Ph. D. Course Work-2023-24

Research Methodology (UNIT-II)

Dr. Prabal Pratap Singh (9889159477)

(Department of Physics)

CHHATRAPATI SHAHU JI MAHARAJ UNIVERSITY

KANPUR-208024 (UP) INDIA

prbalpratapsingh@csjmu.ac.in

SAMPLING ERROR

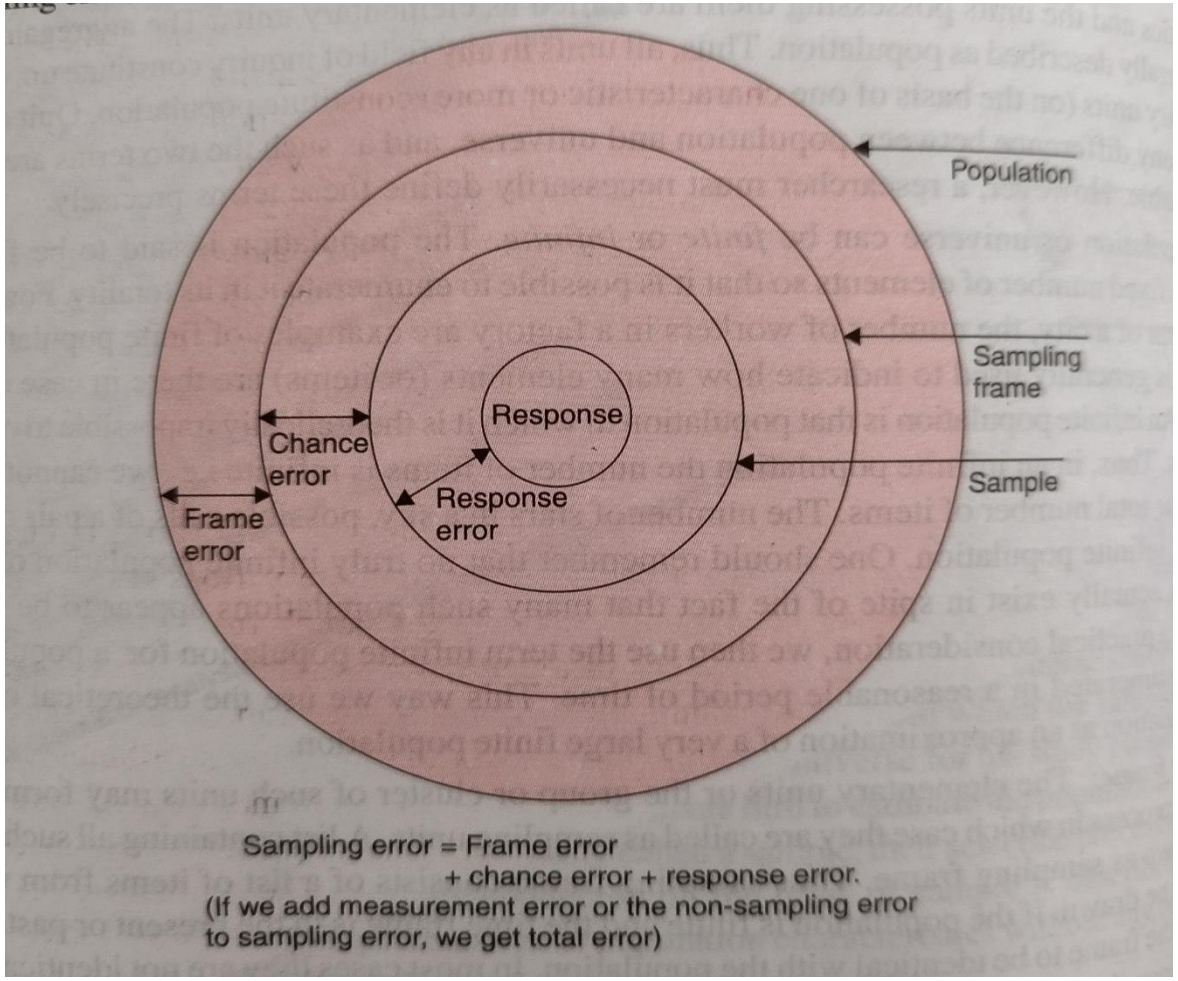
Outlines:-

- What is sampling error?
- Types of Sampling errors
- Eliminating Sampling errors
- Example of Sampling errors
- Sampling Errors Vs Non-sampling Error
- How do you find the sampling error?
- Non-sampling error
- Sampling and non sampling errors
- Important Sampling distribution
- Sampling theory
- Standard Error

SAMPLING ERROR

Sample surveys do imply the study of a small portion of the population and as such there would naturally be a certain amount of inaccuracy in the information collected. This inaccuracy may be known as sampling error or error variance.

- A sampling error occurs when the sample used in the study is not representative of the whole population.
- Sampling is an analysis performed by selecting a number of observations from a large population.
- Even randomized samples will have some degree of sampling error because a sample is only an approximation of population from which it is drawn.
- The diffusion of sampling errors can be reduced by increasing the sample size.



If we add non-sampling error to sampling error, we get total error.

SAMPLING ERROR

- Sampling error occurs randomly and are equally likely to be in either direction.
- The magnitude of the sampling error depend upon the nature of the universe; the more homogeneous the universe, the smaller the sample error.
- Sampling error is inversely related to the size of the sample.
- A measure of the random sampling error can be calculated for a given sample design and size and this measure is called precision(purity) of the sampling plan.
- The sampling error formula is used to calculate the overall sampling error in statistical analysis. This sampling error is calculated by dividing the standard deviation of the population by the square root of the size of sample and then multiplying the result with the Z-score value, which is based on the confidence interval.

Types of Sampling error

There are different categories of sampling errors.

- 1- Population- specific error

  Population- specific error occurs when a researcher doesn't understand who to survey.

- 2- Selection error

  Selection error occurs when the survey is self-selected, or when only those participants who are interested in the survey respond to the questions. Researchers can attempt to overcome selection error by finding ways to encourage participation.

- 3- Sample frame error

  A sample frame error occurs when a sample is selected from the wrong population data.

- 4- Non-response error

  Non-response error occurs when a useful response is not obtain from the surveys.

ELIMINATING SAMPLING ERRORS

The spreading of sampling errors can be reduced by increasing the sample size. As the sample size increases, the sample gets closer to the actual population, which decreases the potential for deviations fro the actual population.

Consider that the average of a sample of 10 varies more than the average of a sample of 100. Steps can also be taken to ensure that the sample adequately represents the entire population.

Researchers might attempt to reduce sampling errors by replicating their study. This could be accomplished by taking the same measurements repeatedly, using more than one subject or multiple groups or by undertaking multiple studies.

Random sampling is an additional way to minimize the occurrence of sampling errors. Random sampling establishes a systematic approach to selecting the sample.

For example:- Rather than choosing participants to be interviewed haphazardly, a researcher might choose those whose names appear first, 10^{th} , 20^{th} , 30^{th} , 40^{th} and so on, on the list.

EXAMPLES OF SAMPLING ERRORS

- Assume that XYZ Company provides a subscription-based service that allows consumers to pay a monthly fee to stream videos and other types of programming via an Internet connection.
- The firm wants to survey homeowners who watch at least 10 hours of programming via the Internet per week and that pay for an existing video streaming service. XYZ wants to determine what percentage of the population is interested in a lower-priced subscription service. If XYZ does not think carefully about the sampling process, several types of sampling errors may occur.

EXAMPLES OF SAMPLING ERRORS

- A population specification error would occur if XYZ Company does not understand the specific types of consumers who should be included in the sample. For example, if XYZ creates a population of people between the ages of 15 and 25 years old, many of those consumers do not make the purchasing decision about a video streaming service because they may not work full-time. On the other hand, if XYZ put together a sample of working adults who make purchase decisions, the consumers in this group may not watch 10 hours of video programming each week.

- Selection error also causes distortions in the results of a sample. A common example is a survey that only relies on a small portion of people who immediately respond. If XYZ makes an effort to follow up with consumers who don't initially respond, the results of the survey may change. Furthermore, if XYZ excludes consumers who don't respond right away, the sample results may not reflect the preferences of the entire population.

- **Sampling Error vs. Non-sampling Error**

- There are different types of errors that can occur when gathering statistical data. Sampling errors are the seemingly random differences between the characteristics of a sample population and those of the general population. Sampling errors arise because sample sizes are inevitably limited. (It is impossible to sample an entire population in a survey or a census.)
- A sampling error can result even when no mistakes of any kind are made; sampling errors occur because no sample will ever perfectly match the data in the universe from which the sample is taken.
- Company XYZ will also want to avoid non sampling errors. Non-sampling errors are errors that result during data collection and cause the data to differ from the true values. Non-sampling errors are caused by human error, such as a mistake made in the survey process.
- If one group of consumers only watches five hours of video programming a week and is included in the survey, that decision is a non-sampling error. Asking questions that are biased is another type of error.

Why Is Sampling Error Important?

Being aware of the presence of sampling errors is important because it can be an indicator of the level of confidence that can be placed in the results. Sampling error is also important in the context of a discussion about how much research results can vary.

How do you find the sampling error?

In survey research, sampling errors occur because all samples are representative samples: a smaller group that stands in for the whole of your research population. It's impossible to survey the entire group of people you'd like to reach.

It’s not usually possible to quantify the degree of sampling error in a study since it's impossible to collect the relevant data from the entire population you are studying. This is why researchers collect representative samples (and representative samples are the reason why there are sampling errors).

Non-Sampling Error

Non-Sampling Error is an umbrella term which comprises of all the errors, other than the sampling error. They arise due to a number of reasons, i.e. error in problem definition, questionnaire design, approach, coverage, information provided by respondents, data preparation, collection, tabulation, and analysis.

- There are two types of non-sampling error:

a- Response Error: Error arising due to inaccurate answers were given by respondents, or their answer is misinterpreted or recorded wrongly. It consists of researcher error, respondent error and interviewer error which are further classified as under.

- Researcher Error

  1- Surrogate Error 2-Sampling Error 3- Measurement Error 4- Data Analysis Error 4- Population Definition Error

- Respondent Error

  1- Inability Error 2- Unwillingness Error

- Interviewer Error

  1- Questioning Error 2- Recording Error 3- Respondent Selection Error 4-Cheating Error

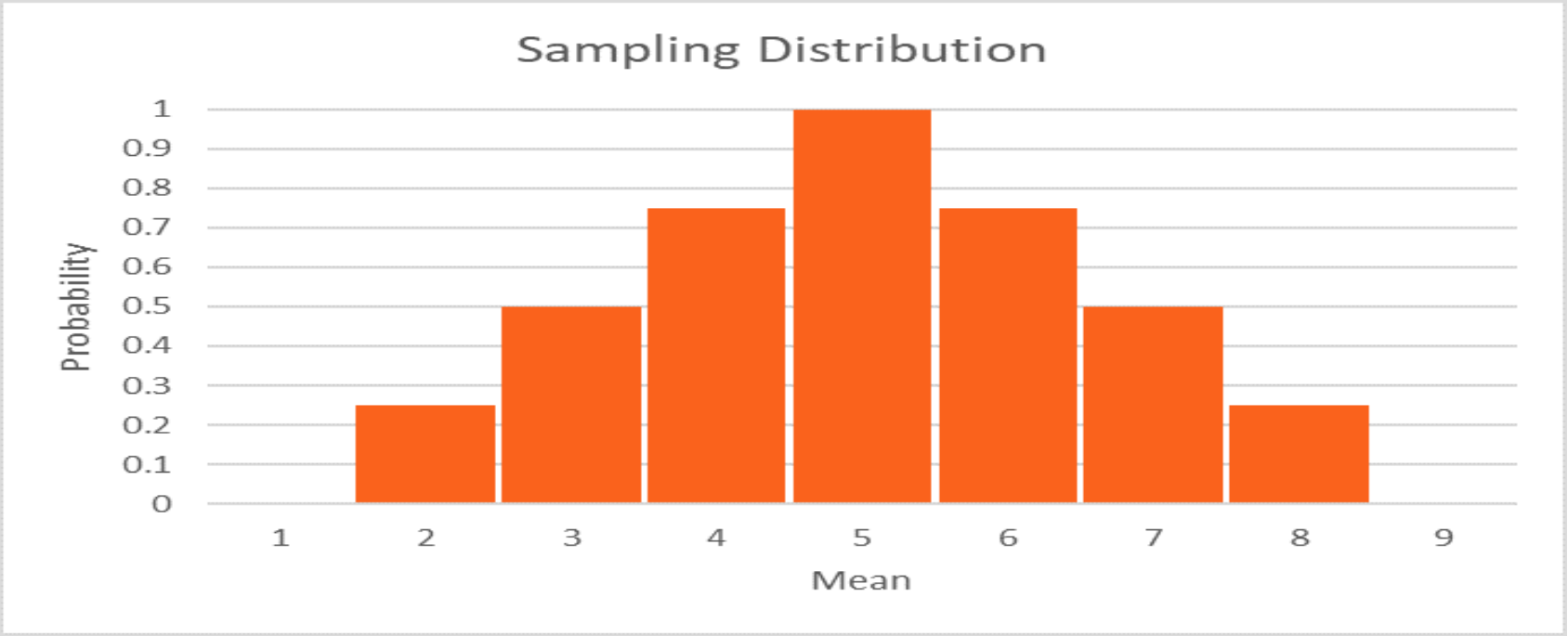
b- Non-Response Error: Error arising due to some respondents who are a part of the sample do not respond.

DIFFERENCES BETWEEN SAMPLING AND NON-SAMPLING ERROR

BASIS FOR COMPARISON	SAMPLING ERROR	NON-SAMPLING ERROR
Meaning	Sampling error is a type of error, occurs due to the sample selected does not perfectly represents the population of interest.	An error occurs due to sources other than sampling, while conducting survey activities is known as non sampling error.
Cause	Deviation between sample mean and population mean	Deficiency and analysis of data
Type	Random	Random or Non-random
Occurs	Only when sample is selected.	Both in sample and census.
Sample size	Possibility of error reduced with the increase in sample size.	It has nothing to do with the sample size.

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.



- **SAMPLING DISTRIBUTION**
- A sampling distribution is a graph of a statistic for your sample data. While, technically, you could choose any statistic to paint a picture, some common ones you'll come across are:
- Mean
- Mean absolute value of the deviation from the mean
- Range Standard deviation of the sample
- Unbiased estimate of variance
- Variance of the sample

Mean

Often in statistics, we tend to represent a set of data by a representative value which would approximately define the entire collection. This representative value is called the measure of central tendency, and the name suggests that it is a value around which the data is centered. These central tendencies are mean, median and mode.

- Mean is the most commonly used measure of central tendency. It actually represents the average of the given collection of data. It is applicable for both continuous and discrete data.
- It is equal to the sum of all the values in the collection of data divided by the total number of values.
- Suppose we have *n* values in a set of data namely as x_1, x_2, x_3, …, x_n, then the mean of data is given by:
- $\overline{X} = \frac{X_1+X_2+X_3+\cdots\ldots}{n}$
- It can also bee denoted as:
- $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$

For grouped data, we can calculate the mean using three different methods of formula.

Direct method	**Assumed mean method**	**Step deviation method**
Mean $\overline{X} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$ Here, $\sum f_i$ = Sum of all frequencies	Mean $\overline{X} = a + \frac{\sum f_i d_i}{\sum f_i}$ Here, a = Assumed mean $d_i = x_i - a$ $\sum f_i$ = Sum of all frequencies	Mean $\overline{X} = a + h\frac{\sum f_i d_i}{\sum f_i}$ Here, a = Assumed mean $u_i = (x_i - a)/h$ h = Class size $\sum f_i$ = Sum of all frequencies

Example: The given table shows the scores obtained by different players in a match. What is mean, median and mode of the given data?

S.No	Name	Runs Scored
1	Sachin	80
2	Yuvraj	52
3	Virat	40
4	Sehwag	52
5	Rohit	70
6	Harbhajan	1
7	Dhoni	6

Solution:

i) The mean is given by: $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} = \frac{80+52+40+52+70+1+6}{7} = 43$

The mean of the given data is 43

Sampling Theory

Sampling theory is a study of relation ship between a population and sample drawn from population. This theory is applicable only on random samples. A universe is the complete group of items about which knowledge is search. Universe may be finite or infinite.

- Finite universe is one which has a definite and certain number of items.
- If number of items is uncertain and infinite, the universe is said to be an infinite universe.

Similarly, the universe may be hypothetical or existent.

- In hypothetical case the universe in fact does not exist and we can only imagine the items constituting it.

Example: Tossing of coin or throwing a dice

Sampling Theory

- Existent universe is a universe of concrete(solid).

Example: the universe where the items constituting it really exist.

The theory of sampling studies the relationships that exist between the universe and the sample or sample drawn from it.

❖The main problem of sampling theory is the problem of relationship between a parameter and a statistic. The theory of sampling measure the purity of the estimate. This mix-up of movement from sample towards universe is what is known as statistical induction or statistical conclusion.

❖In more clear terms “from the sample we attempt to draw conclusion concerning the universe.

Sampling Theory

In order to be able to follow this inductive method, we first follow a deductive argument which is that we imagine a population(universe) and investigate the behaviour of the sample drawn from this universe applying the laws of probability. “ The methodology dealing with all this known as sampling theory”.

- **What Is Sampling Theory?**
- Sampling theory refers to the statistical branch dealing with the selection, gathering, and evaluation of data obtained from a sample population. It provides efficient and trusted methods to estimate and draw inferences, plus estimate characteristics of the whole population along with testing hypotheses.
- Sampling theory is fundamental in various fields, including **market research**, political polling, quality control in manufacturing, and scientific research. Moreover, it aids surveys, observational studies, and experiments. Researchers use it for **data collection** of animals, plants, and other organisms’ populations. It finds usage in testing and inspecting processes and products.

- Sampling theory is a branch of statistics that provides a framework for making conclusion about a population based on a subset of that population, called a sample.
- Its types include simple random, systematic, stratified, cluster, non-probability, convenience, judgmental, snowball, and quota sampling.
- Moreover, businesses, marketers, financiers, educators, economists, quality control specialists, agriculturists, forest managers, healthcare professionals, environmental researchers, and social scientists all apply it.

Hence, it is essential for efficient data collection, cost-effectiveness, informed decision-making, population representation, practicality in complex situations, and statistical analysis support

Sampling Theory in Statistics Explained

- Sampling theory in statistics means the practice of choosing a subset entity or individual out of a larger population to collect data to make inferences about the whole population. Moreover, it works by selecting a **representative sample** of market participants or financial data. The samples could contain investor behaviour, stock prices, or **bond yields**.
- Furthermore, researchers employ a random or **systematic sampling** method to ensure unbiased pictures. They then analyze these samples and draw appropriate conclusions concerning the larger financial landscape. Additionally, sampling theory bias arises when the sample selection is more or less likely to include specific individuals, resulting in inaccurate or skewed results.
- Therefore, **financial analysts** and researchers use it to draw inferences about entire market participants' stock prices or bond yields in a short time and affordable cost. Hence, using sampling, they are able to estimate vital financial elements like **market trends**, average returns, and **volatility** with confidence.

- **Sampling Theory in Statistics Explained**
- In the financial world, it has led to the development of different models like the Black-Scholes-Merton model and the **Capital Asset Pricing Model** (CAPM). These models forecast the return and risk of markets by utilizing statistical sampling techniques.
- In finance, professionals utilize it for **risk assessment**, market research, and **portfolio management**. It encompasses the sampling theory definition and its practical implementation, ensuring that researchers select representative samples for accurate data collection. Besides, attribute sampling theory is particularly relevant in auditing, quality control, and compliance testing.
- In summary, it has advanced financial research, modeling, and decision-making processes by enabling effective data collecting and estimating. Hence, the critical goal of sampling theory in research is to ensure that the sample is representative of the population, allowing researchers to draw accurate and meaningful conclusions.

- **Types**
- Sampling theory encompasses various types of sampling methods, each with its own set of principles and procedures. Here are some common types of sampling methods:
- **Probability sampling:** Every member of the dataset has an equal chance of getting selected for the sample, making it the most reliable sampling type. It can be further subdivided into.
- **Simple random sampling:** A random number generator conducts the sampling, providing every member with an equal opportunity for selection.
- **Systematic sampling:** It is applicable to a population organized in any way, like a customer list. Here, every kth member of the list is selected, originating from a random starting point.
- **Stratified sampling:** It is applicable in a heterogeneous population by dividing it into different groups and choosing a random sample from every group.
- **Cluster sampling:** It is done for widely scattered populations, and it isn't easy to recognize every member. It divides the population into clusters from which any single cluster is selected randomly.

Types

- **Non-probability sampling:** Here, every participant of a population does not have an equal chance to get selected in the sample.
- **Convenience sampling**: It selects easy-to-access members of a population when resources and time are limited.
- **Judgmental sampling:** Here, researchers select only those samples which they believe represent the entire population.
- **Snowball sampling:** Researchers begin the sampling from smaller samples and reach out to larger ones through them.
- **Quota sampling:** In it, first, small quotas are assigned, and then samples are selected till the quota of each group gets fulfilled.

- **Applications**

Many fields widely use it to benefit from it. Hence, listed below are some of its applications:

- **Business and marketing**: Used in surveys of market research, evaluating customer satisfaction, and marketing campaigns.
- **Finance:** It gets applied in assessing transactions and financial audits.
- **Education:** It allows evaluation of student performance.
- **Economics**: It aids in data collection on economic indicators, consumer spending, and employment rates.
- **Quality Control:** In manufacturing and production, it helps in quality control.
- **Forest Management:** Habit data, species composition, and timber volume deploy it.
- **Healthcare**: It assists in monitoring disease outbreaks, conducting clinical trials, and assessing patient health.
- **Environmental Studies:** Environmental pollution levels, monitoring air and water quality, and biodiversity apply.
- **Social Sciences:** Demographic studies, surveys, and opinion polls use it.
- **Epidemiology**: Here, it helps in studying the prevalence and distribution of diseases in populations—random sampling of individuals to estimate disease rates and identify risk factors.

- **Importance**
- Owing to its wide applications, as discussed in the previous section, it has many importance. Hence, let's list some of its major importance below:
- **Efficient Data Collection:** The use of smaller subgroups of populations simplifies data collecting and helps researchers save time and money.
- **Cost-Effective Research:** It is economical since examining a whole population may be quite expensive, particularly in fields like **market research** and healthcare.
- **Informed Decision-Making:** It offers trustworthy insights into industries like market research and healthcare, plus quality control, enabling well-informed decision-making.
- **Population Representation:** Ensuring that insights are typical of the total population is achieved by pulling data from samples, thereby increasing the validity of findings.
- **Practicality in Complex Situations:** It is crucial in situations when it is impracticable to examine the entire population, for example, when evaluating enormous forests or complex manufacturing processes.
- **Statistical Analysis Support:** It is an indispensable instrument for research, analysis, and effective resource management since it facilitates **statistical analysis** and helps with parameter estimates.

- **1 According to sampling theory, researchers can generalize findings to what?**

In line with sampling theory, a meticulously crafted sample should faithfully mirror the characteristics of the entire population. This theory substantiates the possibility of applying findings to a broader population, allowing researchers to draw valid conclusions from sample data.

- **2. How is sampling theory linked to probability?**

Sampling theory closely intertwines with probability by employing randomization in sample selection. Probability sampling forms the bedrock of statistical inference by ensuring equitable inclusion of every population element in the sample, fostering unbiased analysis.

- **3. What is sampling theory in machine learning?**

In machine learning, this theory involves techniques for creating representative data subsets for model training. To ensure the training dataset accurately mirrors the overall data distribution, it employs methods such as random sampling and stratified sampling.

- **4. What is sampling theory in research methodology?**

In research methodology, it pertains to the systematic selection of a subset (sample) from a larger population for data collection. It guides the choice of sampling methods, such as non-probability or random sampling, ensuring the sample faithfully represents the population for dependable research outcomes.

- **What Is 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.

- **Why standard error matters?**
- In statistics, data from samples is used to understand larger populations. Standard error matters because it helps you estimate how well your sample data represents the whole population.
- With probability sampling, where elements of a sample are randomly selected, you can collect data that is likely to be representative of the population. However, even with probability samples, some sampling error will remain. That's because a sample will never perfectly match the population it comes from in terms of measures like means and standard deviations.
- By calculating standard error, you can estimate how representative your sample is of your population and make valid conclusions.
- A high standard error shows that sample means are widely spread around the population mean—your sample may not closely represent your population. A low standard error shows that sample means are closely distributed around the population mean—your sample is representative of your population.
- You can decrease standard error by increasing sample size. Using a large, random sample is the best way to minimize sampling bias.

- **standard deviation vs Standard error**
- Standard error and standard deviation are both measures of variability:
- The **standard deviation** describes variability **within a single sample**.
- The standard deviation is a descriptive statistic that can be calculated from sample data.
- Example: Standard error vs standard deviationIn a random sample of 200 students, the mean math SAT score is 550. In this case, the sample is the 200 students, while the population is all test takers in the region.

- **standard deviation vs Standard error**
- the standard error is an inferential statistic that can only be estimated (unless the real population parameter is known).
- The **standard error** estimates the variability **across multiple samples** of a population.
- The standard error of the math scores, on the other hand, tells you how much the sample mean score of 550 differs from other sample mean scores, in samples of equal size, in the population of all test takers in the region

- **Standard error formula**
- The standard error of the mean is calculated using the standard deviation and the sample size.
- From the formula, you'll 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 orderly.

$$SE = \frac{\sigma}{\sqrt{n}}$$

SE is standard error, n is number of element in sample, σ is population standard deviation

- **When population parameters are unknown**
- 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

$$SE = \frac{s}{\sqrt{n}}$$

SE is standard error, n is number of element in sample, and *S* is sample standard deviation

- **Example:**
- Standard error vs standard deviation In a random sample of 200 students, the mean math SAT score is 550. In this case, the sample is the 200 students, while the population is all test takers in the region. The standard deviation of the math scores is 180. This number reflects on average how much each score differs from the sample mean score of 550**.**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-
- $\sqrt{n}$=200, than n=$\sqrt{200}$ =14.1
- Next, divide the sample standard deviation by the number you found in step one.

$$SE = \frac{s}{\sqrt{n}}$$

s=180 and n=14.1

Than *SE*=$\frac{180}{14.1}$= *12.8*

- **How should you report the standard error?**
- You can report the standard error alongside the mean or in a confidence interval to communicate the uncertainty around the mean.
- Example: Reporting the mean and standard errorThe mean math SAT score of a random sample of test takers is 550 ± 12.8 (*SE*).
- The best way to report the standard error is in a confidence interval because readers won't have to do any additional math to come up with a meaningful interval.
- A confidence interval is a range of values where an unknown population parameter is expected to lie most of the time, if you were to repeat your study with new random samples.
- With a 95% confidence level, 95% of all sample means will be expected to lie within a confidence interval of ± 1.96 standard errors of the sample mean.
- Based on random sampling, the true population parameter is also estimated to lie within this range with 95% confidence.

- **Other standard errors**
- Aside from the standard error of the mean (and other statistics), there are two other standard errors you might come across: the standard error of the estimate and the standard error of measurement.
- The **standard error of the estimate** is related to regression(growth) analysis. This reflects the variability around the estimated regression line and the accuracy of the regression model. Using the standard error of the estimate, you can construct a confidence interval for the true regression coefficient.
- The **standard error of measurement** is about the reliability of a measure. It indicates how variable the measurement error of a test is, and it's often reported in standardized testing. The standard error of measurement can be used to create a confidence interval for the true score of an element or an individual.

T

H

A

N

K

S