**Sampling Techniques**

So far we have learnt about analysis, interpretation and analysis of statistical data already collected. But how is to collect statistical data? There are broadly two modes of collection of statistical data – through survey or experiment.

**Survey:** When we are ***interested in natural variation*** without exercising any control on factors that may influence it, we conduct what may be called a survey of the phenomena of interest. For example, we may carry out a survey of socio-economic conditions in a section of population; estimating product quality of a process etc.

**Experiments:** When the object is to study the ***variation produced in response to changes purposefully made*** on influencing factors we conduct an experiment, e.g. studying the effect of temperature and pressure on yield of a chemical process.

**Sampling:** Whatever be the mode of data collection, because of many reasons, often it becomes impossible to collect data on the aggregate of items of interest. The reason may be high cost, loss of time, physical barrier etc. Hence we take the help of sampling.

The single most basic assumption in the sampling concept is that if sample items are chosen ***at random*** from the total population, the sample will tend to have the same characteristics as the entire population. Notice the emphasis on random selection. If the sample of size is selected really in a random manner from the aggregate of items, i.e. population, we can place great confidence in the basic assumption.

The aggregate of items in a population may be finite or infinite.

**Sampling bias:** In statistics, **sampling bias** is a bias in which a sample is collected in such a way that some members of the intended population have a lower or higher sampling probability than others. It results in a **biased sample** of a population in which all individuals, or instances, were not equally likely to have been selected.

Some important potential sources of bias that should be taken into account when selecting a sample are:

* People in hard-to-reach groups are omitted
* Selected individuals are replaced with others, for example if they are difficult to contact
* An out-of-date list is used as the sample frame (for example, if it excludes people who have recently moved to an area)

# Methods of sampling from a population

* **Probability sampling:** It is a sampling technique in which samples are chosen from a population using a method based on the theory of probability.
* This sampling method considers every member of the population and forms samples based on a fixed process. Therefore, it start with a complete sampling frame of all eligible individuals from which we select the sample.
* All statistical samplings are basically probability sampling, and in these samplings each member of the population gets a *definite probability* of being included in the sample.
* In this sampling method, the *bias* in the sample derived from a population is negligible.
* This sampling method leads to higher quality data collection as the sample *appropriately represents* the population, i.e. the sample is truly *representative* of the population.
* **Non-probability sampling:** It is a sampling technique in which samples are chosen from a population without taking into consideration the theory of probability.
* In this sampling methods, we do not start with a complete sampling frame, so some individuals have no chance of being selected.
* Consequently, we cannot estimate the ***effect of sampling error***and there is a significant risk of ending up with a ***non-representative*** sample.
* This sampling method tends to be ***cheaper and more convenient*** and therefore, they are sometimes used for exploratory research *(****pilot survey) and hypothesis generation***.
* This sampling method is ***not recognized as statistical samplings*** and therefore, should be avoided.

**Four types of non-probability sampling techniques**

* Convenience sampling
* Quota sampling
* Judgement (or Purposive) Sampling
* Snowball sampling

**Convenience sampling:** This method is dependent on the ease of access to subjects such as surveying customers at a mall or passers-by on a busy street. It is usually termed as [convenience sampling](https://www.questionpro.com/blog/convenience-sampling/), because of the researcher’s ease of carrying it out and getting in touch with the subjects. Researchers have nearly no authority to select the sample elements, and it’s purely done based on proximity and not representativeness. This non-probability sampling method is used when there is time and cost limitations in collecting feedback. In situations where there are resource limitations such as the initial stages of research, convenience sampling is used.

**Quota sampling**

This method of sampling is often used by market researchers. Interviewers are given a quota of subjects of a specified type to attempt to interview. For example, an interviewer might be told to go out and select 20 adult men, 20 adult women, 10 teenage girls and 10 teenage boys so that they could interview them about their television viewing. Ideally the quotas chosen would proportionally represent the characteristics of the underlying population.

Whilst this has the advantage of being relatively straightforward and potentially representative, the chosen sample may not be representative of other characteristics that weren’t considered.

**Judgement (or Purposive) Sampling**

This is also known as selective or subjective sampling. This technique relies on the judgement of the researcher when choosing who to ask to participate. Researchers purely consider the purpose of the study, along with the understanding of the target audience. For instance, when researchers want to understand the thought process of people interested in studying for their master’s degree. The selection criteria will be: “Are you interested in doing your masters in …?” and those who respond with a “No” are excluded from the sample. This approach is often used by the media when canvassing the public for opinions and in qualitative research.

**Snowball sampling**

This method is commonly used in social sciences when investigating hard-to-reach groups. Existing subjects are asked to nominate further subjects known to them, so the sample increases in size like a rolling snowball. For example, when carrying out a survey of risk behaviours amongst intravenous drug users, participants may be asked to nominate other users to be interviewed.

Snowball sampling can be effective when a sampling frame is difficult to identify. For example, it will be extremely challenging to survey shelterless people or illegal immigrants. Researchers also implement this sampling method in situations where the topic is highly sensitive and not openly discussed—for example, surveys to gather information about HIV Aids. Not many victims will readily respond to the questions. Still, researchers can contact people they might know or volunteers associated with the cause to get in touch with the victims and collect information.

**Four types of probability sampling techniques**

* Simple random sampling (Random sampling)
* Cluster sampling
* Stratified random sampling
* Systematic sampling

**Simple random sampling:** One of the best probability sampling techniques that helps in saving time and resources, is the [Simple Random Sampling](https://www.questionpro.com/blog/simple-random-sampling/) method. It is a reliable method of obtaining information where every single member of a population is chosen randomly, merely by chance. Each individual has the same probability of being chosen to be a part of a sample.

One way of obtaining a simple random sample is to give each individual in a population a number, and then use a table of **random numbers** to decide which individuals to include. For example, if we have a sampling frame of 1000 individuals, labelled 0 to 999, use groups of three digits from the random number table to pick the sample. So, if the first three numbers from the random number table were 094, select the individual labelled “94”, and so on.

**Cluster sampling:** In a clustered sample, subgroups of the population are used as the sampling unit, rather than individuals. The population is divided into subgroups, known as clusters, which are randomly selected to be included in the study. Each cluster must be representative of the population. In single-stage cluster sampling, all members of the chosen clusters are then included in the study. In two-stage cluster sampling, a selection of individuals from each cluster is then randomly selected for inclusion.

The General Household survey is a good example of a (one-stage) cluster sample. All members of the selected households (clusters) are included in the survey.

**Stratified random sampling:** In this method, the population is first divided into subgroups (or strata) who all share a similar characteristic. It is used when we might reasonably expect the measurement of interest to vary between the different subgroups, and we want to ensure representation from all the subgroups.

For example, a researcher looking to analyze the characteristics of people belonging to different annual income divisions, will create strata (groups) according to the annual family income, e.g. less than ₹5,000, ₹5,000 – ₹10,000, ₹11,000 -₹15,000, ₹16,000 -₹20,000, etc. By doing this, the researcher concludes the characteristics of people belonging to different income groups. In a study of stroke outcomes, we may stratify the population by sex, to ensure equal representation of men and women.

**Systematic sampling:** Individuals are selected at regular intervals from the sampling frame. The intervals are chosen to ensure an adequate sample size. If we need a sample size  from a population of size , we should select every  individual for the sample.

For example, if we wanted a sample size of 100 from a population of 1000, we would select every 1000/100 = 10th member of the sampling frame.

**Importance of probability sample:** Using the probability sampling method, the bias in the sample derived from a population is negligible to non-existent. Probability sampling leads to higher quality [data collection](https://www.questionpro.com/blog/data-collection/) as the sample appropriately represents the population.

**Sampling frame**

A **sampling frame** is a list of all the items in the [population](https://www.statisticshowto.com/what-is-a-population/).**It’s a complete list of everyone or everything we want to study.** The difference between a population and a sampling frame is that the population is general and the frame is specific. For example, the population could be “Students of ISI”. The frame would name all of those students of ISI.

## Qualities of a good Sampling Frame

A good sampling frame for a project on living conditions would:

* Include all individuals in the [target population](https://www.statisticshowto.com/target-population-definition-examples/).
* Exclude all individuals **not**in the target population.
* Includes accurate information that can be used to contact selected individuals.

Other general factors that we would want to make sure we have:

* A unique identifier for each member. This could be a simple numerical identifier (e.g. from 1 to 1000).
* A logical organization to the list. For example, put them in alphabetical order.
* Up to date information. This may need to be periodically checked (e.g. for address changes).

## Sampling Frame vs. Sample Space

A *sampling frame* is a list of things from which we draw a sample. A sample space is a list of all possible outcomes for an experiment. For example, we might have a sampling frame of names of people in a certain town for a survey we are going to be conducting on family size. The [sample space](https://www.statisticshowto.com/sample-space/) is all possible outcomes from our survey: 1 person, 2 people, 3 people…10 or more.

**Random numbers**

Random numbers are numbers that occur in a sequence such that two conditions are met:

(1) The values are uniformly distributed over a defined interval or set, and

(2) It is impossible to predict future values based on past or present ones.

**Random Number Generators**

Random number generators can be of two types:

* Hardware based random number generators
* Pseudo-random number generators

**Hardware based random number generators**

* It can involve the use of a dice, a coin for flipping, or many other devices.
* A common scheme for hardware based random number generation is the selection of numbered ping-pong balls from a set of 10, one bearing each digit, as the balls are blown about in a container by forced-air jets. After each number is selected, the ball with that number is returned to the set, the balls are allowed to blow around for a minute or two, and then another ball is allowed to escape.
* Number picker wheel or random number wheel is also a hardware based means for generation of random numbers.

**Pseudo-random Number Generators**

* Several algorithms have been devised that supposedly generate random numbers. Digits or numbers generated in this manner are called ***pseudo-random***. A pseudo-random number generator is an algorithm for generating a sequence of numbers whose properties approximate the properties of sequences of random numbers.
* The problem with these methods is that they violate condition (2) in the definition of randomness. The existence of any number-generation algorithm produces future values based on past and/or current ones.
* Computer based random number generators are almost always pseudorandom number generators.

**Example**: ***Middle-square method***

To generate a sequence of *n*-digit pseudorandom numbers, an *n*-digit starting value is created and squared, producing a 2*n*-digit number. If the result has fewer than 2*n* digits, leading zeros are added to compensate. The middle *n* digits of the result would be the next number in the sequence and returned as the result. This process is then repeated to generate more numbers.

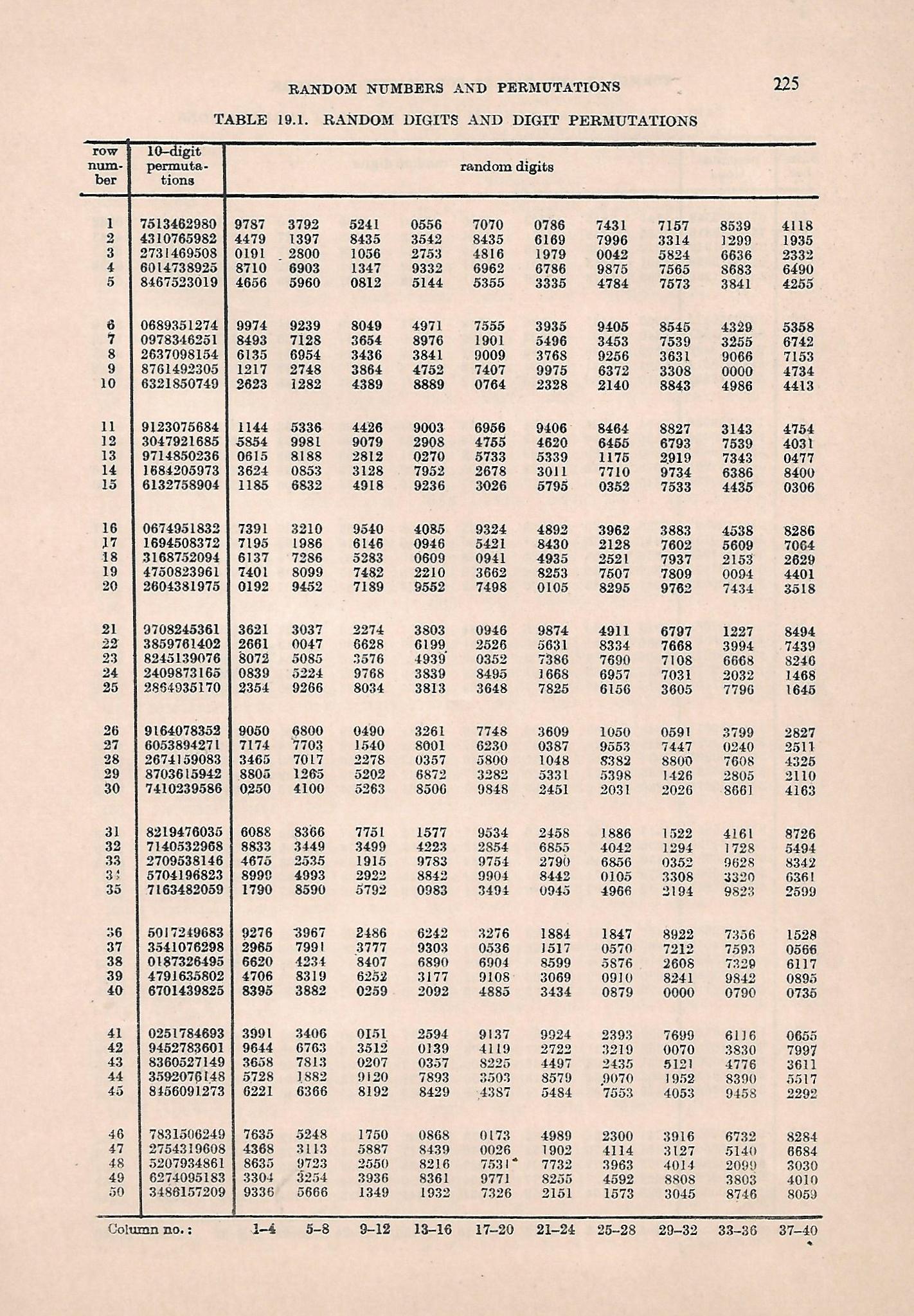
**Random Number Table**

The lottery method is a clumsy physical process for choosing random samples, particularly if the size of population is large. Often it is convenient to use a ready-made table of random numbers. ***A random number table is a table of digits. The digit given in each position in the table was originally chosen randomly from the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 by a random process in which each digit is equally likely to be chosen.*** A possible mechanism for doing this is the lottery method with replacement using ten labeled balls. The digits are often grouped in four or five.

The first such table was published by [Tippett](https://en.wikipedia.org/wiki/L.H.C._Tippett" \o "L.H.C. Tippett) in 1927, and since then a number of other such tables were developed. The first "testing" of random numbers for statistical randomness was developed by Kendall and Smith in the late 1930s, and was based upon looking for certain types of probabilistic expectations in a given sequence. The simplest test looked to make sure that roughly equal numbers of 1s, 2s, 3s, etc. were present; more complicated tests looked for the number of digits between successive 0s and compared the total counts with their expected probabilities.

**Many statistics and research books contain random number tables. We discuss here the usage of random number table taking into consideration the ‘Random Number Table’ available in the ‘Formulae and Tables for Statistical Work’ published by Rao, Mitra, Matthai and Ramamurthy. This book containing various tables for statistical work is popularly known as RMMR tables.**

The first page of the random numbers in RMMR tables is shown below:



It may be noted that each row of digits in the random number table contains a serial number of row and a random permutation of numbers 0, 1, 2,....,9 followed by 40 random digits in 40 columns arranged in sets of 4. The serial numbers of the columns of random digits are indicated in the bottom line of each page so that each random digit can be identified by a row number and a column number. ***There are altogether 5000 four digited random numbers.*** They have been compiled from a number of existing random number tables. The random numbers so compiled have been examined through standard tests of randomness.

In using the table, we need a starting point identified by a row and a column. ***There are no set rules for the choice of a starting point except that no preference is shown to particular page, row or column and the choice is made without prior inspection of the numbers themselves.***

**Different methods for simple random sampling from a list using random number table**

* 1. **A straightforward method:** Suppose we have to sample 5 households from a list of 23, serially numbered 0, 1, ..., 22.

Locate a starting point of random digits and consider two adjacent columns. Read two digited numbers either upwards or downwards or diagonally and record the first five numbers that lie in 0-22. If sampling is without replacement continue reading till five distinct numbers are obtained.

Suppose we start from row 11 and read downwards the two digited numbers in columns 3 and 4. Then the selected households are 15, 1, 21, 5, 20.

* 1. **The method of inflated range:** In the above method, we have to reject all numbers greater than 22, which on an average amounts to 77% of the numbers read. To reduce the number of rejections, consider the range of numbers from 0 to -1, where is chosen such that is nearest to, but does not exceed a power of 10.

In the present example, gives the range 0-91. Choosing two columns as before select the first five two digited numbers in the range 0-91.

Each number chosen is, then, replaced by the remainder after dividing by 23 to obtain a number in the range 0-22. Thus using the same starting point as in (i) above the numbers are 44, 54, 15, 24, 85 which give the sample 21, 8, 15, 1, 16.

Alternatively, when is small as the present example the number chosen could be divided by and the quotient taken as the number finally selected. Thus in the example considered above, the numbers 44, 54, 15, 24, 85 on division by , lead to the sample 11, 13, 3, 6, 21.

* 1. **Independent choice of the first digit:** The method of inflated range reduces the rejection of random numbers at the expense of a tedious operation of repeated division by a given number. An alternative method due to Matthai is as follows:

To select five numbers at random from 0-383, locate a starting point and record two digited numbers (one less than the number of digits in the given number). To each of these numbers prefix a digit at random from 0 to 3. This could be done, for example, by considering the first number from among 0-3 in the random permutation that appear in the same row. A three digited number, so obtained, is rejected if it exceeds 383.

Thus with the columns 3 and 4 from row 11 as the starting point and reading downwards the numbers selected are as follows: 144, 354, 115, 124185 where in, the digits underlined are prefixed as indicated.