**Sampling**

In research, sampling refers to the selection of a smaller group of participants from the population of interest. While it would be ideal for the entire population you are researching to take part in your study, logistically this may not be feasible. Therefore, by researching a smaller and representative group obtained from your population of interest, we are able to generalise the findings back to and make inferences about the whole population.

**Sampling frame**

A sampling frame is a list of the elements that make up the universe to be studied, which are known as sampling units. Each of these elements share characteristics that allow them to be identified.

From the sampling frame it is possible to extract the study sample, that is, the population to be investigated. The elements that make it up can be clients, employees, homes, institutions, cities or anything that needs to be studied.

**Types of Sampling:**

Samples can be obtained in a variety of ways, and can include:

**Haphazard:** haphazard sampling refers to the selection of a sample of participants using ‘trial and error’ or ‘hit and miss’ approaches. Haphazard sampling does not rely on any specific criteria or approaches, and often means that your results will be unpredictable and prone to error because there is no systematic approach used.

**Purposive:** purposive sampling is the selection of a non-random sample, whereby participants are specifically chosen because of a particular reason (i.e.they were chosen on purpose because they meet a certain criteria). While purposive samples are advantageous because they focus on a specific group, if the researcher makes an error in judgement about whether an individual meets criterion for inclusion this will influence the results.

**Convenience:** convenience sampling is another form of non-random sampling, where participants are chosen because of convenience-related reasons, such as accessibility or availability. Such samples are advantageous because participants are often readily available and easy to access, however this approach often increases the risk of bias in the results.

While each of the above approaches have their place, a common limitation to all three approaches is that they are quite subjective in nature, and heavily rely on researcher discretion to determine them. Another limitation is that since these approaches are targeted (i.e. non-random), this raises questions about whether we can generalise our findings to the broader population of interest. While how you sample will be based on your specific research question and study design, generally the preferred method of sampling is the probability approach.

### Probability Sampling:

Probability sampling, also referred to as random sampling, is the independent and random selection of participants based on probability theory, in that it is controlled by chance alone. Sampling based on probability is advantageous because it increases the likelihood of obtaining a sample that is more representative of the population you are interested in. For a sample to be genuinely random, each participant drawn from the population of interest must have an equal chance of being selected, and one participant being selected must occur independently of any other participant being selected.

There are several subtypes of probability sampling, and include systematic, simple random, stratified random and cluster samples. We shall use the topic

***“Does pineapple belong on pizza?”***

### Systematic Samples

A systematic sample is a type of probability sampling, however systematic samples are not random. In systematic sampling, a rule for selecting participants is pre-determined and applied. A common form of this is to select every ‘nth’ person to be part of the sample.

For our pizza example, suppose you wanted to select a sample of 25 guests out of 100 who are attending a pizza party, by surveying every 4th guest who arrives. Each guest is allocated a number from 1 to 100, and as they enter you ask every 4th person whether pineapple belongs on pizza. By choosing every 4th person, you obtain a probability sample because 25% of guests have been selected, however this sample is not random because any guest who was not number 4, 8, 12 and so on had zero chance of being selected.

**Simple Random Samples:**

A simple random sample is a subset of a statistical population in which each member of the subset has an equal probability of being chosen. A simple random sample is meant to be an unbiased representation of a group. Simple random selection is often used when the population you are sampling is relatively homogenous, or similar.

**A few take aways**

* A simple random sample takes a small, random portion of the entire population to represent the entire data set, where each member has an equal probability of being chosen.
* Researchers can create a simple random sample using methods like lotteries or random draws.
* A sampling error can occur with a simple random sample if the sample does not end up accurately reflecting the population it is supposed to represent.
* Simple random samples are determined by assigning sequential values to each item within a population, then randomly selecting those values.

**Steps**

### Step 1: Define the Population

The origin of statistical analysis is to determine the population base. This is the group in which you wish to learn more about, confirm a [hypothesis](https://www.investopedia.com/terms/h/hypothesistesting.asp), or determine a statistical outcome. This step is to simply identify what that population base is and to ensure that group will adequately cover the outcome you are trying to solve for.

***Example:***I wish to learn how the stocks of the largest companies in the United States have performed over the past 20 years. My population is the largest companies in the United States as determined by the S&P 500.

### Step 2: Choose Sample Size

Before picking the units within a population, we need to determine how many units to select This sample size may be constrained based on the amount of time, [capital rationing](https://www.investopedia.com/terms/c/capitalrationing.asp), or other resources available to analyze the sample. However, be mindful to pick a sample size large enough to be truly representative of the population. In the example above, there are constrains in analyzing the performance for every stock in the S&P 500, so we only want to analyze a sub-set of this population.

***Example:*** My sample size will be 20 companies from the S&P 500.

### Step 3: Determine Population Units

In our example, the items within the population are easy to determine as they've already been identified for us (i.e. the companies listed within the S&P 500). However, imagine analyzing the students currently enrolled at a [university](https://www.investopedia.com/colleges-and-universities-4689808) or food products being sold at a grocery store. This steps entails crafting the entire list of all items within your population.

### Step 4: Assign Numerical Values

The simple random sample process call for every unit within the population receiving an unrelated numerical value. This is often assigned based on how the data may be filtered. For example, I could assign the numbers 1 to 500 to the companies based on [market cap](https://www.investopedia.com/terms/m/marketcapitalization.asp), alphabetical, or company formation date. How the values are assigned doesn't entirely matter; all that matters is each value is sequential and each value has an equal chance of being selected.

***Example:***I assign the numbers 1 through 500 to the companies in the S&P 500 based on alphabetical order of the current CEO, with the first company receiving the value '1' and the last company receiving the value '500'.

### Step 5: Select Random Values

In step 2, we selected the number of items we wanted to analyze within our population. For the running example, we choose to analyze 20 items. In the fifth step, we randomly select 20 numbers of the values assigned to our variables. In the running example, this is the numbers 1 through 500. There are multiple ways to randomly select these 20 numbers discussed later in this article.

***Example:***Using the random number table, I select the numbers 2, 7, 17, 67, 68, 75, 77, 87, 92, 101, 145, 201, 222, 232, 311, 333, 376, 401, 478, and 489.

### Step 6: Identify Sample

The last step of a simple random sample is the bridge step 4 and step 5. Each of the random variables selected in the prior step corresponds to a item within our population. The sample is selected by identifying which random values were chosen and which population items those values match.

***Example:***My sample consists of the 2nd item in the list of companies alphabetically listed by CEO's last name. My sample also consists of company number 7, 17, 67, etc.

## **Random Sampling Techniques**

* **Random lottery.**Whether by ping-pong ball or slips of paper, each population number receives an equivalent item that is stored in a box or other indistinguishable container. Then, random numbers are selected by pulling or selecting items without view from the container.
* **Physical Methods.** Simple, early methods of random selection may use dice, flipping coins, or spinning wheels. Each outcome is assigned a value or outcome relating to the population.
* **Random numbers from**[**Excel**](https://www.investopedia.com/articles/personal-finance/032415/importance-excel-business.asp)**.** Numbers can be selected in Excel using the =RANDBETWEEN formula. A cell containing =RANDBETWEEN(1,5) will selected a single random number between 1 and 5.

### Stratified Random Samples:

A stratified random sample is a random sample where two or more groups are represented from your population of interest. Stratified random sampling is more often used when the population you are sampling is relatively heterogeneous, or there are notable subgroups present. This involves dividing your population into the smaller groups and then randomly selecting a sample from each - in essence, you are treating it as if there are two populations. Common examples include stratifying by age, sex or ethnicity.

In relation to our pizza example, let’s assume that at our pizza party there are 55 females and 45 males, and you have reason to believe that both sexes will respond differently to the question of whether pineapple belongs on pizza. To be representative, and for a sample of 25%, you determine that you will need to survey 14 females and 11 males. You randomly allocate each female a number from 1 to 55, and each male a number from 1 to 45. Starting with the female group, as before you hop onto the random number generator and generate a number between 1 and 55, and ask the female guest who corresponds with the generated number whether pineapple belongs on pizza. You repeat this 14 times until you have completed your female sample. Once this has been completed, you repeat the entire process for the males. We can then compare the responses of females to males to see if one sex prefers pineapple on pizza more than the other sex. Like with simple random sampling, this example is a probability sample because 25% of guests from each subgroup have been selected, and it is random because there is an equal chance of being selected at random. Stratified random selection was used because the sample was heterogenous, in that there were males and females.

### Random Cluster Sampling:

## What is Cluster Sampling?

In statistics, cluster sampling is a sampling method in which the entire population of the study is divided into externally, homogeneous but internally, heterogeneous groups called clusters. Essentially, each cluster is a mini-representation of the entire [population](https://corporatefinanceinstitute.com/resources/knowledge/other/statistics/).

After identifying the clusters, certain clusters are chosen using simple random sampling while the others remain unrepresented in a study. After selecting the clusters, a researcher must choose the appropriate method to sample the elements from each selected group.

### Primary Sampling Methods

There are primarily two methods of sampling the elements in the cluster sampling method: **one-stage** and **two-stage**.

In one-stage sampling, all elements in each selected cluster are sampled. In two-stage sampling, simple random sampling is applied within each cluster to select a subsample of elements in each cluster.

The cluster method must not be confused with stratified sampling. In stratified sampling, the population is divided into mutually exclusive groups that are externally heterogeneous but internally homogeneous.

For example, in stratified sampling, a researcher may divide the population into two groups: males vs. females. Conversely, in cluster sampling, the clusters are similar to each other but with different internal composition.

### Advantages of Cluster Sampling

The cluster method comes with a number of advantages over simple random sampling and stratified sampling. The advantages include:

#### 1. Requires fewer resources

Since cluster sampling selects only certain groups from the entire population, the method requires fewer resources for the sampling process. Therefore, it is generally cheaper than simple random or stratified sampling as it requires fewer [administrative and travel expenses](https://corporatefinanceinstitute.com/resources/knowledge/accounting/what-is-sga/).

#### 2. More feasible

The division of the entire population into homogenous groups increases the feasibility of the sampling. Additionally, since each cluster represents the entire population, more subjects can be included in the study.

### Disadvantages of Cluster Sampling

Despite its benefits, this method still comes with a few drawbacks, including:

#### 1. **Biased samples**

The method is prone to [biases](https://corporatefinanceinstitute.com/resources/knowledge/other/sample-selection-bias/). If the clusters representing the entire population were formed under a biased opinion, the [inferences](https://stat.duke.edu/~fab2/inference_talk.pdf) about the entire population would be biased as well.

#### 2. **High sampling error**

Generally, the samples drawn using the cluster method are prone to higher sampling error than the samples formed using other sampling methods.