BUSINESS STATISTICS

UNIT-IV

SAMPLING CONCEPT

When you conduct research about a group of people, it is rarely possible to collect data from every person in that group. Instead, you select a **sample**. The sample is the group of individuals who will actually participate in the research.

To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. This is called a **sampling method**. There are two primary types of sampling methods that you can use in your research:

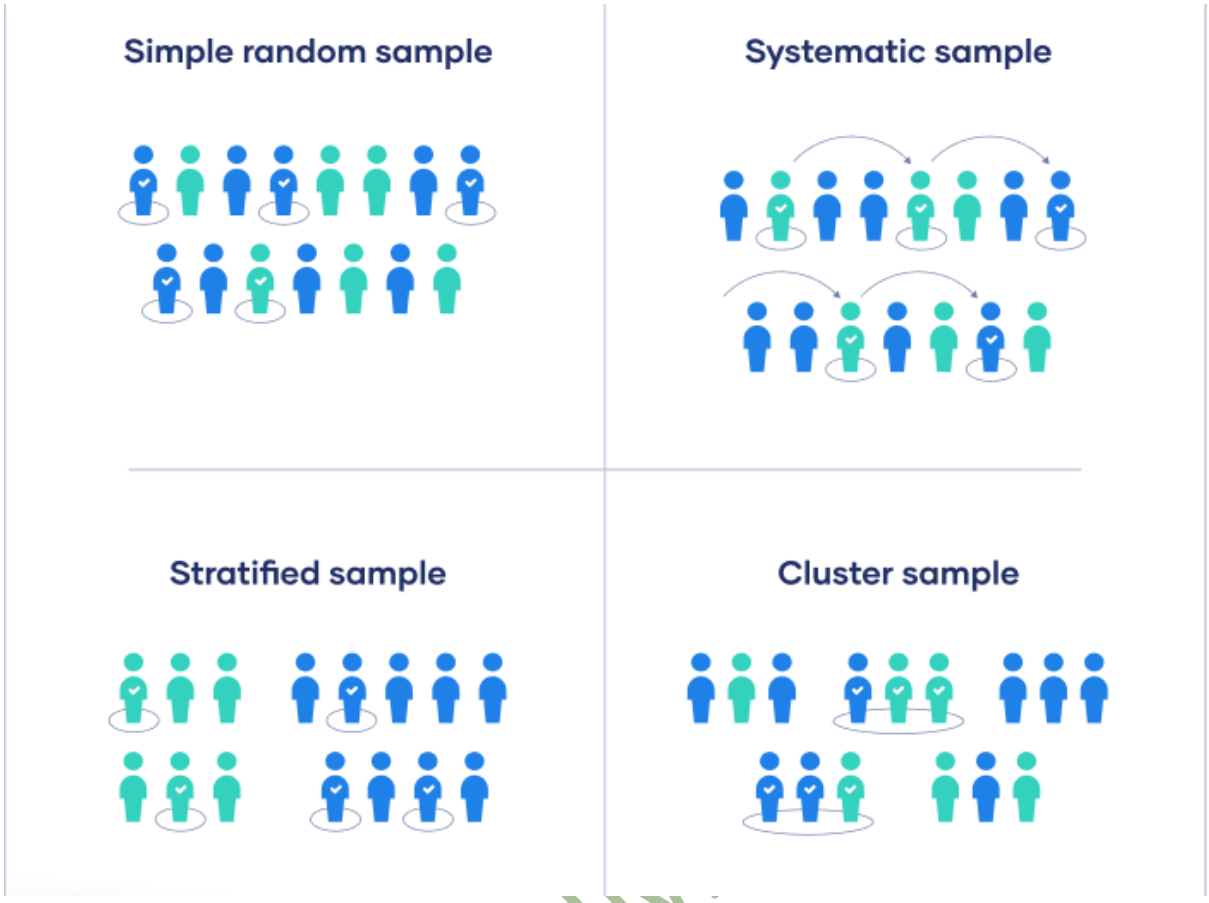
- **Probability sampling** involves random selection, allowing you to make strong statistical inferences about the whole group.
- **Non-probability sampling** involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

Probability sampling methods

Probability sampling means that every member of the population has a chance of being selected. It is mainly used in quantitative research. If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.

There are four main types of probability sample.

DR. PANKAJ KUMAR SHARMA



DR. PANKAJ KUM

1. Simple random sampling

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

Example: Simple random sampling- You want to select a simple random sample of 1000 employees of a social media marketing company. You assign a number to every employee in the company database from 1 to 1000 and use a random number generator to select 100 numbers.

2. Systematic sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Example: Systematic sampling- All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6

onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

3. Stratified sampling

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

Example: Stratified sampling- The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

4. Cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It is difficult to guarantee that the sampled clusters are really representative of the whole population.

Example: Cluster sampling- The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You do not have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.

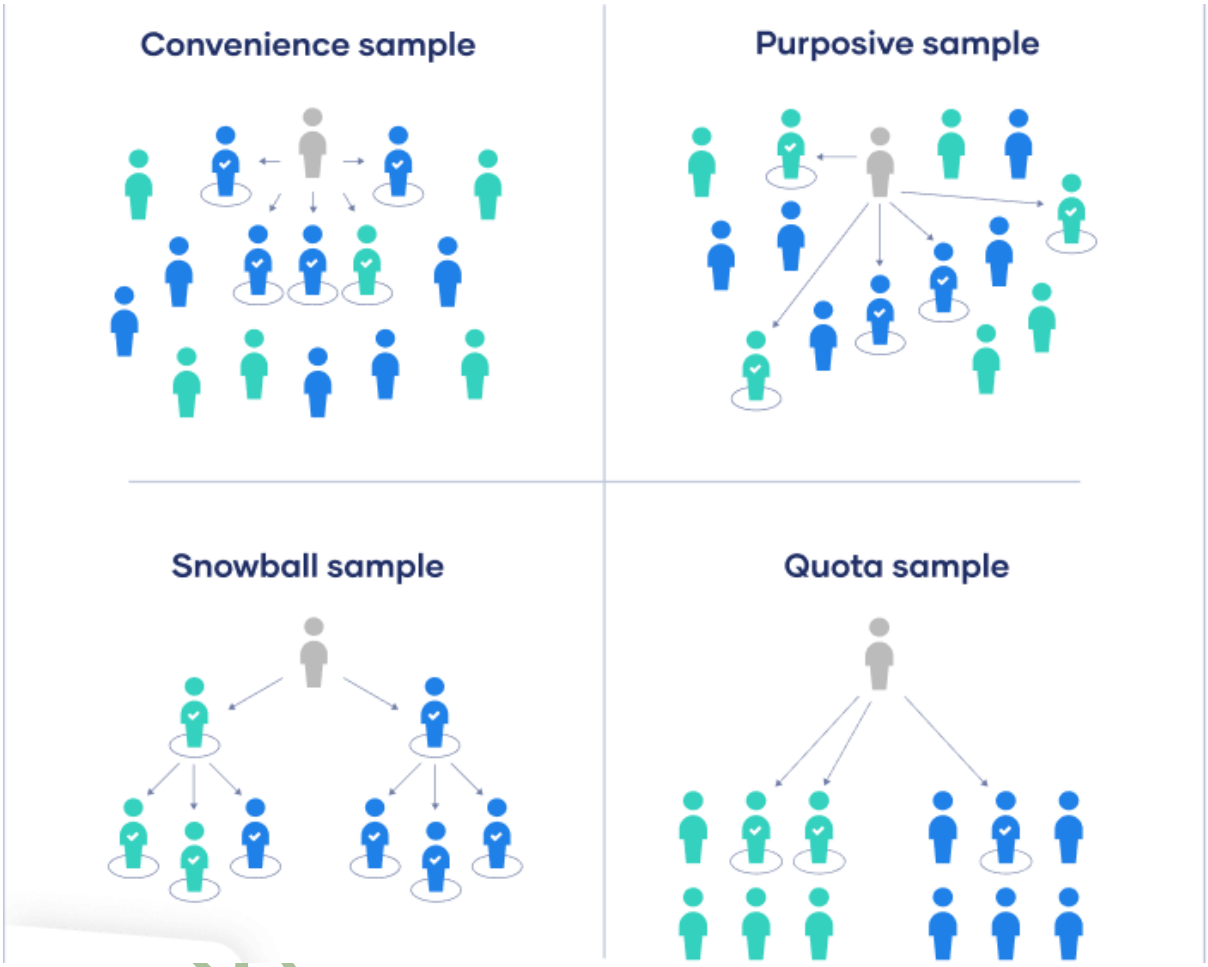
Non-probability sampling methods

In a non-probability sample, individuals are selected based on non-random criteria, and not every individual has a chance of being included.

DR. PANKAJ KUMAR SHARMA

This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias. That means the inferences you can make about the population are weaker than with probability samples, and your conclusions may be more limited. If you use a non-probability sample, you should still aim to make it as representative of the population as possible.

Non-probability sampling techniques are often used in exploratory and qualitative research. In these types of research, the aim is not to test a hypothesis about a broad population, but to develop an initial understanding of a small or under-researched population.



1. Convenience sampling

A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can't produce generalizable results. Convenience samples are at risk for both sampling bias and selection bias.

Example: Convenience sampling- You are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.

2. Quota Sampling

Quota Sampling In quota sampling the interviewers are interested to interview a specified number of persons from each category. The required numbers of elements from each category are determined in the office ahead of time according to the number of elements in each category. Thus, an interviewer would need to contact a specified number of men and specified number of women, from different age categories from different religious or social groups etc. The basis purpose of quota sampling is the selection of a sample that no true replace of the population about which one wants to generalize.

3. Purposive sampling

This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion. Always make sure to describe your inclusion and exclusion criteria and beware of observer bias affecting your arguments.

Example: Purposive sampling- You want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student services.

4. Snowball sampling

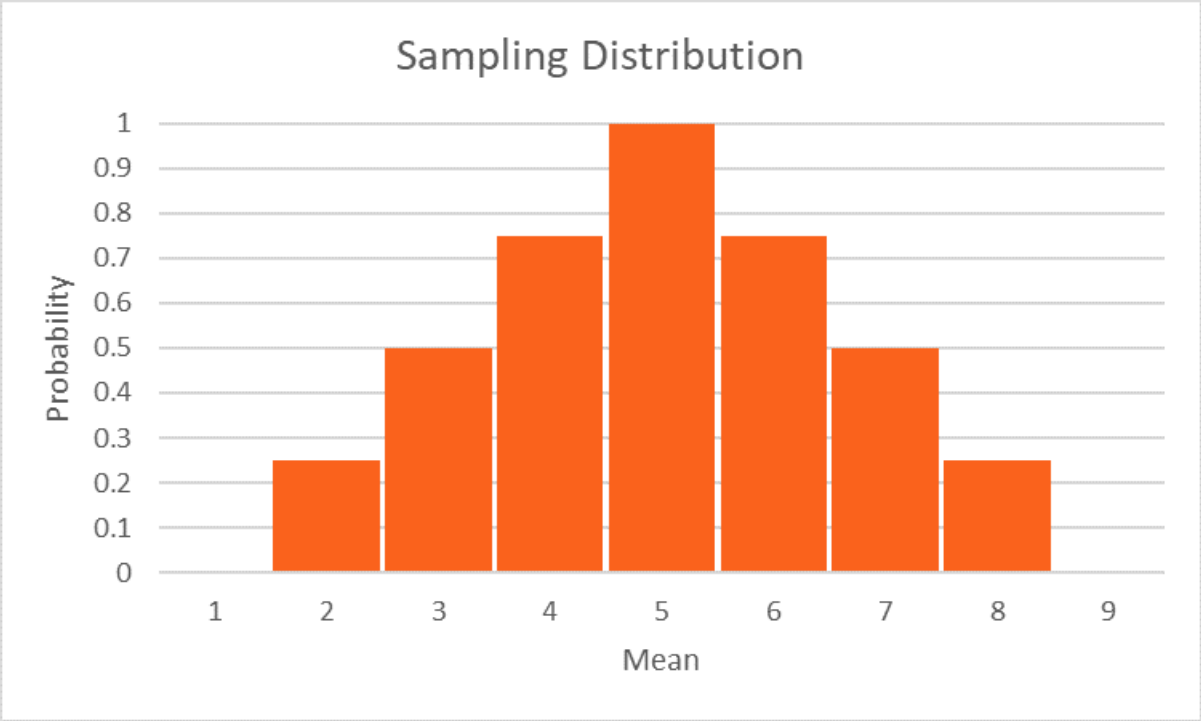
If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to "snowballs" as you get in contact with more people. The downside here is also representativeness, as you have no way of knowing how representative your sample is due to the reliance on participants recruiting others. This can lead to sampling bias.

Example: Snowball sampling- You are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling is not possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.

SAMPLING DISTRIBUTION

A sampling distribution refers to a probability distribution of a statistic that comes from choosing random samples of a given population. Also known as a finite-sample distribution, it represents the distribution of frequencies on how spread apart various outcomes will be for a specific population.

DR. PANKAJ KUMAR SHARMA



The sampling distribution depends on multiple factors – the statistic, sample size, sampling process, and the overall population. It is used to help calculate statistics such as means, ranges, variances, and standard deviations for the given sample.

Types of Sampling Distribution

1. Sampling distribution of mean

As shown from the example above, you can calculate the mean of every sample group chosen from the population and plot out all the data points. The graph will show a normal distribution, and the centre will be the mean of the sampling distribution, which is the mean of the entire population.

2. Sampling distribution of proportion

It gives you information about proportions in a population. You would select samples from the population and get the sample proportion. The mean of all the sample proportions that you calculate from each sample group would become the proportion of the entire population.

Example

Suppose you want to find the average height of children at the age of 10 from each continent. You take random samples of 100 children from each continent, and you compute the mean for each sample group.

For example, in South America, you randomly select data about the heights of 10-year-old children, and you calculate the mean for 100 of the children. You also randomly select data from North America and calculate the mean height for one hundred 10-year-old children.

As you continue to find the average heights for each sample group of children from each continent, you can calculate the mean of the sampling distribution by finding the mean of all

DR. PANKAJ KUMAR SHARMA

the average heights of each sample group. Not only can it be computed for the mean, but it can also be calculated for other statistics such as standard deviation and variance.

Central Limit Theorem

The central limit theorem helps in constructing the sampling distribution of the mean. The theorem is the idea of how the shape of the sampling distribution will be normalized as the sample size increases. In other words, plotting the data that you get will result closer to the shape of a bell curve the more sample groups you use.

The more sample groups you use, the less variable the means will be for the sample groups. When the sample size increases, the standard error decreases. Therefore, the centre of the sampling distribution is fairly close to the actual mean of the population.

STANDARD ERROR

The standard error of the mean, or simply **standard error**, indicates how different the population mean is likely to be from a sample mean. It tells you how much the sample mean would vary if you were to repeat a study using new samples from within a single population.

The standard error of the mean (SE or SEM) is the most commonly reported type of standard error. But you can also find the standard error for other statistics, like medians or proportions. The standard error is a common measure of sampling error—the difference between a population parameter and a sample statistic.

Standard error formula

The standard error of the mean is calculated using the standard deviation and the sample size.

From the formula, you will see that the sample size is inversely proportional to the standard error. This means that the larger the sample, the smaller the standard error, because the sample statistic will be closer to approaching the population parameter.

Different formulas are used depending on whether the population standard deviation is known. These formulas work for samples with more than 20 elements ($n > 20$).

When population parameters are known

When the population standard deviation is known, you can use it in the below formula to calculate standard error precisely.

Formula	Explanation
$SE = \frac{\sigma}{\sqrt{n}}$	• SE is standard error • σ is population standard deviation • n is the number of elements in the sample

When population parameters are unknown

DR. PANKAJ KUMAR SHARMA

When the population standard deviation is unknown, you can use the below formula to only estimate standard error. This formula takes the sample standard deviation as a point estimate for the population standard deviation.

Formula	Explanation
$SE = \frac{s}{\sqrt{n}}$	• SE is standard error • s is sample standard deviation • n is the number of elements in the sample

Example: Using the standard error formula- To estimate the standard error for math SAT scores, you follow two steps.

First, find the square root of your sample size (*n*).

Formula	Calculation
$\sqrt{n}$	$n = 200$ $\sqrt{n} = \sqrt{200} = 14.1$

Next, divide the sample standard deviation by the number you found in step one.

Formula	Calculation
$SE = \frac{s}{\sqrt{n}}$	$s = 180$ $\sqrt{n} = 14.1$ $\frac{s}{\sqrt{n}} = \frac{180}{14.1} = 12.8$

The standard error of math SAT scores is 12.8.

Types of Estimation

Estimators are two different types

- Point Estimates
- Interval Estimates

Point Estimates

A **point estimate** is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean.

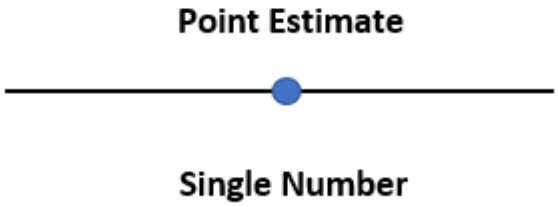
A point estimate is a sample statistic calculated using the sample data to estimate the most likely value of the corresponding unknown population parameter. In other words, we derive the point estimate from a single value in the sample, and we use it to estimate the population value.

DR. PANKAJ KUMAR SHARMA

For instance, if we use a value of $\bar{x}$ to estimate the mean μ of a population.

$\bar{x} = \Sigma x/n$

For example, 62 is the average ($\bar{x}$) marks achieved by a sample of 15 students randomly collected from a class of 150 students is considered to be the mean marks of the entire class. Since it is in the single numeric form, it is a point estimator.



The basic drawback of point estimates is that no information is available regarding their reliability. In fact, the probability that a single sample statistic is equal to the population parameter is very less.

Take a sample, find $\bar{x}$. $\bar{x}$ is a close approximation of μ. But, depending on the size of your sample that may not be a good point estimate. s is a good approximation of σ. So, if we want stronger confidence in what range our estimate lies, we need to do a confidence interval.

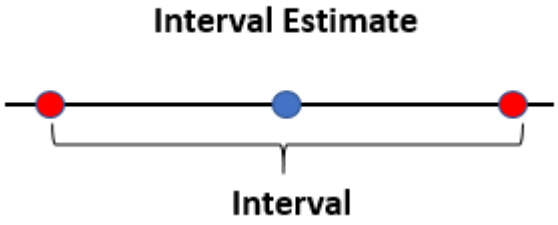
Interval Estimates

An **interval estimate** gives you a range of values where the parameter is expected to lie. A confidence interval is the most common type of interval estimate.

A confidence interval estimate is a range of values constructed from sample data so that the population parameter is likely to occur within the range at a specified probability. Accordingly, the specified probability is the level of confidence.

- Broader and probably more accurate than a point estimate
- Used with inferential statistics to develop a confidence interval – where we believe with a certain degree of confidence that the population parameter lies.
- Any parameter estimate that is based on a sample statistic has some amount of sampling error.

In statistics, interval estimation is the use of sample data to calculate an interval of possible values of an unknown population parameter.



DR. PANKAJ KUMAR SHARMA