Damak Campus Journal 2023, **12 (1): 34-45**
ISSN : 2565-4772 (Print)

The Sampling Distribution of Sample Means

Shantiram Subedi
Department of Physics
Damak Multiple Campus
E-mail: *subedishantiramdamk@gmail.com*

Abstract

This article explains what sampling distribution is, how you can use it and the factors that influence its calculation, and the types of sampling and types of sampling distribution. We further define variance, describe how to calculate it and explain the advantages and disadvantages of using a variance.

Introduction

Professionals often gather statistics and evaluate them to help a company know more about its market, products or processes. Sampling distribution, a statistical tool, helps calculate the probability of an event by repeatedly sampling a small group of subjects rather that sampling an entire population.

In statistics, we use a random sample from the population of interest to draw conclusions and make inferences about the population. If the sample does not represent the population of interest, then inferences from data derived from the sample might not be valid. For example, determining the effect of an intervention for adolescents based on the results from a sample of adults could yield the wrong conclusion.

Research Objectives

What is sampling distribution?

Sampling distribution is a statistic that determines the probability of an event based on data from a small group within a large population. It is a probability distribution of a statistics obtained from large number of samples drawn from a specific population. Any statistics that can be computed for a sample has a sampling distribution.

Its primary purpose is to establish representative results of small samples of a comparatively larger population. Since the population is too large to analyze, you can select a smaller group and repeatedly sample or analyze them. The gathered data, or statistic, is used to calculate the likely occurrence, or probability, of an event.

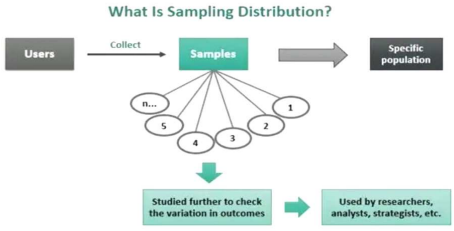

Figure 1, showing what is sampling distribution.

Using a sampling distribution simplifies the process of making inferences, or conclusions, about large amounts of data**.**

Materials and Methods

Understanding sampling distribution

The idea behind a sampling distribution is that when you have a large amount of data (gathered from a large group) the value of a statistic from random samples of a small group can inform you of that statistic's value for the entire group.

Each random sample selected may have a different value assigned to the statistic being studied. For example, if you randomly sample data three times and determine the mean, or the average, of each sample, all three means are likely to be different and fall somewhere along the graph. That's variability. You do that many times, and eventually the data you plot may look like a bell curve. That process is a sampling distribution. Once we plot the data on a graph, the values of any given statistic in random samples may make a normal distribution from which we can draw inferences.

Here's a simple example of the theory: when you roll a single die, your odds of getting any number (1,2,3,4,5, or 6) are the same (1/6). The mean for any roll is (1 + 2 + 3 + 4 + 5 + 6) / 6 = 3.5. The results from a one-die roll are shown in the first figure below: it looks like uniform distribution. However, as the sample size is increased (two dice, three dice…), the distribution of the mean looks more and more like a normal distribution as shown in second, third, fourth and fifth figures below. That is what the central limit theorem predicts.

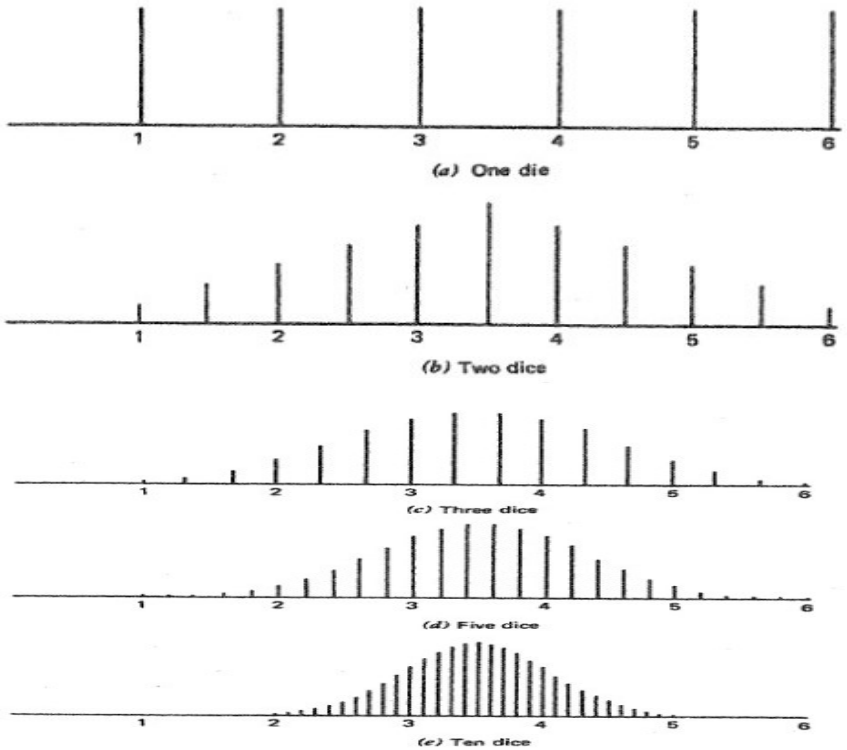

Figure-3 showing the distribution of mean [face].

As the sample size increases, distribution of the mean will approach the population mean of μ, and the variance will approach to $\frac{\sigma^2}{n}$, where n is the sample size.

Factors that influence sampling distribution

The sampling distribution depends on multiple factors – the statistic, sample size, sampling process, and the overall population. We can measure the sampling distribution's variability either by standard deviation, also called "standard error of the mean," or population variances, depending on the context and inferences you are trying to draw. They both are mathematical formulas that measure the spread of data points in relation to the mean.

There are three primary factors that influence the variability of a sampling distribution. They are:

- **The number observed in a population:** The symbol for this variable is "N." It is the measure of observed activity in a given group of data.
- **The number observed in the sample:** The symbol for this variable is "n." It is the measure of observed activity in a random sample of data that is part of the larger grouping.
- **The method of choosing the sample:** How you chose the samples can account for variability in some cases.

Types of distributions

There are three standard types of sampling distributions in statistics:

Sampling distribution of mean

The most common type of sampling distribution is the mean. It focuses on calculating the mean of every sample group chosen from the population and plotting the data points. The graph shows a normal distribution where the center is the mean of the sampling distribution, which represents the mean of the entire population.

Sampling distribution of proportion

It gives us information about proportions in a population. We would select samples from the population and get the sample proportion. This sampling distribution focuses on proportions in a population. The mean of all the sample proportions that we calculate from each sample group would become the proportion of the entire population.

T-distribution

T-distribution is used when the sample size is very small or not much is known about the population. It is used to estimate the mean of the population and other statistics such as confidence intervals, statistical differences and linear regression. The T-distribution uses a t-score to evaluate data that wouldn't be appropriate for a normal distribution. The formula for t-score is:

$$t = \frac{x - \mu}{\frac{s}{\sqrt{n}}}$$

In the formula, "x" is the sample mean and "μ" is the population mean and *s* indicates standard deviation.

Example of a sampling distribution

Here is an example of a sampling distribution using a fictional scenario with a data set and a graph:

A professor is interested in understanding the sampling distribution of their students' test scores. This professor thinks this may help determine a suitable curve for the previous tests their students completed. The professor recorded test scores from the previous three tests and created a data table and a sampling distribution graph.

	Test one	Test two	Test three
Student one	80	82	84
Student two	78	76	80
Student three	74	86	80
Student four	75	81	80
Student five	76	81	78
Student six	88	81	89
Student seven	72	79	75
Student eight	94	95	99

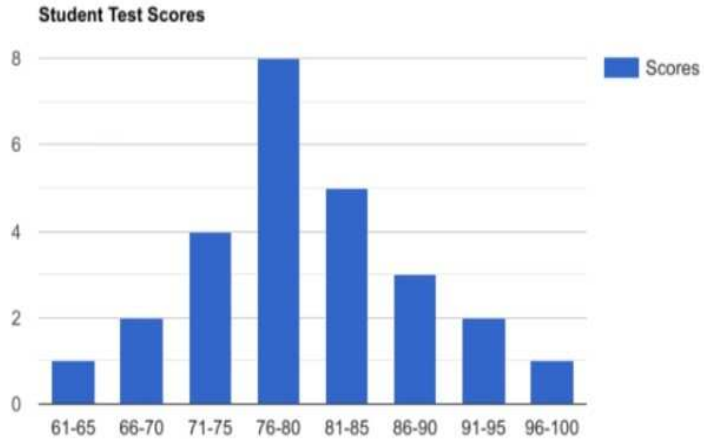

Figure-3 Data table and its sampling distribution

How to Calculate Sample Mean?

When statisticians study populations, they may take a sampling of a larger population to apply statistical calculations to figure out trends and predict outcomes about the larger population. The sample mean is one calculation that can tell statisticians the average of a given set of data. Statisticians use the sample mean of a data set to make projections about a standard

of normality within a given population, and the sample mean can also be used to find the variance, deviation and standard error within a data set.

What is the sample mean?

A sample mean is an average of a set of data. The sample mean can be used to calculate the central tendency, standard deviation and the variance of a data set. The sample mean can be applied to a variety of uses, including calculating population averages.

How to calculate the sample mean

Calculating sample mean is as simple as adding up the number of items in a sample set and then dividing that sum by the number of items in the sample set. To calculate the sample mean, you can use the formula:

$$\bar{x} = \frac{\sum_i^n x_i}{n}$$

Here, $\bar{x}$ represents the sample mean, $\sum_i^n x_i$ tells us to add of all the X-values and n stands for the number of items in the data set.

How to calculate sample variance?

The following steps will show how to calculate the sample variance of a data set:

i. Write the equation for variance

To find the variance, you can use the equation below:

$$s^2 = \frac{\sum_i^n (x_i - x)}{n-1}$$

Where:

- s^2 is the sample variance of a data set
- $\sum_i^n (x_i - x)$represents the sum of your X values minus the mean of your data set
- $\bar{x}$represents the mean of your sample
- n is the sample size

ii. Replace symbols with values from your data set

iii. Simplify and solve the equation

Findings and Discussion

What is the variance of the sampling distribution of the mean?

The variance of sample mean in simple random sample without replacement is given by

$$var(\bar{x}) = \frac{s^2}{n}$$

Proof: By definition,

$$\begin{aligned} var(\bar{x}) &= E[\bar{x} - E(\bar{x})]^2 \\ &= E[\bar{x} - \bar{X}]^2 \\ &= E[\tfrac{x_1 + x_2 + \cdots + x_n}{n} - \bar{X}]^2 \\ &= E[\tfrac{x_1 + x_2 + \cdots + x_n - n\bar{X}}{n}]^2 \end{aligned}$$

$= \frac{1}{n^2}\, E[(x_1 - \bar{X}) + (x_2 - \bar{X}) + \cdots + (x_n - \bar{X})]^2$

$= \frac{1}{n^2}\, E\left[\sum_i^n (x_i - \bar{X})^2 + 2\sum_{i<j}^{n-1}\sum_j^n (x_i - \bar{X})(x_j - \bar{X})\right]$

$= \frac{1}{n^2}\left[\sum_{i=1}^n E(x_i - \bar{X})^2 + 2\sum_{i<j}^{n-1}\sum_j^n E(x_i - \bar{X})(x_j - \bar{X})\right]$

$= \frac{1}{n^2}\left[\sum_{i=1}^n \frac{1}{N}\sum_i^N (X_i - \bar{X})^2 + 2\sum_{i<j}^n \sum_j^n \frac{1}{N(N-1)}\sum_{i<j}^{N-1}\sum_j^n (X_i - \bar{X})\,(X_j - \bar{X})\right]$

$= \frac{1}{n^2}\left[\sum_{i=1}^n \sigma^2 + \frac{(n-1)n}{(N-1)N}\,(-N\sigma^2)\right]$ ---------(*)

$= \frac{1}{n^2}\left[n\sigma^2 - \frac{(n-1)n}{(N-1)}\sigma^2\right]$

$= \frac{n}{n^2}\left[\frac{N-1}{N}s^2 - \frac{n-1}{N-1}\frac{N-1}{N}s^2\right]$ ----------(**)

$= \frac{1}{nN}\ [N - 1 - \text{n+1}]\, s^2$

$= \dfrac{N-n}{N}\dfrac{S^2}{n}$

$= \left(1 - \dfrac{n}{N}\right)\dfrac{S^2}{n}$

$= \frac{S^2}{n}$, when $\frac{n}{N} \to 0\ for\ large\ N.$

In equation (*), $\sum_{i=1}^N (X_i - \bar{X}) = 0$ gives $(X_1 - \bar{X}) + (X_2 - \bar{X}) + \cdots + (X_N - \bar{X}) = 0.$

Squaring both sides and writing in compact form, we get

$$\sum_{i=1}^{N}(X_i - \bar{X})^2 + 2\sum_{i<j}^{N-1}\sum_{j}^{N}(X_i - \bar{X})\left(X_j - \bar{X}\right) = 0$$

$$2\sum_{i<j}^{N-1}\sum_{j}^{N}(X_i - \bar{X})\left(X_j - \bar{X}\right) = -\sum_{i=1}^{N}(X_i - \bar{X})^2 = -N\,\sigma^2$$

In equation (*), $s^2 = \frac{1}{N-1}\sum_{i=1}^n (X_i - \bar{X})^2$ &$\sigma^2 = \frac{1}{N}\sum_{i=1}^n (X_i - \bar{X})^2$gives

$$s^2(N-1) = \sigma^2\, N\ \ so, \sigma^2 = \frac{N-1}{N}s^2$$

The variance of a data set refers to the spread of the items within the sample set. When statisticians calculate variance, they are trying to figure out how far apart the items are from each other when representing data on a graph. Variance can tell you how different each item in a sample set is. Additionally, the sample mean, variance, standard deviation and error can be analyzed to assume and predict outcomes and trends about a population as well as a sampling of that population. Standard deviation is an important calculation because it allows companies and

individuals to understand whether their data is in proximity to the average or if the data is spread over a wider range.

Central limit theorem

The discussion on sampling distribution is incomplete without the mention of the central limit theorem, which states that the shape of the distribution will depend on the size of the sample. The central limit theorem helps in constructing a sampling distribution.

The **Central Limit Theorem (CLT)** states the principle that: the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.

More specifically, if $X_1, X_2, \ldots, X_n$ are n identical and independently distributed random variables with mean μ and standard deviation σ, then the distribution of their sample mean, $\frac{X_1+X_2+\cdots+X_n}{n}$, as n gets large, is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

The theorem says a normal distribution depends on the sample size. As the number of sample groups increases, the number of variables or standard error decreases. According to this theorem, the increase in the sample size will reduce the chances of standard error, thereby keeping the distribution normal. When users plot the data on a graph, the shape will be close to the bell-curve shape. In short, the more sample groups one studies, the better and more normal is the result/representation.

What is the standard error of the sample mean?

The standard error of the mean (SEM), or standard deviation, represents how far the sample mean is from the true population mean. Standard error of the mean is measure of dispersion of the distribution of the sample mean. In other word, the standard error of mean measures the extent to which we expect the means from the different samples to vary because of chance error in the sampling process. A distribution of sample means that less spread out (that has a small standard error) is a better estimator of the population mean than a distribution sample means that is widely spread and has a larger standard error.

Verification

Here, we are interested to verify the sample mean and sample variance are unbiased estimator of population mean and population variance. Suppose a population consists of four units with values 1,4,6 and 9. In simple random sample with replacement, the total number of random samples of size two which can be drawn from the population of size four is equal to 4^2 =16. The 16 simple random samples of size 2 are:

(1,1), (1,4), (1,6), (1,9), (4,1), (4,4), (4,6), (4,9), (6,1), (6,4), (6,6), (6,9), (9,1), (9,4), (9,6),(9,9).

Calculation of sample mean and variance:

Random Samples	Sample Mean ($\bar{x}_i$)	Sample Variance $s_i{}^2 = \frac{1}{n-1}\sum_{i=1}^{2}(x_i - \bar{x})^2$
(1,1)	1	0
(1,4)	2.5	4.5
(1,6)	3.5	12.5
(1,9)	5	32
(4,1)	2.5	4.5
(4,4)	4	0
(4,6)	5	2
(4,9)	6.5	12.5
(6,1)	3.5	12.5
(6,4)	5	2
(6,6)	6	0
(6,9)	7.5	4.5
(9,1)	5	32
(9,4)	6.5	12.5
(9,6)	7.5	4.5
(9,9)	9	0
	$\sum_{i=1}^{16} x_i = 90$	$\sum_{i=1}^{16} s_i{}^2 = 136$

Now the population mean (μ)

$$= \frac{\sum_{i=1}^{4} X_i}{N}$$

$= \frac{1+4+6+9}{4}$

$= 5.$

Probability of $\bar{x}\ is\ P(\bar{x}) = \frac{1}{N^n} = \frac{1}{16}$.

So, $E(\bar{x}) = \sum_{i=1}^{16} x_i P_i(x_i)$

$= 80 \times \frac{1}{16} = 5.$

Hence, $E(\bar{x}) = \mu = 5$, which implies that the sample $\bar{x}$ is an unbiased estimator of population mean μ.

Next, the population variance (σ^2)

$$= \frac{1}{N}\sum_{i=1}^{4}(X_i - \bar{X})^2$$

$= \frac{1}{4}[(1-5)^2 + (4-5)^2 + (6-5)^2 + (9-5)^2] = 8.5.$

And sample variance $E(s^2) = \sum_{i=1}^{16} {s_i}^2 P_i({s_i}^2)$

$= 136 \times \frac{1}{16}$ $= 8.5$

Hence, $E(s^2) = \sigma^2 = 8.5$, which implies that the sample variance s^2 is an unbiased estimator of population variance σ^2.

Further, the variance of sample means,

$$s^2 = V(\bar{x}) = \frac{1}{n}(\bar{x}_i - \mu)^2$$

$$= \frac{1}{16}[(1-5)^2 + (2.5-5)^2 + (3.5-5)^2 + (5-5)^2 + (2.5-5)^2 + (4-5)^2 + (5-5)^2 + (6.5-5)^2 + (3.5-5)^2 + (5-5)^2 + (6-5)^2 + (7.5-5)^2 + (5-5)^2 + (6.5-5)^2 + (7.5-5)^2 + (9-5)^2$$

$= 4.25$

and, $\frac{\sigma^2}{n} = \frac{8.5}{2} = 4.25$

Hence, $s^2 = \frac{\sigma^2}{n}$ is also verified.

Finally, the standard error$= \sqrt{V(\bar{x})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} = \sqrt{4.25} = 2.0$

Interpretation

If the standard error is 2.06, it means that the estimate or statistic we are considering is expected to have an average deviation of 2.06 units from the true population mean. The standard error represents the variability of the estimate. With a standard error of 2.06, you can expect the estimate to vary by approximately 2.06 units on average.

What is a bell curve?

A bell curve is a graph that is a normal distribution. The graph is a bell-shaped line where the curve's highest point shows the most probable event in a number (or series) of data. There are other occurrences that are equally scattered around this highest point on the curve.

Why is the curve important?

The bell curve has a lot of features, uses and relevance, and some of them are as follows:

- It is important in the field of statistics because they model many real-world data like test results and performance reviews of employees.
- The bell curve has one mode, and it coincides with the mean and median. This mode is the center of the bell curve, and it is the highest point.
- When the bell curve is folded into two parts along the vertical line, the two parts are mirror images of each other. This shows that it is symmetric.
- For a bell curve, exactly 68% of the data is within one standard deviation of the mean.
- For a bell curve, exactly 95% of the data lies within the two standard deviations of the mean.
- For a bell curve, exactly 99.7% of the data is within three standard deviations of the mean.
- The bell curve graph is useful for repeated measurements of equipment.
- What is bell curve distribution?

• The bell curve distribution is a means to figure out how data are distributed when plotted in a graph. You will always arrive at a bell-shaped curve if the data is evenly distributed.
• For a bell curve normal distribution, a small percentage of the points (about 5%) of the data would fail on the tails of the graph while about 90% of the data would fall in between the graph.

What is the relationship between the standard deviation and a bell curve?

• There is a significant relationship between the standard deviation and the bell curve.The spread of the bell curve normal distribution is controlled by the standard deviation.
• A larger standard deviation value shows that the data is spread out around the mean of the data. In this situation, the bell curve graph will be flatter and wider.
• However, if there is a smaller standard deviation value, then it shows that the data is tightly concentrated around the mean of the data. In this situation, the bell curve graph will be taller and thinner. This is illustrated in the figure below:

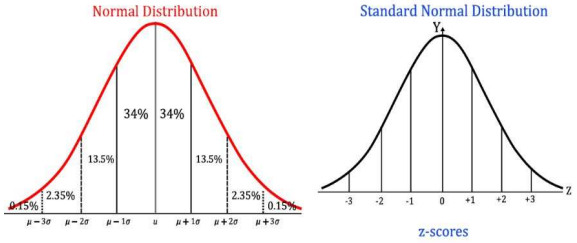

Figure 4. Showing Normal curve

What is sampling?

Sampling is the selection of subjects in a statistical study to represent a larger population. Because testing every member of a given population isn't always feasible, researchers select samples to make testing more efficient and cost-effective.

How researchers develop samples can have a significant impact on the quality of the study's results. The following elements determine a sample's efficiency:

• **Accuracy:** Accuracy refers to how accurate sample responses are. Researchers should try to eliminate bias and influence from both researchers and participants.
• **Precision:** Samples should provide answers to the specific research question researchers are asking. Answers should be relevant to the study.
• **Representativeness:** A research sample should seek to provide the most representative group of subjects for the population as a whole. For instance, if researchers want to estimate how a city's residents feel about the government instituting a curfew, the sample should match the city's demographic percentages as closely as possible.

What are the types of sampling?

All types of sampling fall into one of these two fundamental categories:

• **Probability sampling:** if the small units of the population are drawn according to the results of probability is called probability sampling. In probability sampling, researchers can calculate the probability of any single person in the population being selected for the study. These studies provide greater mathematical precision and analysis.

• **Non-probability sampling:** In no probability sampling, researchers cannot calculate the probability of being in the study for individuals within the population. If the small units of population are drawn according to the judgment (view) of the researcher is called non-probability sampling. These samples tend to be less accurate and less representative of the larger population.

Types of probability sampling

Here are the five types of probability sampling that researchers use:

Simple random sampling

Simple random sampling, or SRS, occurs when each sample participant has the same probability of being chosen for the study. Consider a lottery method. You can place all possible respondents in a pool and randomly, or blindly, select participants. Every person in the pool has the same likelihood that you will choose them.

Simple random sampling is a sampling technique that uses a table of random numbers or an electronic random number generator for selecting samples. Researchers may also use computer programs that generate random numbers from a set. This method can be efficient for drawing samples from a large and diverse population.

For example, researchers might assemble a large group of individuals, assign numbers to each person and randomly choose numbers through an automated process. Researchers may also use lottery system or computer software to conduct the automated selection process.

Simple random sampling provides less opportunity for bias and influence by researchers in participant selection. However, true random sampling can be challenging because it requires a list of every potential participant.

Stratified Random Sampling

In stratified random sampling, the population is divided into subgroups or strata based on certain characteristics. Random samples are taken from each stratum, ensuring representation from all subgroups.

Systematic Sampling

Systematic sampling involves selecting every k^{th} individual from a list after starting with a random sample. For example, if every 5^{th} person is chosen, the first person is randomly selected, and then every 5^{th} person thereafter is included in the sample.

Cluster Sampling

For example, suppose we have to study the academic performance of students in a large city. The population is all students in the city. The city is divided into several school districts are called cluster, and each district consists of multiple schools. Instead of sampling individual students, we will use random cluster sampling to select entire school a clusters.

In cluster sampling, the population is divided into clusters, and a random sample of cluster is selected. Then, all individuals within the chosen cluster are included in the sample.

The advantage of cluster sampling in this scenario is that it simplifies the sampling process. Instead of individually sampling students from each school, we are treating entire schools as

clusters and randomly selecting a few clusters to study. This approach can be more cost-effective and logistically feasible, especially when dealing with a large population.

Multistage Sampling

Multistage sampling is a combination of different sampling methods especially a combination of cluster and stratified sampling. This method is often used to collect data from a large, geographically spread group of people in national surveys.

Conclusion

In conclusion, understanding the sampling distribution of the mean is essential for making reliable statistical inferences about population parameters. It allows us to make statements about the precision of sample estimates and provides a foundation for constructing confidence intervals and performing hypothesis tests. The central limit theorem assures us that, under certain conditions, the mean of a sufficiently large random sample will be normally distributed, reinforcing the robustness and generalizability of our statistical methods.

Researchers and practitioners can use the knowledge of the sampling distribution of the mean to make informed decisions based on sample data, contributing to the validity and reliability of statistical analyses. As statistical techniques continue to play a crucial role in various fields, a solid understanding of the principles underlying the sampling distribution of the mean remains indispensable play a crucial role in various fields, a solid understanding of the principles underlying the sampling distribution of the mean remains indispensable.

References

Bangert, Q.W.,& Baumberger, J.P. (2055).”Research and Statistical Techniques used in the journal of Counseling & Development: *Journal of Counseling & Development, 83,480-487.*Agresti, A. 2007. “An Introduction to Categorical Data Analysis.” *The Statistical Analysis of Composition Data, Chapman & Hall.*

Douglas A.Lind, Willam G. Marchal and Samuel A. Wathen. (2012).” Statistical Techniques in Business & Economics.” *MaGraw-Hill/Irwin.*

D.R. Cox, G. Gudmundsson, et.al., (1981).” Statistical Analysis of Time Series. “*Scandinavian Journal of Statistics,93-115.*

F.A. Adesoji and M.A. Babatonde. (2009).“Basic Statistical Techniques in Research.”*ResearchGate,4-20.*

GianlucaMalato. (2020). “Statistical analysis of Stock price.”*Toward Data Science.*

J. Karthikeyan, H.P. Horizan, et.al., (2018). “Statistical Techniques and Tools for Describing and Analyzing data in Elt Research.” *International Journal of Civil engineering and Technology.*

Karim Elbahloul. (2019). “Stock Market Prediction Using Various Statistical Methods Volume I.”*ResearchGate.*

Lyle F. Bachman. (2005). “Statistical Analyses for Language Assessment Workbook and CD-ROM.” *Cambridge University Press, 29-54.*

Nasser Said G.A., Muhhamad A.S. et. al., (2022). “Statistical Analysis Tools: A Review of Implementation and effectiveness of Teaching English.” *ResearchGate, 2—4.*

Refael A. Irizarry. (2019). “Introduction to Data Science: Data Analysis and Prediction Algorithm with R.” *Chapman and Hall/CRC;1st edition.*

Susan Trocoso Skidmore and Bruce Thompson. (2010). “Statistical Techniques Used in Published Article: A Historical Review of Reviews.”*SageJournals.*