SAMPLING
and
SAMPLING METHODS

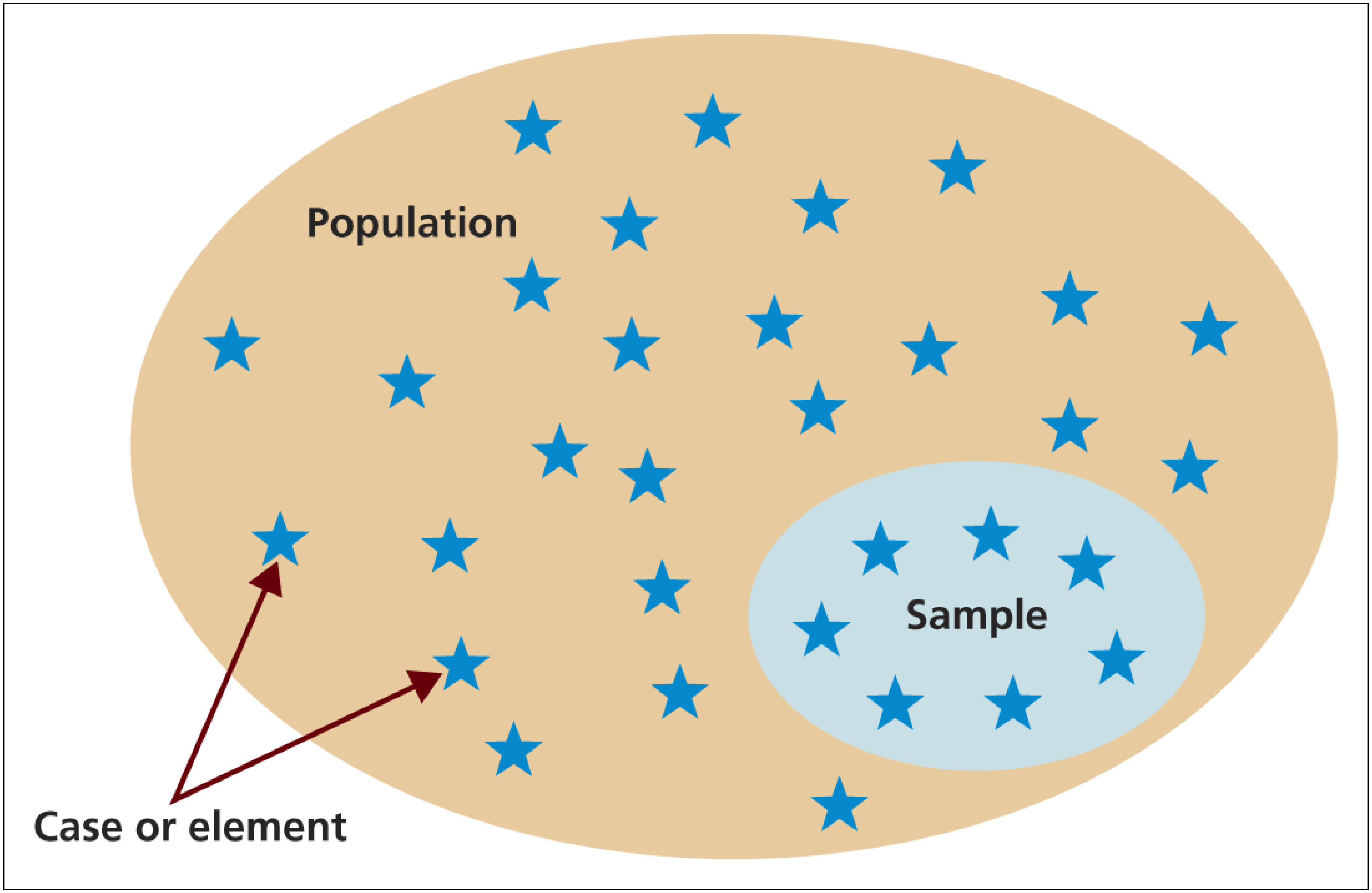
SAMPLING

- If the data you collect really are the same as you would get from the rest, then you can draw conclusions from those answers which you can relate to the whole group.

- This process of selecting just a small group of cases from out of a large group is called **sampling**.

SAMPLING

- A sample is “a smaller (but hopefully representative) collection of units from a population used to determine truths about that population” (Field, 2005)
- Why sample?
  - Resources (time, money) and workload
  - Gives results with known accuracy that can be calculated mathematically
- The sampling frame is the list from which the potential respondents are drawn
  - Registrar’s office
  - Class rosters
  - Must assess sampling frame errors


Population
Sample
Case or element

The need to sample

Sampling- a valid alternative to a census when;

1. A survey of the entire population is impracticable

2. Budget constraints restrict data collection

3. Time constraints restrict data collection

4. Results from data collection are needed quickly

When doing a survey, the question inevitably arises:

- How representative is the sample of the whole population, in other words;
- How similar are characteristics of the small group of cases that are chosen for the survey to those of all of the cases in the whole group?

(i) The Census Method or Complete Enumeration.

(ii) The Sample Method or Partial Enumeration.

Population in Research

- It does not necessarily mean a number of people, it is a collective term used to describe the total quantity of things (or cases) of the type which are the subject of your study.

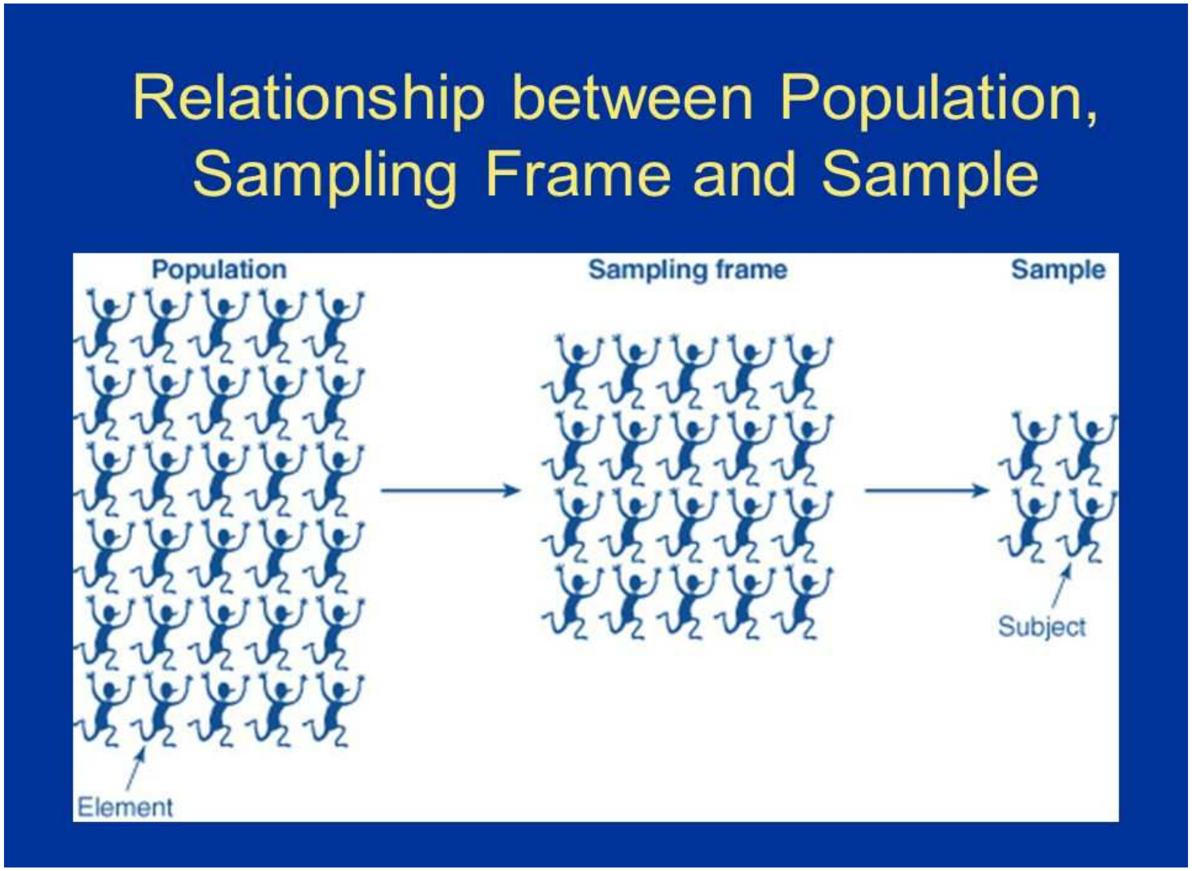
- So a ***population*** can consist of certain types of objects, organizations, people or even events.

Population

- Note also that the population from which the sample is drawn may not be the same as the population about which we actually want information.
- Often there is large but not complete overlap between these two groups due to frame issues etc .
- Sometimes they may be entirely separate - for instance, we might study rats in order to get a better understanding of human health, or we might study records from people born in 2008 in order to make predictions about people born in 2009.

Sampling Frame

- Within this population, there will probably be only certain groups that will be of interest to your study, this ***selected category*** is your sampling frame.



Relationship between Population, Sampling Frame and Sample

Populations can have the following characteristics:

Characteristics	Explains	Examples
homogeneous	all cases are similar	bottles of beer on a production line
stratified	contain strata or layers	people with different levels of income: low, medium, high
proportional stratified	contains strata of known proportions	percentages of different nationalities of students in a university
grouped by type	contains distinctive groups	of apartment buildings – towers, slabs, villas, tenement blocks
grouped by location	different groups according to where they are	animals in different habitats – desert, equatorial forest, savannah, tundra

SAMPLING……

- What is your population of interest?
    - To whom do you want to generalize your results?
        - All doctors
        - School children
        - Indians
        - Women aged 15-45 years
        - Other
- Can you sample the entire population?

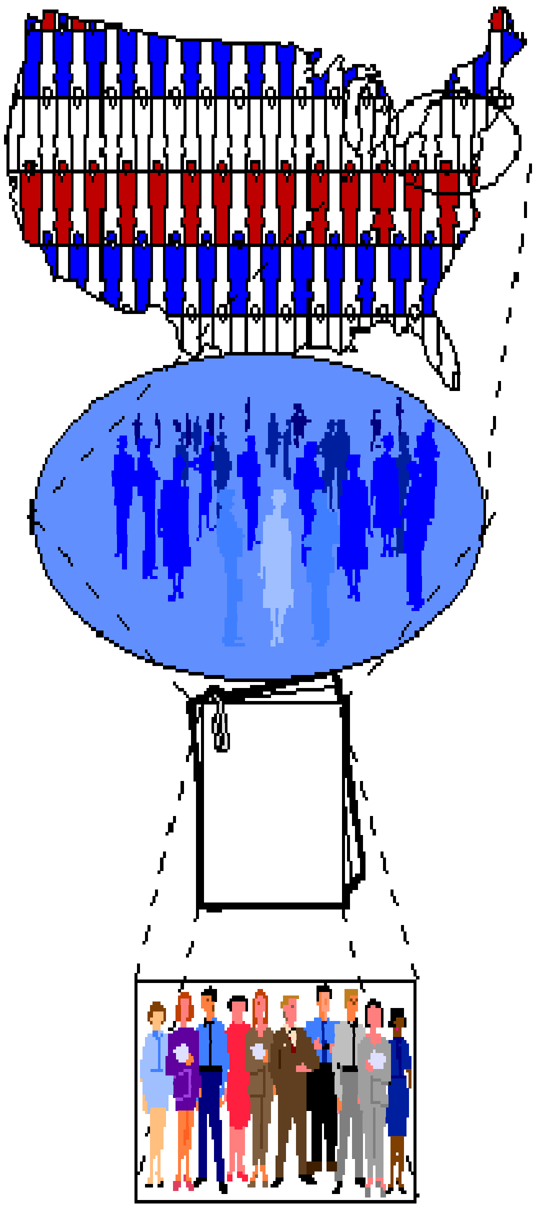
SAMPLING…….

- 3 factors that influence sample representative-ness
  - Sampling procedure
  - Sample size
  - Participation (response)

- When might you sample the entire population? (Census Method)
  - When your population is very small
  - When you have extensive resources
  - When you don’t expect a very high response

Who do you want to generalize to?

The Theoretical Population



What population can you get access to?

The Study Population

How can you get access to them?

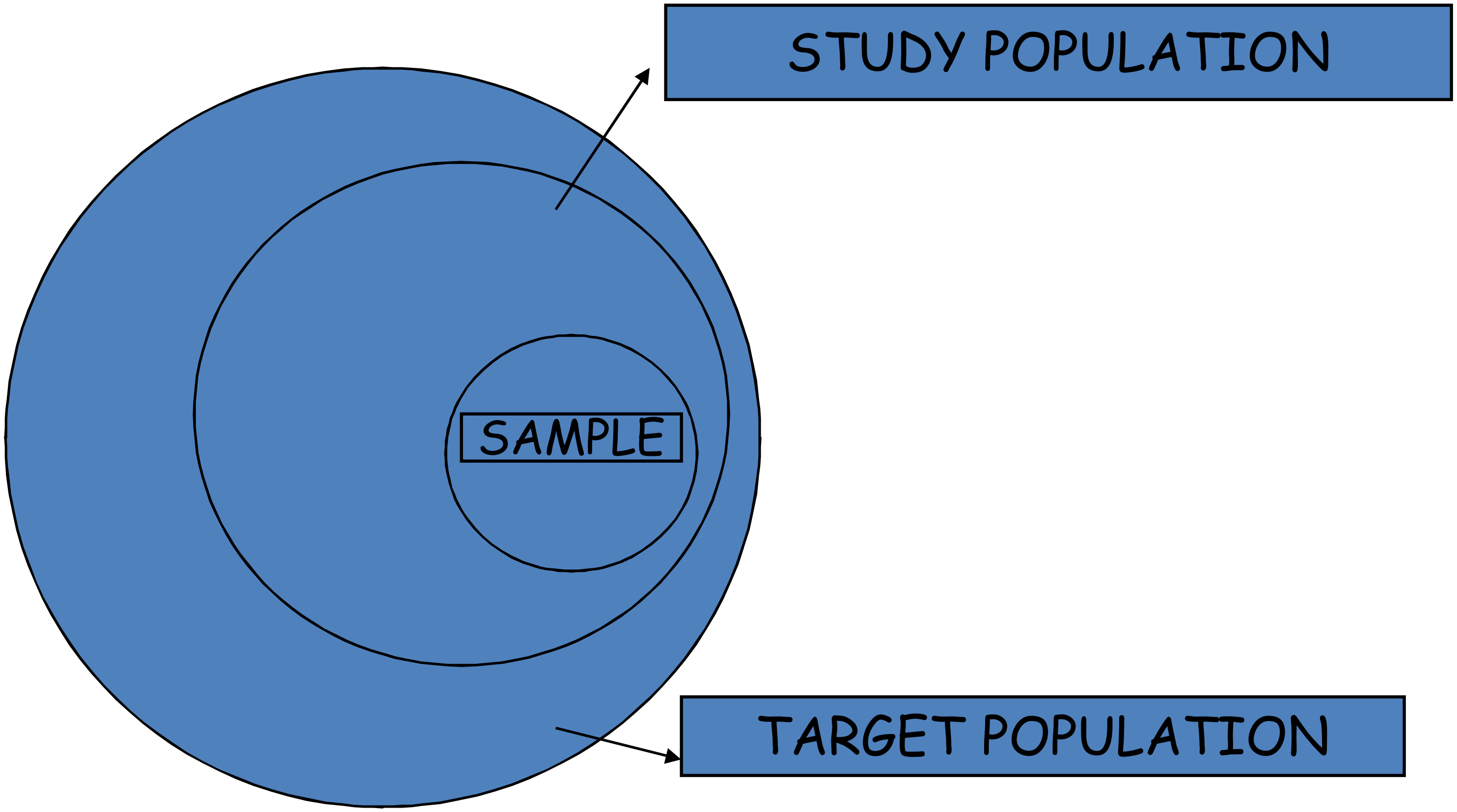
The Sampling Frame

Who is in your study?

The Sample

SAMPLING BREAKDOWN

SAMPLING.......



STUDY POPULATION

SAMPLE

TARGET POPULATION

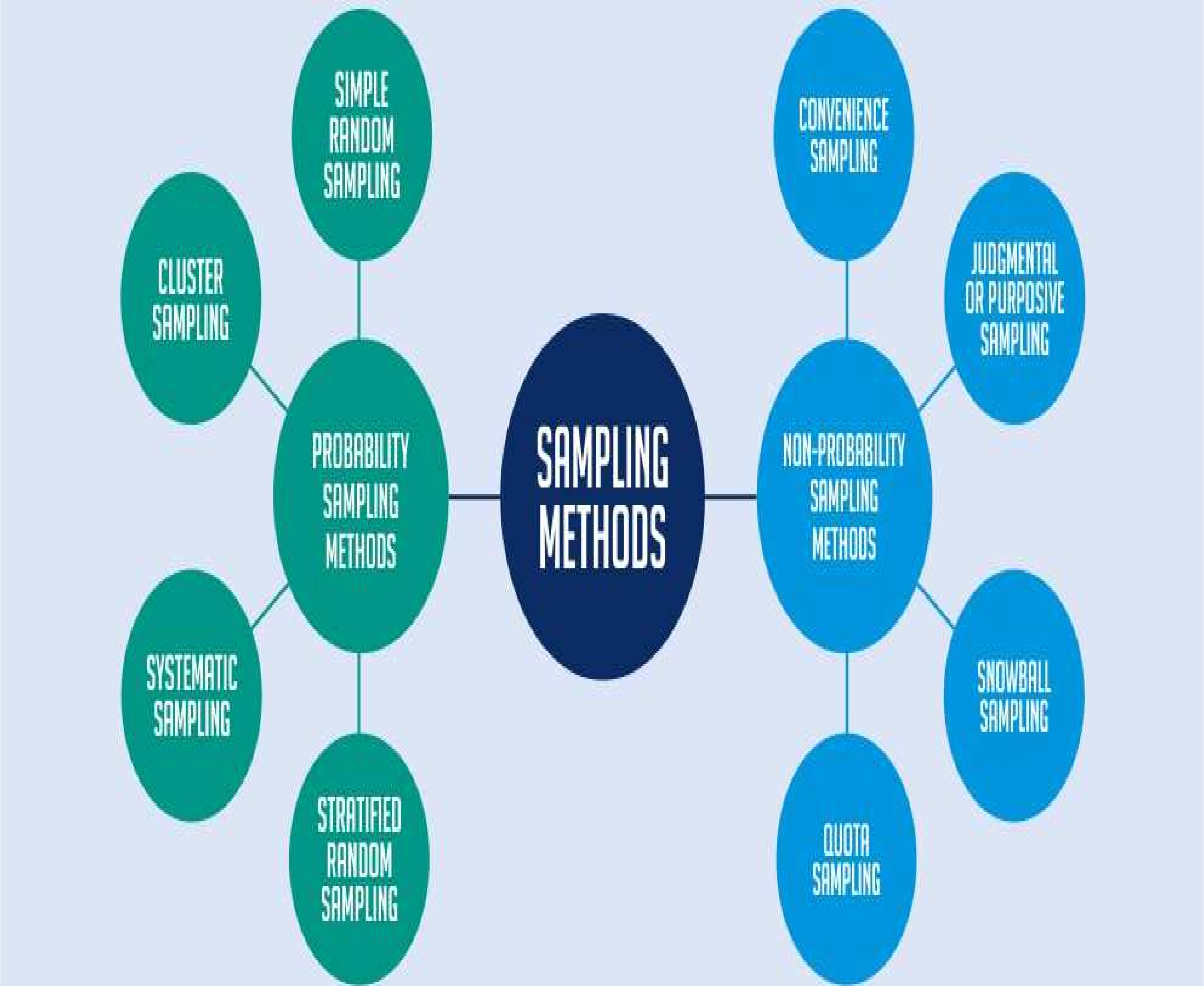
SAMPLING METHODS

Process

The sampling process comprises several stages:

1. Defining the population of concern
2. Specifying a sampling frame, a set of items or events possible to measure
3. Specifying a sampling method for selecting items or events from the frame
4. Determining the sample size
5. Implementing the sampling plan
6. Sampling and data collecting
7. Reviewing the sampling process

Sampling techniques



- ✓ **Probability sampling** techniques give the most reliable representation of the whole population.

- ✓ **Non-probability techniques**, relying on the judgment of the researcher or on accident, cannot generally be used to make generalizations about the whole population.

Probability Sampling

- It is a sampling technique in which sample from a larger population are chosen using a method based on the theory of probability.
- For a participant to be considered as a probability sample, he/she must be selected using a random selection.
- The most important requirement of probability sampling is that everyone in your population has a known and an equal chance of getting selected.
- Equal Probability of Selection' (EPS)
- Probability sampling uses statistical theory to select randomly, a small group of people (sample) from an existing large population and then predict that all their responses together will match the overall population.

Types of Probability Sampling

Four main techniques used for a probability sample:

- Simple random sampling
- Stratified random sampling
- Cluster sampling
- Systematic sampling

Simple random sampling

- As the name suggests is a completely random method of selecting the sample. This sampling method is ***as easy as assigning numbers to the individuals (sample)*** and then randomly choosing from those numbers through an automated process.

Simple random sampling

- Applicable when population is small, homogeneous & readily available
- Estimates are easy to calculate.
- Simple random sampling is always an EPS design, but not all EPS designs are simple random sampling.

Disadvantages

- If sampling frame large, this method impracticable.
- Minority subgroups of interest in population may not be present in sample in sufficient numbers for study.

Selection of a Simple Random Sample

(i) Lottery Method.

(ii) Use of Table of Random Numbers.

Replacement of selected units

Sampling schemes may be

- *without replacement* ('WOR' - no element can be selected more than once in the same sample)
- *with replacement* ('WR' - an element may appear multiple times in the one sample).

For example, if we catch fish, measure them, and immediately return them to the water before continuing with the sample, this is a WR design, because we might end up catching and measuring the same fish more than once. However, if we do not return the fish to the water (e.g. if we eat the fish), this becomes a WOR design.

Example

Draw a random sample without replacement of 15 students from a class of 450 students

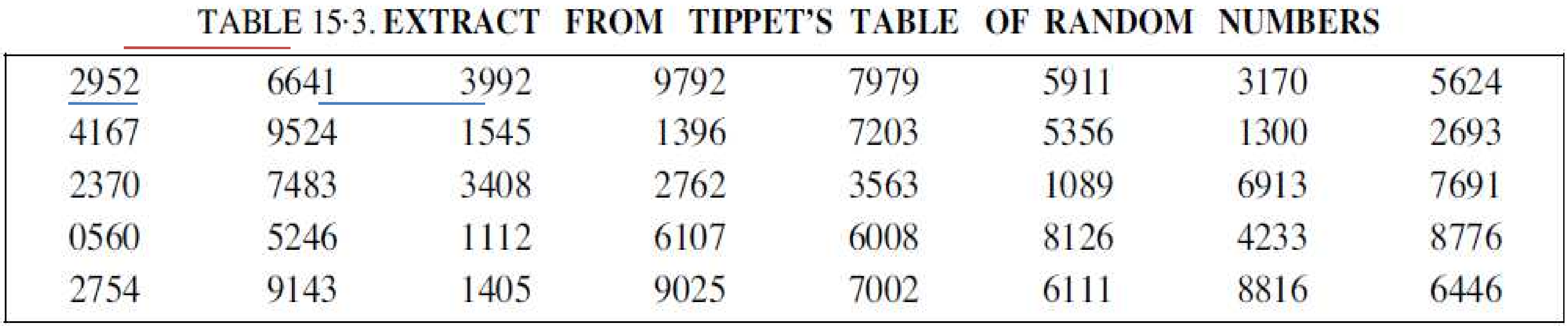


TABLE 15·3. EXTRACT FROM TIPPET'S TABLE OF RANDOM NUMBERS

2952	6641	3992	9792	7979	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2370	7483	3408	2762	3563	1089	6913	7691
0560	5246	1112	6107	6008	8126	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

Numbers grouped in three's are : 295, 266, 413, 992, 979, 279, 795, 911, 317, 056, 244, 167, 952, 415, 451, 396, 720, 353, 561, 300, 269, 323, 707, 483, 340 …

295, 266, 413, 279, 317, 56, 244, 167,
415 396, 353, 300, 269, 323, and 340,

Example

Draw a random sample (without replacement) of size 5 students from a class of 450 students from a population of 24 units

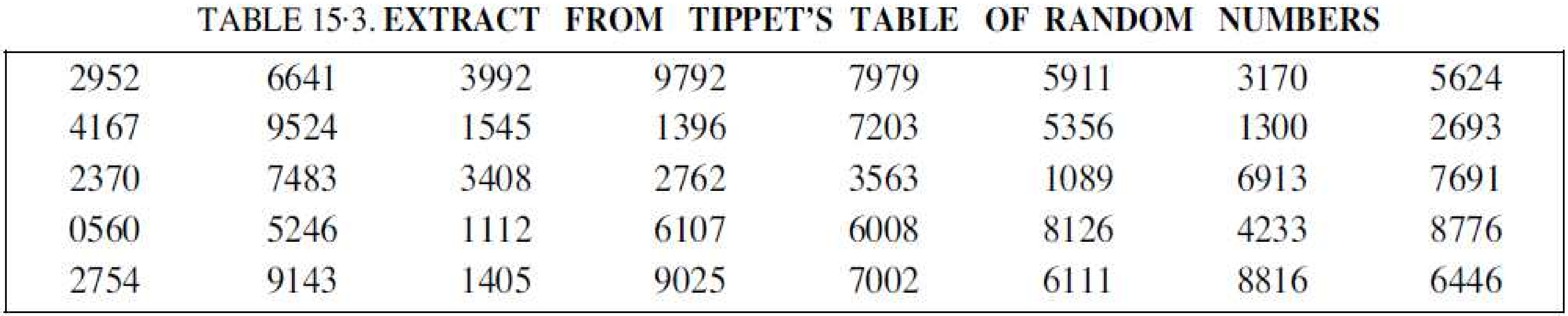


TABLE 15·3. EXTRACT FROM TIPPET'S TABLE OF RANDOM NUMBERS

2952	6641	3992	9792	7979	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2370	7483	3408	2762	3563	1089	6913	7691
0560	5246	1112	6107	6008	8126	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

11,24,15,13,03

Example

- It may even happen that extract given from the table of random numbers is so small that we are not able to draw a random sample of the desired size.
- This difficulty can be overcome by assigning more than one number to each of the sampling units.
- For previous ex, the first unit may be assigned the numbers :

1 , 1 + 24, 1 + 2 × 24, 1 + 3 × 24, and so on

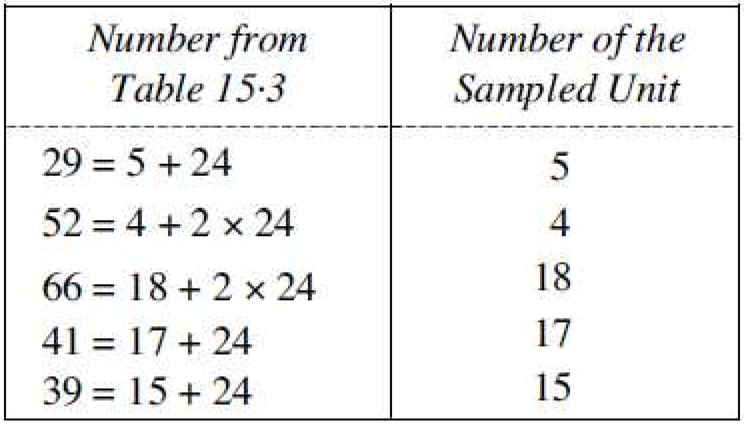
1, 25, 49, 73, 97, 121, … and so on.

Similarly the second unit may be assigned the numbers :

2, 26, 50, 74, 98, 122, … and so on.

Finally, the last unit may be assigned the

0, 24, 48, 72, 96, 120, …



Number from Table 15·3	*Number of the Sampled Unit*
$29 = 5 + 24$	5
$52 = 4 + 2 \times 24$	4
$66 = 18 + 2 \times 24$	18
$41 = 17 + 24$	17
$39 = 15 + 24$	15

Probability of Selection of a Unit

Let the size of the population is *N*.

One out of *N* sampling unit is to be chosen.

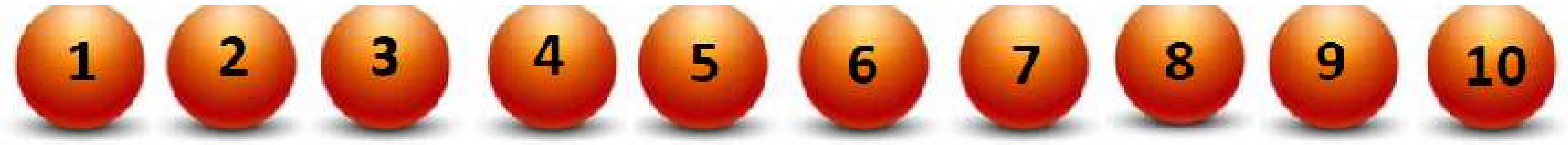
SRSWOR

The probability of drawing a sampling unit = $\mathbf{1/N}$

SRSWR

The probability of drawing a sampling unit = $\mathbf{1/N}$

Probability of Selection of a Unit



Probability of drawing ball 1= 1/10

Probability of drawing ball 2= 1/10

...

Probability of drawing ball 10= 1/10

Probability of Selection of a Sample SRSWOR

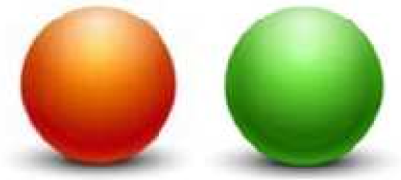
- Total number of combinations to choose n sampling units out of N
- sampling unit = $\boldsymbol{N}\mathrm{C}_{\boldsymbol{n}}$
- The probability of drawing a sample = $\mathbf{1}/\ \boldsymbol{N}\mathrm{C}_{\boldsymbol{n}}$

Probability of Selection of a Sample SRSWOR

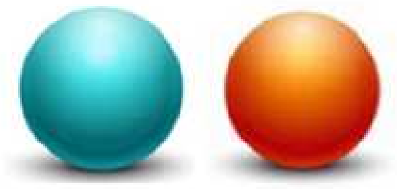
Suppose N =3, n = 2

Total samples = $\binom{3}{2} = 3$

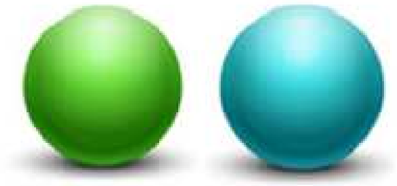
Sample 1



Sample 2



Sample 3



Probability of Selection of a Sample SRSWR

SRSWR

Total number of combinations to choose n sampling units out of N

sampling unit = N^n

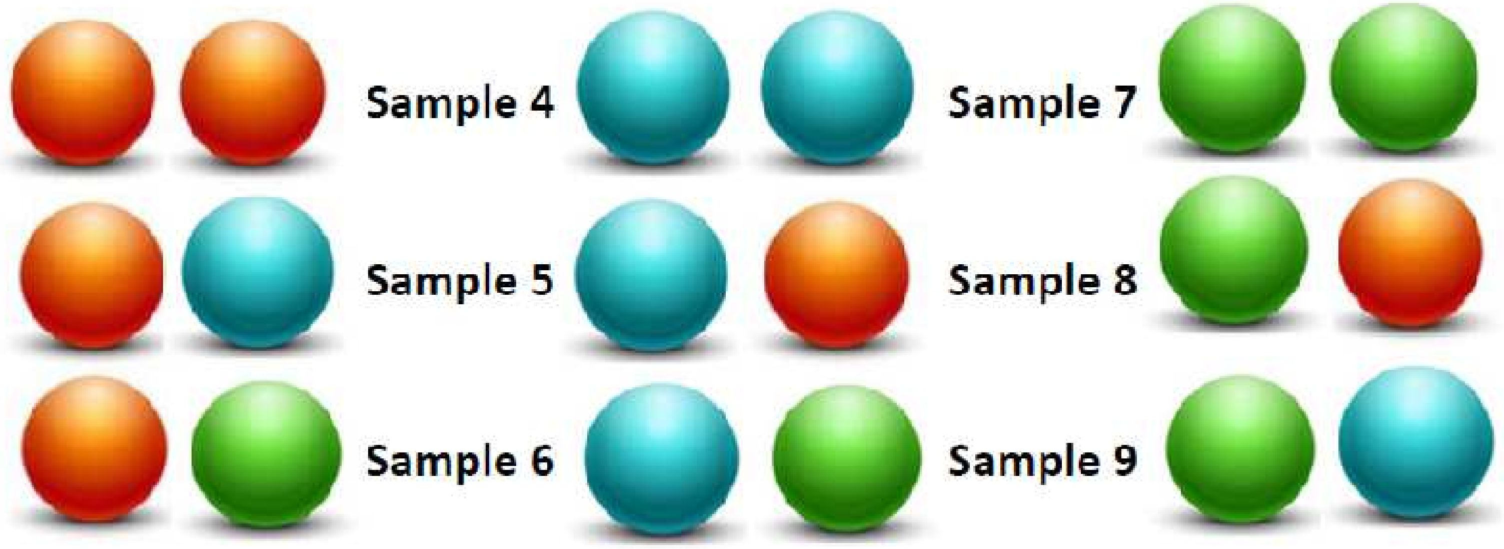
The probability of drawing a sample = $\mathbf{1}/N^n$

Probability of Selection of a Sample SRSWR

SRSWR

Suppose N =3,

Total samples N=3, n=2, $N^n = 3^2 = 9$



Sample 1

Sample 4

Sample 7

Sample 2

Sample 5

Sample 8

Sample 3

Sample 6

Sample 9

Probability of drawing a sample = $\frac{1}{9}$

Drawing of sample using R

```
> heightdata
name height
1 A 151
2 B 152
3 C 153
4 D 154
5 E 155
6 F 156
7 G 157
8 H 158
9 I 159
10 J 160
> names=heightdata$name
> names
[1] A B C D E F G H I J
Levels: A B C D E F G H I J
> heights=heightdata$height
> heights
[1] 151 152 153 154 155 156 157
158 159 160
```

Drawing of sample using R

```
sample(names, size=5, replace = FALSE)
> sample(names, size=5, replace = FALSE)
[1] G F A B H
Levels: A B C D E F G H I J
```

Suppose we want this sample in terms of heights of persons.

```
sample(heights, size=5, replace = FALSE)
> sample(heights, size=5, replace = FALSE)
[1] 152 156 154 155 158
```

```
> sample(names, size=5, replace = TRUE)
[1] F F I E A
Levels: A B C D E F G H I J
```

Stratified Random sampling

- It involves a method ***where a larger population can be divided into smaller groups***, that usually don't overlap but represent the entire population together. While sampling these groups can be organized and then draw a sample from each group separately. A common method is to arrange or classify by sex, age, ethnicity and similar ways.

Stratified Random Sampling

Stratified Random sampling

- Where population embraces a number of distinct categories, the frame can be organized into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected.
- Every unit in a stratum has same chance of being selected.
- Using same sampling fraction for all strata ensures proportionate representation in the sample.
- Adequate representation of minority subgroups of interest can be ensured by stratification & varying sampling fraction between strata as required.

Stratified Random sampling

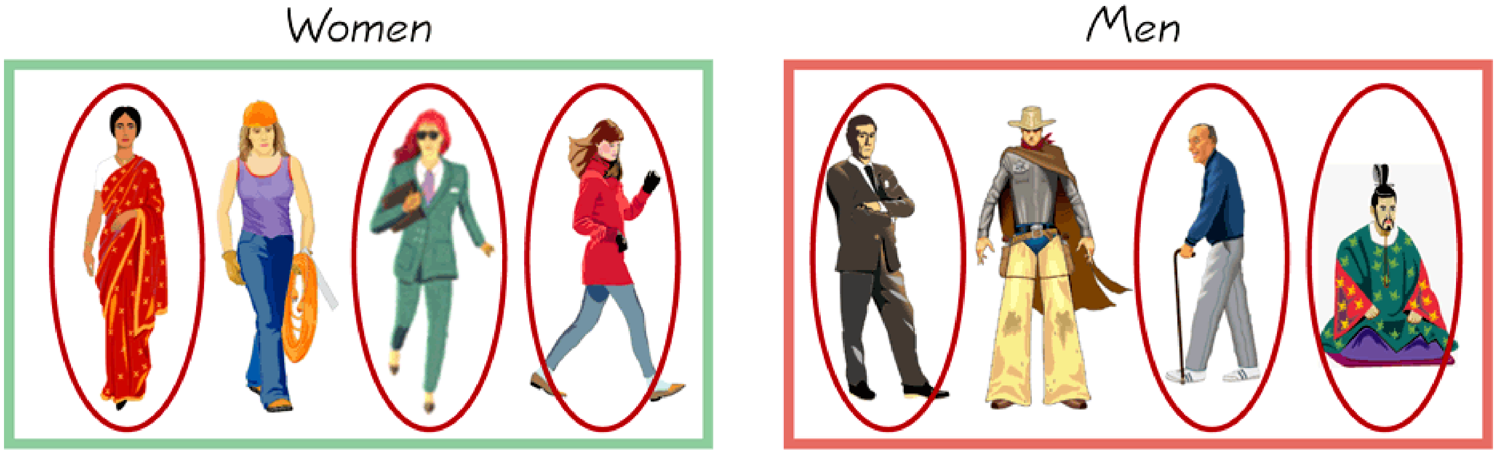
- Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata.

Drawbacks to using stratified sampling.

- Sampling frame of entire population has to be prepared separately for each stratum
- When examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata.
- In some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods

STRATIFIED SAMPLING……

Draw a sample from each stratum

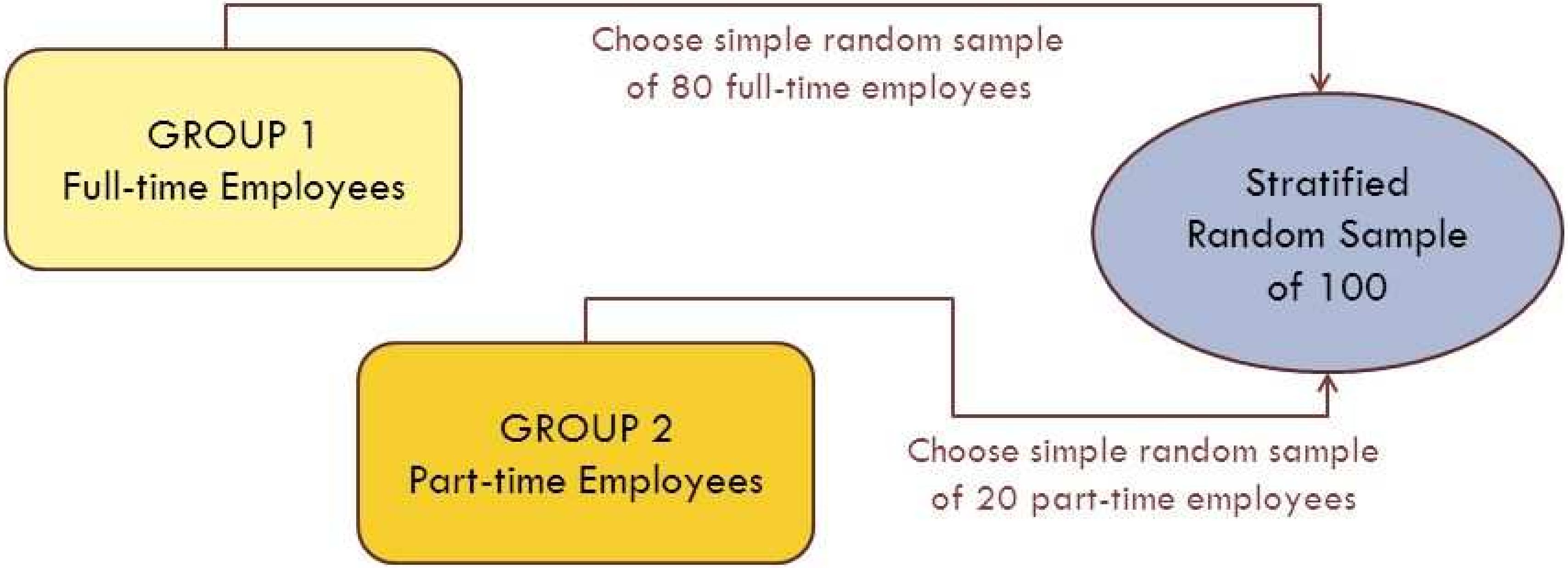


Women

Men

Example – Stratified Random Sampling

A company has 800 full-time and 200 part-time employees. To draw a sample of 100 employees, a simple random sample of 80 full-time employees is selected and a simple random sample of 20 part-time employees is selected.


GROUP 1
Full-time Employees
Choose simple random sample of 80 full-time employees
Stratified Random Sample of 100
GROUP 2
Part-time Employees
Choose simple random sample of 20 part-time employees

Stratified Random sampling

Let’s say, 100 (N_h) students of a school having 1000 (N) students were asked questions about their favorite subject.

It’s a fact that the students of the 8th grade will have different subject preferences than the students of the 9th grade.

For the survey to deliver precise results, the ideal manner is to divide each grade into various strata.

Grade	Number of students (n)
5	150
6	250
7	300
8	200
9	100

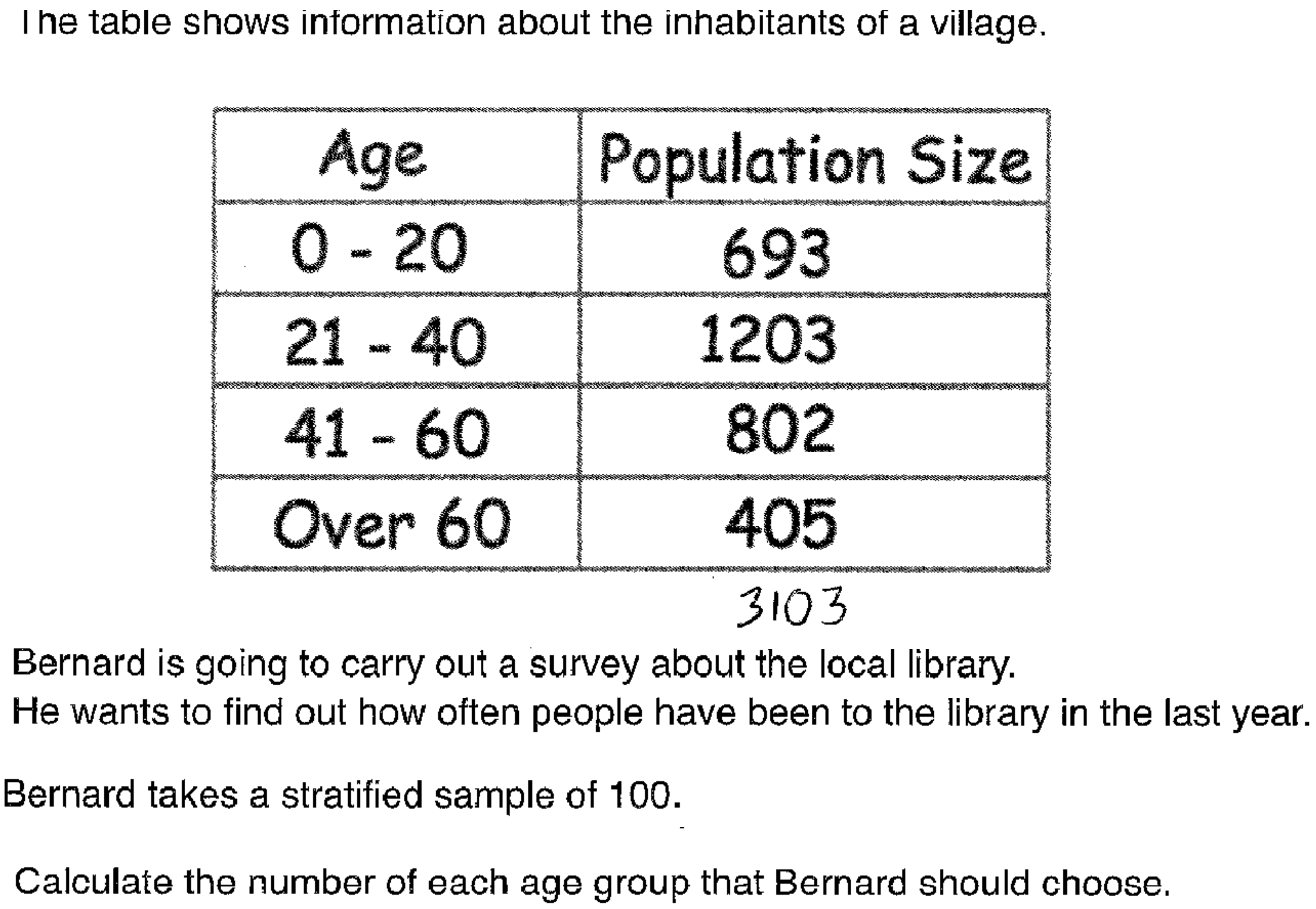
Stratified Random sampling

$$n_h = (N_h / N) * n$$

n_h= Sample size for hth stratum
N_h= Population size for hth stratum
N = Size of entire population
n = Size of entire sample

Stratified Sample (n_5) = 100 / 1000 * 150 = 15
Stratified Sample (n_6) = 100 / 1000 * 250 = 25
Stratified Sample (n_7) = 100 / 1000 * 300 = 30
Stratified Sample (n_8) = 100 / 1000 * 200 = 20
Stratified Sample (n_9) = 100 / 1000 * 100 = 10

Stratified Random sampling



I he table shows information about the inhabitants of a village.

Age	Population Size
0 - 20	693
21 - 40	1203
41 - 60	802
Over 60	405

3103

Bernard is going to carry out a survey about the local library.
He wants to find out how often people have been to the library in the last year.

Bernard takes a stratified sample of 100.

Calculate the number of each age group that Bernard should choose.

Stratified Random sampling

- 0-20 (693/3103)*100=22
- 21-40 (1203/3103)*100=39
- 41-60 (802/3103)*100=26
- 61-80 (405/3103)*100=13

Stratified Random sampling

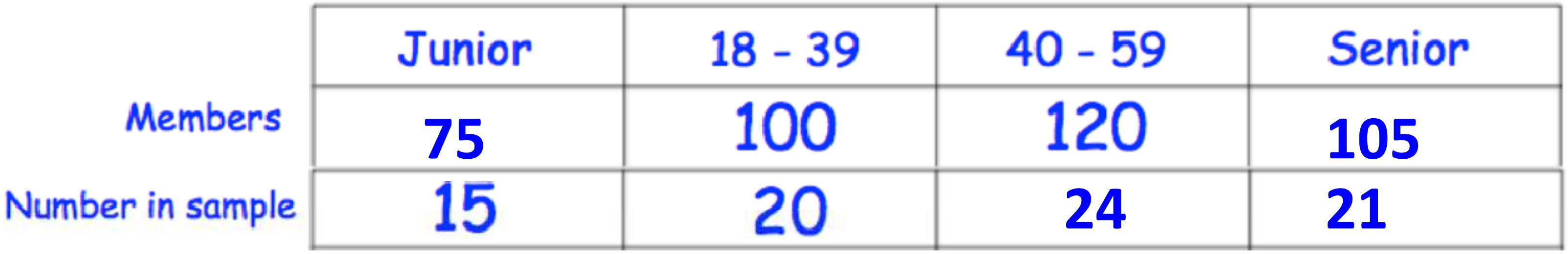
A cricket club has 400 members.
A stratified sample of member is taken, by age group.

The table shows some information.

	Junior	18 - 39	40 - 59	Senior
Members		100	120	
Number in sample	15	20		

Complete the table.

Stratified Random sampling



	Junior	18 - 39	40 - 59	Senior
Members	75	100	120	105
Number in sample	15	20	24	21

Cluster sampling

- It is a way to randomly select participants when they are geographically spread out. Cluster sampling usually analyzes a particular population in which the sample consists of more than a few elements, for example, city, family, university etc. The clusters are then selected by dividing the greater population into various smaller sections.

Stratified & Cluster Sampling

Stratified

- Population divided into few subgroups
  - Each subgroup has many elements in it.
  - Subgroups are selected according to some criterion that is related to the variables under study.
- Homogeneity within subgroups
- Heterogeneity between subgroups
- Choice of elements from within each subgroup

Cluster

- Population divided into many subgroups
  - Each subgroup few elements in it.
  - Subgroups are selected according to some criterion of ease or availability in data collection.
- Heterogeneity within subgroups
- Homogeneity between subgroups
- Random choice of subgroups

Cluster sampling

- Cluster sampling is an example of 'two-stage sampling' .
- First stage a sample of areas is chosen;
- Second stage a sample of respondents *within* those areas is selected.
- Population divided into clusters of homogeneous units, usually based on geographical contiguity.
- Sampling units are groups rather than individuals.
- A sample of such clusters is then selected.
- All units from the selected clusters are studied.

Cluster sampling

Advantages :

- Cuts down on the cost of preparing a sampling frame.
- This can reduce travel and other administrative costs.

Disadvantages

- Sampling error is higher for a simple random sample of same size.
- Often used to evaluate vaccination coverage in EPI

Cluster sampling

- **Identification of clusters**
  - List all cities, towns, villages & wards of cities with their population falling in target area under study.
  - Calculate cumulative population & divide by 30, this gives sampling interval.
  - Select a random no. less than or equal to sampling interval having same no. of digits. This forms 1st cluster.
  - Random no.+ sampling interval = population of 2nd cluster.
  - Second cluster + sampling interval = 4th cluster.
  - Last or 30th cluster = 29th cluster + sampling interval

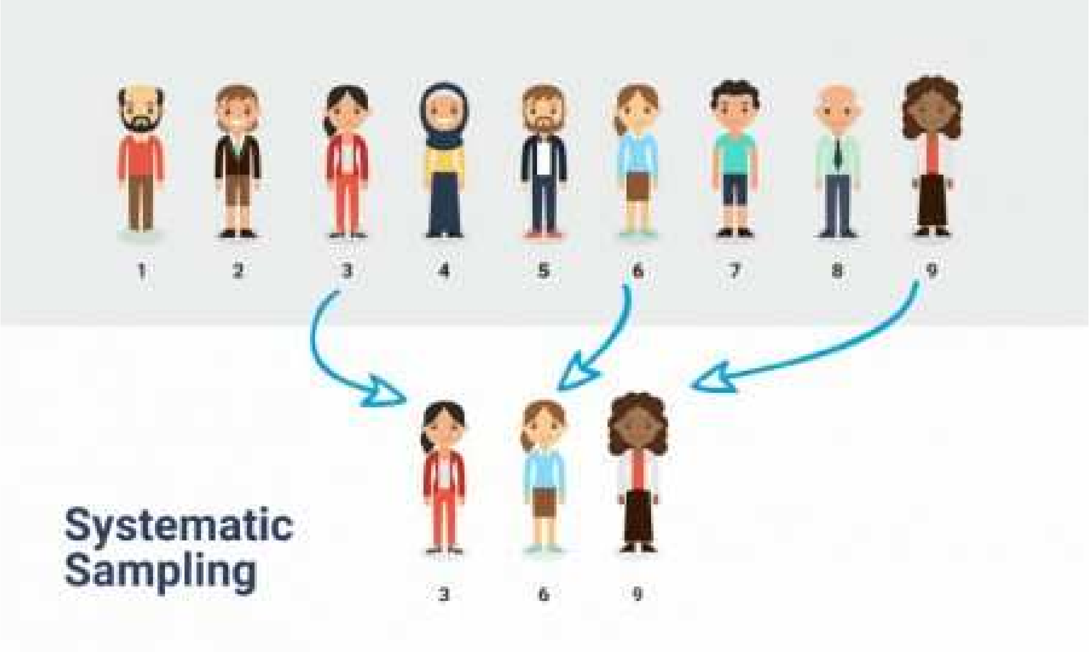
Cluster sampling

	Freq	c f	cluster
I	2000	2000	1
II	3000	5000	2
III	1500	6500	
IV	4000	10500	3
V	5000	15500	4, 5
VI	2500	18000	6
VII	2000	20000	7
VIII	3000	23000	8
IX	3500	26500	9
X	4500	31000	10
XI	4000	35000	11, 12
XII	4000	39000	13
XIII	3500	44000	14,15
XIV	2000	46000	
XV	3000	49000	16
XVI	3500	52500	17
XVII	4000	56500	18,19
XVIII	4500	61000	20
XIX	4000	65000	21,22
XX	4000	69000	23
XXI	2000	71000	24
XXII	2000	73000	
XXIII	3000	76000	25
XXIV	3000	79000	26
XXV	5000	84000	27,28
XXVI	2000	86000	29
XXVII	1000	87000	
XXVIII	1000	88000	
XXIX	1000	89000	30
XXX	1000	90000	

90000/30 = 3000 sampling interval

Systematic Sampling

- It is when you choose every "**nth" individual to be a part of the sample**. For example, you can choose every 5th person to be in the sample. Systematic sampling is an extended implementation of the same old probability technique in which each member of the group is selected at regular periods to form a sample. There's an equal opportunity for every member of a population to be selected using this sampling technique.



Systematic Sampling

- Systematic sampling relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list.
- Systematic sampling involves a random start and then proceeds with the selection of every *k*th element from then onwards. In this case, *k*=(population size/sample size).
- It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the *k*th element in the list.
- A simple example would be to select every 10th name from the telephone directory (an 'every 10th' sample, also referred to as 'sampling with a skip of 10').

Systematic Sampling

ADVANTAGES:

- Sample easy to select
- Suitable sampling frame can be identified easily
- Sample evenly spread over entire reference population

DISADVANTAGES:

- Sample may be biased if hidden periodicity in population coincides with that of selection.
- Difficult to assess precision of estimate from one survey.

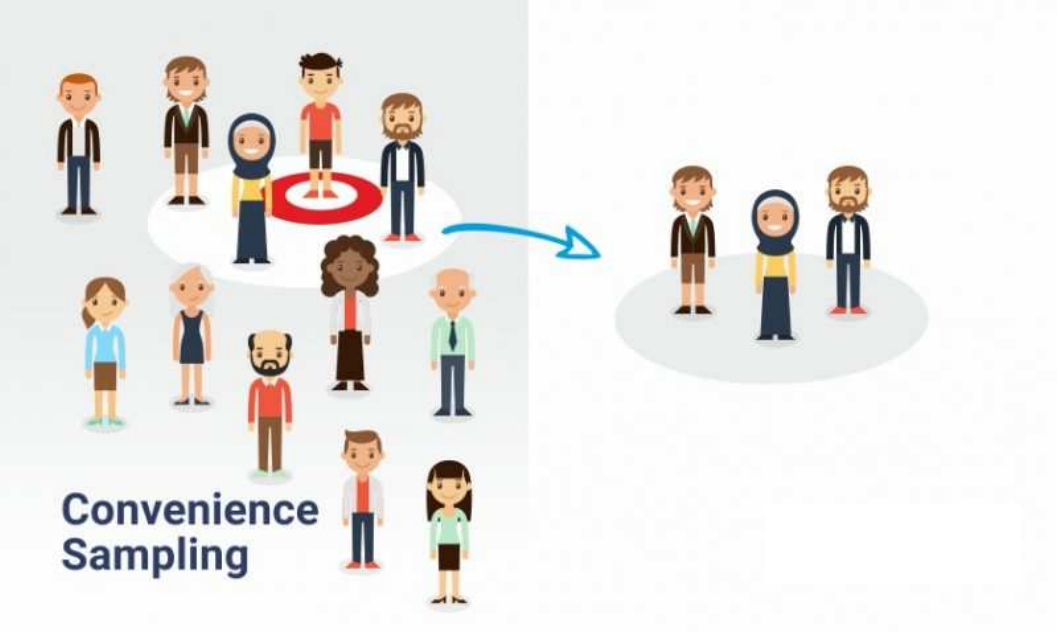
Types of Non-probability Sampling

Four main techniques used for a non-probability sample:

- ➢ Convenience
- ➢ Judgemental
- ➢ Snowball
- ➢ Quota

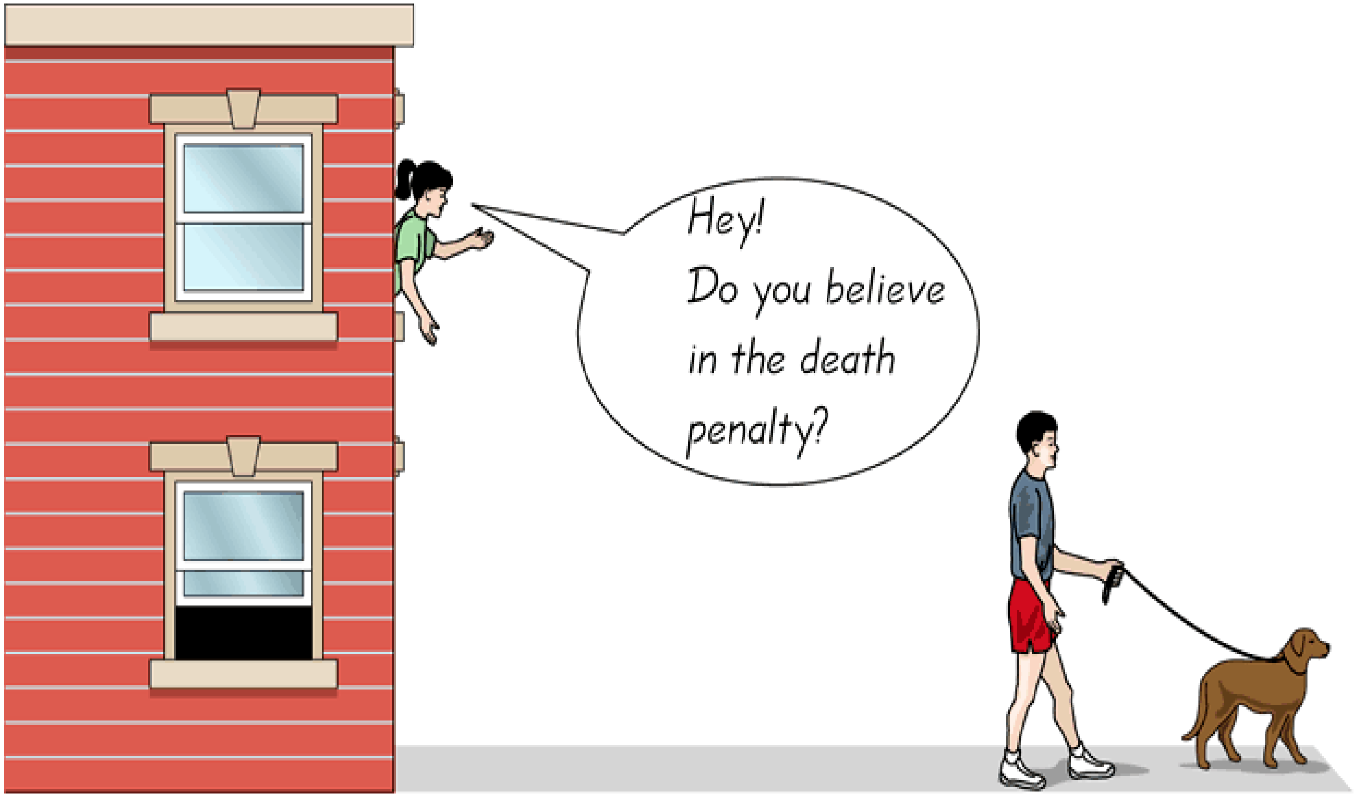
Convenience Sampling

- It is a non-probability sampling technique used to create sample as per ease of access, readiness to be a part of the sample, availability at a given time slot or any other practical specifications of a particular element.
- Convenience sampling involves selecting haphazardly those cases that are easiest to obtain for your sample, such as the person interviewed at random in a shopping center for a television program.



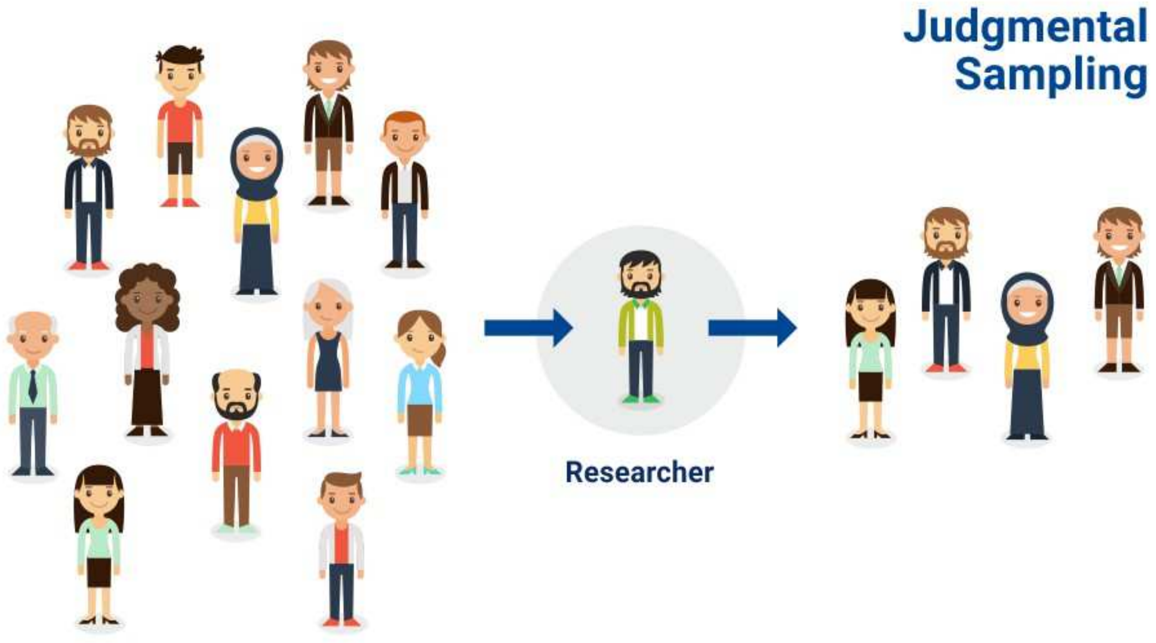
CONVENIENCE SAMPLING……

- Use results that are easy to get



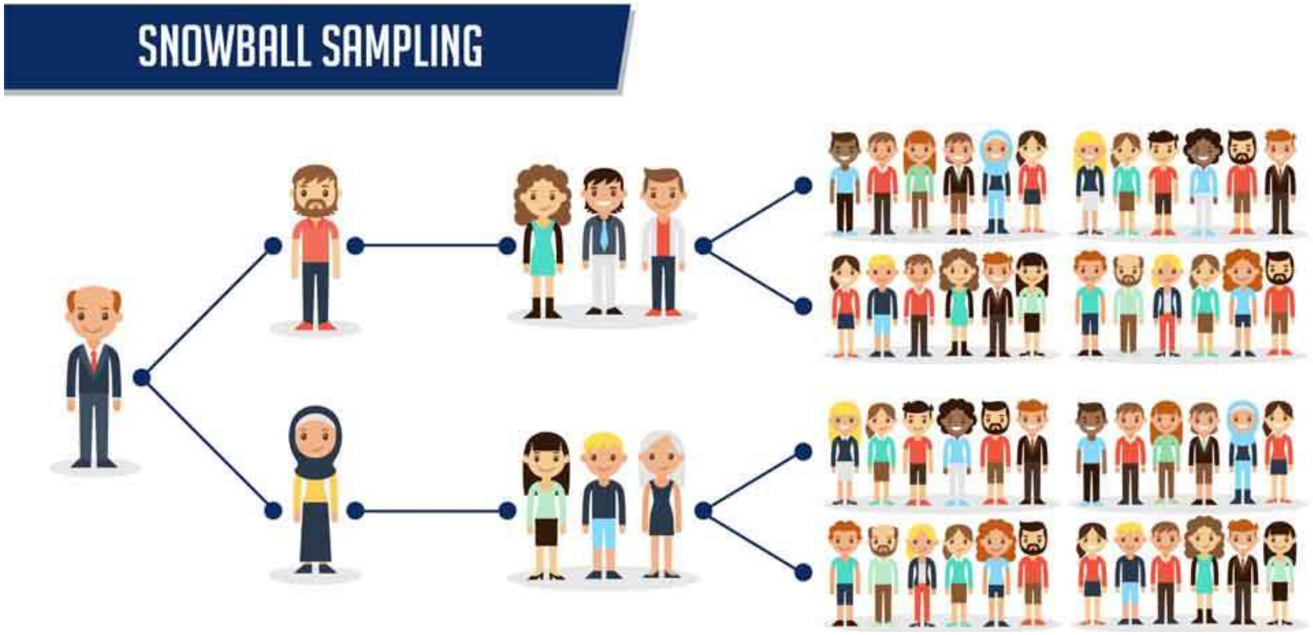
Judgmental Sampling

- In the judgmental sampling, also called purposive sampling, the sample members are chosen only on the basis of the researcher's knowledge and judgment.
- It enables you to select cases that will best enable you to answer your research question(s) and to meet your objectives.



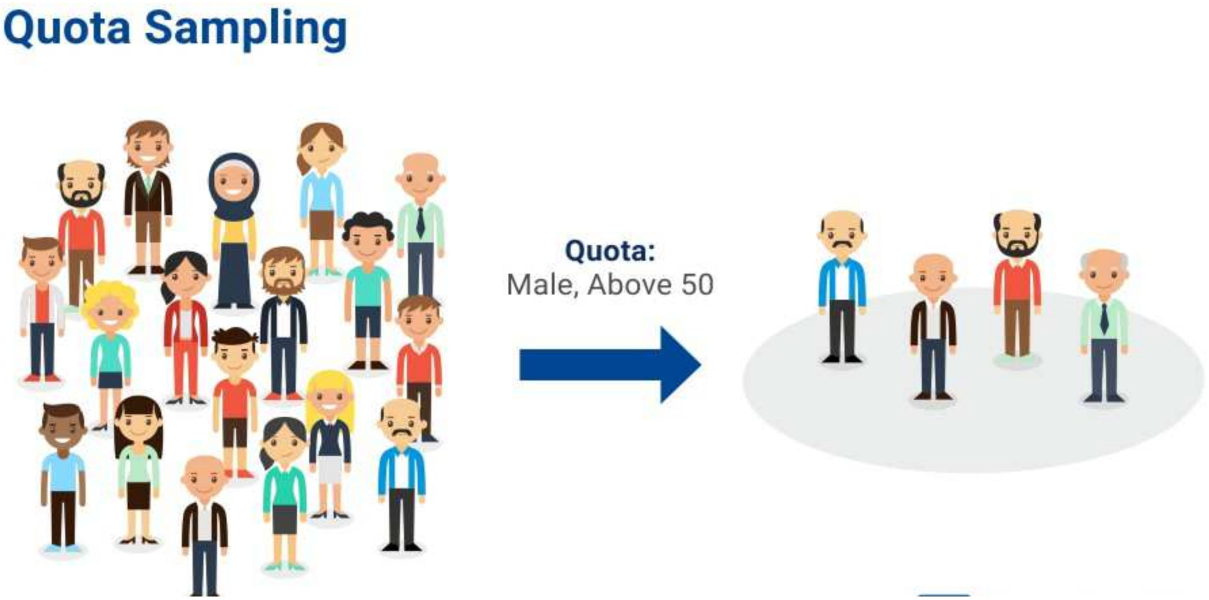
Snowball Sampling

- Snowball sampling method is purely based on referrals and that is how a researcher is able to generate a sample. Therefore this method is also called the chain-referral sampling method.
- This sampling technique can go on and on, just like a snowball increasing in size (in this case the sample size) till the time a researcher has enough data to analyze, to draw conclusive results that can help an organization make informed decisions.



Quota Sampling

- Selection of members in this sampling technique happens on basis of a pre-set standard. In this case, as a sample is formed on basis of specific attributes, the created sample will have the same attributes that are found in the total population. It is an extremely quick method of collecting samples.
- Quota sampling is therefore a type of stratified sample in which selection of cases within strata is entirely non-random.



Estimation of Population Mean and Population Variance

- One of the main objectives after the selection of a sample is to know about the tendency of the data to cluster around the central value and the scatterdness of the data around the central value in the population.
- Popular choices are arithmetic mean and variance.
- Population mean is generally measured by arithmetic mean (or weighted arithmetic mean)

Estimation of Population Mean and Variance: Notations

$Y_1, Y_2, \ldots, Y_N$: **Population**

$y_1, y_2, \ldots, y_n$: **Sample**

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$: **Population mean**

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$: **Sample mean**

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right)$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N}\left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right)$$

Population Variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)$$

Sample Variance

Estimation of Population Mean

Let us consider the sample arithmetic mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ as an estimator of population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$

Estimate population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ by sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

$\bar{y}$ is an unbiased estimator of $\bar{Y}$ under SRSWR and SRSWOR cases.

$$E(\bar{y}) = \frac{1}{N}\sum_{i-1}^{N} y_i = \bar{Y}.$$

Estimation of Population Mean

Population: $X1 = 1$, $X2 = 3$, $X3 = 5$

Population mean = 3

Number of Samples of size 2= $\mathbf{3}C_2$ n!/(n-r)!*r!

=3

Suppose the population mean is unknown.

Use sample arithmetic mean to estimate the population mean.

Estimation of Population Mean

Sample arithmetic mean is an unbiased estimator of population mean.

Sample 1=(1,3) Sample mean $\overline{x1}$ = 2

Sample 2=(3,5) Sample mean $\overline{x2}$ = 4

Sample 3=(1,5) Sample mean $\overline{x3}$ = 3

$$\overline{x} = \frac{\overline{x}_1 + \overline{x}_2 + \overline{x}_3}{3} = \frac{2+4+3}{3} = 3 = \text{Population mean}$$

- $\bar{y}$ is an unbiased estimator of $\bar{Y}$ in SRSWOR

$$E(\bar{y}) = \frac{1}{n}\sum_{j=1}^{n} E(y_j)$$

- $\bar{y}$ is an unbiased estimator of $\bar{Y}$ in SRSWR

$$E(\bar{y}) = \frac{1}{n} E\left(\sum_{i=1}^{n} y_i\right)$$

Sample Mean Example

Y: Height of students in a class
N = 10 : Number of students in the class (Population size)
n = 3 : Number of students in the sample (Sample size)

Name of Student	Y_i = Height of students (in Centimeters)
A	Y_1= 151
B	Y_2= 152
C	Y_3 = 153
D	Y_4= 154
E	Y_5 = 155
F	Y_6= 156
G	Y_7 = 157
H	Y_8= 158
I	Y_9 = 159
J	Y_{10}= 160

Sample 1: 3rd , 7th and 9th student
$y_1 = Y_3$ = 153 cms., $y_2 = Y_7$ = 157 cms., $y_3 = Y_9$ = 159 cms
Sample mean 1($\bar{y}_1$) = (153 + 157 + 159)/3= 156.33 cms

Sample 2: 2nd , 5th and 4th student
$y_1 = Y_2$ = 152 cms., $y_2 = Y_5$ = 155 cms., $y_3 = Y_4$ = 154 cms
Sample mean 2 ($\bar{y}_2$) = (152 + 155 + 154)/3= 153.66 cms

Sample 3: 1st , 6th and 10th student
$y_1 = Y_1$ = 151 cms., $y_2 = Y_6$ = 156 cms., $y_3 = Y_{10}$ = 160 cms
Sample mean 3 ($\bar{y}_3$) = (151 + 156 + 160)/3= 155.66 cms

Population mean =$\bar{Y}$ 155.5

The total number of samples = $\mathbf{10C_3=120}$

Variance of Sample Mean

- Population variability is generally measured by variance.
- Several sample can be drawn by SRSWR as well as SRSWOR from a population.
- Each sample will have different sample mean.
- Sample mean is a statistic, i.e., a function of random variables.
- So sample mean will also have variance.

Variance of Sample Mean

Variance of sample mean under SRSWOR

$$V(\bar{y}_{WOR}) = E(\bar{y}-\bar{Y})^2 = \frac{N-n}{Nn}S^2$$

var($\bar{Y}$)=(N-n)/(N-1). (σ^2/n)

Variance of sample mean under SRSWR

$$V(\bar{y}_{WR}) = E(\bar{y}-\bar{Y})^2 = \frac{N-1}{Nn}S^2$$

var($\bar{Y}$)=σ^2/n

Sampling and Non-Sampling Errors

- In a sample survey, since only a small portion of the population is studied, its results are bound to differ from the census results and thus have a certain amount of error.
- This error would always be there, no matter that the sample is drawn at random and that it is highly representative.
- This error is attributed to fluctuations of sampling and is called *sampling error.*
- *Sampling error is due to the fact that* only a subset of the population (*i.e., sample) has been used to estimate the population parameters and draw* inferences about the population.
- Thus, sampling error is present only in a sample survey and is completely absent in census method.

Sampling Errors

Reasons :

1. *Faulty selection of the sample*

   Purposive or judgment sampling, Random sampling can be used

2. *Substitution*

   Bias

3. *Faulty demarcation of sampling units*

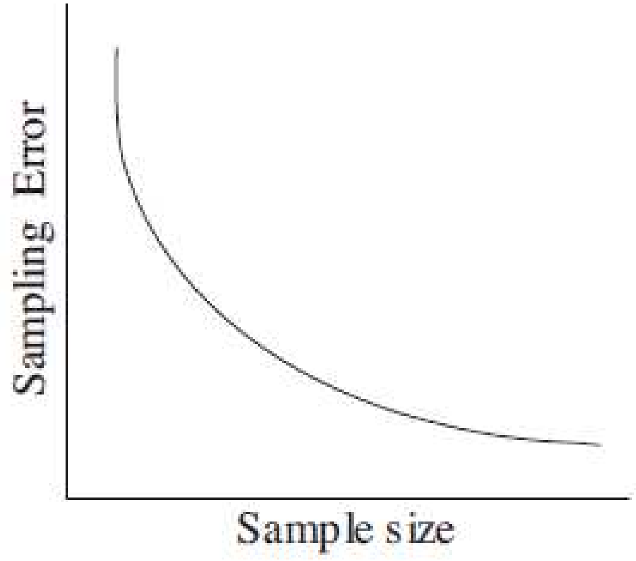
   Eg. Crop cutting surveys, border line cases

4. *Error due to bias in the estimation method*
5. *Variability of the population*

   Population and sample variance

Sampling Errors

- Sampling errors follow random or chance variations and tend to cancel out each other on averaging.
- A measure of the sampling error is provided by the standard error of the estimate.
- Standard error of the estimate is inversely proportional to the square root of the sample size



Non Sampling Errors

- Non-sampling errors are not attributed to chance and are a consequence of certain factors which are within human control.
- These errors can be traced and may arise at any stage of the enquiry, *viz., planning and execution of the survey and collection,* processing and analysis of the data.
- Non-sampling errors are thus present both in census surveys as well as sample surveys
- Large magnitude in a census survey than in a sample survey

Non Sampling Errors

1. Faulty planning, including vague and faulty definitions of the population or the statistical units to be used, incomplete list of population-members

2. Vague and imperfect questionnaire which might result in incomplete or wrong information.

3. Defective methods of interviewing and asking questions.

4. Vagueness about the type of the data to be collected.

5. Exaggerated or wrong answers to the questions which appeal to the pride or prestige or self-interest

   of the respondents.

6. Personal bias of the investigator.

7. Lack of trained and qualified investigators and lack of supervisory staff.

8. Failure of respondents' memory to recall the events or happenings in the past.

9. *Non-response and Inadequate or Incomplete Response.*

10. *Improper coverage.*

Non Sampling Errors

- In a census, sampling error is completely absent so the total error is non-sampling error.
- A sample survey, on the other hand, contains both sampling and non-sampling errors.
- In a sample survey non-sampling errors can be controlled by :

(i) Employing qualified and trained personnel

(ii) Using more sophisticated statistical techniques and equipment

(iii) Providing adequate supervisory checks on the field work

(*iv) Pre-testing or conducting a pilot survey*

(*v) Through editing and scrutiny of the results*

(*vi) Effective checking of all the steps in the processing and analysis of the data*

(*vii) More effective follow up of non-response cases*

(*viii) Imparting thorough training to the investigators for efficient conduct of the enquiry*

Biased and Unbiased Errors

Biased Errors

(*i) Bias on the part of the enumerator or investigator whose personal beliefs and prejudices are likely to* affect the results of the enquiry

(*ii) Bias in the measuring instrument or the equipment used for recording the observations.*

(*iii) Bias due to faulty collection of the data, and in the statistical techniques and the formulae used for* the analysis of the data.

(*iv) Respondents' bias.* wrong answers with or without purpose

(*v) Bias due to Non-response*

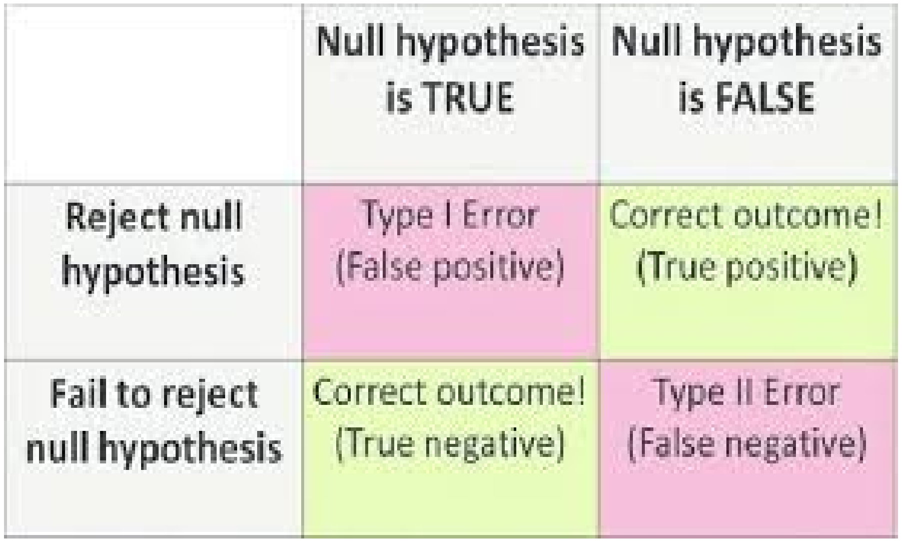
(*vi) Bias in the Technique of Approximations rounding off*

Biased and Unbiased Errors

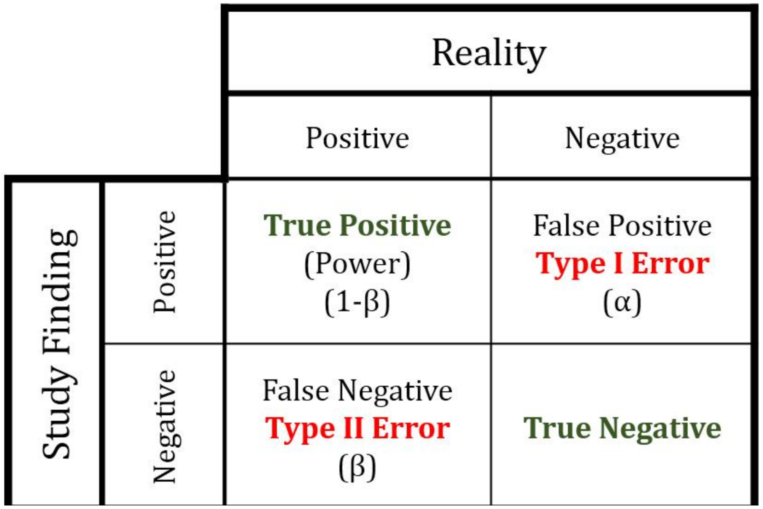
Unbiased Errors

- If the estimated or approximated values are likely to err on either side, *i.e., if the chances of making an over-estimate is almost same as the chance* of making an under-estimate.
- 385, 415, 355, 445 rounded to the *nearest complete unit i.e., hundred*
- Since these errors move in both the directions, the errors in one direction are more or less neutralised by the errors in the opposite direction and consequently the ultimate result is not much affected.
- *Compensatory Errors*

Type I, Type II Errors



	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)



		Reality	
		Positive	Negative
Study Finding	Positive	**True Positive** (Power) (1-β)	False Positive **Type I Error** (α)
	Negative	False Negative **Type II Error** (β)	**True Negative**

Example

For the following population, consider all SRSWOR samples of size 3

i	1	2	3	4	5
y_i	5	8	3	11	9

1. Show that $\bar{y}$ is unbiased estimator of $\bar{Y}$
2. Show that s^2 is unbiased estimator of S^2
3. Calculate sampling variation $\bar{y}$ and show that it agrees with the formula

   $(N\text{-}n/nN)\ S^2$
4. Verify that $\mathrm{Var}_{srswr}(\bar{y})>\mathrm{Var}_{srswor}(\bar{y})$

N=5

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

(Population Mean)

$\bar{Y}$=(5+8+3+11+9)=7.2

(Population Mean Square)

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$$

Y_i	1	2	3	4	5	
$(y_i-\bar{Y})^2$	$(5-7.2)^2$	$(8-7.2)^2$	$(3-7.2)^2$	$(11-7.2)^2$	$(9-7.2)^2$	40.80

S^2=1/4*40.80=10.2

From a population of 5 a sample of size 3 can be drawn in $5C_3$ ways= 10 ways

No	Values	Mean	S^2	$(y_i-\bar{Y})$	$(y_i-\bar{Y})^2$
1	1,2,3	16/3	19/3	-1.87	3.48
2					
3					
4					
5					
6					
7					
8					
9					
10					

mean=(5+8+3)=16/3

Y_i	1	2	3			
$(y_i-\bar{Y})^2$	$(5-16/3)^2$	$(8-16/3)^2$	$(3-16/3)^2$			12.66

S^2=1/2*12.66=6.33=19/3

(16/3-7.2)=-1.87

No	Values	Mean	S^2	$(y_i-\bar{Y})$	$(y_i-\bar{Y})^2$
1	1,2,3	16/3	19/3	-1.87	3.48
2					
3					
4					
5					
6					
7					
8					
9					
10					
		216/3	306/3		13.60

$$E(\bar{y}) = \frac{1}{NC_n}\sum_{i=1}^{N} \bar{y}_i \quad = 1/10 * 216/3 = 7.2 = \bar{Y}$$

$$E(s^2) = \frac{1}{NC_n}\sum_{i=1}^{N} s_i^2 \quad = 1/10 * 306/3 = 10.2 = S^2$$

In SRSWOR

Var(ȳ)=$\frac{1}{NC_n}\sum_{i=1}^{N}(y_i-\bar{Y})^2$ =1/10*13.6=1.36

Var(ȳ)=$\frac{N-n}{Nn}S^2$ =(5-3/5*3) *10.2=(2/15)*10.2=1.36

In SRSWR

Var(ȳ)=$\frac{N-1}{Nn}S^2$ =(5-1/5*3) *10.2=(4/15)*10.2=2.72

$\text{Var}_{srswr}(\bar{y})>\text{Var}_{srswor}(\bar{y})$

- Explain why a random sample of size 25 is to be preferred to a random sample of size 20 to estimate population mean.

$\mathrm{var}(\bar{y})=\sigma^2/n$

$\mathrm{S.E.}(\bar{y})=\sigma/\mathrm{sqrt}(n)$

Larger the sample smaller is the error