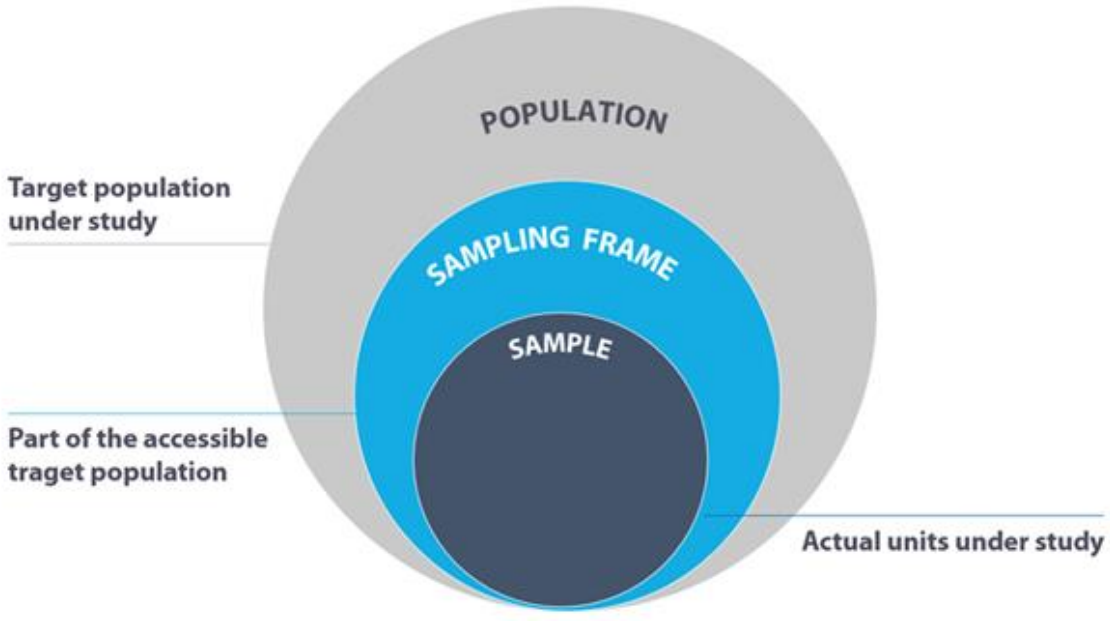
WHAT IS SAMPLING

In sampling, we select a group of individuals from a target population. This group of individuals forms a sample. Why? As the population is large (say, all people in the country), it will not be possible to study each individual in the population. To make it manageable, we select individuals that represent the population. By studying and analysing this sample, we want to characterize the whole population. In machine learning, all the models we build are based on the analysis of the sample. Then it follows, if we do not select the sample properly, the model will not learn properly.

Before we proceed further, let's understand the key terms in sampling — The population, sampling frame, and sample.



POPULATION: Based on the scope of the study, the population includes all possible outcomes.

SAMPLING FRAME: Contains the accessible target population under study. We derive a sample from the sampling frame.

SAMPLE: Subset of a population, selected through various techniques that we will cover in this guide.

ADVANTAGES OF SAMPLING

Sampling brings many advantages in terms of speed and accuracy. While we are inclined to think that studying each individual on the whole population will lead to accuracy, we tend to overlook the many sources of errors that can happen in a study of the whole population. Further, in most cases, it is just not feasible to study the whole population.

A sample can provide accuracy as we will be able to deploy trained field workers on whom we can rely to collect the observations, scientifically monitor the biases and remove them and since we are collecting limited observations, we reduce the possibility of mistakes that come from processing the data. Moreover, the smaller size of the sample means that we can supervise with efficacy and have clean, usable data.

As our analysis will rely on the sample, it's important that we scientifically approach how we go about selecting samples. However, before we go into sample selection methodologies, let's look at the errors that can happen while selecting samples.

ERRORS IN SAMPLE SELECTION

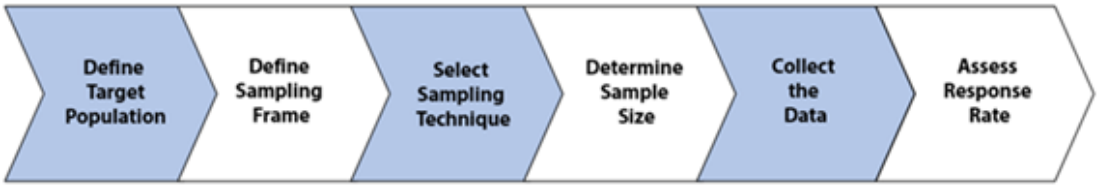
Selecting a sample that closely represents the population is critical to business problem-solving. Here are some of the errors:

● CYCLICAL BUSINESS INDUCED ERRORS — If we are looking at buying behaviour, taking samples around Christmas and Diwali will not be representative of the overall behaviour.

● SPECIFICATION ERROR — If the study is around sales of toys, and we survey the mothers only, that may not be accurate as children influence the buying behaviour.

● SAMPLE FRAME ERROR — This error happens when we select the wrong sub-population. For instance, our study was to understand if the population favours a new policy that has been introduced in India. We survey everyone who speaks English. It may not be accurate as ~90% of the country's population does not speak English.

Let's understand the sampling process



Sampling Process Flow

Define target population: Based on the objective of the study, clearly scope the target population. For instance, if we are studying a regional election, the target population would be all people who are domiciled in the region that are eligible to vote.

DEFINE SAMPLING FRAME: The sampling frame is the approachable members from the overall population. In the above example, the sampling frame would consist of all the people from the population who are in the state and can participate in the study.

SELECT SAMPLING TECHNIQUE: Now that we have the sampling frame in place, we want to select an appropriate sampling technique. We will discuss this in detail in the next section.

DETERMINE SAMPLE SIZE: To ensure that we have an unbiased sample, free from errors and that closely represents the whole population, our sample needs to be of an appropriate size. What is an appropriate size? Well, this is dependent on factors like the complexity of the population under study, the researcher's resources and associated constraints. Also, it's important to keep in mind that not all individuals we approach for the study will respond. Researchers like Bartlett et al. suggest that we should increase the number of individuals we approach initially, by as much as 50%, to factor in the non-response rate.

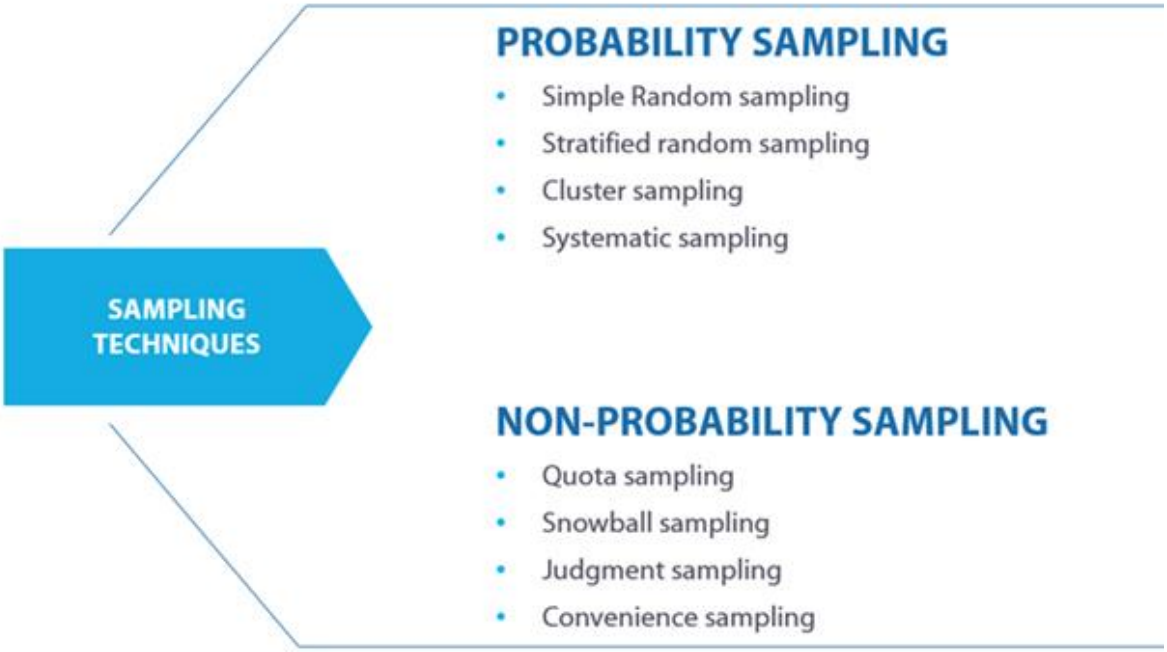
COLLECT THE DATA: Data collection is critical to solving the business case. We should attempt to ensure that we don't have too many empty fields in our data, and we document the reasons in cases where the data is missing. This helps in analysis, as this gives us perspective on how to treat the missing data when we perform analysis.

ASSESS RESPONSE RATE: It is important to closely monitor the response rate to ensure you make timely changes to your sample collection approach and ensure you achieve your determined sample collection.

Now that we have developed an understanding of the overall sampling process, let's deep-dive into the heart of sampling and look at the sampling techniques

POPULAR SAMPLING TECHNIQUES

As we know, taking a subset from the sampling frame forms the act of sampling.



The various ways in which we can select samples can be divided into two types:

PROBABILITY SAMPLING: Some researchers refer to this as random sampling.

NON-PROBABILITY SAMPLING: This is also referred to as non-random sampling.

Whether you decided to go for a probability or a non-probability approach depends on the following factors:

Goal and scope of the study

Data collection methods that are feasible

Duration of the study

Level of precision you wish to have from the results

Design of the sampling frame and viability to maintain the frame

Probability sampling

SIMPLE RANDOM SAMPLING:

Here, as the name suggests, we pick the sample, at random. There is no pattern, and it's a purely random selection. For instance, you wanted to survey vaccination uptake. You could put 100 names of all eligible people in a hat and pull out a few to sample them. For instance, in machine learning, when you split your data into a training set and a test set you use the principle of simple random sampling.

Let's look at the two subtypes of simple random sampling:

SIMPLE RANDOM SAMPLING WITH REPLACEMENT

Here, in a sample size N, you select an element of the population and return it to the population. This implies that each element of the population could theoretically be selected more than once. Each time we select an individual, we have the whole selected population available to select from. Typically, when the population itself is small, we use this technique.

SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT

Here, once you select an individual from the population, you don't return it. With each passing selection, the available population decreases by 1. This also implies that for a sample size N, we repeat the selection process N times. When the population size is large, we go for this without-replacement method of simple random sampling.

STRATIFIED RANDOM SAMPLING

When we have supplemental information available to aid with the sample design, we can consider using stratified random sampling. As the name suggests, we divide the population into strata or groups based on certain characteristics by which we can identify the groups. Now, we select the elements from these groups to create a sample. This way, we can ensure the representation of the overall population. These subgroups are formed based on attributes like a particular age group, gender, occupation. If your population has a lot of variation, you want to use stratified random sampling.

For instance, suppose the government wants feedback on a new education policy they are going to pursue. It will not be sufficient to survey only the stakeholders of government schools, which might be easier to accomplish. The sample would need representation from all strata on which the policy might have implications like private, semi-private, minority, international schools, in addition to government schools.

We have three types of stratified random sampling:

PROPORTIONATE STRATIFIED RANDOM SAMPLING

Here, we divide each stratum in proportion to its representation in the whole population under study. For instance.

Milestone	Strata 1	Strata 2	Strata 3
Population representation	300	700	1000
Individuals in the sample (sampling fraction 10%)	30	70	100

DISPROPORTIONATE STRATIFIED RANDOM SAMPLING

In disproportionate stratified random sampling, we do not go by sampling fraction. The intent here is to ensure that all groups in the population find representation in the sample, irrespective of the proportion of their representation in the population.

OPTIMAL STRATIFIED SAMPLING

In optimal stratified sampling, we form groups in the proportion to the standard deviation of the observations. This is also known as Neyman optimal allocation. The allocation becomes optimal as it considers the size of strata as well as variability within the population.

CLUSTER SAMPLING

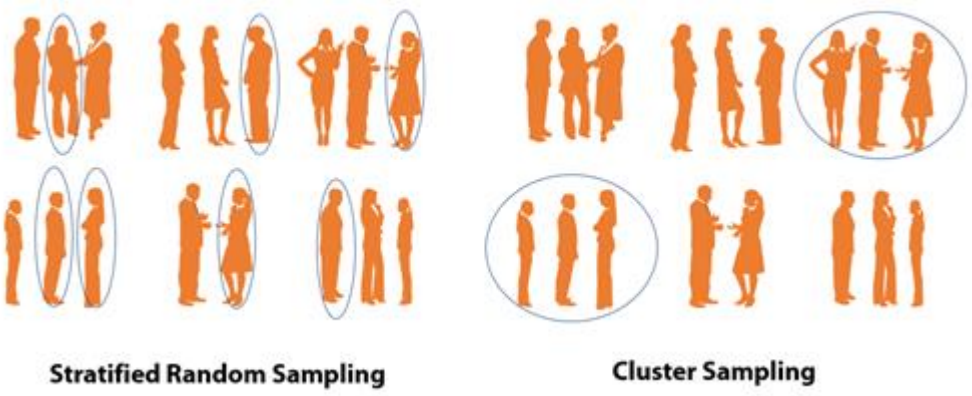
In this, first, we divide the population into clusters. Next, we pick a random sample of these clusters to form our sample. If your subjects are sporadic, spread over a large geographical area, cluster sampling can save your time and be more prudent financially. Here are the stages of cluster sampling:

SAMPLING FRAME – Choose your grouping, like the geographical region in the sampling frame.

Tag each cluster with a number.

Perform a random selection of these clusters.

STRATIFIED RANDOM SAMPLING VS CLUSTER SAMPLING



Before we cover the next probabilistic sampling technique, let's understand the difference between stratified random sampling and cluster sampling. In stratified random sampling, first, we use common characteristics to divide the whole population into strata and next we select elements from each stratum. In clustering, we divide the whole population into clusters and then randomly pick clusters to form a sample and not elements within clusters.

SYSTEMATIC SAMPLING

The last probabilistic technique is systematic sampling. Here, we start at a certain point in the population and keep selecting elements at a regular, fixed interval. In statistical terminology, we essentially select every k-th element, also known as the sampling interval, in the population.

For instance, when you are checking for quality control or auditing, you extensively deploy systematic sampling. It is not feasible to test the quality of all the products or accounting entries. By deploying systematic sampling, you get a strong sample on which you can base your inferences about the population.

NON-PROBABILITY SAMPLING:

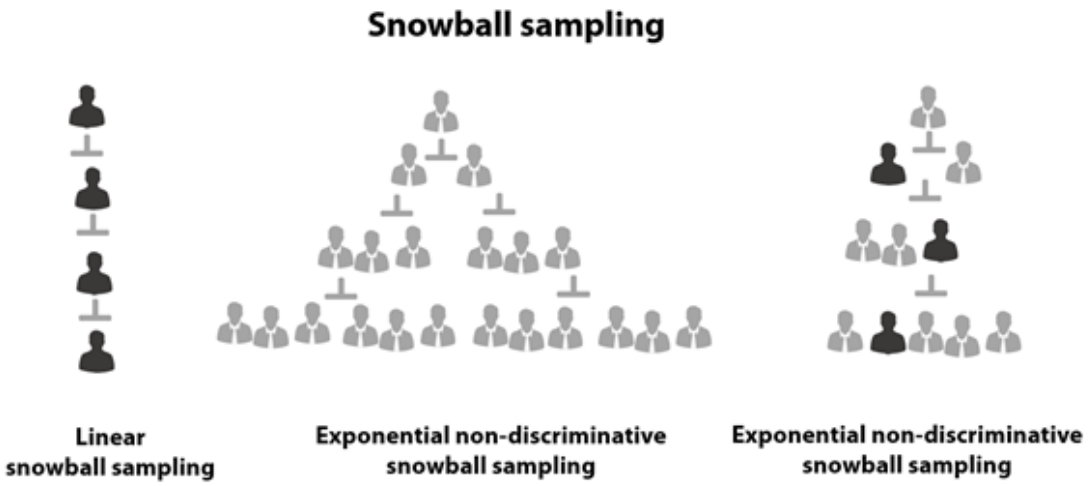
In this kind of sampling, we intentionally do not assign importance to each element in the population having an equal chance of being picked up in the sample.

QUOTA SAMPLING:

In this, we divide the population into quotas that represent the population, and this forms the basis of the elements we select in the sample. This might look similar to random sampling, but the important difference is that we first divide the population into fixed quotas. From these fixed quotas, we select the sample. Quota could be something like all males above 20 or children between 12 and 18 years of age. Using quota sampling saves time and resources and is a quick way to get the study started.

SNOWBALL SAMPLING

This is one of the most interesting non-probabilistic techniques. You first select, at random, members for the sample. Suppose you selected 3 members. Now, these three will suggest more names for the study, and this creates a chain effect. Snowball sampling is useful in cases where it is difficult to locate people, or they do not wish to be identified. For instance, in medical research where you are studying a rare disease, you might find that snowball sampling is the only way you can get to the desired sample size.



There are three sub-categories in snowball sampling:

LINEAR SNOWBALL SAMPLING:

The chain grows linearly. Each member in the sample refers to one more member.

EXPONENTIAL NON-DISCRIMINATIVE SNOWBALL SAMPLING:

One to many relationships. Each member in the study refers to multiple members, and all are selected in the study. As you can imagine, this creates an exponential effect on the size of the sample. As you might have guessed, this may introduce bias into the sampling and researchers have no idea if the sample is representative of the population under study.

EXPONENTIAL DISCRIMINATIVE SNOWBALL SAMPLING

Here, while we will request the member to provide multiple referrals, we will select only one out of these and nullify the remaining referrals. By doing this, researchers attempt to reduce the chances of bias in the sampling technique.

JUDGMENT SAMPLING

Here, the researcher brings forth their qualified opinion and judgment on who should be part of the sample. This is typically used where you want to select experts or highly intellectual individuals in your sample. The best approach is to identify the experts and form the sample.

CONVENIENCE SAMPLING

Here, we prioritize the accessibility of the element above other considerations. The researcher selects the elements based on convenience. This is typically used in the initial phases of the survey, where the researcher intends to gain quick feedback on the design of the survey. It helps to quickly prototype the survey design.

APPLICATION OF SAMPLE

Here are key industry use cases where your knowledge and understanding of sampling techniques would be critical

VALIDATING ASSUMPTIONS THROUGH MARKET RESEARCH

Suppose your company wants to launch a bike ride-sharing service. This service relies on people having smartphones with sufficiently charged batteries and sufficient mobile data. Now you wish to evaluate the market size. To do so, you will have to get a sample that represents people from various income levels, mobility needs, access to data, type of devices, willingness to adopt the bike-sharing model, etc. By doing so, you can arrive at a reasonable estimate of the overall market size of your offering.

QUALITY CONTROL

Extensively used in the manufacturing industry. Suppose you wanted to check the quality of injections produced in a factory. Let's say, the company produces 1 million injections a month. In this case, quality assurance becomes critical. However, it may not be possible to check each injection manufactured. So the company will sample a proportion from each batch and, based on the results, make an inference on the quality of the whole quality produced.

USES IN NEW PRODUCT DEVELOPMENT

Suppose you are working on a new service, say a new bike-sharing service. The typical process with you will follow will involve four steps:

Concept creation & testing

Pilot testing

Beta Testing

Launch

In most of these stages, you would find good use of sampling techniques. Essentially, you want to draw inferences about the whole population by studying the responses of the sample. It becomes vital that you steer clear of any biases and under-representation of the population in your sample.

CONCEPT TESTING: Before starting the development, you might want to know the appeal of such an offering. We can accomplish this by asking a few prospective users of such a service. However, a better approach would be to scientifically go about surveying the people. This way, you can ensure that you are getting representation from all groups, both those that are comfortable with newer modes of transport and those that are apprehensive. You might want to understand who much people are willing to spend on such a service. During interpretation of the findings, we can ensure that each stratum of the society finds representation in the sample and also there are sufficient people in each stratum.

This will lead to meaningful feedback and eliminate the scope of false confidence you might get if you say surveyed only people who are in the pro-sharing economy.

PILOT TESTING: This is the phase just before the beta launch, and you want to factor in as much feedback as possible. Here, using the same principles of testing, you can have useful feedback by ensuring that you have factored in cultural and behavioural patterns from your study, by using the sampling techniques.

CONCLUSION

In this guide, we have covered the sampling process, its techniques and the industry use cases. An understanding of these will serve you well when your judgement is required to solve a business use case. In data science and machine learning, this understanding will help you efficiently do accurate model selection, processing of data, and performing predictive analytics that is more likely to pass the test of reality. It will also allow you to explain any deviation that can be expected from the model you would build based on the sample data.