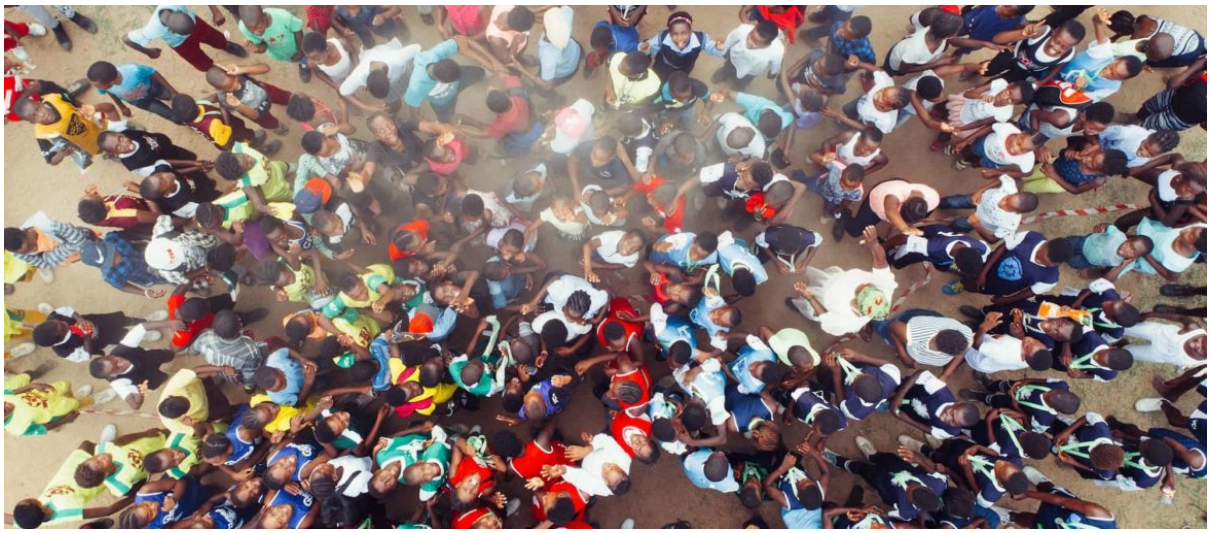
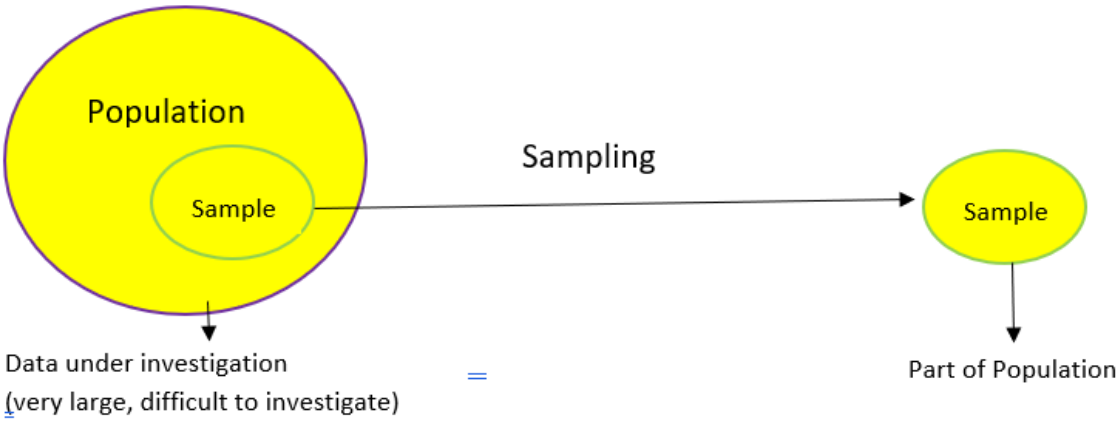
How to Choose a Representative Subset of the Population



What is Sampling?

Let's start by formally defining what sampling is.

Sampling is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual.



The above diagram perfectly illustrates what sampling is. Let's understand this at a more intuitive level through an example.

We want to find the average height of all adult males in Delhi. The population of Delhi is around 3 crore and males would be roughly around 1.5 crores (these are general assumptions for this example so don't take them at face value!). As you can imagine, it is nearly impossible to find the average height of all males in Delhi.

It's also not possible to reach every male so we can't really analyze the entire population. So what *can* we do instead? We can take multiple samples and calculate the average height of individuals in the selected samples.

But then we arrive at another question – how can we take a sample? Should we take a random sample? Or do we have to ask the experts?

Let's say we go to a basketball court and take the average height of all the professional basketball players as our sample. This will not be considered a good sample because generally, a basketball player is

taller than an average male and it will give us a bad estimate of the average male's height.

Here's a potential solution – find random people in random situations where our sample would not be skewed based on heights.

What Is Survey Sampling?

Surveys would be meaningless and incomplete without accounting for the respondents that they're aimed at. The best survey design practices keep the target population at the core of their thought process.

'All the residents of the Dharavi slums in Mumbai', 'every NGO in Calcutta' and 'all students below the age of 16 in Manipur' are examples of a population; they are countable, finite and well-defined.

When the population is small enough, researchers have the resources to reach out to all of them. This would be the best case scenario, making sure that everybody who matters to the survey is represented accurately. A survey that covers the entire target population is called a census.

However, most surveys cannot survey the entire population. This is when sampling techniques become crucial to your survey.

Why Is It Important?

Resource Constraints

If the target population is not small enough, or if the resources at your disposal don't give you the bandwidth to cover the entire population, it is important to identify a subset of the population to work with – a carefully identified group that is representative of the population. This process is called survey sampling, and it is one of the most important aspects of survey design.

Whatever the sample size, there are fixed costs associated with any survey. Once the survey has begun, the marginal costs associated with gathering more information, from more people, are proportional to the size of the sample.

Drawing Inferences About the Population

Researchers are not interested in the sample itself, but in the understanding that they can potentially infer from the sample and then apply across the entire population.

A sample survey usually offers greater scope than a census. Working within a given resource constraint, sampling may make it possible to study the population of a larger geographical area or to find out more about the same population by examining an area in greater depth through a smaller sample.

Before we dive into the survey sampling methods at our disposal it is imperative that we develop a perspective on what an effective sample should look like.

Why do we need Sampling?

I'm sure you have a solid intuition at this point regarding the question.

Sampling is done to draw conclusions about populations from samples, and it enables us to determine a population's characteristics by directly observing only a portion (or sample) of the population.

- Selecting a sample requires less time than selecting every item in a population
- Sample selection is a cost-efficient method
- Analysis of the sample is less cumbersome and more practical than an analysis of the entire population

3 Features to Keep in Mind While Constructing a Sample

Consistency

It is important that researchers understand the population on a case-by-case basis and test the sample for consistency before going ahead with the survey. This is especially critical for surveys that track changes across time and space where we need to be confident that any change we see in our data reflects real change – across consistent and comparable samples.

Diversity

Ensuring diversity of the sample is a tall order, as reaching some portions of the population and convincing them to participate in the survey could be difficult. But to be truly representative of the population, a sample must be as diverse as the population itself and sensitive to the local differences that are unavoidable as we move across the population.

Transparency

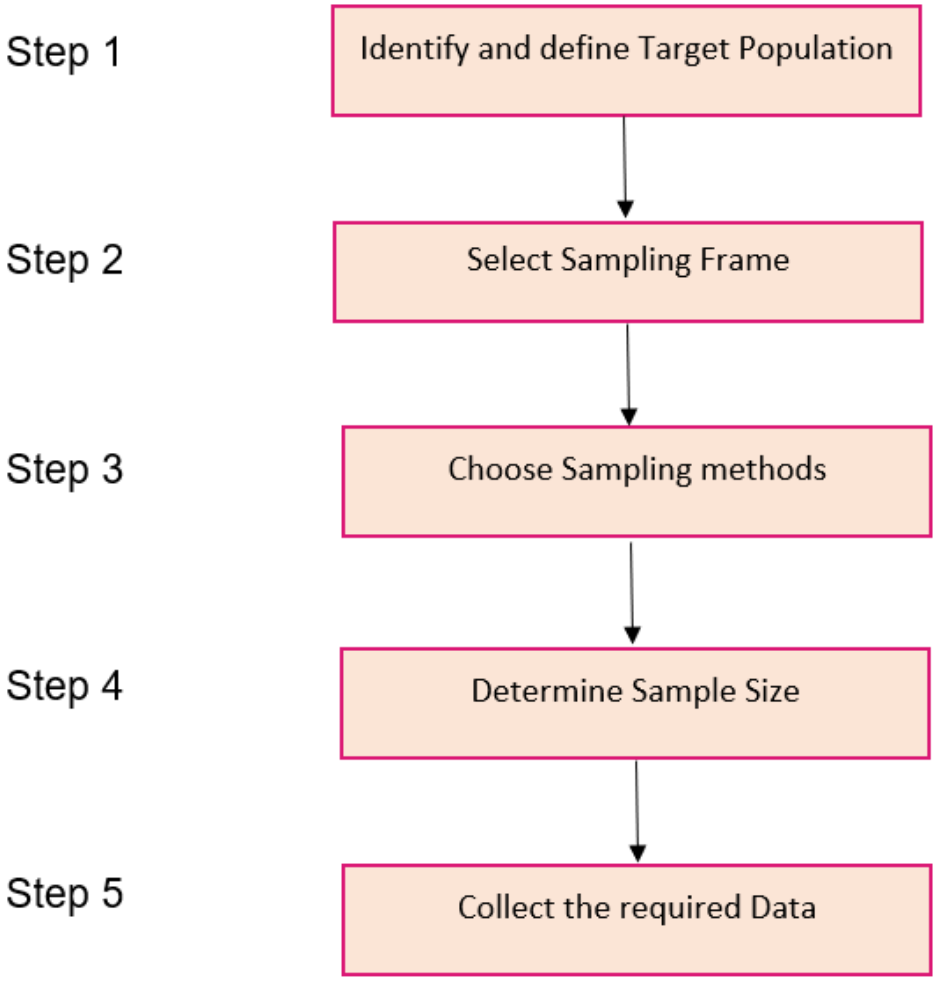
There are several constraints that dictate the size and structure of the population. It is imperative that researchers discuss these limitations and maintain transparency about the procedures followed while selecting the sample so that the results of the survey are seen with the right perspective.

Now that we understand the necessity of choosing the right sample and have a vision of what an effective sample for your survey should be like, let's explore the various methods of constructing a sample and understand the relative pros and cons of each of these approaches.

Sampling methods can broadly be classified as probability and non-probability.

Steps involved in Sampling

I firmly believe visualizing a concept is a great way to ingrain it in your mind. So here's a step-by-step process of how sampling is typically done, in flowchart form!



Let's take an interesting case study and apply these steps to perform sampling. We recently conducted General Elections in India a few months back. You must have seen the public opinion polls every news channel was running at the time:

Were these results concluded by considering the views of all 900 million voters of the country or a fraction of these voters? Let us see how it was done.

Step 1

The first stage in the sampling process is to clearly define the target population.

So, to carry out opinion polls, polling agencies consider only the people who are above 18 years of age and are eligible to vote in the population.

Step 2

Sampling Frame – It is a list of items or people forming a population from which the sample is taken.

So, the sampling frame would be the list of all the people whose names appear on the voter list of a constituency.

Step 3

Generally, probability sampling methods are used because every vote has equal value and any person can be included in the sample irrespective of his caste, community, or religion. Different samples are taken from different regions all over the country.

Step 4

Sample Size – It is the number of individuals or items to be taken in a sample that would be enough to make inferences about the population with the desired level of accuracy and precision.

Larger the sample size, more accurate our inference about the population would be.

For the polls, agencies try to get as many people as possible of diverse backgrounds to be included in the sample as it would help in predicting the number of seats a political party can win.

Step 5

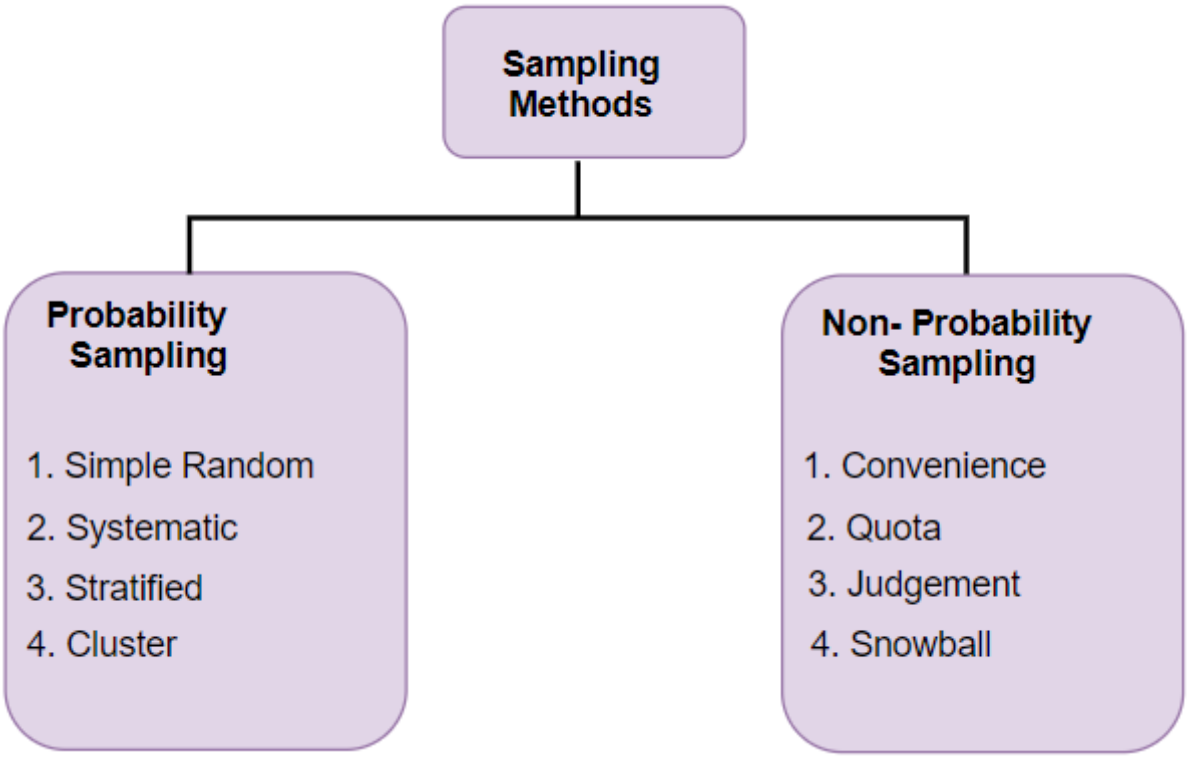
Once the target population, sampling frame, sampling technique, and sample size have been established, the next step is to **collect data from the sample**.

In opinion polls, agencies generally put questions to the people, like which political party are they going to vote for or has the previous party done any work, etc.

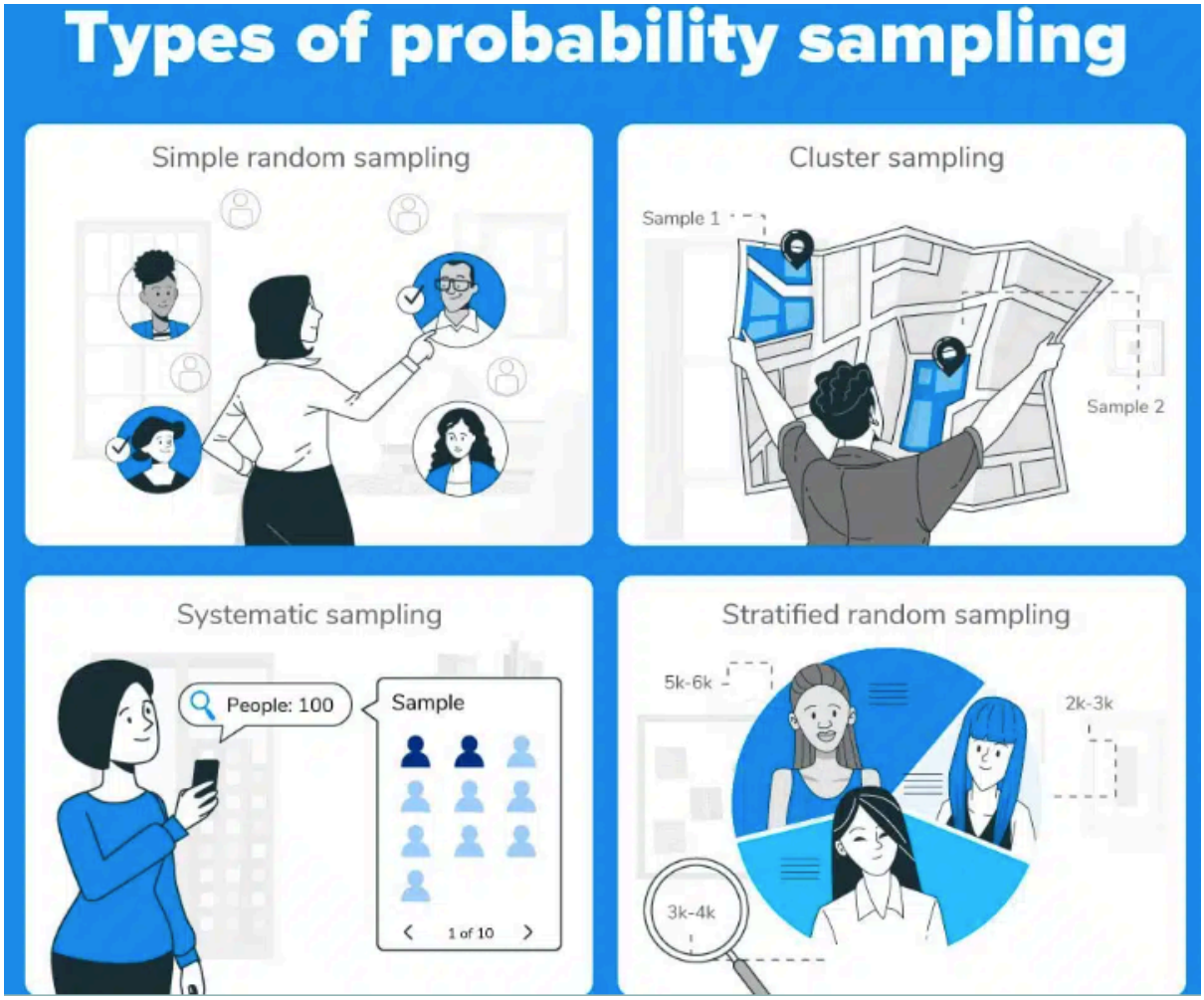
Based on the answers, agencies try to interpret who the people of a constituency are going to vote for and approximately how many seats is a political party going to win. Pretty exciting work, right?!

Different Types of Sampling Techniques

Here comes another diagrammatic illustration! This one talks about the different types of sampling techniques available to us:



- **Probability Sampling:** In probability sampling, every element of the population has an equal chance of being selected. Probability sampling gives us the best chance to create a sample that is truly representative of the population



Random Sampling

When: There is a very large population and it is difficult to identify every member of the population.

How: The entire process of sampling is done in a single step with each subject selected independently of the other members of the population. The term random has a very precise meaning and you can't just collect responses on the street and have a random sample.

Pros: In this technique, each member of the population has an equal chance of being selected as subject.

Cons: When there are very large populations, it is often difficult to identify every member of the population and the pool of subjects becomes biased. Dialing numbers from a phone book for instance, may not be entirely random as the numbers, though random, would correspond to a localized region. A sample created by doing so might leave out many sections of the population that are significant to the study.

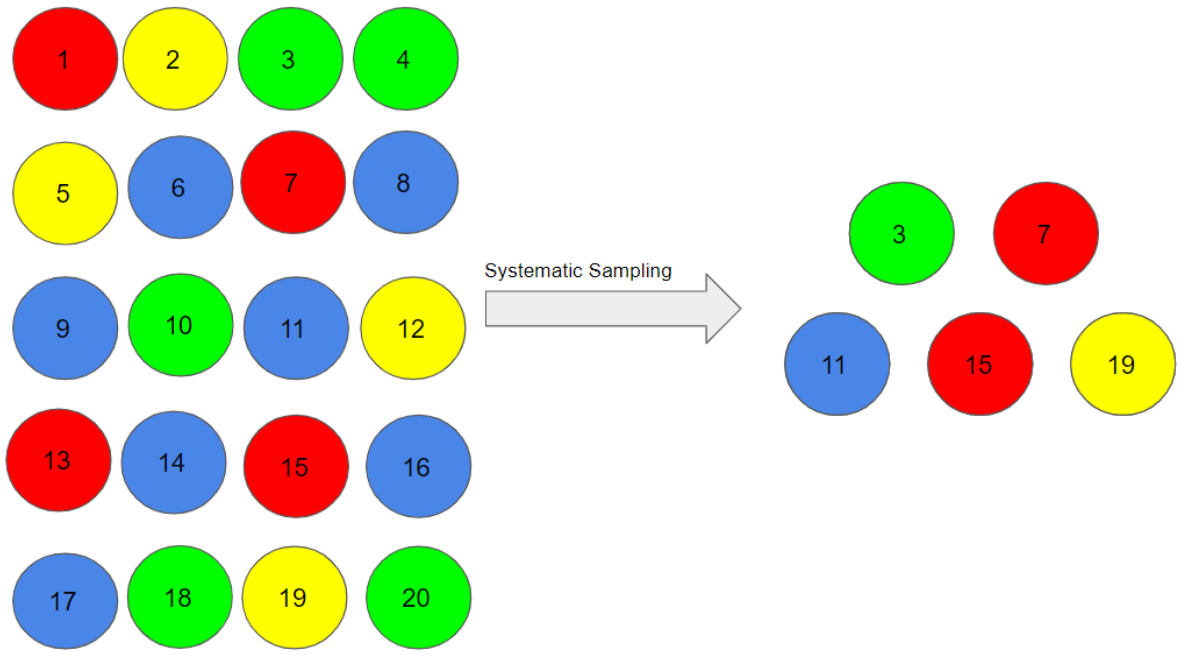
Use case: Want to study and understand the rice consumption pattern across rural India? While it might not be possible to cover every household, you could draw meaningful insights by building your sample from different districts or villages (depending on the scope).

Systematic Sampling

When: Your given population is logically homogenous.

How: In a systematic sample, after you decide the sample size, arrange the elements of the population in some order and select terms at regular intervals from the list.

Example - Say our population size is x and we have to select a sample size of n. Then, the next individual that we will select would be x/nth intervals away from the first individual. We can select the rest in the same way.



Suppose, we began with person number 3, and we want a sample size of 5. So, the next individual that we will select would be at an interval of (20/5) = 4 from the 3rd person, i.e. 7 (3+4), and so on.

3, 3+4=7, 7+4=11, 11+4=15, 15+4=19 = **3, 7, 11, 15, 19**

Systematic sampling is more convenient than simple random sampling. However, it might also lead to bias if there is an underlying pattern in which we are selecting items from the population (though the chances of that happening are quite rare).

Pros: The main advantage of using systematic sampling over simple random sampling is its simplicity. Another advantage of systematic random sampling over simple random sampling is the assurance that the population will be evenly sampled. There exists a chance in simple random sampling that allows a clustered selection of subjects. This can be avoided through systematic sampling.

Cons: The possible weakness of the method that may compromise the randomness of the sample is an inherent periodicity of the list. This can be avoided by randomizing the list of your population entities, as you would randomize a deck of cards for instance, before you proceed with systematic sampling.

Use Case: Suppose a supermarket wants to study buying habits of their customers. Using systematic sampling, they can choose every 10th or 15th customer entering the supermarket and conduct the study on this sample.

Stratified Sampling

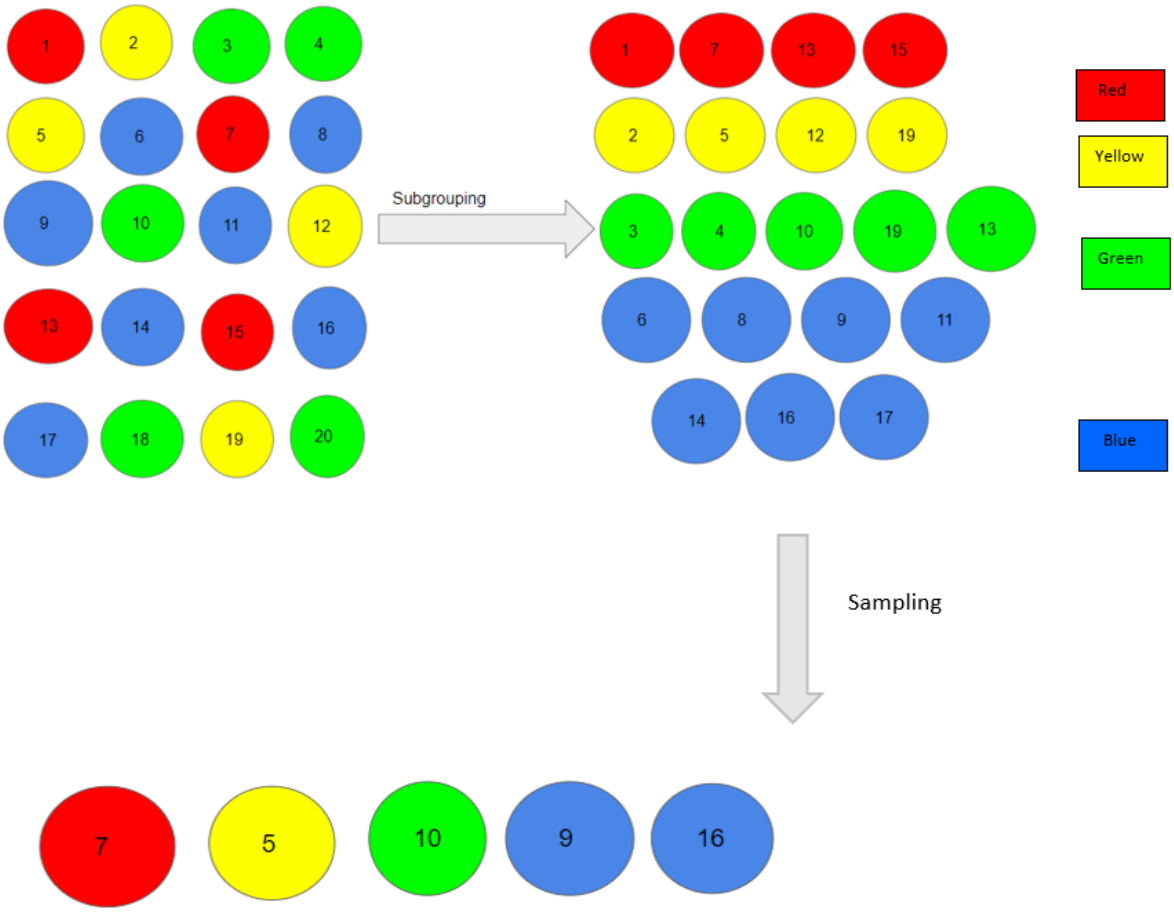
When: You can divide your population into characteristics of importance for the research.

How: A stratified sample, in essence, tries to recreate the statistical features of the population on a smaller scale. Before sampling, the population is divided into characteristics of importance for the research — for example, by gender, social class, education level, religion, etc. Then the population is randomly sampled within each category or stratum. If 38% of the population is college-educated, then 38% of the sample is randomly selected from the college-educated subset of the population.

In this type of sampling, we divide the population into subgroups (called strata) based on different traits like gender, category, etc. And then we select the sample(s) from these subgroups:

Here, we first divided our population into subgroups based on different colors of red, yellow, green and blue. Then, from each color, we selected an individual in the proportion of their numbers in the population.

We use this type of sampling when we want representation from all the subgroups of the population. However, stratified sampling requires proper knowledge of the characteristics of the population.



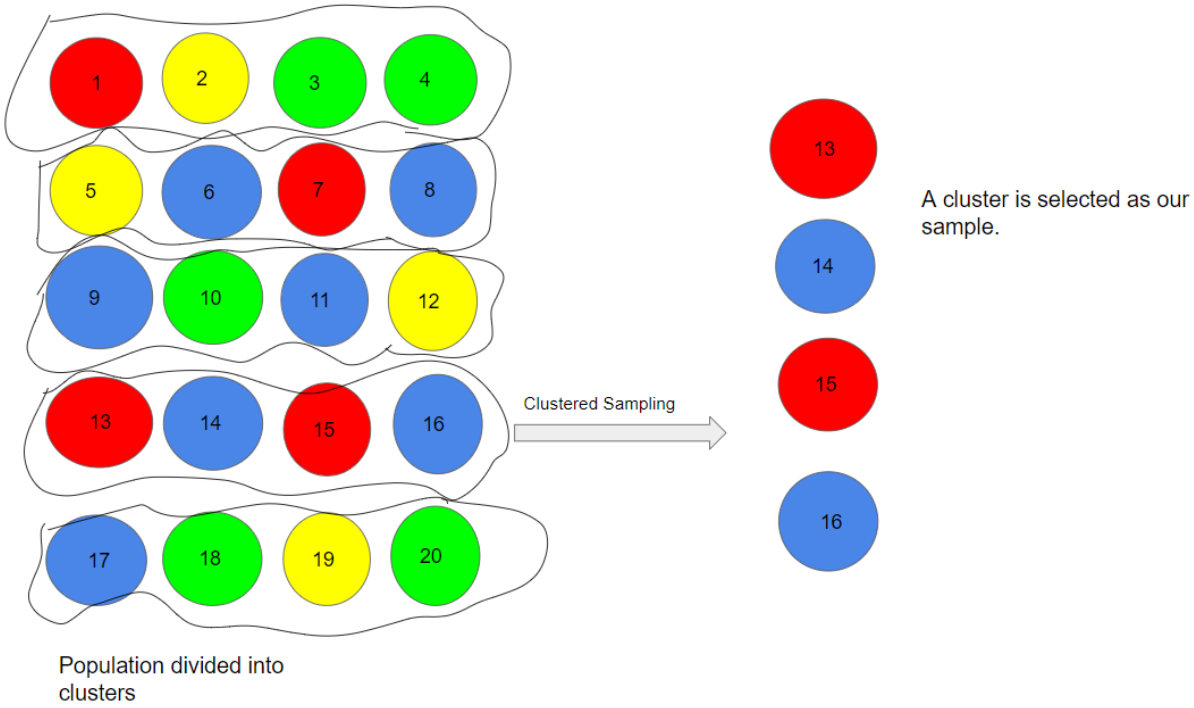
Pros: This method attempts to overcome the shortcomings of random sampling by splitting the population into various distinct segments and selecting entities from each of them. This ensures that every category of the population is represented in the sample. Stratified sampling is often used when one or more of the sections in the population have a low incidence relative to the other sections.

Cons: Stratified sampling is the most complex method of sampling. It lays down criteria that may be difficult to fulfill and place a heavy strain on your available resources.

Use Case: If 38% of the population is college-educated and 62% of the population have not been to college, then 38% of the sample is randomly selected from the college-educated subset of the population and 62% of the sample is randomly selected from the non-college-going population. Maintaining the ratios while selecting a randomized sample is key to stratified sampling.

Cluster Sampling

In a clustered sample, we use the subgroups of the population as the sampling unit rather than individuals. The population is divided into subgroups, known as clusters, and a whole cluster is randomly selected to be included in the study:



In the above example, we have divided our population into 5 clusters. Each cluster consists of 4 individuals and we have taken the 4th cluster in our sample. We can include more clusters as per our sample size.

This type of sampling is used when we focus on a specific region or area.

Uses of probability sampling

There are multiple uses of probability sampling:

- **Reduce Sample Bias:** Using the probability sampling method, the bias in the sample derived from a population is negligible to non-existent. The sample selection mainly depicts the researcher's understanding and inference. Probability sampling leads to higher-quality data collection as the sample appropriately represents the population.
- **Diverse Population:** When the population is vast and diverse, it is essential to have adequate representation so that the data is not skewed toward one demographic. For example, suppose Square would like to understand the people that could make their point-of-sale devices. In that case, a survey conducted from a sample of people across the India from different industries and socio-economic backgrounds helps.
- **Create an Accurate Sample:** Probability sampling helps the researchers plan and create an accurate sample. This helps to obtain well-defined data.

- **Non-Probability Sampling:** In non-probability sampling, all elements do not have an equal chance of being selected. Consequently, there is a significant risk of ending up with a non-representative sample which does not produce generalizable results

  For example, let's say our population consists of 20 individuals. Each individual is numbered from 1 to 20 and is represented by a specific color (red, blue, green, or yellow). Each person would have odds of 1 out of 20 of being chosen in probability sampling.

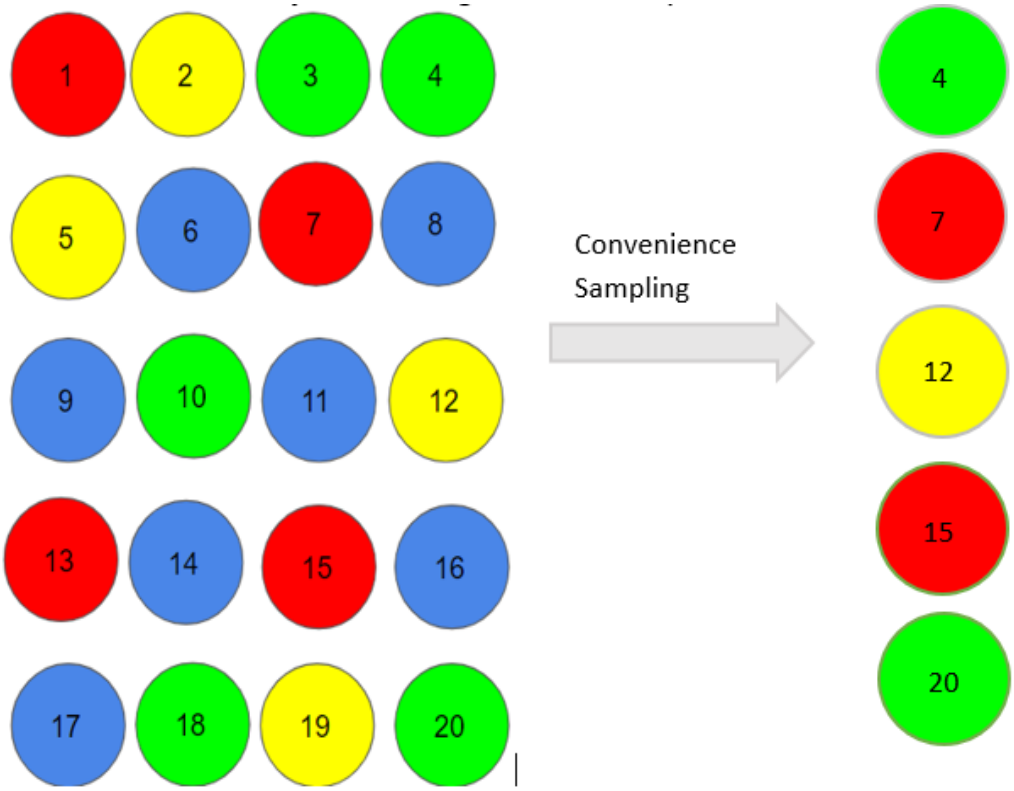
  With non-probability sampling, these odds are not equal. A person might have a better chance of being chosen than others.

  So now that we have an idea of these two sampling types, let's dive into each and understand the different types of sampling under each section.

  Convenience Sampling

  This is perhaps the easiest method of sampling because individuals are selected based on their availability and willingness to take part.

  Here, let's say individuals numbered 4, 7, 12, 15 and 20 want to be part of our sample, and hence, we will include them in the sample.



Convenience sampling is prone to significant bias, because the sample may not be the representation of the specific characteristics such as religion or, say the gender, of the population.

Convenience Sampling

When: During preliminary research efforts.

How: As the name suggests, the elements of such a sample are picked only on the basis of convenience in terms of availability, reach and accessibility.

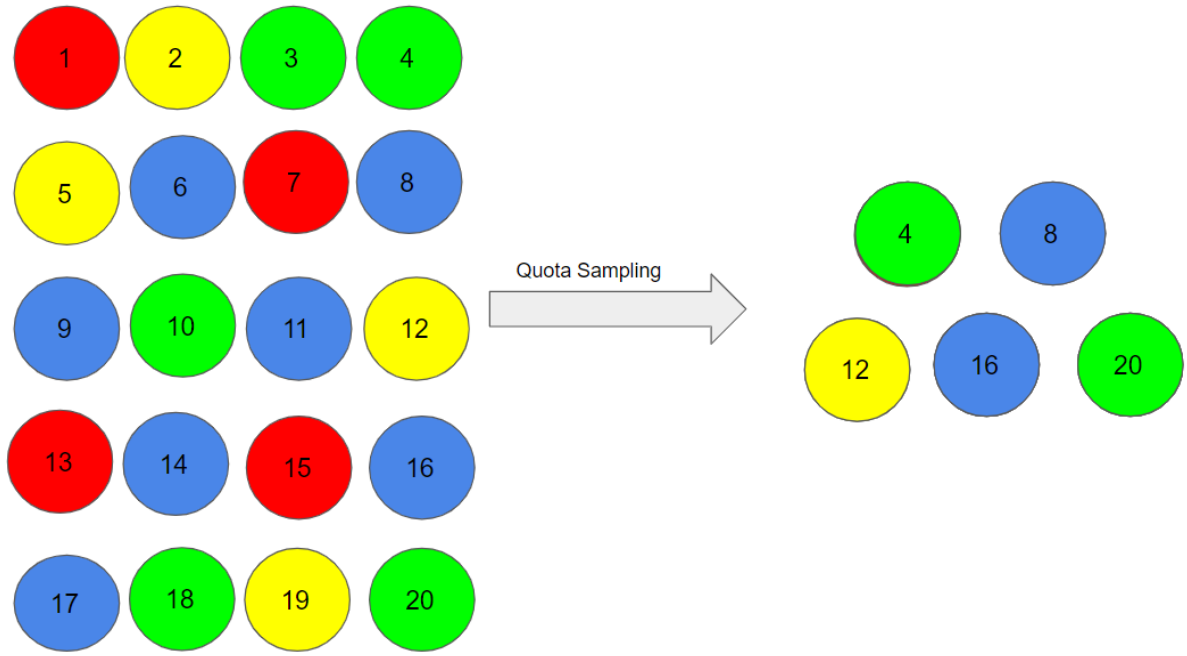
Pros: The sample is created quickly without adding any additional burden on the available resources.

Cons: The likelihood of this approach leading to a sample that is truly representative of the population is very poor.

Use Case: This method is often used during preliminary research efforts to get a gross estimate of the results, without incurring the cost or time required to select a random sample.

Quota Sampling

In this type of sampling, we choose items based on predetermined characteristics of the population. Consider that we have to select individuals having a number in multiples of four for our sample:



Therefore, the individuals numbered 4, 8, 12, 16, and 20 are already reserved for our sample.

In quota sampling, the chosen sample might not be the best representation of the characteristics of the population that weren't considered.

When: When you can characterize the population based on certain desired features.

How: Quota sampling is the non-probability equivalent of stratified sampling that we discussed earlier. It starts with characterizing the population based on certain desired features and assigns a quota to each subset of the population.

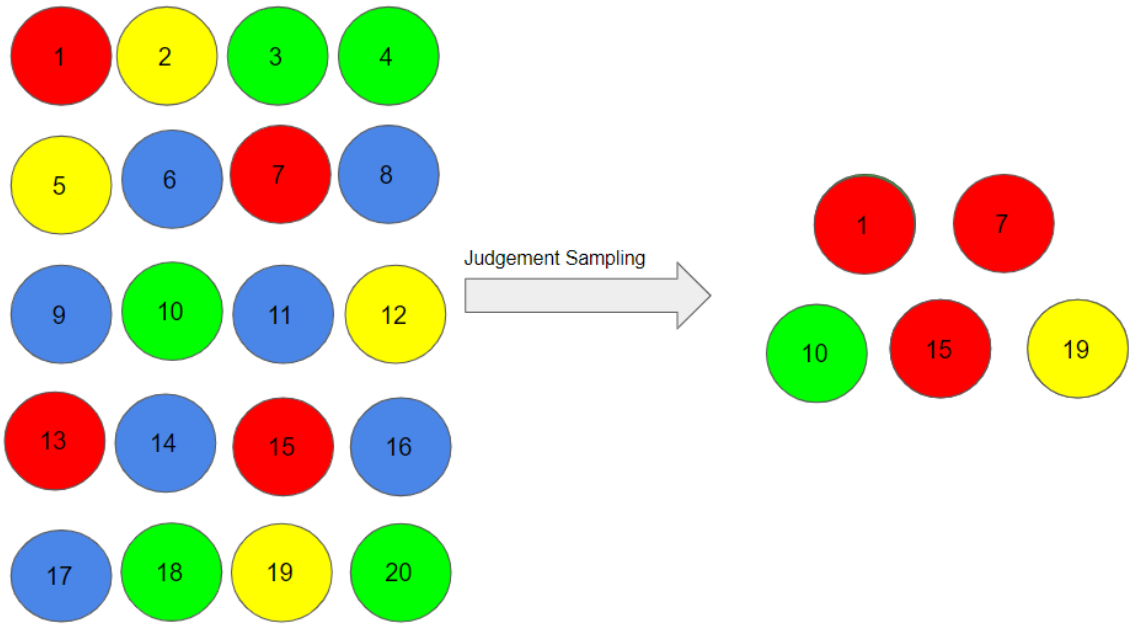
Pros: This process can be extended to cover several characteristics and varying degrees of complexity.

Cons: Though the method is superior to convenience and snowball sampling, it does not offer the statistical insights of any of the probability methods.

Use Case: If a survey requires a sample of fifty men and fifty women, a quota sample will survey respondents until the right number of each type has been surveyed. Unlike stratified sampling, the sample isn't necessarily randomized.

Judgment Sampling

It is also known as selective sampling. It depends on the judgment of the experts when choosing whom to ask to participate.

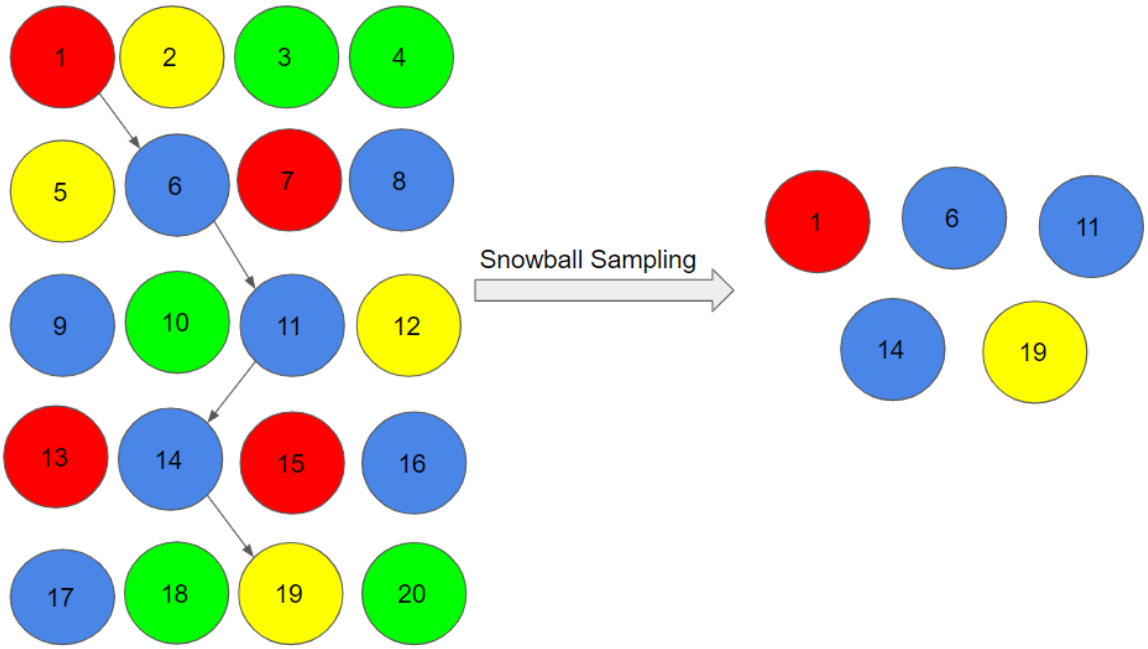


Suppose, our experts believe that people numbered 1, 7, 10, 15, and 19 should be considered for our sample as they may help us to infer the population in a better way. As you can imagine, quota sampling is also prone to bias by the experts and may not necessarily be representative.

Snowball Sampling

I quite like this sampling technique. **Existing people are asked to nominate further people known to them so that the sample increases in size like a rolling snowball.** This method of sampling is

effective when a sampling frame is difficult to identify.



Here, we had randomly chosen person 1 for our sample, and then he/she recommended person 6, and person 6 recommended person 11, and so on.

1->6->11->14->19

There is a significant risk of selection bias in snowball sampling, as the referenced individuals will share common traits with the person who recommends them.

When: When you can rely on your initial respondents to refer you to the next respondents.

How: Just as the snowball rolls and gathers mass, the sample constructed in this way will grow in size as you move through the process of conducting a survey. In this technique, you rely on your

initial respondents to refer you to the next respondents whom you may connect with for the purpose of your survey.

Pros: The costs associated with this method are significantly lower, and you will end up with a sample that is very relevant to your study.

Cons: The clear downside of this approach is that you may restrict yourself to only a small, largely homogenous section of the population.

Use Case: Snowball sampling can be useful when you need the sample to reflect certain features that are difficult to find. To conduct a survey of people who go jogging in a certain park every morning, for example, snowball sampling would be a quick, accurate way to create the sample.

Uses of non-probability sampling

Non-probability sampling is used for the following:

- **Create a hypothesis:** Researchers use the non-probability sampling method to create an assumption when limited to no prior information is available. This method helps with the immediate return of data and builds a base for further research.
- **Exploratory research:** Researchers use this sampling technique widely when conducting qualitative research, pilot studies, or exploratory research.

- **Budget and time constraints:** The non-probability method when there are budget and time constraints, and some preliminary data must be collected. Since the survey design is not rigid, it is easier to pick respondents randomly and have them take the survey or questionnaire.

How do you decide on the type of sampling to use?

For any research, it is essential to choose a sampling method accurately to meet the goals of your study. The effectiveness of your sampling relies on various factors. Here are some steps expert researchers follow to decide the best sampling method.

- Jot down the research goals. Generally, it must be a combination of cost, precision, or accuracy.
- Identify the effective sampling techniques that might potentially achieve the research goals.
- Test each of these methods and examine whether they help achieve your goal.
- Select the method that works best for the research.

Difference between probability sampling and non-probability sampling methods

We have looked at the different types of sampling methods above and their subtypes. To encapsulate the whole discussion, though, the significant differences between probability sampling methods and non-probability sampling methods are as below:

	Probability Sampling Methods	**Non-Probability Sampling Methods**
Definition	Probability Sampling is a sampling technique in which samples from a larger population are chosen using a method based on the theory of probability.	Non-probability sampling is a sampling technique in which the researcher selects samples based on the researcher's subjective judgment rather than random selection.
Alternatively Known as	Random sampling method.	Non-random sampling method
Population selection	The population is selected randomly.	The population is selected arbitrarily.
Nature	The research is conclusive.	The research is exploratory.
Sample	Since there is a method for deciding the sample, the population	Since the sampling method is arbitrary, the population demographics

	demographics are conclusively represented.	representation is almost always skewed.
Time Taken	Takes longer to conduct since the research design defines the selection parameters before the market research study begins.	This type of sampling method is quick since neither the sample nor the selection criteria of the sample are undefined.
Results	This type of sampling is entirely unbiased; hence, the results are also conclusive.	This type of sampling is entirely biased, and hence the results are biased, too, rendering the research speculative.
Hypothesis	In probability sampling, there is an underlying hypothesis before the study begins, and this method aims to prove the hypothesis.	In non-probability sampling, the hypothesis is derived after conducting the research study.