



Hamadan University of Medical Sciences

Familiarity with sampling methods and sample size estimation in clinical research

Sanaz Omidi

PhD in Epidemiology

Department of Epidemiology, School of
Public Health

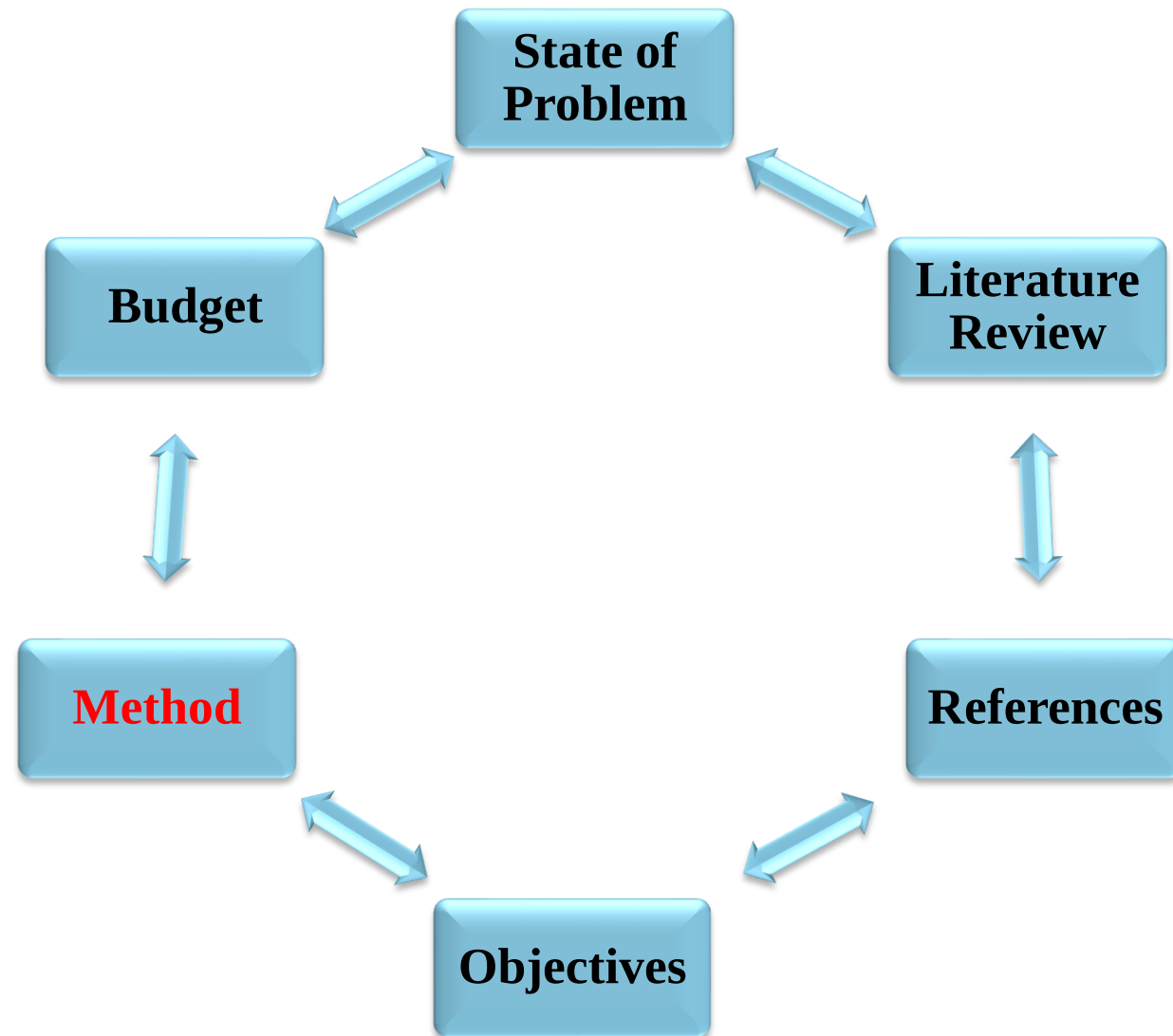
Sanaz.omidi91@gmail.com

Reference



<https://www.aparat.com/poorolajal>

Proposal Components



Method Components

Types of Studies

Variables

Sampling and Sample Size Calculation

Data Collection Techniques and Tools

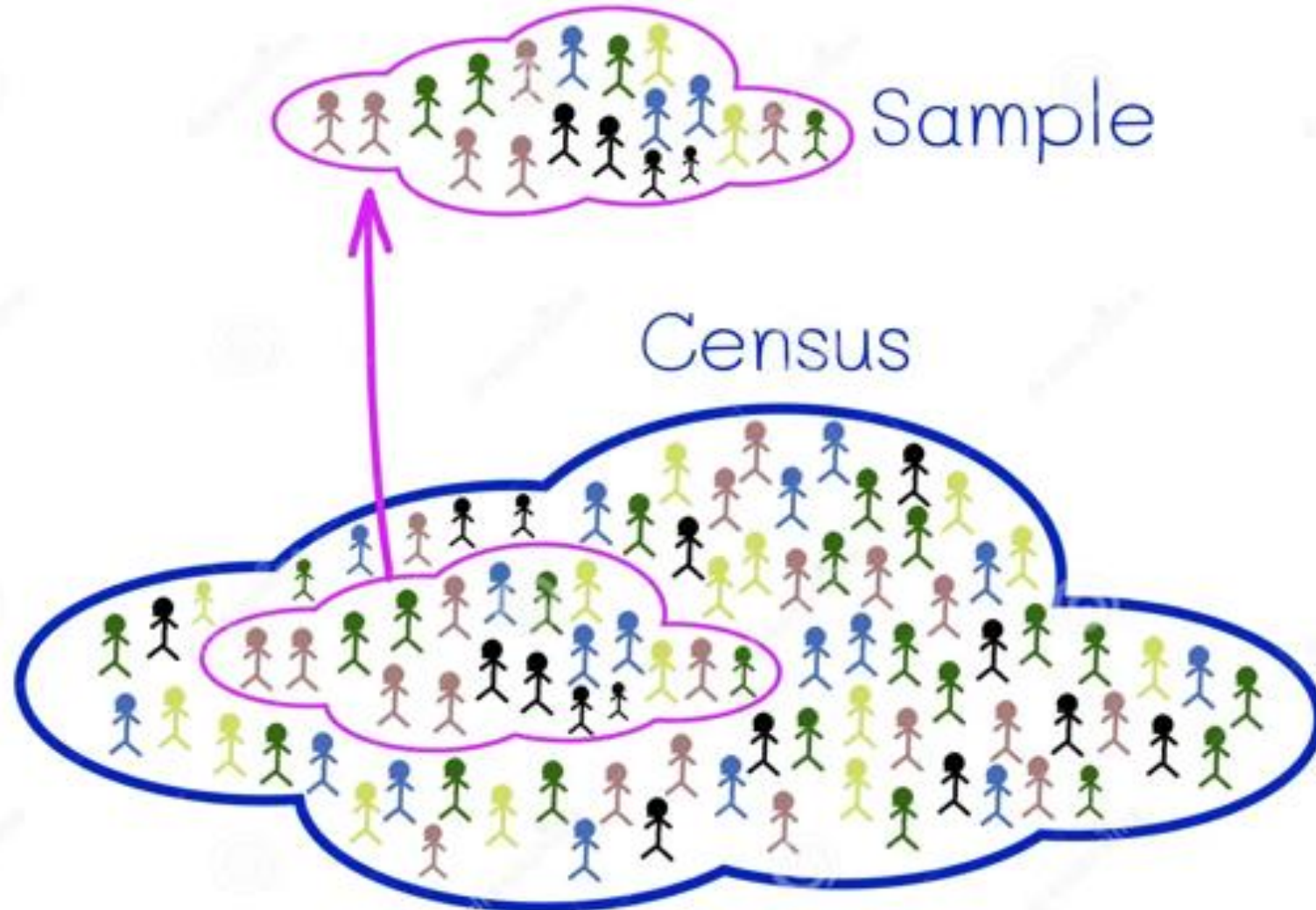
Statistical Methods of Analysis

Ethical Considerations

Census vs. Sampling

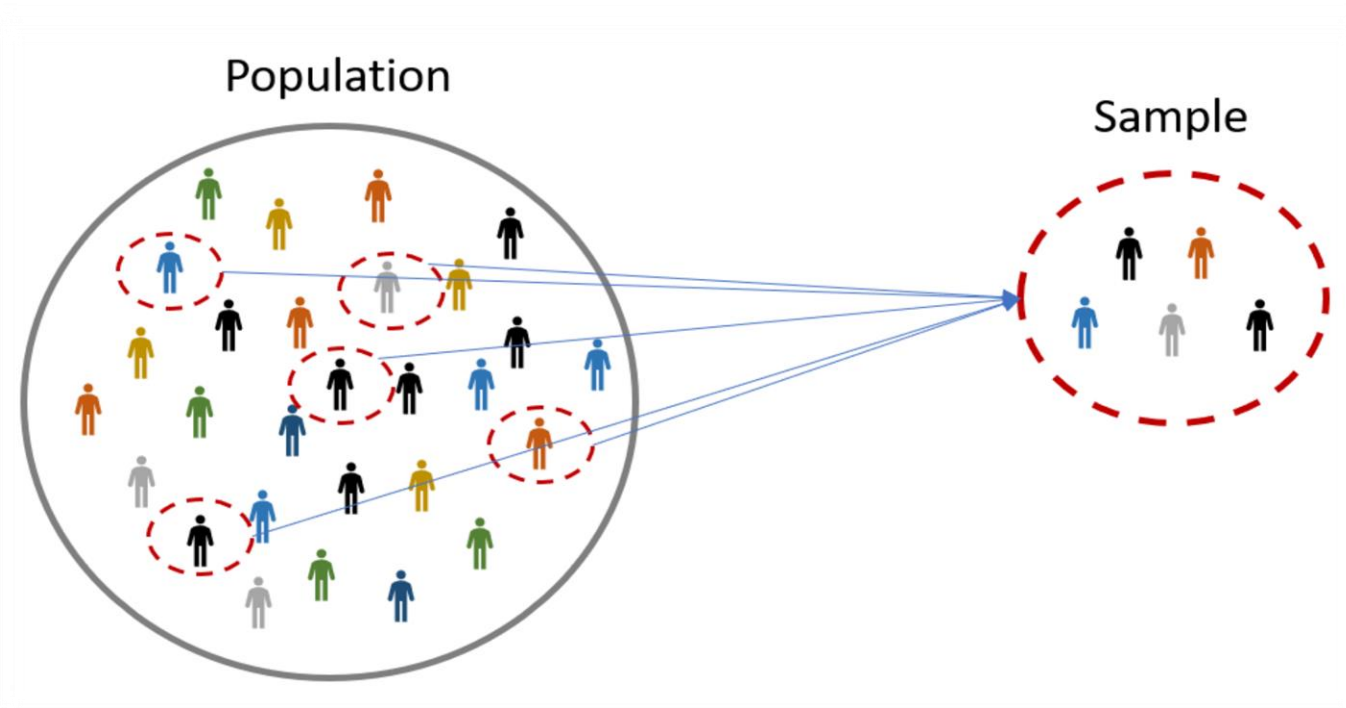
❖ **Census**

❖ **Sample**



Sample

- ❖ A sample is a sub-collection of members selected from a population.
- ❖ **Key Objective:** The primary goal is to make statistical inferences about the whole population based on the data collected from the sample, saving time and resources while maintaining accuracy.



Sample



- ❖ The sample **MUST** be a miniature of the target population.
- ❖ It **MUST** be the same as target population, but with smaller dimensions.



A caricature sample



A representative sample

Randomization vs Random Sampling

Random allocation



Random sampling



Randomization vs Random Sampling

Random allocation

- Critical for clinical trial
- Increase internal validity
- Comparability

Random sampling

- Critical for observational studies
- Increase external validity
- Generalizability

Statistical Population

- ❖ The entire set of individuals, objects, or units that share at least one common characteristic.

The study population can be classified into two main

❖ **Infinite Population:** Members are not fully countable, or access to all of them is impossible.

- ✓ Fish in the sea
- ✓ Butterfly populations

A census is not feasible; therefore, sampling* must be used.

❖ **Finite Population:** Members are

❖ countable.

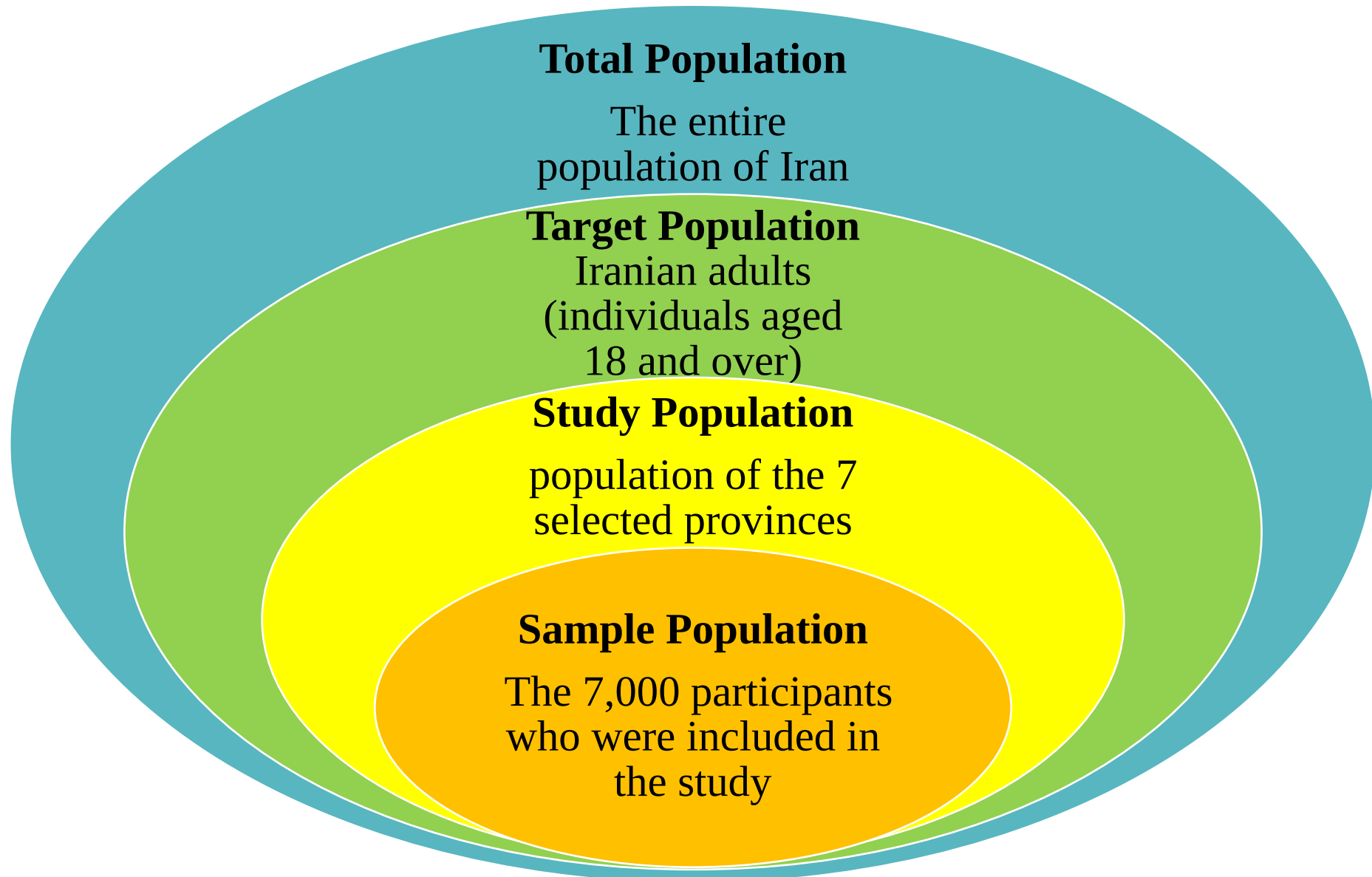
It is possible to examine every individual in the population.

- ✓ Finite Population Workers in a specific factory Students of a university
- If everyone is examined → Census

❖ Why Sampling?

- ✓ Cost Reduction
- ✓ Time-saving
- ✓ Feasibility for large populations
- ✓ Achieving acceptable accuracy through statistical inference
- ✓ Better quality control and monitoring of data collection
- ✓ Essential when a census is impractical or costly

Sampling is a practical, cost-effective, and scientifically valid method for making inferences about a statistical population.



Total Population → Target Population → Study Population → Sample Population

Sampling Frame

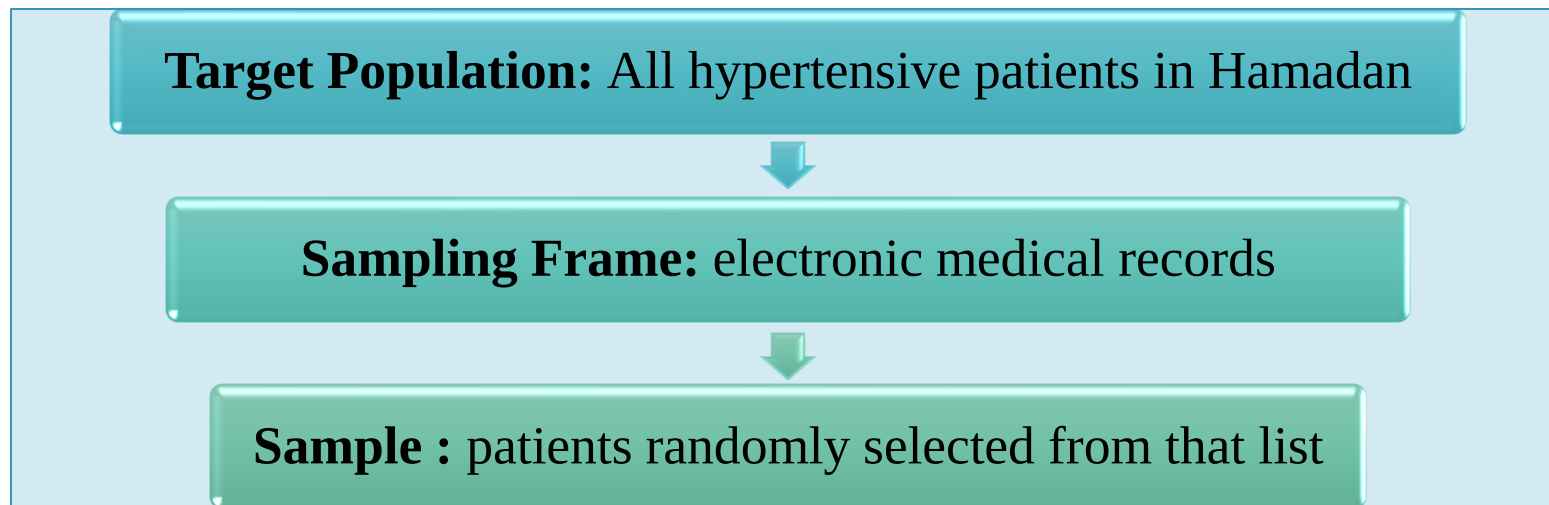
- ❖ A list of population members or the group from which the sample is selected.

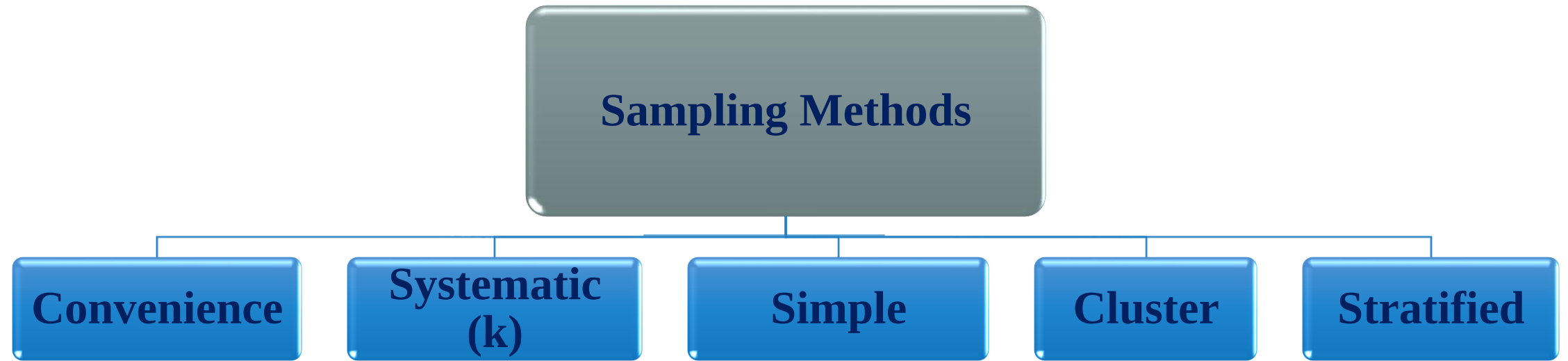
Example: University Students

- ❖ Key Note: In many studies, a complete sampling frame does not exist.

In this case, field sampling is conducted from the community.

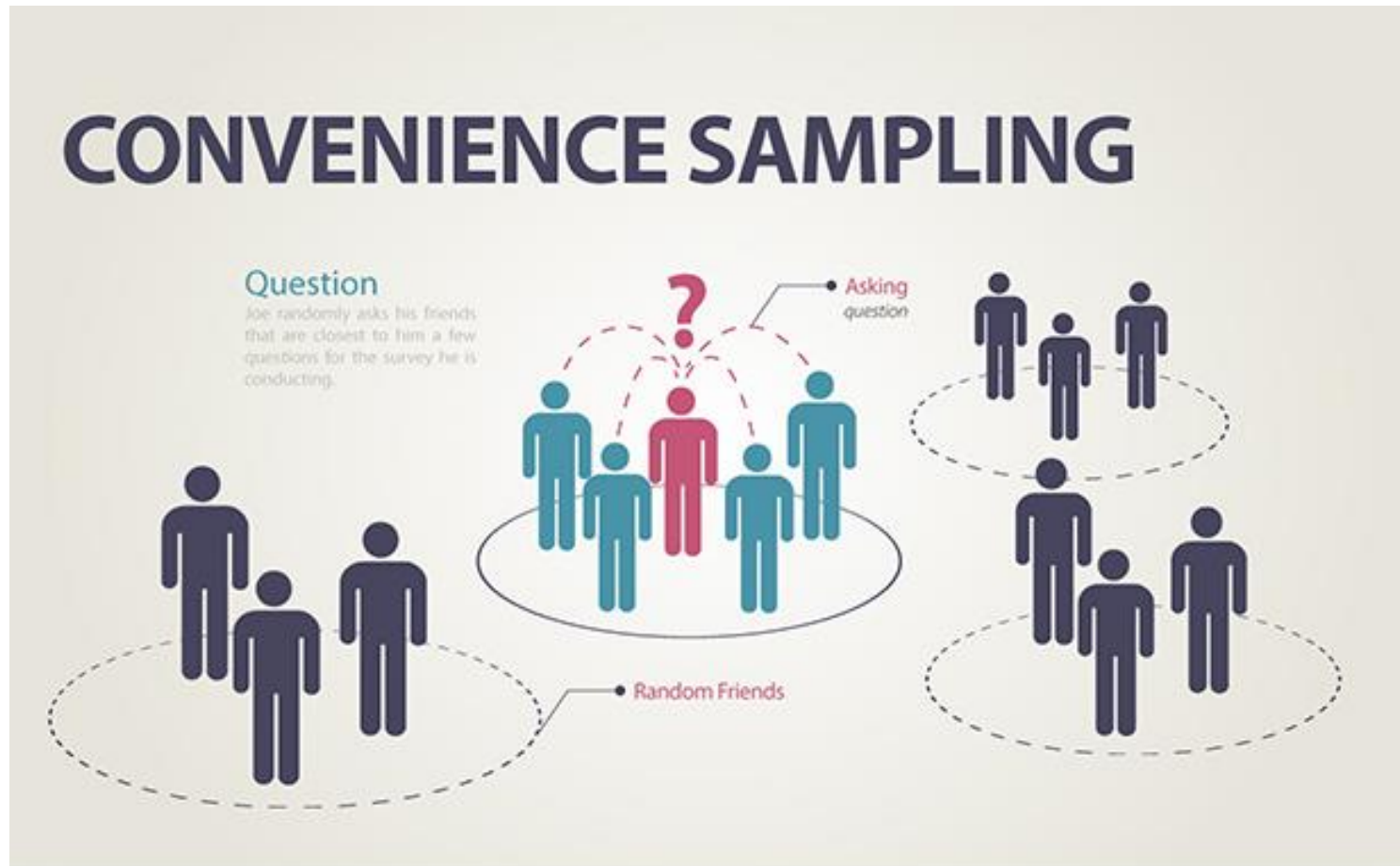
The sampling frame is the basis from which the Study Population.





Convenience Sampling

❖ Convenience sampling collects results that are easiest to obtain.



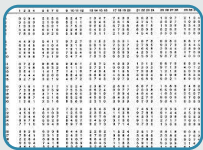
Simple random sampling

- ❖ In this method, every individual in the population has an equal and known chance of being selected.
- ❖ Participants are selected using the following methods:



Computer random number generator

<https://www.randomizer.org/>

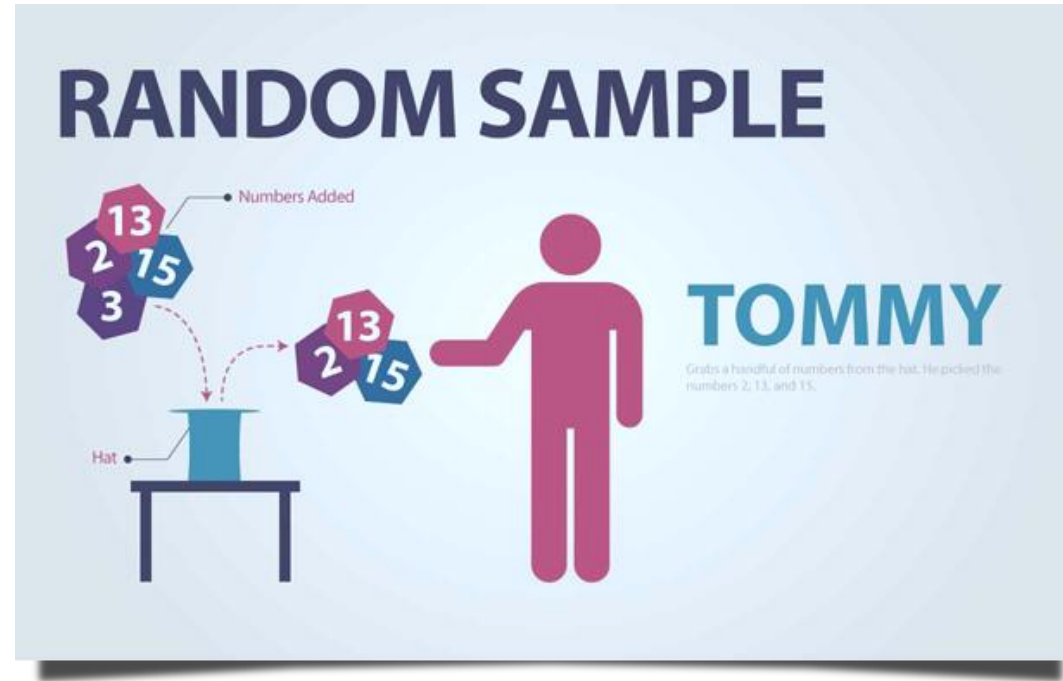


Random number table



Drawing of lots

Simple Random Sampling



In this scenario, every member of the population is assigned a number, and those numbers are placed in a hat. "Tommy" then randomly draws several numbers from the hat. This is a classic example of Simple Random Sampling, as every member of the population has an exactly equal and known chance of being selected.

Simple Random Sampling

- ❑ Sampling can be conducted with or without replacement, depending on the study's objective.
- ✓ **Sampling with Replacement:** After an individual is selected, they are returned to the population; therefore, they may be chosen more than once.
- ✓ **Sampling without Replacement:** Once an individual is selected, they are removed from the list; thus, each person can only be chosen once.

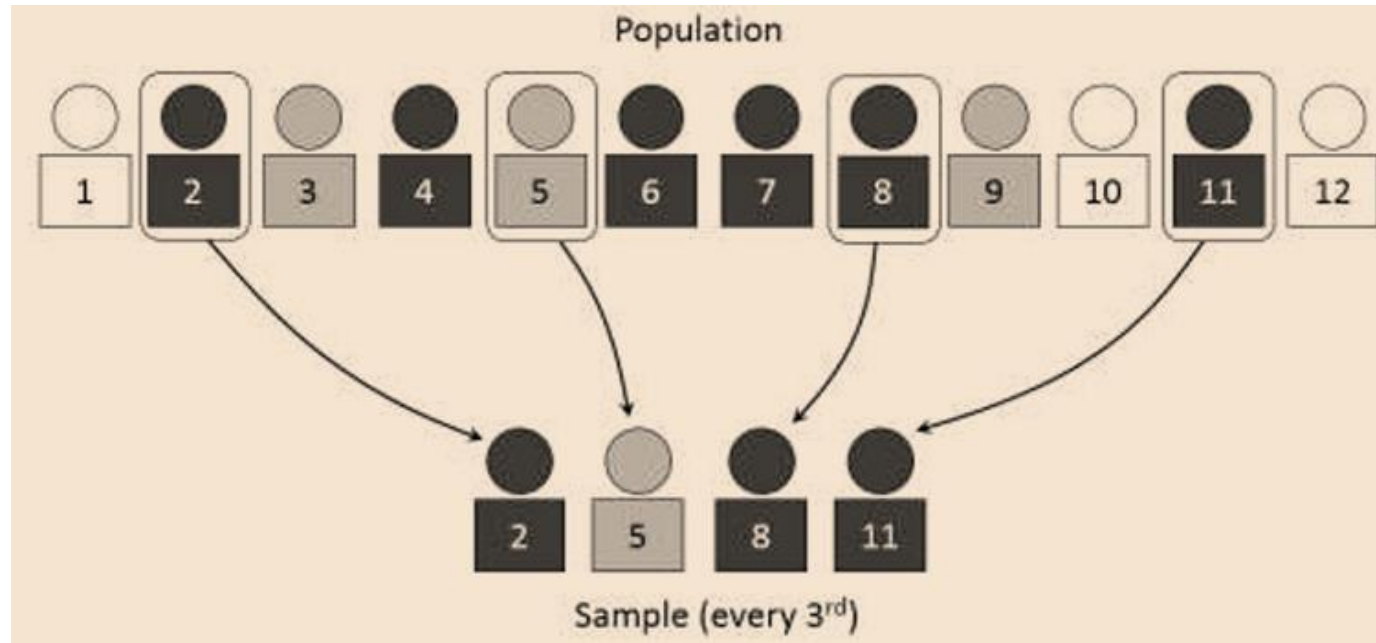
Table of Random Numbers

32 44 44 82 77	59 82 09 61 63	64 65 42 58 43	41 14 54 28 20	03 99 11 04 61
88 57 07 40 15	25 70 49 10 35	01 75 51 47 50	48 96 83 86 03	38 55 59 55 54
95 16 03 92 21	22 30 49 03 14	72 87 71 73 34	39 28 30 41 49	17 54 67 37 04
80 58 04 18 67	17 71 05 96 21	06 55 40 78 50	73 95 07 55 52	32 46 35 28 61
60 47 80 33 43	25 85 25 89 05	57 21 63 96 18	49 85 69 93 26	69 57 26 87 77
95 04 35 26 80	45 78 05 64 87	09 97 15 94 81	37 00 62 21 86	24 12 26 65 91
45 54 77 08 18	59 84 99 61 69	61 45 92 16 47	67 41 71 71 98	61 19 63 02 31
31 73 25 72 60	47 67 00 76 54	45 37 62 53 66	94 74 64 95 80	30 53 22 17 04
85 99 59 03 07	54 30 47 18 03	26 82 50 55 11	12 45 99 13 14	03 78 89 75 99
35 84 18 57 71	08 10 55 99 87	87 11 22 14 76	14 71 37 11 81	48 22 85 33 79
85 30 18 89 77	29 49 06 97 14	73 03 54 12 07	74 69 90 93 10	60 36 39 45 53
54 06 61 52 43	47 72 45 67 33	47 43 14 39 05	31 04 85 66 99	83 79 94 24 02
98 52 52 43 35	24 43 22 48 96	73 27 75 88 84	11 45 61 60 82	32 95 00 76 05
90 39 16 11 05	57 41 10 63 68	53 85 63 07 43	08 67 08 47 41	19 32 25 38 45
26 62 91 90 87	24 47 28 87 79	30 54 02 78 86	61 73 27 54 54	11 22 09 47 47
37 30 14 26 78	45 99 04 32 42	17 37 45 20 03	70 70 77 02 14	31 75 15 72 60
92 00 39 80 85	76 66 87 32 09	59 20 21 19 73	02 90 23 32 50	88 49 29 93 82
16 45 39 45 14	39 01 49 70 66	83 01 20 98 32	25 57 17 76 28	30 93 44 77 44
99 83 70 05 82	81 23 24 49 87	09 50 49 64 12	90 19 37 55 68	22 88 84 88 93
07 36 29 77 03	76 44 74 25 37	98 52 49 78 31	65 70 40 55 14	78 21 21 69 93

Systematic Random sampling

- ❖ Systematic sampling is a method in which units are selected from a sampling frame in a regular and organized manner.
- ❖ Suppose we want to select a sample of size n from a finite population of size N .
- ❖ In systematic sampling, we first calculate the sampling interval $K = \frac{\text{Target Population}}{\text{Sample size}}$
- ✓ Selecting the First Unit Randomly: A number between 1 and K is chosen at random. This number represents the first unit to enter the sample.
- ✓ Selecting Subsequent Units Systematically: Each subsequent unit is selected at an interval of K from the previous one. This process continues until the desired sample size n is reached.

Systematic Random sampling



If the population size is 12 and the sample size is 4 $\rightarrow K=12 \div 4=3$

First, a number between 1 and 3 is chosen at random (for example, the number 2). Then, a selection is made every 3 people: ***2, 5, 8, 11***

In this way, 4 individuals are systematically included in the sample.

Systematic Random sampling

❑ Even if a complete list of the population is not available, systematic sampling can still be used.

Suppose a researcher wants to assess the satisfaction level of clients visiting a center. He decides to interview one person for every five clients.

In this case: $K = 5$

One individual is randomly selected from the first five clients.

Then, every fifth client is included in the study.

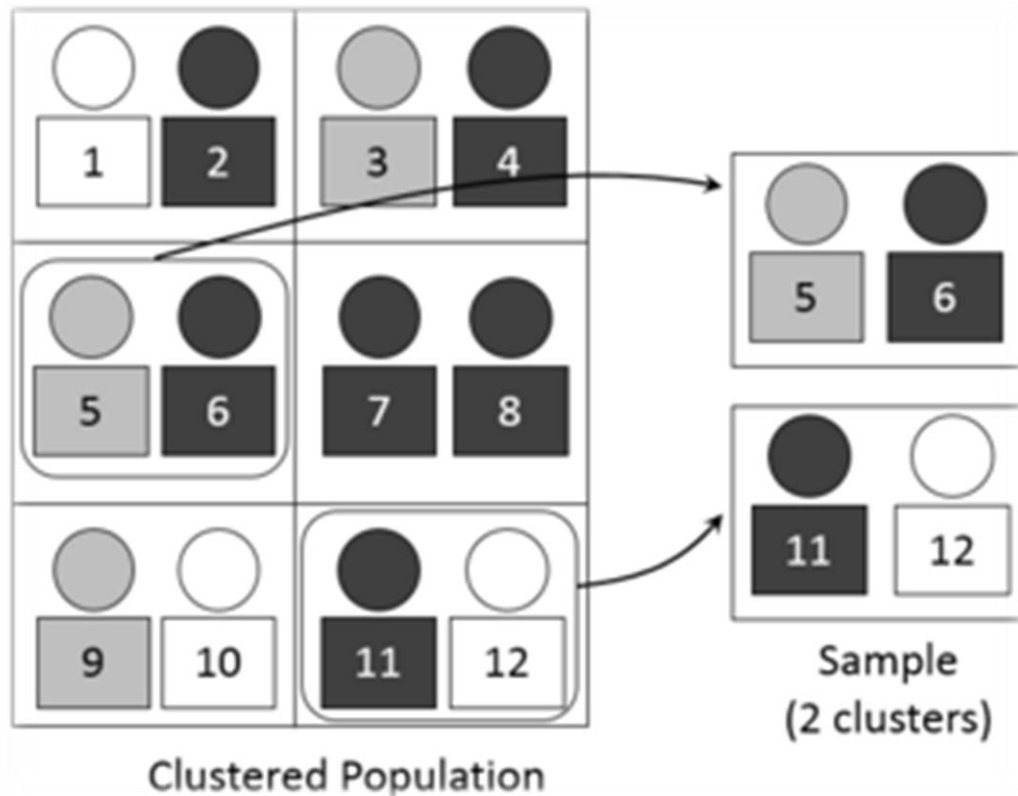
This process continues until the desired sample size is reached.

Cluster Random Sampling

- ❖ In cluster sampling, instead of selecting individual members, the population is divided into groups (clusters), and then a number of clusters are randomly selected.
- ❖ **Applications of Cluster Sampling**
 - **Lack of a complete sampling frame:** When a list of all individuals in the target population is not available.
 - **Geographical dispersion:** When the population is widely dispersed, making simple random sampling very costly, time-consuming, or impractical.
 - **Large-scale studies:** Frequently used in large studies where the goal is to reduce the cost and time required for data collection.

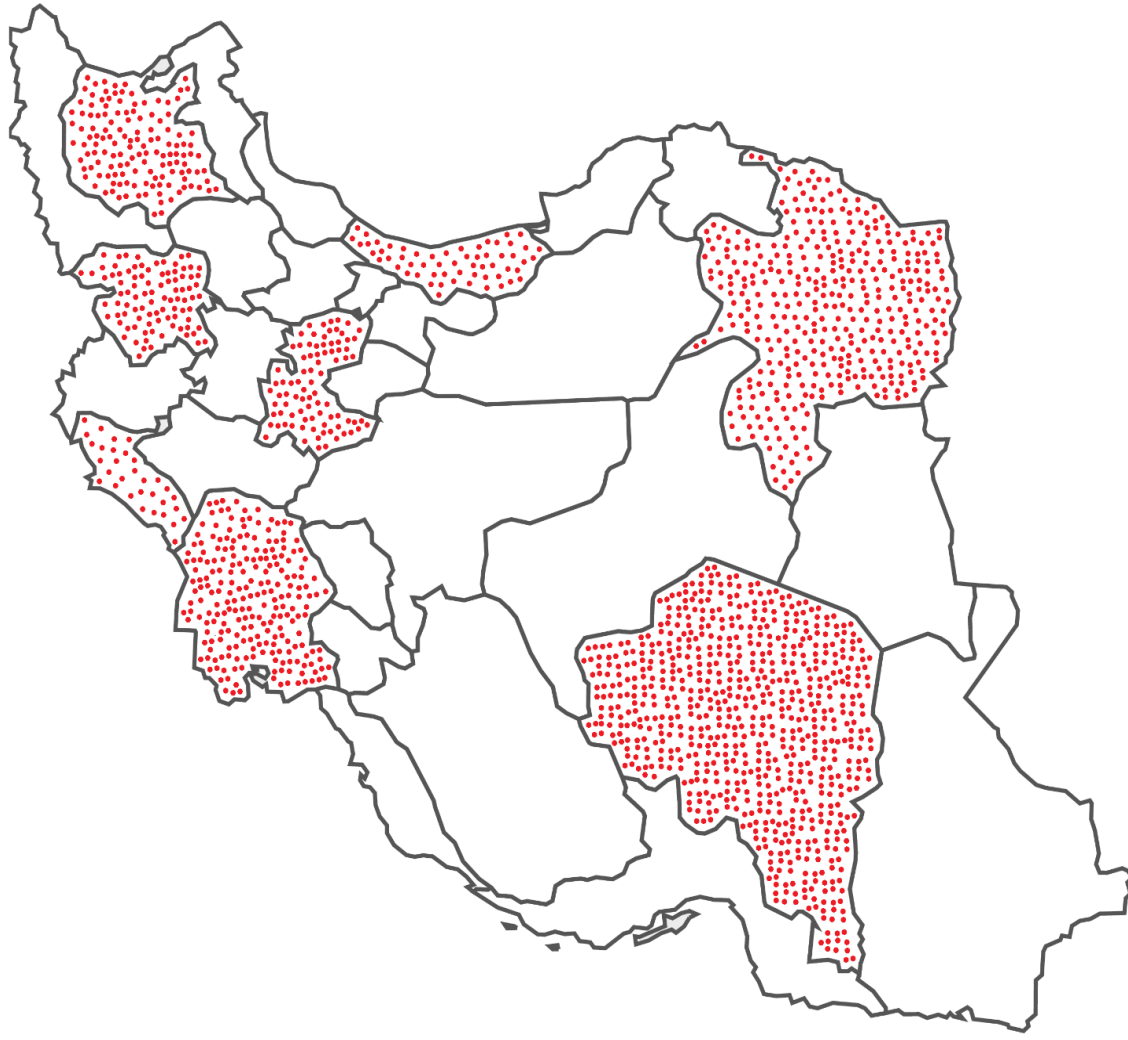
Cluster Random Sampling

In this method, the target population is first divided into several areas or clusters. Then, a number of clusters are randomly selected. After selecting the clusters, the units within those clusters are included in the sample (either all units or a random sample of them).



A group of twelve people is divided into pairs, and then two of these pairs are randomly selected.

Cluster Random Sampling



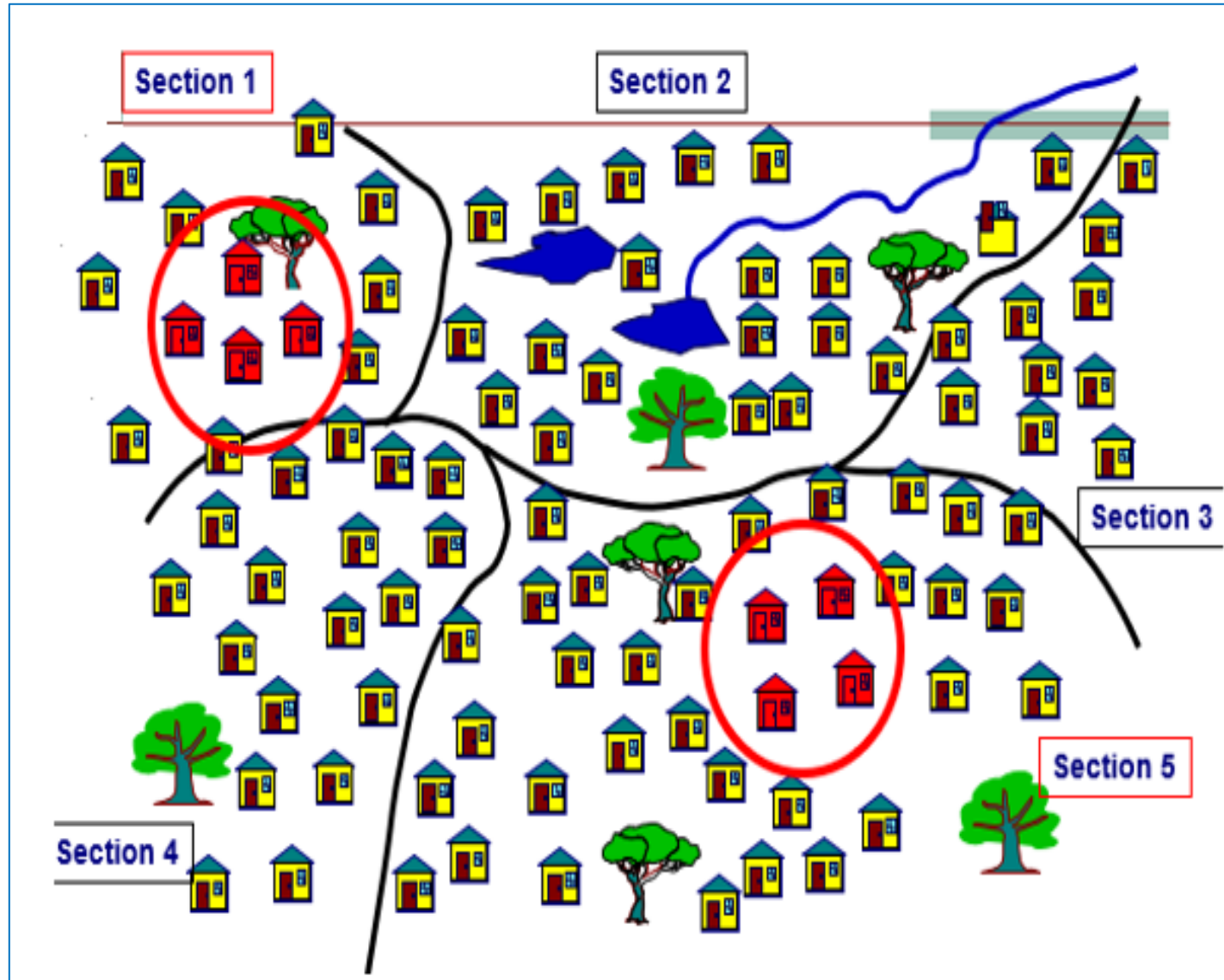
Suppose a researcher wants to conduct a study at the national level.

The researcher aims to select a sample of 4,000 individuals. The country's population is divided into distinct provinces (clusters).

From all the provinces in the country (31 provinces), 8 provinces are randomly selected.

From each of the selected provinces, 500 individuals are then randomly chosen to form the final sample.

Cluster Random Sampling

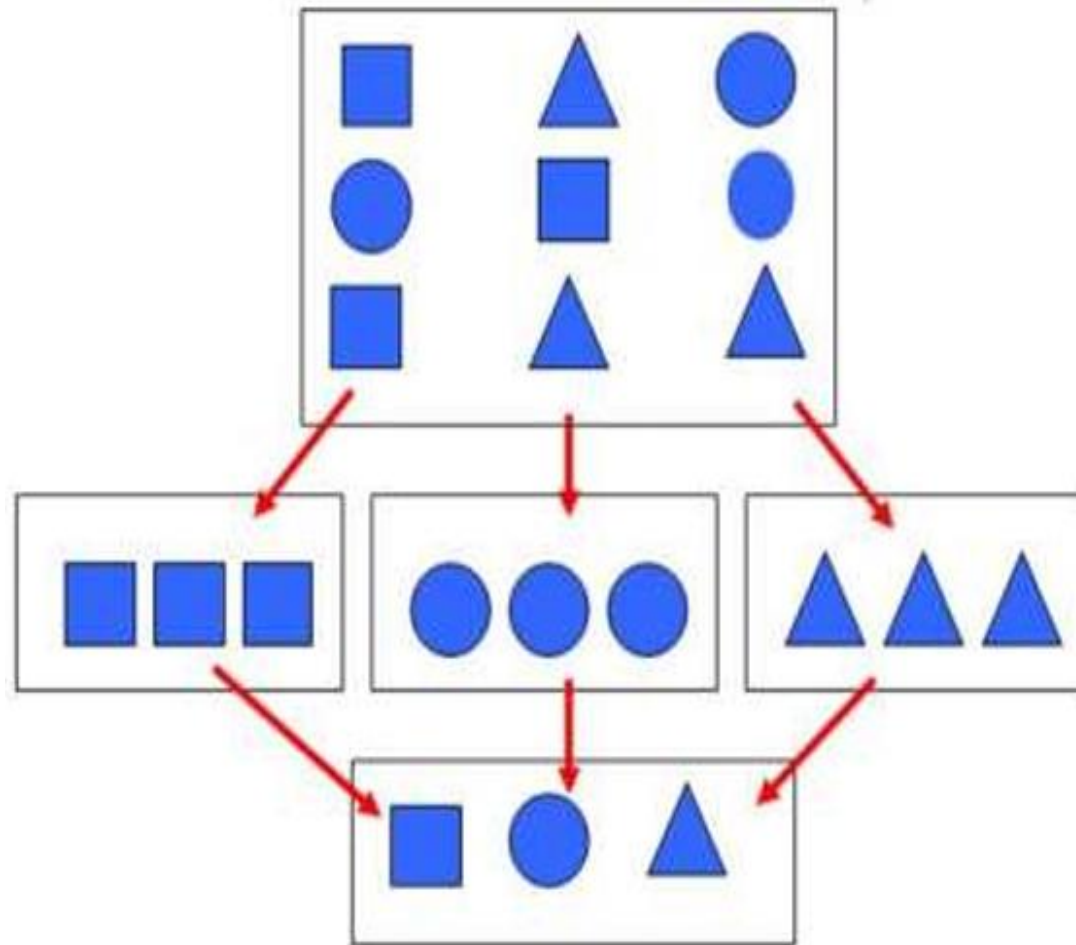


Stratified Random Sampling

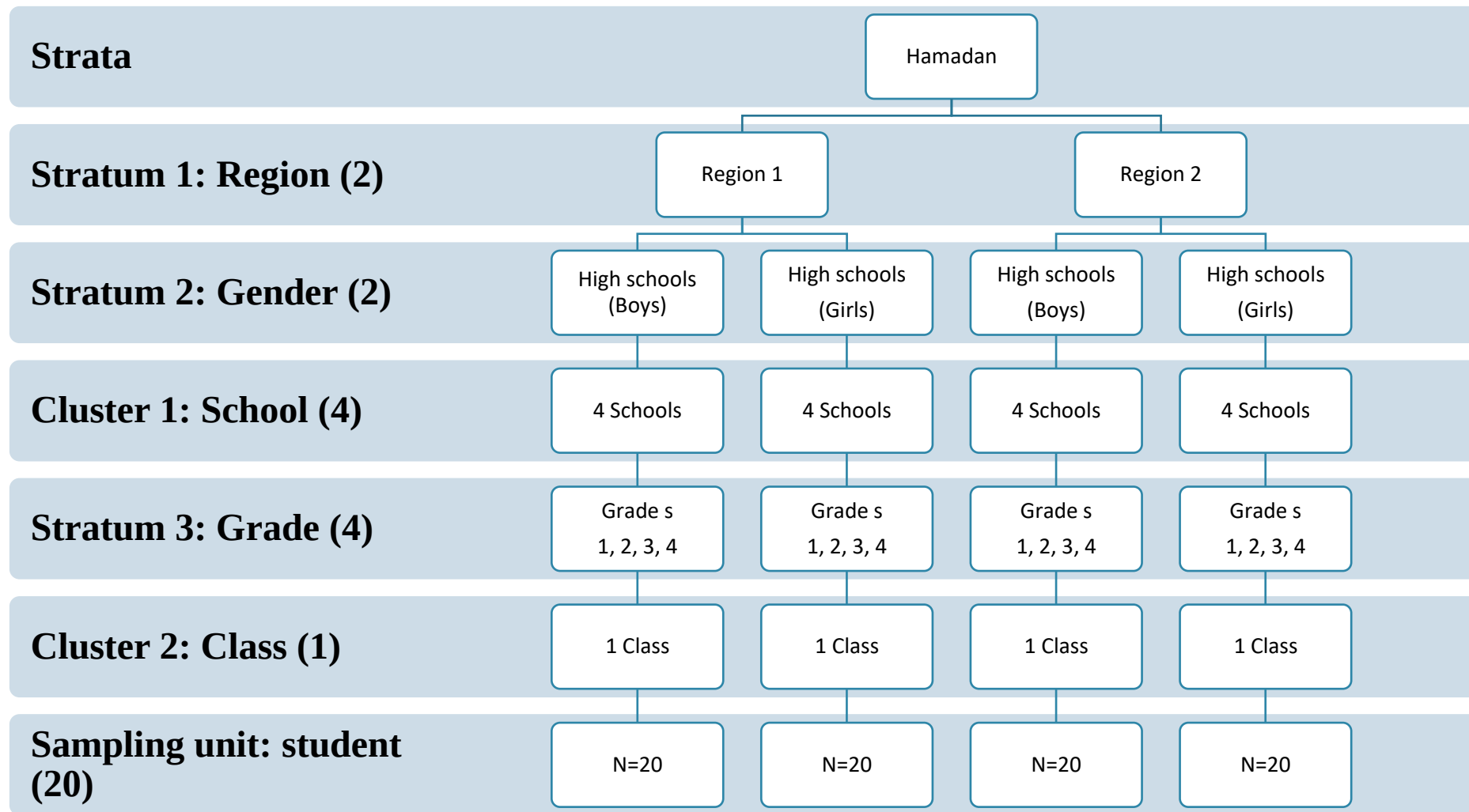
- ❖ Stratified sampling subdivide the population into at least two different groups called strata that share the same characteristics.
- ❖ Then draw a sample from each group.

In **simple sampling**, everything is left to chance; but in **stratified sampling**, we first create order, and then introduce chance within that order.

Stratified Random Sampling



Stratified Cluster Random Sampling



Prevalence of Cigarette Smoking and Associated Risk Factors among Adolescents in Hamadan City, West of Iran In 2010. J Res Health Sci. 2012;12(1):31-7.

Sample Size Calculation



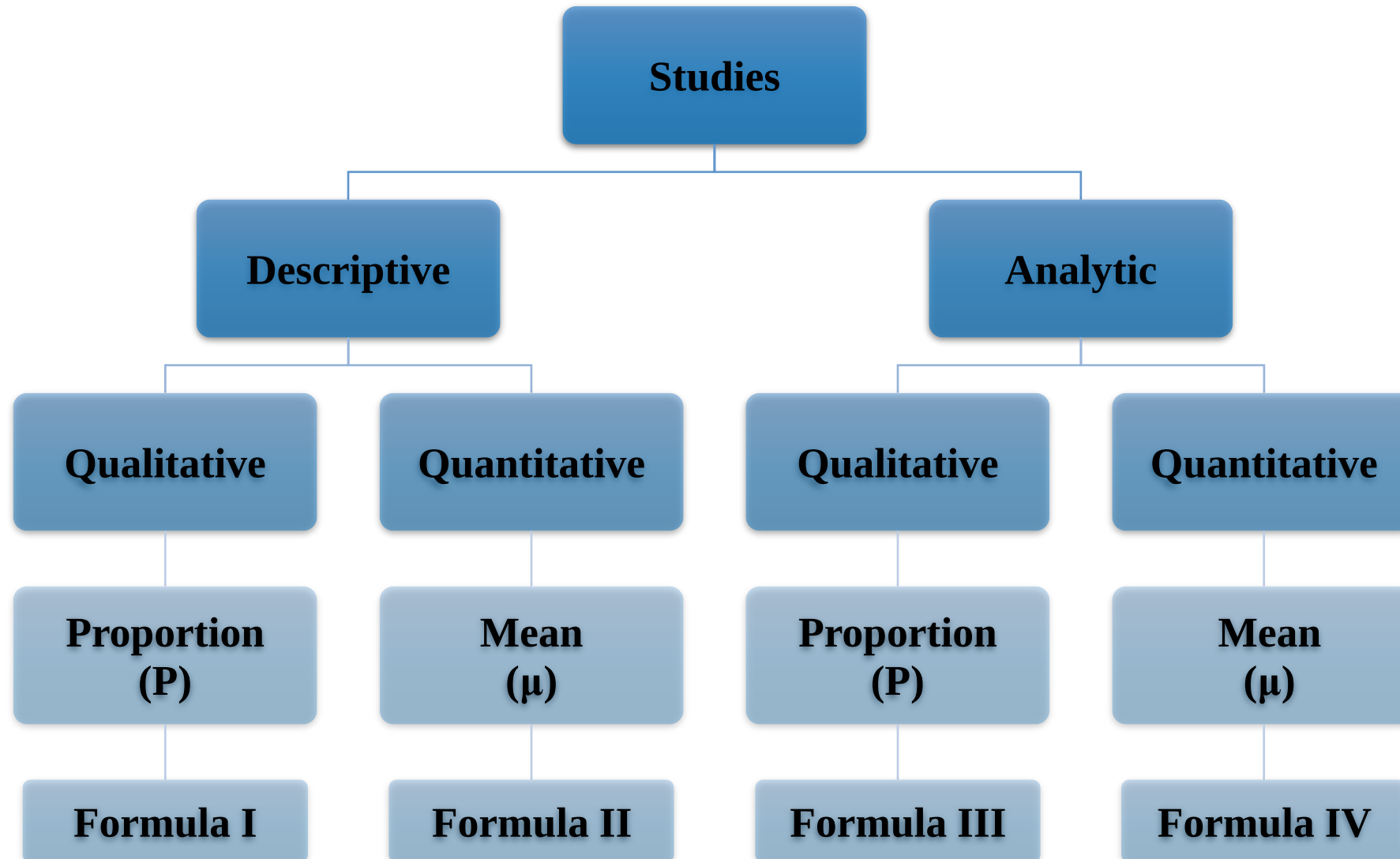
Sample Size Calculation

❖ We have to determine the minimum sample size needed to estimate a parameter, such as:

➤ Mean (μ)

➤ Proportion (P)

Sample Size Formula



Factors effecting sample size

❖ Variation

➤ σ^2

➤ $P(1 - P)$

➤ If variance is zero, a sample of one is enough.

❖ Difference

➤ $\mu_1 - \mu_2$

➤ $P_1 - P_2$

❖ Confidence level $(1 - \alpha)$

➤ 95%, 99%, etc.

❖ Statistical power $(1 - \beta)$

➤ 80%, 90%, etc.

❖ Population size?

Descriptive Studies

❖ Estimation of Probability (P)

❖ **Formula I**

$$n = \frac{Z^2 \times V}{E^2} = \frac{Z^2 \times P (1 - P)}{E^2}$$

❖ Estimation of Mean (μ)

❖ **Formula II**

$$n = \frac{Z^2 \times V}{E^2} = \frac{Z^2 \times \sigma^2}{E^2}$$

Determinants of Sample Size

Descriptive studies

Z

- Confidence level (95%, 98%, 99%, etc.)
- Two-way hypothesis: $Z_{1-\alpha/2}$
- One-way hypothesis: $Z_{1-\alpha}$

V

- Variance for qualitative variables: $P(1 - P)$
- Variance for quantitative variables: σ^2

E

- Margin of error or confidence interval: $\mu \pm E \mid P \pm E$
- Maximum difference between the observed sample estimate ($\bar{x}$) and the true population parameter (μ)
- E is usually considered $\leq 0.2 P$.

Z -values

One-tail	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
Two-tail	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
Z	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

❖ One-tail

➤ $Z_{1-\alpha}$

➤ $Z_{1-\beta}$

❖ Two-tail

➤ $Z_{1-\alpha/2}$

Variance

Parameter P	Parameter σ^2
• Previous studies	• Previous studies
• Pilot study	• Pilot study
• Maximum logically acceptable proportion	• Normal distribution $\sigma = (\text{Max} - \text{Min})/4$
• P is considered 0.5 to maximize variance	• Unknown distribution $\sigma = (\text{Max} - \text{Min})/6$

Example

Estimation of Standard Deviation

- Minimum weight of a newborn: 1500
- Maximum weight of a newborn: 4500

$$\sigma = \frac{4500 - 1500}{4} = \frac{3000}{4} = 750$$

Descriptive Studies

❖ Estimation of Probability (P)

❖ **Formula I**

$$n = \frac{Z^2 \times P \times (1 - P)}{E^2}$$

Example

Cross-sectional Study

❖ Cigarette smoking among adolescents

➤ Prevalence (P) = 14.3%

➤ 95% CI = 1.96

➤ E = 0.2 P

$$n = \frac{1.96^2 \times 0.143 \times 0.857}{(0.2 \times 0.143)^2} = 576$$

Source: Prevalence of cigarette smoking and associated risk factors among adolescent in Hamadan city in 2010. Journal of Research in Health Sciences. 2012;12(1):31-7

Example

Cross-sectional Study

❖ Low birth weight in Iran in 2002

- Prevalence (P) = 7%
- 95% CI = 1.96
- E = 0.2 P

$$n = \frac{Z^2 \times P \times (1-P)}{E^2} = ? \rightarrow n = \frac{1.96^2 \times 0.07 \times 0.93}{(0.2 \times 0.07)^2} = 1276$$

Source: Analyzing socioeconomic related health inequality using the concentration index. Epidemiology, Biostatistics and Public Health. 2014;11(3):e9086-1-10

Example

Cross-sectional Study

❖ The resistance rate of E. coli to co-trimoxazole among outpatients with UTI

- Prevalence (P) = 38.5%
- 95% CI = 1.96
- E = 0.2 P

$$n = \frac{1.96^2 \times 0.385 \times 0.615}{(0.2 \times 0.385)^2} = 154$$

Source: Antibacterial susceptibility of Escherichia coli among outpatients with community-acquired urinary tract infection in Hamadan, Iran. Journal of Global Antimicrobial Resistance. 2015;3:40-3.

Descriptive Studies

❖ Estimation of Mean (μ)

❖ **Formula II**

$$n = \frac{Z^2 \times \sigma^2}{E^2}$$

Example

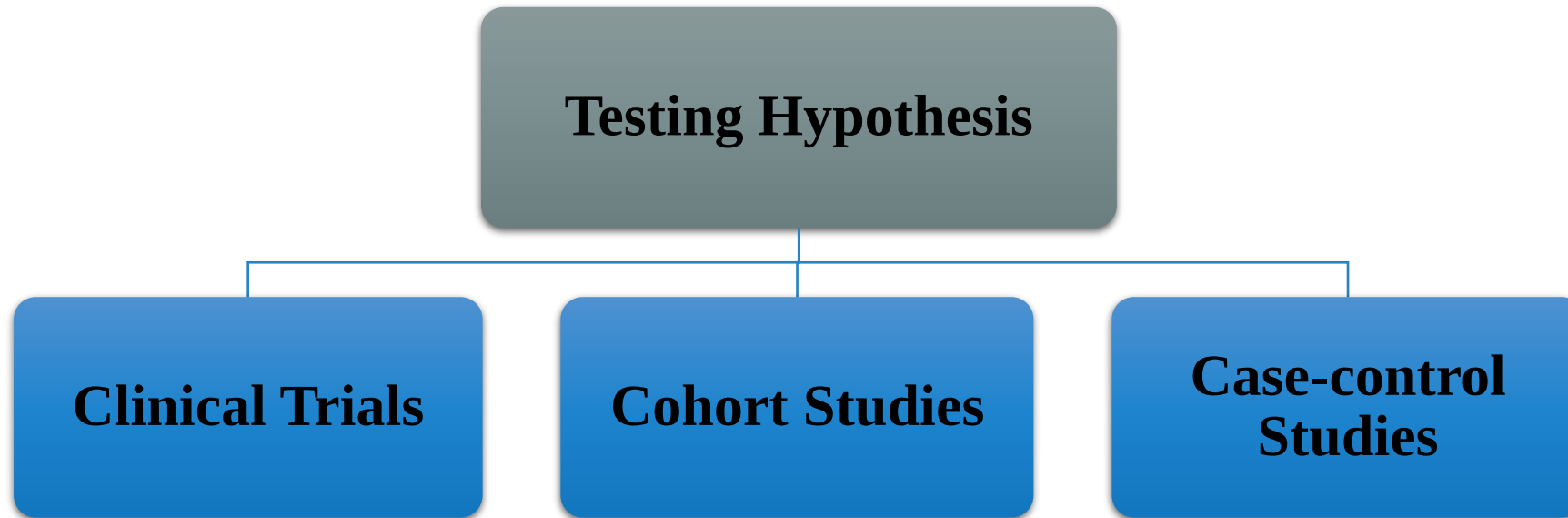
Cross-sectional Study

❖ Height in newborns

- Standard deviation (σ) = 15
- Mean (μ) = 45 cm
- 95% CI = 1.96
- E = 1.5 cm

$$n = \frac{1.96^2 \times 15^2}{(1.5)^2} = 385$$

Analytical Studies



Determinants of Sample Size

Analytic studies

Z

Confidence level (95%, 98%, 99%, etc.)

Two-way hypothesis: $Z_{1-\alpha/2}$

One-way hypothesis: $Z_{1-\alpha}$

V

Variance for qualitative variables: $P(1 - P)$

Variance for quantitative variables: σ^2

E

- $E = P_1 - P_2$
- $E = \mu_1 - \mu_2$

β

- $\beta: Z_{1-\beta}$
- $\beta \leq 4\alpha$

Analytical Studies

Testing Difference in Proportion (P)

Formula III

$$\bar{P} = \frac{P_1 + P_2}{2}$$

$$n = \frac{\left[Z_{1-\alpha/2} \sqrt{2\bar{P}(1-\bar{P})} + Z_{1-\beta} \sqrt{P_1(1-P_1) + P_2(1-P_2)} \right]^2}{(P_1 - P_2)^2}$$

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \times [P_1(1-P_1) + P_2(1-P_2)]}{(P_1 - P_2)^2}$$

$$n = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2 \times \bar{P}(1-\bar{P})}{(P_1 - P_2)^2} \quad \text{For each group}$$

Example

Case-control study

- ❖ Prevalence of periodontitis
- ❖ In rheumatoid subjects: 60%
- ❖ In non-rheumatoid subjects: 16%
- ❖ $\alpha = 0.01$
- ❖ $\beta = 0.01$

n= ?

Example

Case-control study

$$\bar{p} = \frac{0.60 + 0.16}{2} = 0.38$$

$$n = \frac{[2.576\sqrt{2 \times 0.38 \times 0.62} + 2.326\sqrt{0.60 \times 0.40 + 0.16 \times 0.84}]^2}{(0.60 - 0.16)^2} = 53$$

$$n = \frac{(2.576 + 2.326)^2 \times (0.60 \times 0.40 + 0.16 \times 0.84)}{(0.60 - 0.16)^2} = 47$$

$$n = \frac{2(2.576 + 2.326)^2 \times (0.38 \times 0.62)}{(0.60 - 0.16)^2} = 59$$

For each group

Example

Clinical Trial

❖ Quality of the surgical field during endoscopic sinus surgery

➤ Score of bleeding in the intervention group: 14.3%

➤ Score of bleeding in the control group: 42.9%

➤ $\alpha = 0.05$ (one-way)

➤ $\beta = 0.20$

❖ $n = ?$

$$\bar{P} = \frac{0.143 + 0.429}{2} = 0.286$$

$$n = \frac{[1.645\sqrt{2 \times 0.286 \times 0.714} + 0.842\sqrt{0.143 \times 0.857 + 0.429 \times 0.571}]^2}{(0.143 - 0.429)^2} = 30$$

Source: Effect of Topical Tranexamic Acid on Bleeding and Quality of Surgical Field During Functional Endoscopic Sinus Surgery in Patients with Chronic Rhinosinusitis: A Triple Blinded Randomized Controlled Trial. Plos One. 2014;9(8):e104477.

Analytical Studies

Testing Mean (μ) Difference

Formula IV

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \times (\sigma_1^2 + \sigma_2^2)}{(\mu_1 - \mu_2)^2}$$

$$n = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2 \times \sigma^2}{(\mu_1 - \mu_2)^2}$$

Example

Case-control Study

❖ Association between Early Childhood Caries (ECC) and level of alpha-amylase

- Mean (SD) in ECC-active children: 71.8 U/L (48.6)
- Mean (SD) in ECC-free children: 122.8 U/L (101.5)
- $\alpha = 0.01$
- $\beta = 0.05$

❖ $n = ? \rightarrow n = \frac{(2.576+1.645)^2 \times (48.6^2 + 101.5^2)}{(71.8 - 122.8)^2} = 87$ for each group

Analytical Studies

Estimation of Proportion (P) Using OR

$$P_2 = P_1 \frac{OR}{1 + P_1(OR - 1)}$$

Example

Case-control Study

- ❖ Prevalence of unintended pregnancy was in Iran: 24.1%
- ❖ Considering the OR to be at least 2 for the potential risk factors.
- ❖ $\alpha = 0.05$
- ❖ $\beta = 0.10$

$$P_2 = 0.241 \frac{2}{1 + 0.241 (2 - 1)} = 0.388$$

$$\bar{P} = \frac{0.241 + 0.388}{2} = 0.3145$$

$$n = \frac{2}{[1.645\sqrt{2 \times 0.3145 \times 0.6855} + 1.282\sqrt{0.241 \times 0.759 + 0.388 \times 0.612}]^2} = 208$$

Source: Predictors of unintended pregnancy among married women in Hamadan, Eastern Iran: a case-control study. Iranian Journal of Public Health. 2013;42(8):854-9.

Imbalance Sample Size

“r” is the Proportion of Controls to Cases

- ❖ Risk factors for unwanted pregnancy
- ❖ P_1 for unwanted pregnancy: 24.1%
- ❖ Sample size ($\alpha=0.05$, $\beta=0.10$): 208
- ❖ $n_{cases} = \frac{r+1}{2r} n = \frac{2+1}{2 \times 2} \times 208 = 156$
- ❖ $n_{controls} = r \times n_{cases} = 2 \times 156 = 312$

Source: Predictors of unintended pregnancy among married women in Hamadan, Eastern Iran: a case-control study. Iranian Journal of Public Health. 2013;42(8):854-9.

Imbalance Sample Size

“r” is the Proportion of Controls to Cases

r	Cases	Controls
1	208	208
2	156	312
3	139	417
4	130	520

Correction of Sample Size

Dropouts

- ❖ If we estimate that, for instance, 20% of the participants will be withdrawal.
- ❖ We should correct the sample size for these dropouts as follows.

$$n_c = \frac{1}{1 - 0.2} \times n$$

Correction of Sample Size

Dropouts

❖ Sample size: 105

❖ Estimated dropouts: 5%

$$n' = \frac{1}{1 - 0.05} n = \frac{1}{0.95} \times 105 = 110$$

Source: Effect of Topical Sucralfate vs Clindamycin on Posttonsillectomy Pain in Children Aged 6 to 12 Years A Triple-Blind Randomized Clinical Trial. JAMA Otolaryngology - Head & Neck Surgery. 2014;140(8):698-703

Correction of Sample Size

Finite Population

❖ When population is limited and estimated sample size (n) is more than 5% of total population (N); we can correct the sample size:

$$n > 5\% N$$

$$n_c = \frac{n}{1 + \frac{n}{N}}$$

Example 1

Finite Population

- ❖ Prevalence of brucellosis in HIV+ subjects: 6%
- ❖ $\alpha = 0.05$
- ❖ $d = 0.2P$
- ❖ Number of HIV+ subjects in Hamadan = 546

$$n = \frac{1.96^2 \times 0.06 \times 0.94}{(0.2 \times 0.06)^2} = 1505 \rightarrow n > 5\% N$$

$$n_c = \frac{n}{1 + \frac{n}{N}} = \frac{1505}{1 + \frac{1505}{546}} = 401$$

A Control Group for Multiple Comparison

- ❖ We should use a larger control group in comparison with other intervention or case groups
- ❖ n = Sample size
- ❖ k = Number of groups

$$n_{\text{control}} = n\sqrt{k}$$



THANK YOU