

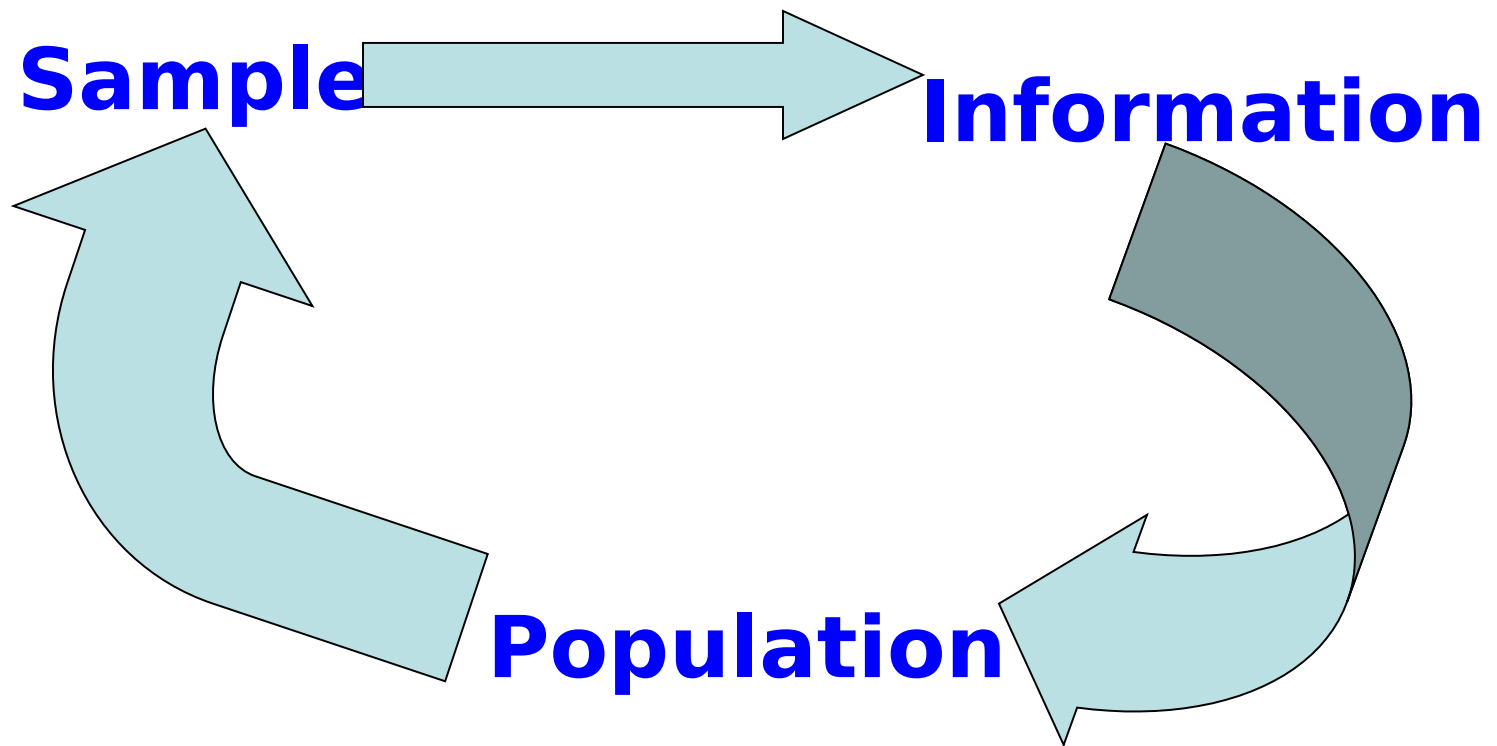
Sampling Methods

- Researchers often use sample survey methodology to obtain information about a larger population by selecting and measuring *a sample from that population*.
- Since population is too large, we rely on the information collected from the sample.
- Cost minimization.

- **Inferences about the population are based on the information from the sample drawn from that population.**
- However, due to the variability in the characteristics of the population, scientific sample designs should be applied to select a *representative sample*.
- If not, there is a high risk of *distorting* the view of the population.

- A sample is a collection of individuals selected from a larger population.
- For example, we may have a single sample composed of 50 individuals, representing a population of 1000 people.

- **Sampling** enables us to estimate the characteristic of a population by directly observing a portion of the population.
- Researchers are not interested in the **sample itself**, but in what can be learned from the **sample**—and how this information can be applied to the **entire population**.



- Therefore, it is essential that a sample should be correctly defined and organized.
- If the *wrong questions* are posed to the *wrong people*, reliable information will not be received and lead to a *wrong conclusion* when applied to the entire population.

Steps needed to select a sample and ensure that this sample will fulfill its goals.

1. Establish the study's objectives

- The first step in planning a useful and efficient survey is to specify the objectives with as much detail as possible.
- Clarifying the aims of the survey is critical to its ultimate success.
- Without objectives, the survey is unlikely to generate valuable results.
- The initial users and uses of the data should be identified at this stage.

2. Define the target population

- The *target population* is the total population for which the information is required.
- Specifically, the target population is defined by the following characteristics:
 - Nature of data required
 - Geographic location
 - Reference period
 - Other characteristics, such as socio-demographic characteristics

3. Decide on the data to be collected

- The data requirements of the survey must be established.
- To ensure that the requirements are operationally sound, the necessary data terms and definitions also need to be determined.

4. Set the level of precision

- There is a level of uncertainty associated with estimates coming from a sample.
- Researchers can estimate the sampling error associated with a particular sampling plan, and try to minimize it.
- Sample-to-sample variation causes sampling error

↑ Sample size ≡ ↑ Precision ≡ ↑ Cost

- Acceptable precision is important

5. Decide on the methods of measurement

- Choose measuring instrument and method of approach to the population
- Data about a person's state of health may be obtained from statements that he/she makes or from a medical examination
- The survey may employ a self-administered questionnaire or an interview

6. Preparing Frame

- List of all members of the population from which the sample will be taken
- The elements must not overlap

The sample design _

- **Sample design**: how the sample will be collected.
- **Estimation techniques**: how the results from the sample will be extended to the whole population.
- **Measures of precision**: how the sampling error will be measured.

Other Considerations

- Sample size determination
- Questionnaire development
- Pretest
- Organization of the field work
- Data collection
- Summary and analysis of the data
 - Edit the completed questionnaires
 - Decide on computation procedures

Sampling theory in public health

- **A health survey (sampling)** is a planned study to investigate the health characteristics of a population

A health survey is used to:

- Measure the total amount of illness in the population;
- Measure the amount of illness caused by a specified disease;
- Examine the utilization of existing health care facilities and demand for new ones;
- Measure the distribution of a particular characteristic, e.g.. breast-feeding practice in the population;
- Examine the role and relationship of one or more factors in the aetiology of a disease.

Sampling

- The process of selecting a portion of the population to represent the entire population.
- A main concern in sampling:
 - Ensure that the sample represents the population, and
 - The findings can be generalized.

Advantages of sampling:

- **Feasibility**: Sampling may be the only feasible method of collecting information.
- **Reduced cost**: Sampling reduces demands on resource such as finance, personnel, and material.
- **Greater accuracy**: Sampling may lead to better accuracy of collecting data
- **Sampling error**: Precise allowance can be made for sampling error
- **Greater speed**: Data can be collected and summarized more quickly

Disadvantages of sampling:

- There is always a *sampling error*.
- Sampling may create a feeling of discrimination within the population.

Errors in sampling

1) Sampling error: Errors introduced due to errors in the selection of a sample.

- They cannot be avoided or totally eliminated.

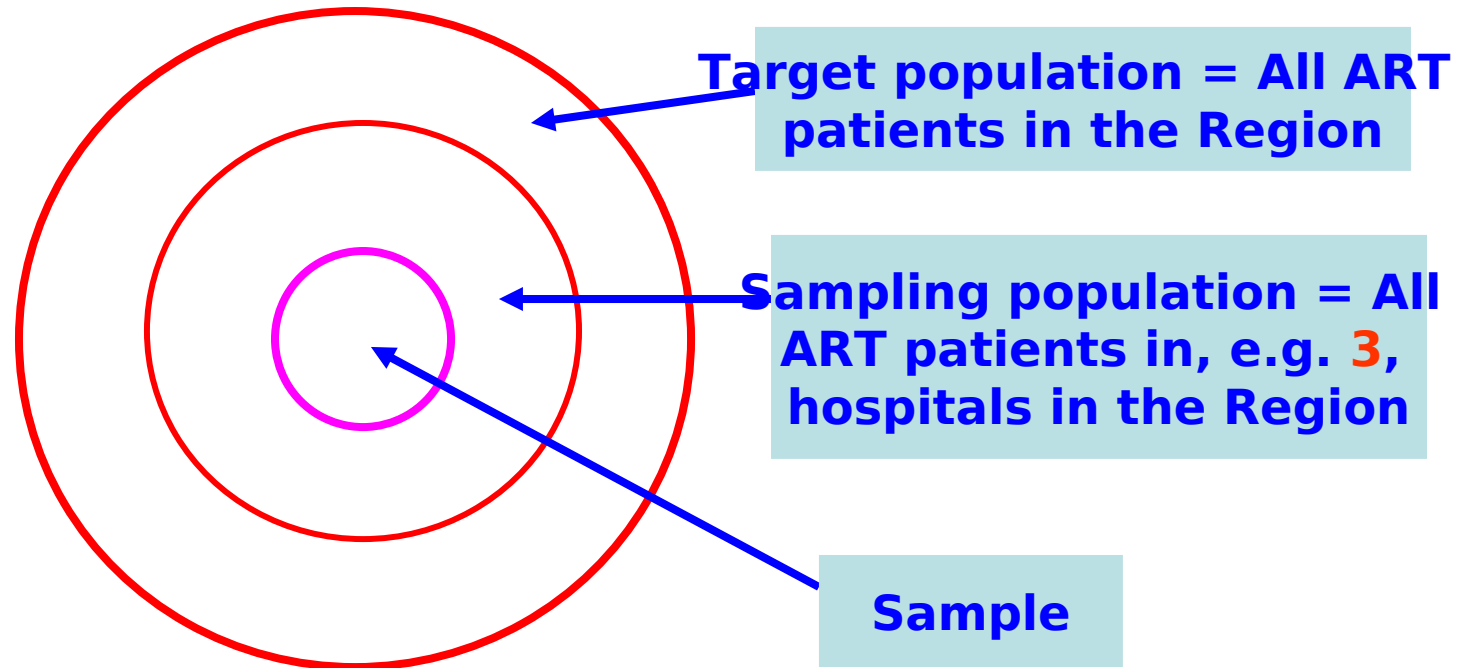
2) Non-sampling error:

- Observational error
- Respondent error
- Lack of preciseness of definition
- Errors in editing and tabulation of data

- While selecting a SAMPLE, there are basic questions:
 - What is the group of people (STUDY POPULATION) from which we want to draw a sample?
 - How many people do we need in our sample?
 - How will these people be selected?

- **Reference population (or target population):** the population of interest to whom the researchers would like to make generalizations.
- **Sampling population:** the subset of the target population from which a sample will be drawn.
- **Study population:** the actual group in which the study is conducted = **Sample**
- **Study unit:** the units on which information will be collected: persons, housing units, etc.

Researchers are interested to know about factors associated with ART use among HIV/AIDS patients attending certain hospitals in a given Region



Sampling Methods

Two broad divisions:

A. Probability sampling methods

**B. Non-probability sampling
methods**

A. Probability sampling

- Involves random selection of a sample
- Every sampling unit has a known and non-zero probability of selection into the sample.
- Involves the selection of a sample from a population, based on chance.

- **Probability sampling is:**
 - more complex,
 - more time-consuming and
 - usually more costly than non-probability sampling.
- **However, because study samples are randomly selected and their probability of inclusion can be calculated,**
 - reliable estimates can be produced and
 - inferences can be made about the population.

- There are several different ways in which a probability sample can be selected.
- The method chosen depends on a number of factors, such as
 - the available sampling frame,
 - how spread out the population is,
 - how costly it is to survey members of the population

Most common probability sampling methods

- 1. Simple random sampling**
- 2. Systematic random sampling**
- 3. Stratified random sampling**
- 4. Cluster sampling**
- 5. Multi-stage sampling**
- 6. Sampling with probability proportional to size**

1. Simple random sampling

- The required number of individuals are selected at random from the sampling frame, a list or a database of all individuals in the population
- Each member of a population has an equal chance of being included in the sample.

- **To use a SRS method:**

- Make a numbered list of all the units in the population
- Each unit should be numbered from 1 to N (where N is the size of the population)
- Select the required number.
- The randomness of the sample is ensured by:
 - Use of “lottery” methods
 - Table of random numbers
 - Computer programs

Random number table

- It is a table of random numbers constructed by a process that
 1. In any position in the table, each of the numbers 0 through 9 has a probability $1/10$ of occurring.
 2. The occurrence of any number in one part of the table is independent of the occurrence of any number in any other part of the table.

Random numbers

....	8094	2525	8247	1347	7433	3620	1897	
....	3563	2198	8211	9045	2618	2751	2627	
....	1330	6331	3753	9693	8738	6815	1538	
....	3565	0016	2243	6432	4796	6095	5283	
....	7850	5925	5588	7311	2192	4545	3530	
....	4490	5417	9727	6153	5901	4878	9980	
....	6545	9104	9318	8819	7537	2785	9373	

Example

- Suppose your school has 500 students and you need to conduct a short survey on the quality of the food served in the cafeteria.
- You decide that a sample of 10 students should be sufficient for your purposes.
- In order to get your sample, you assign a number from 1 to 500 to each student in your school.

- To select the sample, you use a table of randomly generated numbers.
- Pick a starting point in the table (a row and column number) and look at the random numbers that appear there.
- In this case, since the data run into three digits, the random numbers would need to contain three digits as well.

- Ignore all random numbers after 500 because they do not correspond to any of the students in the school.
- Remember that the sample is without replacement, so if a number recurs, skip over it and use the next random number.
- The first 10 different numbers between 001 and 500 make up your sample.

- **SRS has certain limitations:**

- Requires a sampling frame.
- Difficult if the reference population is dispersed.
- Minority subgroups of interest may not be selected.

2. Systematic random sampling

- Sometimes called *interval sampling*
- Selection of individuals from the sampling frame systematically rather than randomly
- Individuals are taken at regular intervals down the list.
- The starting point is chosen at random.

- Important if the reference population is arranged in some order:
 - Order of registration of patients
 - Numerical number of house numbers
 - Student's registration books
- Taking individuals at fixed intervals (every k^{th}) based on the sampling fraction, eg. if the sample includes 20%, then every fifth.

Steps in systematic random sampling

1. Number the units on your frame from 1 to **N** (where **N** is the total population size).
2. Determine the sampling interval (**K**) by dividing the number of units in the population by the desired sample size.
3. Select a number between one and **K** at random. This number is called *the random start* and would be the first number included in your sample.
4. Select every **Kth** unit after that first number

Note: *Systematic sampling should not be used when a cyclic repetition is inherent in the sampling frame*

Example

- To select a sample of 100 from a population of 400, you would need a sampling interval of $400 \div 100 = 4$.
- Therefore, **K** = 4.
- You will need to select one unit out of every four units to end up with a total of 100 units in your sample.
- Select a number between 1 and 4 from a table of random numbers.

- If you choose 3, the third unit on your frame would be the first unit included in your sample;
- The sample might consist of the following units to make up a sample of 100: 3 (the random start), 7, 11, 15, 19...395, 399 (up to **N**, which is 400 in this case).

- Using the above example, you can see that with a systematic sample approach there are only four possible samples that can be selected, corresponding to the four possible random starts:

A. 1, 5, 9, 13...393, 397

B. 2, 6, 10, 14...394, 398

C. 3, 7, 11, 15...395, 399

D. 4, 8, 12, 16...396, 400

- Each member of the population belongs to only one of the four samples and each sample has the same chance of being selected.
- The main difference with SRS, any combination of 100 units would have a chance of making up the sample, while with systematic sampling, there are only four possible samples.

3. Stratified random sampling

- It is done when the population is known to have heterogeneity with regard to some factors and those factors are used for stratification.
- Using **stratified sampling**, the population is divided into homogeneous, mutually exclusive groups called **strata**, and
- A population can be stratified by any variable that is available for all units prior to sampling (e.g., age, sex, province of residence, income, etc.).

- A separate sample is taken independently from each stratum.
- Any of the sampling methods mentioned in this section (and others that exist) can be used to sample within each stratum.

Why do we need to create strata?

- It can make the sampling strategy more efficient.
- A larger sample is required to get a more accurate estimation if a characteristic varies greatly from one unit to the other.
- For example, if every person in a population had the same salary, then a sample of one individual would be enough to get a precise estimate of the average salary.

- This is the idea behind the efficiency gain obtained with stratification.
 - If you create strata within which units share similar characteristics (e.g., income) and are considerably different from units in other strata (e.g., occupation, type of dwelling) then you would only need a small sample from each stratum to get a precise estimate of total income for that stratum.
 - Then you could combine these estimates to get a precise estimate of total income for the whole population.

- If you use a SRS approach in the whole population without stratification, the sample would need to be larger than the total of all stratum samples to get an estimate of total income with the same level of precision.

- Stratified sampling ensures an adequate sample size for sub-groups in the population of interest.
- When a population is stratified, each stratum becomes an independent population and you will need to decide the sample size for each stratum.

- **Equal allocation:**

- Allocate equal sample size to each stratum

- **Proportionate allocation:**

$$n_j = \frac{n}{N} N_j$$

- n_j is sample size of the j^{th} stratum
 - N_j is population size of the j^{th} stratum
 - $n = n_1 + n_2 + \dots + n_k$ is the total sample size
 - $N = N_1 + N_2 + \dots + N_k$ is the total population size

Example: Proportionate Allocation

• Village Total	A	B	C	D
• HHs 500	100	150	120	130
• S. size 60	?	?	?	?

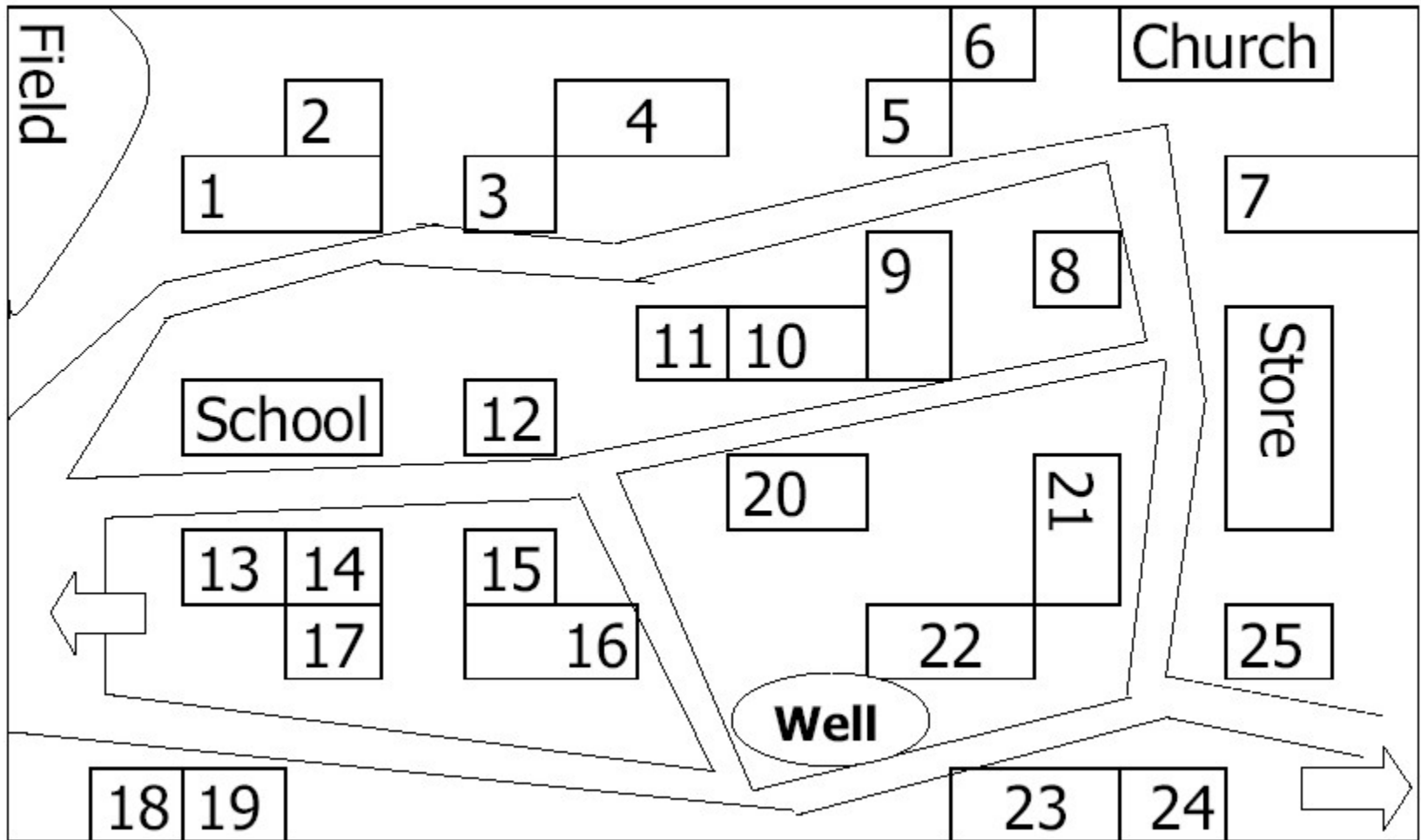
4. Cluster sampling

- Sometimes it is too expensive to carry out SRS
 - Population may be large and scattered.
 - Complete list of the study population unavailable
 - Travel costs can become expensive if interviewers have to survey people from one end of the country to the other.
- **Cluster sampling** is the most widely used to reduce the cost
- The clusters should be **homogeneous**, unlike stratified sampling where the strata are heterogeneous

Steps in cluster sampling

- Cluster sampling divides the population into groups or clusters.
- A number of clusters are selected randomly to represent the total population, and then all units within selected clusters are included in the sample.
- No units from non-selected clusters are included in the sample—they are represented by those from selected clusters.
- This differs from stratified sampling, where some units are selected from each group.

Cluster Map



Example

- In a school based study, we assume students of the same school are homogeneous.
- We can select randomly sections and include all students of the selected sections only

Advantages

- Cost reduction
- It creates 'pockets' of sampled units instead of spreading the sample over the whole territory.
- Sometimes a list of all units in the population is not available, while a list of all clusters is either available or easy to create.

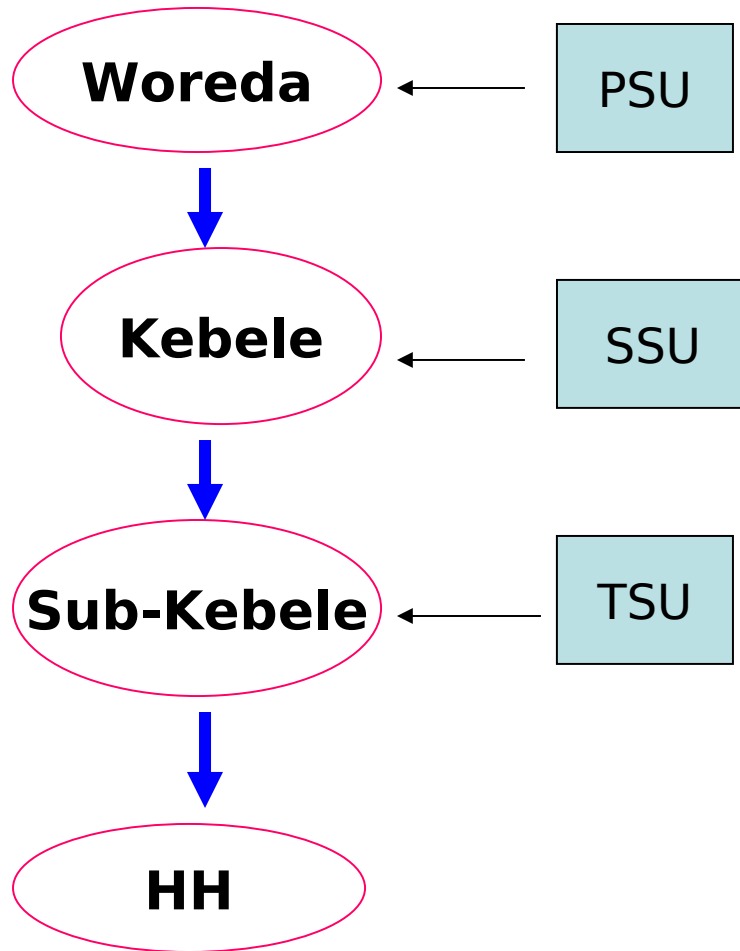
Disadvantages

- Creates a loss of efficiency when compared with SRS.
- It is usually better to survey a large number of small clusters instead of a small number of large clusters.
 - This is because neighboring units tend to be more alike, resulting in a sample that does not represent the whole spectrum of opinions or situations present in the overall population.

- Another drawback to cluster sampling is that you do not have total control over the final sample size.
- Since not all schools have the same number of (say Grade 11) students and city blocks do not all have the same number of households, and you must interview every student or household in your sample, as an example, the final size may be larger or smaller than you expected.

5. Multi-stage sampling

- Similar to the cluster sampling, except that it involves picking a sample from within each chosen cluster, rather than including all units in the cluster.
- This type of sampling requires at least two stages.
- The primary sampling unit (PSU) is the sampling unit in the first sampling stage.
- The secondary sampling unit (SSU) is the sampling unit in the second sampling stage, etc



- In the first stage, large groups or clusters are identified and selected. These clusters contain more population units than are needed for the final sample.
- In the second stage, population units are picked from within the selected clusters (using any of the possible probability sampling methods) for a final sample.

- If more than two stages are used, the process of choosing population units within clusters continues until there is a final sample.
- With multi-stage sampling, you still have the benefit of a more concentrated sample for cost reduction.
- However, the sample is not as concentrated as other clusters and the sample size is still bigger than for a simple random sample size.

- Also, you do not need to have a list of all of the units in the population. All you need is a list of clusters and list of the units in the selected clusters.
- Admittedly, more information is needed in this type of sample than what is required in cluster sampling.
- However, multi-stage sampling still saves a great amount of time and effort by not having to create a list of all the units in a population.

6. Sampling with probability proportional to size

- Probability sampling requires that each member of the survey population has a chance of being included in the sample, but it does not require that this chance be the same for everyone.
- Requires that a sampling frame of clusters with measures of size be available or developed
- This information can be used in the sampling selection in order to increase the efficiency.
- This is known as **sampling with probability proportional to size (PPS)**

- With this method, the bigger the size of the unit, the higher the chance it has of being included in the sample.
- For this method to achieve increased efficiency, the measure of size needs to be accurate.

Steps in PPS

- List all Kebeles/clusters with their population size/HH size
- Calculate the cumulative frequency
- Calculate the sampling interval by dividing the total population size by the sample size, say K
- Randomly choose a number between 1 and K , say j
- Kebeles/clusters with cumulative frequency containing the j^{th} , $(j+K)^{\text{th}}$, ... will be included in the sample

Example

- Planned clusters to be included in the study = 40
- Cumulative size of the HHs = 17,219
- Sampling interval = $17,219/40 = 430$
- Random start between 1 and 430 = 73
- Clusters selected = 001, 005, 008, etc.

Cluster No.	HH size	Cum. size	Sampling No.	Cluster selected
001	120	120	73	001
002	105	225		
003	132	357		
004	96	453		
005	110	563	503	005
006	102	665		
007	165	839		
008	98	937	933	008
009	115	1,052		
.	.	.	.	.
.	.	.	.	.
170	196	17,219		
(last)				

When measures of cluster size are not available

- When the measures of size (population of HH size) are not available, all clusters will have the same chance or probability of selection
- This is **equal probability**
- Decide on the number of clusters to be included in the study
- Use SRS or systematic sampling to select them

B. Non-probability sampling

- In non-probability sampling, every item has an unknown chance of being selected.
- In non-probability sampling, there is an assumption that there is an even distribution of a characteristic of interest within the population.
- For probability sampling, random is a feature of the selection process.

- This is what makes the researcher believe that any sample would be representative and because of that, results will be accurate.
- For probability sampling, random is a feature of the selection process, rather than an assumption about the structure of the population.

- In non-probability sampling, since elements are chosen arbitrarily, there is no way to estimate the probability of any one element being included in the sample.
- Also, no assurance is given that each item has a chance of being included, making it impossible either to estimate sampling variability or to identify possible bias

- Reliability cannot be measured in non-probability sampling; the only way to address data quality is to compare some of the survey results with available information about the population.
- Still, there is no assurance that the estimates will meet an acceptable level of error.
- Researchers are reluctant to use these methods because there is no way to measure the precision of the resulting sample.

- Despite these drawbacks, non-probability sampling methods can be useful when descriptive comments about the sample itself are desired.
- Secondly, they are quick, inexpensive and convenient.
- There are also other circumstances, such as researches, when it is unfeasible or impractical to conduct probability sampling.

The most common types of non-probability sampling

- 1. Convenience or haphazard sampling**
- 2. Volunteer sampling**
- 3. Judgment sampling**
- 4. Quota sampling**
- 5. Snowball sampling technique**

1. Convenience or haphazard sampling

- **Convenience sampling** is sometimes referred to as *haphazard* or *accidental sampling*.
- It is not normally representative of the target population because sample units are only selected if they can be **accessed easily and conveniently**.

- The obvious advantage is that the method is easy to use, but that advantage is greatly offset by the presence of **bias**.
- Although useful applications of the technique are limited, it can deliver accurate results when the population is **homogeneous**.

- For example, a scientist could use this method to determine whether a lake is polluted or not.
- Assuming that the lake water is well-mixed, any sample would yield similar information.
- A scientist could safely draw water anywhere on the lake without bothering about whether or not the sample is representative

2. Volunteer sampling

- As the term implies, this type of sampling occurs when people volunteer to be involved in the study.
- In psychological experiments or pharmaceutical trials (drug testing), for example, it would be difficult and unethical to enlist random participants from the general public.
- In these instances, the sample is taken from a group of volunteers.

- Sometimes, the researcher offers payment to attract respondents.
- In exchange, the volunteers accept the possibility of a lengthy, demanding or sometimes unpleasant process.

- Sampling voluntary participants as opposed to the general population may introduce **strong biases**.
- Often in opinion polling, only the people who care strongly enough about the subject tend to respond.
- **The silent majority** do not typically respond, resulting in large **selection bias**.

3. Judgment sampling _

- This approach is used when a sample is taken based on certain judgments about the overall population.
- The underlying assumption is that the investigator will select units that are characteristic of the population.
- The critical issue here is **objectivity**: how much can judgment be relied upon to arrive at a typical sample?

- ***Judgment sampling*** is subject to the researcher's biases and is perhaps even more biased than haphazard sampling.
- Since any preconceptions the researcher may have are reflected in the sample, large biases can be introduced if these preconceptions are inaccurate.

- Researchers often use this method in exploratory studies like pre-testing of questionnaires and focus groups.
- They also prefer to use this method in laboratory settings where the choice of experimental subjects (i.e., animal, human) reflects the investigator's pre-existing beliefs about the population.

- One advantage of **judgment sampling** is the **reduced cost and time** involved in acquiring the sample.

4. Quota sampling

- This is one of the most common forms of non-probability sampling.
- Sampling is done until a specific number of units (quotas) for various sub-populations have been selected.

- Since there are no rules as to how these quotas are to be filled, ***quota sampling*** is really a means for satisfying sample size objectives for certain sub-populations.

- As with all other non-probability sampling methods, in order to make inferences about the population, it is necessary to assume that persons selected are similar to those not selected.
- Such strong assumptions are rarely valid.

- The main argument against quota sampling is that it does not meet the basic requirement of randomness.
- Some units may have no chance of selection or the chance of selection may be unknown.
- Therefore, the sample may be **biased**.

- **Quota sampling** is generally less expensive than random sampling.
- It is also easy to administer, especially considering the tasks of listing the whole population, randomly selecting the sample and following-up on non-respondents can be omitted from the procedure.

- **Quota sampling** is an effective sampling method when information is urgently required and can be conducted without sampling frames.
- In many cases where the population has no suitable frame, quota sampling may be the only appropriate sampling method.

5. Snowball sampling

- A technique for selecting a research sample where existing study subjects recruit future subjects from among their acquaintances.
- Thus the sample group appears to grow like a rolling snowball.

- This sampling technique is often used in hidden populations which are difficult for researchers to access; example populations would be drug users or commercial sex workers.
- Because sample members are not selected from a sampling frame, snowball samples are subject to numerous biases. For example, people who have many friends are more likely to be recruited into the sample.