Auditing Course Material

Part 51 of 61 (Chapters 5001-5100)

10. Normal Distribution

Illustration 1

From actual tests with the bulbs, a company estimated that the mean life of bulbs is μ = 36,500 hours and that the standard deviation is σ = 5000. In addition, the data collected indicate that a normal distribution is a reasonable assumption. What percentage of the bulbs can be expected to last more than 40,000 hours? In other words, what is the probability that the life of bulb will exceed 40,000 hours?

Solution:

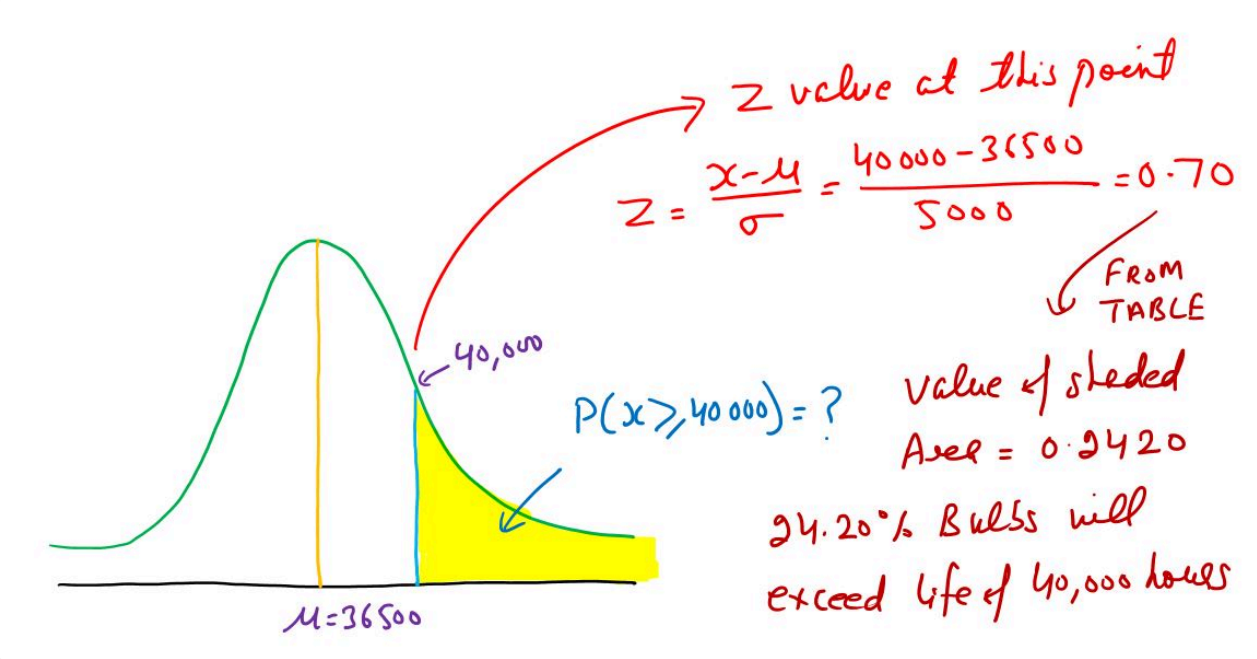


Illustration 2

The download time of Youtube videos is assumed to be normally distributed with a mean of 7 seconds and standard deviation of 2 seconds. Calculate the followings:

(i) Probability that the download time will be less than 9 seconds.

(ii) Probability that the download time will be more than 9 seconds.

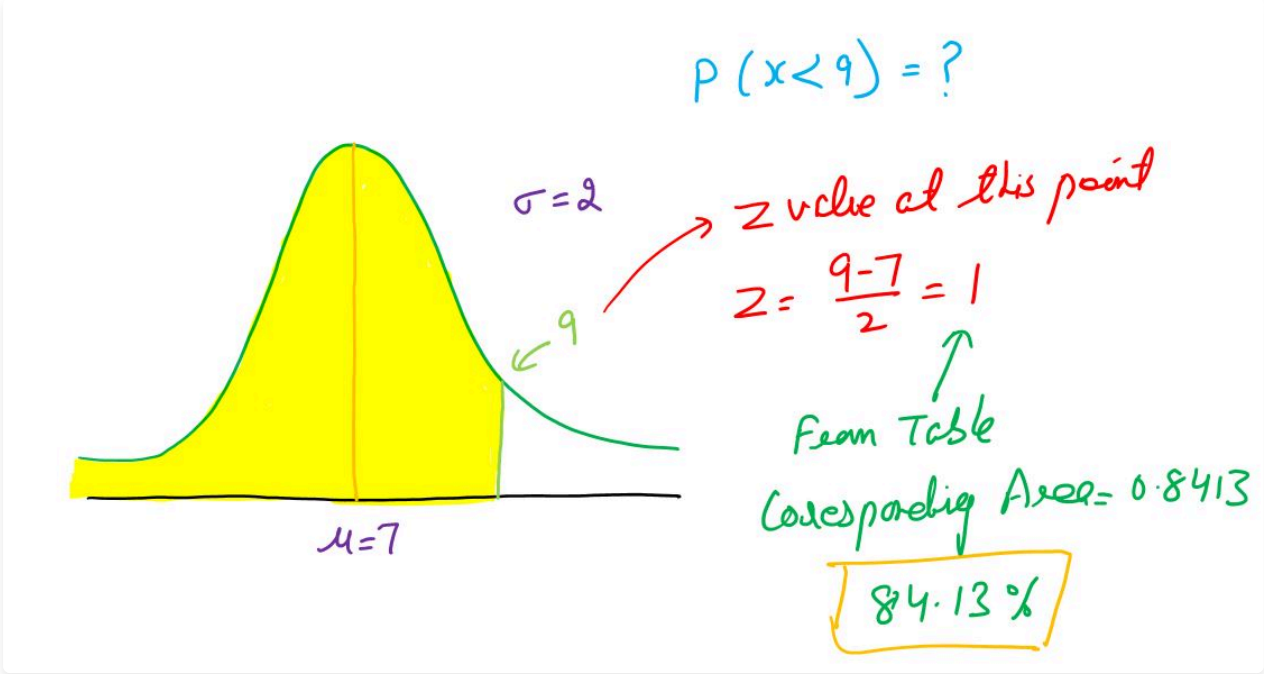
(iii) Probability that the download time will be less than 7 seconds or over 9 seconds.

(iv) Probability that the download time will be between 5 to 9 seconds.

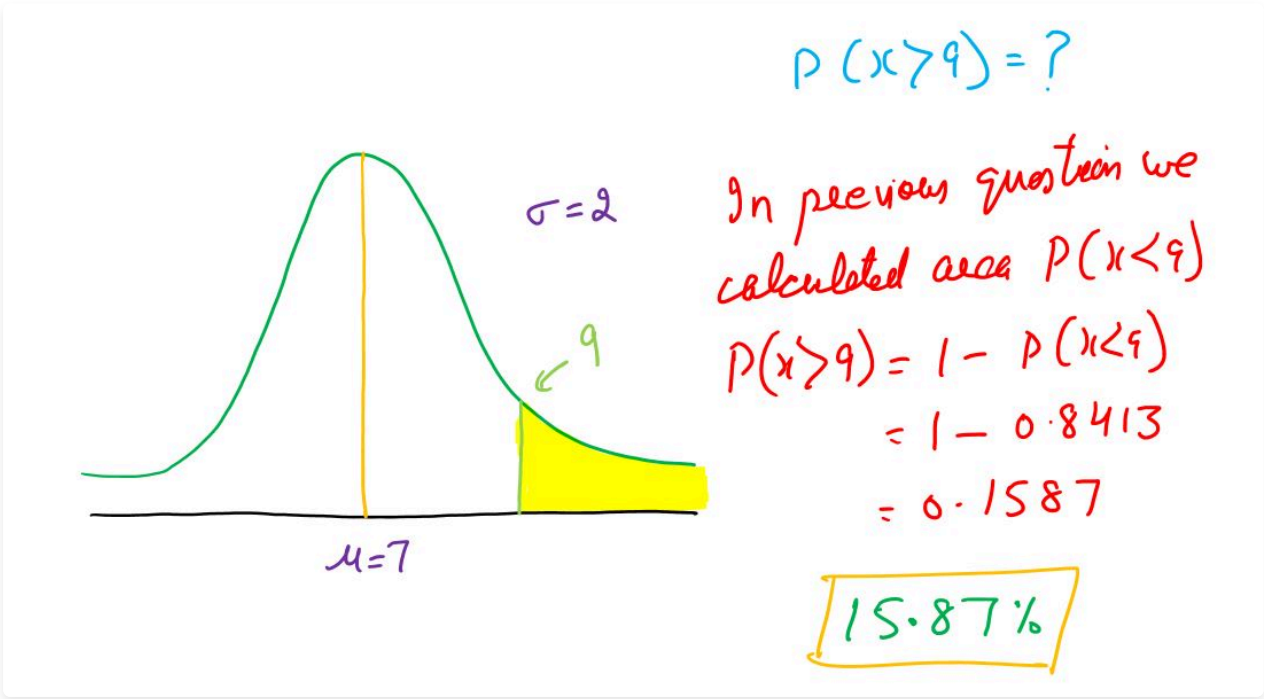
(v) How much time (in seconds) will elapse before the fastest 10% of the downloads are complete.

Solutions:

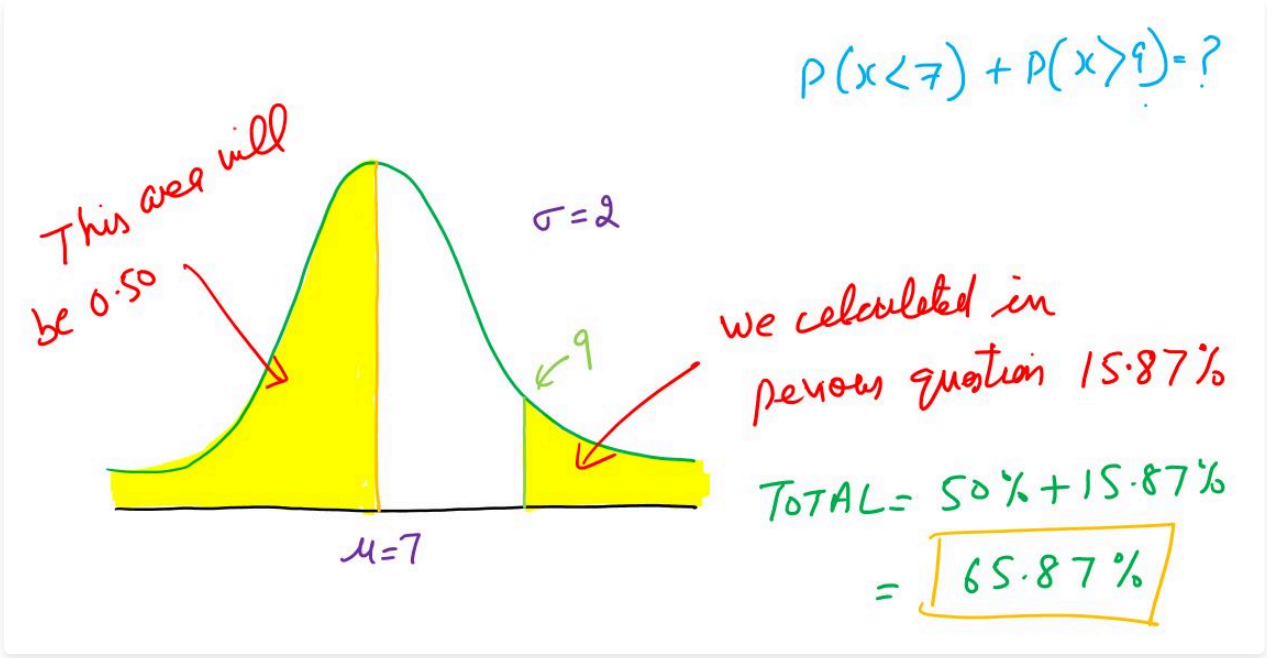
(i) Probability that the download time will be less than 9 seconds.



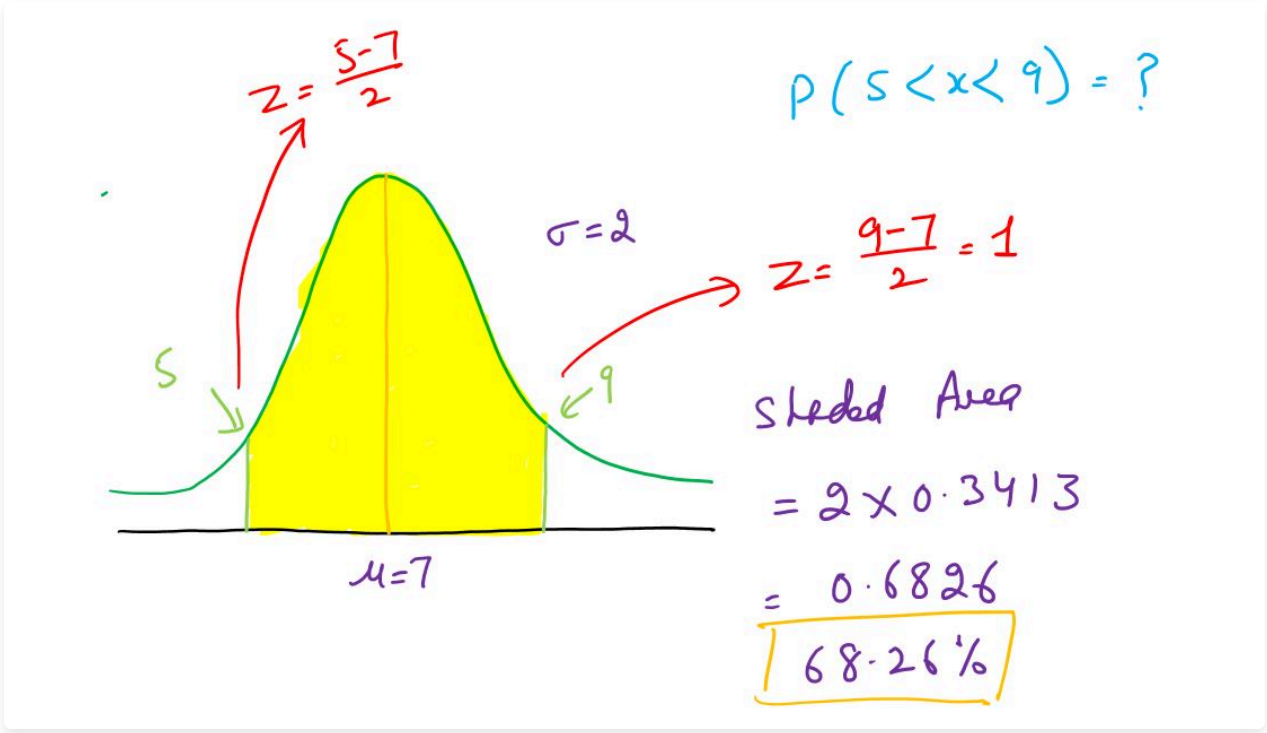
(ii) Probability that the download time will be more than 9 seconds.



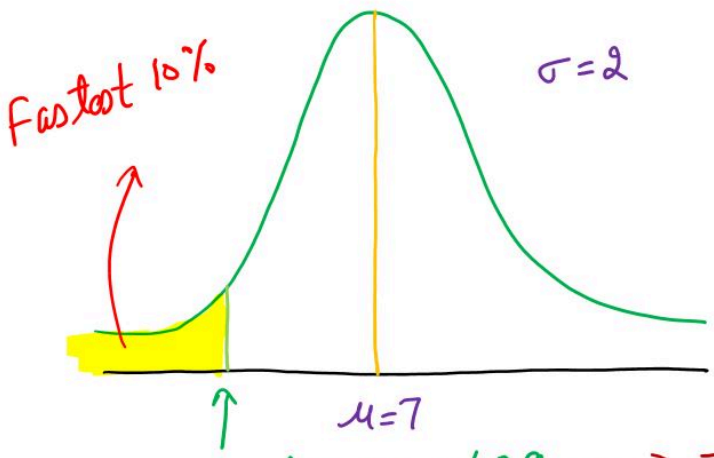
(iii) Probability that the download time will be less than 7 seconds or over 9 seconds.



(iv) Probability that the download time will be between 5 to 9 seconds.



(v) How much time (in seconds) will elapse before the fastest 10% of the downloads are complete.

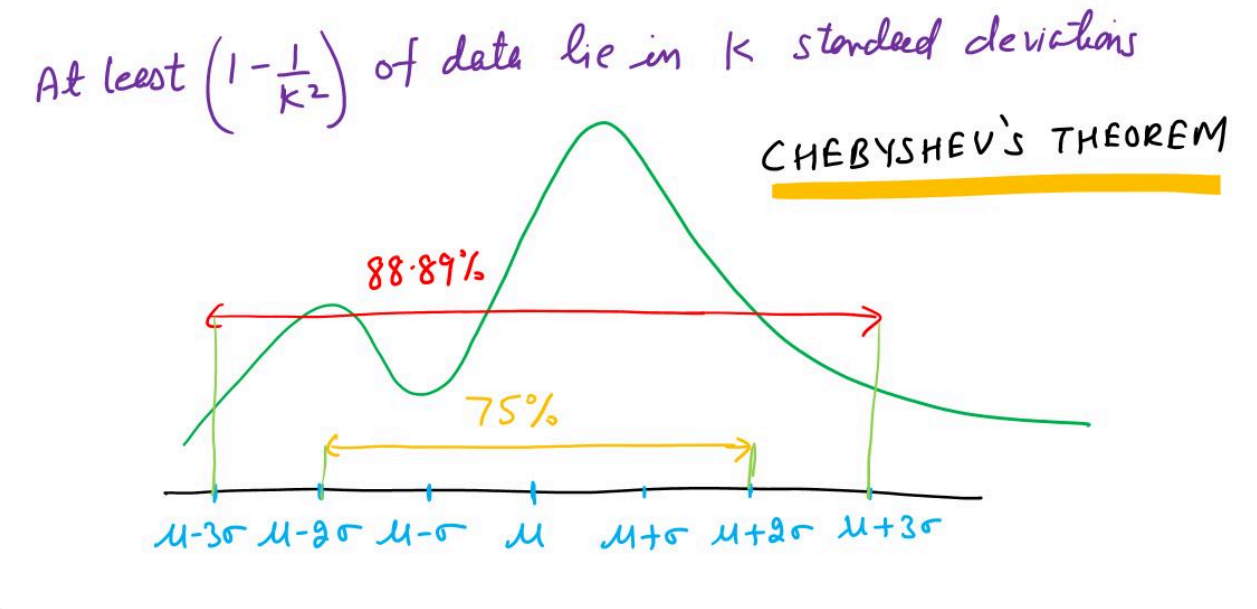


Corresponding $z = -1.28 \implies -1.28 = \frac{x-7}{2}$

$\implies x =$ 4.44 Seconds

10. Normal Distribution

When dealing with data that doesn't follow a normal distribution, **Chebyshev's Theorem** becomes a valuable tool as it applies universally to all shapes of distributions.



A result that applies to every data set is known as Chebyshev's Theorem:

- at least 3/4 of the data lie within two standard deviations of the mean, that is, in the interval with endpoints μ±2σ for populations.
- at least 8/9 of the data lie within three standard deviations of the mean, that is, in the interval with endpoints μ±3σ for populations.

We can generalize that, at least (1 – $\frac{1}{k^2}$), of the data lie within k standard deviations of the mean, that is, in the interval with endpoints μ ± kσ for populations, where k is any positive whole number that is greater than 1.

Illustration 1

The mean level of work experience of employees in KPMG is 8 years and the standard deviation is 1 year. What is the probability that a randomly selected employee from the KPMG will have had between 6 and 10 years of work experience?

Solution:
Since we have not been told the form of the distribution, we can use Chebyshev's theorem, which applies to any distribution.

With μ= 8 years and σ=1 year, 6 years of work experience is 2 standard deviations below μ and 10 years of work experience is 2 standard deviations above μ.

The probability of an individual picked at random from the population will be within 2 standard deviations from the mean is

$1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = \frac{3}{4}$ or 75%

Therefore, the probability that the individual will have had work experience between 6 to 10 years is 75%.

---

11. Exponential Distribution

Exponential distribution is a **Continuous random distribution.**

The exponential distribution is a type of continuous probability distribution used to model the time between independent events that occur at a constant average rate within a specific interval.

This distribution is commonly applied to random variables representing various scenarios such as the time intervals between successive arrivals at a car wash, the time taken to load a truck, the distances between major defects on a highway, and similar situations where events occur independently and at a consistent average rate.

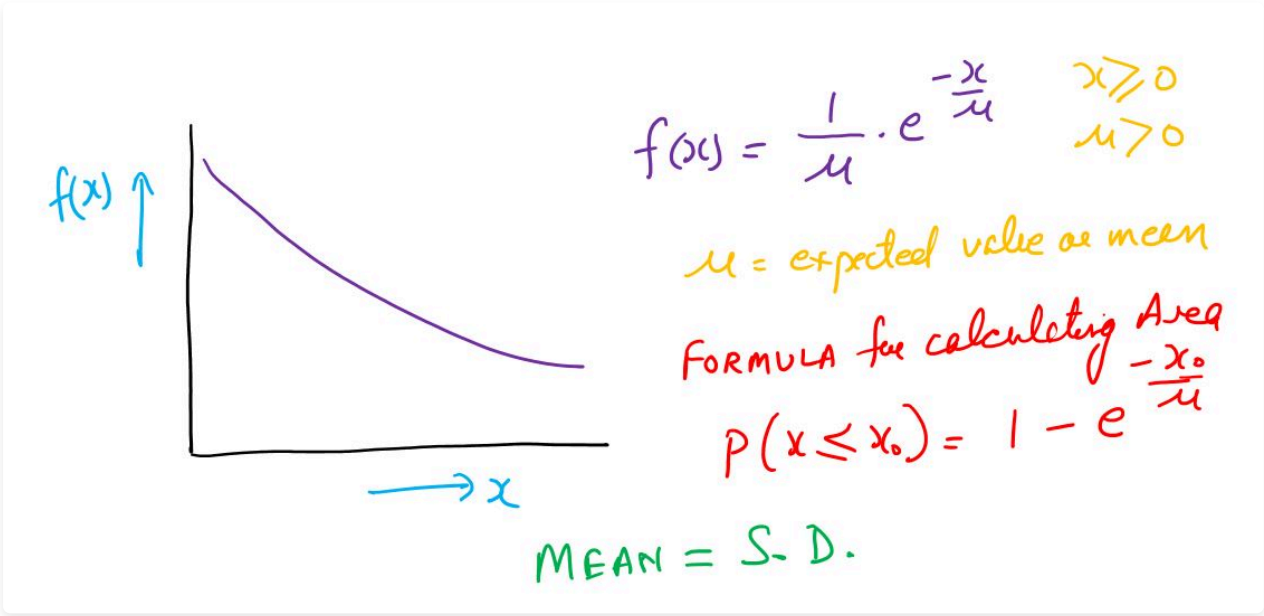
---

11. Exponential Distribution

The exponential probability density function is given by following formula.

$f(x) = \frac{1}{\mu} e^{\frac{-x}{\mu}}$, for $x \geq 0$

where μ is Expected Value of Mean. For $x < 0$, $f(x) = 0$.



As with any continuous probability distribution, the area under the curve corresponding to an interval provides the probability that the random variable assumes a value in that interval.

To compute exponential probabilities such as those just described, we use the following formula. It provides the cumulative probability of obtaining a value for the exponential random variable of less than or equal to some specific value denoted by x_0.

$P(x \leq x_0) = 1 - e^{\frac{-x_0}{\mu}}$

A property of the exponential distribution is that the mean of the distribution and the standard deviation of the distribution are equal. Thus:

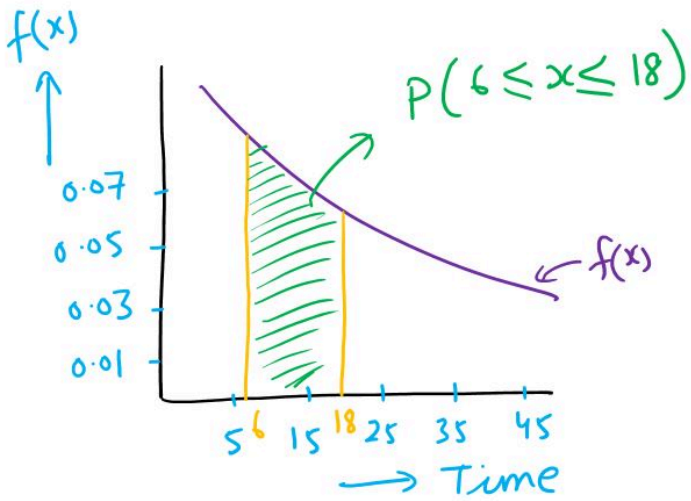
Expected value or Mean of Exponential Distribution, $E(x) = \mu = \frac{1}{\lambda}$

Variance of Exponential Distribution, $Var(x) = \sigma^2 = \frac{1}{\lambda^2}$

Standard Deviation of Exponential Distribution = $\sigma = \frac{1}{\lambda}$

Example

As an example of the exponential distribution, suppose that x represents the time taken for one person to get vaccination for Covid and follows exponential. If the mean, or average, vaccination time is 15 minutes (μ=15), the appropriate probability density function for x is
$f(x) = \frac{1}{15} e^{\frac{-x}{15}}$



As with any continuous probability distribution, the area under the curve corresponding to an interval provides the probability that the random variable assumes a value in that interval. In this example, the probability that vaccinating one person will take 6 minutes or less P (x ≤ 6) is defined to be the area under the curve in the figure from x = 0 to x = 6. Similarly, the probability that the vaccination time will be 18 minutes or less P (x ≤ 18) is the area under the curve from x = 0 to x = 18. Note also that the probability that the vaccination time will be between 6 minutes and 18 minutes P (6 ≤ x ≤ 18) is given by the area under the curve from x = 6 to x = 18.

Hence, the probability that vaccinating one person will take 6 minutes or less is:

P(x≤6) = 1 - \(e^\frac{-6}{15} \) = 0.3297

Using same equation, we calculate the probability of vaccinating one person in 18 minutes or less.

P(x≤18) = 1 - \(e^\frac{-18}{15} \) = 0.6988

Thus, the probability that vaccinating one person will take between 6 minutes and 18 minutes is equal to 0.6988 - 0.3297 = 0.3691. Probability for any other interval can be computed similarly.

---

11. Exponential Distribution

Illustration 1

Service times for customers at a library information desk can be modeled by an exponential distribution with a mean service time of 5 minutes. What is the probability that a customer service time will take longer than 10 minutes?

Solution:

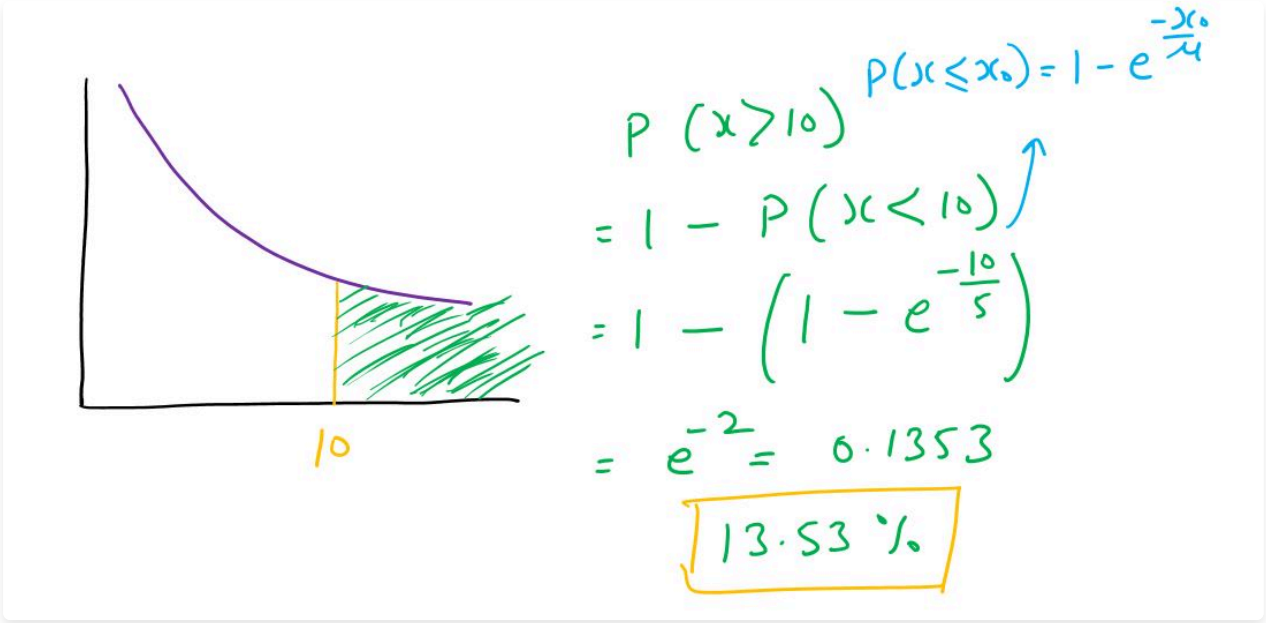


Illustration 2

A manufacturing firm has been involved in statistical quality control for several years. As part of the production process, parts are randomly selected and tested. From the records of these tests, it has been established that a defective part occurs in a pattern that is Poisson distributed on the average of 1.38 defects every 20 minutes during production runs. Use this information to determine the probability that less than 15 minutes will elapse between any two defects.

Solution:

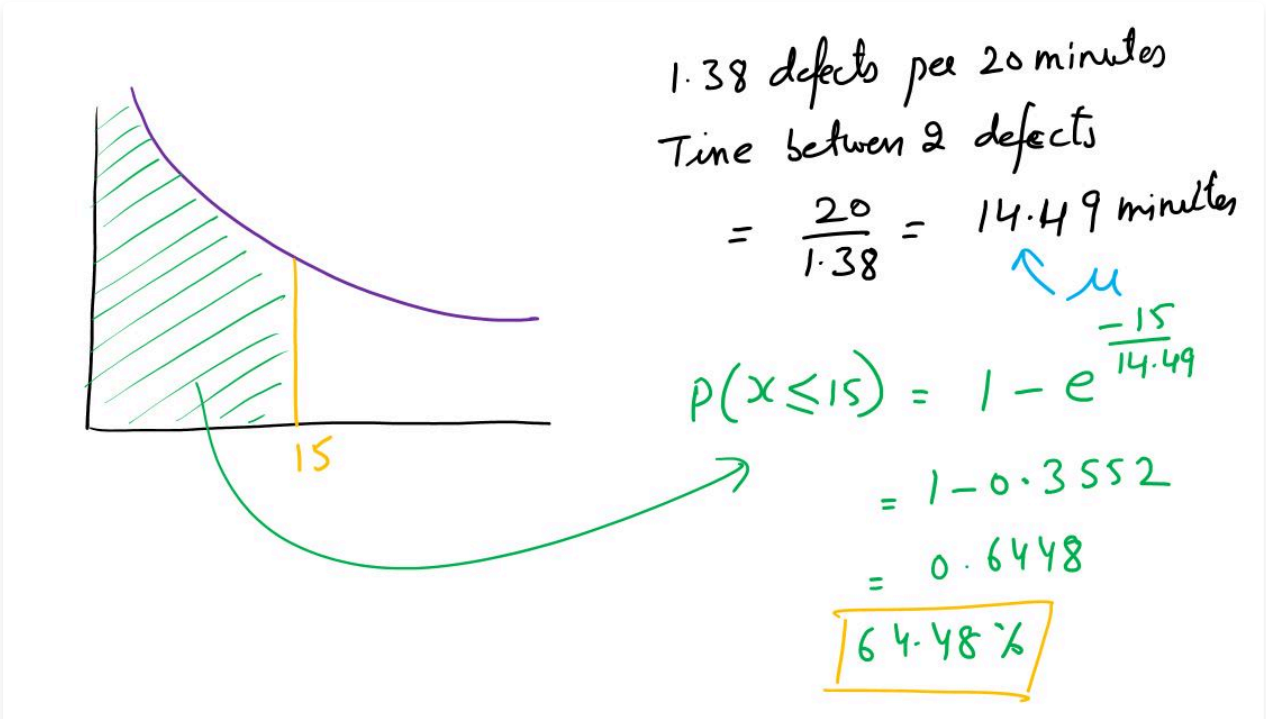


Illustration 3

If jobs arrive every 15 seconds on average, i.e. λ = 4 jobs per minute, what is the probability of waiting less than or equal to 30 seconds (0.5 minutes)?

Solution:

Here X_0 = 0.5, λ = 4,

We calculate μ = $\frac{1}{\lambda} = \frac{1}{4}$ = 0.25

P(t≤0.5) = 1 - $e^{\frac{-0.5}{0.25}}$ = 0.86

---

11. Exponential Distribution

The continuous exponential probability distribution is related to the discrete Poisson distribution. If the Poisson distribution provides an appropriate description of the number of occurrences per interval, the exponential distribution provides a description of the length of the interval between occurrences.

Consider following two examples to understand this.

Arrival Times at a Service Center

Consider a service center where customers arrive according to a Poisson process, meaning the number of arrivals per unit of time follows a Poisson distribution. The Poisson distribution here describes the number of customers arriving in a specific time frame, let's say per hour. Now, the time between successive customer arrivals can be modeled using an exponential distribution. The exponential distribution describes the waiting time between arrivals at the service center.

Defects in Manufacturing

In a manufacturing process, defects might occur randomly, following a Poisson distribution for the number of defects in a certain area or timeframe. The time between two successive defects could then be modeled using an exponential distribution. This distribution characterizes the time interval between individual defect occurrences.

Numerical Interpretation

To understand this, consider that, in a Poisson process, if events occur on average at the rate of λ per unit of time, then there will be on average λt occurrences per t units of time. The Poisson distribution describing this process is therefore P(x) = $\frac{e^{- \lambda t} \times(\lambda t)^x}{x!}$, from which P (x = 0) = $e^{- \lambda t}$ is the probability of no occurrences in 't' units of time.

Another interpretation of P (x = 0) = $e^{- \lambda t}$ is that this is the probability that the time, t, to the first occurrence is greater than 't', i.e. P (T > t) = P (x = 0 | μ = λt) = $e^{- \lambda t}$.

Conversely, the probability that an event does occur during t units of time is given by P (T ≤ t) = 1−P (x = 0 | μ = λt) = 1− $e^{- \lambda t}$

Note that this is the cumulative exponential distribution which, when differentiated with respect to t, produces the probability density function of the exponential distribution f(t) = λ $e^{- \lambda t}$.

---

1. Introduction

Sampling is a method used to gather information from a subset of a larger group or population. It involves selecting a smaller portion (called sample), from the entire population. We use the characteristics of this sample to draw conclusions or make estimations about the entire population.

Let us understand this with few examples.

Quality Check of Tyres by Honda Cars

Honda Cars is buying 1,00,000 tyres from MRF tyres. It doesn't mean Honda will check each and every tyre for quality; it's impractical and time-consuming. Instead, they might select a representative sample, say a few hundred or thousand tyres, randomly from the entire lot. This sample set undergoes rigorous testing to assess the overall quality. The results obtained from this sample will then be used to estimate the quality of the entire batch. If the sample shows good quality, it's assumed the rest of the tyres in the lot are of similar quality.

Survey on Chief Minister's Popularity

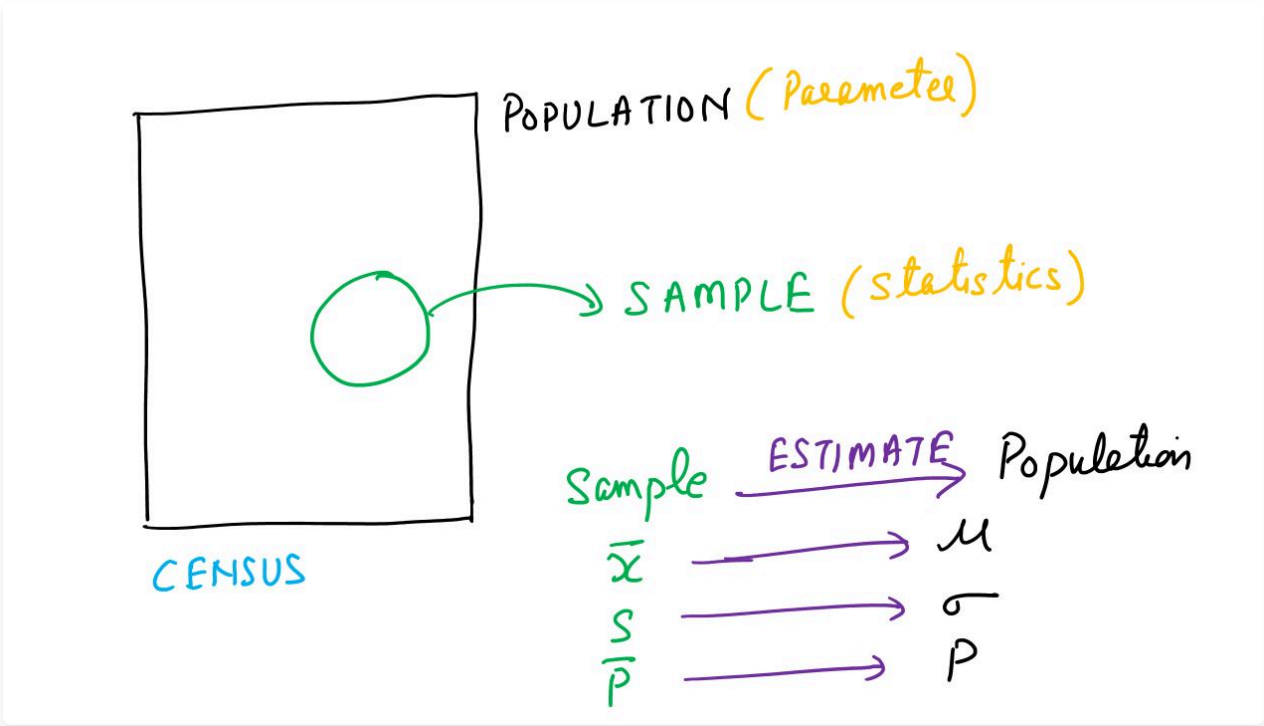
Suppose a political party is interested in finding public opinion about the current Chief Minister. Asking all 2.5 crore people (population of state) in the state is impractical. Instead, the political party selects a smaller representative sample, maybe a few thousand individuals, from different regions, age groups, and demographics. The opinions collected from this sample will provide an estimate of the overall sentiment towards the Chief Minister within the entire state population.

In both cases, sampling helps in drawing conclusions about the larger population without the need to examine every single item or person, making the process more feasible, cost-effective, and efficient. The key is to ensure that the sample chosen is representative of the entire population to provide accurate estimations or insights.

---

2. Key Terms

Before moving further, let us understand key concepts, used in process of Sampling. We will use an example to explain concepts.



Imagine you're in charge of quality control at a fluorescent tube manufacturing company. You're tasked with assessing the average lifespan of the tubes produced.

Population
The population in this scenario would be all the fluorescent tubes manufactured by your company. This includes every single tube ever produced.

Sampling
You cannot test every single tube to determine its lifespan; it's impractical and time-consuming. Instead, you select a representative sample, say 15 tubes, randomly from the entire stock. These 15 tubes represent a smaller part of the larger population.

Sample
The 15 tubes chosen for testing are your sample. This smaller group represents the entire population of tubes.

Parameter
The average lifespan of all the tubes manufactured by your company is a parameter. It's a characteristic of the entire population. This is what you aim to estimate or infer through your study of the sample.

Statistics
The mean lifespan of the 15 tubes in your sample is a statistic. It's a characteristic or measurement obtained from your smaller sample group.

Census
If you were to inspect and test every single fluorescent tube ever manufactured by your company to determine their lifespans, that would be a census. However, a census might not be feasible due to time, cost, and practical constraints.

In practice, you rely on the sample's statistics, like the mean lifespan of the 15 tubes, to estimate or infer the parameter, the average lifespan of all the tubes in the population. This process of studying a smaller part (sample) to draw conclusions about the larger group (population) is the essence of sampling.

It is similar to assessing a few grains of rice to infer the quality of an entire bag, just as we do at the grocery store.

3. Finite and Infinite Population

A population refers to the entire group or collection of individuals, items, or units that share at least one common characteristic. There are two types of populations:

Finite Population

A finite population is a set of elements that has a countable and limited size. This means that there is a specific and definite number of individual units in the population.

Examples:

- The number of students in a classroom.
- The quantity of laptops in a store.
- The total count of houses in a neighborhood.

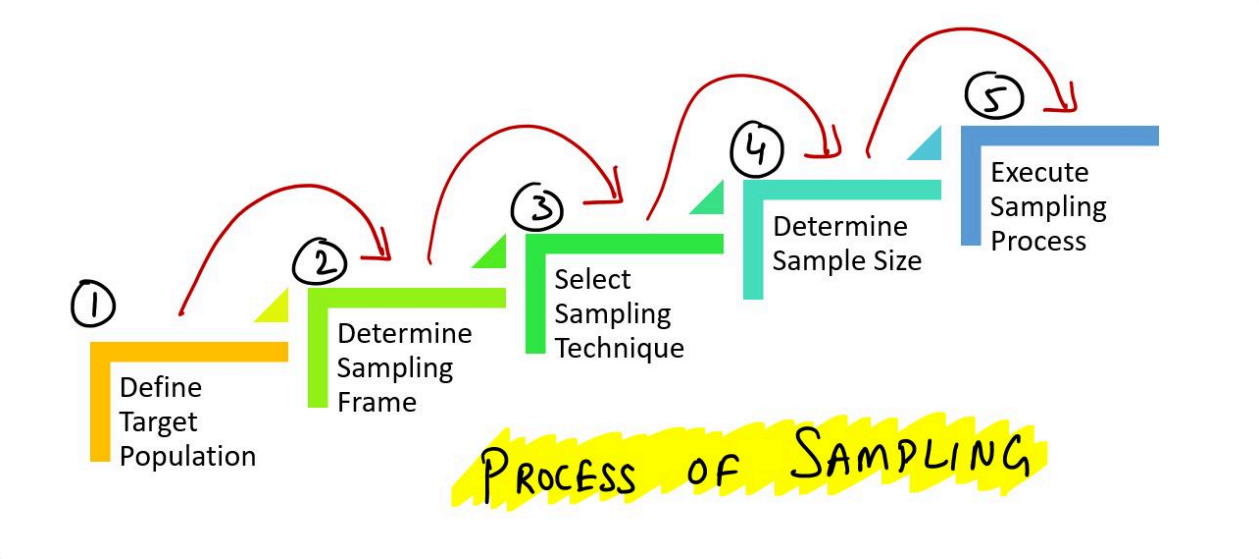
Infinite Population

An infinite population, as the name suggests, is a population that is practically limitless and cannot be counted or exhaustively identified. It's characterized by an uncountable number of elements, making it impossible to reach an endpoint or determine the exact size of the population.

Examples:

- All social media users worldwide.
- The weights of all adults in a country.
- The heights of all trees in a forest.

4. Process of Sampling



The process of sampling can be divided into following 5 steps:
(i) Define the Target Population
(ii) Determine the Sampling Frame
(iii) Select a Sampling Technique
(iv) Determine the Sample Size
(v) Execute the Sampling Process

Let us understand there steps, one by one.

---

4. Process of Sampling

The **target population** is the complete collection of entities that possess the specific information required by the researcher. Defining the target population involves considerations related to elements, sampling units, geographical extent, and the relevant time frame.

An **element** represents the fundamental object or subject from which the desired information is sought. For instance, in market research, elements could be respondents, products, stores, companies, or families.

Sampling units, on the other hand, are the distinct units containing these elements that are available for potential selection during various stages of the sampling process. For example, when conducting surveys, households might be considered as the initial sampling unit, followed by selecting the male or female head of the household responsible for most shopping at department stores.

Extent refers to the geographical boundaries within which the target population exists. This can be defined as narrowly as a specific city or as broadly as a continent or worldwide.

Time signifies the specific time period or duration under consideration for the study. For instance, when studying consumer behavior, the time frame could be March 2025.

An example:
Element: Chemical engineers
Sampling units: Companies purchasing over Rs 300 Crores of chemicals annually; then chemical engineers within these companies
Extent: Punjab
Time: 2024

Another example:
Element: Hospital Patients
Sampling Units: Hospital departments (e.g., Cardiology, Pediatrics); then patients in those departments
Extent: Gwalior
Time: July 2023

---

4. Process of Sampling

The **sampling frame** serves as a representation or guide comprising the elements within the target population. It is sub-set of working population. It is also called Working Population.

Examples of Sampling Frames:
A Telephone Directory: Lists of individuals or households in a specific area or region.
Industry Association Directories: Catalogs or databases containing the names of companies within a particular industry.
Mailing Lists: Purchased databases from commercial entities providing contact details of individuals or organizations.
Geographic Maps: Spatial representations indicating areas or regions for sampling purposes.

---

4. Process of Sampling

Selecting a sampling technique is a crucial step in the process of sampling. There are various methods to choose from, each with its own advantages and applicability based on the research objectives and characteristics of the population.

Some common sampling techniques include:

Simple Random Sampling: This technique involves randomly selecting elements from the population, where each element has an equal chance of being chosen. For example, using a random number generator to pick names from a list.

Stratified Sampling: The population is divided into homogeneous subgroups or strata, and then random samples are taken from each stratum proportionate to their size. This method ensures representation from various subgroups.

Systematic Sampling: Selecting elements at regular intervals from an ordered list after randomly selecting a starting point. For instance, selecting every 5th participant from a list of students.

Cluster Sampling: Dividing the population into clusters or groups, randomly selecting some clusters, and then including all elements within those selected clusters in the sample. This method is useful when the population is widely dispersed.

Convenience Sampling: Involves selecting readily available subjects for the study, often based on ease of access or availability. However, this method may not represent the entire population accurately due to biases.

Snowball Sampling: Used when members of a population are hard to find. Initial participants refer or lead the researcher to other potential participants, forming a chain.

Each sampling technique has its strengths and limitations, and the choice of method depends on factors such as the research objectives, resources available, and the nature of the population being studied.

The goal is to select a method that provides a representative sample to draw accurate inferences about the entire population.

---

4. Process of Sampling

Determining the sample size is a crucial step in research, balancing the need for precision without unnecessary resource utilization. The size of the sample chosen impacts the accuracy and credibility of study findings. It relies on various factors like population size, the desired level of accuracy or confidence interval, variability within the population, and the study's objectives.

Researchers often use statistical formulas or software to calculate the sample size. These calculations involve a trade-off between achieving statistical significance and managing practical constraints. A larger sample generally reduces the margin of error but might increase costs and resource utilization. Conversely, a smaller sample might save resources but could risk lowered precision and limited generalizability of the findings.

Optimal sample size determination aims to strike a balance, ensuring the research outcomes are reliable and applicable within reasonable resource limits.

---

4. Process of Sampling

Executing the sampling process involves implementing the chosen sampling technique to select elements from the sampling frame. This step necessitates precision and adherence to the predetermined sampling plan. The process typically follows a well-defined protocol to ensure accuracy and representativeness of the sample.

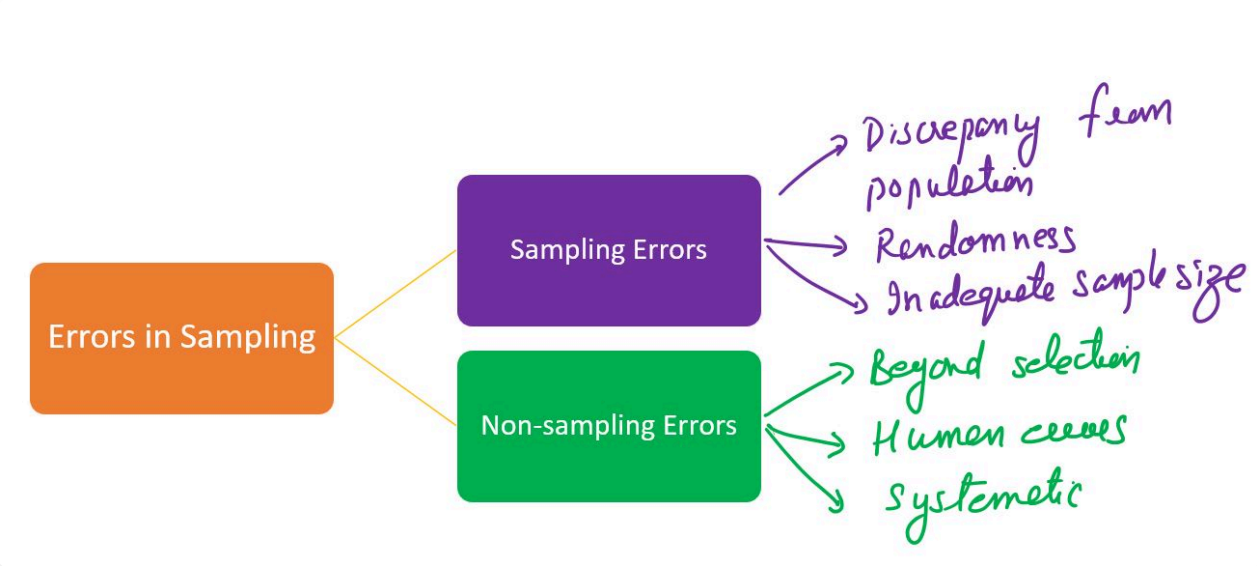
Firstly, researchers utilize the selected sampling method, whether it's simple random sampling, stratified sampling, cluster sampling, or others, to pick elements from the sampling frame systematically. This step often involves random selection or predefined criteria to maintain fairness and avoid bias.

Once the elements are chosen, the researchers gather data or conduct the necessary observations, surveys, or experiments according to the study's objectives. It's essential to adhere to the predetermined sample size and method to maintain consistency and reliability in the results.

Throughout this process, meticulous documentation and record-keeping are crucial. Researchers should meticulously record details of the selected sample, any deviations from the initial plan, and any unforeseen challenges or biases encountered during the sampling process. This comprehensive documentation aids in transparency, replicability, and validity of the research outcomes.

---

5. Errors in Sampling



Sampling error and non-sampling error are two distinct types of errors that can occur in the process of sampling during research or data collection.

Sampling Error

Sampling error refers to the variation or discrepancy between the sample statistic and the true population parameter that arises because a sample is used to estimate information about a larger population. It occurs due to the randomness involved in selecting a sample rather than the entire population.

Causes: Sampling error emerges when the sample chosen is not a perfect representation of the entire population. It occurs due to chance, random fluctuations, or inadequate sample size. It is also called Random Sampling Error.

Impact: Sampling error is inherent and expected in sampling processes. Larger sample sizes generally reduce sampling error as they tend to better represent the population, whereas smaller sample sizes can lead to larger sampling errors.

Mitigation: Researchers often attempt to minimize sampling error by employing appropriate sampling methods and ensuring random and representative sample selection.

Non-Sampling Error

Non-sampling error encompasses all other potential sources of error in a research study besides sampling error. It includes errors arising from data collection, processing, analysis, and interpretation. It is a type of Systematic Error.

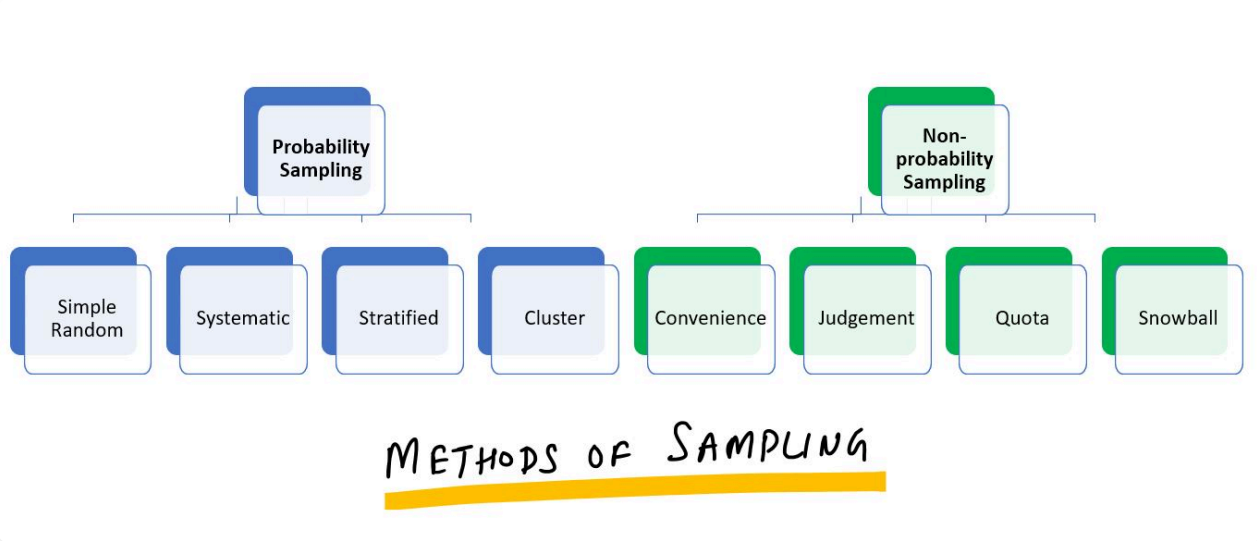
Causes: Non-sampling errors can arise due to various factors such as measurement errors, data entry mistakes, respondent errors in surveys, faulty instruments, biased sampling methods, human errors during data collection, and errors in analysis or interpretation.

Impact: Non-sampling errors can significantly affect the accuracy, reliability, and validity of research findings. Unlike sampling error, non-sampling error is more controllable and can often be minimized through careful planning, rigorous data collection procedures, and thorough validation processes.

Mitigation: Minimizing non-sampling errors involves employing robust data collection techniques, ensuring data accuracy and completeness, training data collectors adequately, using standardized measurement tools, and conducting quality checks throughout the research process.

---

6. Methods of Sampling



Sampling methods can be broadly classified into two categories based on the control over sample selection:

1. Probability Sampling

In probability sampling, each element or member of the population has a known and non-zero chance of being selected for the sample. This method ensures that every individual or unit in the population has an equal opportunity of being included in the sample. Various techniques such as simple random sampling, systematic sampling, stratified sampling, and cluster sampling fall under probability sampling. Probability sampling methods are advantageous as they offer statistical theories for estimating accuracy and allowing generalizations to the entire population.

2. Non-probability Sampling

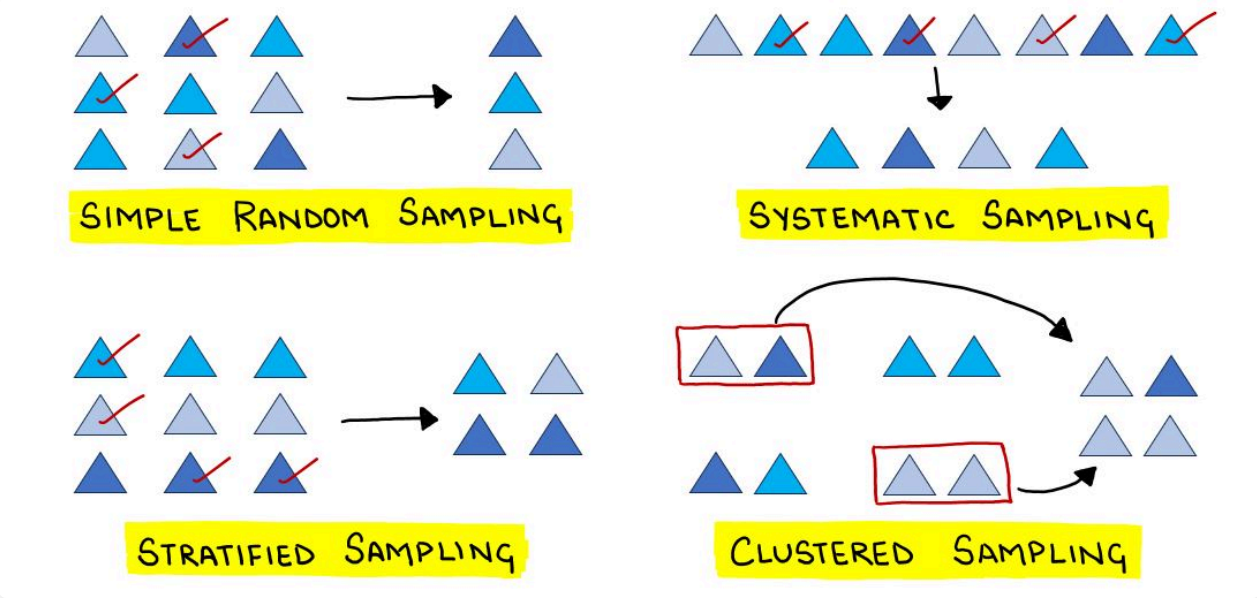
In non-probability sampling, the selection of elements from the population is not based on random selection or known probabilities. This method relies on the judgment of the researcher or other subjective criteria. Examples of non-probability sampling techniques include convenience sampling, quota sampling, purposive sampling, and snowball sampling. Non-probability sampling methods are less reliable for generalizing findings to the entire population due to the lack of known probabilities associated with sample selection.

Let us discuss various sub-types of these methods.

---

6. Methods of Sampling

Various types of probability sampling methods involve random selection, where each element in the population has a known and non-zero chance of being included in the sample. These methods aim to provide unbiased representations of the population.



Let's explain and provide an example for each type:

1. Simple Random Sampling

Simple random sampling is a method where every individual in the population has an equal probability of being chosen for the sample. This is achieved by selecting participants purely by chance, without any bias or specific criteria.

Example: A researcher interested in the opinions of college students about a new campus policy could assign each student in the college a unique number and use a random number generator to select a fixed number of students to participate in the survey.

2. Systematic Sampling

Systematic sampling involves selecting every kth element from the population, where k is a constant interval known as the sampling ratio. The starting point is randomly chosen, and then every kth element is included in the sample until the desired sample size is reached.

Example: In a population of 1000 students, a researcher wants a sample of 100. The sampling ratio is 1000/100 = 10. The researcher could randomly select a number between 1 and 10 as the starting point and then include every 10th student in the sample.

3. Stratified Sampling

Stratified sampling divides the population into distinct subgroups or strata based on certain characteristics. Then, a random sampling method, like simple random sampling, is applied independently within each stratum. This ensures representation from different strata in the final sample.

Example: A company wants to assess employee satisfaction in different departments (e.g., IT, HR, Marketing). They divide all employees into strata based on their department and then randomly select a sample from each department to ensure representation from all areas.

4. Cluster Sampling

Cluster sampling involves dividing the population into clusters or groups, and then randomly selecting entire clusters as the sampling units. Unlike stratified sampling, where individual members are selected, cluster sampling focuses on selecting entire groups. The Clusters from which, the sample is drawn randomly are heterogenous in nature.

Example: A health researcher wants to study the prevalence of a disease in a large city. They divide the city into several neighborhoods, randomly select a few neighborhoods, and then include all individuals within the selected neighborhoods in the study.

---

6. Methods of Sampling

Various types of non-probability sampling methods are used when it is not feasible to ensure a non-zero probability of inclusion for each element in the population. These methods may introduce bias in the study as the samples are selected based on judgment, convenience, or specific needs.

Let's explain each type and provide an example for better understanding:

1. Convenience Sampling

Convenience sampling involves selecting the sample based on what is most convenient for the researcher or interviewer. The researcher might choose participants who are easily accessible or readily available. This method is quick and easy to implement but can lead to a biased sample.

Example: A researcher studying the demand for non-aerated beverages might conduct interviews with shoppers in a few retail shops located nearby for convenience.

2. Purposive Sampling

Purposive sampling is used when the researcher has a specific purpose or target group in mind. The selection of participants is done intentionally to serve that particular need. It is non-representative and focuses on a specialized subset of the population.

Example: A researcher interested in studying the behavior of high-level business executives might specifically target and interview CEOs and top-level managers from various companies.

3. Snowball (Opportunity) Sampling

Snowball sampling is a method used when it is difficult to identify and reach members of a certain population. The researcher starts with a small number of participants and then asks them to refer others who fit the criteria. This process continues, like a snowball accumulating more snow, until the desired sample size is achieved.

Example: A researcher investigating drug users in a community might identify one willing participant who can refer other drug users in the same community, forming a snowball sample.

4. Judgment Sampling

In judgment sampling, the sample is selected based on the judgment or opinion of experts or experienced individuals. These experts identify participants they believe are representative of the population or possess the desired characteristics.

Example: A TV researcher seeking quick opinions about a political announcement might approach a diverse group of people on the street, based on their judgment, to get a cross-section of views.

5. Quota Sampling

Quota sampling involves dividing the population into subgroups based on specific characteristics and then setting a quota for each subgroup. The researcher then selects participants from each subgroup until the quota is filled. This method ensures representation of different subgroups but may not be fully representative of the entire population.

Example: A researcher interested in studying attitudes towards the death penalty in a city might set a quota to include a specific percentage of people from different religious backgrounds.

6. Dimensional Sampling

Dimensional sampling is an extension of quota sampling, where the researcher considers multiple characteristics such as gender, age, income, residence, and education. The aim is to ensure representation from each category or dimension of interest.

Example: A researcher conducting a survey about consumer preferences might ensure that they interview a specific number of participants from different age groups, income levels, and educational backgrounds to capture a diverse perspective.

---

7. SRSWOR and SRSWR

SRSWOR and SRSWR are different types of sampling techniques.

SRSWOR (Simple Random Sampling Without Replacement)

In SRSWOR, a sample is selected randomly without replacement from the population. It means that once an element is selected into the sample, it is not returned to the population, and thus, it cannot be chosen again.

This method ensures each element has an equal chance of being selected, but as elements are not replaced, subsequent selections are made from a reduced population.

SRSWR (Simple Random Sampling With Replacement)

SRSWR involves the random selection of elements from a population, but after each selection, the chosen element is returned to the population before the next selection.

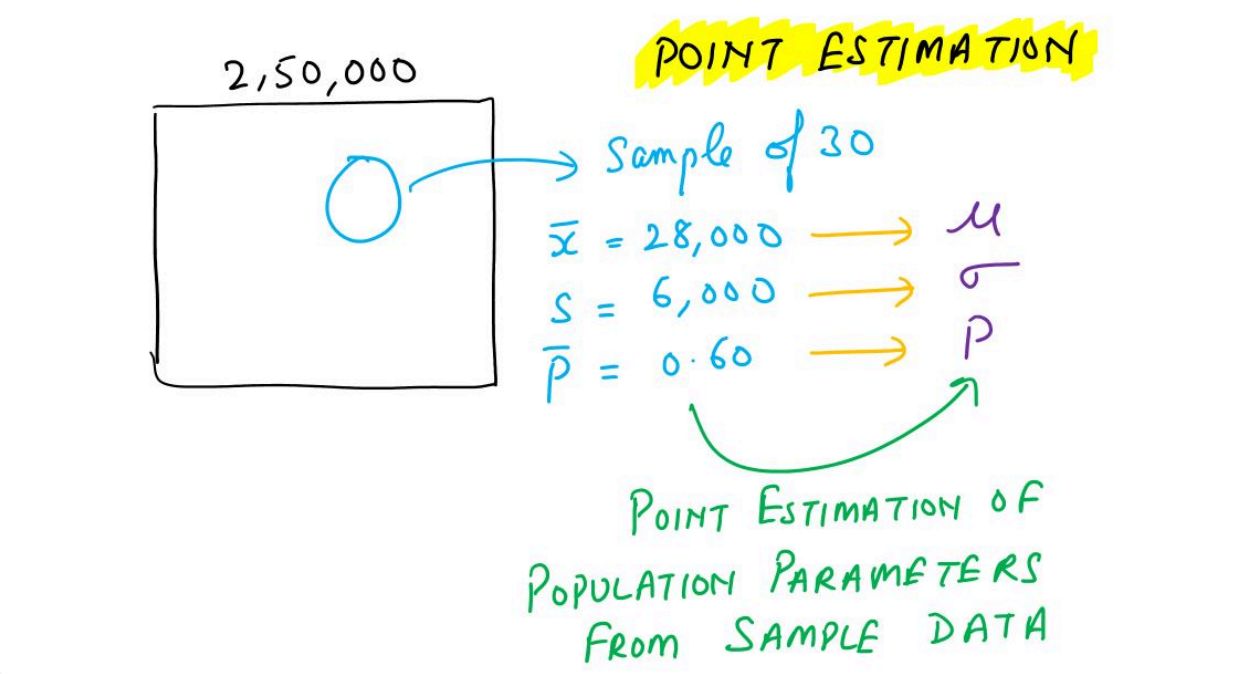
Unlike SRSWOR, elements can be selected more than once in SRSWR. Each selection is independent, and the chances of selecting an element remain the same in each draw.

Both techniques aim to ensure randomness in selecting samples from a larger population, but they differ in whether the selected elements are returned to the population for subsequent selections.

---

1. Point Estimation

Point estimation involves the use of sample data to estimate unknown population parameters. In situations where it's impractical or impossible to collect data from an entire population, a sample is taken, and specific statistics from that sample are used as estimators for the corresponding population parameters.



For instance, let us consider Infosys, a company with 2,50,000 employees. It is unfeasible to collect information from every employee, so a sample of 30 employees is taken to estimate three population parameters:
(i) average salary
(ii) standard deviation of salaries
(iii) proportion of employees who like Infosys.

From this sample of 30 employees, specific values are obtained:
Average salary: Rs 28,000
Standard deviation in salary: Rs 6000
Proportion of employees who like Infosys: 0.60 (18 out of 30 like Infosys)

By making the preceding computations, we perform the statistical procedure called point estimation. We refer to the sample mean $\bar{x}$ as the point estimator of the population mean μ, the sample standard deviation s as the point estimator of the population standard deviation σ, and the sample proportion $\bar{p}$ as the point estimator of the population proportion p. The numerical value obtained for $\bar{x}$, s, or $\bar{p}$ are called the point estimates.

These values derived from the sample are termed as **point estimators**, which serve as the point estimates for the respective population parameters.

It is important to note that these point estimates might differ somewhat from the actual population parameters due to the inherent variability in samples. The discrepancy between the point estimates and actual population parameters is expected since the estimates are derived from a sample and not the entire population.

To evaluate how close the point estimates are to the true population parameters, statisticians use the concept of interval estimation, which we will study later.

---

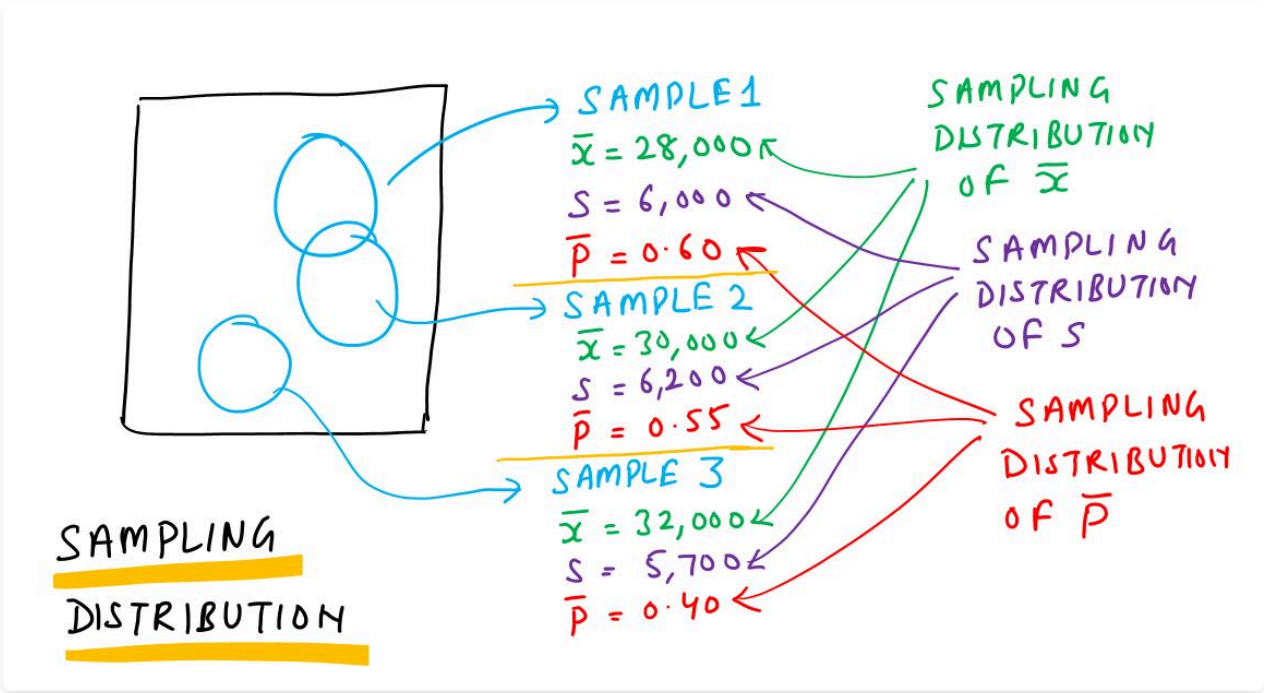
1. Point Estimation

There are 3 Properties of Point Estimators

- **Unbiased:** A property of a point estimator that is present when the expected value of the point estimator is equal to the population parameter it estimates.
- **Relative efficiency:** Given two unbiased point estimators of the same population parameter, the point estimator with the smaller standard error is more efficient.
- **Consistency"** A property of a point estimator that is present whenever larger sample sizes tend to provide point estimates closer to the population parameter.

---

2. Sampling Distribution



In the previous section on Infosys example, we got following values from our sample of 30 employees:
Average salary: Rs 28,000
Standard deviation in salary: Rs 6000
Proportion of employees who like Infosys: 0.60 (18 out of 30 like Infosys)

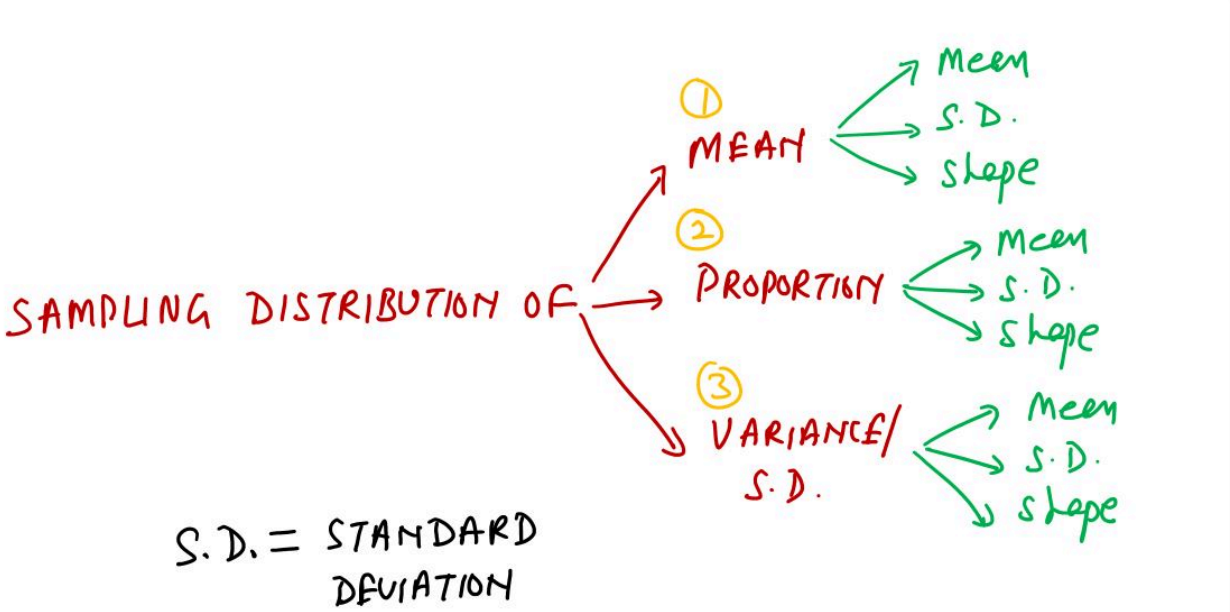
Now, we take another sample of 30 employees, and we get:
Average salary: Rs 30,000
Standard deviation in salary: Rs 6200
Proportion of employees who like Infosys: 0.55

Subsequently, a third sample of 30 employees was taken:
Average salary: Rs 32,000
Standard deviation in salary: Rs 5700
Proportion of employees who like Infosys: 0.40

Now, suppose we repeat the process of selecting a simple random sample of 30 employess over and over again (suppose 500 times), each time computing the values of $\bar{x}$, s and $\bar{p}$.

If we consider the process of selecting a simple random sample as an experiment, the sample mean $\bar{x}$ is the numerical description of the outcome of the experiment. Thus, the sample mean $\bar{x}$ is a random variable. As a result, just like other random variables, $\bar{x}$ has a mean or expected value, a standard deviation, and a probability distribution.

Because the various possible values of $\bar{x}$ are the result of different simple random samples, the probability distribution of $\bar{x}$ is called the **Sampling Distribution of $\bar{x}$ (Sampling Distribution of mean)**. Knowledge of this sampling distribution and its properties will enable us to make probability statements about how close the sample mean $\bar{x}$ is to the population mean μ.



Just like $\bar x$, we see that $\bar p$ is also a random variable. If every possible sample of size 30 were selected from the population and if a value of $\bar p$ were computed for each sample, the resulting probability distribution would be the **sampling distribution of $\bar p$ (sampling distribution of proportion)**

Just like $\bar x$ and $\bar p$, we see that s is also a random variable. If every possible sample of size 30 were selected from the population and if a value of s were computed for each sample, the resulting probability distribution would be the **sampling distribution of s (sampling distribution of standard deviation)**.

Thus, we understand that the probability distribution of any particular sample statistic is called the **sampling distribution of the statistic.** Similar to any other sampling distribution, it will possess its own Mean, Standard Deviation, and Shape.

Our subsequent focus will delve into examining the mean, standard deviation, and shape for the sampling distribution of various statistics.

---

3. Sampling Distribution of Mean

In the previous section we said that the sample mean $\bar x$ is a random variable and its probability distribution is called the sampling distribution of $\bar x$. Just as with other probability distributions we studied, the sampling distribution $\bar x$ has
(a) an expected value or mean,
(b) a standard deviation, and
(c) a characteristic shape or form.

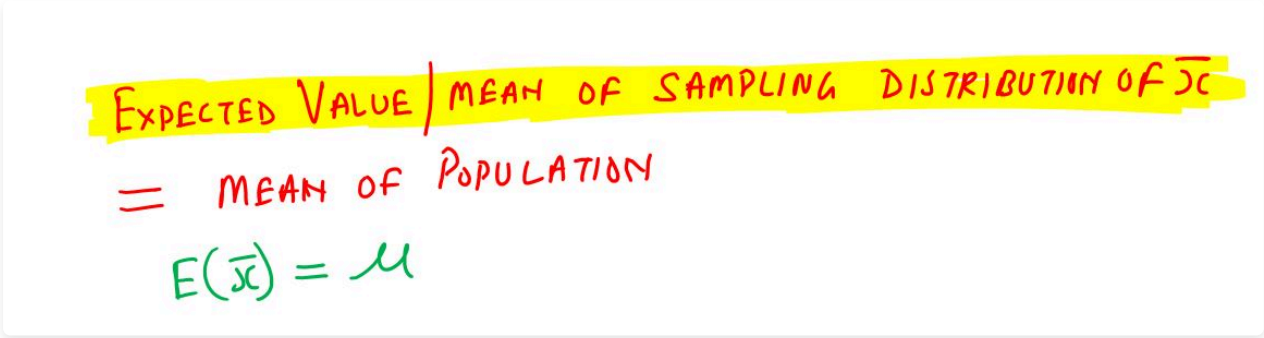
Let us discuss all three of them, one by one.

---

3. Sampling Distribution of Mean

Expected Value (Mean) of Sampling Distribution of Mean (\(\bar x \))

The expected Value of \(\bar x \) equals the mean of the population (μ) from which the sample is selected.



E (\(\bar x \))= μ, where E (\(\bar x \)) is Expected Value of \(\bar x \) and μ is Mean of Population.

---

3. Sampling Distribution of Mean

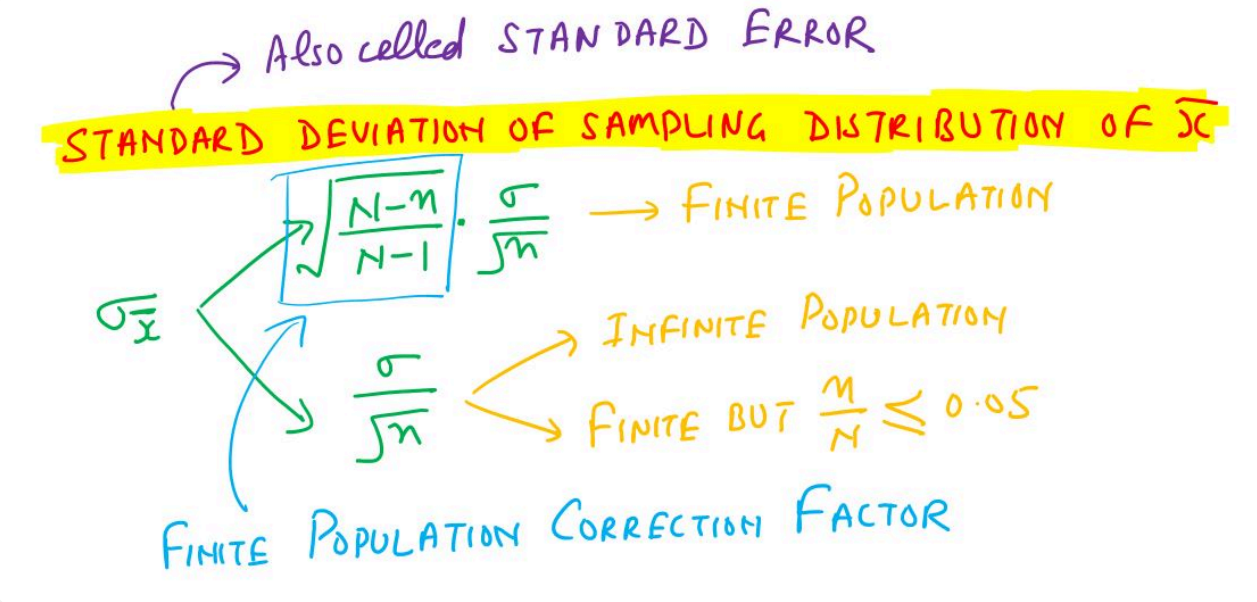
Standard Deviation of Sampling Distribution of Mean ($\bar x$)

The Standard Deviation of Samling Distribution of $\bar x$ (represented by $\sigma_{\bar x}$)is given by following formulas:

$\sigma_{\bar x} = \sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}}$ in case of finite population with size N.

$\sigma_{\bar x} = \frac{\sigma}{\sqrt{n}}$ in case of infinite population or when the population is finite, but the sample size is less than or equal to 5% of the population size ($\frac{n}{N} \leq 0.05$).

Here n is size of sample in all these formulas.



$\sqrt{\frac{N-n}{N-1}}$ is called **Finite Population Correction Factor.** If population is large and relatively, the sample is samll, its value tends to be 1.

Standard Deviation of Sampling Distribution of $\bar x$ (represented by $\sigma_{\bar x}$) is also called **Standard Error of Mean.**

In general, the "term standard error" refers to the "standard deviation of a point estimator". The value of the standard error of the mean is helpful in determining how far the sample mean may be from the population mean.

The standard error of mean also depends upon, whether sampling is done with or without replacement.

When sampling is done with replacement
$\sigma_{\bar x} = \frac{\sigma}{\sqrt{n}}$

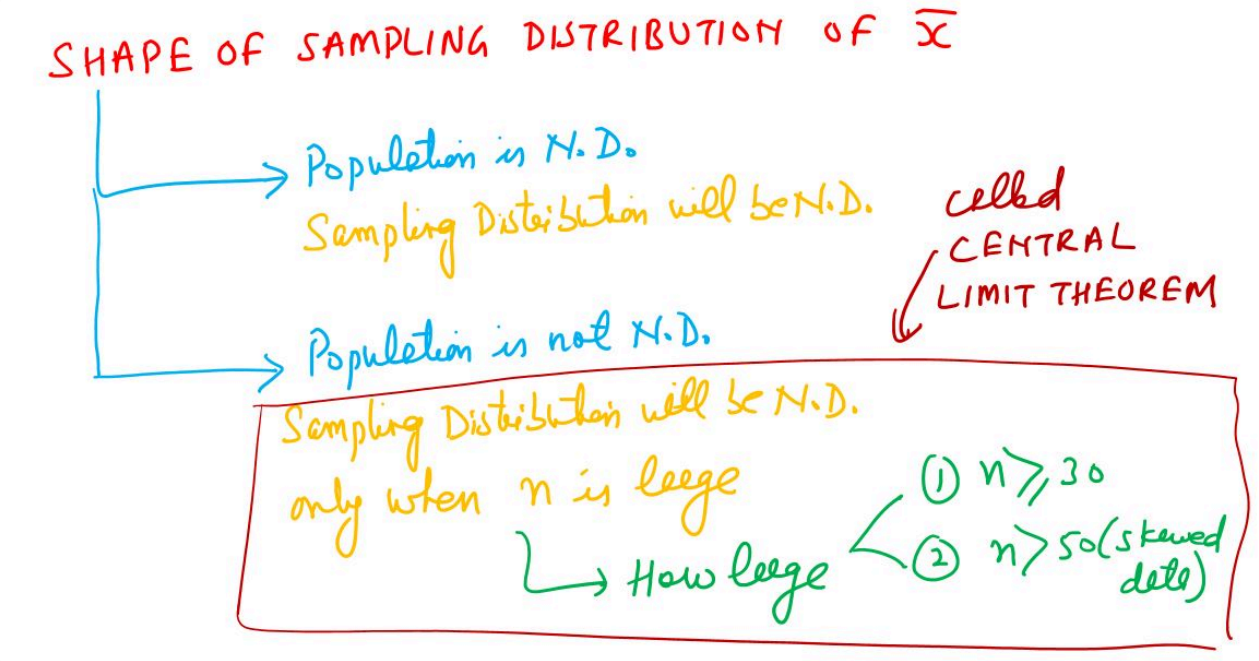
When sampling is done without replacement
$\sigma_{\bar x} = \sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}}$

---

3. Sampling Distribution of Mean

Shape of Sampling Distribution of Mean (\(\bar x \))

Let us now determine the form or shape of the sampling distribution of \(\bar x \).



We will consider two cases:

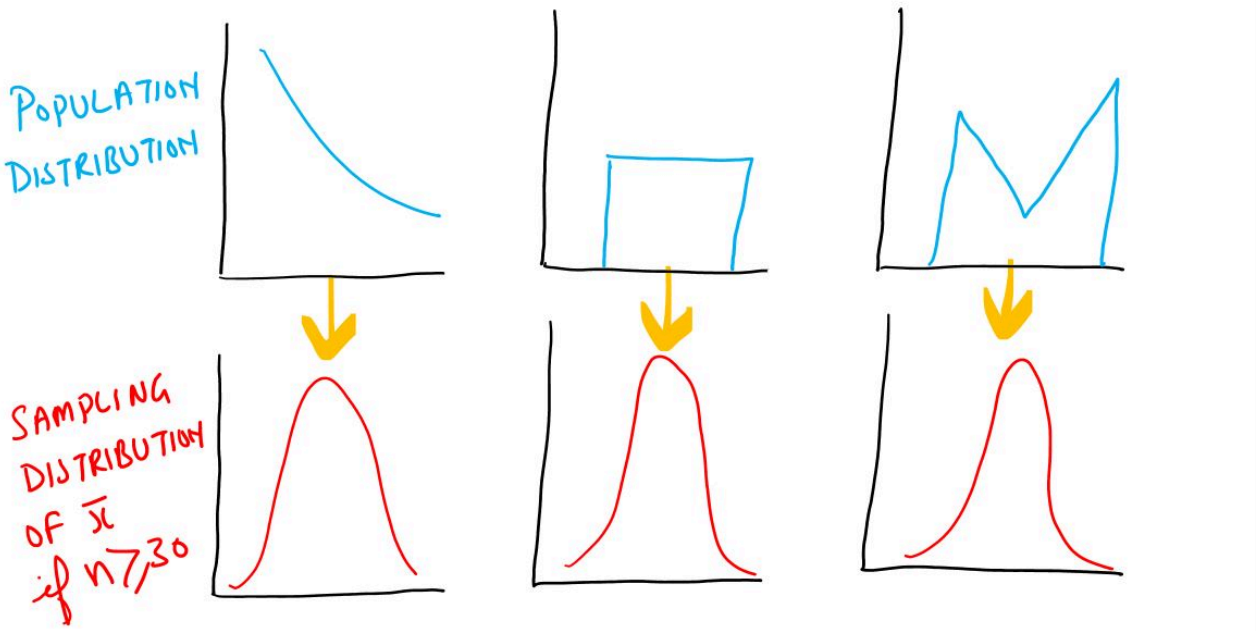
(i) Population has a normal distribution: When the population has a normal distribution, the sampling distribution of \(\bar x \) is normally distributed for any sample size.

(ii) Population does not have a normal distribution: When the population from which we are selecting a random sample does not have a normal distribution, the central limit theorem (CLT) is helpful in identifying the shape of the sampling distribution of \(\bar x \).

Central Limit Theorem

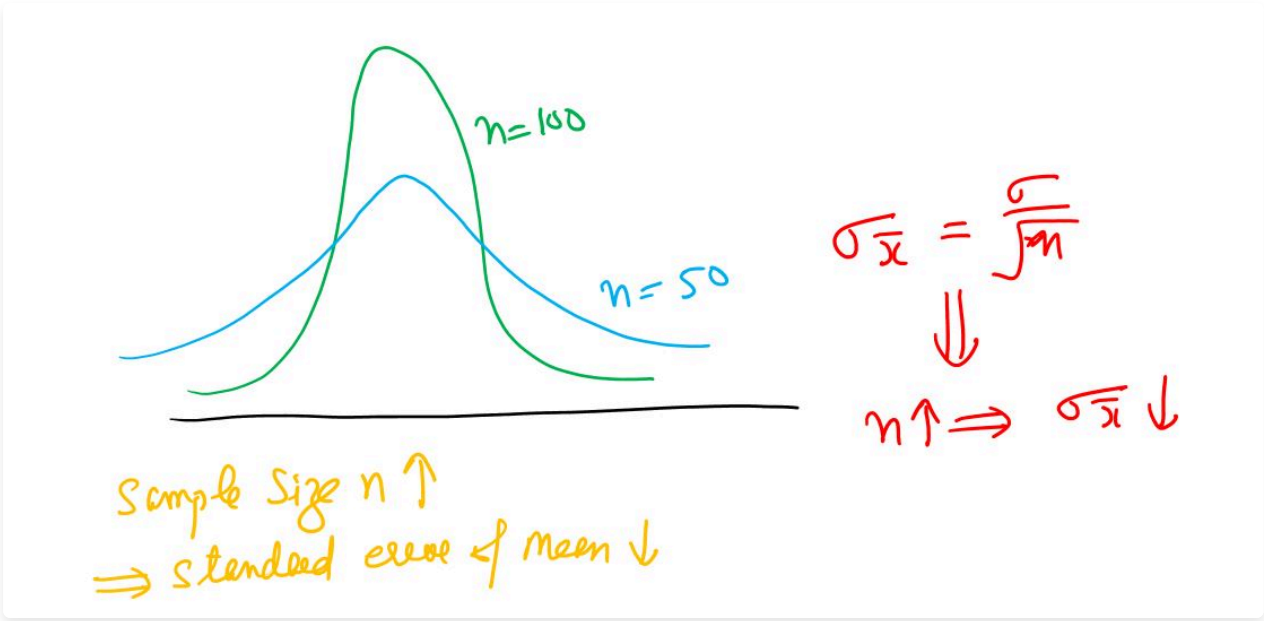
Central Limit Theorem states that "In selecting random samples of size n from a population, the sampling distribution of the sample mean \(\bar x \) can be approximated by a normal distribution as the sample size becomes large".

Please note that, for most applications, the sampling distribution of \(\bar x \) can be approximated by a normal distribution whenever the **sample is size 30 or more**. In cases where the population is highly skewed or outliers are present, samples of size 50 may be needed.


POPULATION DISTRIBUTION
SAMPLING DISTRIBUTION OF $\bar{x}$ if $n \geq 30$

3. Sampling Distribution of Mean

Let us examine the influence of sample size,n on the Expected Value (E (\(\bar x \))) and Standard Deviation of the sampling distribution of the Mean, \(\sigma_ \bar x \) (Standard Error).



The Expected Value of the sampling distribution of the Mean (E (\(\bar x \))) remains unaffected by the sample size,n. Irrespective of the sample size, the Expected Value of the sampling distribution of the Mean (E (\(\bar x \))) remains equal to the Population Mean (μ).

However, the Standard Deviation of the sampling distribution of the Mean, \(\sigma_ \bar x \) (Standard Error) is contingent on the sample size.

\(\sigma_ \bar x \) = \(\frac{ \sigma }{ \sqrt{n}} \)

With an increase in sample size, n, the Standard Deviation of the sampling distribution of the Mean, \(\sigma_ \bar x \) decreases. This occurs because a larger sample size enhances the likelihood of the sample mean falling within a specified range of the population mean.

---

3. Sampling Distribution of Mean

Illustration 1

A spark plug manufacturer claims that the lives of its plugs are normally distributed with a mean of 60,000 miles and a standard deviation of 4,000 miles. A random sample of 16 plugs had an average life of 58,500 miles. If the manufacturer's claim is correct, what is the probability of finding a sample mean of 58,500 or less?

Solution:

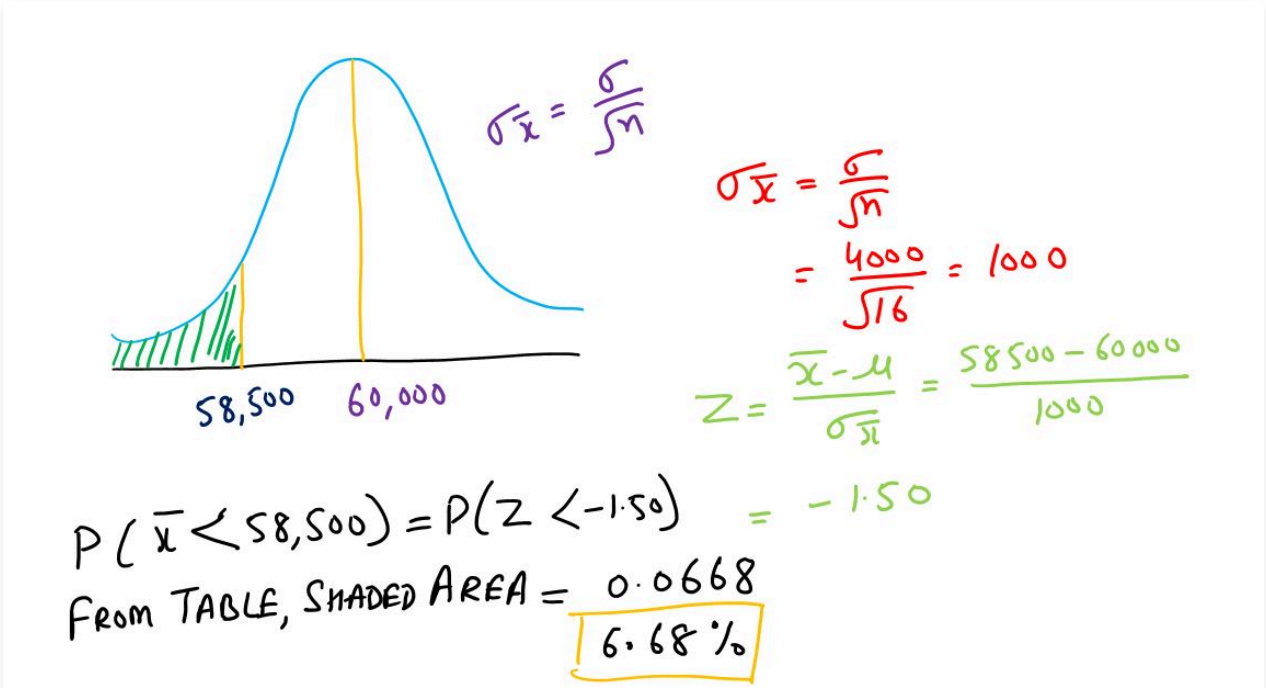


Illustration 2

Suppose that during any hour in a large department store, the average number of shoppers is 448, with a standard deviation of 21 shoppers. What is the probability that a random sample of 49 different shopping hours will yield a sample mean between 441 and 446 shoppers?

Solution:

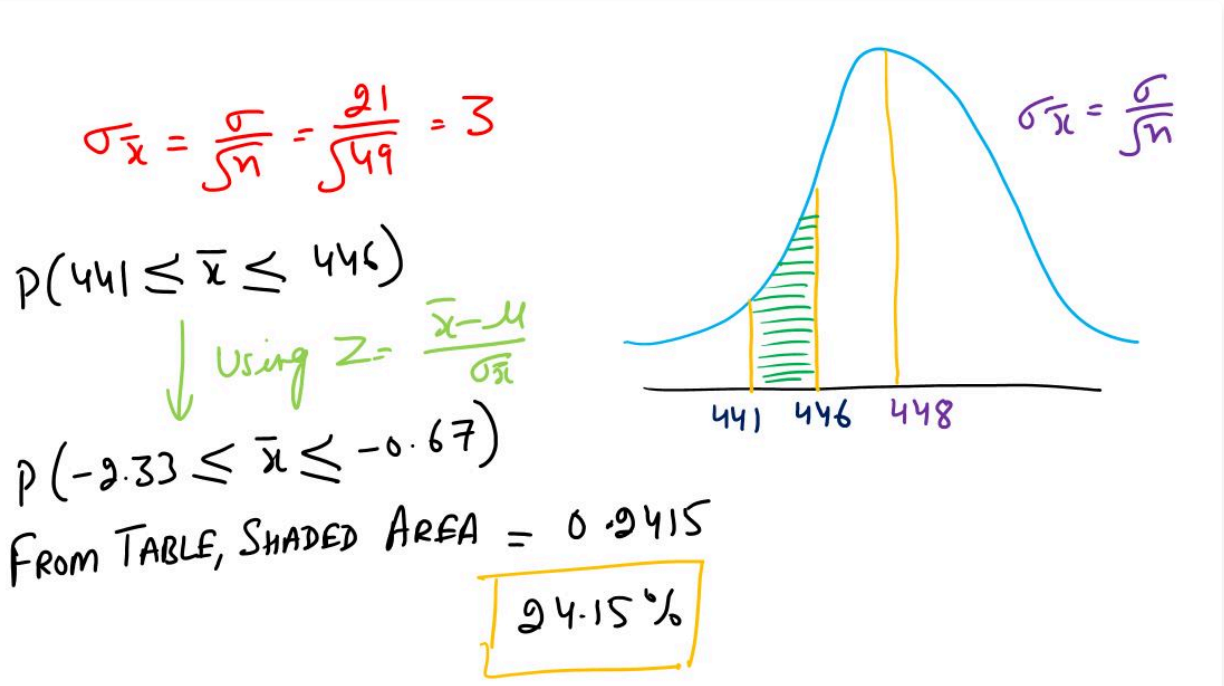


Illustration 3

A production company's 350 hourly employees average 37.6 years of age, with a standard deviation of 8.3 years. If a random sample of 45 hourly employees is taken, what is the probability that the sample will have an average age of less than 40 years?

Solution:

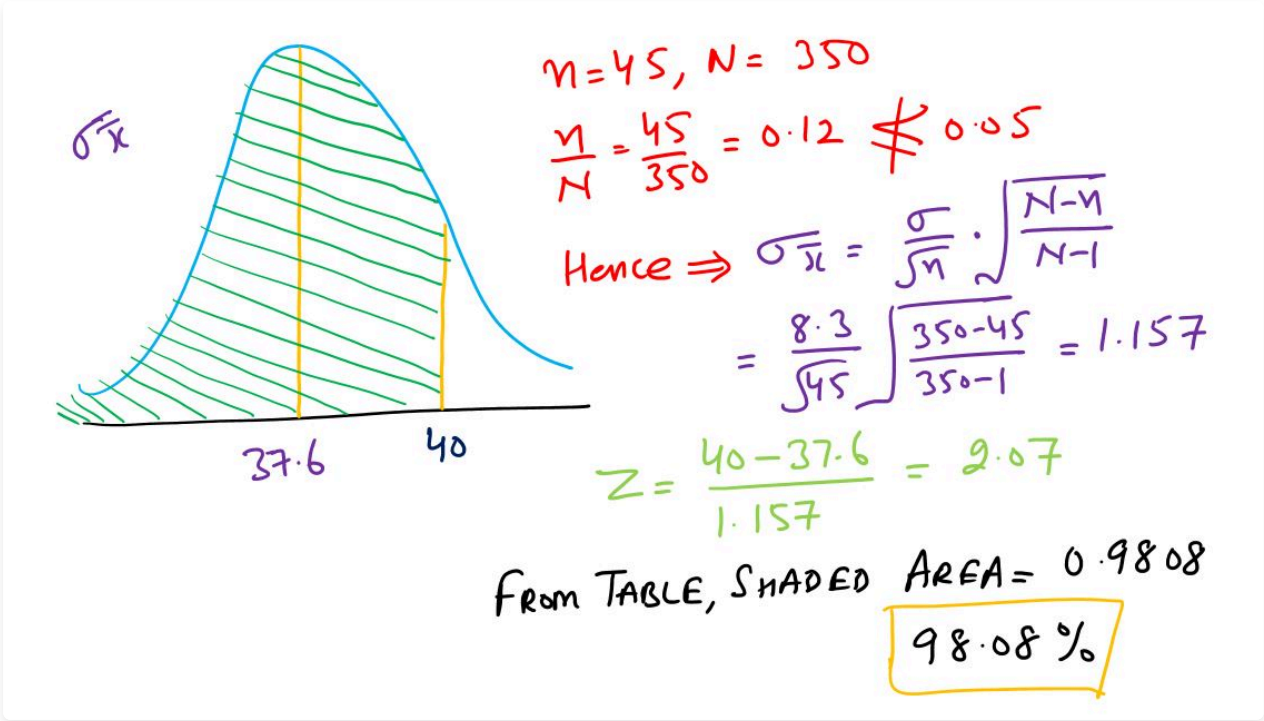


Illustration 4

The standard deviation of population of 2500 employees is σ = 4000. The sample size is 30. Calculate Standard Error of Mean.

Solution:
We calculate $\frac{n}{N} = \frac{30}{2500} = 0.012$, which is less than 0.05.

So, the Formula for calculating Standard Deviation of Sampling distribution of $\bar x$ (also called Standard Error of Mean) is:

$\sigma_{\bar x} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{30}} = 730.3$

---

4. Sampling Distribution of Proportion

Just like sample mean $\bar x$, sample proportion $\bar p$ is also a random variable and its probability distribution is called the sampling distribution of proportion $\bar p$. The sampling distribution of $\bar p$ is the probability distribution of all possible values of the sample proportion $\bar p$.

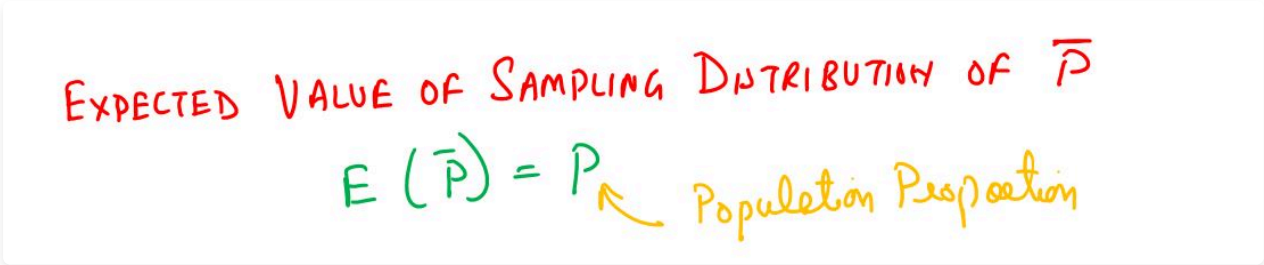
Just as with other probability distributions we studied, the sampling distribution of $\bar p$ will have:
(a) an expected value or mean,
(b) a standard deviation, and
(c) a characteristic shape or form.

Let us discuss all three of them, one by one.

---

4. Sampling Distribution of Proportion

Expected Value of sampling Distribution of \(\bar p \)

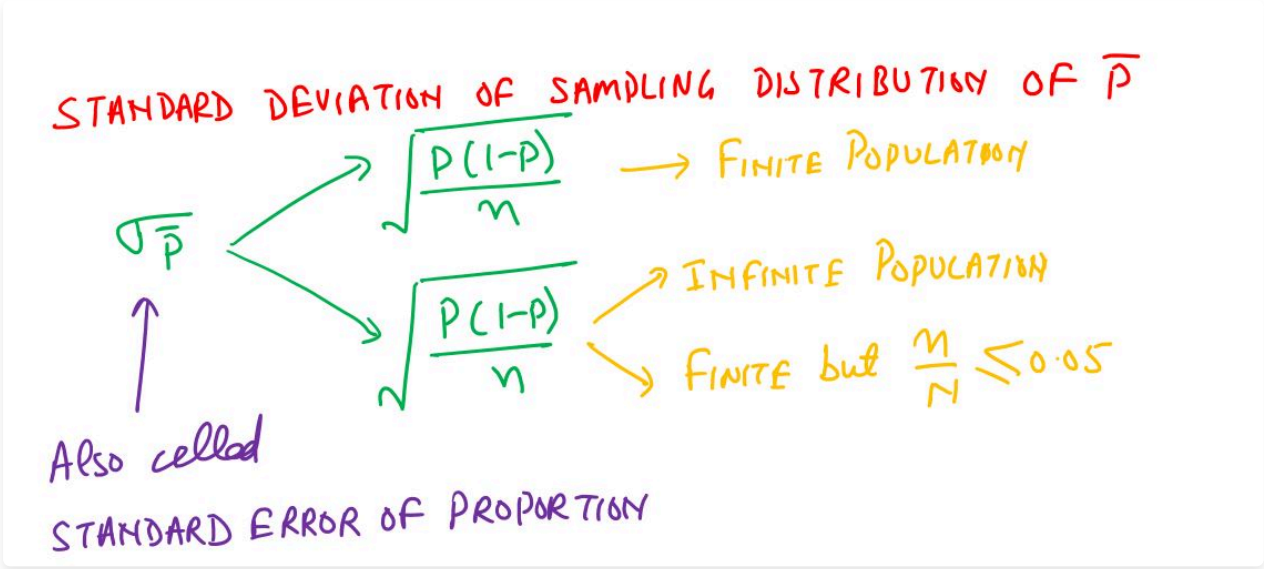


The Expected Value of sampling distribution of \(\bar p \) equals the population proportion from which the sample is selected.

E (\(\bar p \))= p
where E (\(\bar p \)) is Expected Value of Sampling distribution of \(\bar p \)and 'p' is Population Proportion.

---

4. Sampling Distribution of Proportion



The Standard Deviation of Sampling Distribution of \(\bar p \) (represented by \(\sigma_\bar p \)) is given by following formulas:

\(\sigma_\bar p \) = \(\sqrt{ \frac{N-n}{N-1}} \times \sqrt{ \frac{p(1-p)}{n}} \) .....in case of Finite Population with Size N

\(\sigma_\bar p \) = \(\sqrt{ \frac{p(1-p)}{n} } \) .....in case of Infinite Population or when the population is finite and but the sample size, n is less than or equal to 5% of the population size (N).

Standard Error

Standard Deviation of Sampling Distribution of \(\bar p \) (represented by \(\sigma_\bar p \)) is also called **Standard Error of Proportion.**

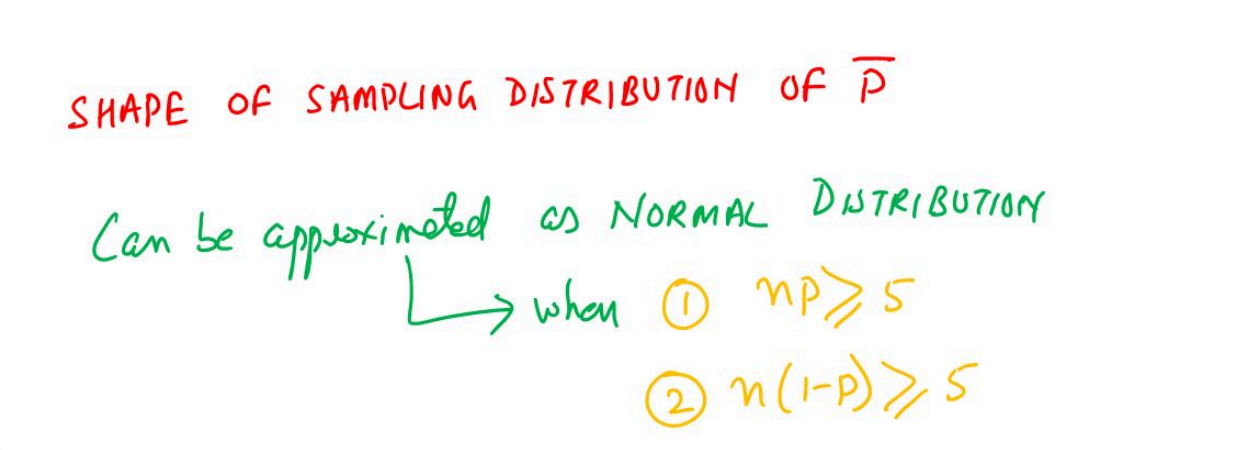
---

4. Sampling Distribution of Proportion

Shape of sampling Distribution of $\bar p$

Let us now discuss the form or shape of the sampling distribution of $\bar p$.

For a simple random sample from a large population, the value of x is a binomial random variable indicating the number of elements in the sample with the characteristic of interest. Because n is a constant, the probability of $\frac{x}{n}$ is the same as the binomial probability of x, which means that the sampling distribution of $\bar p$ is also a discrete probability distribution and that the probability for each value of x/n is the same as the probability of x.



A binomial distribution can be approximated by a normal distribution whenever the sample size is large enough to satisfy the following two conditions:
(i) np ≥ 5 and
(ii) n(1 - p) ≥ 5.

Hence, the sampling distribution of $\bar p$ can be approximated by a normal distribution whenever **np ≥ 5 and n(1 - p) ≥ 5.** Both conditions should be satisfied.

---

4. Sampling Distribution of Proportion

Illustration 1

It has been estimated that 43% of employees at IBM are happy with food quality in office. Find the probability that more than one-half of a random sample of 80 employees are happy with food quality in office.

Solution:

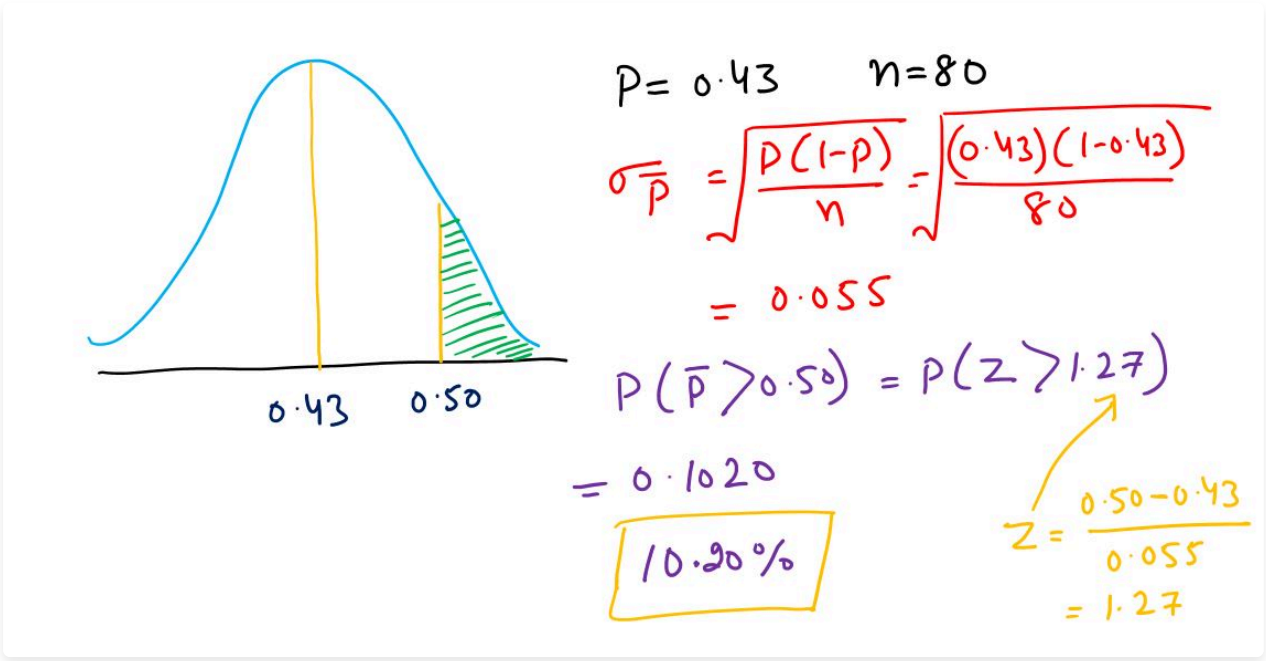


Illustration 2

Delhi Fire Service did a random sample of 270 homes from a large population of homes to estimate the proportion of homes with no fire safety measures. If, in fact, 20% of the homes have no fire safety measures, what is the probability that the sample proportion will be between 16% and 24%?

Solution:

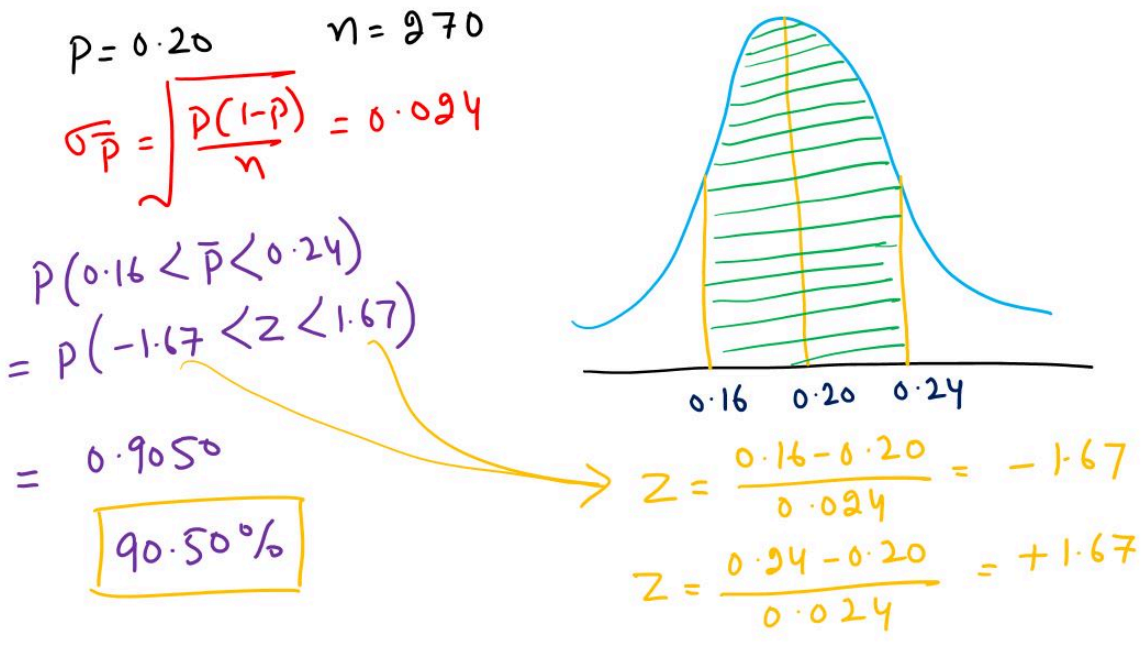


Illustration 3

In the last election, one politician received 52% of the votes cast. One year after the election, the he organized a survey that asked a random sample of 300 people whether they would vote for him in the next election. If we assume that his popularity has not changed, what is the probability that more than half of the sample would vote for him?

Solution:

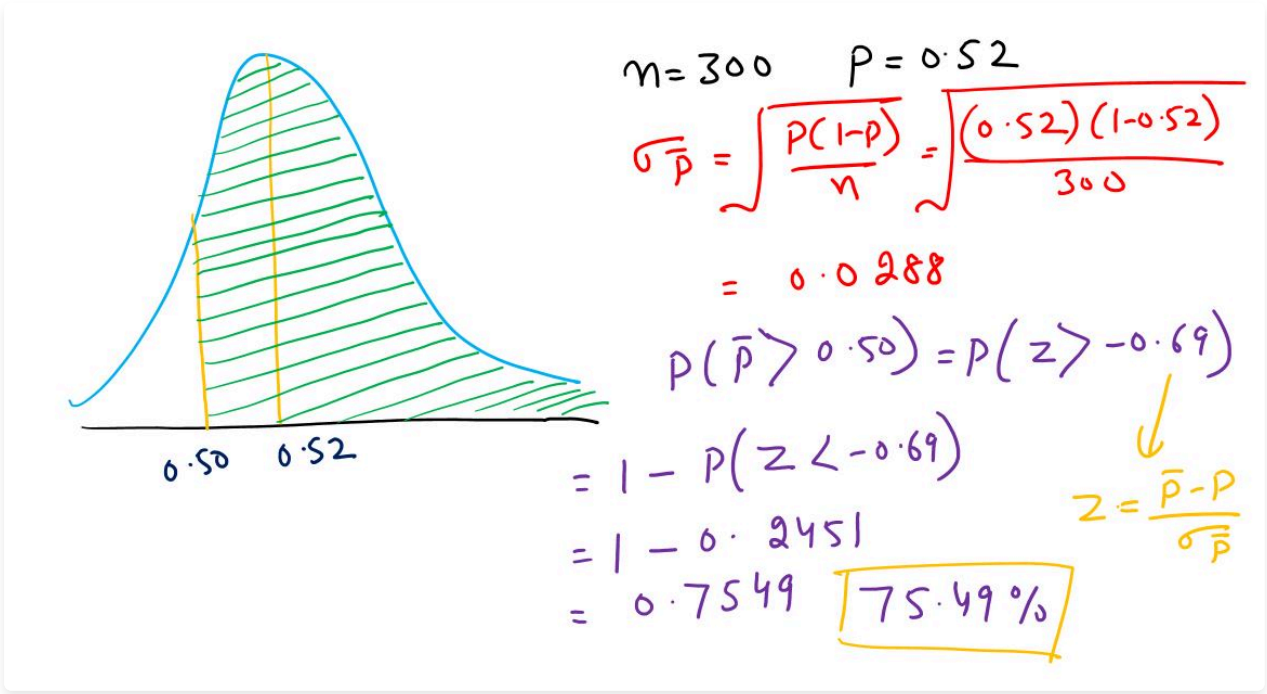


Illustration 4

The Population Proportion for a population of 2500 employees is p = 0.60. The sample size, n is 30. Calculate the standard error.

Solution:

We calculate $\frac{n}{N} = \frac{30}{2500}$ = 0.012, which is less than 0.05.

So, the Formula for calculating Standard Deviation of $\bar p$ (also called Standard Error of Proportion) is:

$\sigma_{\bar p}$ = $\sqrt{\frac{p(1-p)}{n}}$

= $\sqrt{\frac{0.60\ (1-0.60)}{30}}$ = 0.0894

---

1. Interval Estimate

Because a point estimator cannot be expected to provide the exact value of the population parameter, an **interval estimate** is often computed by adding and subtracting a value, called the margin of error, to the point estimate. The general form of an interval estimate is as follows:

Interval Estimate = Point estimate ± Margin of error

The purpose of an interval estimate is to provide information about how close the point estimate, provided by the sample, is to the value of the population parameter. The interval estimate is also called Confidence Interval or simply Estimate.

The difference between the point estimate and actual parameter value (true value of population) is called **Sampling error**. It is the error that results, because we take one sample rather than taking a census of the entire population. It is also called Estimation error.

In this section, we will learn how to calcute:

(1) Interval estimate of the population mean (we will discuss 2 sub-cases, when σ is known and when σ is not known)

(2) Interval estimate of the population proportion.

---

2. Interval estimate of population mean - σ known

In order to develop an interval estimate of a population mean, there are two cases:
(i) population standard deviation σ is known
(ii) population standard deviation 'σ' is not known and instead sample standard deviation 's' is used.

Let us start with interval estimate of a population mean, when σ is known. Let us understand this with an example.

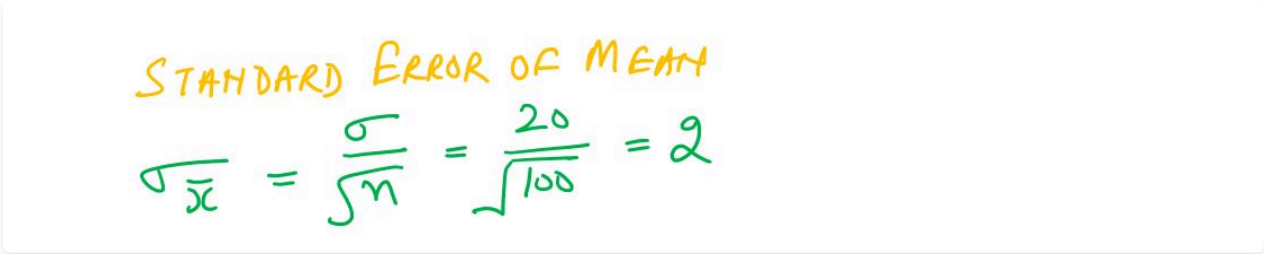
Suppose we are analyzing average spending of customers in a shopping mall. Let x be average spending per customer. From historic data, we know that standard deviation of population, σ = Rs 20.

On one fine day, we took sample of 100 customers, n =100 and found that average spending is $\bar{x}$ = Rs 82.

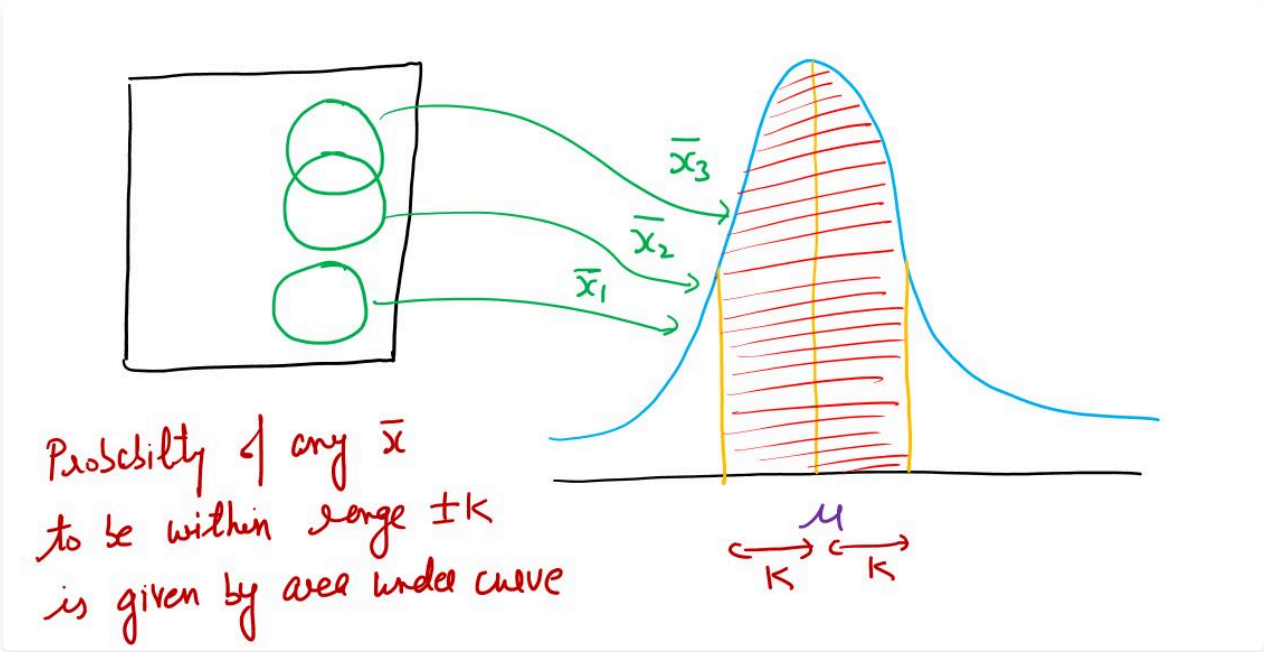
This sample mean amount (Rs 82) spent provides a point estimate of the population mean amount spent per shopping trip, μ.

Let us learn to compute the **margin of error** for this estimate and develop an **interval estimate** of the population mean, μ.

As learnt earlier, the Standard Error of Mean will be:



Using the standard normal probability table, we find that 95% of the values of any normally distributed random variable are within ±1.96 standard deviations of the mean. Thus, when the sampling distribution of $\bar{x}$ is normally distributed, 95% of the values must be within ± 1.96 × $\sigma_{\bar{x}}$ of the mean μ.



In our example, we just calculated that the sampling distribution of $\bar{x}$ is normally distributed with a standard error of $\sigma_{\bar{x}}$ = 2.

±1.96 × $\sigma_{\bar{x}}$ = ±1.96 × 2 = ±3.92

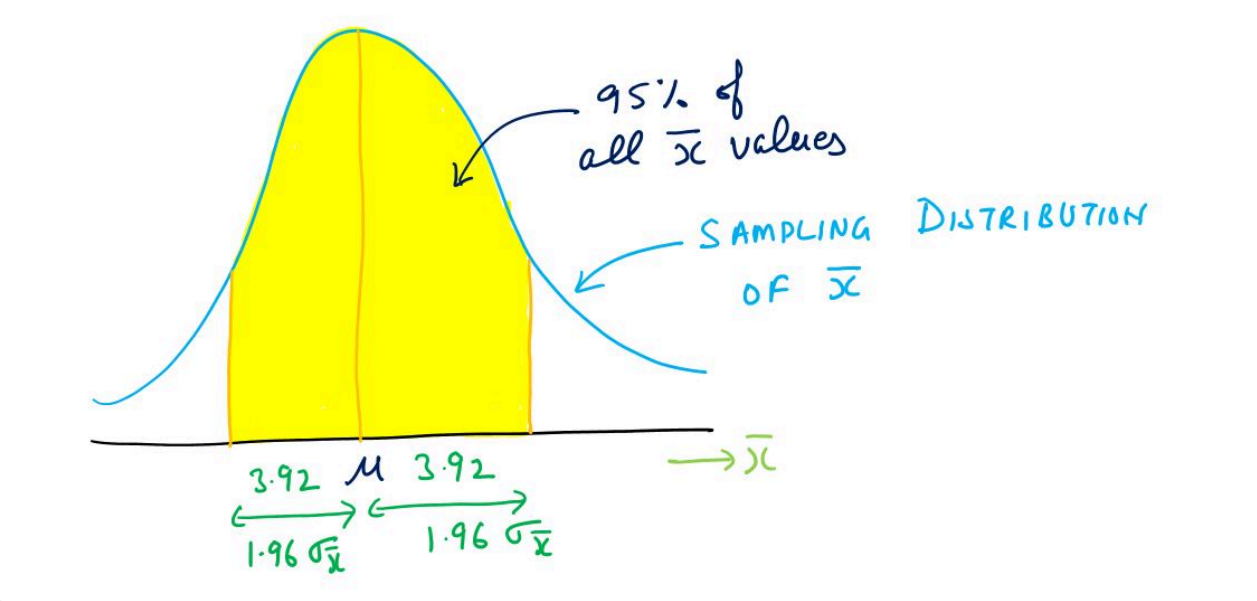
Thus, we can conclude that 95% of all $\bar{x}$ values obtained using a sample size of n = 100 will be within ± 3.92 of the population mean μ.

The **margin of error**, in the example, is equal to 3.92. Using $\bar{x}$ ± 3.92 to construct the interval estimate, we obtain 82±3.92 (average spending of sample of 100 customers is 82).

Thus, the **interval estimate** of μ based on the data from the sample of 100 customers is 82 - 3.92 = 78.08 to 82 + 3.92 = 85.92.

Because 95% of all the intervals constructed using $\bar x$ ± 3.92 will contain the population mean, we say that we are 95% confident that the interval 78.08 to 85.92 includes the population mean μ.

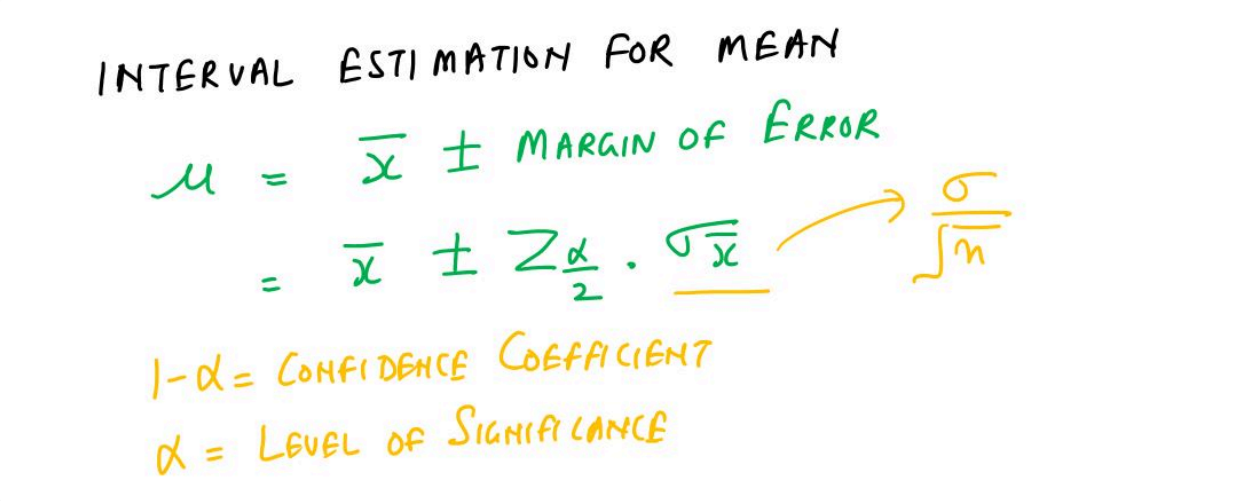
In other words, we say that this interval has been established at the **95% confidence level**. The value 0.95 is referred to as the **confidence coefficient**, and the interval 78.08 to 85.92 is called the 95% confidence interval. Please look at the figure to understand the area covered under 95%.



In this example, we calculated Margin of Error for 95% confidence level as ± 1.96 × $\sigma_{\bar x}$, which may also be written as ± 1.96 × $\frac{\sigma}{\sqrt{n}}$.

The **Margin of error** for any other percentage of confidence level can be calculated using Formula ± $z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$

Thus, general formula of an **interval estimate** of a population mean is given by $\bar x$ ± $z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$



where (1 - α) is the confidence coefficient and $z_{\frac{\alpha}{2}}$ is the z value providing an area of $\frac{\alpha}{2}$ in the upper tail of the standard normal probability distribution.

Although a 95% confidence level is frequently used, other confidence levels such as 90% and 99% may be considered. Look at the table to understand α and $z_{\frac{\alpha}{2}}$ value of other Confidence levels.

Confidence Level	α	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$
90%	0.10	0.05	1.645

95%	0.05	0.025	1.960
99%	0.01	0.005	2.576

In our example, 99% confidence interval can be calculated as
= 82 ± 2.576 × \(\frac{20}{ \sqrt{100} } \)
= 82 ± 5.15

Thus, at 99% confidence, the **margin of error** is 5.15 and the **confidence interval** is 82 - 5.15 = 76.85 to 82 + 5.15 = 87.15.

---

2. Interval estimate of population mean - σ known

Illustration 1

Suppose that shopping times for customers at a local mall are normally distributed with known population standard deviation of 20 minutes. A random sample of 64 shoppers in the local grocery store had a mean time of 75 minutes. Find the standard error, margin of error, and the upper and lower confidence limits of a 95% confidence interval for the population mean.

Solution:

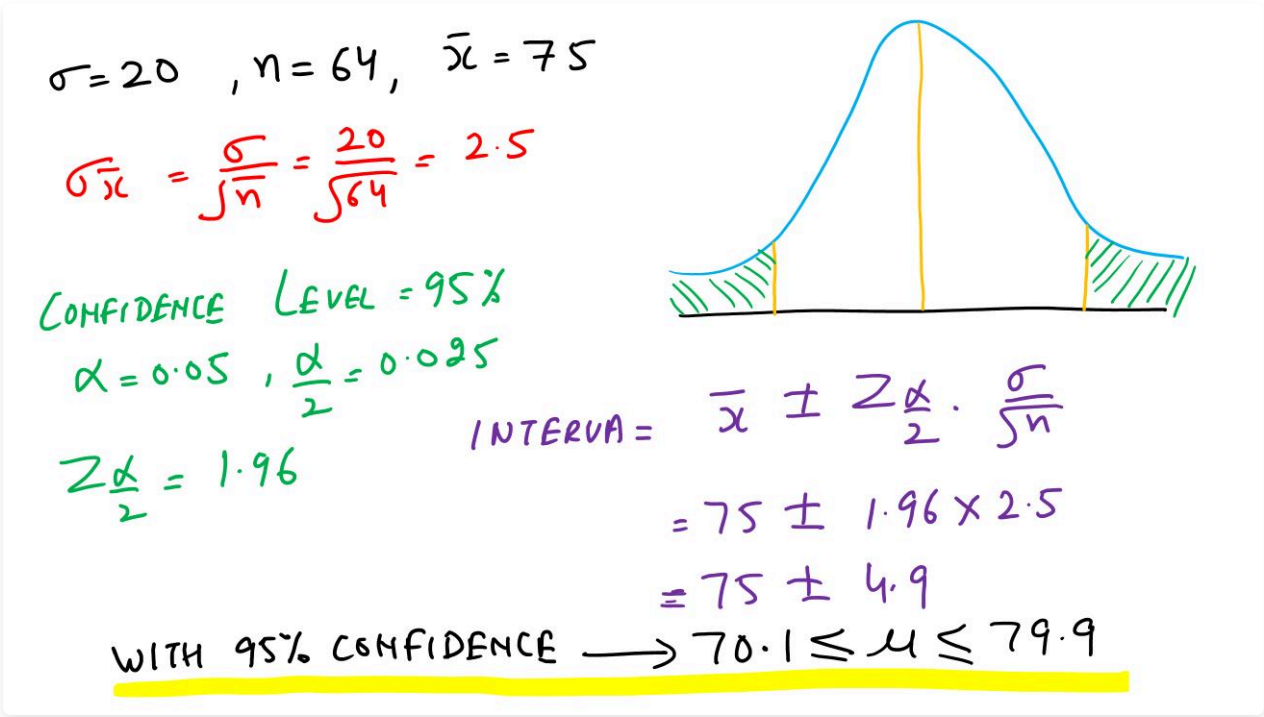
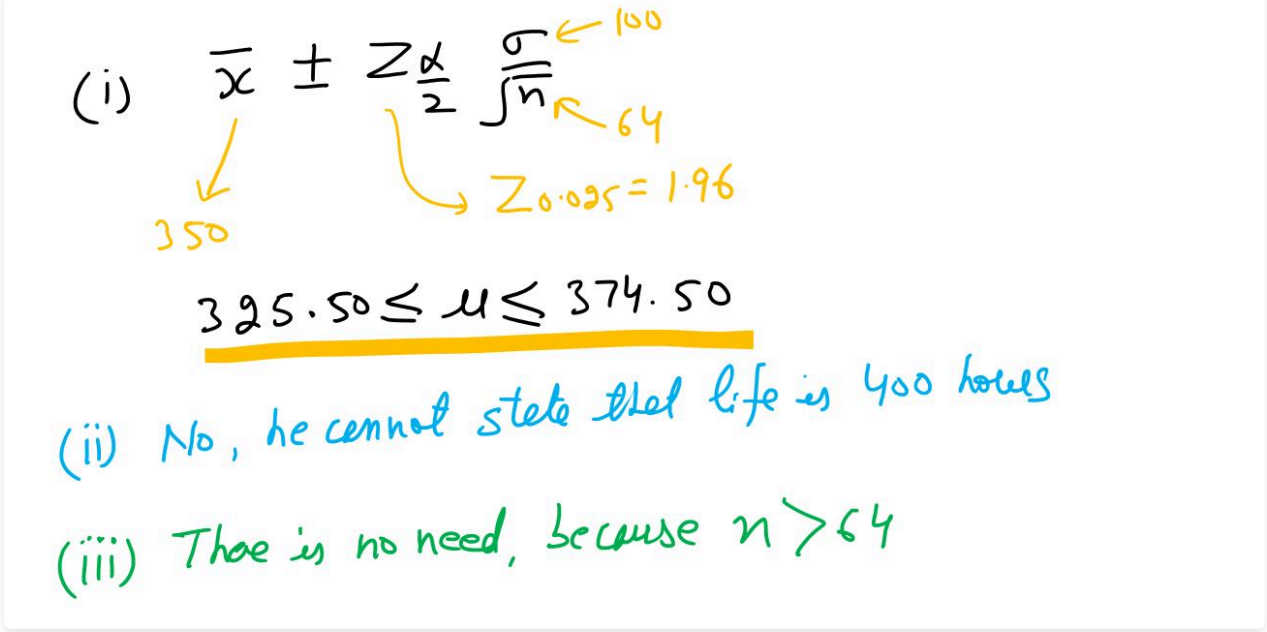


Illustration 2

The quality control manager at a light bulb factory needs to estimate the mean life of a large shipment of light bulbs. The standard deviation is 100 hours. A random sample of 64 light bulbs indicated a sample mean life of 350 hours.
(i) Construct a 95% confidence interval estimate for the population mean life of light bulbs in this shipment.
(ii) Do you think that the manufacturer has the right to state that the light bulbs have a mean life of 400 hours? Explain.
(iii) Must you assume that the population light bulb life is normally distributed? Explain.

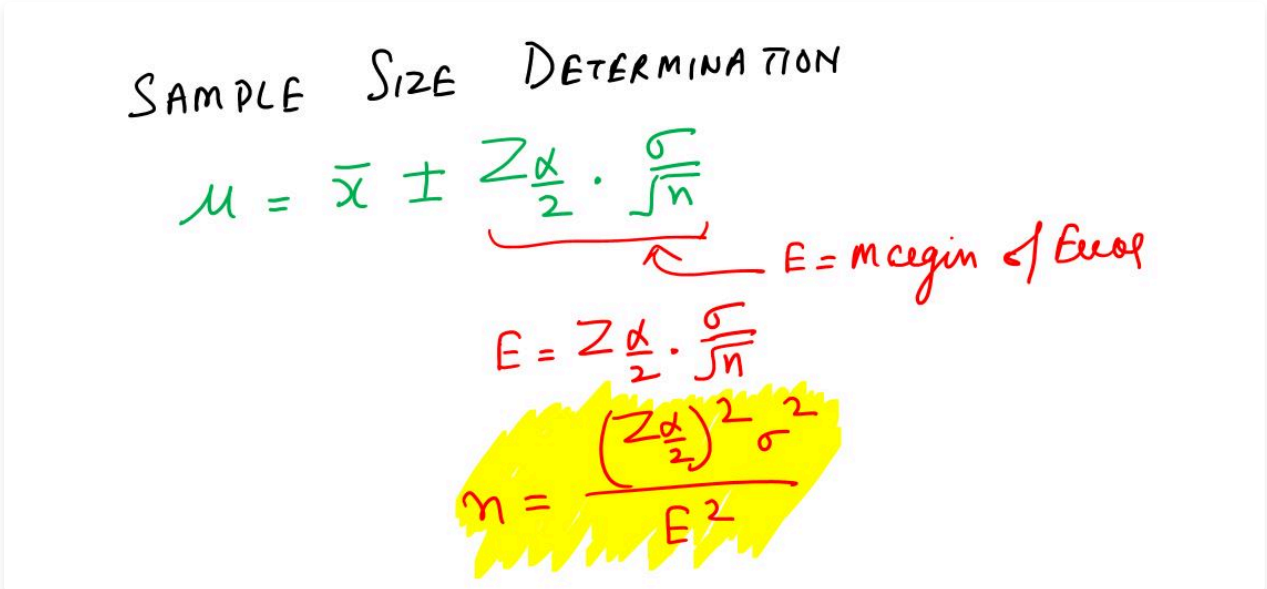
Solution:



---

2. Interval estimate of population mean - σ known

Let us learn, how to choose a sample size large enough to provide a desired margin of error (E).



Let E be desired Margin of Error. Then E is given by

\(E = z_ \frac{ \alpha}{2} \times \frac{ \sigma}{ \sqrt{n} } \)

Solving for n, we get

\(n = \frac{(z_ \frac{ \alpha}{2})^2 \times \sigma^2}{E^2} \)

This sample size provides the desired margin of error at the chosen confidence level.

---

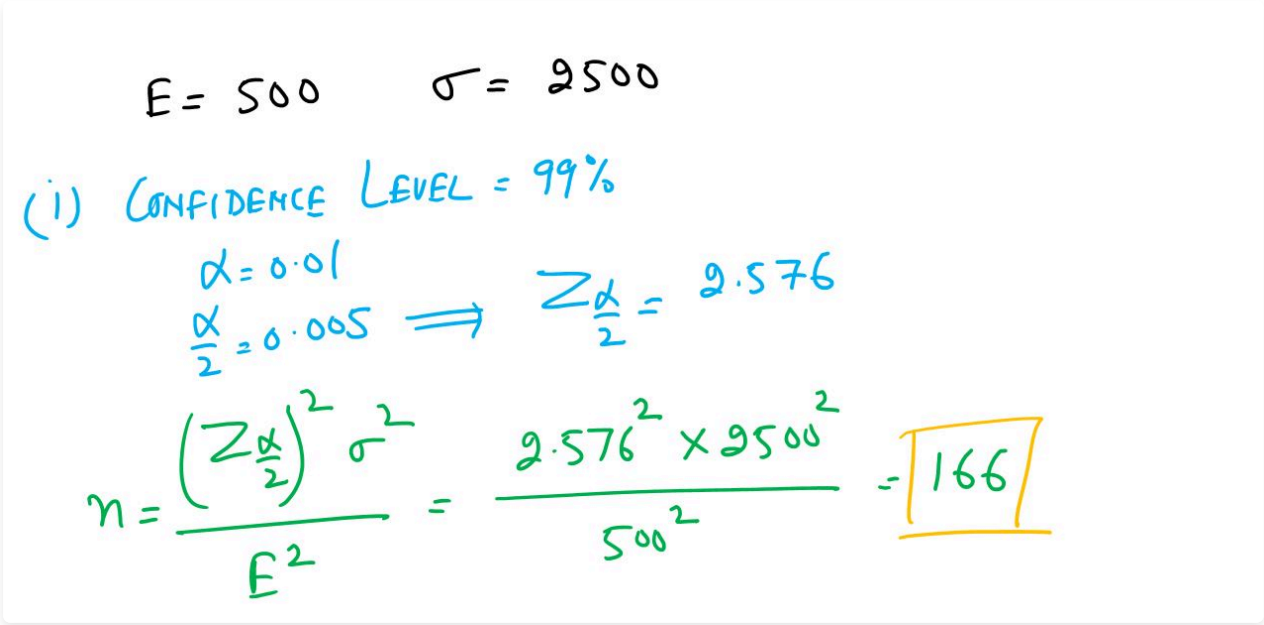
2. Interval estimate of population mean - σ known

Illustration 1

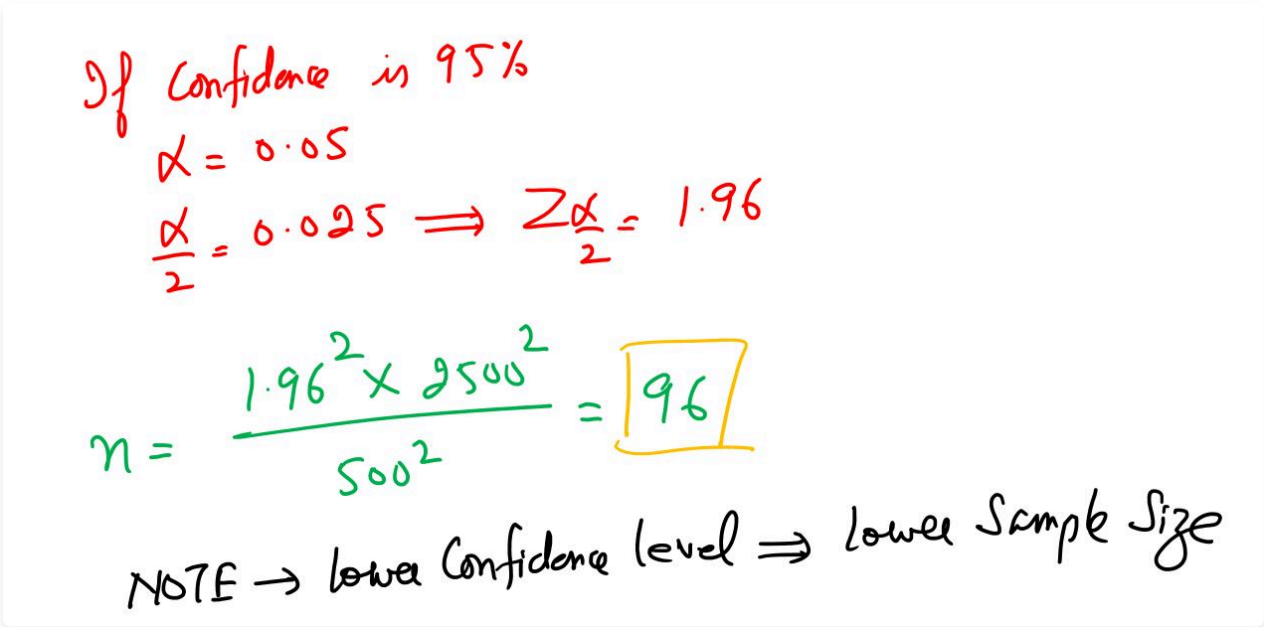
Aam Aadmi Party wants to estimate the mean electric bill for single-family homes in a large city of Gujarat. Based on studies conducted in other cities, the standard deviation is assumed to be Rs 2500. The Party wants to estimate, with 99% confidence, the mean bill to within ±500.
(i) What sample size is needed?
(ii) If 95% confidence is desired, how many homes need to be selected?

Solution:

(i) What sample size is needed?



(ii) If 95% confidence is desired, how many homes need to be selected?



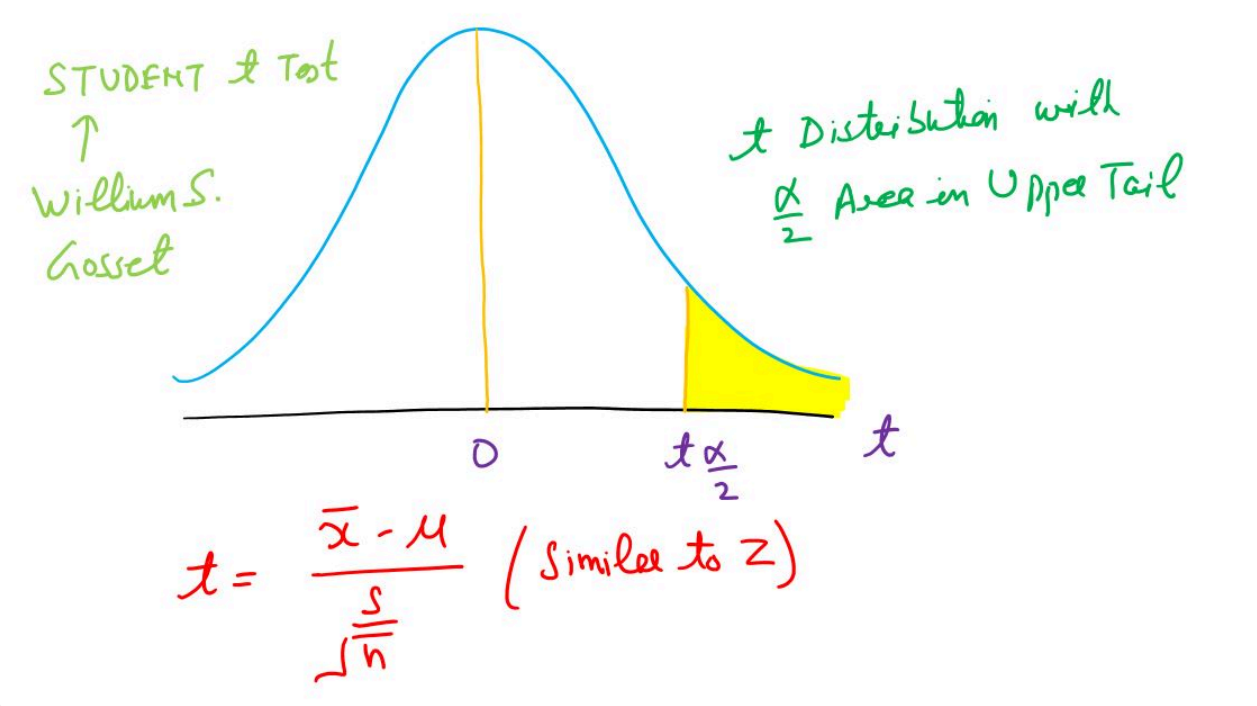
3. Interval estimate of population mean - σ not known

Sometimes we do not have a good estimate of the standard deviation of population (σ). In these cases, we must use the same standard deviation of sample (s) to estimate both μ and σ. This situation is called the **σ unknown case.**

We use standard deviation of sample (represented by s) to estimate standard deviation of population (represented by σ). In this case, the margin of error and the interval estimate for the population mean are based on a probability distribution known as **t distribution.**

The t distribution is a family of similar probability distributions, with a specific t distribution depending on a parameter known as the degrees of freedom. Please note that a distribution with more degrees of freedom exhibits less variability and more closely resembles the standard normal distribution (look at figure below). Note also that the mean of the t distribution is zero.

We place a subscript on t to indicate the area in the upper tail of the t distribution. We will use the notation $t_{\frac{\alpha}{2}}$ to represent a t value with an area of $\frac{\alpha}{2}$ in the upper tail of the t distribution.

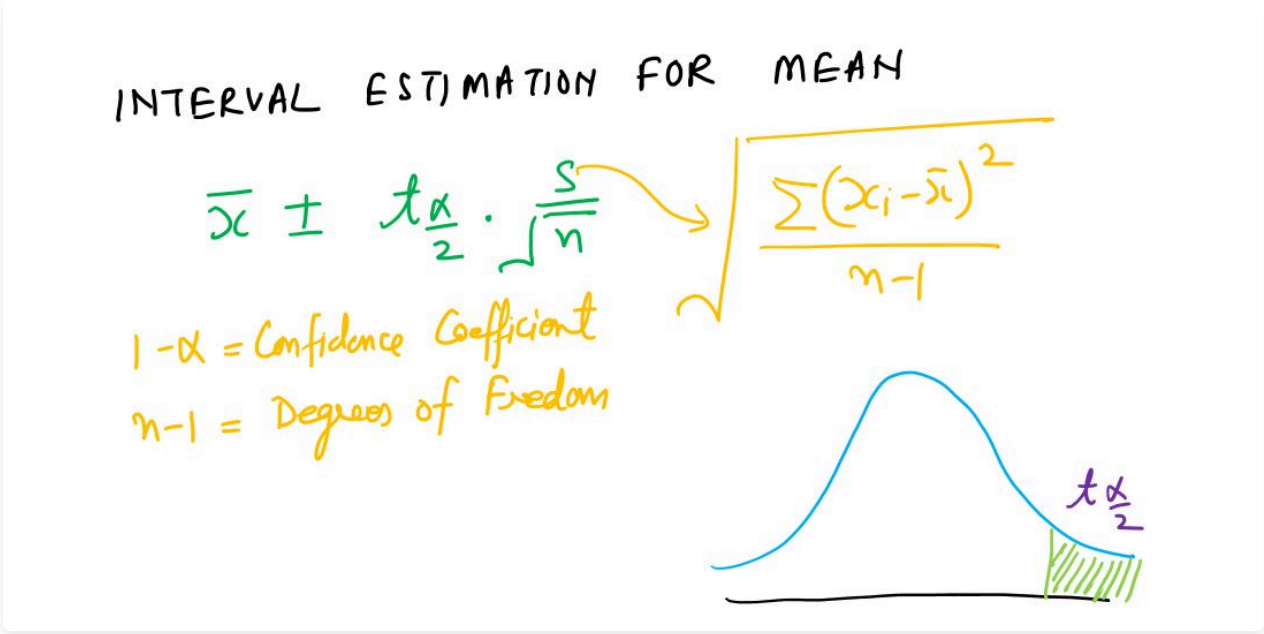


In this case, the **Margin of Error** is then given by $t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$

With this margin of error, the general expression for an **interval estimate** of a population mean (when σ is unknown), is given by

$\bar{x}$ $\pm$ $t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$

where:
s is the sample standard deviation
(1-α) is the confidence coefficient
$t_{\frac{\alpha}{2}}$ is the t value providing an area of $\frac{\alpha}{2}$ in the upper tail of the t distribution with n - 1 degrees of freedom.



Please note that sample standard deviation (s) is calculated using the formula

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

Summary

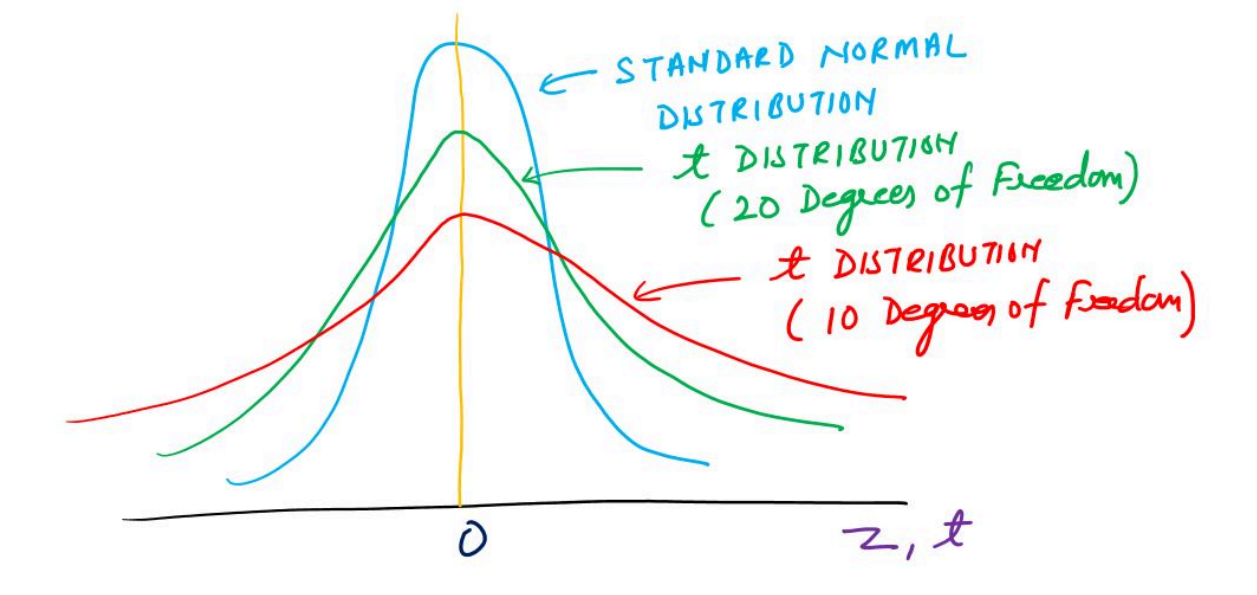
We just learnt two approaches to developing an interval estimate of a population mean.

(i) **For the σ known case**, population standard deviation σ and the standard normal distribution are used to compute the margin of error and to develop the interval estimate.
(ii) **For the σ unknown case,** the sample standard deviation s and the t distribution are used to compute the margin of error and to develop the interval estimate.

---

3. Interval estimate of population mean - σ not known

The t distribution isn't a singular distribution like the normal distribution but a family of distributions. It varies based on sample size, which determines the degrees of freedom.



Proposed by William S. Gosset, a student of Karl Pearson, the t distribution was published under the pseudonym 'Student,' hence often referred to as the 'Student t Test.'

Some of key features of t Distribution are:

1. **Degrees of Freedom**: The degrees of freedom for the t distribution are calculated as n - 1, where 'n' represents the sample size.

2. **Continuous Distribution**: Similar to the normal distribution, the t distribution is continuous, allowing for a range of values.

3. **Symmetrical and Bell-Shaped:** The t distribution, like the normal distribution, exhibits symmetry and a bell-shaped curve.

4. **Mean of Zero:** The mean of the t distribution is zero, analogous to the standard normal distribution.

5. **Greater Spread in Tails:** The t distribution has a broader spread compared to the standard normal distribution. Consequently, it has more area in its tails.

6. **Convergence with Sample Size:** As the sample size (n) increases, the behavior of the t distribution approaches that of the standard normal distribution. For values of n above 100, both distributions become nearly similar in behavior.

3. Interval estimate of population mean - σ not known

Illustration 1

The owner of a fast food joint wants to make a rather quick estimate of the average number of hours, it takes to deliver an order at home. The owner has records of all historic data, but the amount of time required to conduct an analysis of all previous orders would be prohibitive. The owner decides to take a random sample of Fourteen orders yielding the following data (in hours). He uses these data to construct a 99% confidence interval to estimate the average number of hours that an order takes to get delivered at home. It is assumed that the number of hours taken for each order is normally distributed in the population.

3 1 3 2 5 1 2 1 4 2 1 3 1 1

Solution:

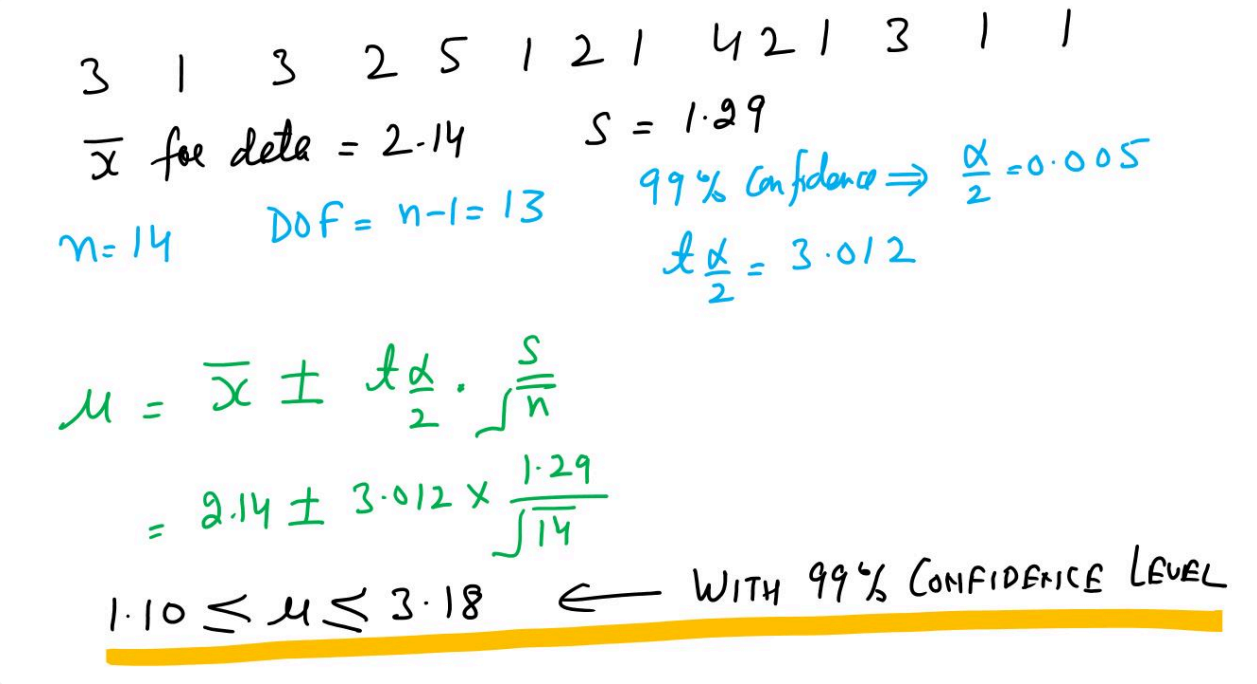


Illustration 2

The mean project completion time at IBM Technologies was 68 days. The company conducted a quality improvement project with the objective of reducing the project completion time. After conducting the quality improvement project, the quality improvement team collected a sample of 50 projects. In this sample, the mean project completion time was 32 days, with a standard deviation of 9 days. Construct a 95% confidence interval estimate for the population mean project completion time. Interpret the interval constructed. Do you think the quality improvement project was a success?

Solution:

$\bar{x} = 32 \quad s = 9 \quad n = 50 \quad DOF = 50 - 1 = 49$

95% Confidance level $\Rightarrow \alpha = 0.05 \quad \frac{\alpha}{2} = 0.025 \rightarrow t_{\frac{\alpha}{2}} = 2.009$

$$\mu = \bar{x} \pm t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 32 \pm 2.009 \times \frac{9}{\sqrt{50}}$$

$29.44 \leq \mu \leq 34.56 \leftarrow$ 95% CONFIDENCE LEVEL

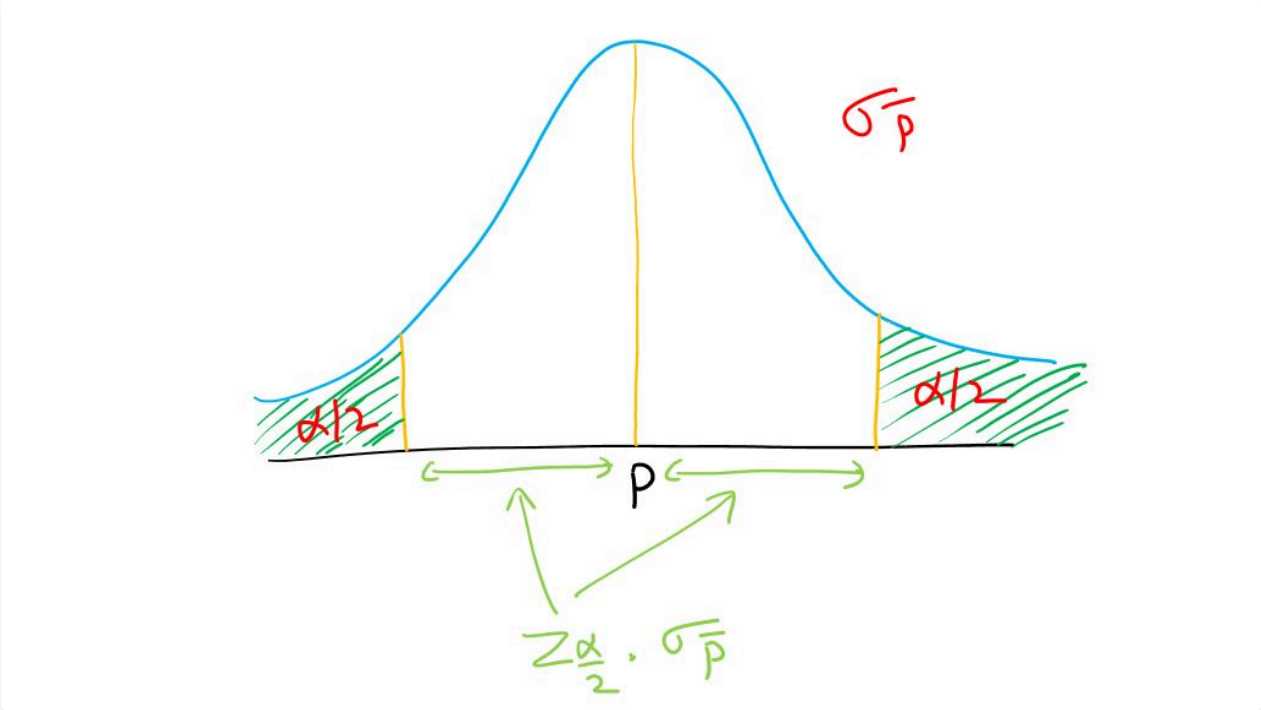
Yes, it was success, as we have reduced the time (in comparison to 68)

4. Interval estimate of the population proportion

The general form of an interval estimate of a population proportion p is given by:

= \(\bar p \ \pm \) Margin of Error

We learnt earlier that sampling distribution of proportions can be approximated by a normal distribution whenever np ≥ 5 and n(1-p) ≥ 5.



Because the sampling distribution of is normally distributed, if we choose \(z_ \frac{ \alpha}{2} \times \sigma_\bar p \) as the **margin of error** in an interval estimate of a population proportion.

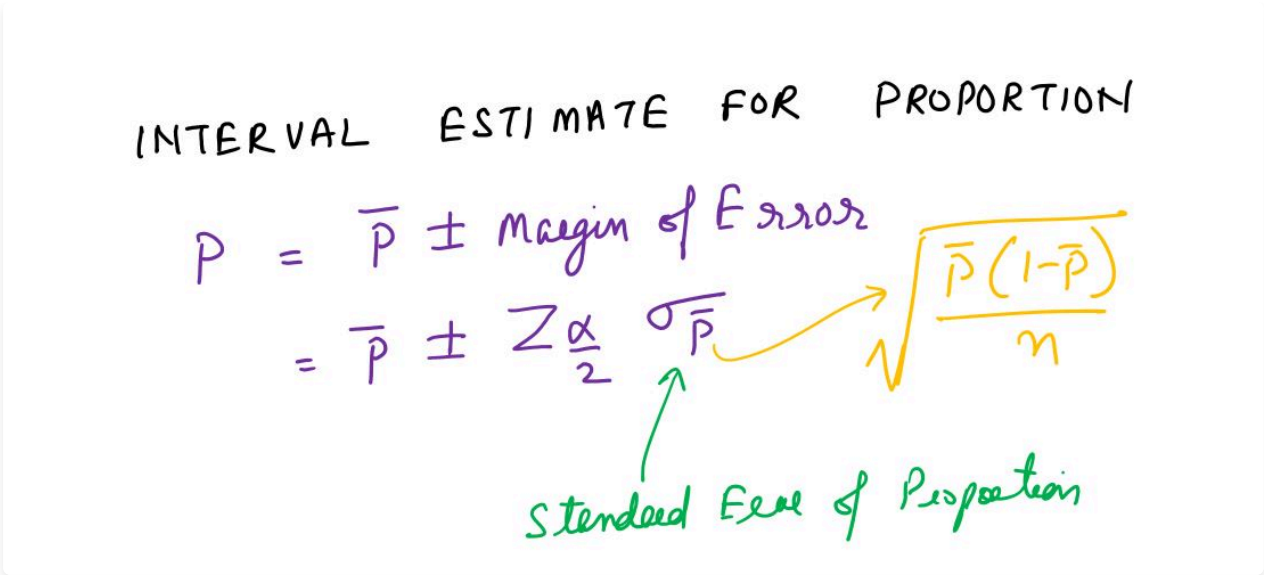
We know that the formula for calculation of \(\sigma_\bar p \) is

\(\sigma_\bar p = \sqrt{ \frac{p(1-p)}{n}} \)

But we do not have value of p, available with us (In fact it is p, that we are trying to estimate). So we substitute it with \(\bar p \).

Finally, the Margin of Error is given by following formula

Margin of Error = \(z_ \frac{ \alpha }{2} \times \sqrt{ \frac{\bar p(1-\bar p)}{n} } \)



And thus, formula for calculation of an interval estimate of a population proportion becomes:

= \(\bar p \) \(\pm \) \(z_ \frac{ \alpha }{2} \times \sqrt{ \frac{\bar p(1-\bar p)}{n} } \)

---

4. Interval estimate of the population proportion

Illustration 1

The Congress Party is worried upcoming election results in a state. The representative of the party surveyed 200 people, out of which 35 people are likely to vote for the party. To analyze the data, you need to construct and interpret a 90% confidence interval for the proportion of people, which are likely to vote for the Congress Party.

Solution:

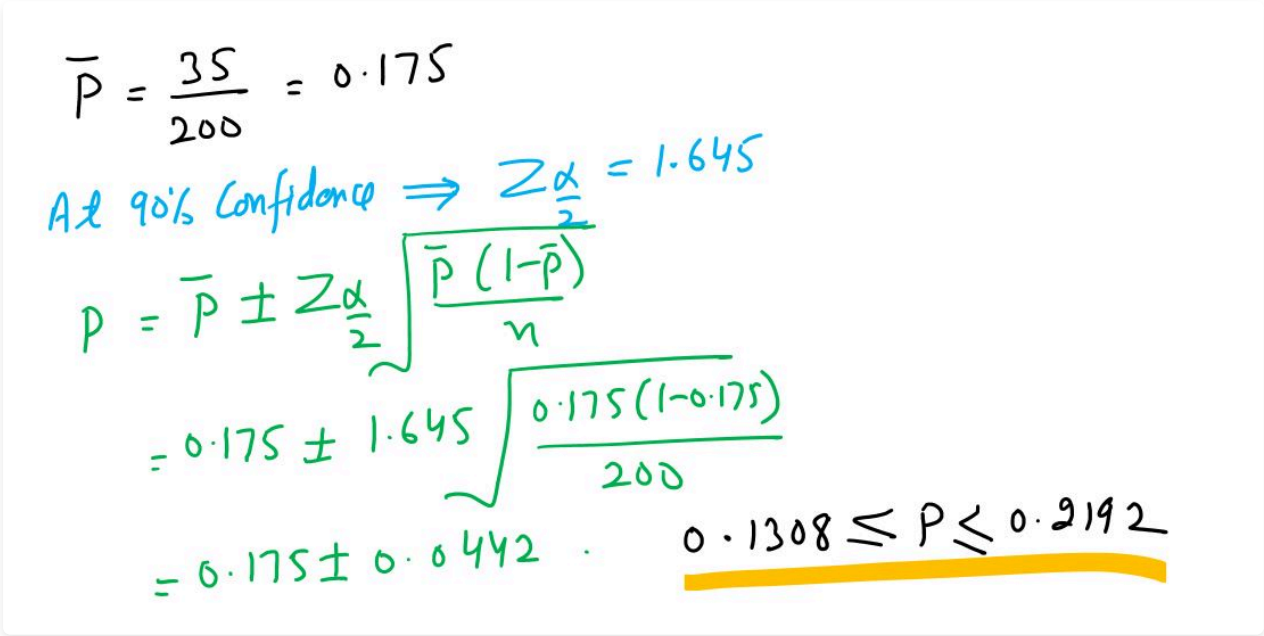


Illustration 2

Labour Department of a state Government surveyed 210 chief executives of small companies. Only 51% of these executives had a canteen facility at workplace. Use the data given to compute a 92% confidence interval to estimate the proportion of all small companies that have a canteen facility at workplace.

Solution:

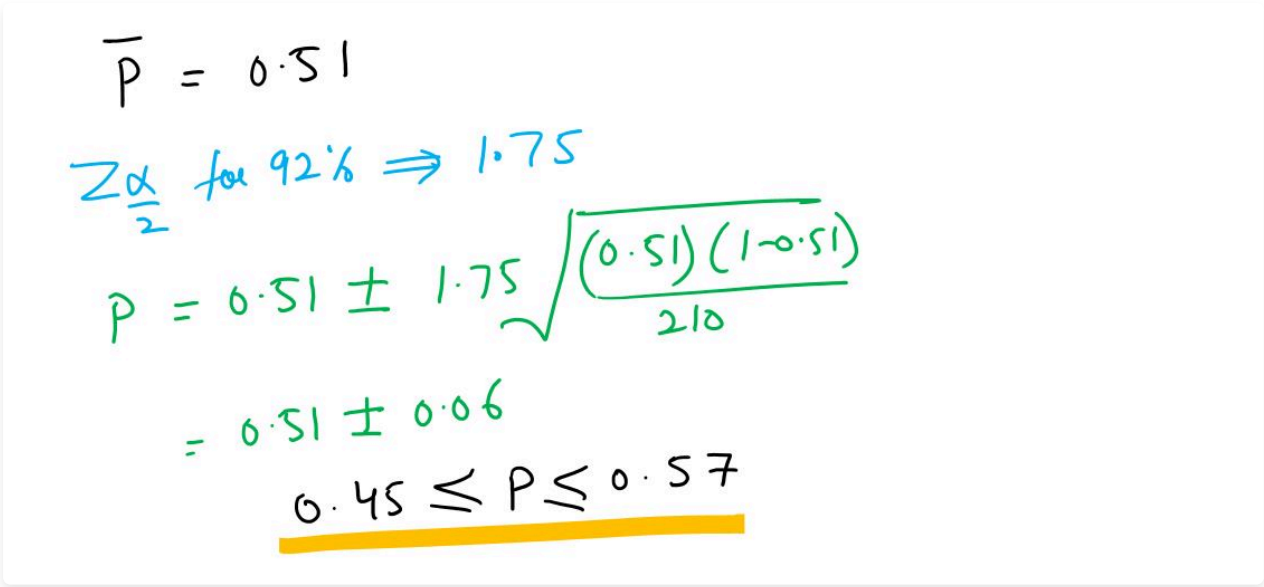


Illustration 3

A national survey of 900 students was conducted to learn how many of them are satisfied with the Government. The survey found that 396 were satisfied. Build the interval estimate at 95% Confidence level.

Solution:

The point estimate of the proportion of satisfied students is $\bar p$, $\frac{396}{900} = 0.44$

We find Interval Estimate using the formula

$\bar p \pm z_{\frac{\alpha}{2}} \times \sqrt{\frac{\bar p(1-\bar p)}{n}}$

$= 0.44 \pm 1.96 \times \sqrt{\frac{0.44(1-0.44)}{900}} = 0.44 \pm 0.0324$

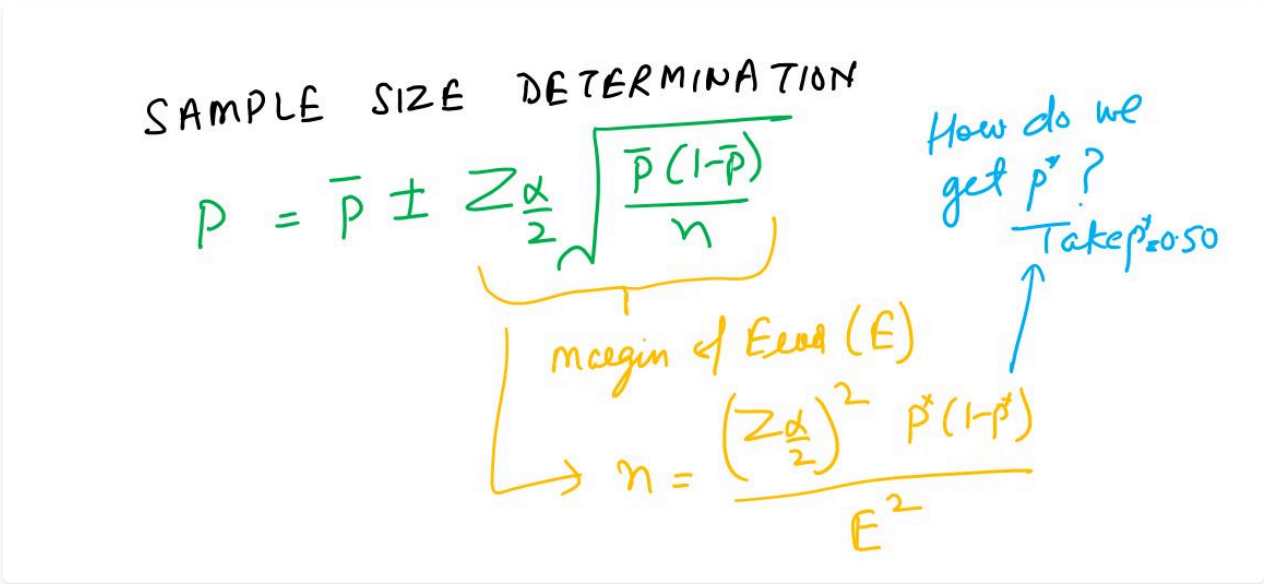
Thus, the margin of error is 0.0324 and the 95% confidence interval estimate of the population proportion is 0.4076 to 0.4724.

Using percentages, the survey results enable us to state with 95% confidence that between 40.76% and 47.24% of all students are satisfied with the Government.

---

4. Interval estimate of the population proportion

Let us learn, how to choose a sample size large enough to provide a desired margin of error.



Let E be desired Margin of Error. Then E is given by

$E = z_{\frac{\alpha}{2}} \times \sqrt{\frac{\bar p(1-\bar p)}{n}}$

Solving for n, we get

$n = \frac{(z_{\frac{\alpha}{2}})^2 \times \bar p (1-\bar p)}{E^2}$

This sample size provides the desired margin of error at the chosen confidence level.

---

4. Interval estimate of the population proportion

Illustration 1

You want to have 90% confidence of estimating the proportion of office workers who respond to e-mail within an hour to within ±0.05. Because you have not previously undertaken such a study, there is no information available from past data. Determine the sample size needed.

Solution:

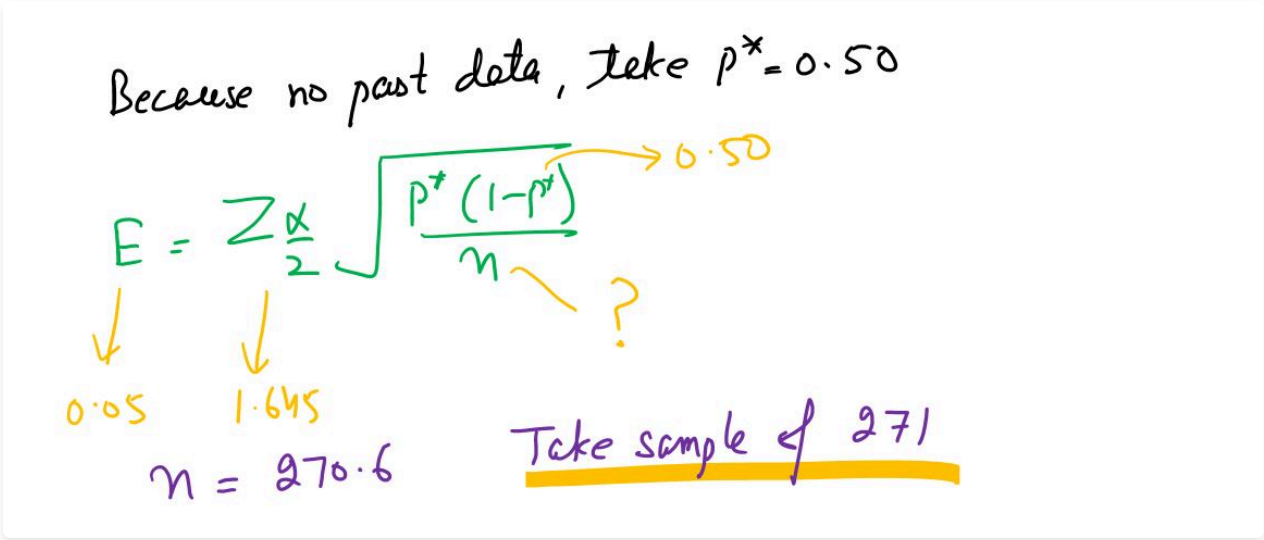
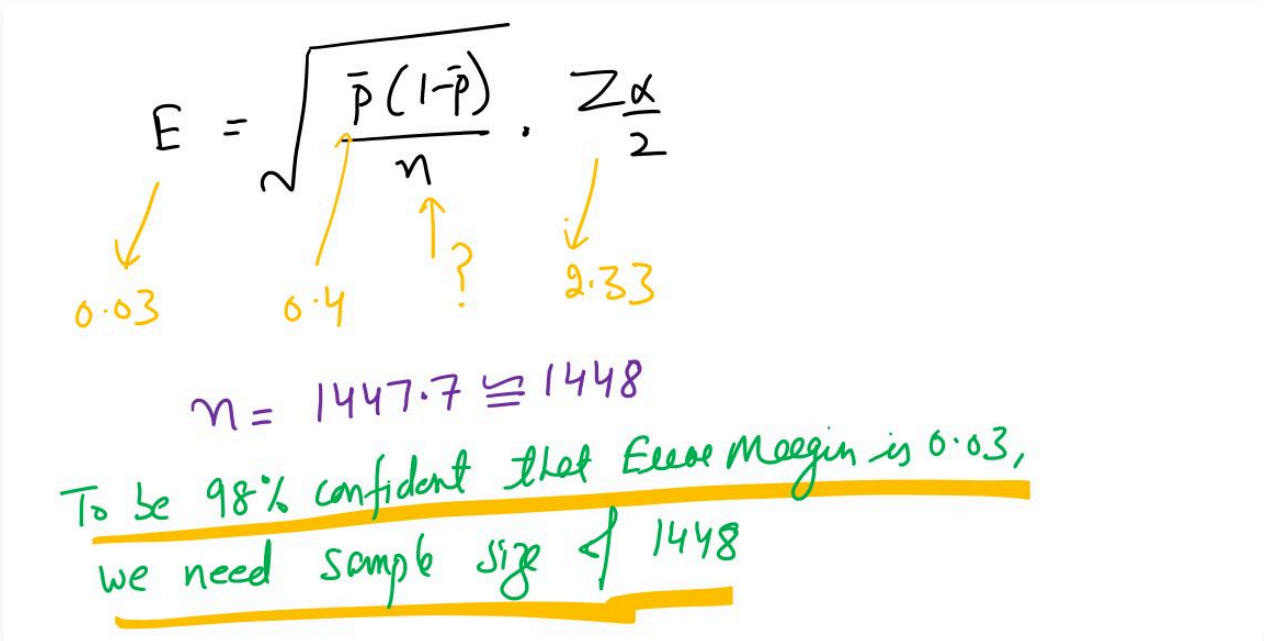


Illustration 2

Ministry of AYUSH conducted a national survey to determine the extent to which schools are promoting health and fitness among their students. One of the questions asked was, Does your school offer Yoga classes? Suppose it was estimated before the study that no more than 40% of the schools would answer Yes. How large a sample would Ministry representatives have to take in estimating the population proportion to ensure a 98% confidence in the results and to be within 0.03 of the true population proportion?

Solution:



5. Law of large numbers

Law of large numbers, in statistics, is a theorem that says, as the number of identically distributed, randomly generated variables increases, their sample mean (average) approaches their theoretical mean.

The law of large numbers is closely related to what is commonly called the **Law of Averages.** In coin tossing, the law of large numbers stipulates that the fraction of heads will eventually be close to $\frac{1}{2}$. Hence, if the first 10 tosses produce only 3 heads, it seems that some mystical force must somehow increase the probability of a head, producing a return of the fraction of heads to its ultimate limit of $\frac{1}{2}$. Yet the law of large numbers requires no such mystical force. Indeed, the fraction of heads can take a very long time to approach $\frac{1}{2}$. For example, to obtain a 95 percent probability that the fraction of heads falls between 0.47 and 0.53, the number of tosses must exceed 1,000. In other words, after 1,000 tosses, an initial shortfall of only 3 heads out of 10 tosses is swamped by the results of the remaining 990 tosses.

---

1. Sampling Distribution of Variance

Let us explore how to infer the variance of a population from the variance of a sample.

The formula for sample variance is:

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

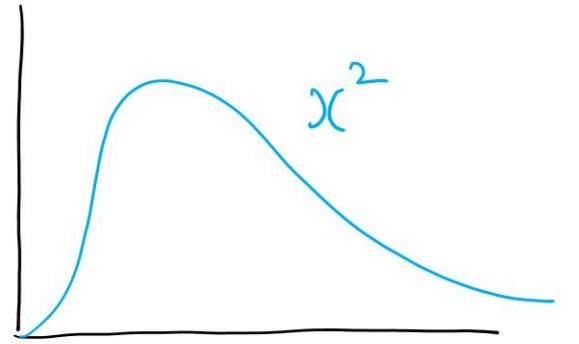
↳ VARIANCE OF SAMPLE

When the population follows a normal distribution (a necessary condition to infer population variance from a sample), the sampling distribution of variance adheres to the chi-square distribution, represented by the following formula:

$$\chi^2_{n-1} = \frac{(n-1)\,s^2}{\sigma^2}$$

s^2 = Variance of Sample

σ^2 = Variance of Population



CHI-SQUARE DISTRIBUTION

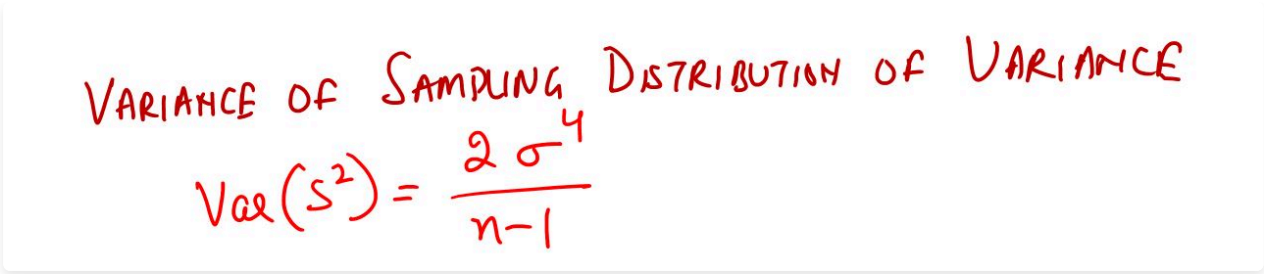
This denotes the relationship between the variance of a sample and the variance of a population for degrees of freedom (n - 1).

The expected value or mean of the sampling distribution of variance is:

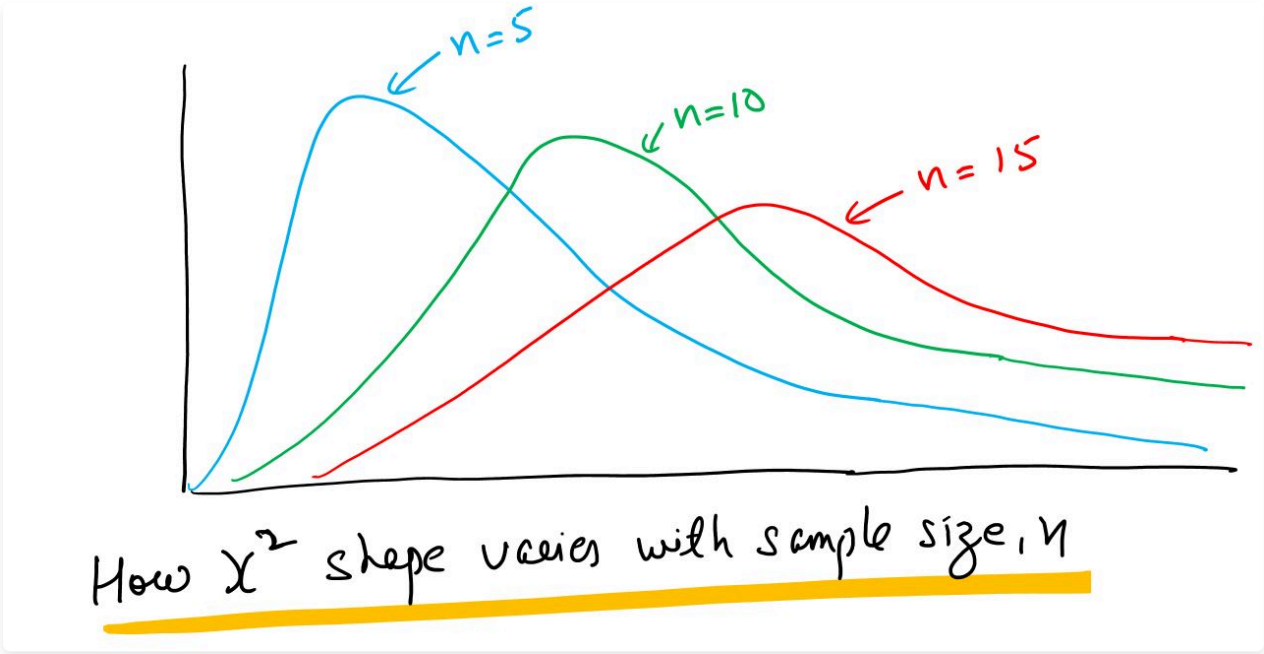
MEAN OF SAMPLING DISTRIBUTION OF VARIANCE

$$E(s^2) = \sigma^2$$

Additionally, the variance of the sampling distribution of variance is given by:

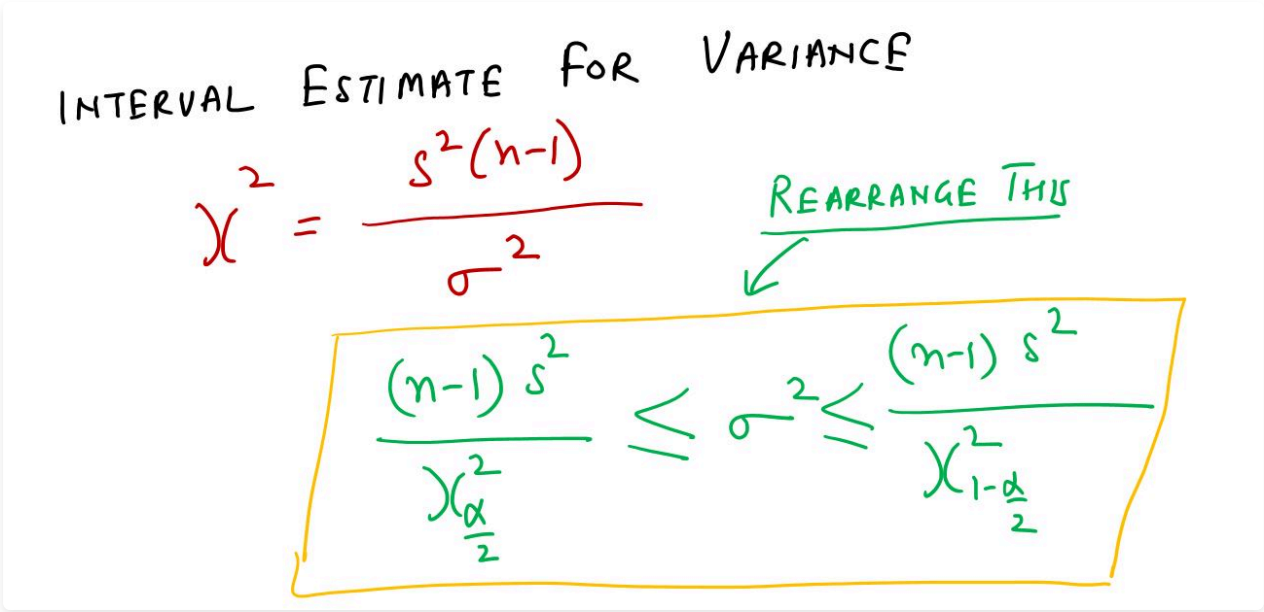


Similar to the t distribution, the chi-square distribution also varies with sample size. The alteration in the shape of the chi-square distribution concerning sample size is depicted in the figure.

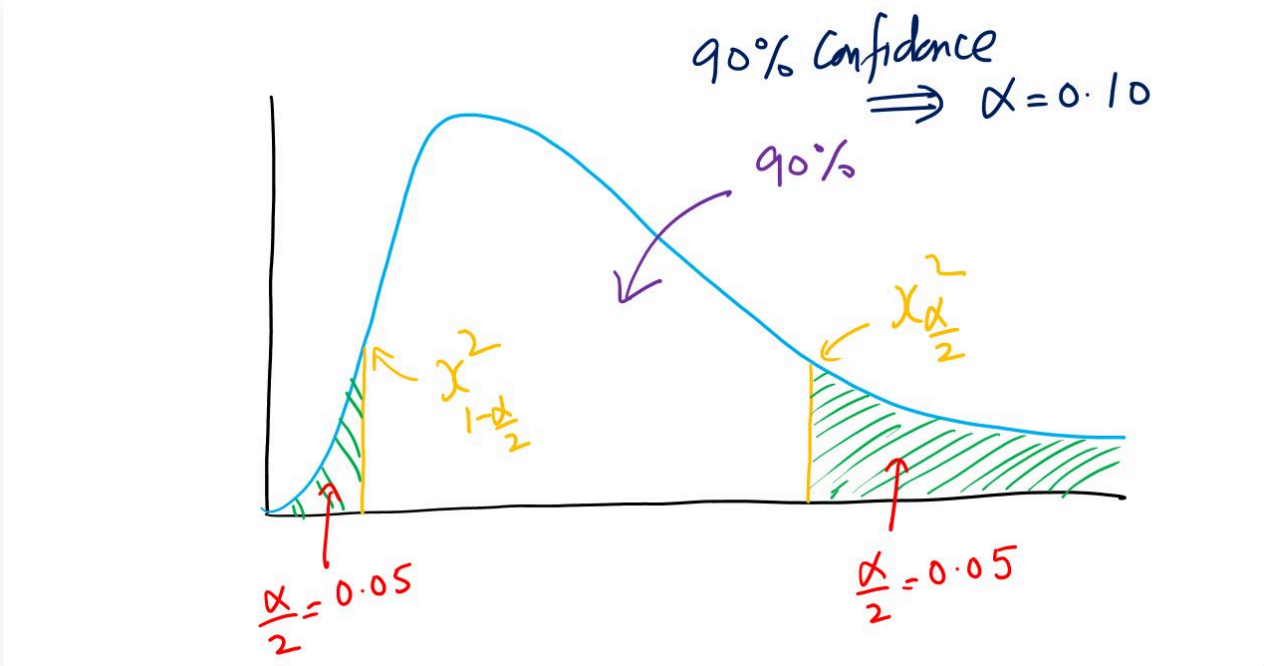


2. Interval Estimation of Variance

Interval Estimation for Variance is given by following formula.



If we consider 90% Confidence level, the corresponding area under the curve is shown by below figure.



3. Illustration on Variance

Illustration 1

The NASSCOM says the average workweek in the India for software engineers is down to only 35 hours, largely because of a rise in part-time workers. Suppose this figure was obtained from a random sample of 20 software engineers and that the standard deviation of the sample was 4.3 hours. Assume hours worked per week are normally distributed in the population. Use this sample information to develop a 98% confidence interval for the population variance of the number of hours worked per week for software engineers. What is the point estimate?

Solution:

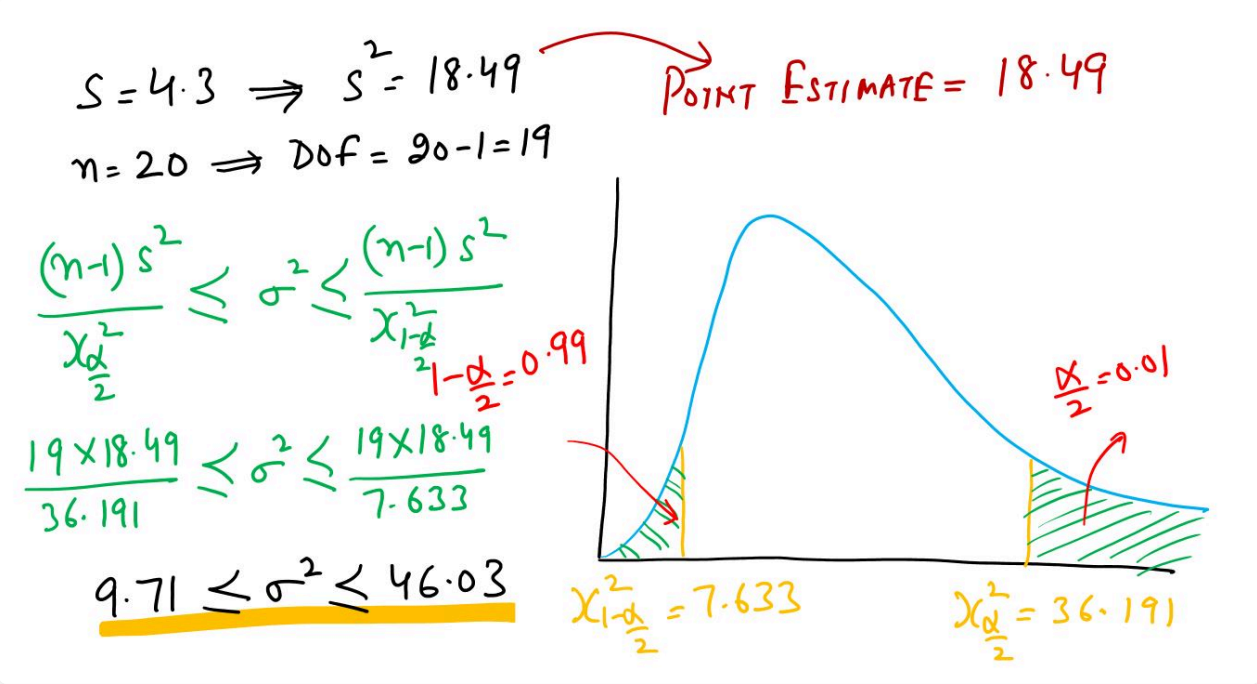
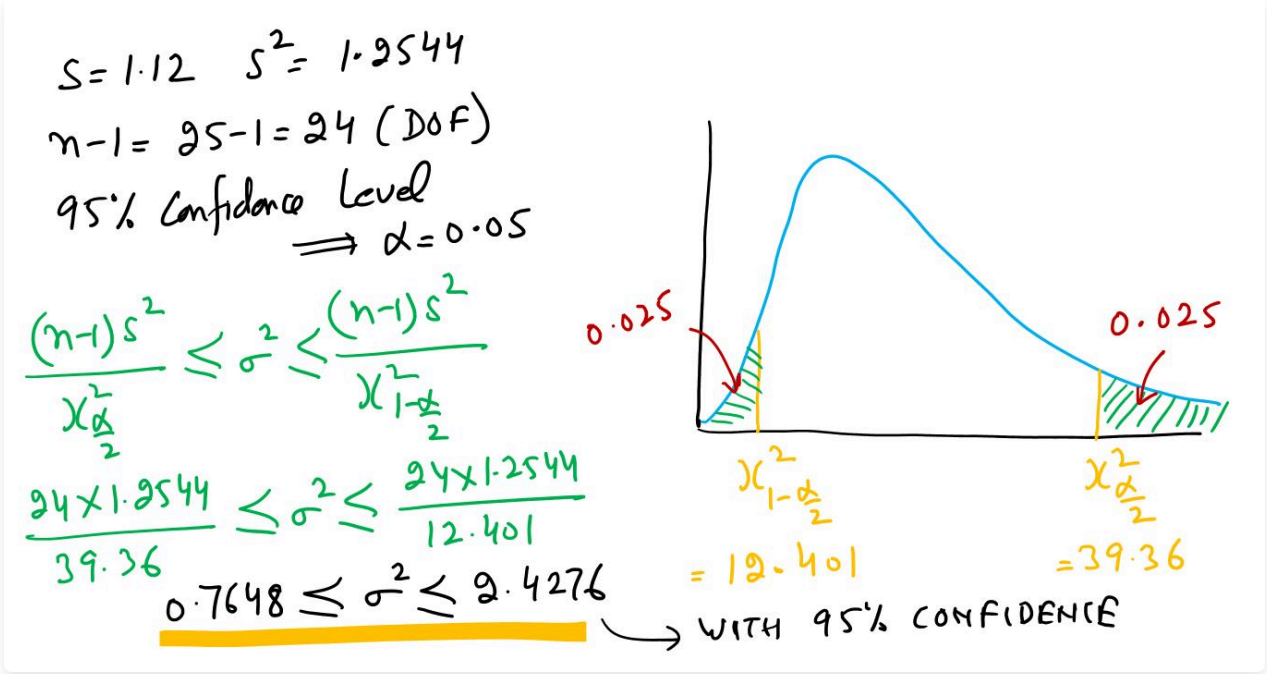


Illustration 2

An NGO for social security publishes data on the hourly wages for security guards in a city. The latest figures published show that the hourly wages for security guards is Rs 16.10. The NGO wants to know how consistent this figure is. They randomly select 25 security guards and determine that the standard deviation of hourly wages for such guards is Rs 1.12. Use this information to develop a 95% confidence interval to estimate the population variance for the hourly wages for security guards. Assume that the hourly wages for security guards are normally distributed in that city.

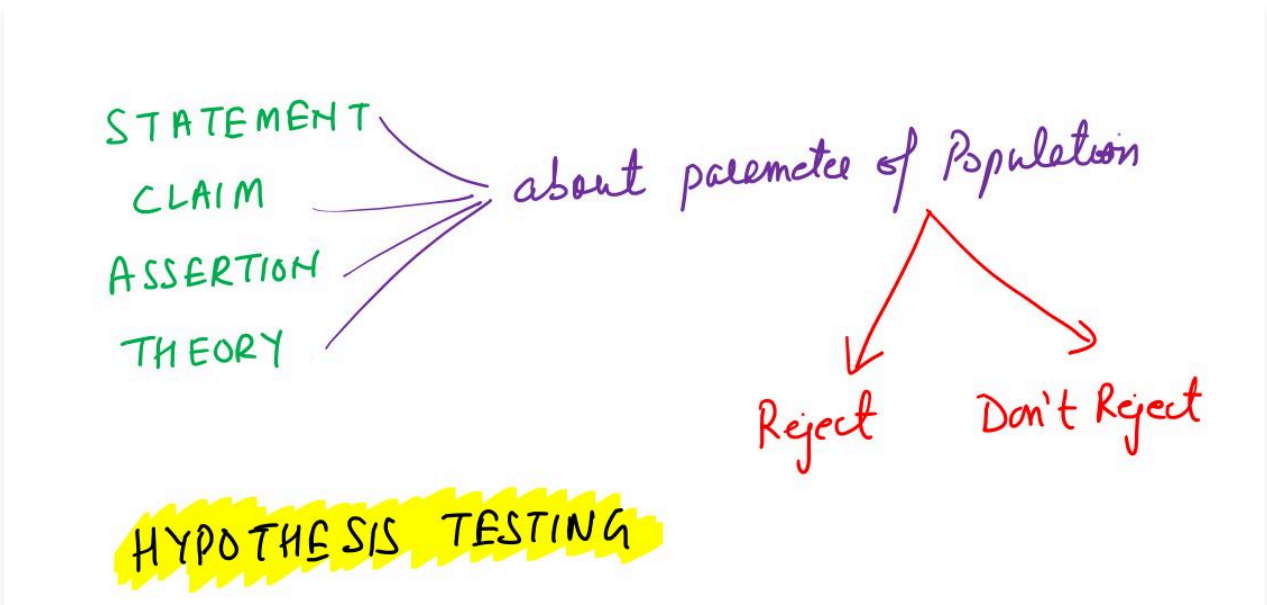
Solution:



---

1. Introduction

Hypothesis Testing is the most important concept of inferential statistics.



Previously, we have understood the methodology of constructing interval estimations about population parameters using sample data, with a specified confidence level.

Now, our focus shifts to forming a claim, assertion, or theory regarding the population parameters. Subsequently, through analysis of sample data, we determine whether to reject or accept this claim, assertion, or theory. This formulated claim or theory is referred to as a Hypothesis, and the entire process is called **Hypothesis Testing**. It is also called Significance Testing.

---

2. Null and Alternate Hypothesis

In hypothesis testing, we begin by making a tentative assumption about a population parameter.

Thus, Hypothesis testing uses sample data to determine whether a statement about the value of a population parameter
(a) should be rejected or
(b) should not be rejected.

This tentative assumption is called the **null hypothesis** and is denoted by H_0. We then define another hypothesis, called the **alternative hypothesis**, which is the opposite of what is stated in the null hypothesis. The alternative hypothesis is denoted by H_a or H_1.

The Null Hypothesis represents the prevailing belief or default stance, standing as the current assumption until evidence surfaces to indicate otherwise. It embodies the existing theory or status quo.

On the other hand, the Alternate Hypothesis is acknowledged only when compelling evidence substantiates it. It bears the responsibility of proof and is also known as the research hypothesis. This hypothesis embodies what we aim to demonstrate or validate through the process of hypothesis testing.

In some situations, it is easier to identify the alternative hypothesis first and then develop the null hypothesis. In other situations, it is easier to identify the null hypothesis first and then develop the alternative hypothesis.

Let us learn framing of null and alternate hypothesis, with few examples.

Illustration 1

The present average mileage of the Skoda Slavia car is 8 kilometers per liter. Our investigation aims to ascertain whether the adoption of a new type of petrol leads to an increase in this average. We seek evidence to support the conclusion that the new type of petrol enhances the average mileage.

Solution:

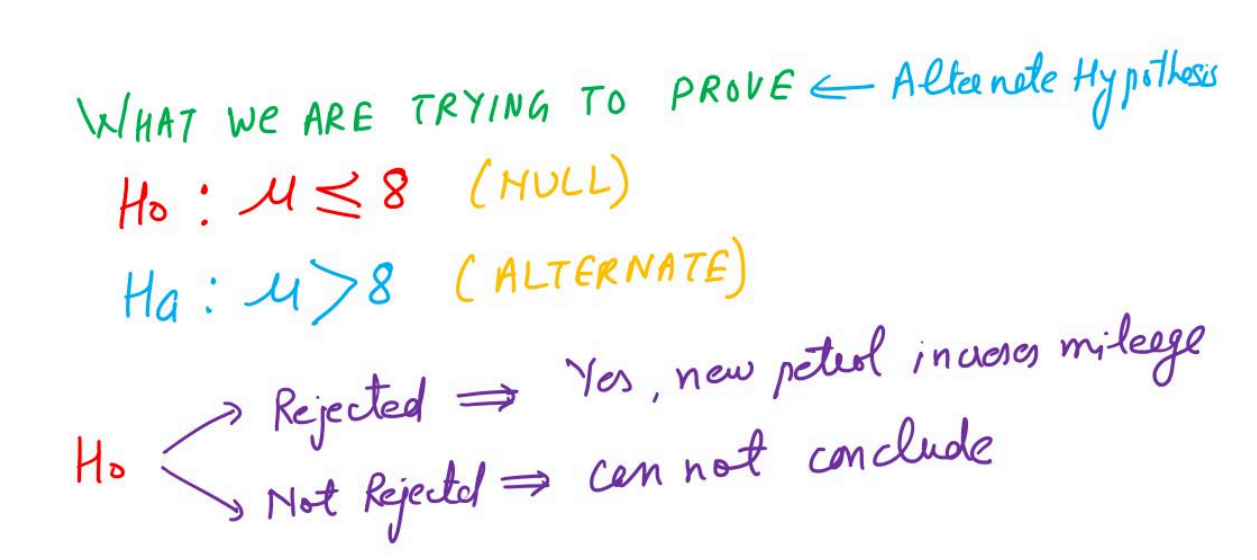


Illustration 2

PepsiCo claims that a packet of Lays contains 25 grams of chips. Our task is to verify the accuracy of this claim through testing.

Solution:

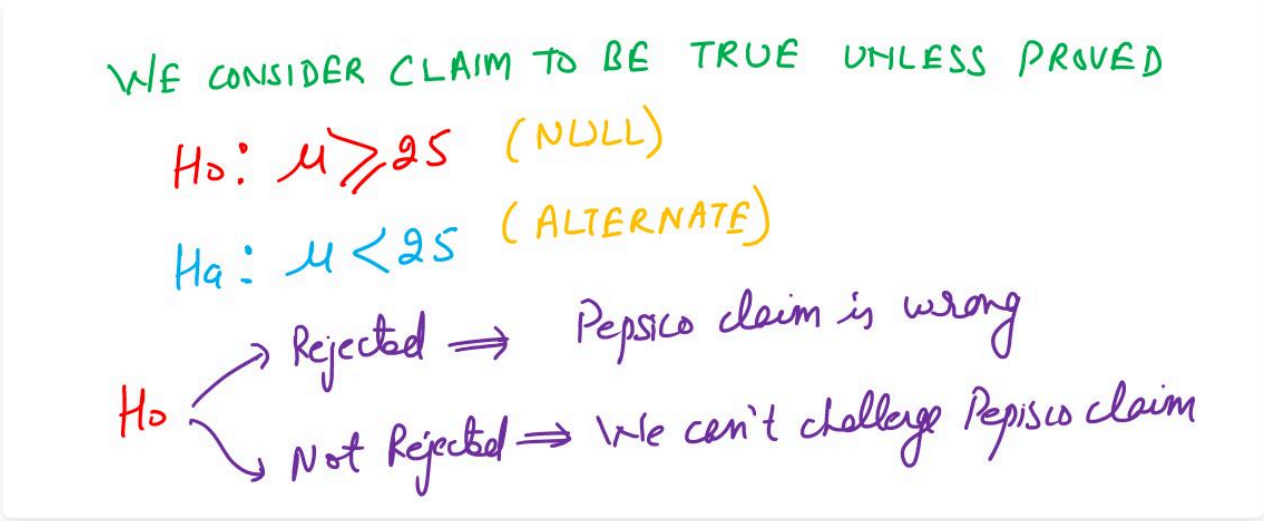


Illustration 3

Maruti cars have requested a supply of tyres from MRF tyres with a specified requirement for the mean radius of the tyre to be 18 centimeters. Any deviation from this 18 centimeters would lead to assembly operation quality issues. Based on a sample of tyres from the recently received shipment, a quality control inspector at Maruti needs to determine whether to approve the shipment or return it to MRF tyres due to a deviation from the specified radius.

Solution:

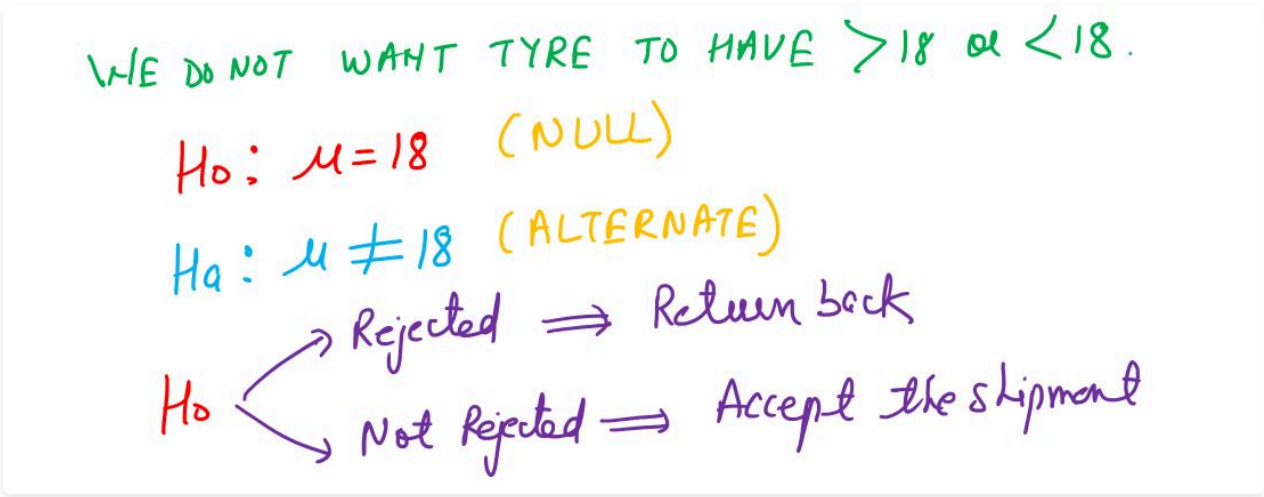


Illustration 4

The manager of a company is considering a new bonus plan designed to increase sales volume. Currently, the mean sales volume is Rs 14 lakh per month. The manager wants to conduct a research study to see whether the new bonus plan increases sales. To collect data on the plan, a sample of sales personnel will be allowed to sell under the new bonus plan for a one-month period.

Solution:

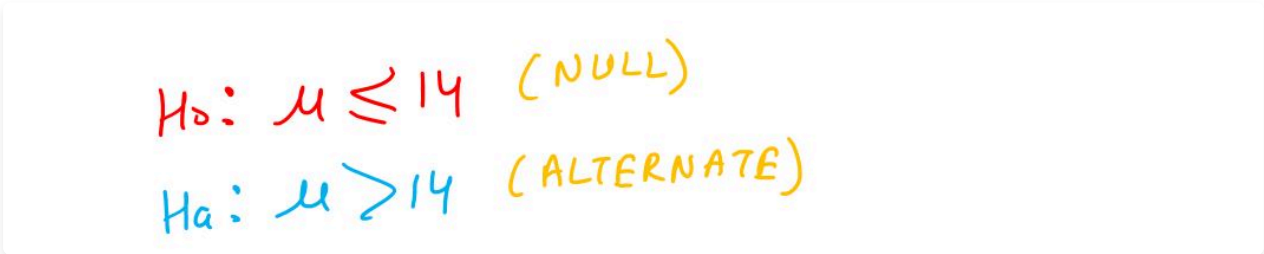


Illustration 5

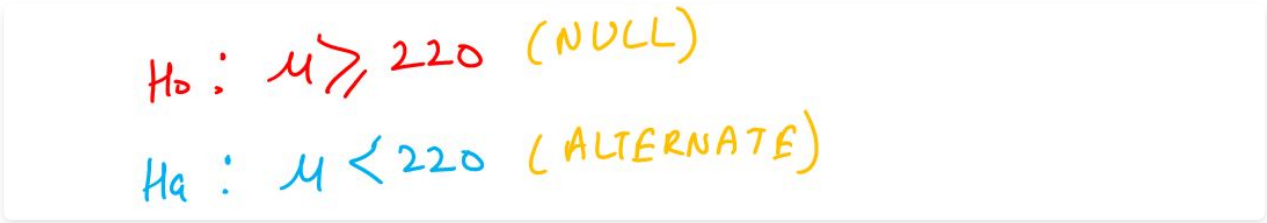
You are the manager of a fast-food restaurant. You want to determine whether the waiting time to place an order has changed in the past month from its previous population mean value of 4.5 minutes.

Solution:

Illustration 6

Because of high production-changeover time and costs, a director of manufacturing must convince management that a proposed manufacturing method reduces costs before the new method can be implemented. The current production method operates with a mean cost of Rs 220 per hour. A research study will measure the cost of the new method over a sample production period.

Solution:



Depending on the situation, hypothesis tests about a population parameter may take one of following 3 forms; out of which 2 use inequalities in the null hypothesis; the third uses an equality in the null hypothesis. Here μ_0 denoted hypothesized value. Not that, we will never have equality sign in Alternate Hypothesis, it will always be in Null Hypothesis.

Type	Null Hypothesis (H_0)	Alternative Hypothesis (H_a)
Type 1	H_0: $\mu \geq \mu_0$	H_a: $\mu < \mu_0$
Type 2	H_0: $\mu \leq \mu_0$	H_a: $\mu > \mu_0$
Type 3	H_0: $\mu = \mu_0$	H_a: $\mu \neq \mu_0$

In selecting the proper form of H_0 and H_a, keep in mind that the alternative hypothesis is often what the test is attempting to establish. Hence, asking whether the user is looking for evidence to support $\mu < \mu_0$,$\mu > \mu_0$, or $\mu \neq \mu_0$ will help determine H_a.

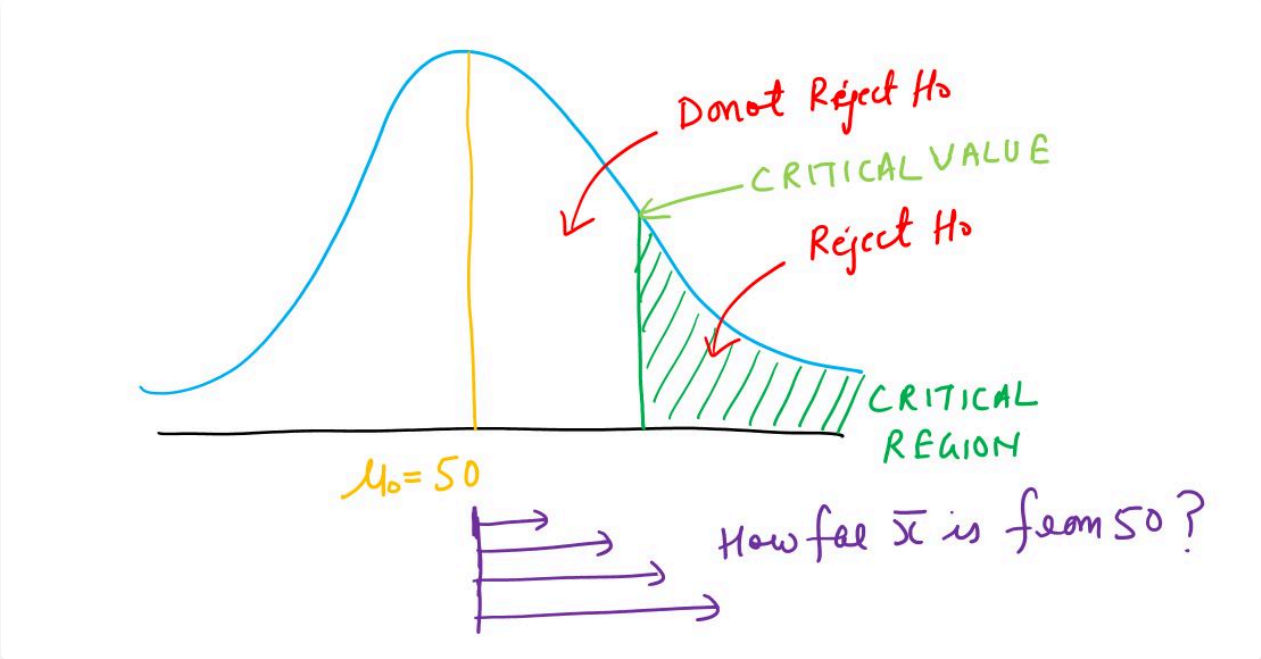
3. Fundamental logic behind Hypothesis Testing

Suppose, the ABC Electronics has placed an order for microchips with XYZ Tech. The specifications necessitate a mean performance value of 50 units for these microchips. Any deviation from this standard value would result in operational inefficiencies.

$$H_0: \mu = 50$$
$$H_a: \mu \neq 50$$

Now, based on a sample of microchips received from the recent shipment, a quality control analyst at ABC Electronics must decide whether to accept the shipment or return it. The sample mean comes out to be $\bar{x}$.

In Hypothesis Testing, we compare calculated value of sample mean $\bar{x}$ and hypothesized value of 50 (μ_0). If $\bar{x}$ and μ_0 are close to each other, we accept (rather not reject) the Null Hypothesis, otherwise, we reject the Null hypothesis.



Thus, when conducting hypothesis testing, accepting the Null Hypothesis is contingent upon the sample mean being in close proximity to the hypothesized mean.

However, the pivotal question arises: what exactly defines "close enough"?

This closeness is determined by the division between rejection and non-rejection regions within the curve of the sampling distribution. The critical value separate the region where we accept the Null Hypothesis from the region where we reject it.

The critical value is a threshold that delineates the limits of acceptance or rejection based on the chosen level of significance or risk. This level of risk or significance, often denoted by the Greek letter alpha (α), dictates how much risk we're willing to tolerate in making a conclusion about the population parameter from the sample data.

Bear in mind that relying solely on sample data to make inferences about the entire population introduces a level of uncertainty due to the presence of a margin of error. This margin of error signifies the potential discrepancy between the sample statistic and the true population parameter.

Therefore, in hypothesis testing, the decision to accept or reject the Null Hypothesis hinges on the positioning of the sample mean concerning the hypothesized mean, guided by the critical value chosen based on the level of risk deemed acceptable.

$\bar{x}$ AND μ_0 ARE CLOSE

Accept H_0

Do not accept H_a

$\bar{x}$ is not statistically different from reference value μ_0

Difference between $\bar{x}$ and μ is due to random chance

Result is not statistically significant

$\bar{x}$ AND μ_0 ARE NOT CLOSE

Accept H_a

Reject H_0

$\bar{x}$ is statistically different from reference value μ_0

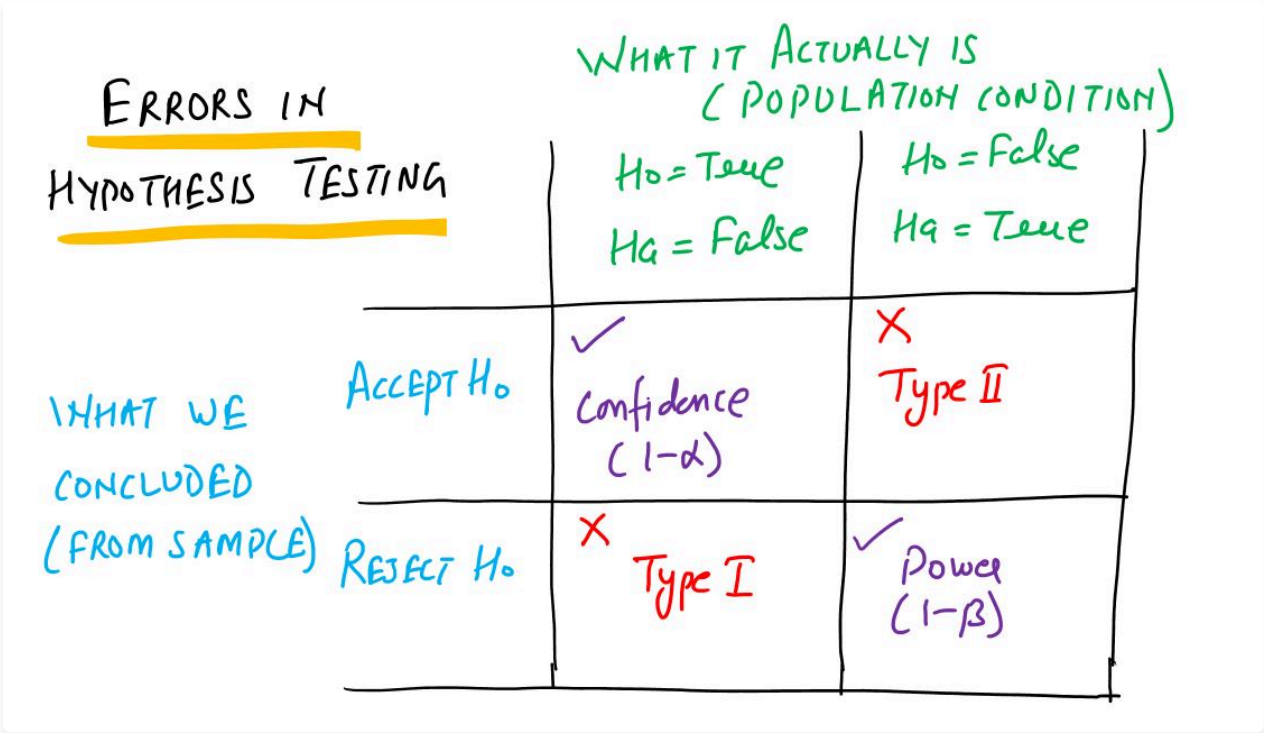
Difference between $\bar{x}$ and μ is not due to random chance

Result is statistically significant

4. Type I and Type II Errors

Since we assess a claim (hypothesis) concerning a population parameter based solely on a sample's data, the possibility of errors alawys exists.

Ideally the hypothesis testing procedure should lead to the acceptance of H_0 when H_0 is true and the rejection of H_0 when H_a is true. Unfortunately, the correct conclusions are not always possible. Two kinds of errors, that can be made in hypothesis testing, as shown in the figure.



The first row of the figure shows what can happen if the conclusion is to accept H_0. If H_0 is true, this conclusion is correct. However, if H_a is true, we make a **Type II error**; that is, we accept H_0 when it is false.

The second row of the figure shows what can happen if the conclusion is to reject H_0. If H_0 is true, we make a **Type I error**; that is, we reject H_0 when it is true. However, if H_a is true, rejecting H_0 is correct.

The probability of making a Type I error, when the null hypothesis is true as an equality is called the **level of significance**. The Greek symbol α (alpha) is used to denote the level of significance, and common choices for α are 0.05 and 0.01. We know that (1-α) is **Confidence Coefficient**. Thus the Confidence Coefficient is the probability that you will not reject H_0, when it is True.

In practice, the person responsible for the hypothesis test specifies the level of significance. By selecting α, that person is controlling the probability of making a Type I error. If the cost of making a Type I error is high, small values of α are preferred. If the cost of making a Type I error is not too high, larger values of α are typically used. Applications of hypothesis testing that only control for the Type I error are called **significance tests**.

The probability of making a Type II error is denoted by beta, β. **Power of test** is given by (1-β). The Power of test is the probability that you will reject H_0, when it is False.

Thus, we conclude that:

- A Type I error (α) is the probability of rejecting a true null hypothesis.
- A Type II error (β) is the probability of failing to reject a false null hypothesis.

Or simply:

- A Type I error (α) is the probability of telling you things are wrong, given that things are correct.
- A Type II error (β) is the probability of telling you things are correct, given that things are wrong.

Few examples of Type I and Type II errors are given below:

Scenario: A diagnostic test for a disease shows a positive result when the patient is actually healthy (false alarm).
Tpe I Error (False Positive): Concluding that the patient has the disease when they do not.
Type II Error (False Negative): Failing to diagnose a disease in a patient who is actually ill, resulting in a missed treatment opportunity.

Scenario: Convicting an innocent person (false accusation).
Type I Error (False Positive): Incorrectly concluding guilt and punishing an innocent individual.
Type II Error (False Negative): Failing to identify the actual perpetrator, resulting in the guilty party going unpunished.

Scenario: Rejecting a batch of products as defective when they are actually fine.
Type I Error (False Positive): Mistakenly identifying a good batch as faulty and taking unnecessary action.
Type II Error (False Negative): Failing to detect actual defects in a product batch, allowing defective products to reach consumers.

Although most applications of hypothesis testing control for the probability of making a Type I error, they do not always control for the probability of making a Type II error. Hence, if we decide to **accept H_0**, we cannot determine how confident we can be with that decision. Because of the uncertainty associated with making a Type II error when conducting significance tests, statisticians usually recommend that we use the statement "**do not reject H_0**" instead of "**accept H_0**". If value of β is defined, the yes, we can even use "**accept H_0**".

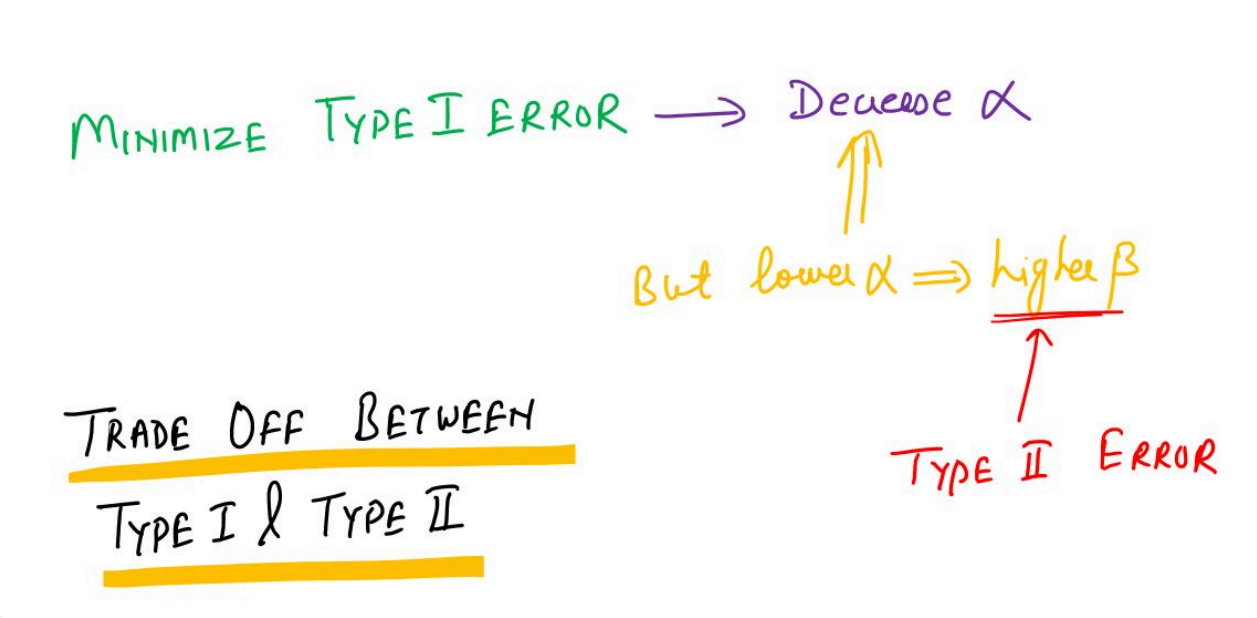
Type I errors are also called:
1) Producer's risk
2) False alarm
3) False negative
4) α error

Type II errors are also called:
1) Consumer's risk
2) Misdetection
3) False positive
4) β error

---

4. Type I and Type II Errors

In a court trial scenario, a Type I error involves convicting an innocent individual and sending them to jail. Conversely, a Type II error occurs when a guilty person, such as a murderer, is set free.



Now, if the legal system emphasizes significantly on preventing Type I errors, the consequence might lead to a system where more individuals, including some guilty of serious crimes, could be acquitted. This situation arises from the cautiousness in making convictions due to the fear of making mistakes and convicting innocent people.

By attempting to minimize the occurrence of Type I errors (α), the risk of committing Type II errors (β) tends to increase. In other words, the more stringent the standards to ensure innocent individuals are not wrongly convicted (lowering α), the higher the possibility that guilty individuals might be acquitted (increasing β).

This scenario underscores the trade-off between Type I and Type II errors in the legal system. Striking a balance between these errors is crucial. Too much focus on reducing one type of error may inadvertently lead to an increase in the other type, highlighting the delicate and intricate balance that exists between avoiding false convictions (Type I) and ensuring guilty parties face appropriate consequences (Type II). Finding an optimal balance minimizes the chances of both types of errors occurring.

Thus, if you wish to reduce the level of Type I error, then reduce the significance level to a very low level, perhaps to $\alpha = 0.01$, or even to $\alpha = 0.001$. Remember though that, this implies a higher level of Type II error. Since the negative consequences of Type I error are not so negative, then it is preferable to provide a better balance of Types I and II error by adopting a significance level (α) such as 0.05 or 0.10.

---

5. Steps of Hypothesis Testing

Now, we will learn steps of Hypothesis Testing, using an example.

Suppose the Consumer Protection Court is interested in finding, if Haldiram company is packing 300 grams of chips in every packet (as written on the cover of packet).

The Judge knows that the company's production process cannot ensure exactly 300 grams in every packet, however, as long as the population mean filling weight is at least 300 grams per packet, the rights of consumers will be protected.

We will show how the Judge can check it by conducting a **lower tail hypothesis test.** Note that the process of doing Upper tailed test is similar to the process of lower tailed test.

The students may carefully note down steps of Hypothesis Testing, while we are doing this example. All questions of Hypothesis Testing follow similar steps.

Step 1. Develop null and alternate hypotheses

The first step is to develop the null and alternative hypotheses for the test. If the population mean filling weight is at least 300 grams per packet, the company's claim is correct. This establishes the null hypothesis for the test. However, if the population mean weight is less than 300 grams per packet, then the company's claim is incorrect.

With μ denoting the population mean filling weight, the null and alternative hypotheses are as follows, where hypothesized value of the population mean is $\mu_0 = 300$.

If the sample data indicate that H_0 cannot be rejected, then the consumer court should not take any action against the company. However, if the sample data indicate H_0 can be rejected, we will conclude that the alternative hypothesis, H_a: $\mu < 300$, is true. In this case, punitive action against the company would be justified.

Step 2. Specify level of significance

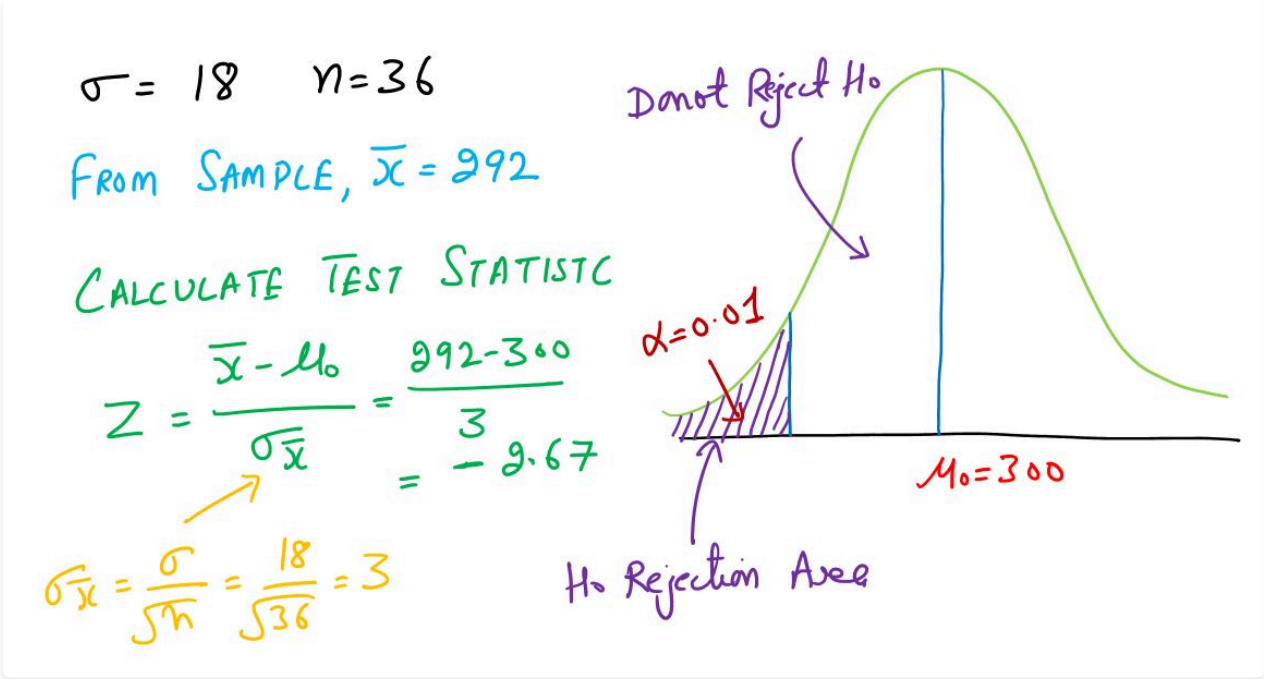
The decision maker must specify the level of significance. If the cost of making a Type I error is high, a small value should be chosen for the level of significance. If the cost is not high, a larger value is more appropriate. The Judge might say that "I am willing to risk a 1% chance of making Type I error".

Thus, we set the level of significance for the hypothesis test at $\alpha = 0.01$.

Step 3. Collect sample data and compute test statistic

For hypothesis tests about a population mean in the **σ known case**, we use the standard normal random variable z as a test statistic to determine whether deviates from the hypothesized value of μ enough to justify rejecting the null hypothesis.

Suppose Population Standard Deviation (from historic data) is, $\sigma = 18$ and we take sample size, $n = 36$.



Suppose mean of sample comes out to be, $\bar x$ = 292 litres

The test statistic (z) is as follows:
$z = \frac{\bar x- \mu_0 }{ \frac{ \sigma }{ \sqrt{n} } }$

where the standard error of mean is given by
$\sigma_\bar x = \frac{ \sigma }{ \sqrt{n} } = \frac{18}{ \sqrt{36} }$

We calculate Z Statistic is below:
$z = \frac{\bar x- \mu_0 }{ \frac{ \sigma }{ \sqrt{n} } }$ = $\frac{292-300}{ \frac{18}{ \sqrt{36} } }$ = -2.67

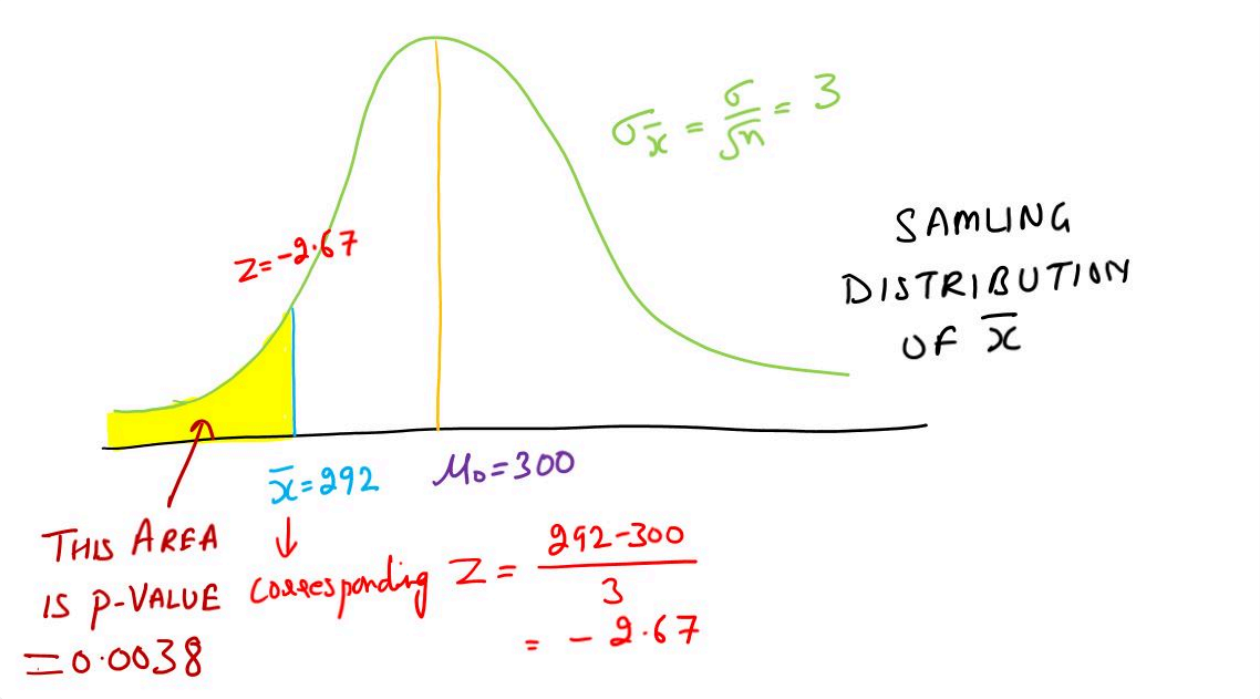
The key question for a lower tail test is, how small must the test statistic z be before we choose to reject the null hypothesis?

Two approaches can be used to answer this question: the p-value approach and the critical value approach.

First, we will carry on with the problem with p-value approach. Then later, we will do with critical value approach.

Step 4. Compute p-value

The p-value is used to determine whether the null hypothesis should be rejected. A p-value is a probability that provides a measure of the evidence against the null hypothesis provided by the sample. Smaller p-values indicate more evidence against H_0.

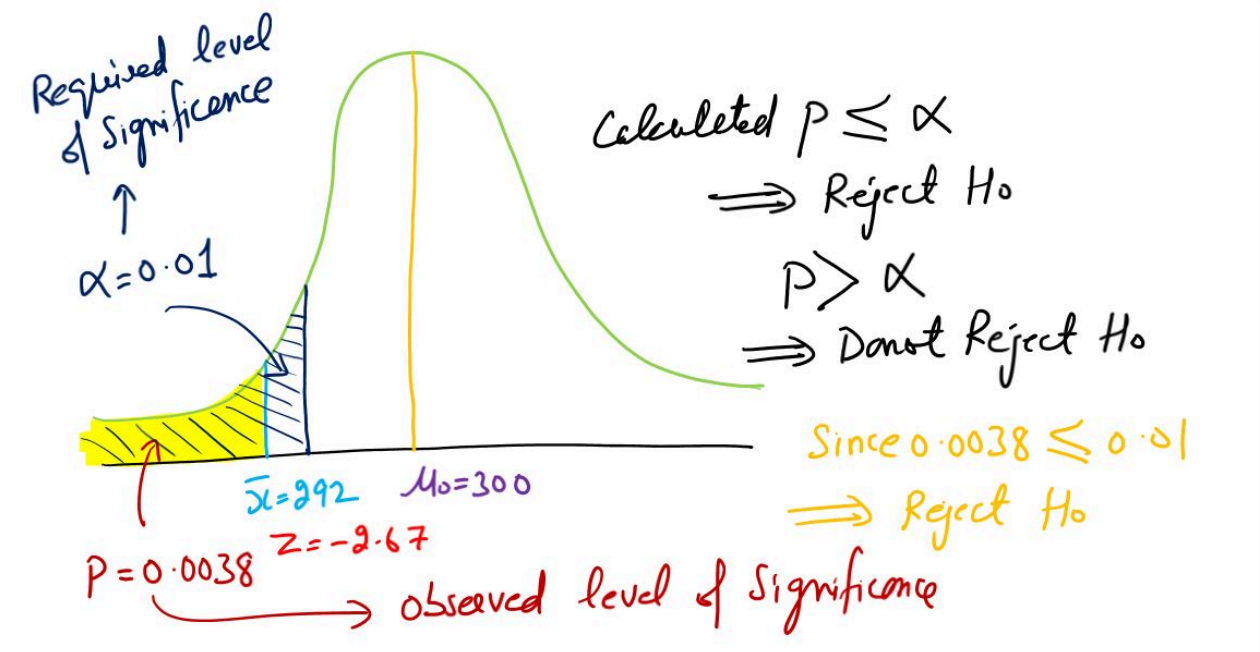


Using the standard normal probability table, we find that the lower tail area (p-value) at z = -2.67 is 0.0038. Look at the figure to appreciate relation in z and p.

Step 5. 'Reject' or 'Not Reject' Null Hypothesis

Our chosen, level of significance is α = 0.01.

The sample of 36 packets resulted in a p-value = 0.0038, which means that the probability of obtaining a value of \(\bar x \)= 2.92 or less when the null hypothesis is true as an equality is .0038. Because .0038 is less than or equal to α = 0.01, we reject H_0. Therefore, we find sufficient statistical evidence to reject the null hypothesis at the 0.01 level of significance.



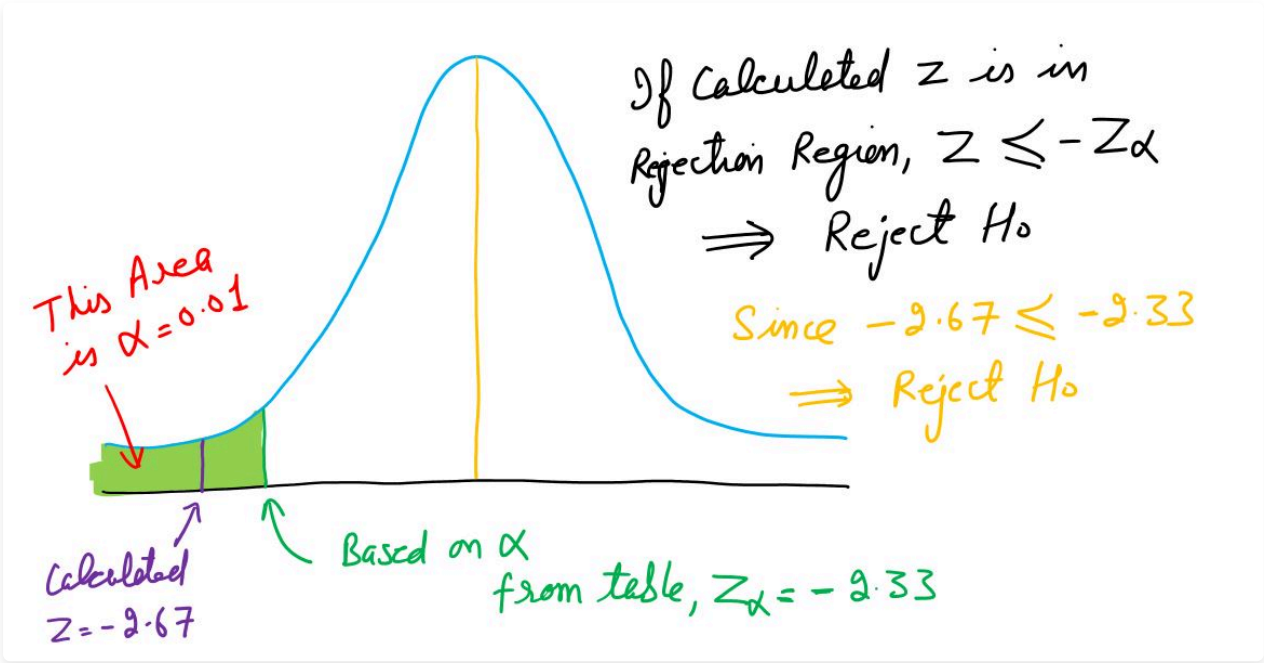
Thus we conclude that, we always Reject **Null Hypothesis, H_0 if p-value ≤ α**

The region of rejection of the null hypothesis is called the **critical region** for the hypothesis test. The critical region is sometimes referred to as the **region of rejection of H_0**, and the two terms are synonymous.

Critical Value Approach

Now, let us go through step 4 and Step 5 again, but with Critical Value Approach.

The critical value is the value of the test statistic that corresponds to an area of α = 0.01 in the lower tail of a standard normal distribution. Using the standard normal probability table, we find that z = -2.33 provides an area of 0.01 in the lower tail. Thus, if the sample results in a value of the test statistic that is less than or equal to -2.33, the corresponding p-value will be less than or equal to 0.01; in this case, we should reject the null hypothesis.



Thus the rejection rule is:

Reject H_0 if $z \leq -z_\alpha$

where $-z_\alpha$ is the critical value; that is, the z value that provides an area of α in the lower tail of the standard normal distribution.

The p-value approach and the critical value approach will always lead to the same rejection decision.

CONCLUSION

Since the Consumer Protection Court rejected the null hypothesis in this lower-tailed hypothesis test regarding the weight of chips packed by Haldiram company, it means that there is sufficient statistical evidence to support the claim that the population mean filling weight of the packets is less than 300 grams. The court should take action against the company.

---

5. Steps of Hypothesis Testing

In our example on consumer court, we understood how to conduct a lower tail test.

The upper tailed tests take following form:
H_0: $\mu \leq \mu_0$
H_a: $\mu > \mu_0$

We can use the same general approach to conduct an upper tail test. The test statistic z is still computed using same equation.

But, for an upper tail test, the p-value is the probability of obtaining a value for the test statistic as large as or larger than that provided by the sample. Thus, to compute the p-value for the upper tail test, we must find the area under the standard normal curve to the **right of the test statistic.**

For lower tail tests, the null hypothesis is rejected if the value of the test statistic is less than or equal to the critical value. For upper tail tests, the null hypothesis is rejected if the value of the test statistic is greater than or equal to the critical value.

In other words, for Upper Tailed Tests.

Reject H_0 if \(z > z_ \alpha \)

---

6. Two Tailed Tests

After having learnt, one tailed test, let us understand, how same steps of Hypothesis Testing can be applied to two tailed tests.

The general form of a two tailed test is
H_0: $\mu = \mu_0$
H_a: $\mu \neq \mu_0$

Suppose a supplier supplies a metallic panel to Maruti, with mean length of 295 cm. Since the panel will fit in a car, Maruti will NOT prefer length to be either greater or lesser than 295 cm.

Step 1. Develop Null Alternative Hypotheses

The first step is to develop the null and alternative hypotheses for the test. The hypothesized value of the population mean is $\mu_0 = 295$.
H_0: $\mu = 295$
H_a: $\mu \neq 295$

If the sample mean $\bar x$ is significantly less than 295 cm or significantly greater than 295 cm, we will reject H_0.

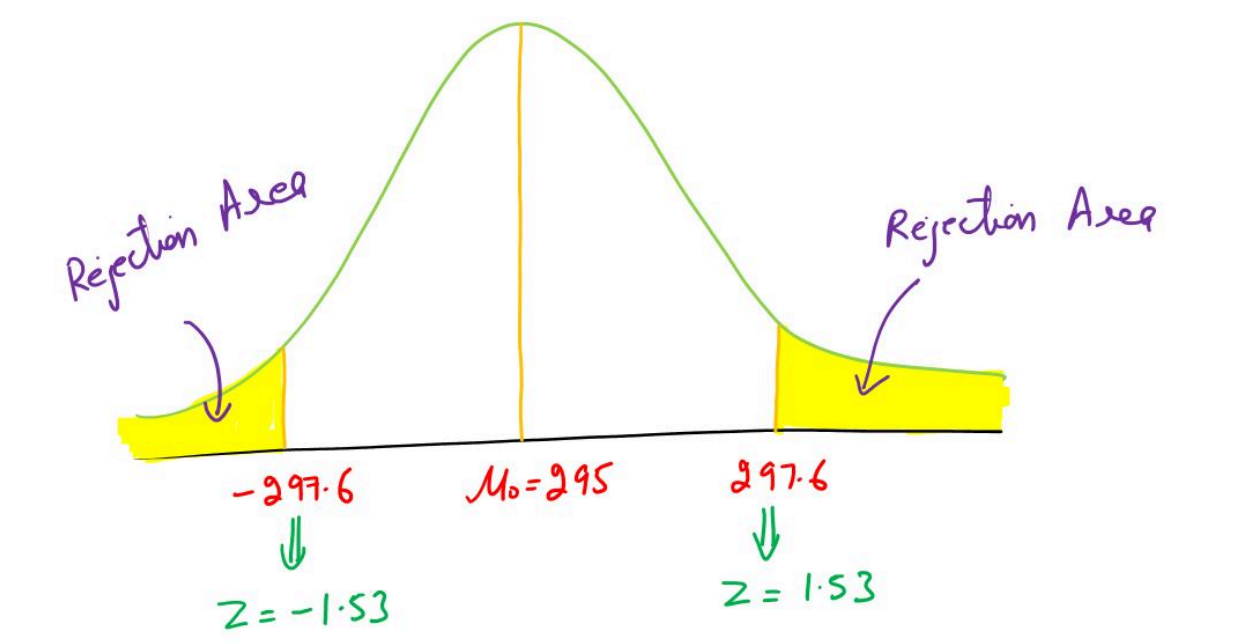
Step 2. Specify level of significance

The decision maker must specify the level of significance. The Maruti might say that "I am willing to risk a 5% chance of making Type I error".

Thus, we set the level of significance for the hypothesis test at $\alpha = 0.05$.

Step 3. Collect sample data and compute test statistic

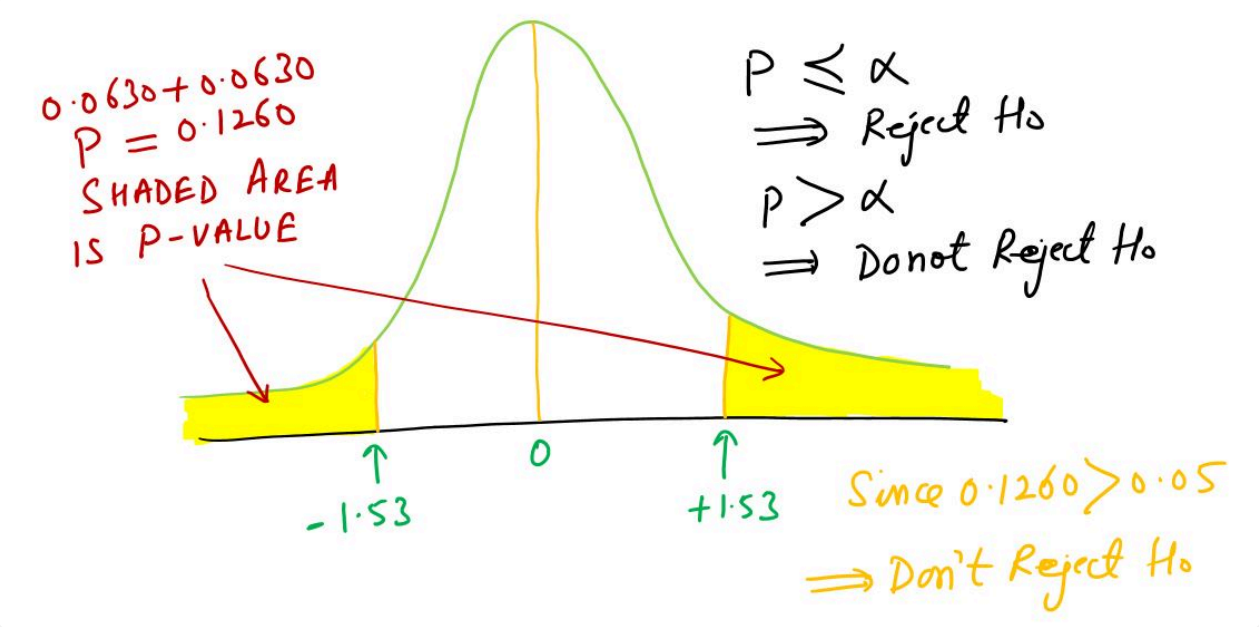
Suppose Population Standard Deviation (from historic data) is, $\sigma = 12$ and we take sample size, $n = 50$.
Then the standard error of mean is given by
$\sigma_{\bar x} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{50}} = 1.7$



We calculate Z Statistic is below:
$z = \frac{\bar x- \mu_0 }{ \frac{ \sigma }{ \sqrt{n} } } = \frac{297.6-295}{ \frac{12}{ \sqrt{50} } } = 1.53$

Step 4. Compute the p-value

Recall that the p-value is a probability used to determine whether the null hypothesis should be rejected. For a two-tailed test, values of the test statistic in either tail provide evidence against the null hypothesis.



Now to compute the p-value we must find the probability of obtaining a value for the test statistic at least as unlikely as z = 1.53. Clearly values of z ≥ 1.53 are at least as unlikely. But, because this is a two-tailed test, values of z ≤-1.53 are also at least as unlikely as the value of the test statistic provided by the sample. As shown in the Figure, we note that the two-tailed p-value in this case is given by P (z ≥ 1.53) + P (z ≤ - 1.53).

The table for the standard normal distribution shows that the area to the left of z = 1.53 is 0.9370. Thus, the area under the standard normal curve to the right of the test statistic z = 1.53 is 1.0000 - 0.9370 = 0.0630. Doubling this, we find the p-value for our example is 2 × (.0630) = 0.1260.

Please note that, if the value of the test statistic is in the upper tail (z > 0), we find the area under the standard normal curve to the right of z. If the value of the test statistic is in the lower tail (z < 0), find the area under the standard normal curve to the left of z.

Step 5. 'Reject' or 'Not Reject' Null Hypothesis

Our chosen, level of significance is α = 0.05.

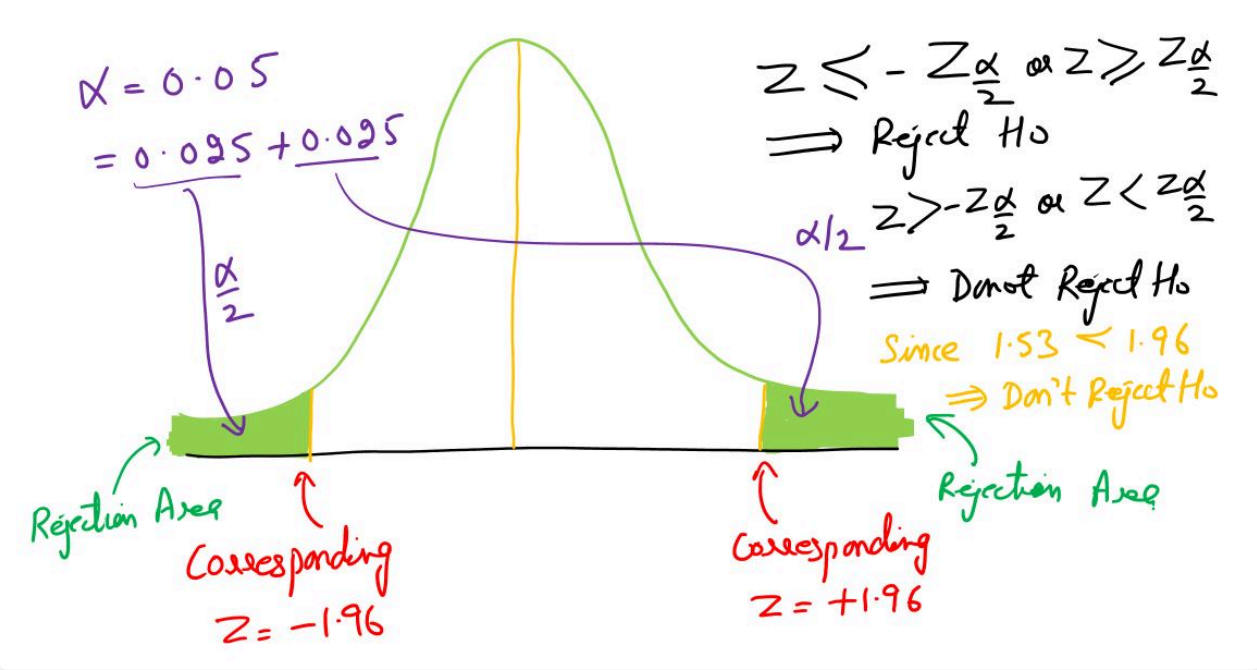
Next we compare the p-value to the level of significance to see whether the null hypothesis should be rejected. With a level of significance of α = 0.05, we do not reject H_0 because the p-value = 0.1260 > 0.05.

Because the null hypothesis is not rejected, no action will be taken on the Supplier by the Maruti and to adjust the manufacturing process.

Thus we conclude that, we Reject Null Hypothesis, H_0 if p-value ≤ α

Critical Value Approach

Let us see how the test statistic z can be compared to a critical value for a two-tailed test. As shown in the Figure, the critical values for the test will occur in both the lower and upper tails of the standard normal distribution.



With a level of significance of α = 0.05, the area in each tail beyond the critical values is \(\frac{ \alpha }{2} = \frac{0.05}{2} \) = 0.025. Using the standard normal probability table, we find the critical values for the test statistic are - $z_{(0.025)}$ = -1.96 and $z_{(0.025)}$ = 1.96.

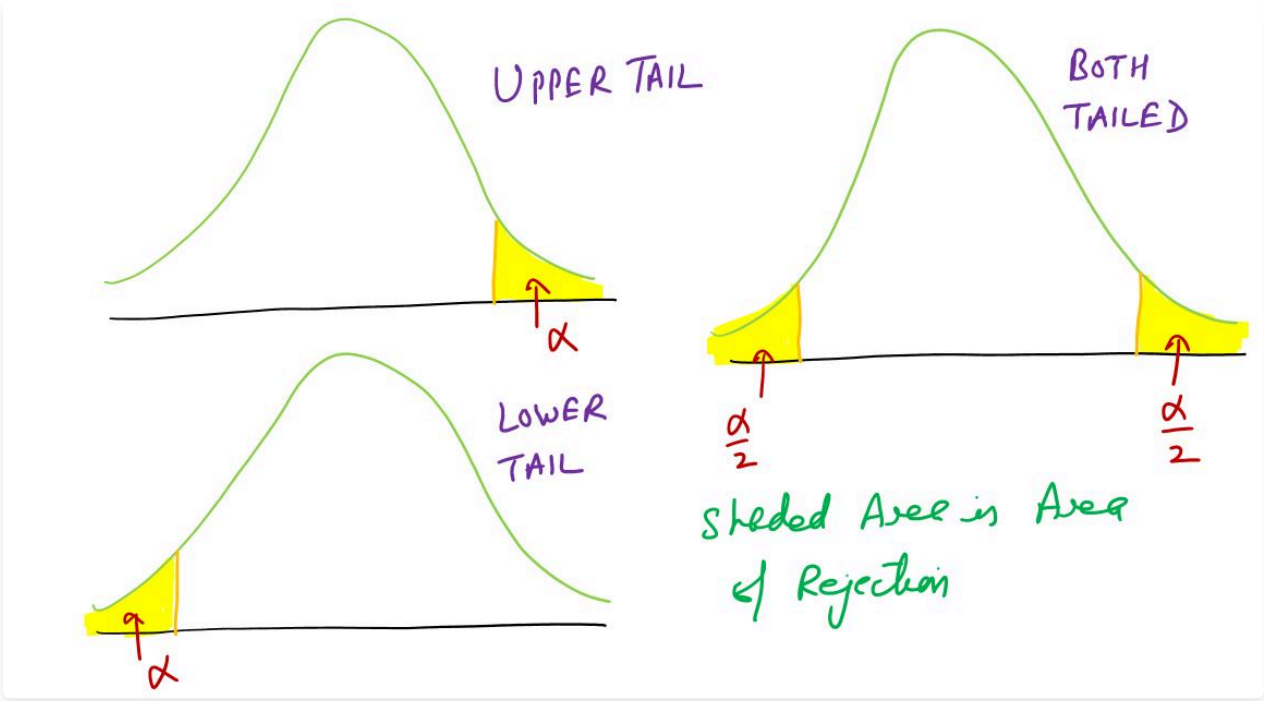
Thus, using the critical value approach, the two-tailed rejection rule is

Reject H_0 if \(z \leq- z_ \frac{ \alpha }{2} \) or if \(z \geq z_ \frac{ \alpha }{2} \)

We do not reject null hypothesis, H_0 , in our example, because 1.53 is less than 1.96.

One Tailed and Two Tailed Tests

Check out below figure to see difference between one-tailed and teo-tailed tests, in identifying rejection araes.



CONCLUSION

Not rejecting the null hypothesis indicates that, based on the sample data and at the chosen level of significance, there isn't enough evidence to suggest that the average length of the metallic panels supplied to Maruti significantly deviates from the specified 295 cm.

---

7. One Tailed and Two Tailed Tests

The choice between a one-tailed or two-tailed test depends on the research hypothesis and the nature of the predicted effect.

1. One-Tailed Test
A one-tailed test (also called a directional test) is used when the research hypothesis predicts the direction of the effect. In this test, the area of rejection (the critical region) is located only on one side of the distribution (either the left or the right), depending on the hypothesis.

Use a one-tailed test when you have a specific direction in mind for the expected relationship or difference between variables. This test is used when the research hypothesis predicts either a positive or negative effect, but not both.

Example:

Hypothesis: "Increasing the amount of study time will increase test scores."

Here, the researcher predicts a positive effect, so the one-tailed test would test whether the study time leads to an increase in test scores (on the right side of the distribution).

Null Hypothesis: There is no change or effect in test scores with study time.

Alternative Hypothesis: Increasing study time increases test scores.

Critical Region: For a one-tailed test, the critical region (the area where you reject the null hypothesis) is only in one tail of the distribution (either left or right). The p-value is calculated for only one side.

2. Two-Tailed Test
A two-tailed test (also called a non-directional test) is used when the research hypothesis does not predict the direction of the effect. In this test, the critical region (the area of rejection) is located in both tails of the distribution (left and right), which allows for the possibility of an effect in either direction.

When to Use: Use a two-tailed test when you are unsure of the direction of the effect or when you want to test for the possibility of an effect in either direction (positive or negative).

Example:

Hypothesis: "There is a relationship between the amount of study time and test scores."

In this case, the researcher does not predict whether study time will increase or decrease test scores. They just predict a relationship exists in either direction.

Null Hypothesis: There is no relationship between study time and test scores.

Alternative Hypothesis: There is a relationship between study time and test scores (can be either an increase or decrease).

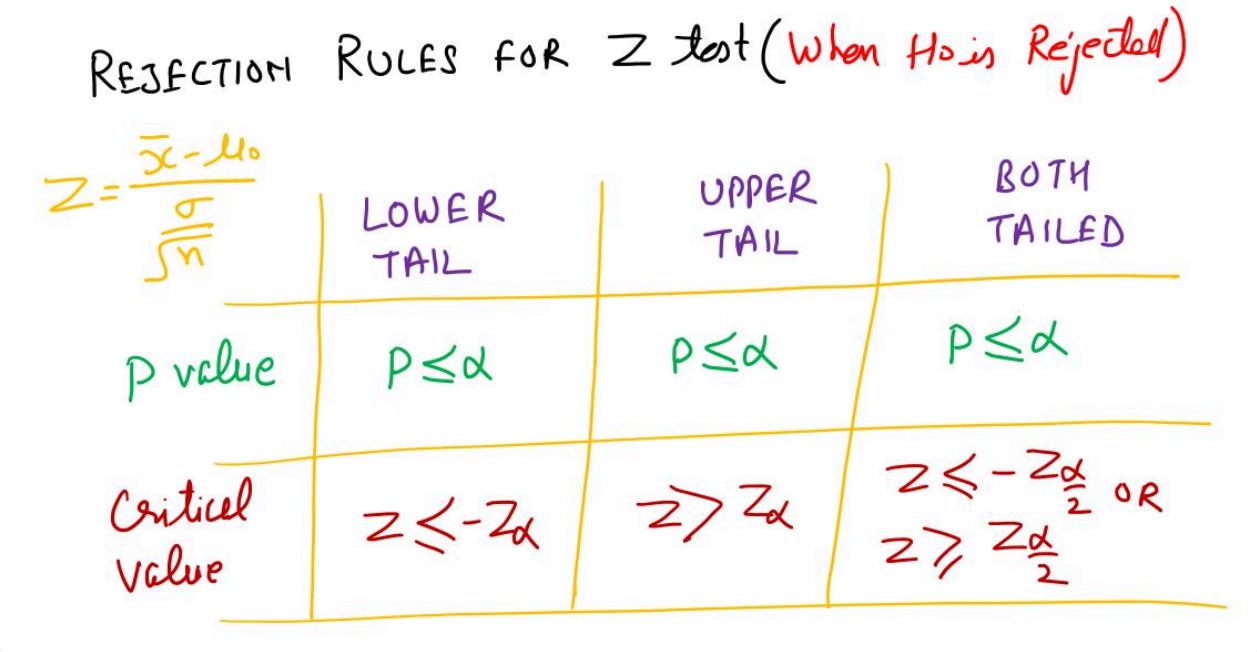
Critical Region: In a two-tailed test, the critical region is split into two equal parts, one in the left tail and one in the right tail of the distribution. The p-value is calculated for both sides.

---

8. Process of Hypothesis Testing

After studying various examples, let us summarize steps of hypothesis testing.

Step 1. Develop the null and alternative hypotheses.
Step 2. Specify the level of significance.
Step 3. Collect the sample data and compute the value of the test statistic.



You can use either of (a) p value approach or (b) critical value approach for Steps 4 and 5.

Approach	Step 4	Step 5
p-Value Approach	Use the value of the test statistic to compute the p-value.	Reject H_0 if the p-value ≤ α.
Critical Value Approach	Use the level of significance to determine the critical value and the rejection rule.	Use the value of the test statistic and the rejection rule to determine whether to reject H_0.

Summary of rejection rules of all three types of tests is tabled next (**using z statistic**).

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypothesis	H_0: μ ≥ μ_0 H_a: μ < μ_0	H_0: μ ≤ μ_0 H_a: μ > μ_0	H_0: μ = μ_0 H_a: μ ≠ μ_0
Test Statistics	\(z = \frac{\bar x- \mu_0}{ \frac{ \sigma}{ \sqrt{n}}} \)	\(z = \frac{\bar x- \mu_0}{ \frac{ \sigma}{ \sqrt{n}}} \)	\(z = \frac{\bar x- \mu_0}{ \frac{ \sigma}{ \sqrt{n}}} \)
Rejection Rule: p-Value Approach	Reject H_0 if p-value ≤ α	Reject H_0 if p-value ≤ α	Reject H_0 if p-value ≤ α
Rejection Rule: Critical Value Approach	Reject H_0 if \(z \leq-z_ \alpha \)	Reject H_0 if \(z \geq z_ \alpha \)	Reject H_0 if \(z \leq- z_ \frac{ \alpha }{2} \) or if \(z \geq z_ \frac{ \alpha }{2} \)

9. Applications of Hypothesis Testing

Till now, we have learnt:

1. Hypothesis Testing for Mean of Population (both one tailed and two tailed), when Standard Deviation of Population (σ) is Known. These are called **z tests.**

In next sections, we will learn:

2. Hypothesis Testing for Mean of Population (both one tailed and two tailed), when Standard Deviation of Population (σ) is **NOT** Known. These are called **t Tests.**

3. Hypothesis Testing for Proportion of Population

4. Hypothesis Testing for Variance of Population

---

10. t test (σ is unknown)

We have understood that when we have knowledge of the population's standard deviation, we can conduct Hypothesis Testing using the z-test statistic (z-test).

However, when the standard deviation of the population is unknown, we employ the t-test statistic instead of the z-test.

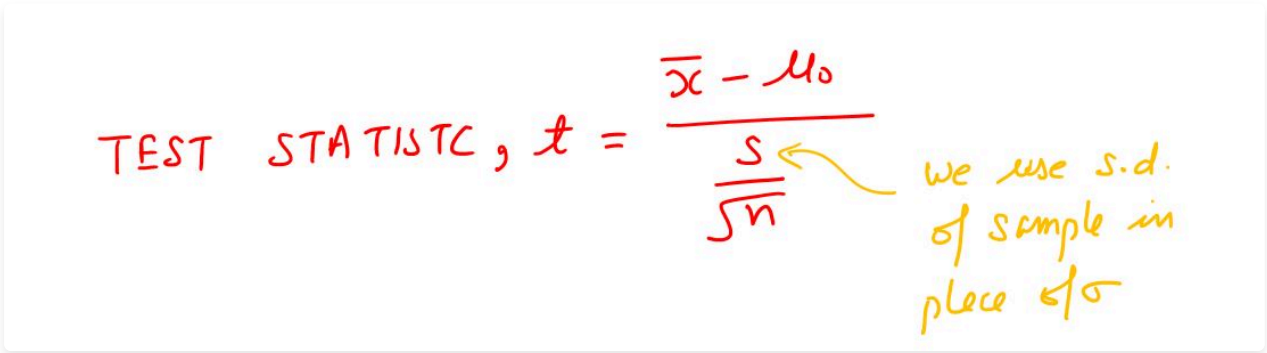
It is important to highlight that utilizing the t-test requires the population to adhere to a normal distribution.

To conduct a hypothesis test about a population mean for the σ unknown case, the sample standard deviation s is used as an estimate of population standard deviation σ.

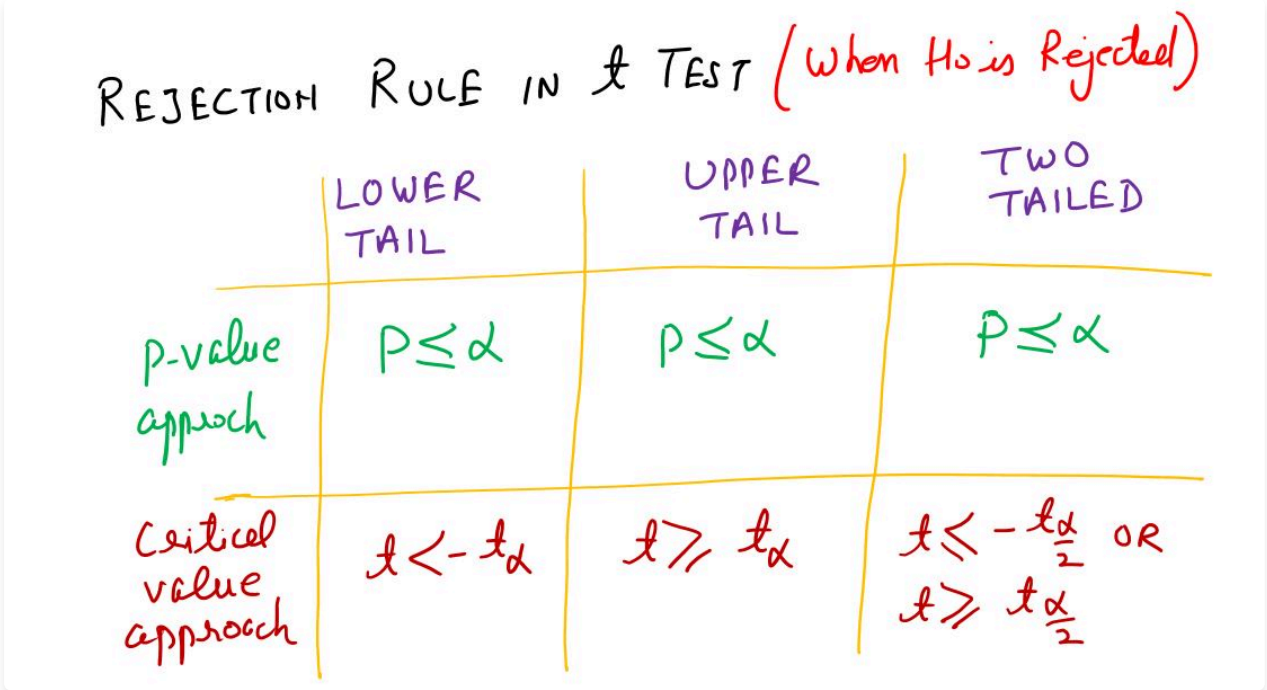
The steps of the hypothesis testing procedure for the "σ unknown case" are the same as those for the "σ known case". Recall that for the σ known case, the sampling distribution of the test statistic has a standard normal distribution.

For the σ unknown case, however, the sampling distribution of the test statistic follows the t distribution (with n - 1 degrees of freedom); it has slightly more variability because the sample is used to develop estimates of both μ and σ.

Test Statisic is calculated as given below:



Rejection Rules are given below.



10. t test (σ is unknown)

Suppose, Uber takes feedback of drivers from the riders on a scale of 1 to 10. All drivers with rating greater than 7 will be Champions and offered high incentives. Mohammad is a driver with Uber. Feedback of 60 riders is obtained about Mohammad. The mean of sample is $\bar x$ = 7.25 and standard deviation of sample data, s = 1.052. Should Mohammad be selected for Champion category?

Step 1. Develop null and alternative hypotheses

We want to develop a hypothesis test for which the decision to reject H_0 will lead to the conclusion that the population mean rating for Mohammad is greater than 7.

Thus, an upper tail test with H_a: $\mu > 7$ is required.

The hypothesized value of the population mean is $\mu_0 = 7$.
H_0: $\mu \leq 7$
H_a: $\mu > 7$

Step 2. Specify the level of significance

We set the level of significance for the hypothesis test at $\alpha = 0.05$.

Step 3. Collect sample data and compute test statistic

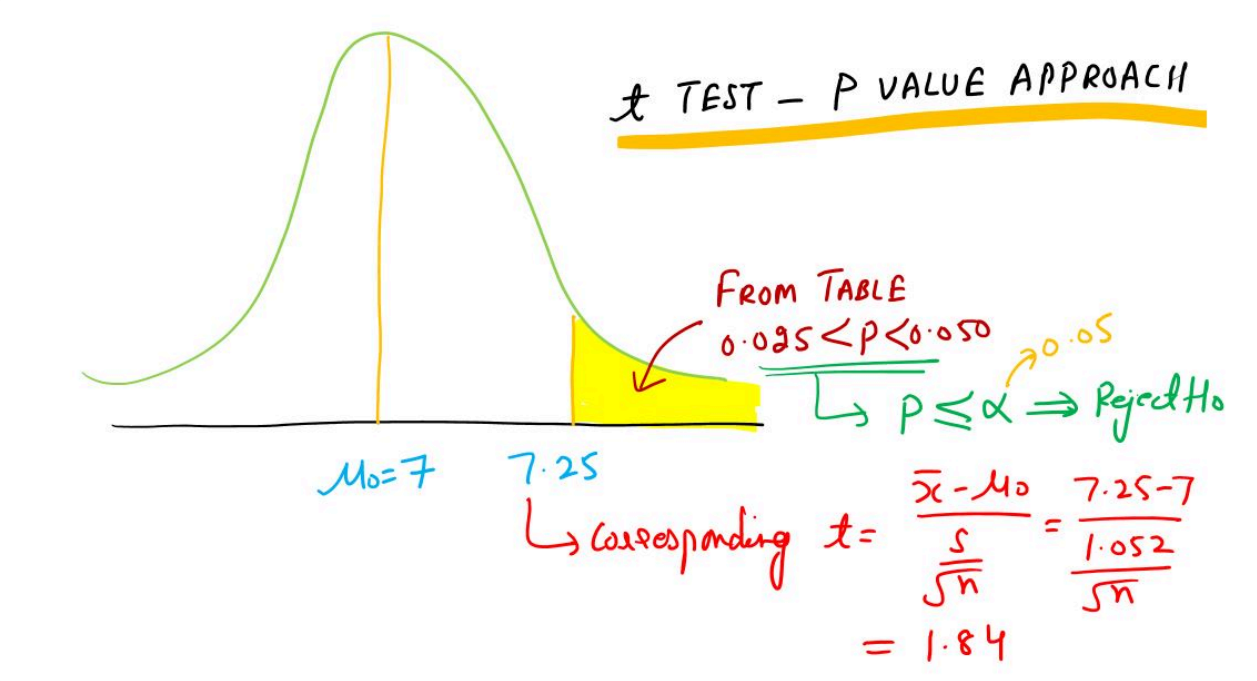
We calculate t Statistic is below:

$t = \frac{\bar x - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{7.25-7}{\frac{1.052}{\sqrt{60}}} = 1.84$

Step 4. Compute the p-value

The sampling distribution of t has 60 – 1 = 59 degrees of freedom. Because the test is an upper tail test, the p-value is the area under the curve of the t distribution to the right of t = 1.84.

Then we use a t table to compute p-value. The p value comes out to be 0.0354 (from t table).



Step 5. 'Reject' or 'Not Reject' Null Hypothesis

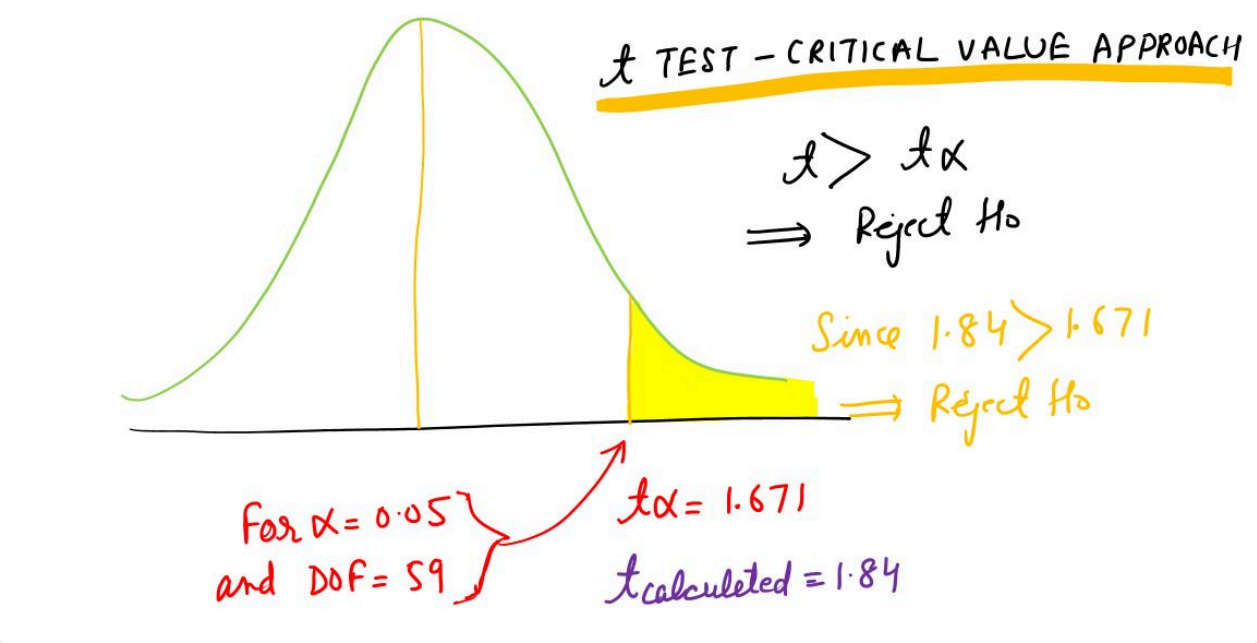
Our chosen, level of significance is $\alpha = 0.05$. Our p value is 0.0354

We Reject Null Hypothesis, H_0 if p-value $\leq \alpha$

Because $0.0354 \leq 0.05$, we reject Null Hypothesis in and Conclude that Mohammad should be made Champion.

Critical Value Appraoch

For the level of significance α = 0.05 and Degrees of freedom = 59, we calculate value of t_α. From t table, the value of t_α comes out to be 1.671.



We Reject Null Hypothesis, H_0 if t \(\geq \) t_α

Since 1.84 is \(\geq \) 1.671, we reject Null Hypothesis.

CONCLUSION

Rejecting the null hypothesis indicates that, based on the sample data and at the chosen level of significance, there is enough evidence to conclude that Mohammad's mean rating from the riders exceeds the threshold of 7, which is the criterion for being selected as a Champion driver by Uber. Therefore, based on the statistical analysis, Mohammad is likely eligible to be categorized as a Champion and offered high incentives by Uber.

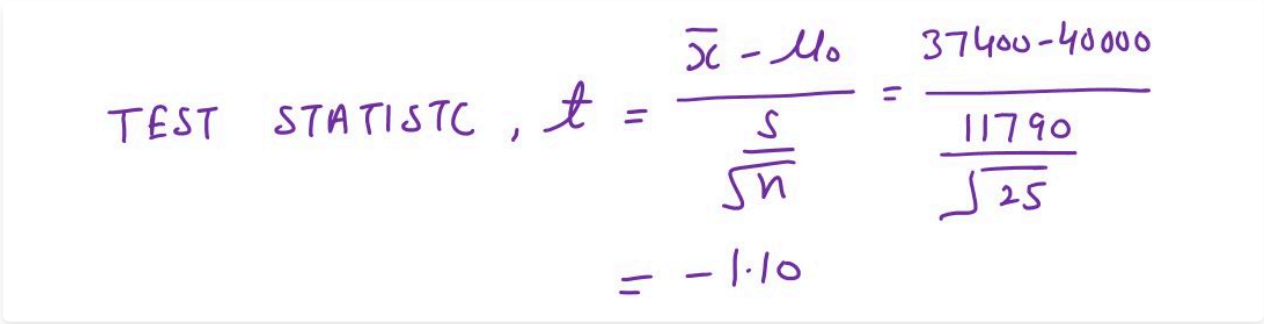
---

10. t test (σ is unknown)

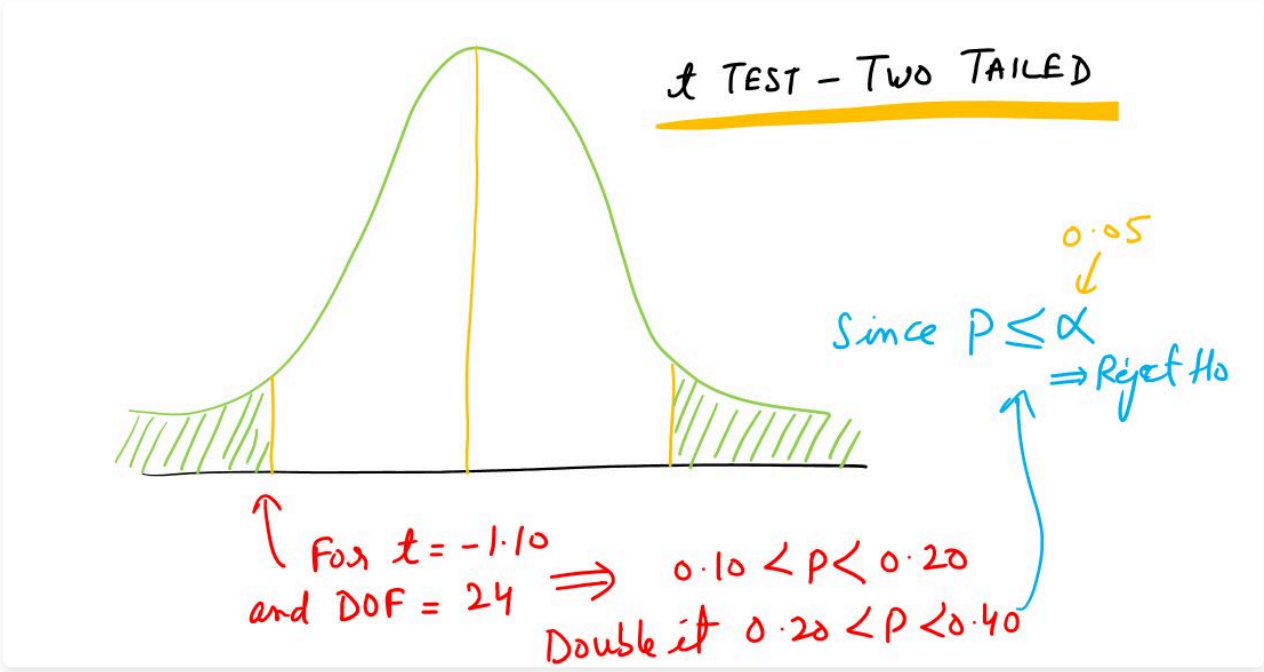
A new product plant is being established with an anticipated market demand of 40,000 units. A sample of 25 retailers was selected to evaluate this expected demand. Upon analysis, the sample mean for the 25 retailers was found to be 37,400 units, with a standard deviation of 11,790 units. Based on this data, conduct a hypothesis test to determine whether there is sufficient evidence to support the claim that the expected market demand is different from 40,000 units. The level of significance is 0.05.

Null and Alternate Hypothesis will be:

The Test Stsitic, t is calculated as below:



Using p-value approach, we find that for t=1.10 and Degree of Freedom=24, the p value is 01.10<p<0.20. But since it is two tailed test, we double it, the p value range come out to be 0.20 < p <0.40.



Since p value range is not less that \(\alpha \), we do not reject Null Hypothesis.

Since the null hypothesis is not rejected, it implies that there isn't sufficient statistical evidence to conclude that the expected market demand significantly differs from the claimed value of 40,000 units based on the sample data.

CONCLUSION

Not rejecting the null hypothesis suggests that, based on the sample data and at the chosen significance level of 0.05, there isn't adequate evidence to conclude that the average market demand, as indicated by the sample of 25 retailers, significantly deviates from the anticipated 40,000 units. Therefore, based on this statistical analysis, there might not be enough grounds to assert that the expected market demand differs from the anticipated value.

---

10. t test (σ is unknown)

The summary of both One tailed tests and Two tailed test presented in the table below (using t statistic):

	Lower Tail Test	**Upper Tail Test**	**Two-Tailed Test**
Hypothesis	H_0: $\mu \geq \mu_0$ H_a: $\mu < \mu_0$	H_0: $\mu \leq \mu_0$ H_a: $\mu > \mu_0$	H_0: $\mu = \mu_0$ H_a: $\mu \neq \mu_0$
Test Statistics	$t = \frac{\bar x- \mu_0}{ \frac{s}{n} }$	$t = \frac{\bar x- \mu_0}{ \frac{s}{n} }$	$t = \frac{\bar x- \mu_0}{ \frac{s}{n} }$
Rejection Rule: p-Value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $t \leq -t_{\alpha}$	Reject H_0 if $t \geq t_{\alpha}$	Reject H_0 if $t \leq -t_{\frac{\alpha}{2}}$ or if $t \geq t_{\frac{\alpha}{2}}$

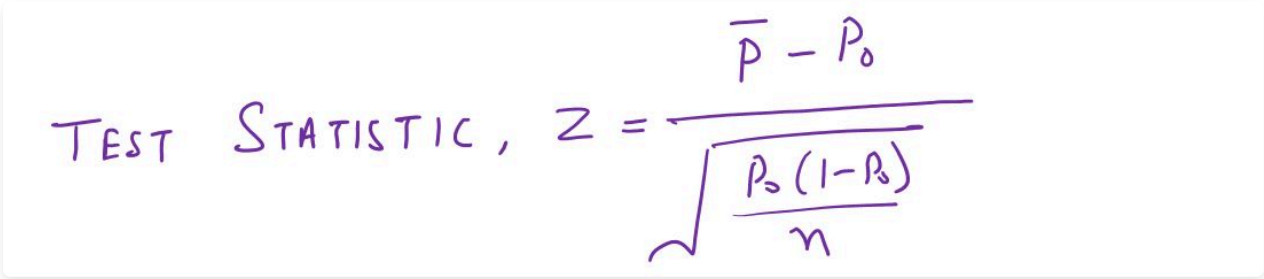
---

11. Test for Population Proportion

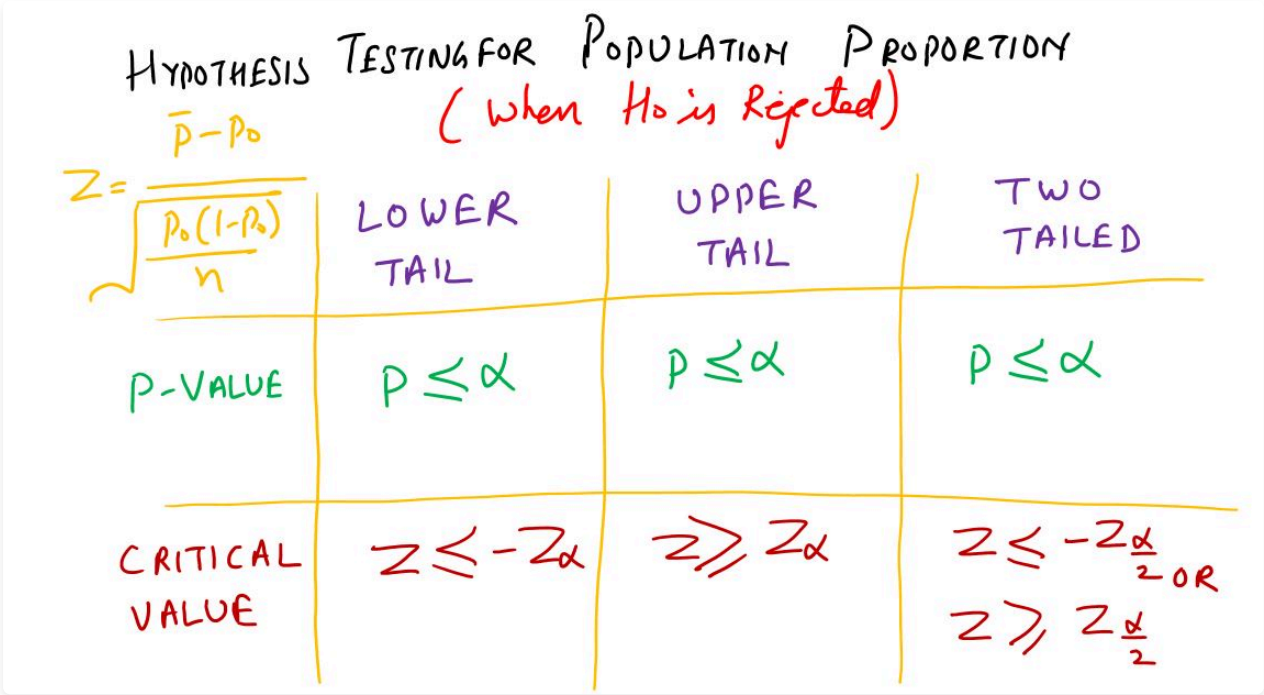
The procedure used to conduct a hypothesis test about a population proportion is similar to the procedure used to conduct a hypothesis test about a population mean.

We assume that np ≥ 5 and n (1-p) ≥ 5; thus the normal probability distribution can be used to approximate the sampling distribution of $\bar p$. These are essential conditions for conducting hypothesis testing for proportion.

Test Ststistics is given below:



The rejection rules are given below:



11. Test for Population Proportion

Suppose 20% of participants are women at Gold Gym. In order to increase participation of women, the Gold Gym started a marketing campaign. After 3 months of campaign, the Gold Gym wants to find out, if proportion of women has increased? A sample of 400 participants was taken, out of which 100 were women.

Step 1. Develop null and alternative hypotheses

Because the objective of the study is to determine whether the proportion of women increased, an upper tail test with H_a: $\mu > 0.20$ is appropriate. The null and alternative hypotheses are:

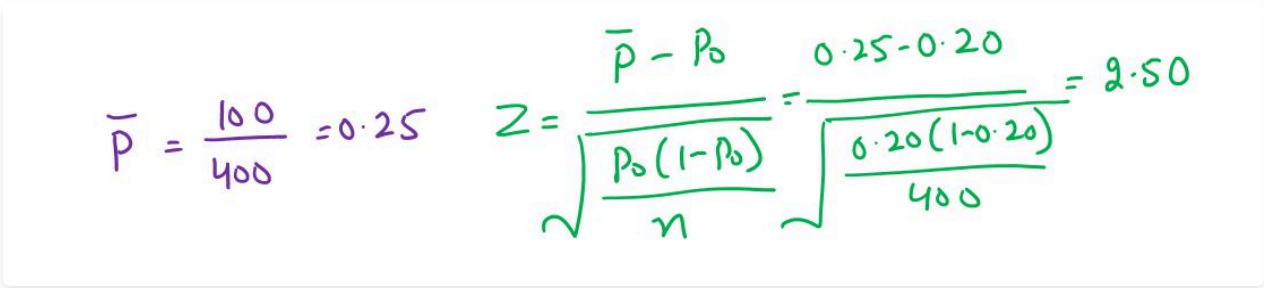
Step 2. Specify level of significance

We set the level of significance for the hypothesis test at $\alpha = 0.05$.

Step 3. Compute test statistic

We calculate z Statistic is below:

$z = \frac{\bar p - p_0}{ \sigma_\bar p}$, where $\sigma_\bar p = \sqrt{ \frac{p_0(1-p_0)}{n} }$



Step 4. Compute the p-value

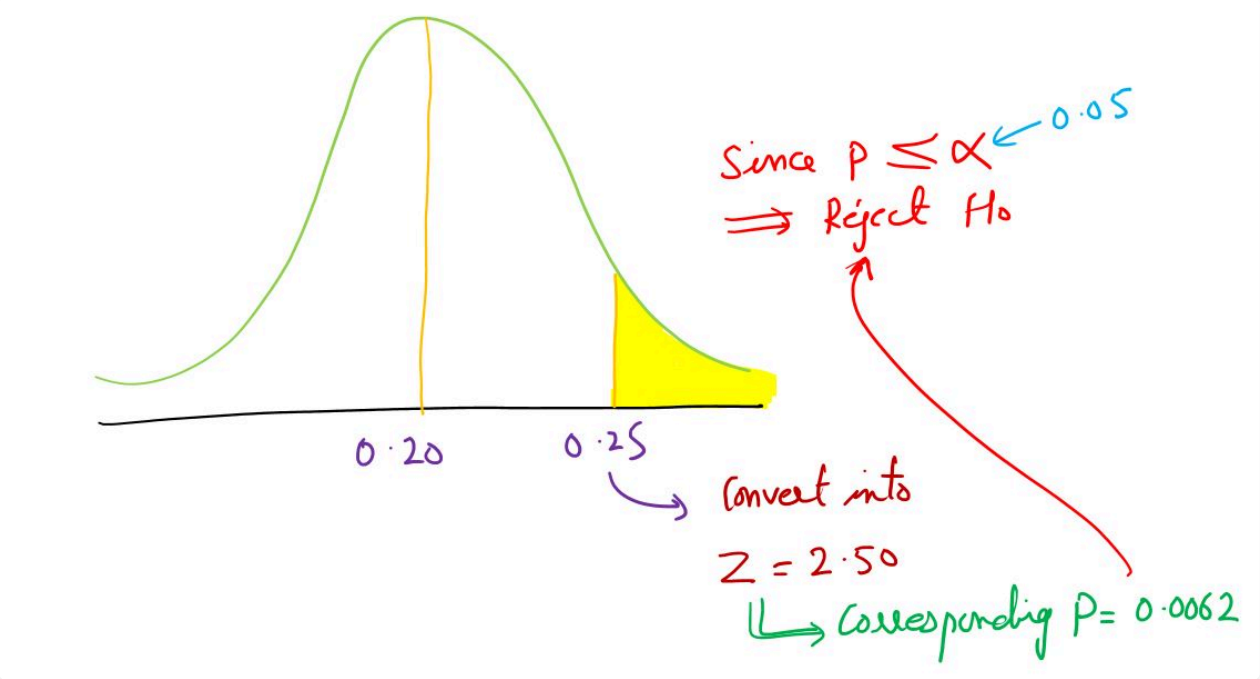
Because it is an upper tail test, the p-value is the probability that z is greater than or equal to z = 2.50; that is, it is the area under the standard normal curve for $z \geq 2.50$.

Using the standard normal probability table, we find that the area to the left of z = 2.50 is 0.9938. Thus, the p-value is 1.0000-0.9938 = 0.0062.

Step 5. 'Reject' or 'Not Reject' Null Hypothesis

Our chosen, level of significance is $\alpha = 0.05$. Our p value is 0.0062

We Reject Null Hypothesis, H_0 if p-value $\leq \alpha$



Because 0.0062 ≤ 0.05, we reject Null Hypothesis in our example.

CONCLUSION

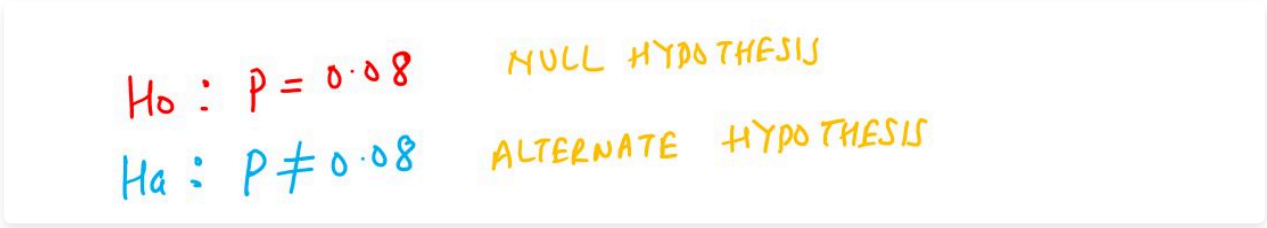
Rejecting the null hypothesis suggests that, based on the sample data and at the chosen level of significance, there is enough evidence to conclude that the proportion of women among the participants at Gold Gym has significantly risen from the initial 20% following the three-month marketing campaign. Therefore, it implies that the efforts made by Gold Gym to increase female participation through the marketing campaign have been successful, leading to a higher proportion of women engaging in the gym activities.

---

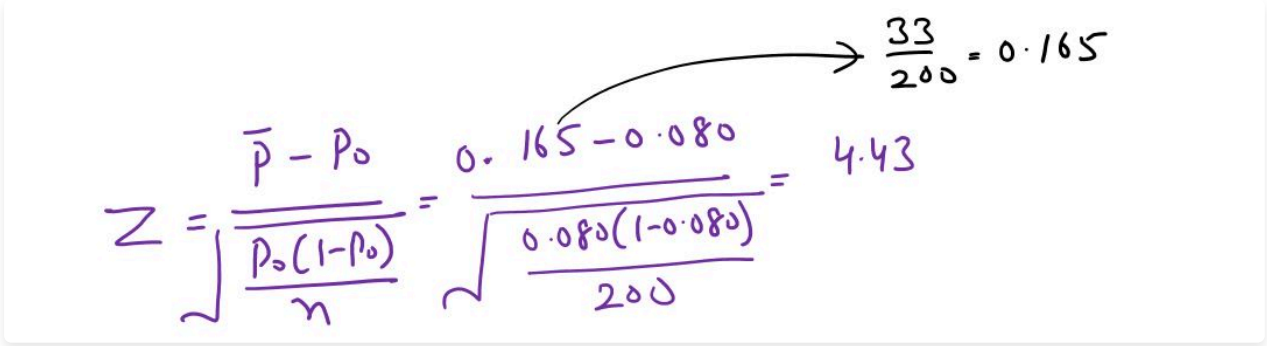
11. Test for Population Proportion

A manufacturer claims that only 8% of their products are defective. To assess this claim, a sample of 200 items was randomly selected, revealing that 33 of them were defective. Using a significance level of 0.10, conduct a hypothesis test to determine whether there is enough evidence to suggest that the actual proportion of defective products differs from the manufacturer's claim of 8%.

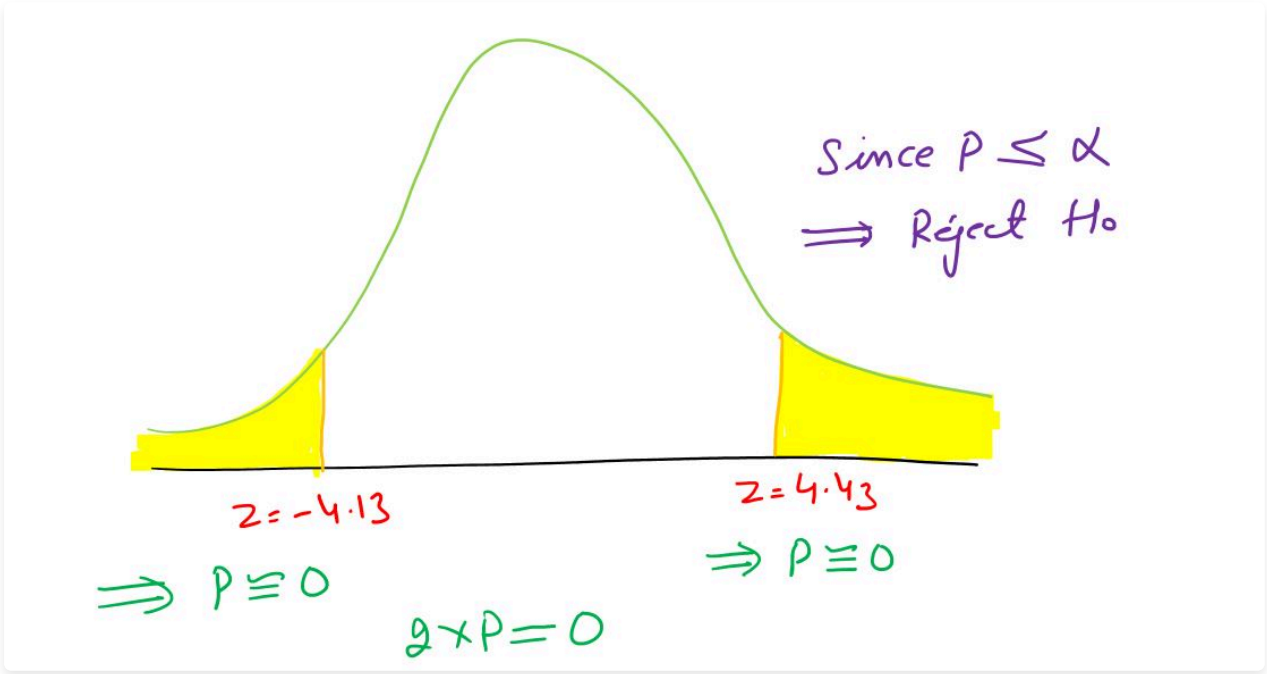
Null and Alternate Hypothesis are given below:



The Test Statistic is given below:



The p values comes out to be close to Zero. Since this p\(\leq \alpha \), we reject the null hypothesis.



CONCLUSION

Since the null hypothesis is rejected based on the hypothesis test results, it implies that there is sufficient statistical evidence to conclude that the actual proportion of defective products differs significantly from the manufacturer's claimed proportion of 8%.

In practical terms, it suggests that the observed sample proportion of defective products (calculated from the sample of 200 items where 33 were defective) is significantly higher or lower than the claimed 8% at the chosen significance level of 0.10.

11. Test for Population Proportion

The summary of one tailed and two tailed tests for proportion of population is given below:

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypothesis	H_0: $\mu \geq \mu_0$ H_a: $\mu < \mu_0$	H_0: $\mu \leq \mu_0$ H_a: $\mu > \mu_0$	H_0: $\mu = \mu_0$ H_a: $\mu \neq \mu_0$
Test Statistics	$z = \frac{\bar p-p_0}{ \sqrt{ \frac{p_0(1-p_0)}{n} } }$	$z = \frac{\bar p-p_0}{ \sqrt{ \frac{p_0(1-p_0)}{n} } }$	$z = \frac{\bar p-p_0}{ \sqrt{ \frac{p_0(1-p_0)}{n} } }$
Rejection Rule: p-Value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $z \leq -z_{\alpha}$	Reject H_0 if $z \geq z_{\alpha}$	Reject H_0 if $z \leq -z_{\frac{\alpha}{2}}$ or if $z \geq z_{\frac{\alpha}{2}}$

12. Test for Population Variance

The procedure for hypothesis testing for population variance closely resembles that for population mean or proportion. The only difference is the utilization of the chi-square distribution (χ^2) instead of the z or t distributions.

However, it may be noted that this test can only be conducted under the presumption that the population follows a normal distribution.

The procedure for conducting a hypothesis test about a population variance uses the hypothesized value for the population variance σ_0^2 and the sample variance s^2 to compute the value of a χ^2 test statistic.

χ^2 = $\frac{(n-1) s^2}{\sigma_0^2}$

After computing the value of the χ^2 test statistic, either the p-value approach or the critical value approach may be used to determine whether the null hypothesis can be rejected.

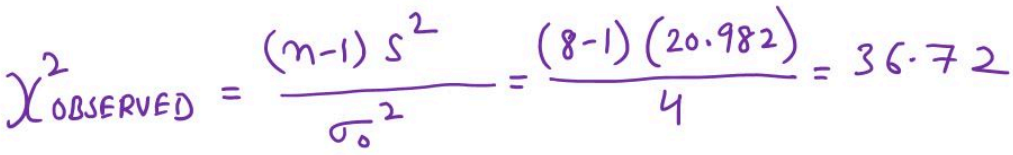
12. Test for Population Variance

The variance in the weight of chips in a packet was claimed to be 4 grams. Following readjustment in the production equipment, an investigation is conducted to determine if the variance has increased. A sample of 8 observations was gathered for this analysis, for which the variance came out to be 20.9821. Using a significance level of 0.05, perform a hypothesis test to assess whether the variance in the weight of chips has indeed increased after the readjustment.

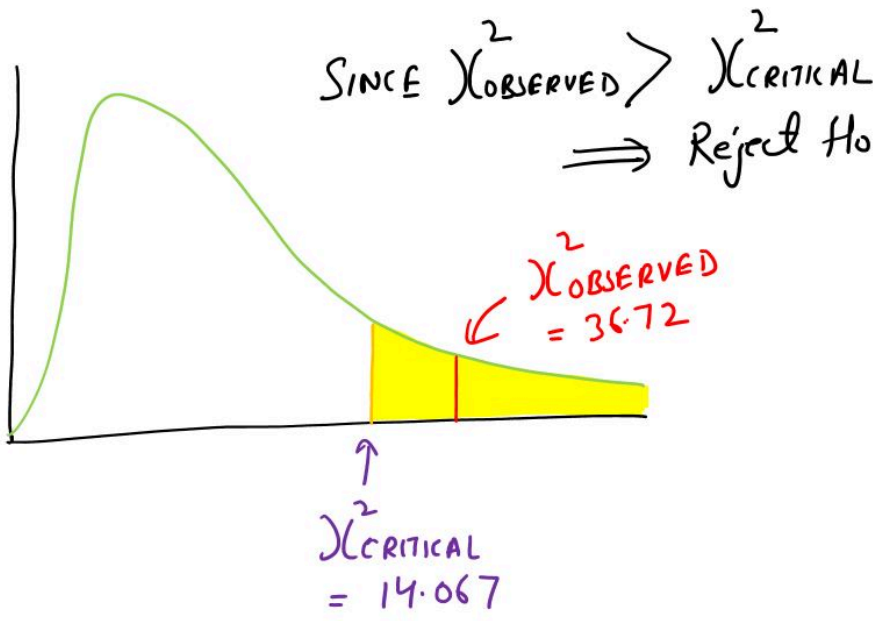
The Null and Alternate Hypothesis are written below:

From Degree of Freedom (n-1) = 7 and \(\alpha \)= 0.05, we calculate critical value from chi square table. It comes out to be 14.067.

Then we calculated observed value of Chi Square, using below formula.



Since Observed Chi Sqaure is > Critical Chi Square, we reject null hypothesis.



CONCLUSION

Since the null hypothesis is rejected, it implies that there is sufficient statistical evidence to support the claim that the variance has indeed increased following the equipment readjustment.

---

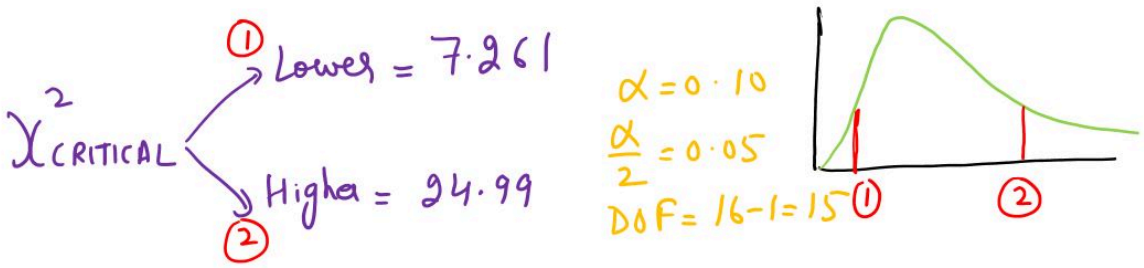
12. Test for Population Variance

The productive hours at Wipro per week were reported to have a variance of 25 hours, with an average of 50 hours. Following an organization-wide holiday of 10 days, an investigation is initiated to determine if there has been a change in the variance of productive hours per week. A sample of 16 employees was randomly selected for this analysis, for which the variance came out to be 28.06. Using a significance level of 0.10, perform a hypothesis test to assess whether the variance in productive hours per week has changed after the holidays.

The Null and Alternate Hypothesis are written below:

$$H_0: \sigma^2 = 25$$
$$H_a: \sigma^2 \neq 25$$

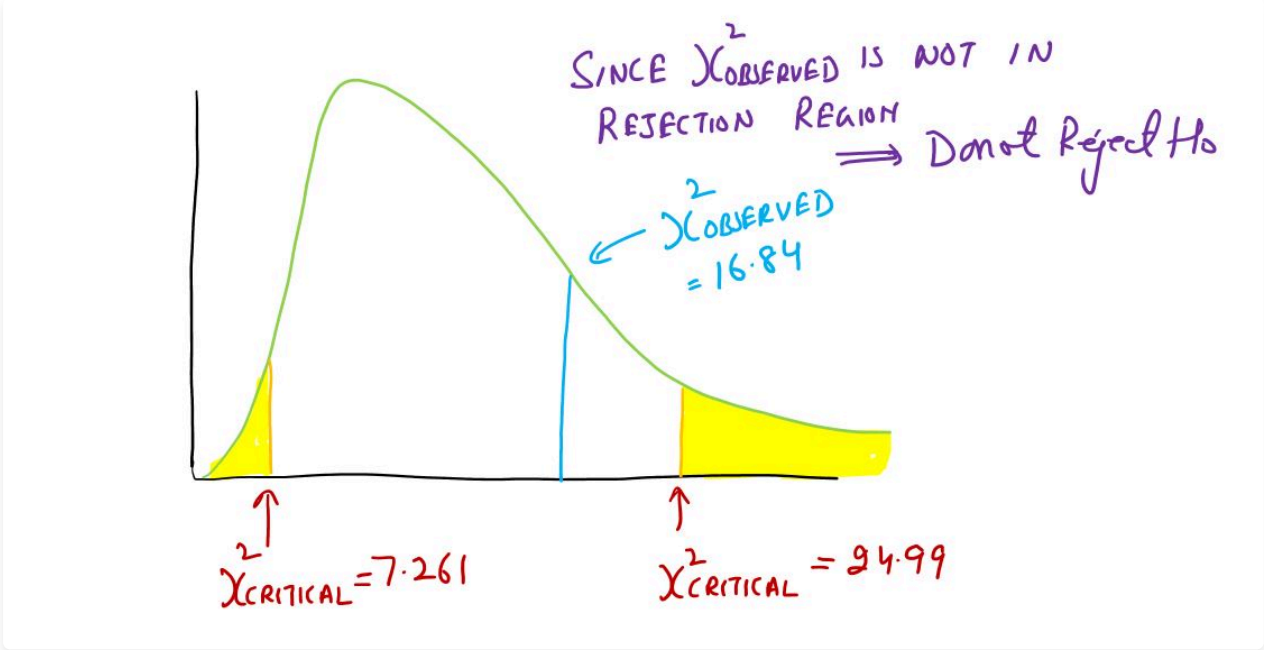
From Degree of Freedom (n-1) = 16 and \(\alpha \)= 0.01, we calculate critical value from chi square table. We will have two values, since it is two-tailed test.



Then we calculated observed value of Chi Square, using below formula.

$$\chi^2_{OBSERVED} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{15 \times 28.06}{25} = 16.84$$

Since Observed Chi Sqaure is > Critical Chi Square, we do not reject null hypothesis.



CONCLUSION

Failing to reject the null hypothesis indicates that, at the chosen significance level of 0.10, there is not enough evidence to conclude that the variance in productive hours per week has significantly deviated from the previously reported variance of 25 hours. Thus, it suggests that the holidays might not have had a substantial impact on altering the variance in productive hours at Wipro.

---

12. Test for Population Variance

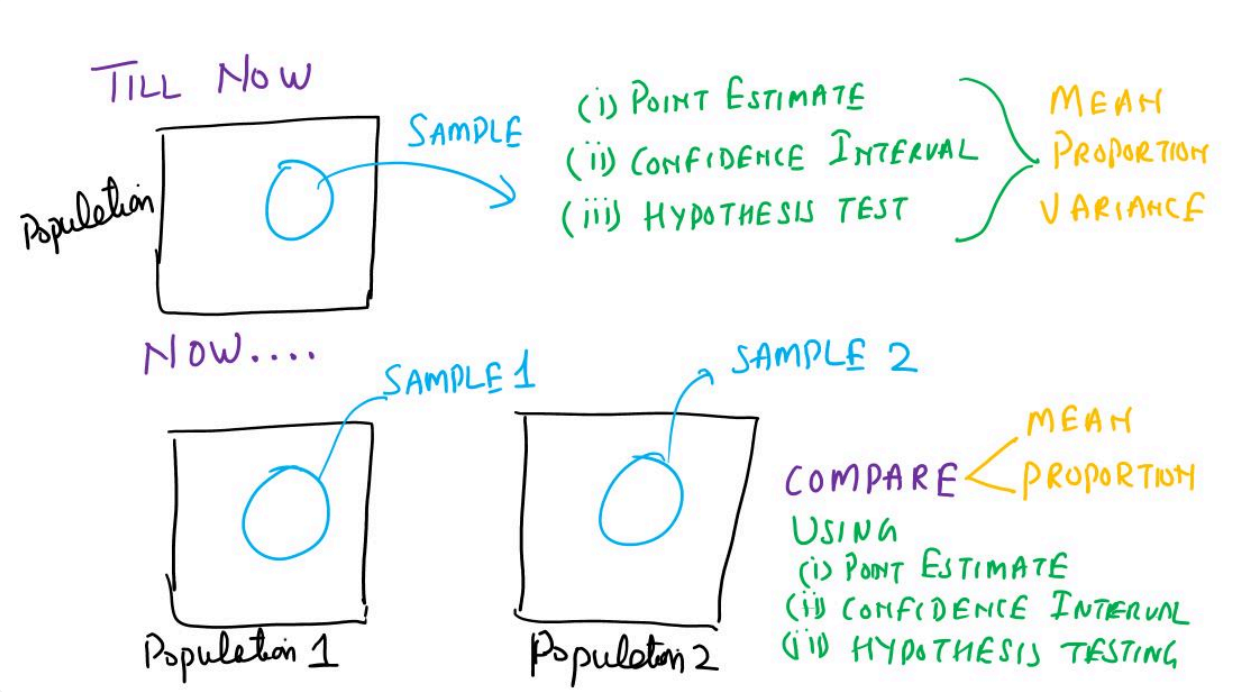
The summary of all hypothesis tests for Population Variance (using chi square) is given below:

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypothesis	$H_0: \sigma^2 \geq \sigma_0^2$ $H_a: \sigma^2 < \sigma_0^2$	$H_0: \sigma^2 \leq \sigma_0^2$ $H_a: \sigma^2 > \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$ $H_a: \sigma^2 \neq \sigma_0^2$
Test Statistics	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$
Rejection Rule: p-Value Approach	Reject H_0 if p-value ≤ α	Reject H_0 if p-value ≤ α	Reject H_0 if p-value ≤ α
Rejection Rule: Critical Value Approach	Reject H_0 if $\chi^2 \leq \chi^2_{(1-\alpha)}$	Reject H_0 if $\chi^2 \geq \chi^2_{\alpha}$	Reject H_0 if $\chi^2 \leq \chi^2_{(1-\frac{\alpha}{2})}$ or if $\chi^2 \geq \chi^2_{\frac{\alpha}{2}}$

---

1. Introduction

So far, we have understood the methods for creating point estimates, interval estimates, and conducting hypothesis tests for Mean, Proportion, and Variance concerning a **single population**.

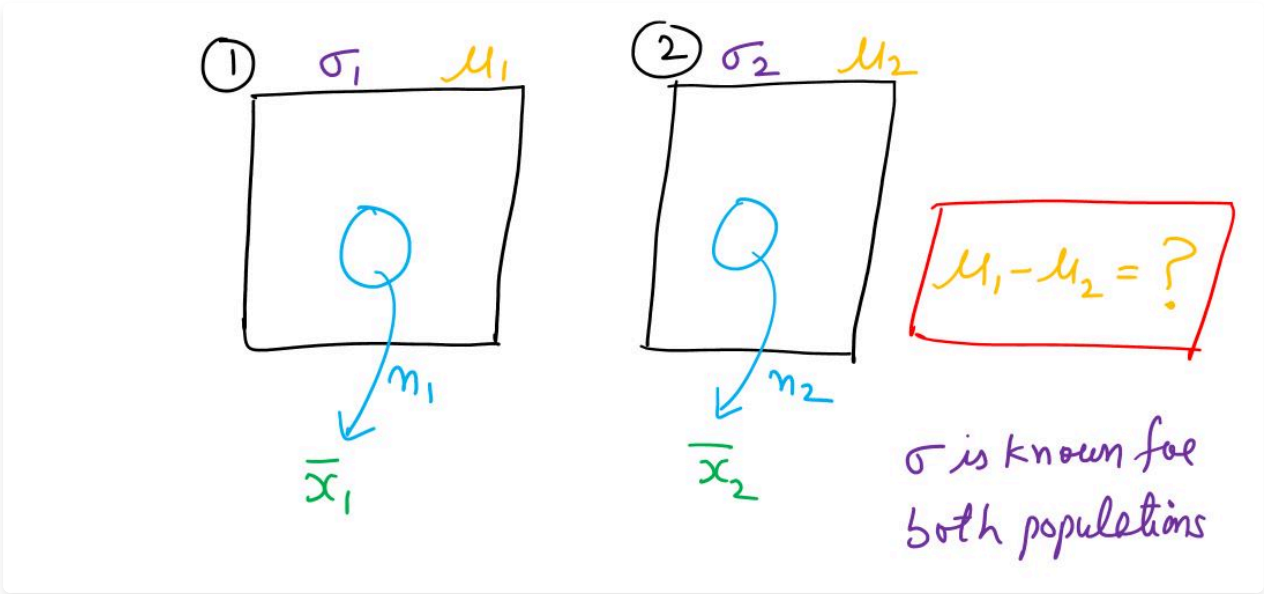


Now, let us delve into understanding how to derive point estimates, interval estimates, and hypothesis tests for scenarios involving **two populations**. This examination involves taking samples from each of these two populations to facilitate the analysis.

---

2. Comparing Means of two populations (σ known)

We will start with the comparing mean of two independent populations (\(\sigma \)), when standard deviation of population is known to us.



The Delhi Public School (DPS) group is running two schools, one at Delhi and other at Noida. The management wants to evaluate differences in education quality between schools. So, a standardized examination is given to few students. The difference between the mean examination scores is used to assess quality differences between two schools. Let Delhi school is represented as 1 and Noida school is represented as 2.

μ_1 = mean of population 1 (Delhi school)
μ_2 = mean of population 2 (Noida school)

Our objective is to find the difference between the means: $\mu_1 - \mu_2$

The standard deviation of two populations, σ_1 and σ_2, are known to us.

To make an inference about this difference ($\mu_1 - \mu_2$), we select a simple random sample of n_1 students from population 1 and a second simple random sample of n_2 students from population 2. The two samples, taken separately and independently, are referred to as independent simple random samples.

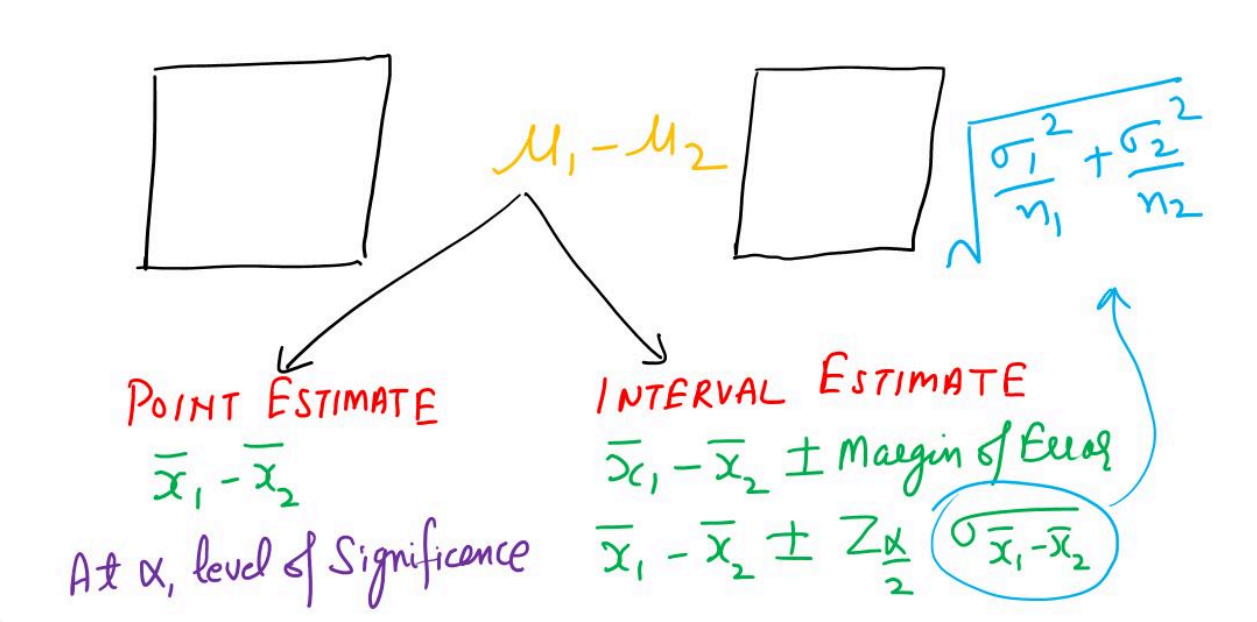
In this case (σ known), it is assumed that if both populations have a normal distribution.

If not, then the sample sizes should be large enough that the central limit theorem enables us to conclude that the sampling distributions of \(\bar x_1 \) and \(\bar x_2 \) can be approximated by a normal distribution.

---

2. Comparing Means of two populations (σ known)

Let us start with Point Estimation and Interval Estimation. Then we will move to Hypothesis Testing.



We then compute the two sample means.
$\bar{x}_1$ = sample mean score for the simple random sample of n_1 students at Delhi school
$\bar{x}_2$ = sample mean score for the simple random sample of n_2 students at Noida school

Point Estimate of the difference between two population means = $\bar{x}_1$ - $\bar{x}_2$

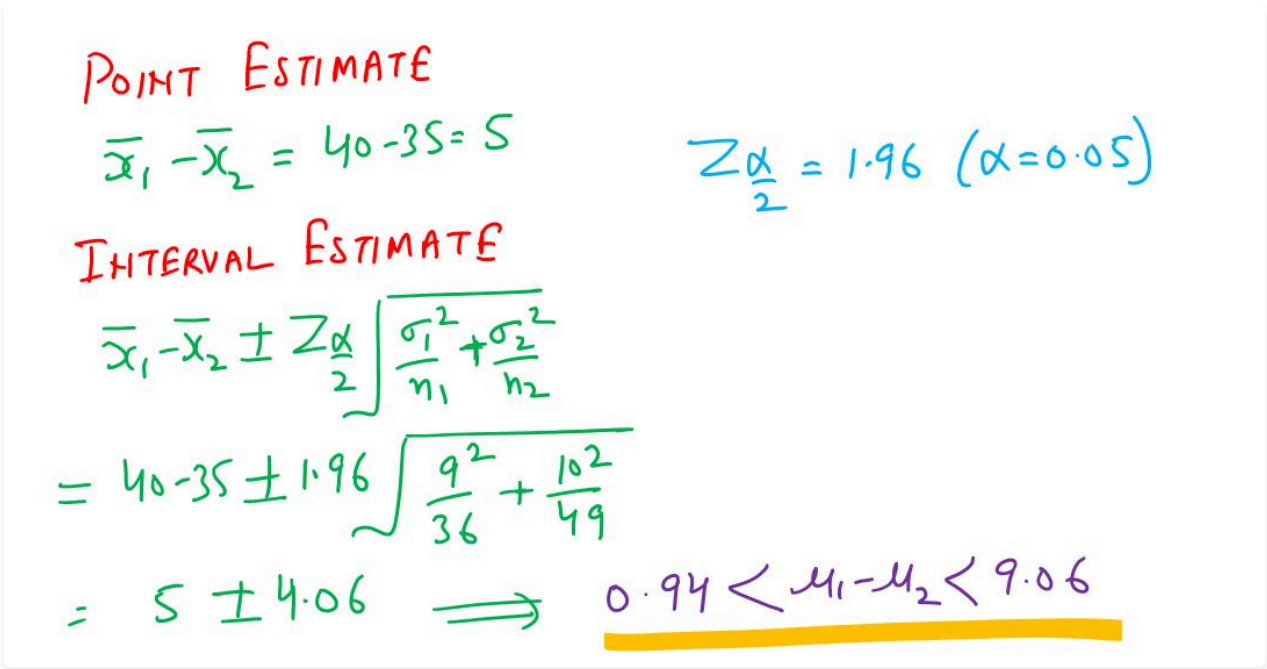
Standard Error of this pointer estimate is given by:
$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Thus, the **Interval Estimate** will take the following form:
$\bar{x}_1$ - $\bar{x}_2$ ± Margin of Error

With two independent simple random samples, the point estimator $\bar{x}_1$ - $\bar{x}_2$ has a standard error $\sigma_{\bar{x}_1 - \bar{x}_2}$ and when the sample sizes are large enough, the distribution of $\bar{x}_1$ - $\bar{x}_2$ can be described by a normal distribution.

With the sampling distribution of $\bar{x}_1$ - $\bar{x}_2$ having a normal distribution, we can write the margin of error as follows:
Margin of error = $z_{\alpha/2}\, \sigma_{\bar{x}_1 - \bar{x}_2} = z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
where (1 – σ) is the confidence coefficient.

The interval estimate of the difference between two population means is re-written as:
$\bar{x}_1$ - $\bar{x}_2$ ± $z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$



Let us return to our example of DPS schools. Based on data from previous exams, the two population standard deviations are known with $\sigma_1 = 9$ and $\sigma_2 = 10$. The data collected from the two independent simple random samples of two schools provided the following results.

	Delhi School	**Noida School**
Sample Size	$n_1 = 36$	$n_2 = 49$
Sample Mean	$\bar{x}_1 = 40$	$\bar{x}_2 = 35$

The point estimate of the difference between the mean scores of the two populations = $\bar{x}_1$ - $\bar{x}_2$ = 40 - 35 = 5.

Thus, we estimate that the students at the Delhi school have a mean score, 5 greater than, the mean score of the Noida school.

Using 95% confidence and $z_{\alpha/2} = z_{0.25} = 1.96$.

Point estimate with margin of error = $\bar{x}_1$ - $\bar{x}_2$ ± $z_{\alpha/2}$ $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

= 40 – 35 ± 1.96 $\sqrt{\frac{9^2}{36} + \frac{10^2}{49}}$ = 5 ± 4.06

Margin of error = 4.06

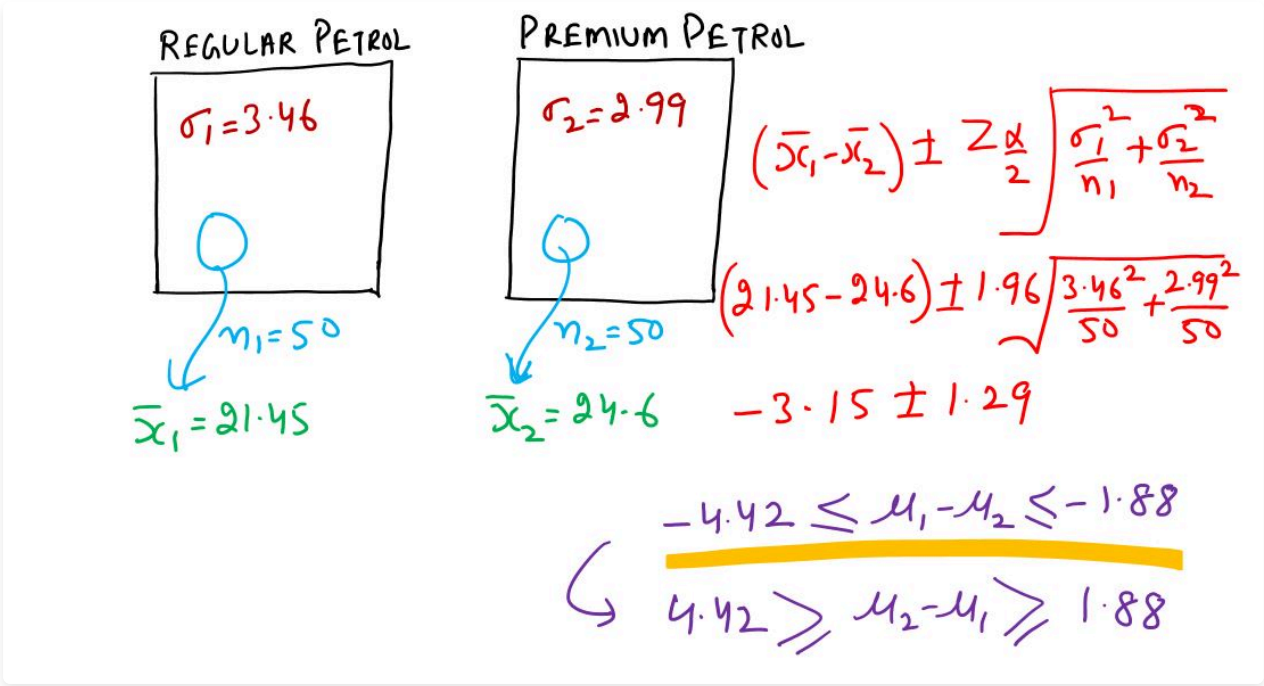
With 95% confidence, the interval estimate of the difference between the two population means is 5 - 4.06 = 0.94 years to 5 + 4.06 = 9.06 years.

---

2. Comparing Means of two populations (σ known)

CarDekho portal wants to determine the difference in mileage of cars using regular petrol and cars using premium petrol. Researchers for the group divided a fleet of 100 cars of the same make in half and tested each car. Fifty of the cars were filled with regular petrol and 50 were filled with premium petrol. The sample average for the regular petrol group was 21.45 Km per litre, and the sample average for the premium petrol was 24.6 Km per litre. Assume that the population standard deviation of the regular petrol population is 3.46 Km per litre, and that the population standard deviation of the premium petrol population is 2.99 Km per litre. Construct a 95% confidence interval to estimate the difference in the mean mileage between the cars using regular petrol and the cars using premium petrol.

Solution:



Interpreting this interval, it implies that, with 95% confidence, the true difference in mean mileage between cars using regular petrol and those using premium petrol falls somewhere within this range. Specifically, the negative values suggest that, on average, cars using premium petrol have a higher mileage (around 1.88 to 4.42 Km per litre higher) compared to cars using regular petrol.

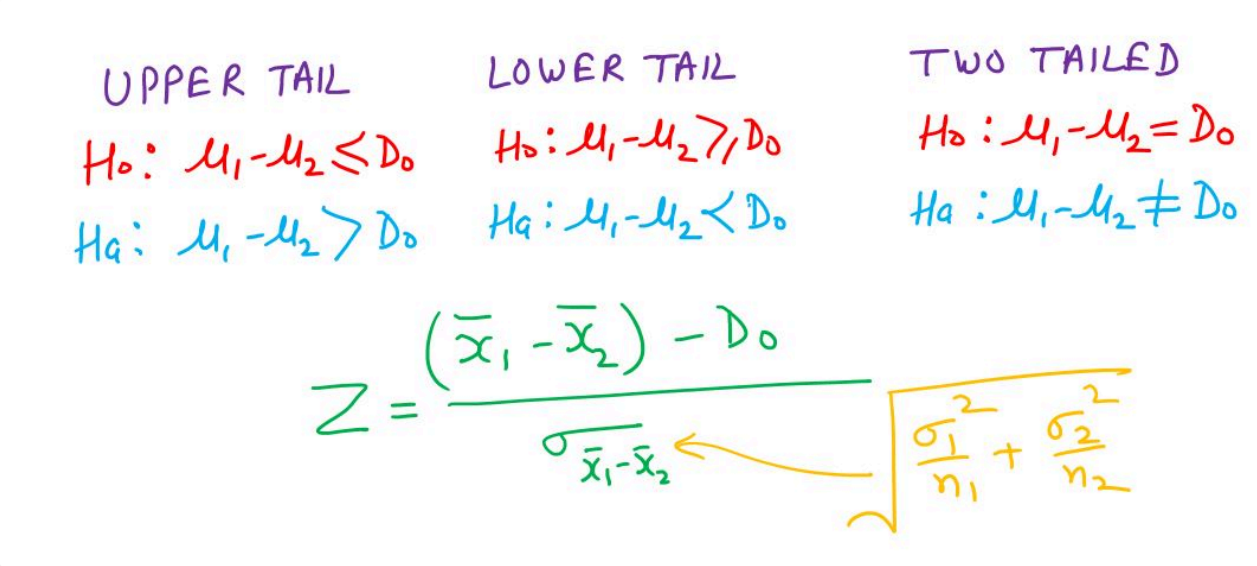
---

2. Comparing Means of two populations (σ known)

Hypothesis is a nothing, but a claim about parameter of population. In this case, we make a hypothesis about the difference of the mean of population (μ_1 - μ_2).

Let D_0 be the hypothesized difference between μ_1 and μ_2 .

The three forms for a hypothesis test are as follows:



Next, we choose a level of significance, compute the value of the test statistic z.

The test statistic for the difference between two population means, when σ_1 and σ_2 are known, is as follows:
z = \(\frac{{\bar{x}_1 - \bar{x}_2 - D_0}}{{\sqrt{\frac{{\sigma_1^2}}{{n_1}} + \frac{{\sigma_2^2}}{{n_2}}}}} \)

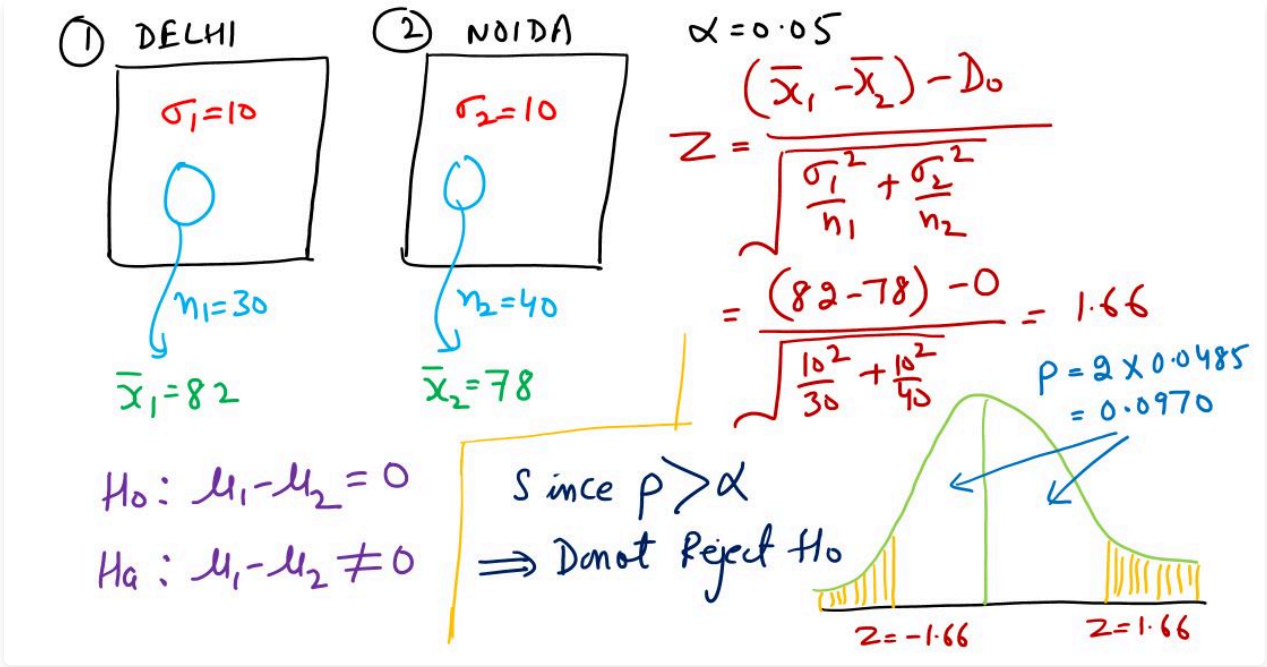
Finally we use p-value to determine whether the null hypothesis should be rejected. Null Hypothesis is rejected of p-value is less than or equal to \(\alpha \).

Let us demonstrate all these steps in our DPS schools example.

We begin with the tentative assumption that no difference exists between the education quality provided at the two schools. Hence, in terms of the mean examination scores, the null hypothesis is that μ_1 - μ_2 = 0.

The null and alternative hypotheses for this two-tailed test are written as follows:
H_0 : μ_1 - μ_2 = 0
H_a : μ_1 - μ_2 ≠ 0

We are given that the population standard deviations are known with σ_1 = 10 and σ_2 = 10. An α = 0.05 level of significance is specified for the study. Independent simple random samples of n_1 = 30 students from Delhi school and n_2 = 40 students from Noida school are taken. The respective sample means are \(\bar x_1 \) = 82 and \(\bar x_2 \) = 78.



We will compute the test statistic using the below equation:
z = \(\frac{{\bar{x}_1 - \bar{x}_2 - D_0}}{{\sqrt{\frac{{\sigma_1^2}}{{n_1}} + \frac{{\sigma_2^2}}{{n_2}}}}} \) = \(\frac{{82 - 78 - 0}}{{\sqrt{\frac{{10^2}}{{30}} + \frac{{10^2}}{{40}}}}} \) = 1.66

Next let us compute the p-value for this two-tailed test. Because the test statistic z is in the upper tail, we first compute the area under the curve to the right of z = 1.66. Using the standard normal distribution table, the area to the left of z = 1.66 is 0.9515. Thus, the area in the upper tail of the distribution is 1.0000 - 0.9515 = 0.0485. Because this test is a two tailed test, we must double the tail area: p-value = 2 × (0.0485) = 0.0970.

Following the usual rule to reject H_0 if p-value ≤ α, we see that the p-value of 0.0970 does not allow us to reject H_0 at the 0.05 level of significance.

Thus, we conclude that the sample results do not provide sufficient evidence to conclude the schools differ in quality.

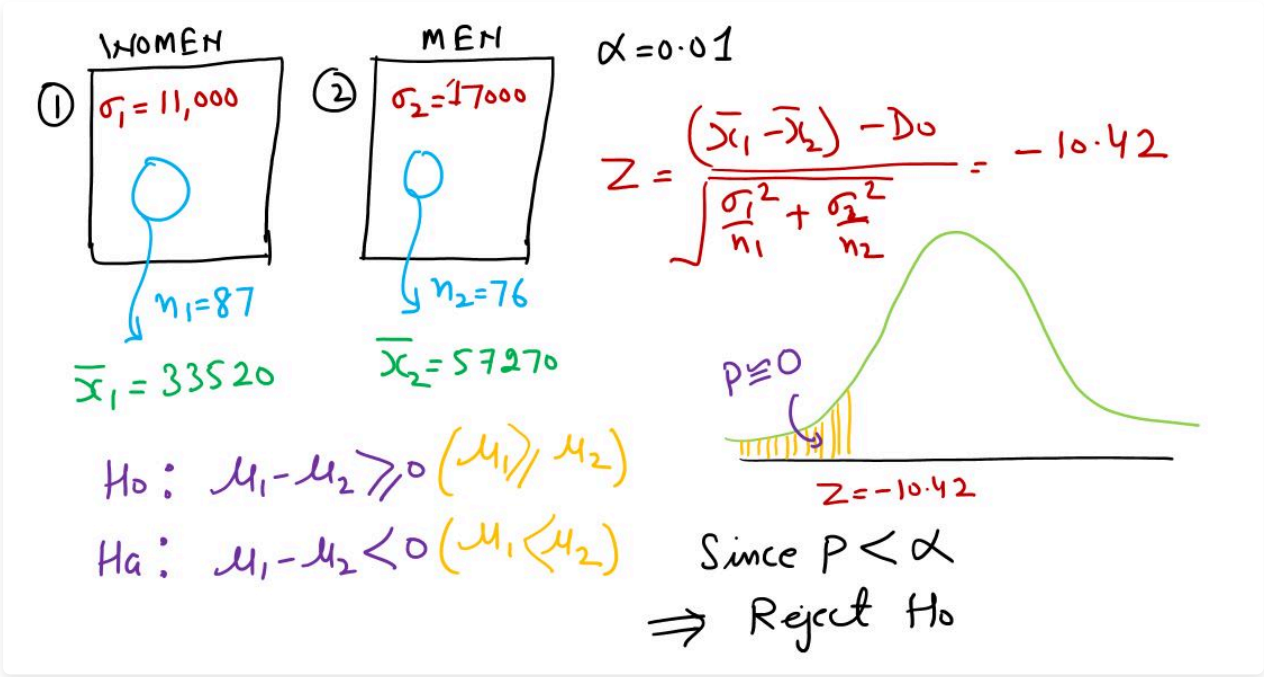
If you prefer, the test statistic and the critical value rejection rule may be used. With α = .05 and $z_{\alpha/2} = z_{0.25} = 1.96$, the rejection rule employing the critical value approach would be reject H_0 if z ≤ -1.96 or if z ≥ 1.96.

With z = 1.66, we reach the same conclusion "do not reject H_0".

2. Comparing Means of two populations (σ known)

At TCS, a sample of 87 professional working women showed that the average amount paid annually is Rs 33,520. The population standard deviation is Rs 11,000. A sample of 76 professional working men showed that the average amount paid annually is Rs 57,270, with a population standard deviation of Rs 17,000. A women's activist group wants to "prove" that women do not get paid as much per year as men. Use the hypothesis-testing process to verify this using α to be 0.01.

Solution:

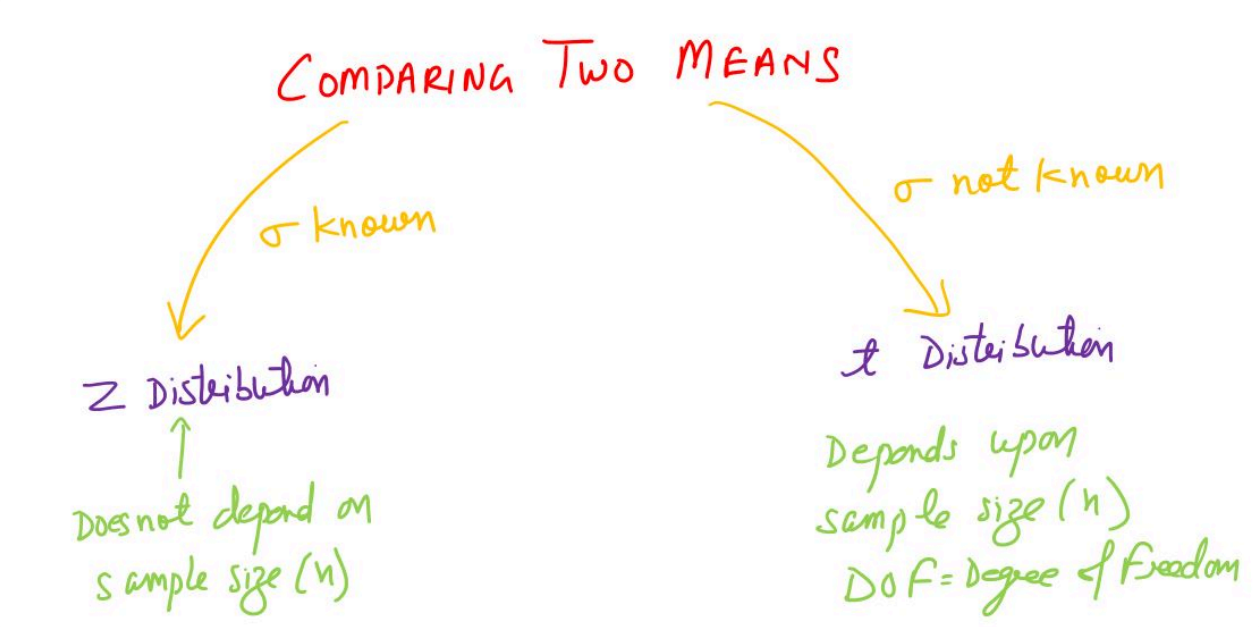


Rejecting the null hypothesis in this context implies that there is sufficient statistical evidence to support the claim made by the women's activist group that women, on average, do not receive the same annual pay as men in the professional workforce at TCS. In other words, the data suggests that there is a significant difference in the annual pay received by women compared to men within the sampled population.

---

3. Comparing Means of two populations (σ not known)

Till now, we have learnt about inferences about the difference between two population means, in case when the two population standard deviations, σ_1 and σ_2, are known to us.



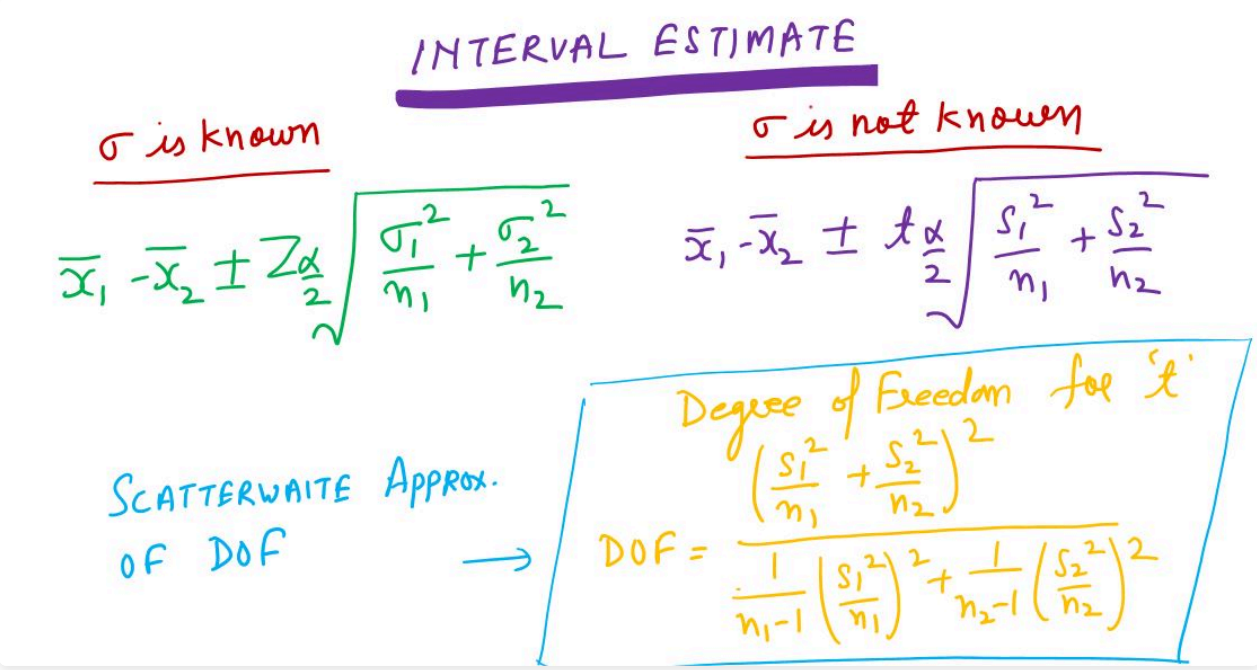
Let us now discuss about the inferences about the difference between two population means to the case when the two population standard deviations, **σ_1 and σ_2, are not known.** In this case, we will use the sample standard deviations, s_1 and s_2, to estimate the unknown population standard deviations.

When we use the sample standard deviations, the interval estimation and hypothesis testing procedures will be based on the **t distribution** rather than the standard normal distribution.

---

3. Comparing Means of two populations (σ not known)

Let us start with Point Estimation and Interval Estimation. Then we will move to Hypothesis Testing.



Point Estimate of the difference between two population means = \(\bar x_1 \) - \(\bar x_2 \)

Since σ_1 and σ_2 unknown, we will use the sample standard deviations s_1 and s_2 to estimate σ_1 and σ_2 and replace $z_{\alpha/2}$ with $t_{\alpha/2}$. As a result, the **interval estimate** of the difference between two population means is given by the following expression:
\(\bar x_1 \) - \(\bar x_2 \) ± $t_{\alpha/2}$ \(\sqrt{ \frac{ s_1^2} {n_1} + \frac{ s_2^2 }{n_2}} \)
where (1 – α) is the confidence coefficient.

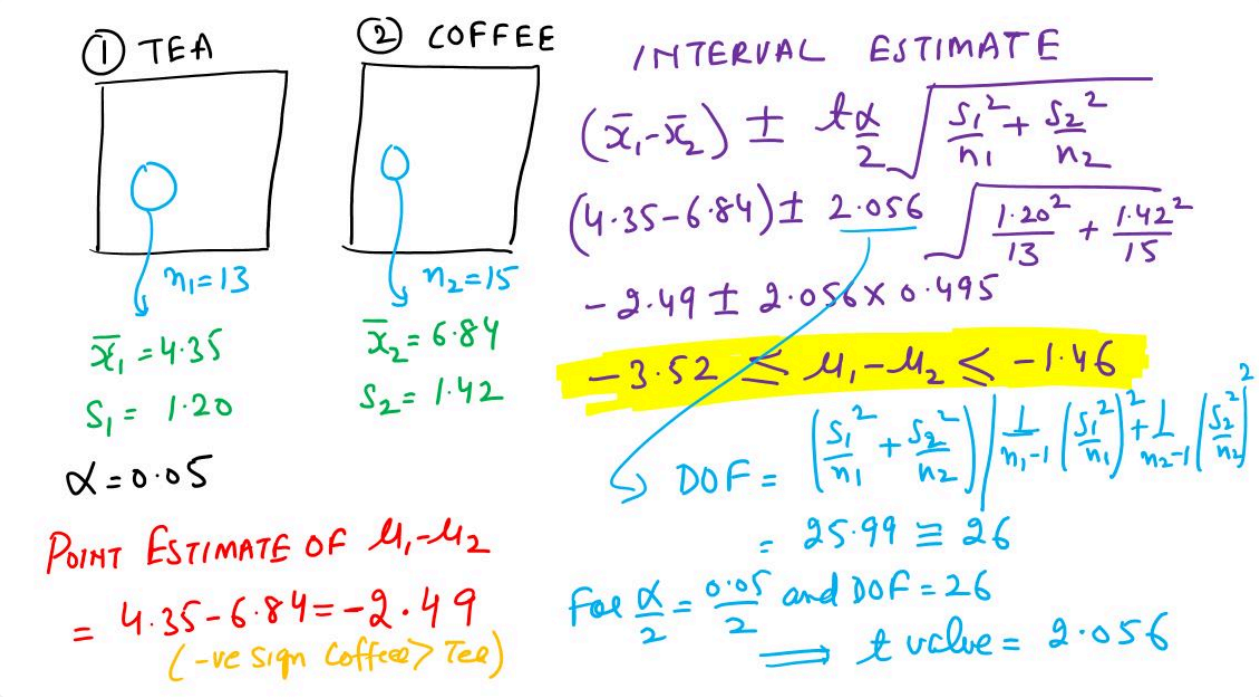
When we use **t distribution**, we need to calculate the degrees of freedom, using following formula:
Degrees of freedom, df = \(\frac{{\left(\frac{{s_1^2}}{{n_1}} + \frac{{s_2^2}}{{n_2}}\right)^2}}{{\frac{1}{{n_1-1}} \left(\frac{{s_1^2}}{{n_1}}\right)^2 + \frac{1}{{n_2-1}} \left(\frac{{s_2^2}}{{n_2}}\right)^2}} \)

---

3. Comparing Means of two populations (σ not known)

A coffee manufacturer is interested in estimating the difference in the average daily consumption of tea drinkers and coffee drinkers. Its researcher randomly selects 13 tea drinkers and asks how many cups of tea per day they drink. He randomly locates 15 coffee drinkers and asks how many cups of coffee per day they drink. The average for the tea drinkers is 4.35 cups, with a standard deviation of 1.20 cups. The average for the coffee drinkers is 6.84 cups, with a standard deviation of 1.42 cups. The researcher assumes, for each population, that the daily consumption is normally distributed, and he constructs a 95% confidence interval to estimate the difference in the averages of the two populations. Also, what is point estimate of difference of population means.

Solution:



The 95% confidence interval for the difference in the average daily consumption of tea and coffee drinkers is from -3.52 to -1.46 cups.

This interval implies that, with 95% confidence, the true difference in the average daily consumption of tea and coffee drinkers lies between -3.52 to -1.46 cups. In specific terms, on average, coffee drinkers consume between approximately 1.46 to 3.52 cups more per day than tea drinkers.

---

3. Comparing Means of two populations (σ not known)

When σ_1 and σ_2 are unknown, we use s_1 as an estimator of σ_1 and s_2 as an estimator of σ_2. Test Statistics for Hypothesis tests about $\mu_1 - \mu_2$ when σ_1 and σ_2 are unknown is shown below:

t = $\frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Let us understand steps of Hypothesis Testing with an example.

Infosys is experimenting a new AI technology in development of softwares for its clients. To evaluate the benefits of the new technology, a random sample of 24 similar projects is selected. The project managers of 12 projects develop softwares with the help of current technology and remaining 12 project managers develop softwares with the help new AI technology. The completion time of these 24 projects is monitored, in days.

μ_1 = mean project completion time using the current technology
μ_2 = mean project completion time using the new AI technology.

Thus, the Infosys is looking for evidence to conclude that μ_2 is less than μ_1; in this case, the difference between the two population means, $\mu_1 - \mu_2$, will be greater than zero. The research hypothesis $\mu_1 - \mu_2 > 0$ is stated as the alternative hypothesis. Thus, the hypothesis test becomes:
$H_0 : \mu_1 - \mu_2 \leq 0$
$H_a : \mu_1 - \mu_2 > 0$

We will use $\alpha = 0.05$ as the level of significance.

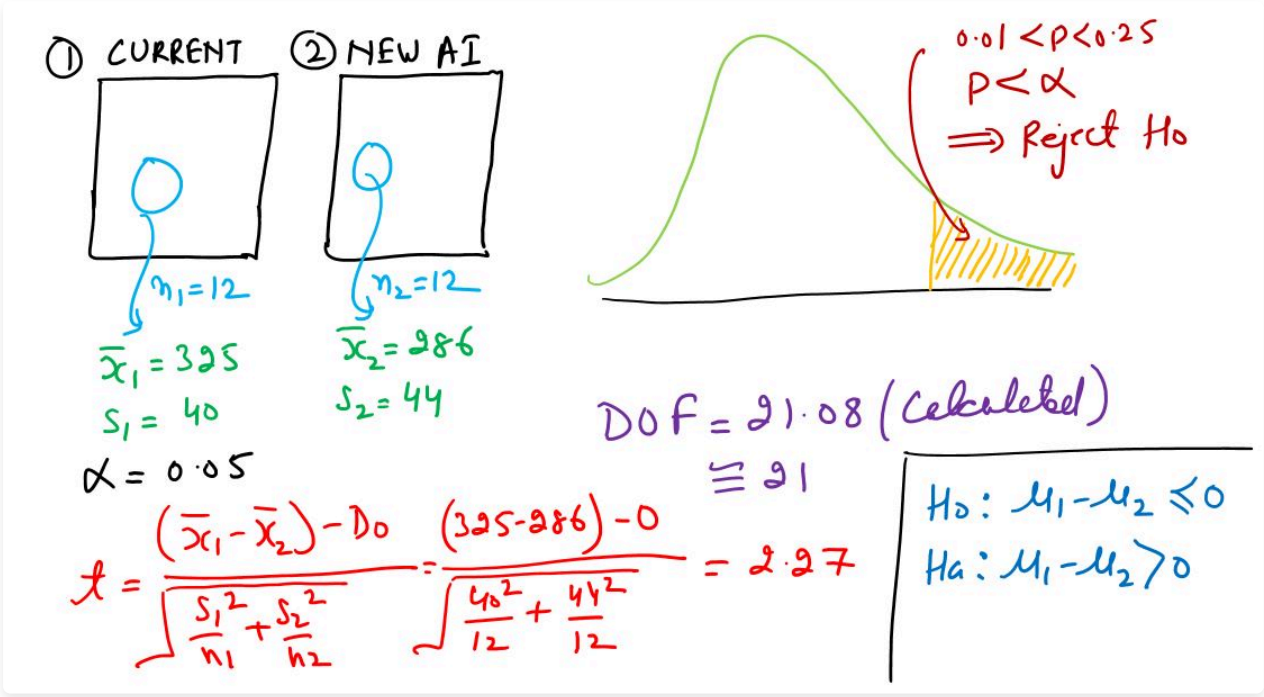
Suppose that 24 project managers complete the study with the results shown below:

Summary Statistics	Current Technology	New Technology
	300	274
	280	220
	344	308
	385	336
	372	198
	360	300
	288	315
	321	258
	376	318
	290	310
	301	332
	283	263
Sample Size	$n_1 = 12$	$n_2 = 12$
Sample Mean	$\bar{x_1} = 325$	$\bar{x_2} = 286$
Sample Standard Deviation	$s_1 = 40$	$s_2 = 44$

Test statistic, t = $\frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(325 - 286) - 0}{\sqrt{\frac{40^2}{12} + \frac{44^2}{12}}} = 2.27$

Degrees of freedom:
df = $\frac{\left(\frac{40^2}{12} + \frac{44^2}{12}\right)^2}{\frac{1}{12-1} \left(\frac{40^2}{12}\right)^2 + \frac{1}{12-1} \left(\frac{44^2}{12}\right)^2} = 21.8$

Rounding down, we will use a t- distribution with 21 degrees of freedom.



With an upper tail test, the p-value is the area in the upper tail to the right of t = 2.27.

From the above results, we see that the p-value is between .025 and .01. Thus, the p-value is less than $\alpha = 0.05$ and H_0 is rejected. The sample results conclude that $\mu_1 - \mu_2$, will be greater than zero or $\mu_1 > \mu_2$.

Thus, the study supports the conclusion that the new AI technology provides a smaller population mean project completion time. Thus, Infosys should consider using new AI technology for development of software.

---

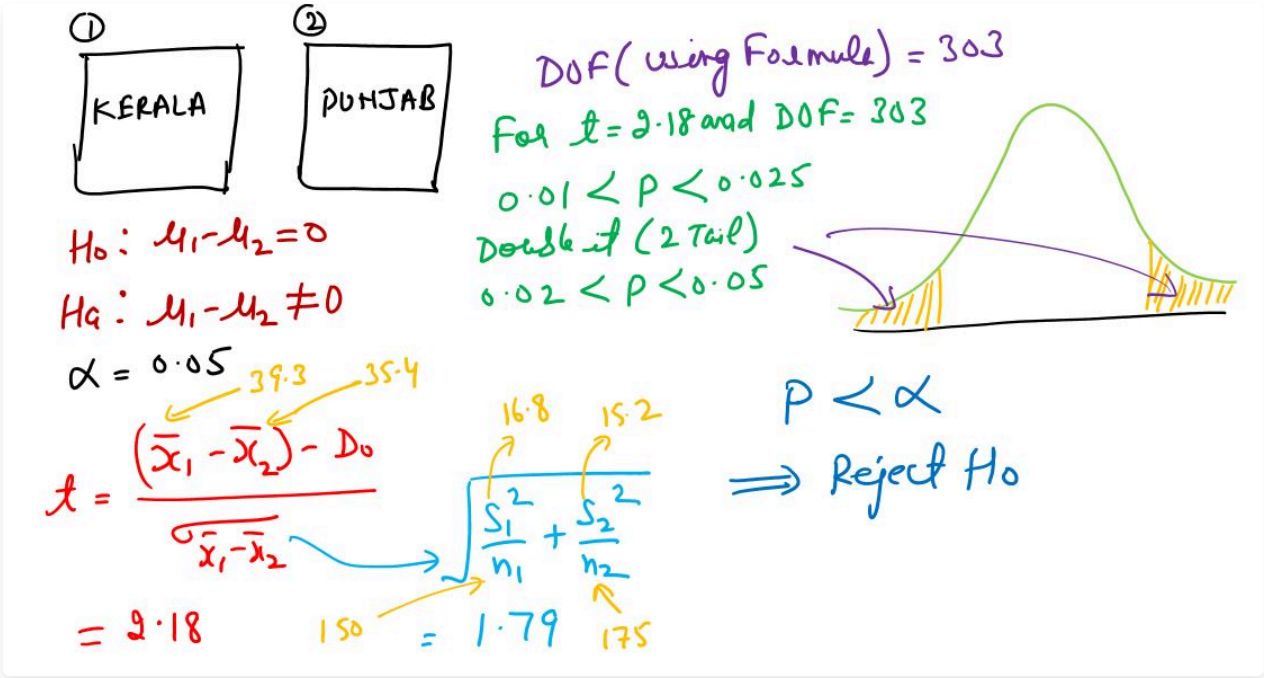
3. Comparing Means of two populations (σ not known)

Data were collected on the mean age of the residents living in the Kerala as well as on residents living in Punjab. The following results were obtained from the samples. Test the hypothesis of no difference between the two population means. Use α = 0.05.

KERALA
$n_1 = 150$
$\bar{x}_1 = 39.3$
$s_1 = 16.8$

PUNJAB
$n_2 = 175$
$\bar{x}_2 = 35.4$
$s_2 = 15.2$

Solution:



Because the null hypothesis for the test comparing the mean ages of residents in Kerala and Punjab was rejected, it implies that there is evidence to suggest a significant difference between the mean ages of the two populations. In other words, the data provide enough statistical support to conclude that the mean ages of residents in Kerala and Punjab are not equal.

4. Comparing Means with Pooled t Test

In the case of the Pooled t-test, the comparison involves two separate groups derived from the same population.

This test is applicable when the standard deviation of the population is unknown, but it's known that the standard deviations of the two groups are identical, akin to the standard deviation of the population.

Here are several scenarios that illustrate the utility of the "Comparing Means with Pooled t Test":

Medical Research: Analyzing the efficacy of distinct drugs in treating a shared medical condition. For instance, assessing the average recovery time of patients administered Drug A versus those given Drug B.

Educational Studies: Contrasting the mean scores of students exposed to different pedagogical methods to determine the approach that yields superior learning outcomes.

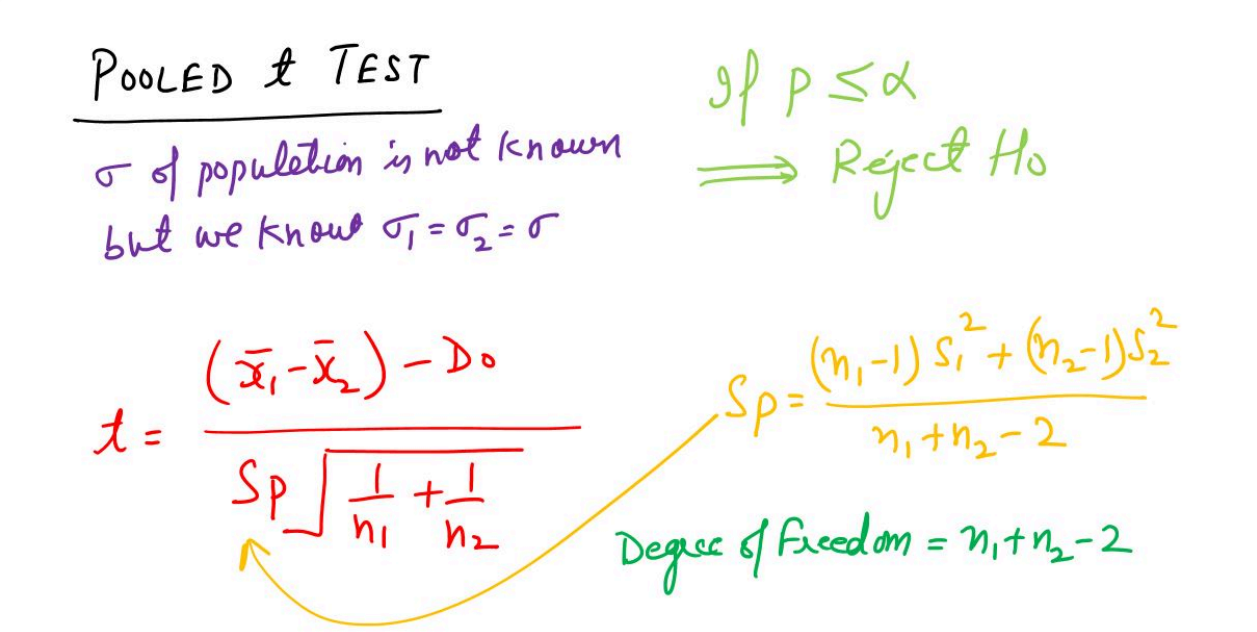
Market Research: Examining the average satisfaction ratings for two distinct products or services offered by a company to ascertain the better-performing option in terms of customer satisfaction.

Employee Studies: Evaluating the average productivity levels of employees engaged in different departments or working under various management styles to discern potential performance variations.

In these scenarios, employing the pooled t-test aids in gauging whether there exists a statistically noteworthy distinction between the means of these independent groups.

Since Standard Deviation of population is not known, we use t distribution.

The test statistic (t) is given by following formula.



The Degree of freedom is n_1+n_2-2.

It is assumed that the population, from which two gropus are derived, is normally disttrubution, or there is enough sample size (more than 30) for CLT to apply.

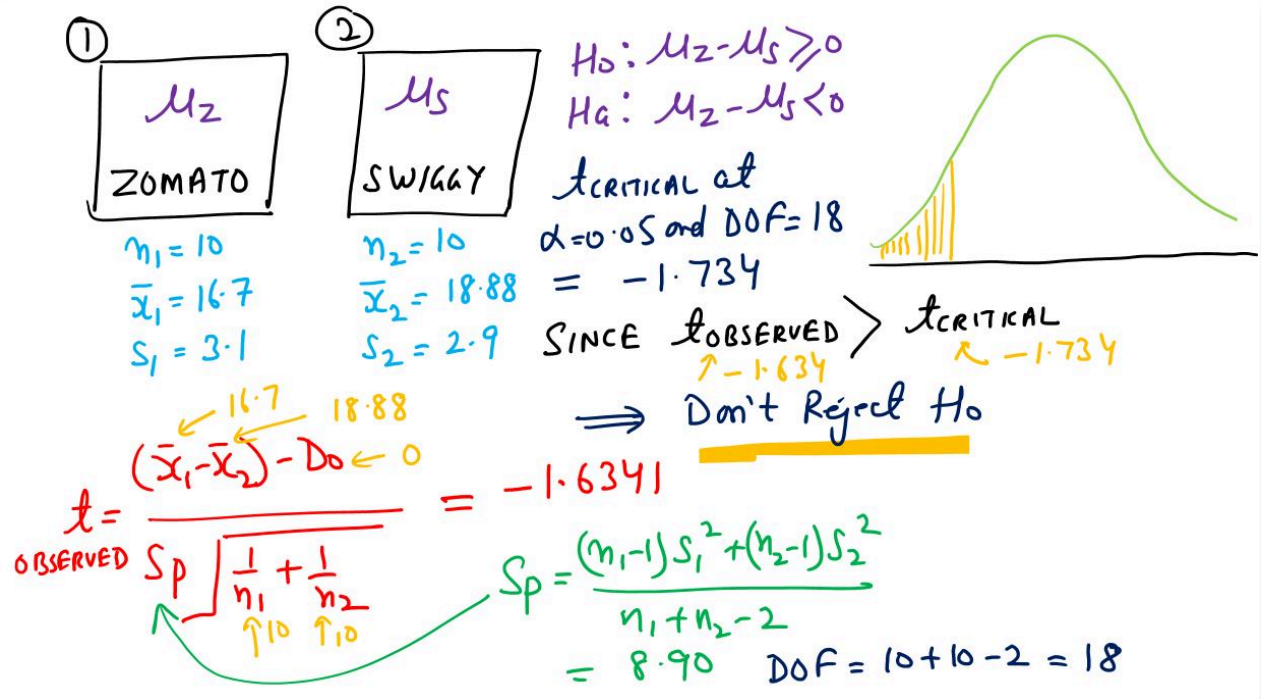
---

4. Comparing Means with Pooled t Test

You and some friends have decided to test the validity of an advertisement by Zomato, which says it delivers to the homes faster than Swiggy. You collect the data by ordering 10 pizzas from Zomato and 10 pizzas from Swiggy. The table shows the delivery times. At the 0.05 level of significance, is there evidence that the mean delivery time for Zomato is less than the mean delivery time for Zomato? Assume that the variance is same in both populations.

ZOMATO		SWIGGY	
16.8	18.1	22.0	19.5
11.7	14.1	15.2	17.0
15.6	21.8	18.7	19.5
16.7	13.9	15.6	16.5
17.5	20.8	20.8	24.0

Solution:

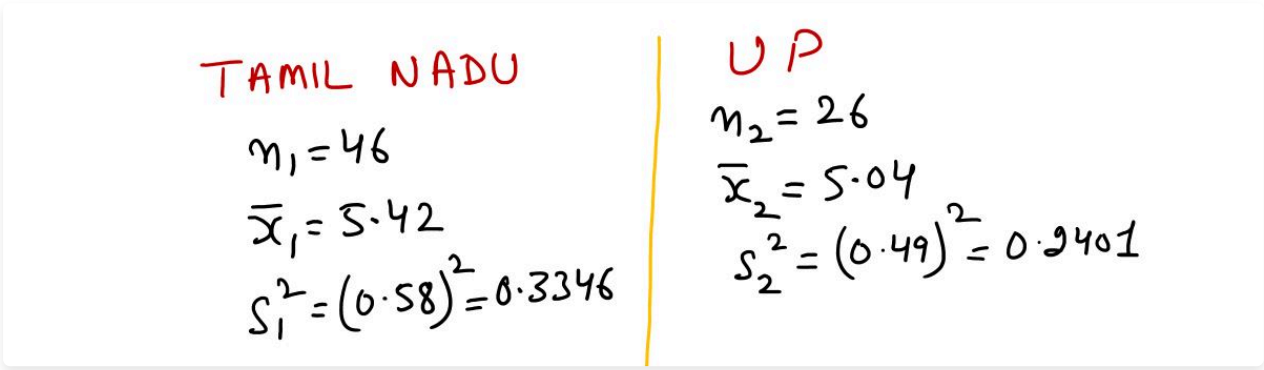


Because we concluded not to reject the null hypothesis, it implies that there isn't sufficient statistical evidence to conclude that the mean delivery time for Zomato is less than the mean delivery time for Swiggy. In other words, based on the collected data and the statistical test performed, you haven't found significant evidence to support the claim that Zomato delivers faster than Swiggy.

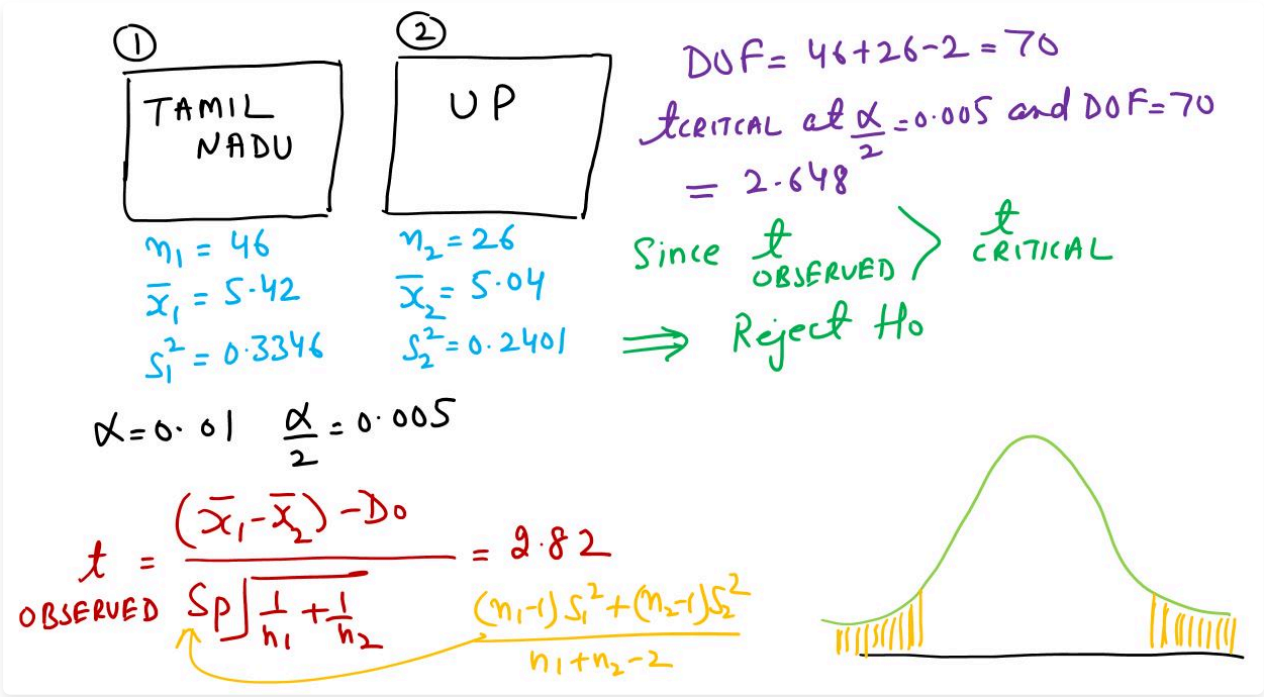
---

4. Comparing Means with Pooled t Test

NDTV wants to determine whether there is a significant difference in the liking of citizens towards current foreign policy between Tamil Nadu and UP. For the study, 46 people from Tamil Nadu and 26 people from UP were contacted and interviewed. They were asked to respond to 35 items using a 9-point scale with possible answers ranging from very bad (1) to very good (9). The resulting statistics for the two groups are shown in the table. Using $\alpha=0.01$ test to determine whether there is a significant difference between liking of citizens in Tamil Nadu and UP. Assume that liking of citizens is normally distributed in the population. Assume that the variance is same in both populations.



Solution:



Because we rejected the null hypothesis, it implies that there is sufficient statistical evidence to suggest that there is a significant difference in the liking of citizens towards the current foreign policy between Tamil Nadu and UP. In other words, the data provides support for the claim that there is a meaningful distinction in the preferences or opinions of citizens from these two regions regarding the current foreign policy.

---

5. Comparing Means using Matched Samples

In one of previous examples at Infosys, 12 project managers developed software using current technology and 12 different project managers developed software using new technology.

What if, we selected only 12 project managers in place of 24, and asked each project manager to do 2 projects, one with current technology and 1 with new technology? For this type of scenarios, we use Matched Sample Tests.

Comparing means with a Matched Sample involves examining data from the same population that undergoes two distinct experiments. Each object or person in the sample contributes two values, representing paired measurements taken under different conditions or treatments. This method, also known as the Paired t Test, acknowledges the dependency between the paired observations.

This analysis, often termed Dependent Samples, Related Populations, Matched Pair Test, or Correlated t-test, focuses on evaluating whether there's a statistically significant difference between the paired measurements. It's applicable when studying the effects of interventions, treatments, or conditions by collecting data before and after applying a change or when comparing related variables within the same group or sample.

Here are few scenarios where a comparison with a matched sample could be applied:

1. **Medical Trials:** Before and after treatment assessments on the same group of patients to evaluate the effectiveness of a new medication.

2. **Educational Studies:** Comparing students' performance in exams before and after implementing a new teaching method or curriculum.

3. **Fitness Training:** Analyzing changes in physical fitness metrics (like weight, endurance, or strength) in individuals before and after a specific workout program.

4. **Psychological Studies:** Assessing the impact of therapy by measuring stress levels in individuals before and after therapy sessions.

5. **Product Testing:** Testing the efficiency of a new manufacturing process by comparing the quality of output before and after process modifications in the same production line.

The Test Statistic is given by following formula:

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} \qquad \mu_1 - \mu_2$$

$$\bar{d} = \frac{\sum d_i}{n} \qquad s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$$

$$DOF = n - 1$$

The Interval Estimate is given by following formula:

$$\text{CONFIDENCE INTERVAL} = \bar{d} \pm t_{\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}}$$

The Degree of Freedom is n-1.

5. Comparing Means using Matched Samples

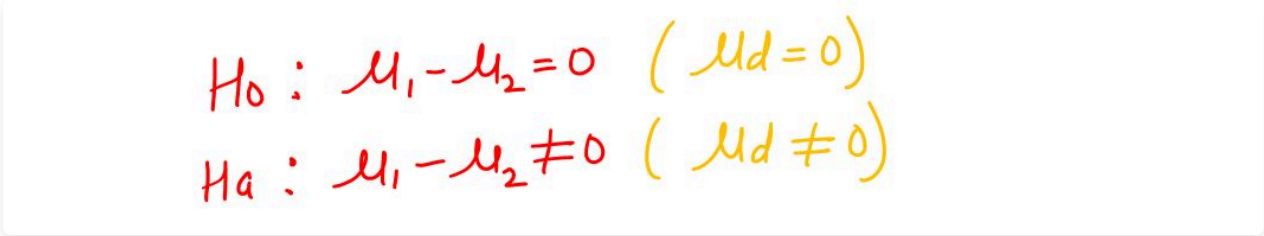
Suppose employees at a manufacturing company can use two different methods to perform a production task. To maximize production output, the company wants to identify the method with the smaller population mean completion time.

A random sample of six workers is used. The data on completion times for the six workers are given below:

Worker	Completion Time for Method 1 (minutes)	Completion Time for Method 2 (minutes)	Difference in Completion Times (d_i)
1	6.0	5.4	0.6
2	5.0	5.2	-0.2
3	7.0	6.5	0.5
4	6.2	5.9	0.3
5	6.0	6.0	0.0
6	6.4	5.8	0.6

The key to the analysis of the matched sample design is to realize that we consider only the column of differences. Therefore, we have six data values (0.6, -0.2, 0.5, 0.3, 0.0, and 0.6) that will be used to analyze the difference between population means of the two production methods.

Let μ_d= the mean of the difference in values for the population of workers. With this notation, the null and alternative hypotheses are rewritten as follows.



If H_0 is rejected, we can conclude that the population mean completion times differ. The 'd' notation is a reminder that the matched sample provides difference data.

The sample mean and sample standard deviation for the six difference values follow:

\(\bar{d} = \frac{{\sum d_i}}{{n}} = \frac{{1.8}}{{6}} \) = 0.30

s_d = \(\sqrt{\frac{{\sum (d_i - \bar{d})^2}}{{n-1}}} = \sqrt{\frac{{0.56}}{{5}}} \) = 0.335

With the small sample of n = 6 workers, we need to make the assumption that the population of differences has a normal distribution. This assumption is necessary so that we may use the t distribution for hypothesis testing and interval estimation procedures.

Based on this assumption, the following test statistic has a t distribution with n - 1 degrees of freedom.

t = \(\frac{{\bar{d} - \mu_d}}{{\frac{{s_d}}{{\sqrt{n}}}}} = \frac{{0.30 - 0}}{{\frac{{0.335}}{{\sqrt{6}}}}} \) = 2.20

Now let us compute the p-value for this two-tailed test. Because t = 2.20 ≥ 0, the test statistic is in the upper tail of the t distribution. With t = 2.20, the area in the upper tail to the right of the test statistic can be found by using the t distribution table with degrees of freedom = n - 1 = 6 - 1 = 5.

From t Table, we see that the area in the upper tail is between 0.05 and 0.025. Because this test is a two-tailed test, we double these values to conclude that the p-value is between 0.10 and 0.05. This p-value is greater than α = 0.05. Thus, the null hypothesis $H_0 : \mu_d = 0$ is not rejected.

In the matched sample design the two production methods are tested under similar conditions (i.e., with the same workers); hence this design often leads to a smaller sampling error than the independent sample design. The primary reason is that in a matched sample design, variation between workers is eliminated because the same workers are used for both production methods.

Interval Estimate

In addition, we can obtain an interval estimate of the difference between the two population means by using the single population methodology:

$\bar{d} \pm t_{0.025} \frac{{s_d}}{{\sqrt{n}}}$ = 0.3 ± 0.35

Thus, the margin of error is 0.35 and the 95% confidence interval for the difference between the population means of the two production methods is -0.05 minutes to 0.65 minutes.

---

5. Comparing Means using Matched Samples

The students were asked to rate a course at Human Peritus both before and after viewing a promotional video. The data is displayed in the table. Use an alpha of 0.05 to test to determine whether there is a significant increase in the ratings of the students after watching the video. Assume that differences in ratings are normally distributed in the population. Also construct constructs a 99% confidence interval.

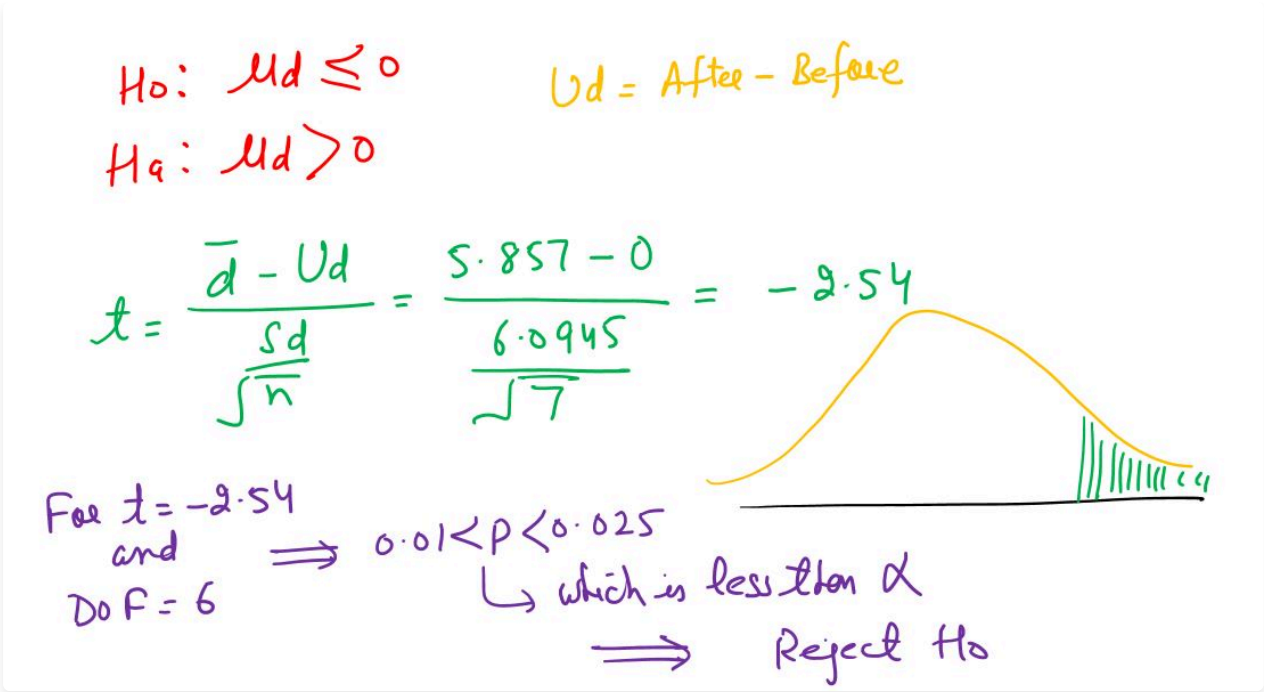
STUDENT NUMBER	BEFORE VIDEO	AFTER VIDEO
1	32	39
2	11	15
3	21	35
4	17	13
5	30	41
6	38	39
7	14	22

Solution:

STUDENT	BEFORE	AFTER	(d_i)	$(d_i-\bar{d})^2$
1	32	39	7	1.99
2	11	15	4	2.51
3	21	35	14	70.78
4	17	13	-4	91.92
5	30	41	11	29.30
6	38	39	1	21.04
7	14	22	8	5.82
			41	223.37

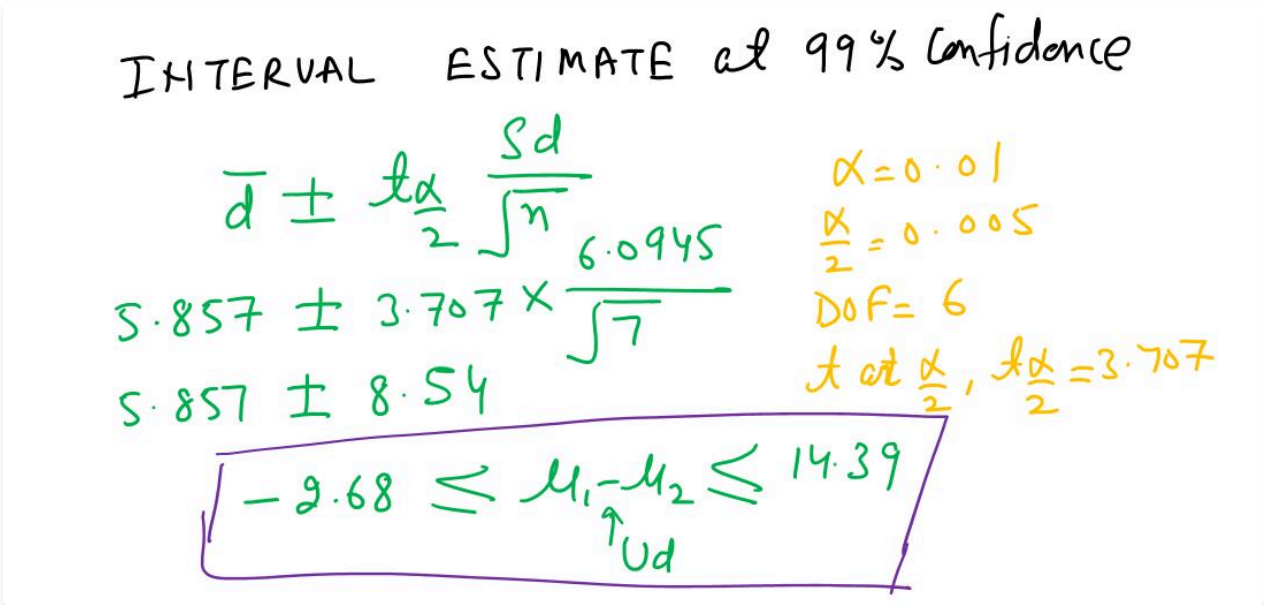
$$\bar{d} = \frac{41}{7} = 5.857$$

$$Sd = \sqrt{\frac{223.37}{7-1}} = 6.0945$$



Rejecting the null hypothesis implies that there is sufficient statistical evidence to support the conclusion that there is a significant increase in the ratings of the students after they watched the video. This suggests that the promotional video had a measurable impact on the students' perceptions or opinions about the course at Human Peritus.

Interval Estimate is calculated as belw:



---

1. Comparing Proportions of two populations

In previous classes, we have understood, how to compare means of two population.

Now, we will learn, how to compare Proportions of two populations.

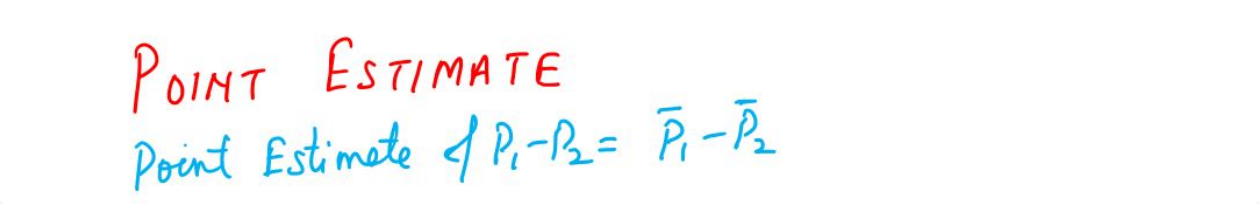
---

2. Point Estimation and Interval Estimation

Let p_1 denote the proportion for population 1 and p_2 denote the proportion for population 2, we next consider inferences about the difference between the two population proportions (p_1 - p_2). To make an inference about this difference, we will select two independent random samples consisting of n_1 units from population 1 and n_2 units from population 2.

Point Estimate of p_1 - p_2

The point estimator of the difference between two population proportions p_1 - p_2 is the difference between the sample proportions of two independent simple random samples \(\bar p_1 \) - \(\bar p_2 \).

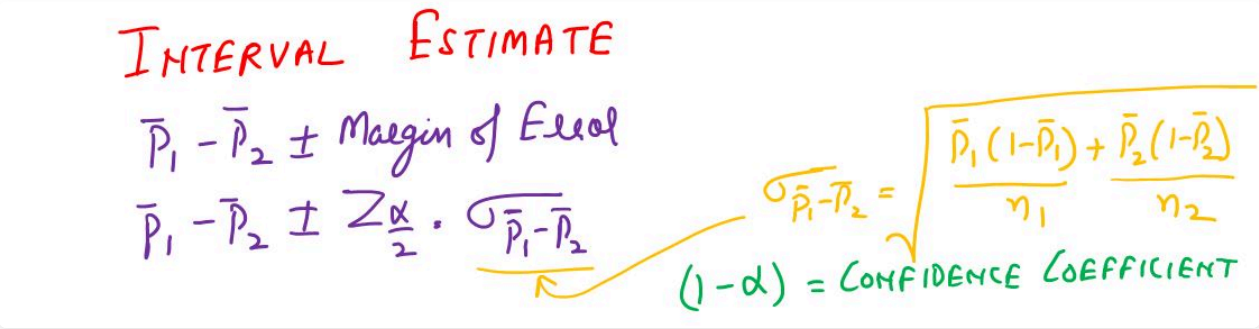


Point Estimate of (p_1 - p_2) is given by \(\bar p_1 \) - \(\bar p_2 \).

where:
p_1 = proportion of population 1
p_2 = proportion of population 2
\(\bar p_1 \) = sample proportion for a simple random sample from population 1
\(\bar p_2 \) = sample proportion for a simple random sample from population 2

Interval Estimation of p_1 - p_2

The standard error of sampling distribution of \(\bar p_1 \) - \(\bar p_2 \) is:
\(\sigma_{(\bar{p}_1 - \bar{p}_2)} = \sqrt{\frac{{p_1(1 - p_1)}}{{n_1}} + \frac{{p_2(1 - p_2)}}{{n_2}} \)



The sampling distribution of \(\bar p_1 \) - \(\bar p_2 \) will have normal distribution, when all of following conditions are met:
(i) n_1p_1 \(\geq \) 5
(ii) $n_1(1-p_1)$\(\geq \) 5
(iii) n_2p_2 \(\geq \) 5
(iv) $n_2(1-p_2)$ \(\geq \) 5

Interval estimate will take the following form:
\(\bar p_1 \) - \(\bar p_2 \) ± Margin of error
where Margin of error = \(z_{\frac{\alpha}{2}} \sqrt{\frac{{\bar{p}_1(1 - \bar{p}_1)}}{{n_1}} + \frac{{\bar{p}_2(1 - \bar{p}_2)}}{{n_2}}} \)
\)
(1 – α) is the confidence coefficient.

---

2. Point Estimation and Interval Estimation

The UIDAI has 8 regional offices all over the country. The UIDAI is interested in comparing the quality of Aadhaar enrollment at two of its regional offices, Lucknow and Ranchi. By randomly selecting samples of Aadhaar enrollment at each office and verifying the enrollment process' accuracy (in terms of faulty enrollment), the UIDAI will be able to estimate the performance of both offices.

Office	Sample Size	Number of Faulty Enrollments
Lucknow	$n_1 = 250$	35
Ranchi	$n_2 = 300$	27

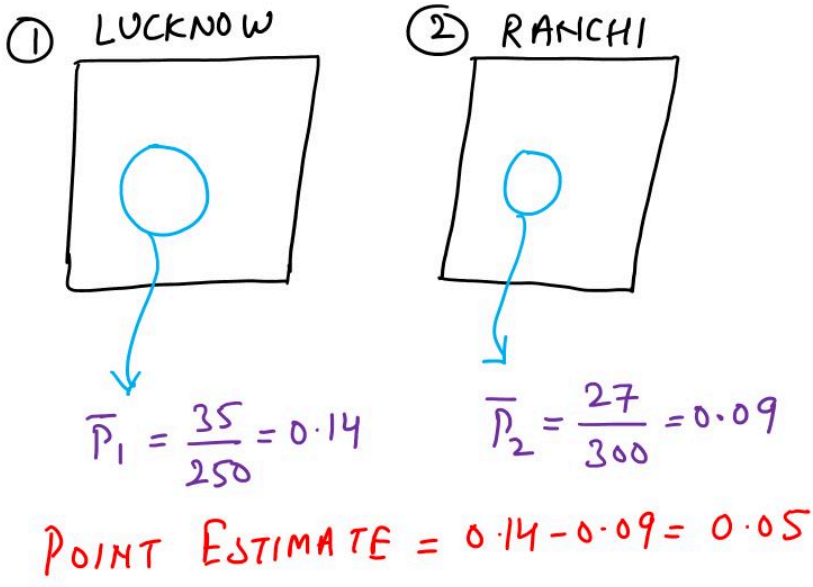
p_1 = proportion of faulty enrollment for Lucknow office, named population 1

p_2 = proportion of faulty enrollment for Ranchi office, named population 2

$\bar p_1$ = sample proportion for a simple random sample from population 1

$\bar p_2$ = sample proportion for a simple random sample from population 2

The sample proportions for the two UIDAI offices:
$\bar p_1$ = $\frac{35}{250}$ = 0.14 ,
$\bar p_2$ = $\frac{27}{300}$ = 0.09



Point estimate of the difference between the proportions of faulty enrollment for the two populations is:
$\bar p_1$ - $\bar p_2$ = 0.14 - 0.09 = 0.05.

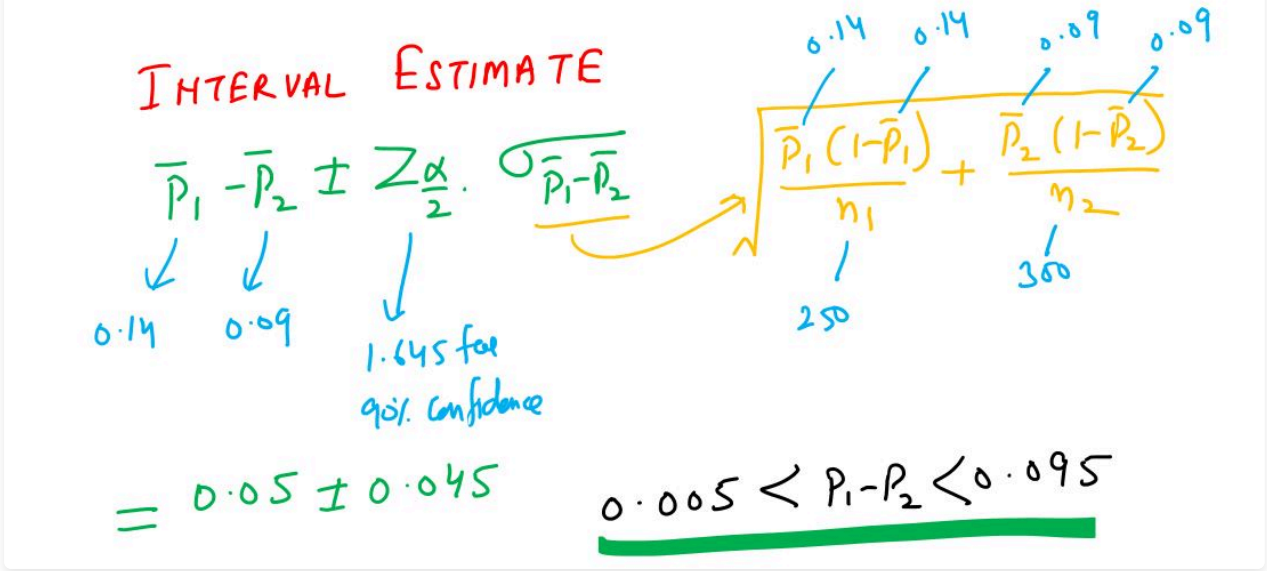
Thus, we estimate that Lucknow office has a 5% greater error rate than Ranchi office.

The **Interval Estimate** of the difference between the two population proportions using a 90% confidence interval with $z_{\alpha/2} = z_{0.05} = 1.645$ is given below:

$$\bar{p}_1 - \bar{p}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{p}_1 (1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2 (1 - \bar{p}_2)}{n_2}}$$

= 0.14-0.09 ± 1.645 $\sqrt{\frac{0.14 \cdot (1 - 0.14)}{250} + \frac{0.09 \cdot (1 - 0.09)}{300}}$ = 0.05 ± 0.045

Thus, the margin of error is 0.045, and the 90% confidence interval is 0.005 to 0.095.



The interval estimate of the difference of proportions between Lucknow and Ranchi being 0.005 to 0.095 indicates the range within which the true difference in the proportions of faulty Aadhaar enrollments between the two regional offices is likely to lie, with a certain level of confidence.

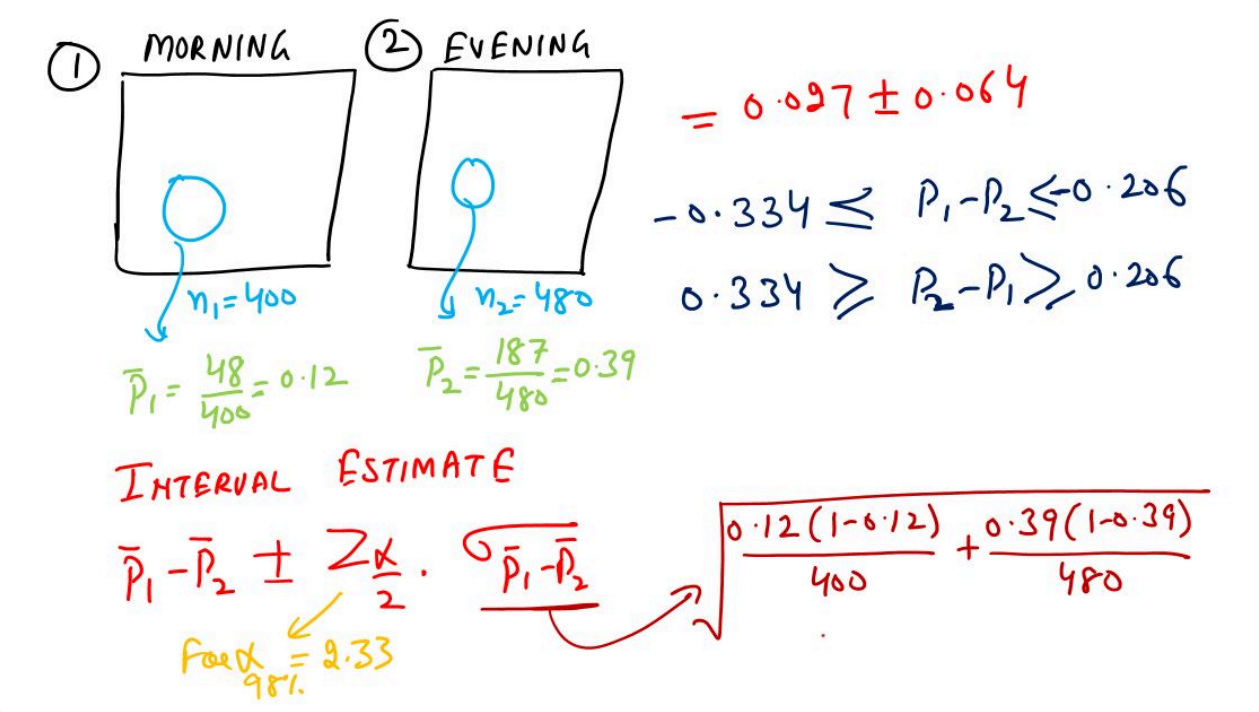
In this context, the interval suggests that, with the given level of confidence, the proportion of faulty Aadhaar enrollments is estimated to be higher in Lucknow compared to Ranchi. The values 0.005 to 0.095 represent the plausible range for this difference.

---

2. Point Estimation and Interval Estimation

At M&S, in an attempt to target its clientele, managers of a supermarket chain want to determine the difference between the proportion of morning shoppers who are men and the proportion of evening shoppers who are men. Over a period of two weeks, the chain's researchers conduct a systematic random sample survey of 400 morning shoppers, which reveals that 352 are women and 48 are men. During this same period, a systematic random sample of 480 evening shoppers reveals that 293 are women and 187 are men. Construct a 98% confidence interval to estimate the difference in the population proportions of men.

Solution:



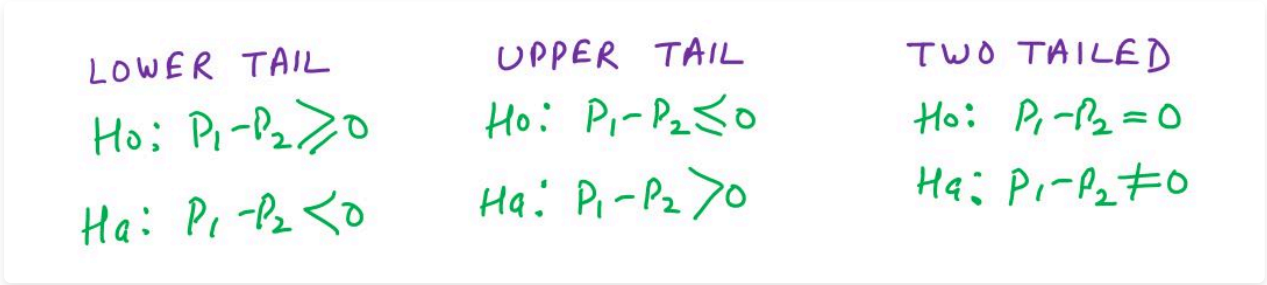
The confidence interval of (-0.334 to -0.206) suggests that, with 98% confidence, the true difference in the population proportions of men between morning and evening shoppers lies within this interval.

Specifically, the negative values in the interval indicate that the proportion of men is higher among evening shoppers compared to morning shoppers.

3. Hypothesis Testing- z test

Let us now consider hypothesis tests about the difference between the proportions of two populations. We focus on tests involving no difference between the two population proportions.

In this case, the three forms for a hypothesis test are as follows:



When we assume H_0 is true as an equality, we have $p_1 - p_2 = 0$, which is the same as saying $p_1 = p_2$.

When $p_1 = p_2 = p$, standard error will become as:

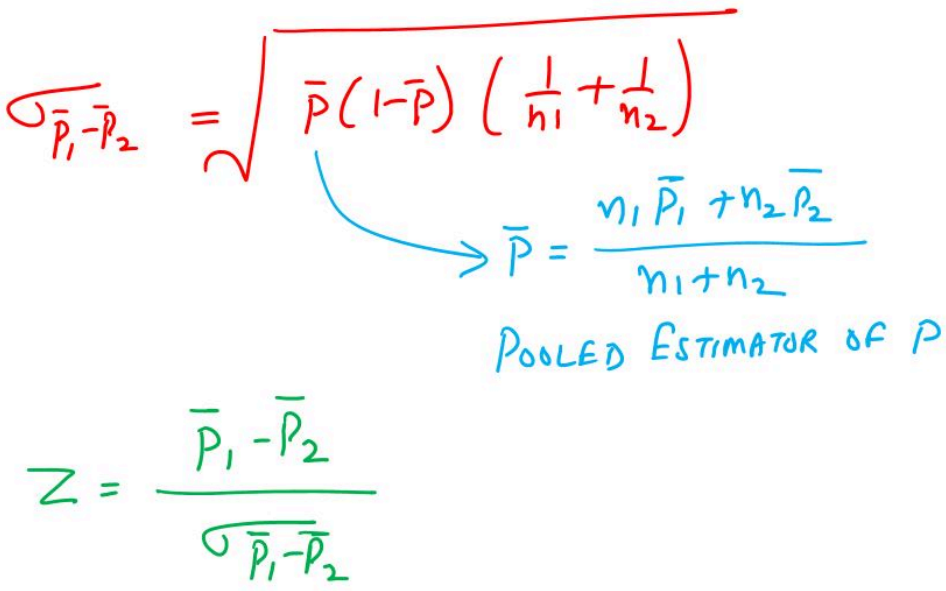
$$\sigma_{(\bar{p}_1 - \bar{p}_2)} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} = \sqrt{\frac{p(1 - p)}{n_1} + \frac{p(1 - p)}{n_2}}$$

$$= \sqrt{p(1 - p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

With p unknown, we pool, or combine, the point estimators from the two samples $\bar p_1$ and $\bar p_2$ to obtain a single point estimator of p as follows:

$$\bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2}$$

This **pooled estimator** of p is a weighted average of $\bar p_1$ and $\bar p_2$.



Substituting these value test statistic for Hypothesis test about p_1 and p_2 is shown below:

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

---

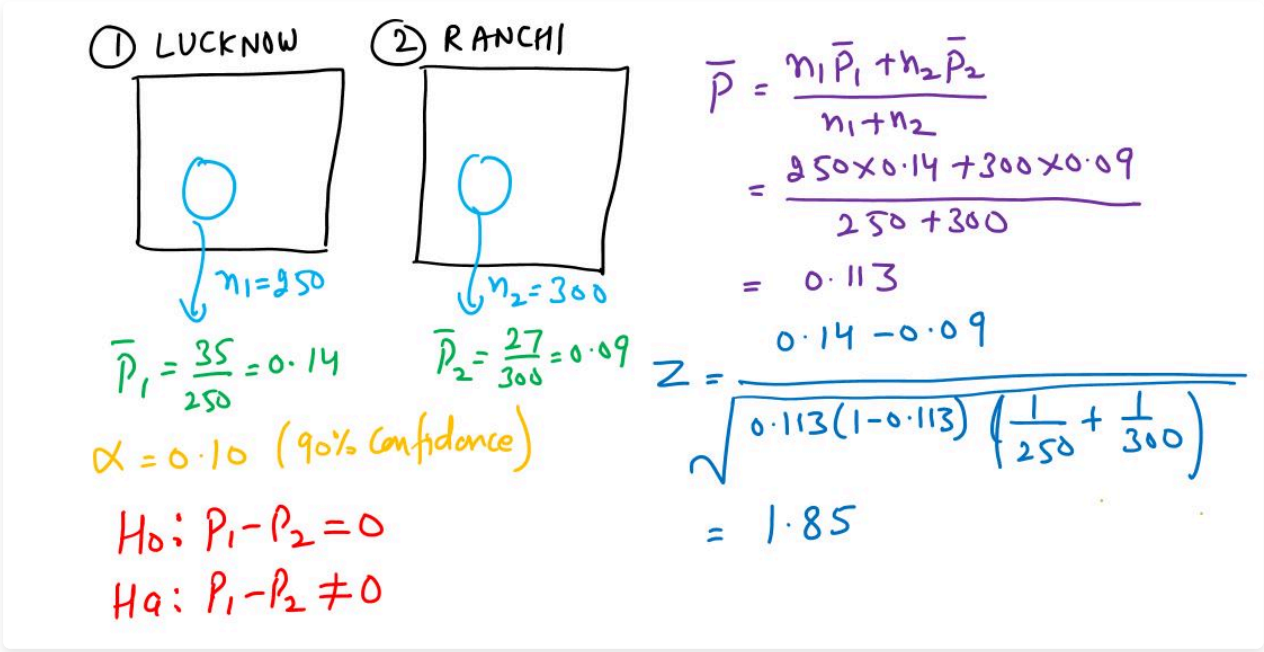
3. Hypothesis Testing- z test

The UIDAI has 8 regional offices all over the country. The UIDAI is interested in comparing the quality of Aadhaar enrollment at two of its regional offices, Lucknow and Ranchi. Is there a statistically significant difference in the proportion of faulty Aadhaar enrollments between the Lucknow and Ranchi regional offices of the UIDAI? Analyze at 90% confidence level.

Office	Sample Size	Number of Faulty Enrollments
Lucknow	$n_1 = 250$	35
Ranchi	$n_2 = 300$	27

Solution:

The sample proportions for the two offices follow:
\(\bar p_1 \) = \(\frac{35}{250} \) = 0.14 ,
\(\bar p_2 \) = \(\frac{27}{300} \) = 0.09



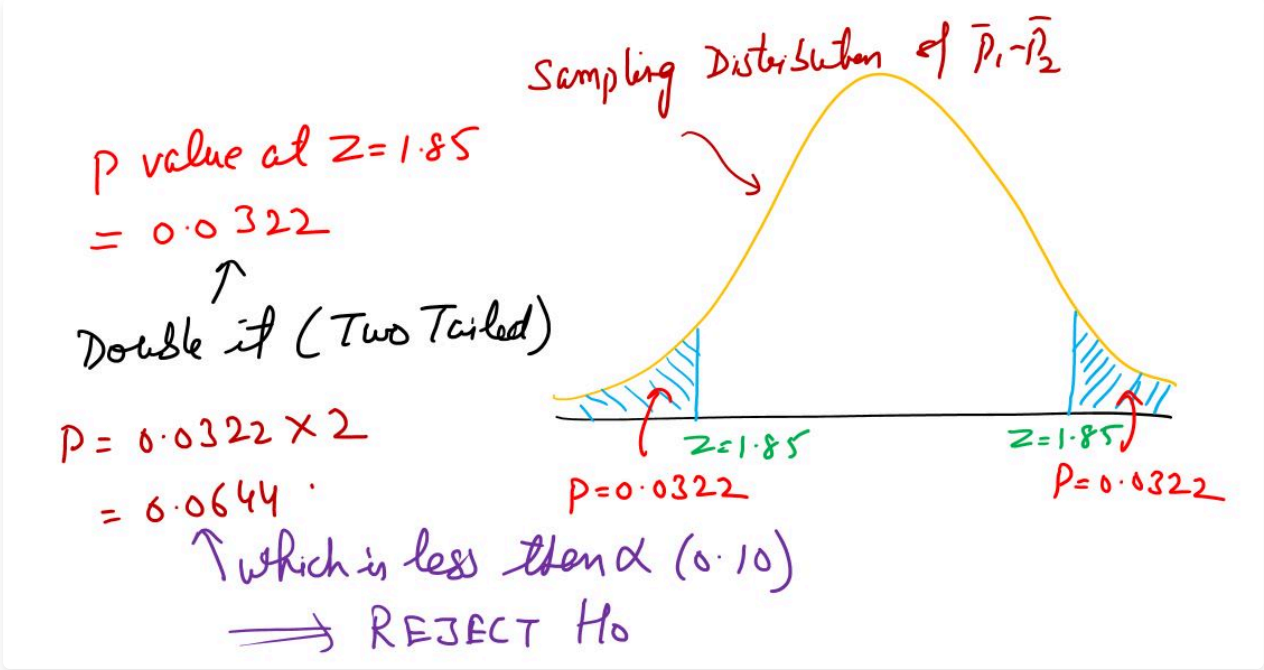
A two-tailed test is required. The null and alternative hypotheses are as follows:
H_0: $p_1 - p_2 = 0$
H_a: $p_1 - p_2$ \(\neq \) 0

If H_0 is rejected, the UIDAI can conclude that the rates of faulty enrollment at the two offices differ. We will use $\alpha = 0.10$ as the level of significance.

Pooled estimate of p, \(\bar p \) = \(\frac{{n_1 \bar{p}_1 + n_2 \bar{p}_2}}{{n_1 + n_2}} \) = \(\frac{{250 (0.14) + 300 (0.09)}}{{250 + 300}} \) = 0.1127

The Test Statistic, z = \(\frac{{\bar{p}_1 - \bar{p}_2}}{{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{{n_1}} + \frac{1}{{n_2}}\right)}}} = \frac{{(0.14 - 0.09)}}{{\sqrt{0.1127(1 - 0.1127)\left(\frac{1}{{250}} + \frac{1}{{300}}\right)}}} \) = 1.85

In computing the p-value for this two-tailed test, we first note that z = 1.85 is in the upper tail of the standard normal distribution.



Using z = 1.85 and the standard normal distribution table, we find the area in the upper tail is 1.0000 - 0.9678 = 0.0322. Doubling this area for a two-tailed test, we find the p-value = 2(0.0322) = 0.0644.

With the p-value less than $\alpha = 0.10$, H_0 is rejected at the 0.10 level of significance.

Thus, UIDAI can conclude that the rate of faulty enrollment differ between the two offices.

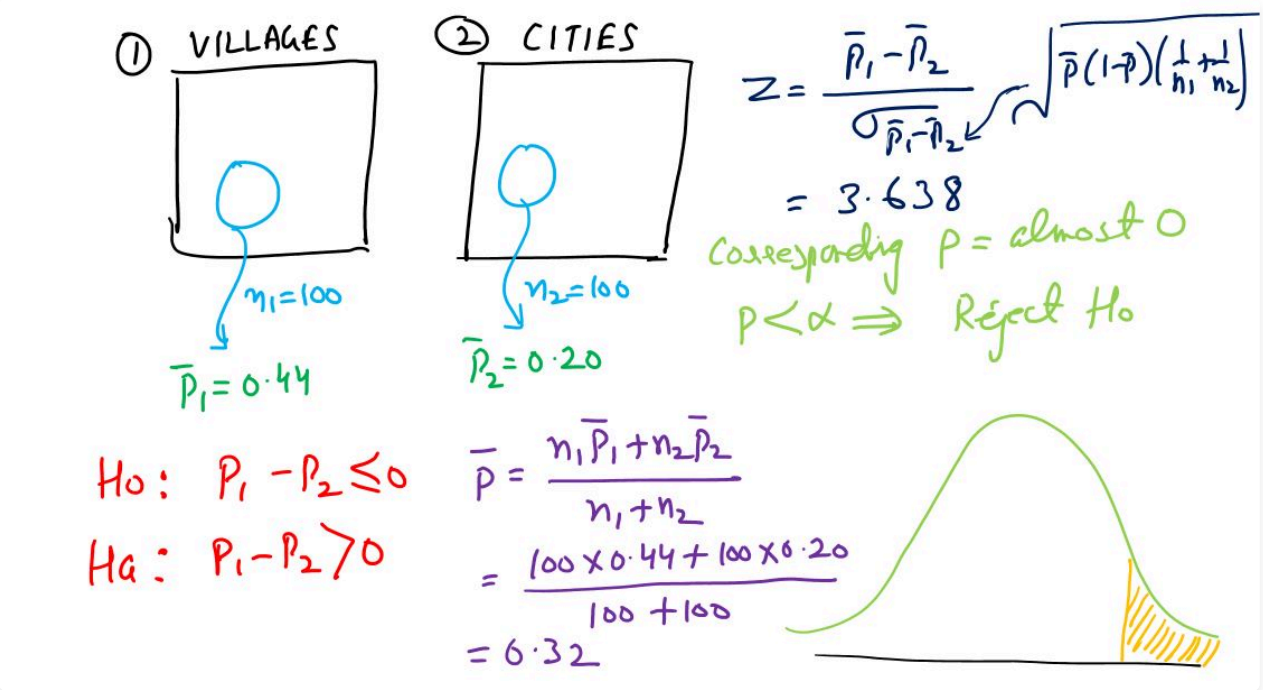
This hypothesis testing conclusion is consistent with the earlier interval estimation results that showed the interval estimate of the difference between the population error rates at the two offices to be 0.005 to 0.095, with Lucknow office having the higher rate of faulty enrollments.

---

3. Hypothesis Testing- z test

A survey was done on privacy concerns while using internet The survey reported that 44% of Internet users from villages are worried about privacy issues as compared to 20% of Internet users from cities. Suppose that the survey consisted of 100 individuals in each age group. At the 0.05 level of significance, is the proportion of worried Internet users from villages is greater than the proportion of Internet users from cities?

Solution:



Because we rejected the null hypothesis, it implies that there is sufficient evidence to suggest that the proportion of worried Internet users from villages is indeed greater than the proportion of Internet users from cities. In other words, the observed difference in privacy concerns between these two groups is statistically significant at the 0.05 level of significance.

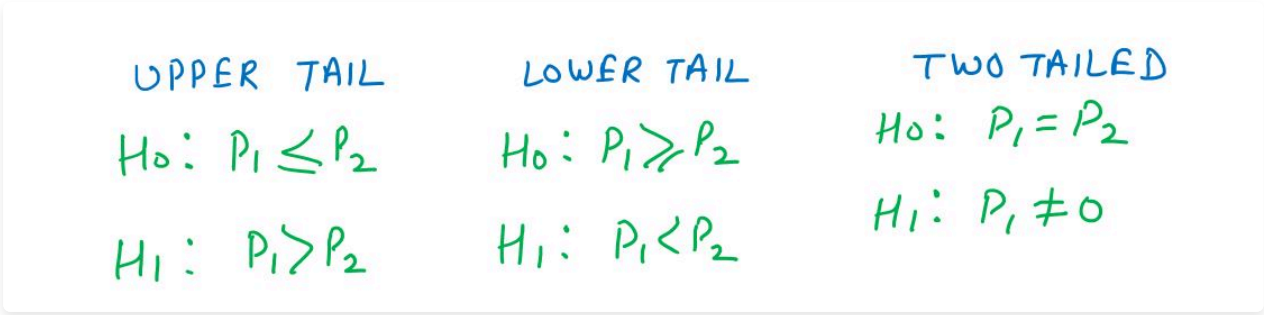
---

4. Hypothesis Testing- Chi Square Test

The Chi-Square test can also be used, if we have to compare proportions of two populations. It is called chi-square test of difference.

Here are the steps for conducting a chi-square test of difference for comparing proportions of 2 populations. Please note that, the chi-square test can be used to compare proportions of more than 2 poulations also, but here we will discuss steps for 2 populations only.

Step 1: Formulate Hypotheses
Null Hypothesis (H_0): There is no difference in proportions between the two populations.
Alternative Hypothesis (H_1): There is a significant difference in proportions between the two populations.

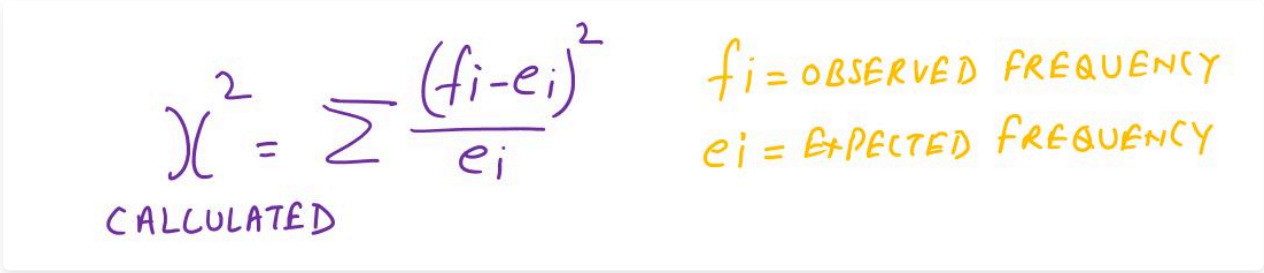


Step 2: Set Significance Level
Choose a significance level (commonly denoted as α).

Step 3: Create Contingency Table
Organize your data into a contingency table. This table should have two rows (representing the two populations) and two columns (representing the two categories being compared).

Step 4: Calculate Expected Frequencies
For each cell in the contingency table, calculate the expected frequency under the assumption that there is no difference in proportions (with assumption that the null hypothesis is True).

Step 5: Calculate Chi-Square Statistic
Compute the chi-square statistic using the formula:



Step 6: Degrees of Freedom
Determine the degrees of freedom for the chi-square distribution. For a 2x2 table, the degrees of freedom is (number of rows - 1) × (number of columns - 1), which is (2-1) × (2-1) = 1.

Step 7: Compare Chi-Square Statistic with Critical Value
Look up the critical chi-square value for your chosen significance level and degrees of freedom in a chi-square distribution table. Compare the calculated chi-square statistic with the critical value.

We can also use p value approach.

Step 8: Make a Decision
If the calculated chi-square statistic is greater than the critical value, reject the null hypothesis and conclude that there is a

significant difference in proportions between the two populations. If it is not, fail to reject the null hypothesis.

Step 9: Interpret Results
Provide a conclusion based on the statistical analysis, considering the context of your study and the results of the test.

We will discuss these steps in the illustrations, next.

---

4. Hypothesis Testing- Chi Square Test

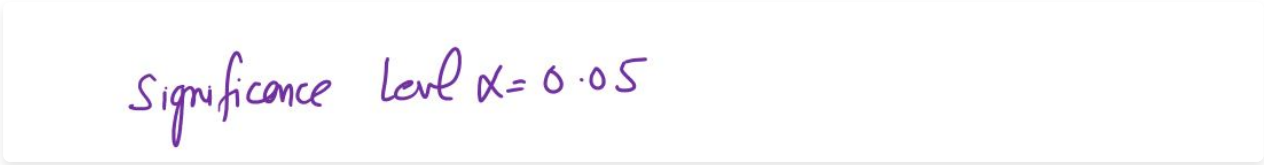
A survey was done on privacy concerns while using internet The survey reported that 44% of Internet users from villages are worried about privacy issues as compared to 20% of Internet users from cities. Suppose that the survey consisted of 100 individuals in each age group. At the 0.05 level of significance, is the proportion of worried Internet users from villages is greater than the proportion of Internet users from cities?

Solution:

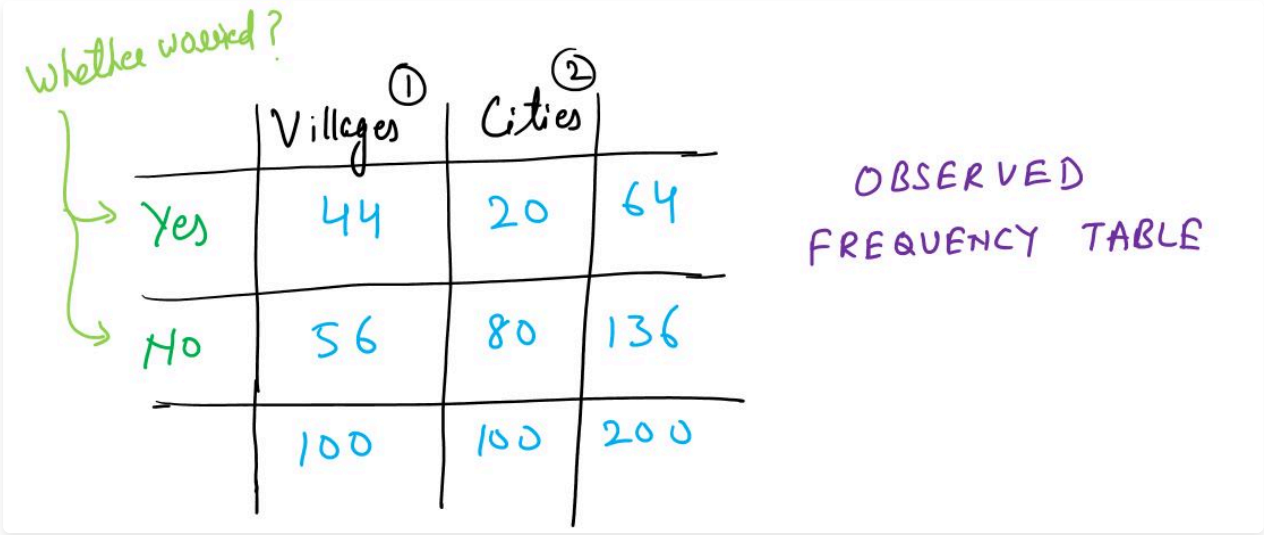
Let Villages is denoted by 1 and Cities by 2.

Step 1: Formulate Hypotheses

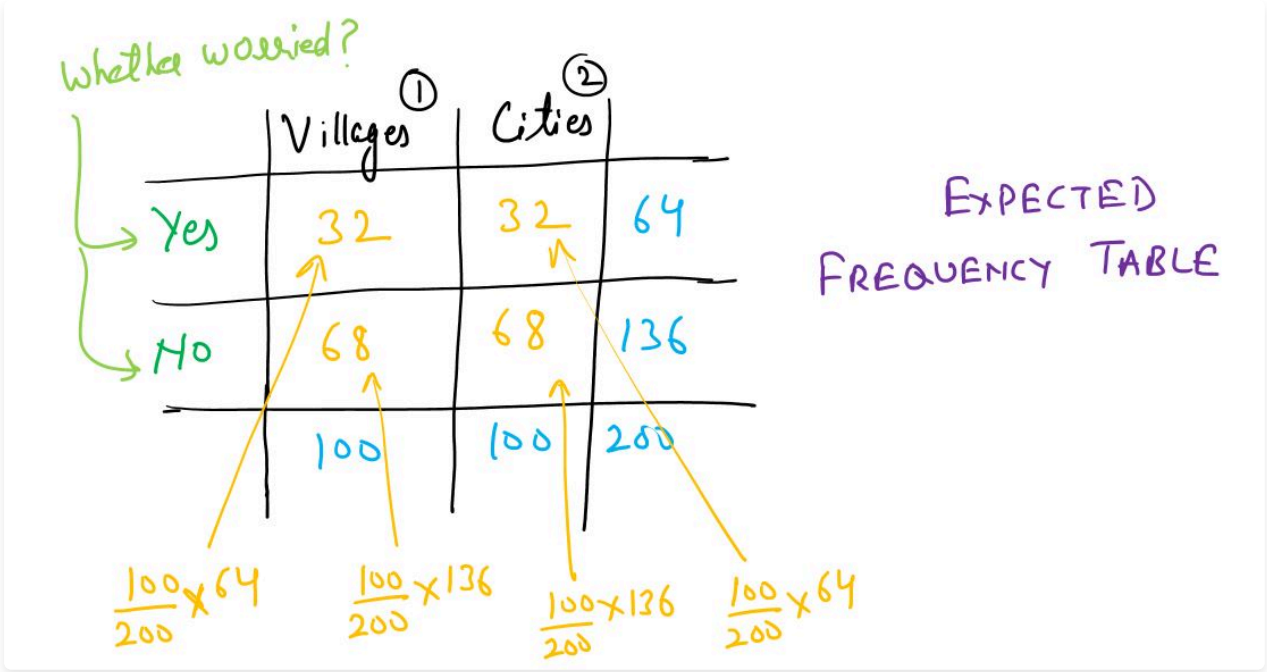
Step 2: Set Significance Level



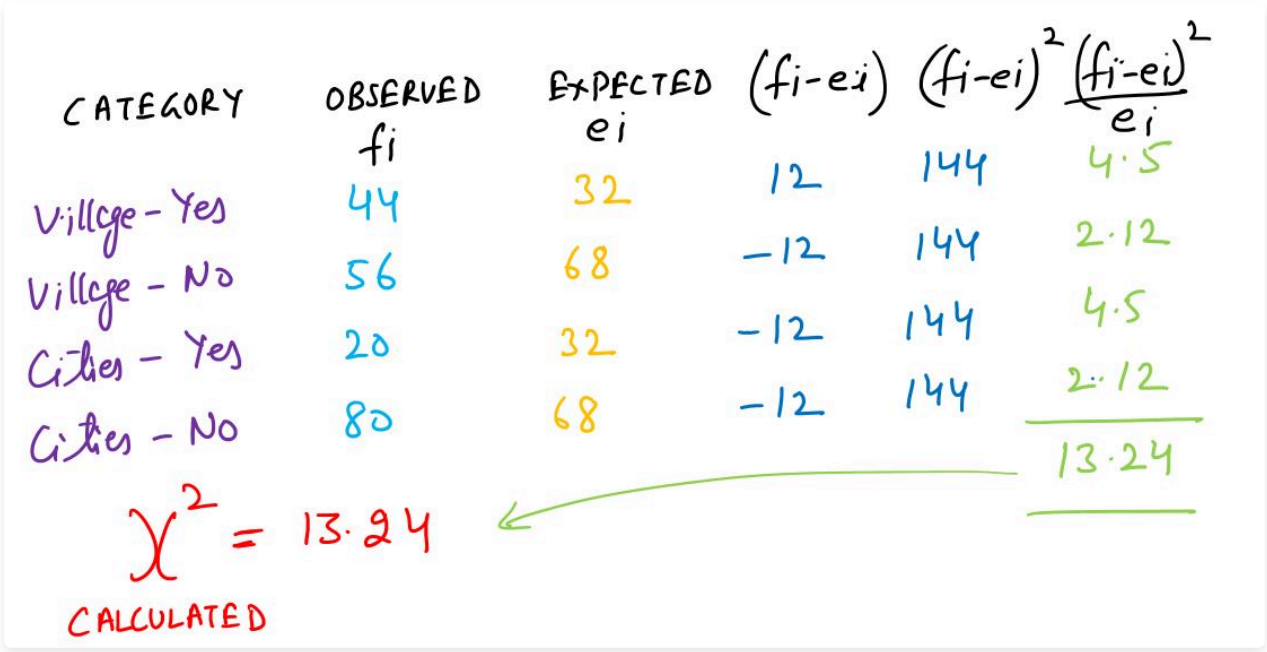
Step 3: Create Contingency Table



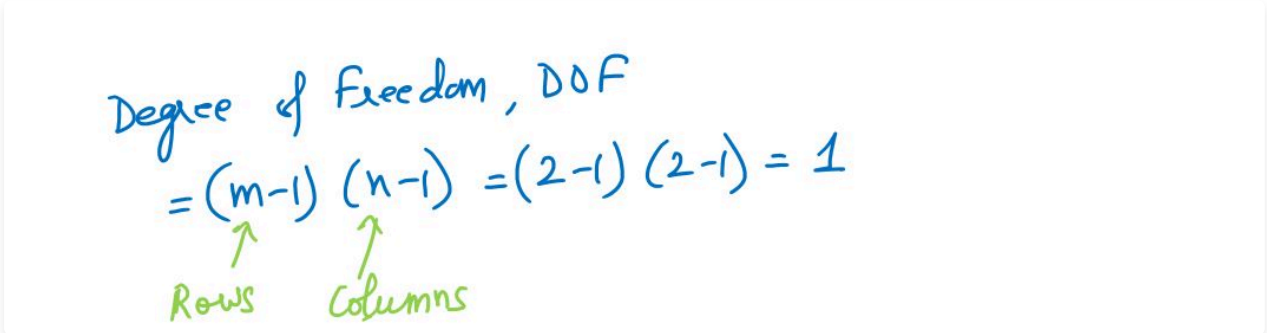
Step 4: Calculate Expected Frequencies



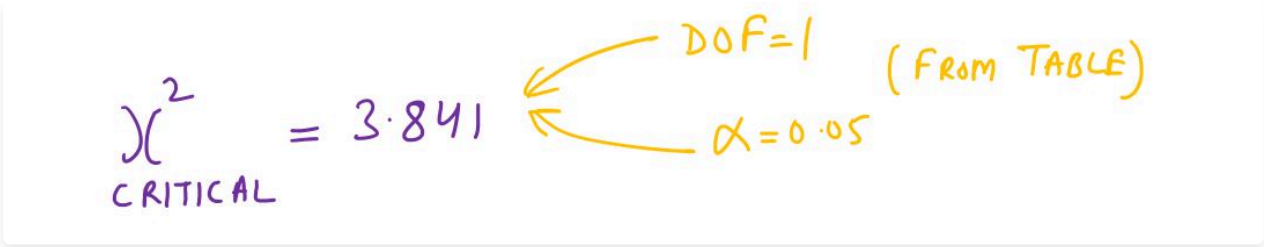
Step 5: Calculate Chi-Square Statistic



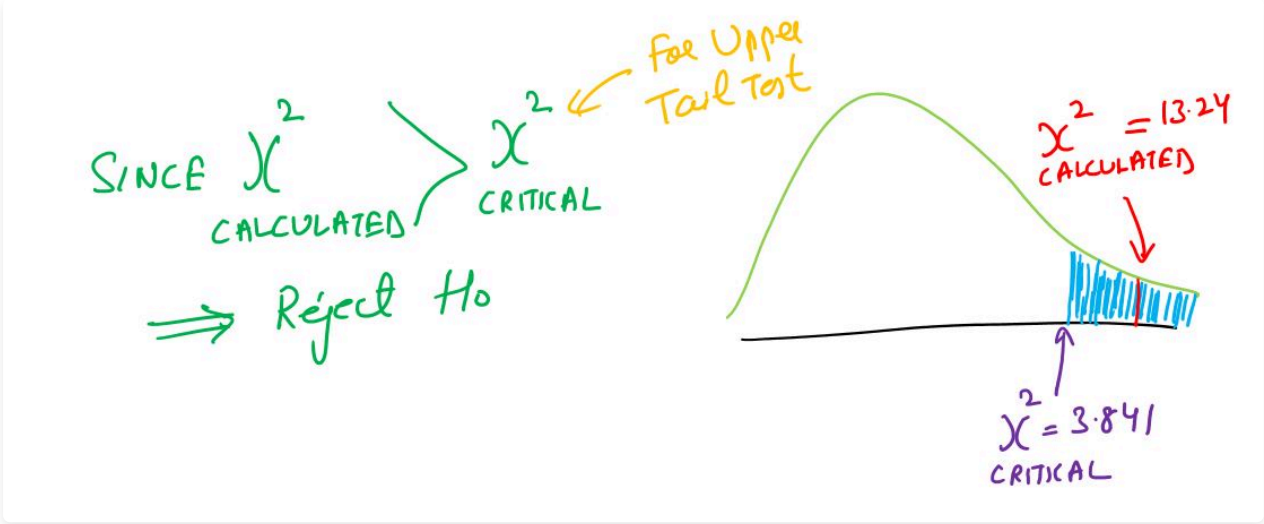
Step 6: Degrees of Freedom



Step 7: Calculate Critical Value of Chi-Square



Step 8: Make a Decision



Step 9: Interpret Results

Since we rejected the null hypothesis, it means that there is enough statistical evidence to support the claim that the proportion of Internet users worried about privacy issues in villages is greater than the proportion in cities.

---

5. z test or chi square ?

We have learned that when comparing proportions between 2 populations, both the z-test and chi-square test can be used. Both methods will yield comparable results.

However, when dealing with the comparison of proportions across more than 2 populations, the chi-square test becomes the only option.

In the context of the chi-square test, it is crucial that each expected frequency in the contingency table is at least 5 for the results to be reliable. If this condition is not met, the Fisher Exact test is an alternative method that can be used to address the issue of small expected frequencies in a more accurate manner

---