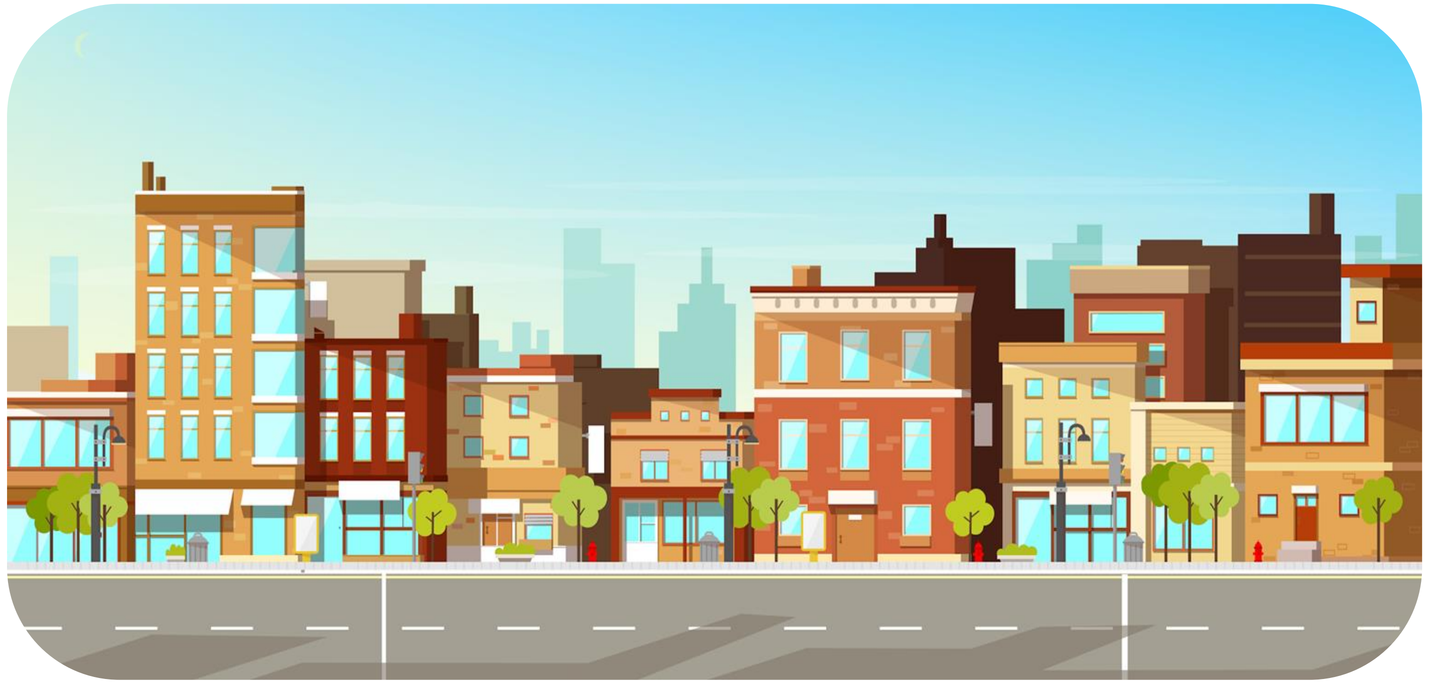
Sample

Simple random sampling or stratified random sampling

All units from these clusters together constitute the required sample.

Cluster Sampling: Example

Expenditure patterns of residents in a city:



Population clusters

Representation of the clusters

Uses of Probability Sampling

Some of the key uses of probability sampling include:

Making informed decisions

Incorporating required quality standards

Establishing rules for rejecting items of poor quality

Discussion: Probability Sampling

Duration: 15 minutes

Consider that you have been given a large set of data about a country's population and have been asked to bifurcate the data based on each city and analyze it to predict the country's population over the next five years.

To do so, perform the following:

- Understand the simple random sampling of data

Answer: Simple random sampling is a technique of sample selection in which every sample has an equal probability of selection.

- Understand probability sampling and systematic sampling

Answer: Probability or random sampling is a sampling approach to select a sample using the theory of probability. In systematic sampling, samples are selected from a sequentially arranged lot, at regular intervals.

Non-Probability Sampling Methods

Discussion

Discussion: Non-Probability Sampling

Duration: 15 minutes

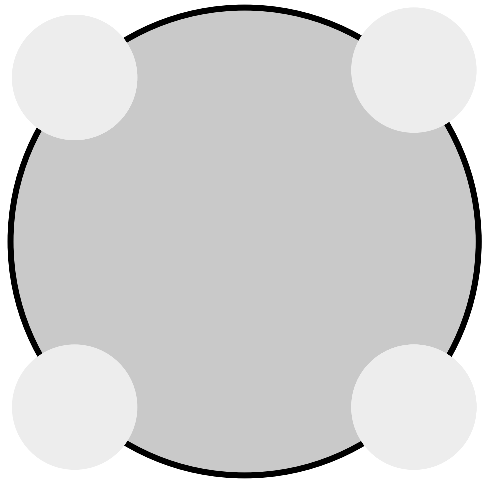
- What is non-probability sampling?
- What is convenience sampling?

Non-Probability Sampling

It is a method of sample selection in which a sample is chosen based on the investigator’s subjective evaluation.

Non-Probability Sampling Methods

The four methods of non-probability sampling are:



Convenience sampling

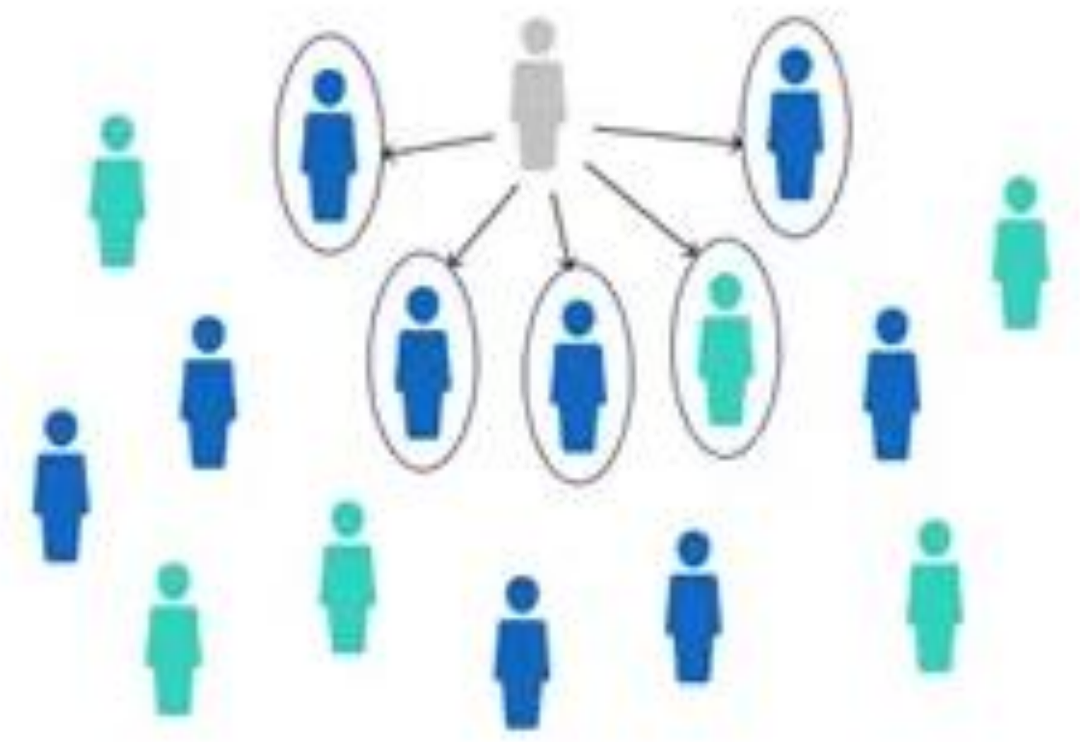
Voluntary response sampling

Snowball sampling

Judgmental or purposive sampling

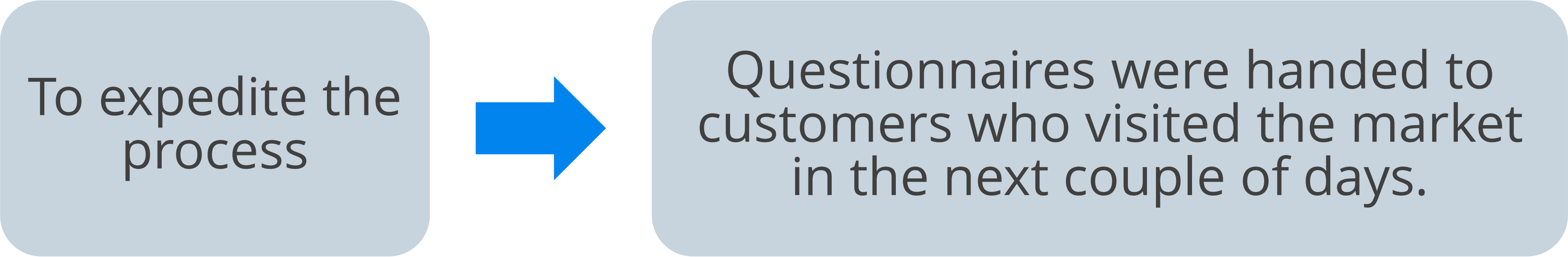
Convenience Sampling

It is a type of non-probability sampling in which the researcher selects the sample based on convenience rather than random selection.



Convenience Sampling: Example

Consider the example of a supermarket chain that has prepared a questionnaire to initiate steps toward improving customer satisfaction.



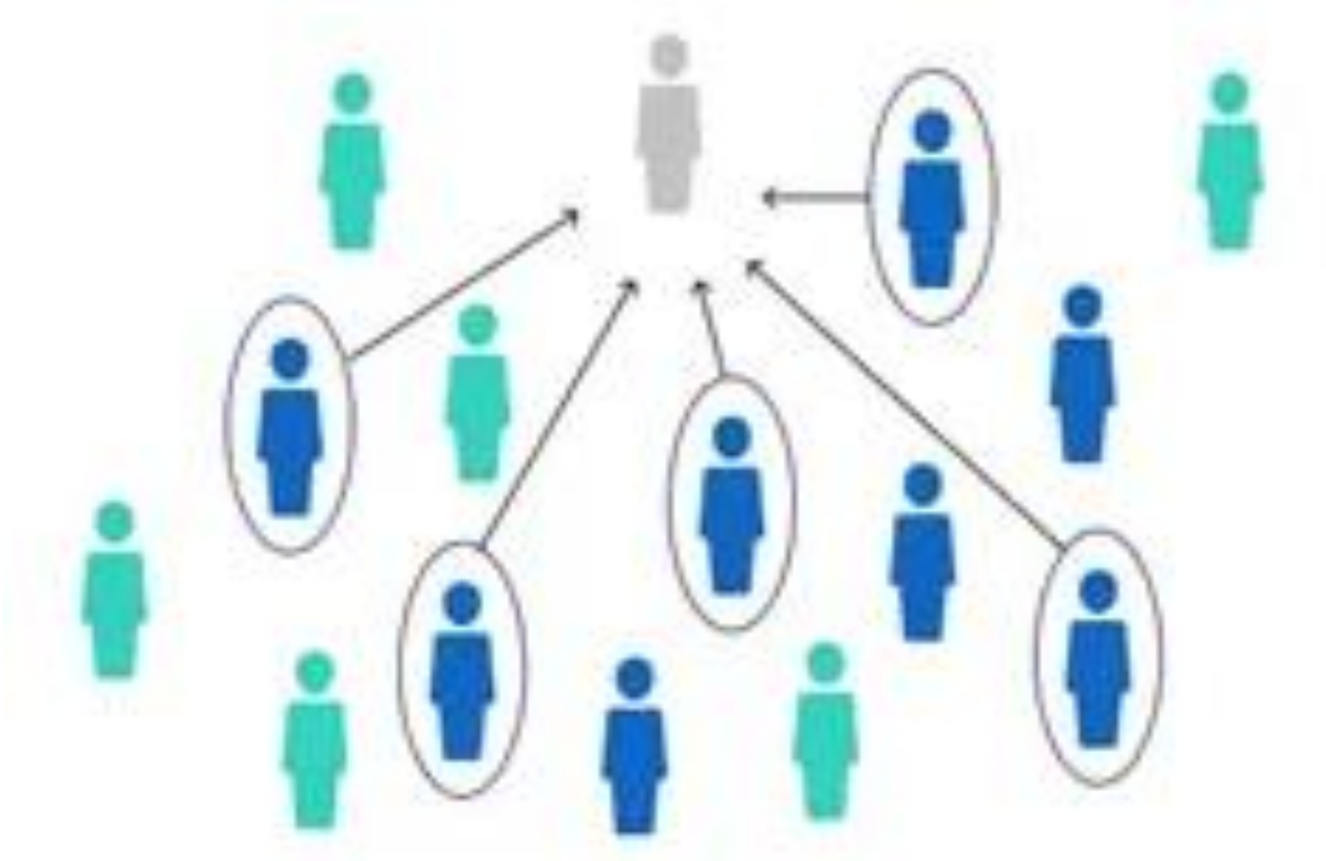
To expedite the process

Questionnaires were handed to customers who visited the market in the next couple of days.

So here, the sample of customers chosen constitutes a convenience sample.

Voluntary Response Sampling

It is a type of non-probability sampling in which the researcher selects the sample from people who volunteer to participate in the study.



Voluntary Response Sampling: Example

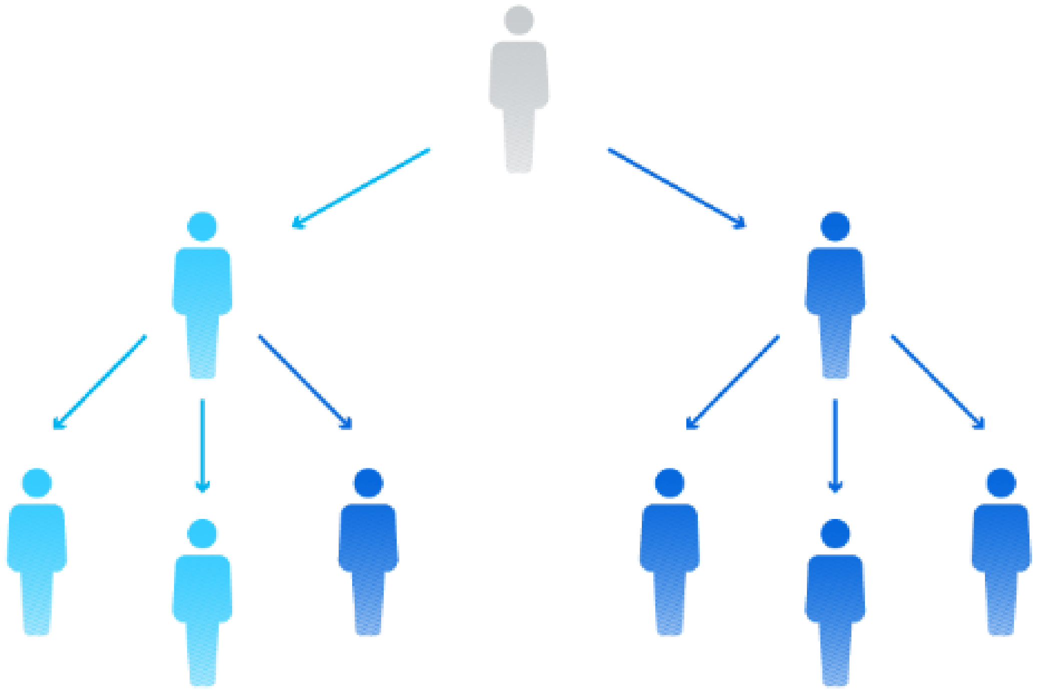
Consider the following example:

- Websites invite individuals to offer responses
- Only a few individuals provide responses

This process of obtaining responses from selected respondents is clearly voluntary response sampling.

Snowball Sampling

Snowball sampling is a research method in which researchers initially recruit participants and then rely on referrals to recruit more, thus creating a chain.



The process is continued till a sample of the desired size is obtained.

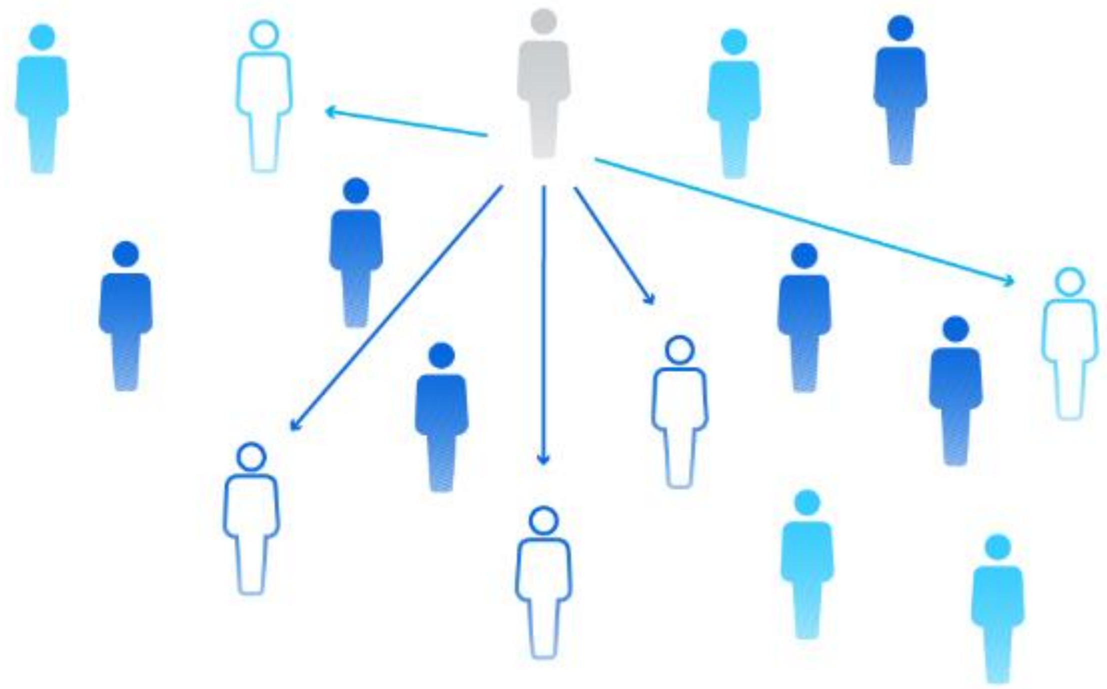
Snowball Sampling: Example

Snowball sampling is popular in business studies that focus on specific organizations.

Some of the employees serve as contacts who provide referrals for further contacts besides serving as respondents for the study.

Purposive or Judgmental Sampling

It is an approach in which the investigator selects units from the population using his knowledge and judgment about various units in the population.



Purposive Sampling: Example

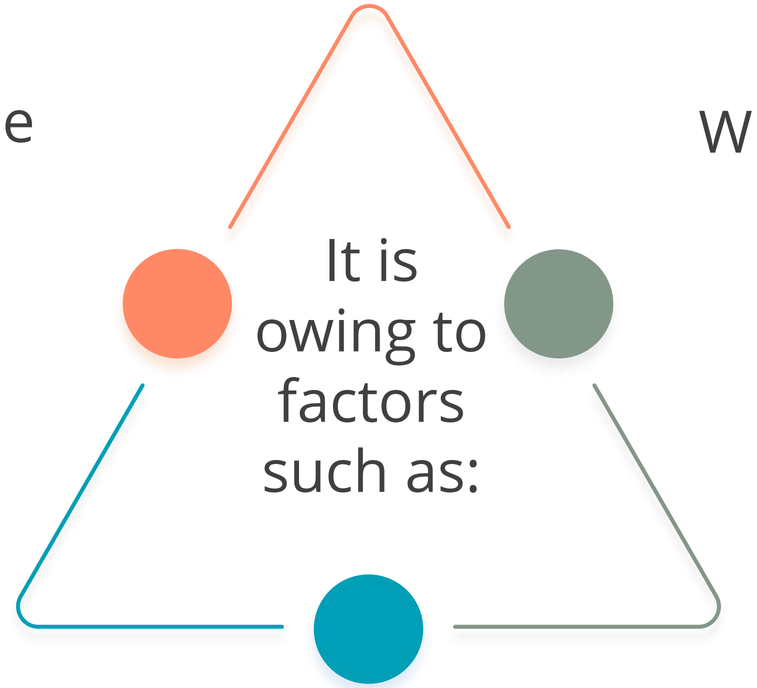
In a qualitative research study, an investigator may target customers who are articulate.

Customer feedback

The investigator could then obtain useful and high-quality information from this sample.

Uses of Non-Probability Sampling

It provides for a quicker and more cost-effective data collection.



It is owing to factors such as:

Convenience

Willingness to respond

Ease of identification

The approach is attractive when time and cost are important considerations.

Non-Probability Sampling

The approach offers greater scope to collect quality data with factors such as:

Results from the sample cannot be generalized to the entire population, as the sample does not necessarily represent the population.

Discussion: Non-Probability Sampling

Duration: 15 minutes

- What is non-probability sampling?

Answer: It is an approach to sample selection wherein a sample is selected based on the subjective judgment of the investigator.

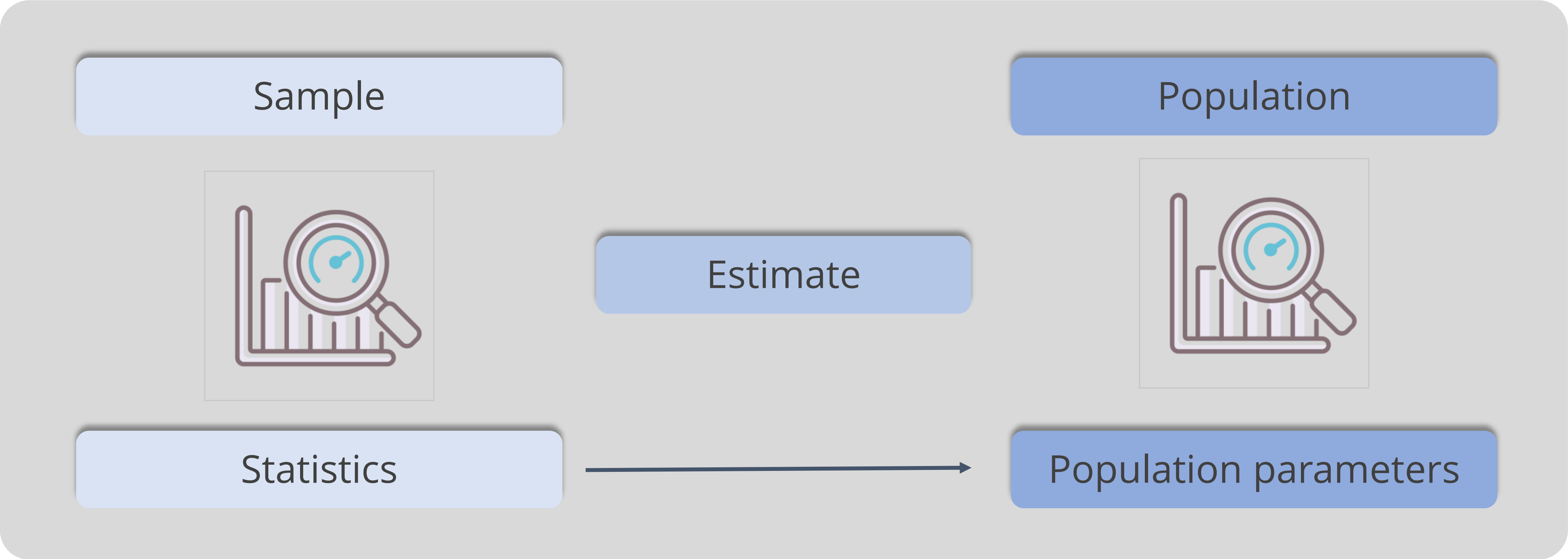
- What is convenience sampling?

Answer: It is an approach to sample selection in which sample units are chosen entirely on considerations of convenience.

Sampling

Sample Indices

The sample indices are usually referred to as statistics, and the population indices are referred to as population parameters.



Sample

Population

Estimate

Statistics

Population parameters

Values of sample statistics clearly vary depending upon the sample chosen.

Sampling: Example

Consider a population comprising four units, with the following values:

Unit ref no.	I	II	III	IV
Value	2	3	3	5

Sampling: Example

The quality of results from the analysis will be poor when several clusters do not adequately represent the population.

Sample no.	Units selected	Sample values	Sample average
1	I, II	2, 3	2.5
2	I, III	2, 3	2.5
3	I, IV	2, 5	3.5
4	II, III	3, 3	3
5	II, IV	3, 5	4
6	III, IV	3, 5	4

Sampling: Example

The probability distribution of the sample average is:

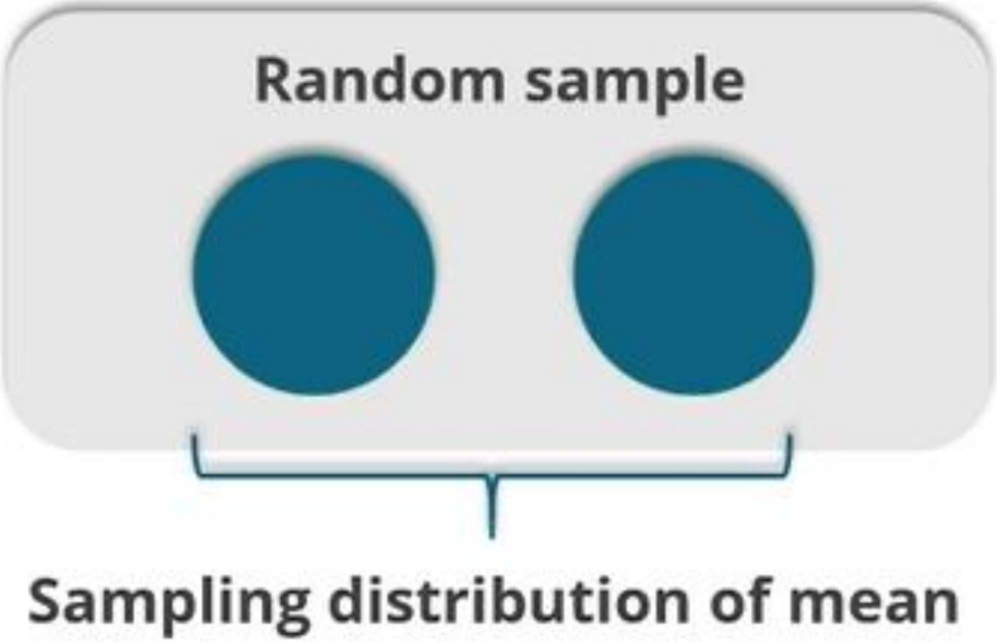
Sample average	2.5	3	3.5	4	TOTAL
Probability	1/3	1/6	1/6	1/3	1

Twice as likely

Sampling Distribution

Sampling Distribution

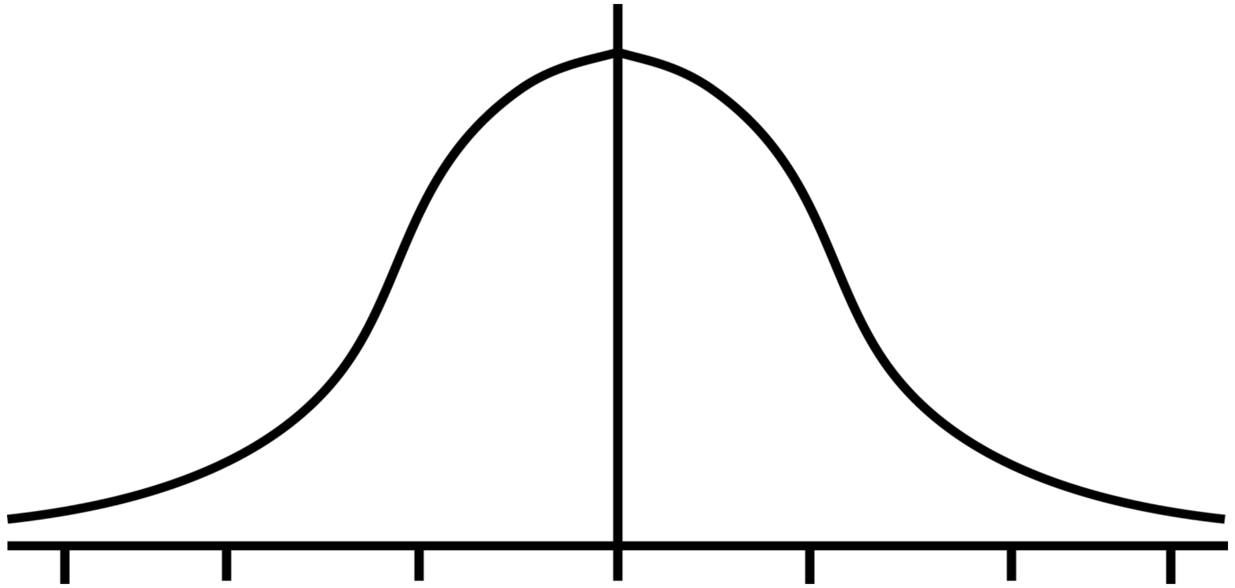
The sampling distribution represents the distribution of a statistic, like mean, proportion, and variance, calculated from multiple random samples taken from the same population.



It provides insights into the characteristics and variability of a statistic across different samples, allowing for population inferences based on sample statistics.

Sampling Distribution

The key characteristics of the sampling distribution of the sample mean are:



The distribution of the sample mean follows a normal distribution with the same mean μ regardless of the sample size.

However, the standard deviation is inversely proportional to the square root of the sample size. Consequently, as the sample size increases, the standard deviation decreases.

Central Limit Theorem (CLT)

Central Limit Theorem

The central limit theorem (CLT) states that independent random variables approach a normal distribution as the sample size increases, regardless of the underlying population distribution. This applies to both the sum and average of the random variables.

If a random sample of size n is drawn from a normal population with parameters μ and σ, denoted as X1, X2, ..., Xk, ..., Xn, the sample mean ($\overline{x}$) follows a normal distribution with parameters μ and σ divided by the square root of n.

Case Study: Sample and Sampling Techniques

Problem Statement

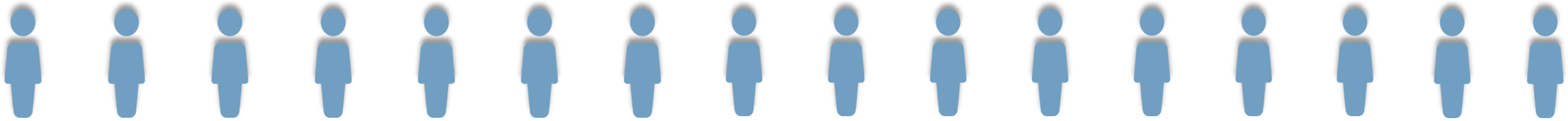
Let's consider a scenario where a significant number of students take exams every year, and the percentage of marks got in mathematics follows a normal distribution. The distribution has a mean of 60 and a standard deviation of 15.

Our objective is to determine the probability of a student scoring 70% or above in the examination.

Problem Statement

Number of samples = 12

Sample size = 16



Determine the sampling distribution of the sample mean

Estimate the proportion of students scoring more than 70%

Calculate the sample mean ($\bar{x}$) for each of the 12 samples.

Data Samples

The data comprises 12 samples, with each sample having a size of n = 16.

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
1	52	87	67	58	54	60	73	62	66	59	50	64
2	57	61	50	61	30	56	64	37	78	62	87	103
3	65	68	90	78	36	73	60	31	83	48	42	40
4	78	68	68	68	78	36	54	84	56	35	47	64
5	96	75	58	64	18	60	66	63	57	80	52	56
6	67	80	66	63	95	67	74	41	45	60	57	39
7	49	46	57	59	87	62	50	66	80	79	72	56
8	39	42	50	48	59	47	69	46	44	48	46	44

Data Samples

The data comprises 12 samples, with each sample having a size of n = 16.

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
9	79	68	83	58	40	77	59	59	60	69	50	52
10	80	47	35	69	44	53	66	62	66	49	51	78
11	43	67	74	57	64	54	75	58	31	66	41	51
12	46	63	68	62	42	58	62	51	74	59	70	75
13	77	79	48	62	64	53	73	46	77	74	52	68
14	58	62	56	64	104	76	75	39	55	70	45	71
15	101	37	85	29	73	38	68	68	49	52	50	76
16	43	38	59	50	38	62	73	75	43	78	67	85

Presentation of Results

Record the results from sample data in the following format:

Sample No.	I	II	III	III	X	XI	XII
Value of $\bar{x}$							
No. of observations ≥ 70							
Proportion of observations ≥70							

Presentation of Results

Record the consolidated results in the following format:

	Average value of sample mean (n=16)	Standard deviation of sample mean (n=16)	Average value of sample proportion (n=16)	Standard deviation of sample proportion (n=16)
Theoretical values				
Estimated values from sample data				

State the observations

Solution

Calculate

Mean

Standard deviation of x̄ values

Mean of the proportions

Standard deviation of proportions

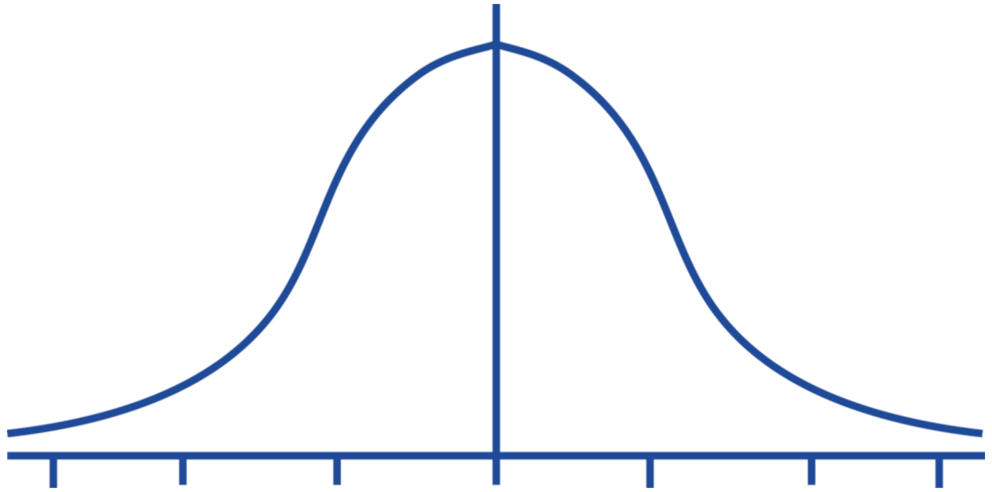
If $\bar{p}$ = average proportion

Standard deviation =

$$\sqrt{((\bar{p}* (1-\bar{p})/n)}$$

Solution

Sampling distribution of the sample mean:



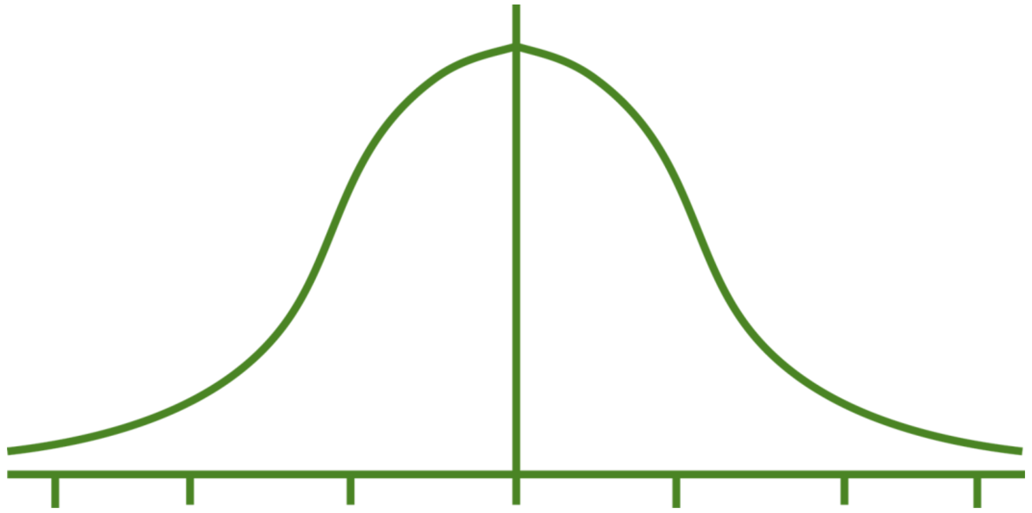
Mean = 60

Standard deviation = 15/ √16 = 15/4 =3.75

The proportion of students scoring > 70 = 0.2525

Solution

Sampling distribution of proportion:



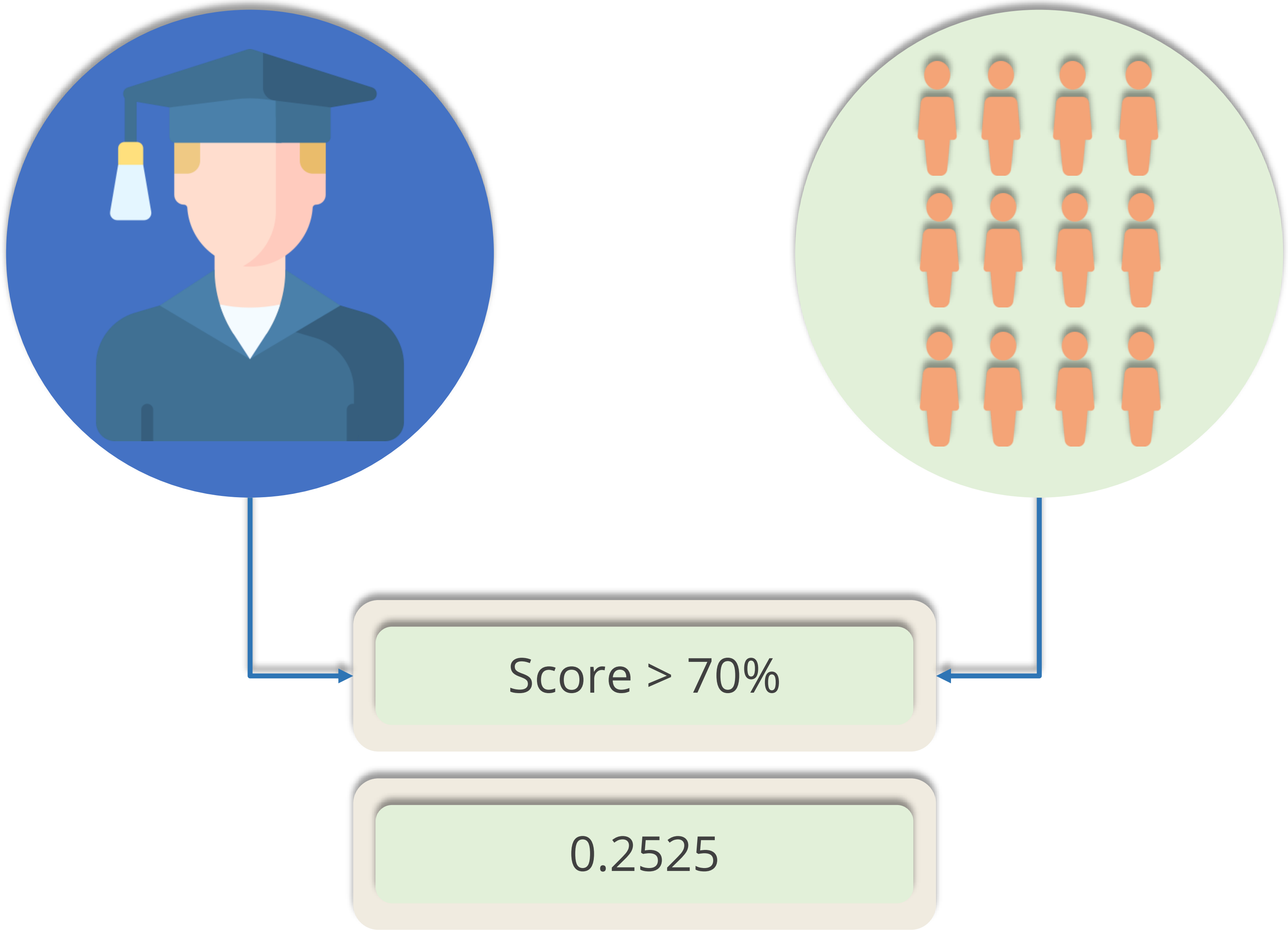
Mean = 0.2525

Standard deviation = √{0.2525*(1- 0.2525)/16} = 0.1086

According to the central limit theorem, the provided approximation holds true when the sample size is 30 or larger.

Solution

The probability of a student scoring above 70 is as follows:

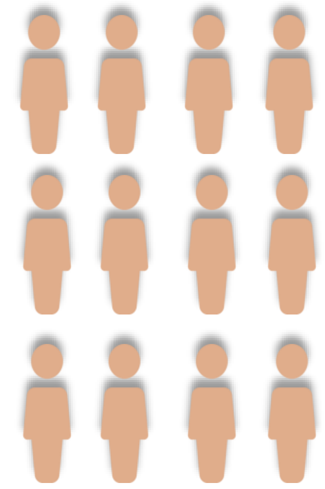


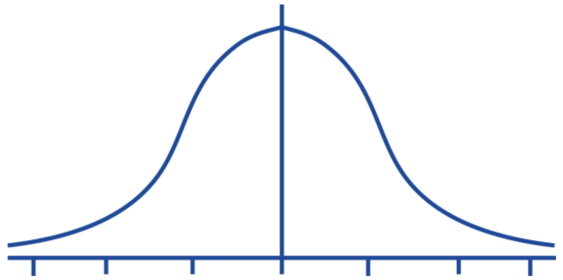
Score > 70%

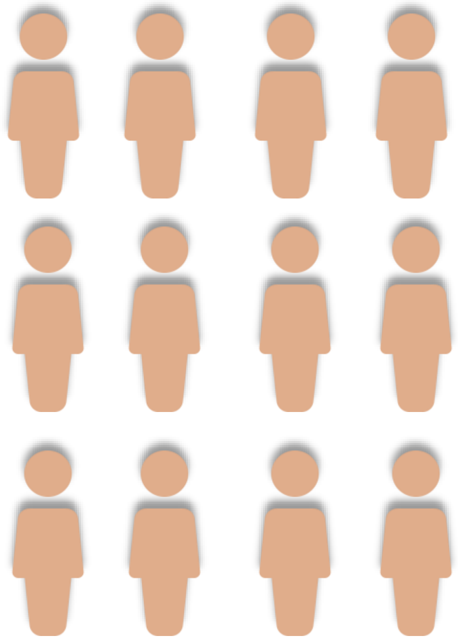
0.2525

Solution

The sample proportion follows a normal distribution.







Sample size = 16

Mean = 0.2525

Standard deviation
= √{0.2525*(1- 0.2525)/16}
= 0.1086

$$\text{Proportion} = \frac{\text{Number of samples} > 70\%}{\text{Total number of samples}}$$

Solution

Example: Set 1 with 6 samples scoring above 70.

78 96 79 80 77 101

$$\text{Proportion} = \frac{6}{16} = 0.375$$

Results

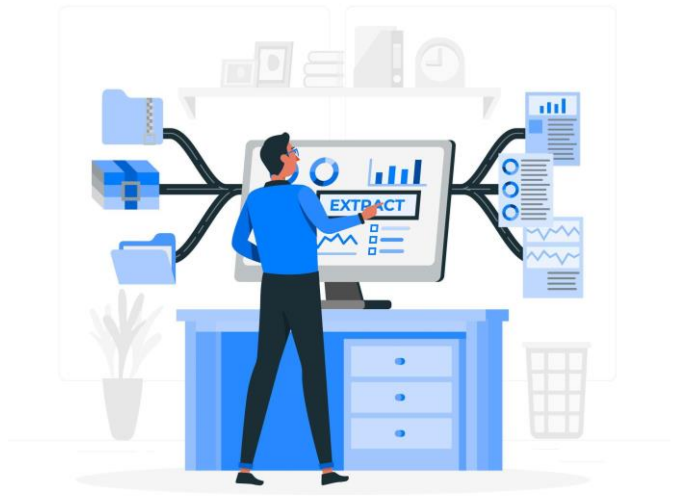
The results from the sample data are as follows:

Sample no.	1	2	3	4	5	6	7	8	9	10	11	12	Avg.	Std. Dev.
Mean	65	62	63	59	58	58	66	55	60	62	55	64	61	3.590815
No. ≥ 70	6	4	4	1	5	3	6	2	5	4	3	6		
Proportion ≥ 70	0.375	0.25	0.25	0.063	0.313	0.188	0.375	0.125	0.313	0.25	0.188	0.375	0.2552	0.108995

Inference

Estimated values of parameters have practical applications such as:

Making predictions



Example: Determining the percentage of students who pass

Making inferences

Example: Evaluating whether the class average exceeds a specified limit

Key Takeaways

- Sampling is a statistical technique that involves performing a predetermined number of observations on a population.
- Known sampling methods can be used to draw conclusions about an entire population from a representative sample.
- In simple random sampling, every sample of a given size has an equal probability of being selected.
- Non-probability sampling involves selecting a sample based on the subjective judgment of the investigator.

Key Takeaways

- The four methods of non-probability sampling include convenience, voluntary response, judgmental or purposeful, and snowball sampling.
- The central limit theorem is applicable in situations involving other distributions.

Knowledge Check

Knowledge Check 1

What is the process of making an inductive inference about an entire population based on a sample?

A. Sampling

B. Sampling error

C. Biases

D. None of the above

Knowledge Check

What is the process of making an inductive inference about an entire population based on a sample?

A. Sampling

B. Sampling error

C. Biases

D. None of the above

The correct answer is **A**

Sampling is the process of making an inductive inference about an entire population based on a sample.

Knowledge Check

2

What is a referral approach to sampling in which individuals provide referrals for further contact?

A. Non-probability sampling

B. Purposive sampling

C. Judgmental sampling

D. Snowball sampling

Knowledge Check

2

What is a referral approach to sampling in which individuals provide referrals for further contact?

A. Non-probability sampling

B. Purposive sampling

C. Judgmental sampling

D. Snowball sampling

The correct answer is **D**

Snowball sampling is a referral approach to sampling in which individuals provide referrals for further contact.

Knowledge Check

3

What is an approach in which the investigator selects units from the population using their knowledge?

A. Non-probability sampling

B. Purposive sampling

C. Snowball sampling

D. None of the above

Knowledge Check

3

What is an approach in which the investigator selects units from the population using their knowledge?

A. Non-probability sampling

B. Purposive sampling

C. Snowball sampling

D. None of the above

The correct answer is **B**

Purposive sampling is an approach in which the investigator selects units from the population using their knowledge.

Thank You