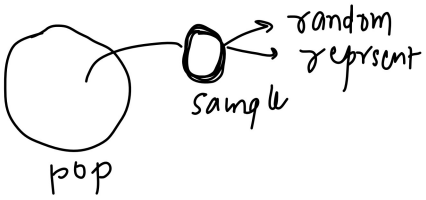
Some Terms

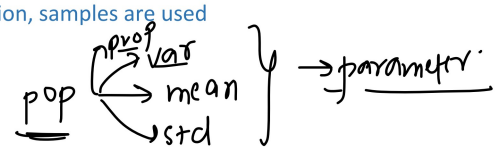
30 March 2023 07:09



Population Vs Sample

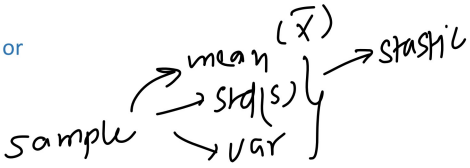
Population: A population is the entire group or set of individuals, objects, or events that a researcher wants to study or draw conclusions about. It can be people, animals, plants, or even inanimate objects, depending on the context of the study. The population usually represents the complete set of possible data points or observations.

Sample: A sample is a subset of the population that is selected for study. It is a smaller group that is intended to be representative of the larger population. Researchers collect data from the sample and use it to make inferences about the population as a whole. Since it is often impractical or impossible to collect data from every member of a population, samples are used as an efficient and cost-effective way to gather information.



Parameter Vs ~~Estimate~~ statistics

Parameter: A parameter is a numerical value that describes a characteristic of a population. Parameters are usually denoted using Greek letters, such as μ (mu) for the population mean or σ (sigma) for the population standard deviation. Since it is often difficult or impossible to obtain data from an entire population, parameters are usually unknown and must be estimated based on available sample data.



$\bar{X} \rightarrow \mu$

Statistic: A statistic is a numerical value that describes a characteristic of a sample, which is a subset of the population. By using statistics calculated from a representative sample, researchers can make inferences about the unknown respective parameter of the population. Common statistics include the sample mean (denoted by $\bar{x}$, pronounced "x-bar"), the sample median, and the sample standard deviation (denoted by s).

sample → population

Inferential Statistics

Inferential statistics is a branch of statistics that focuses on making predictions, estimations, or generalizations about a larger population based on a sample of data taken from that population. It involves the use of probability theory to make inferences and draw conclusions about the characteristics of a population by analysing a smaller subset or sample.

The key idea behind inferential statistics is that it is often impractical or impossible to collect data from every member of a population, so instead, we use a representative sample to make inferences about the entire group. Inferential statistical techniques include hypothesis testing, confidence intervals, and regression analysis, among others.

These methods help researchers answer questions like:

a. Is there a significant difference between two groups?
b. Can we predict the outcome of a variable based on the values of other variables?
c. What is the relationship between two or more variables?

Inferential statistics are widely used in various fields, such as economics, social sciences, medicine, and natural sciences, to make informed decisions and guide policy based on limited data.

Example explained in this video was - suscriber of Nitish bhaiya is 77k , he wants to know the average age of his suscribers , so what he did is he started live stream , where he asked the age of his live suscribers , and he then calculated the mean . now that is what called point estimate

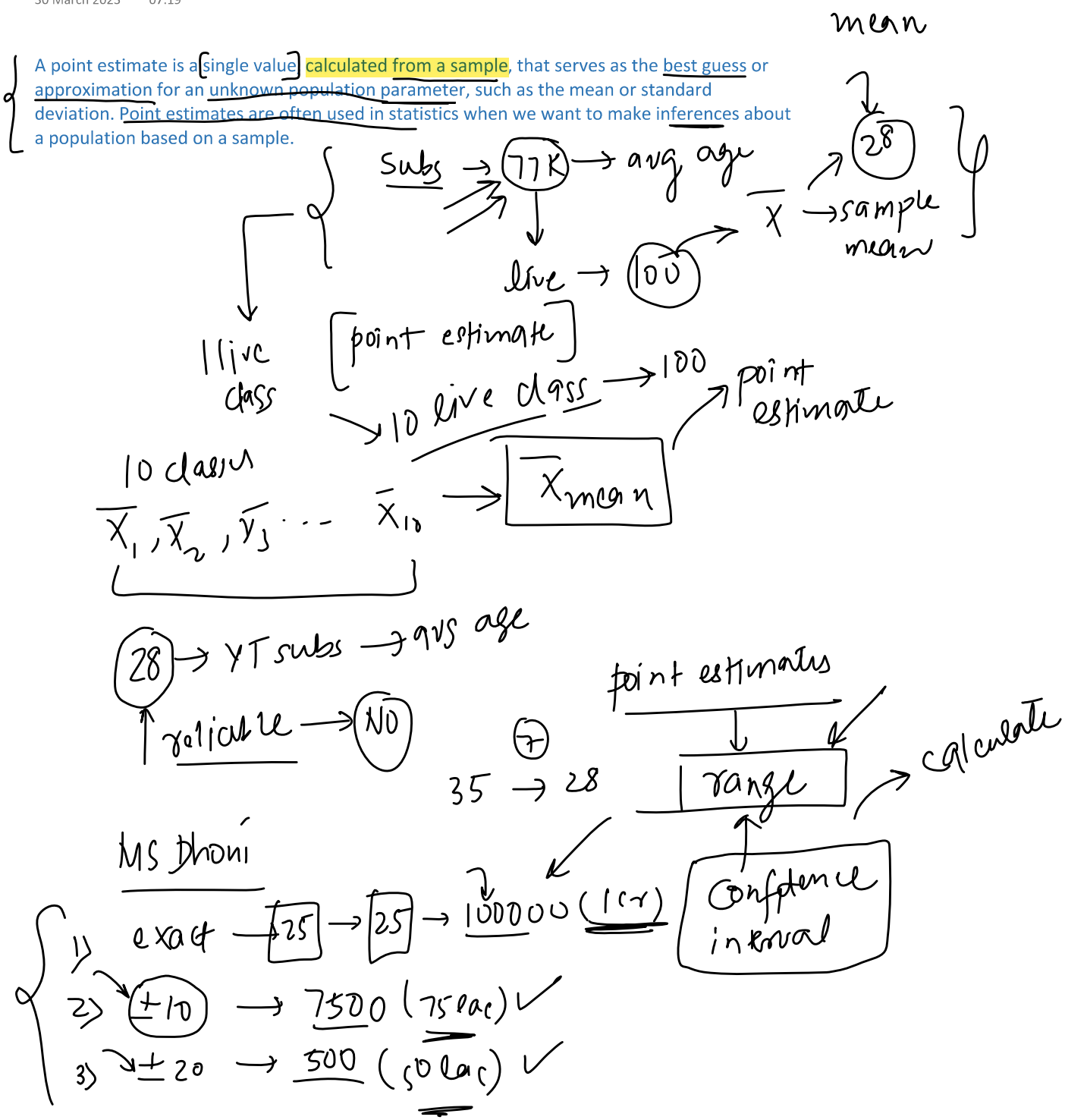
now we can also have a stronger point estimate , you know the ans ie CLT. (Conducting 10 live classes -> 10 means then un means ka mean which is also a point estimate.

Point Estimate

30 March 2023 07:19

now the problem is that point estimates are not reliable , as we are using single number , hence surety will be very less that the number will be same for population. (example of guessing dhoni run in the match , having interval would be more useful than depending on point estimate.) , and that interval is confidence interval

A point estimate is a single value calculated from a sample, that serves as the best guess or approximation for an unknown population parameter, such as the mean or standard deviation. Point estimates are often used in statistics when we want to make inferences about a population based on a sample.



Confidence Interval

30 March 2023 07:18

Confidence interval, in simple words, is a range of values within which we expect a particular population parameter, like a mean, to fall. It's a way to express the uncertainty around an estimate obtained from a sample of data.

Confidence level, usually expressed as a percentage like 95%, indicates how sure we are that the true value lies within the interval.

Confidence Interval = [Point Estimate] ± [Margin of Error]

Ways to calculate CI:

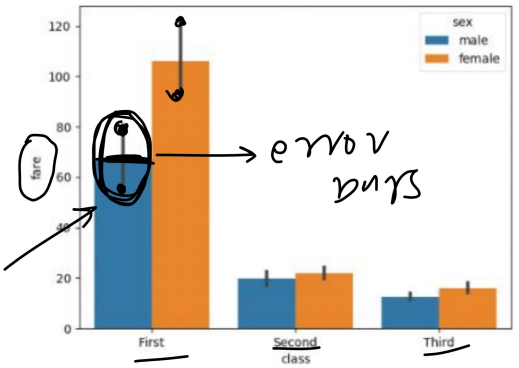
Two ways to calculate CI -
1. Z-Procedure ,
2.T - Procedure

Z- Procedure used when we have population std ,
T-Procedure used when we do not have any population parameters

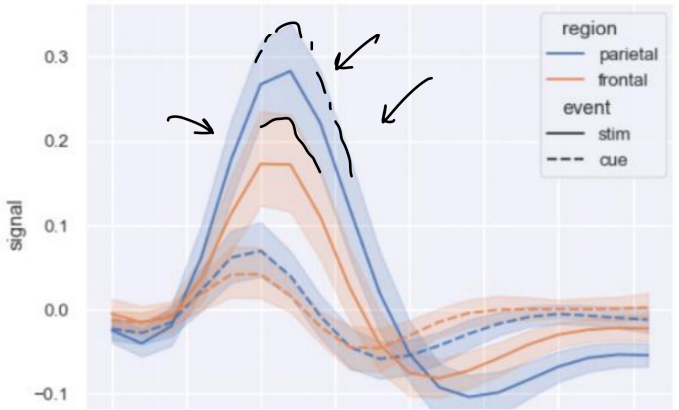
Confidence Interval is created for Parameters and not statistics. Statistics help us get the confidence interval for a parameter.

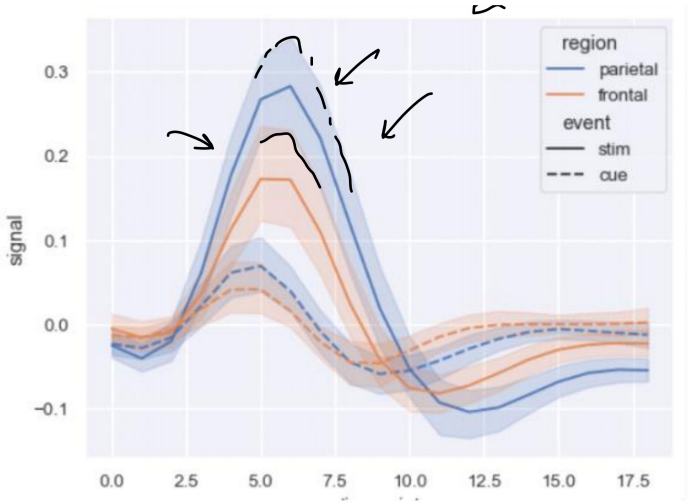
Examples of CT usage

seaborn → bar plot



When we find the fare , the barplot represents avg fare of different class for data we have(sample), but there is also other thing seaborn barplot shows that is confidence interval , the vertical arrow shows that the avg fare for population may range from down to up of arrow




region
parietal
frontal
event
stim
cue
signal
0.3
0.2
0.1
0.0
−0.1
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

Z - Procedure(gererally we dont have population parameter , hence this is used less)

Confidence Interval (Sigma Known)

30 March 2023 07:13

Z procedure → pop → sample random

(σ) pop std available

Assumptions

1. Random sampling: The data must be collected using a random sampling method to ensure that the sample is representative of the population. This helps to minimize biases and ensures that the results can be generalized to the entire population.
2. Known population standard deviation: The population standard deviation (σ) must be known or accurately estimated. In practice, the population standard deviation is often unknown, and the sample standard deviation (s) is used as an estimate. However, if the sample size is large enough, the sample standard deviation can provide a reasonably accurate approximation.
3. Normal distribution or large sample size: The Z-procedure assumes that the underlying population is normally distributed. However, if the population distribution is not normal, the Central Limit Theorem can be applied when the sample size is large (usually, sample size n ≥ 30 is considered large enough). According to the Central Limit Theorem, the sampling distribution of the sample mean will approach a normal distribution as the sample size increases, regardless of the shape of the population distribution.

YT → pop → std (σ)

distribution → normal

77K subs → age → normal dist

CLT

Sample size ≤ 30 → 20 → z procedure

σ = 15

A (1 - alpha)*100% Confidence Interval for mu:

YT → campus X → 77K → 28 ± 14 → [16, 42] ← confidence interval

Confidence level → 95%

formula

CI using Z procedure

$$CI = \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

1) intuition

2) $Z_{\alpha/2}$

Z → ?

(1−α) → confidence level

(1−α) → 95%

σ → std pop

n → sample size → 100

Intuition

Point estimate $(\bar{X})$ → CLT

z-procedure assumes population is normal but if not you can apply CLT , but sample size should be large enough (generally n>=30

Formula = point estimate +- Zalpha/2 * sigma/root(n)

1 - alpha = confidence level (in our case we want 95 percent confidence)

sigma = population std

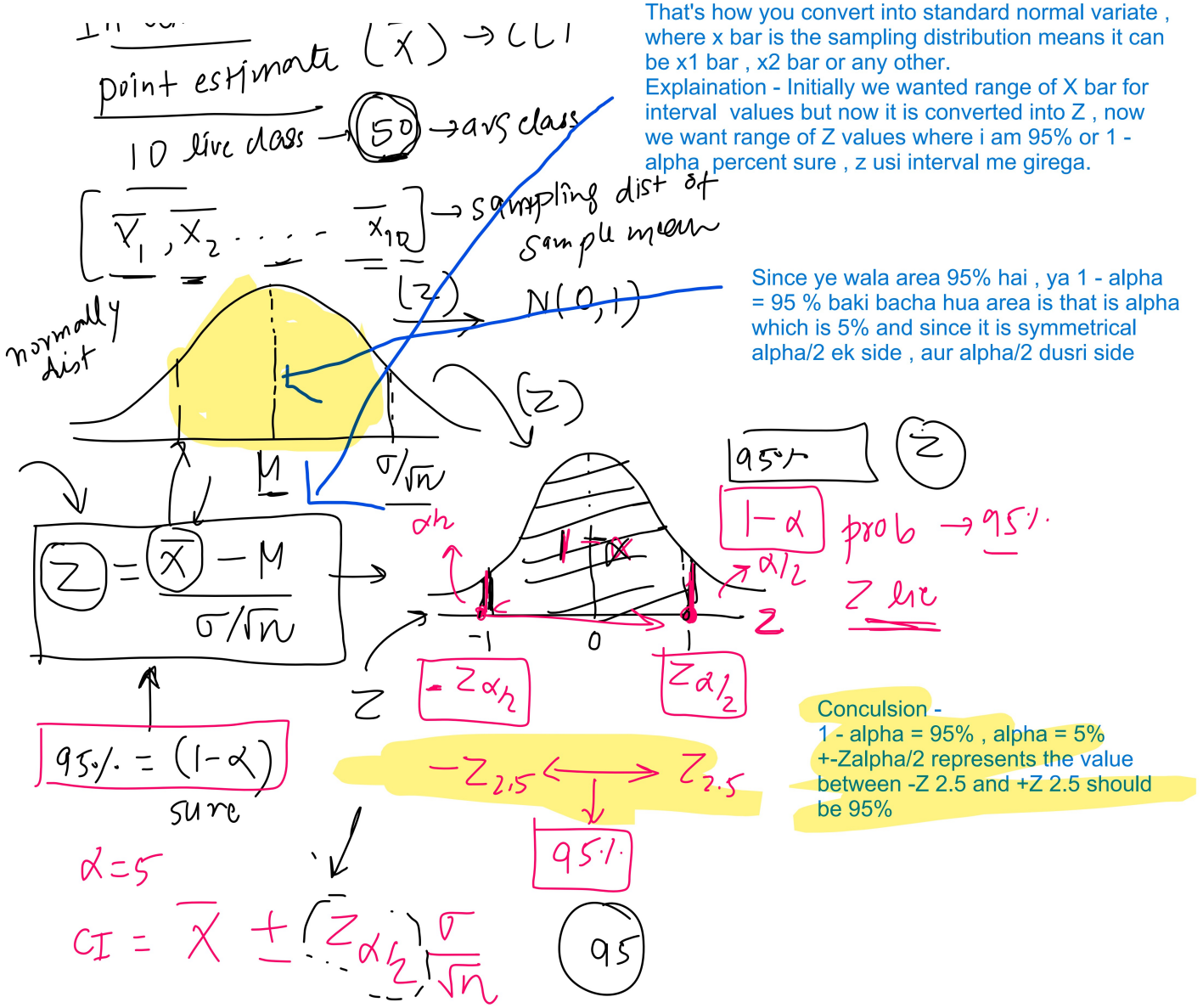
n = sample size

ques1 intuition of this formula

ques2 what is Z alpha/2

Intuition

1. use CLT
2. what if we convert sampling distribution which is normal , into standard normal variate(mean = 0 , std = 1)



That's how you convert into standard normal variate , where x bar is the sampling distribution means it can be x1 bar , x2 bar or any other.
Explaination - Initially we wanted range of X bar for interval values but now it is converted into Z , now we want range of Z values where i am 95% or 1 - alpha percent sure , z usi interval me girega.

Since ye wala area 95% hai , ya 1 - alpha = 95 % baki bacha hua area is that is alpha which is 5% and since it is symmetrical alpha/2 ek side , aur alpha/2 dusri side

Conclusion -
1 - alpha = 95% , alpha = 5%
+-Zalpha/2 represents the value between -Z 2.5 and +Z 2.5 should be 95%

$$P\left(-z_{\alpha/2} < z < z_{\alpha/2}\right) = 1-\alpha$$

we will do some shifting as we want our interval for mean

$$P\left(-z_{\alpha/2} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$$

μ → CI

$$P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{x}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(-\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$P\left(-X - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < \quad + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$

$P\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$

μ → fixed

$\bar{X}$ → sample

95%

$1-\alpha$ → 0.95

$\alpha = 0.50$

$\frac{\alpha}{2}$ → 0.250

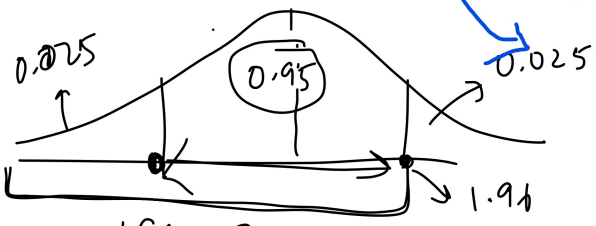
confidence $(1-\alpha)$ → 95%

CI $\mu = \bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

$\mu = \bar{x} \pm Z_{0.250}\frac{\sigma}{\sqrt{n}}$ (Z)

950
025
975

correct value is 0.025



range → confidence interval

$\mu = \bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ → CI with a confidence of 95%.

50%

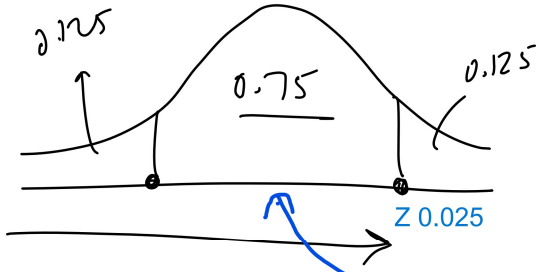
75%

99%

0.87

$Z_{\alpha/2}$

$Z_{\alpha/2}$ → 1.96



Z → 1.13

we know that the population mean is a fixed quantity then why this range , the answer is simple ie because of sample means, range comes because there is variabilty of sample mean , one time we may have different sample , next time we may have differe that's why.

now time to ans how to find Z
The Z 0.025 for the area of .975 is 1.96

and -Z-.025 iis -1.96

hence final formula is this

we know that total area is 1 , area of Z.025 is 0.025 after that , hence we want z value for the area 1 - 0.025 = 0.975 now we have to look for Z value whose area is 0.975 (how - we know that confidence level = 1 - alpha , hence alpha = 5 , since it is symmetrical , 2.5 percent are will be after that 95 percent and 2.5 percent will be before.

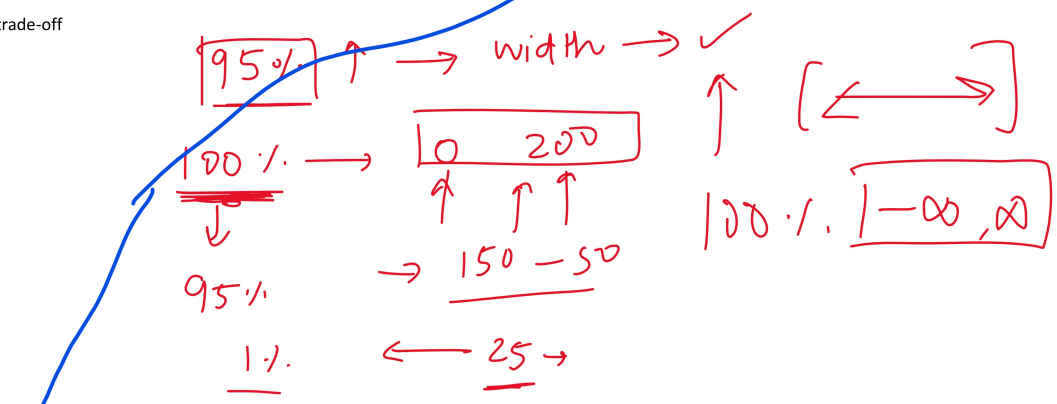
Interpreting Confidence Interval

30 March 2023 08:33

A confidence interval is a range of values within which a population parameter, such as the population mean, is estimated to lie with a certain level of confidence. The confidence interval provides an indication of the precision and uncertainty associated with the estimate. To interpret the confidence interval values, consider the following points:

1. **Confidence level**: The confidence level (commonly set at 90%, 95%, or 99%) represents the probability that the confidence interval will contain the true population parameter if the sampling and estimation process were repeated multiple times. For example, a 95% confidence interval means that if you were to draw 100 different samples from the population and calculate the confidence interval for each, approximately 95 of those intervals would contain the true population parameter.
2. **Interval range**: The width of the confidence interval gives an indication of the precision of the estimate. A narrower confidence interval suggests a more precise estimate of the population parameter, while a wider interval indicates greater uncertainty. The width of the interval depends on the sample size, variability in the data, and the desired level of confidence.
3. **Interpretation**: To interpret the confidence interval values, you can say that you are "X% confident that the true population parameter lies within the range (lower limit, upper limit)." Keep in mind that this statement is about the interval, not the specific point estimate, and it refers to the confidence level you chose when constructing the interval.

What is the trade-off



wrong interpretation - mere suscriber ka age 14 to 42 hone ka probabilty is 95% which is wrong interpretation , note that confidence level is not probability.

Simple meaning Confidence interval - lets say mere 95% confidence ke sath aya yee interval [18 , 42] meaning kya hai ?
meaning - since population is 77k what we did is we make samples of size lets say 50 , 100 times , ab jab un 50 logo ka avg age calculate krenge of 100 sets , 95 sets me avg age [18 , 42] interval me rhega (95%)

use confidence interval simulation visualization tool to have a better understanding

Factors Affecting Margin of Error

30 March 2023 07:15

upper − lower / 2 → margin

1. Confidence Level (1-alpha)
2. Sample Size
3. Population Standard Deviation

$$CI = \text{point estimate} \pm \text{margin of error}$$

$$= \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$\bar{x}$ → sample mean

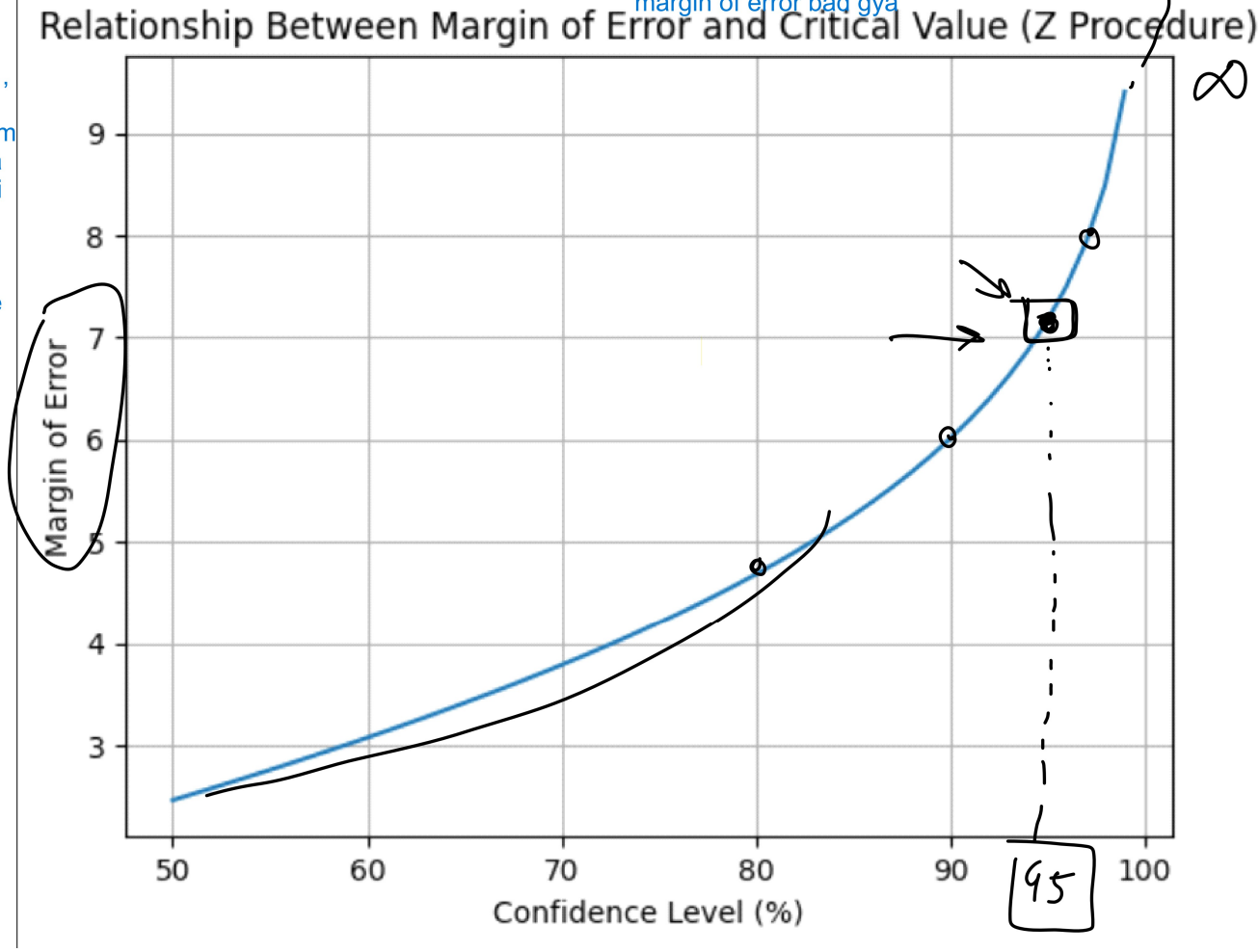
$z_{\alpha/2}$ → (1) critical value

σ → pop std (2)

$\sqrt{n}$ → sample size (3)

relation bwteen population std and margin of error - directly proprtional jiska population me hi zada variance hoga tb usko estimate krne me bhi variance dikgea example - dhoni ut patan run banata hai kbhi 10 , 100 , 200 to since uska variance zada hai , to we know ki 95 confidence me uska run 20 se 80 ke beech hoga see margin of error bad gya

aur aga isi jagah consisent player ho jaise rahul , uska variance km hai to uska interval bhi kum hoga even with high confidence



Relationship between confidence level and margin of error - jaise jaise confidence level badega vaise vaise range badegi

use nitihsh sir notebook to explore relation between factors and margin of error

Confidence Interval (Sigma not known)

30 March 2023 07:15

σ → t-procedure

Using the t procedure

Assumptions

1. Random sampling: The data must be collected using a random sampling method to ensure that the sample is representative of the population. This helps to minimize biases and ensures that the results can be generalized to the entire population.
2. Sample standard deviation: (majboori) The population standard deviation (σ) is unknown, and the sample standard deviation (s) is used as an estimate. The t-distribution is specifically designed to account for the additional uncertainty introduced by using the sample standard deviation instead of the population standard deviation.
3. Approximately normal distribution: The t-procedure assumes that the underlying population is approximately normally distributed, or the sample size is large enough for the Central Limit Theorem to apply. If the population distribution is heavily skewed or has extreme outliers, the t-procedure may not be accurate, and non-parametric methods should be considered.
4. Independent observations: The observations in the sample should be independent of each other. In other words, the value of one observation should not influence the value of another observation. This is particularly important when working with time series data or data with inherent dependencies.

sample ≥ 30 → if its not normal → normal → small sample size (sam) → CT

$t_{\alpha/2} \gg z_{\alpha/2}$

$t_{\alpha/2} \simeq z_{\alpha/2}$

Z-table

$CI = \bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ X → S

pop mean → $\bar{X}$

$CI = \bar{X} + z_{\alpha/2} \frac{s}{\sqrt{n}}$ complexity

$CI = \bar{X} + t_{\alpha/2} \frac{s}{\sqrt{n}}$

ds ≥ 30

95% $z_{\alpha/2}$ → 1.96 $t_{\alpha/2}$ = 2.0008

$\bar{X}$ → $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ → standard normal dist

z' → $\frac{\bar{X} - \mu}{s/\sqrt{n}}$ → Student's t distribution → parameter → df → n−1

n−1

theoritical dist

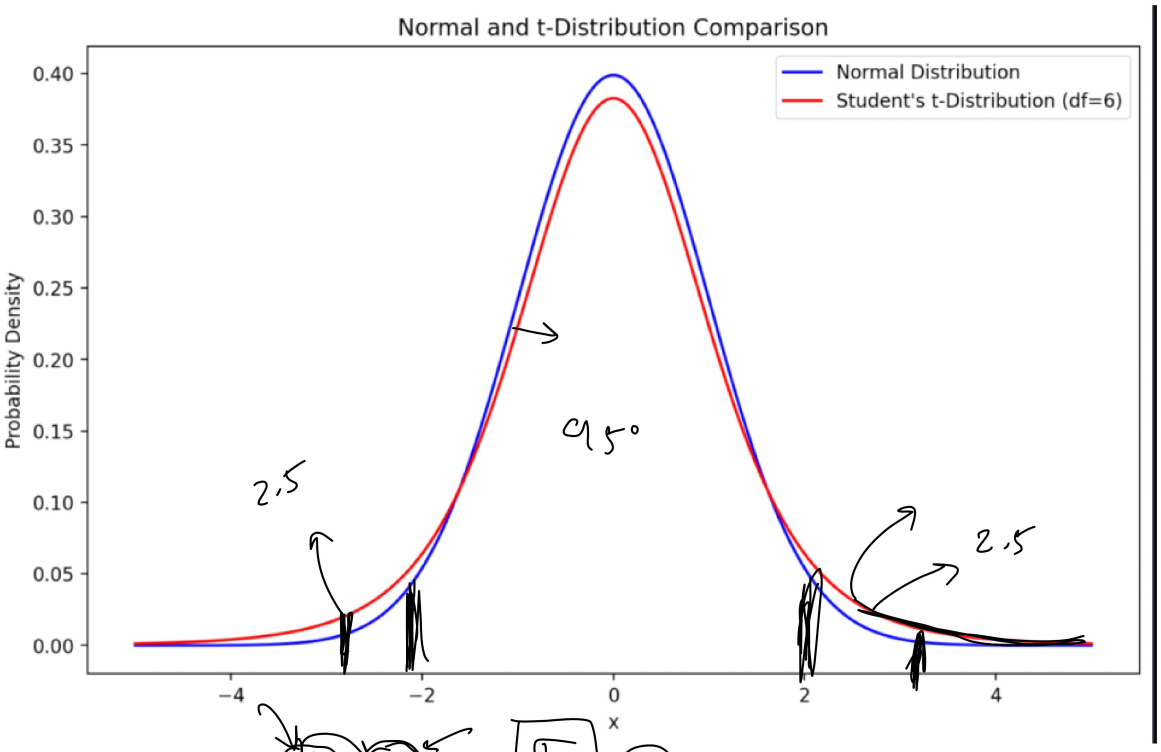
df → ∞ (t → N)

49 30

Student's t ds

M, σ

now since we dont have population std , we will use sample std , but this is not complete statement. When you will use sample std , there will be a complexity.
we know that we can easily convert our X bar distribution into standard normal variate, but now we have replaced population std with sample std , so now formula is this so now problem is this it did not remain normal distribution it changed to some other distribution ie Student's T distribution.Complexity ye aii , ki for mean jo sample to sample vary kr rha hai , uske lie hum confidence interval bana rhe the , but problem ye hai ki humm ab sample std ka use kr rhe hai jo ki sample to sample very krega(itself uncertain)



$CI = \bar{x} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}}$

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

$CI = \bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$ $\longleftrightarrow$ $CI = \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

CLT

$\bar{x}_1$ s_1

s_2

s_3

s_{10}

s_{avg}

diff between normal and T , T is theoritical distribution , it doesn't exist in nature(Used for handling that uncertainity)

parameters of T is only one ie Degree of Freedom (n - 1) if sample size is 50 then degree of freedom is 49.
T have fatter tails than Normal Distribution. Use Visualization tool. Jaise jaise sample size badaoge , vaise vaise your t distribution will look like normal , as you are incrasing samples , now there is less certainity as t distiribution was made because of uncertainty.

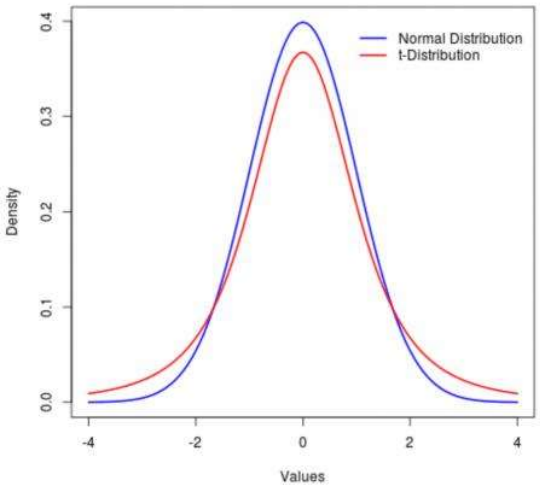
so now the formula updated to T rather than Z because we are using t-distribution

Conclusion - when you use Z score while using sample std , you will get wrong results more . lets say you habe confidence level of 95% but the actual correct will be only 92. so after trying ths experimeny maybe the founder get to know that the assumption of having normal distribution while using sample std is wrong , it must be other distribution.

also not that since sample std , therefore the confidence interval will be more in case of T score , as there s uncertainity

Student's T Distribution

30 March 2023 07:16



Student's t-distribution, or simply the t-distribution, is a probability distribution that arises when estimating the mean of a normally distributed population when the sample size is small and the population standard deviation is unknown. It was introduced by William Sealy Gosset, who published under the pseudonym "Student."

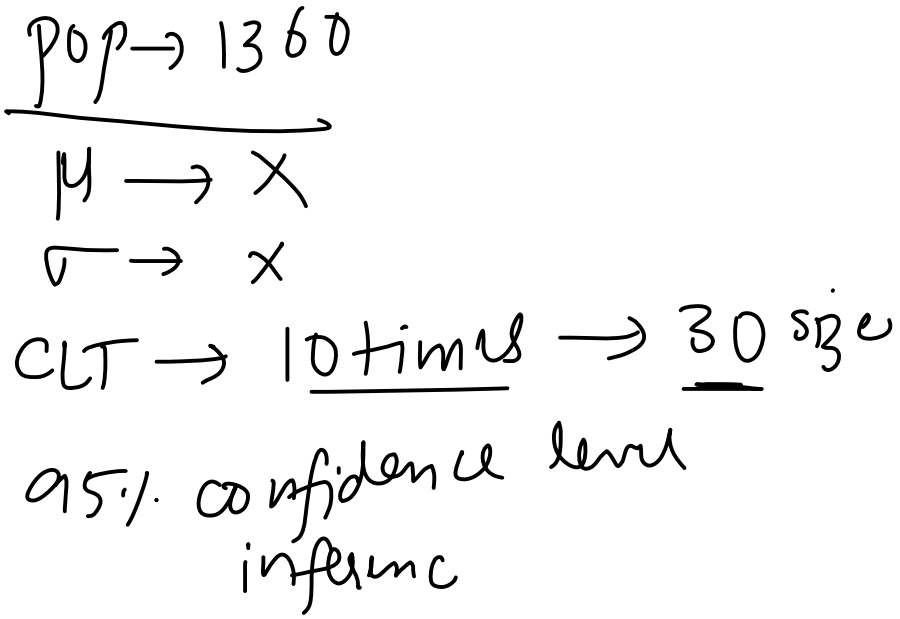
The t-distribution is similar to the normal distribution (also known as the Gaussian distribution or the bell curve) but has heavier tails. The shape of the t-distribution is determined by the degrees of freedom, which is closely related to the sample size (degrees of freedom = sample size - 1). As the degrees of freedom increase (i.e., as the sample size increases), the t-distribution approaches the normal distribution.

In hypothesis testing and confidence interval estimation, the t-distribution is used in place of the normal distribution when the sample size is small (usually less than 30) and the population standard deviation is unknown. The t-distribution accounts for the additional uncertainty that arises from estimating the population standard deviation using the sample standard deviation.

To use the t-distribution in practice, you look up critical t-values from a t-distribution table, which provides values corresponding to specific degrees of freedom and confidence levels (e.g., 95% confidence). These critical t-values are then used to calculate confidence intervals or perform hypothesis tests.

Titanic Case Study

31 March 2023 18:00



Confusioon - since we are using CLT there will be multiple sample std which we will choose , we will find the avg of std