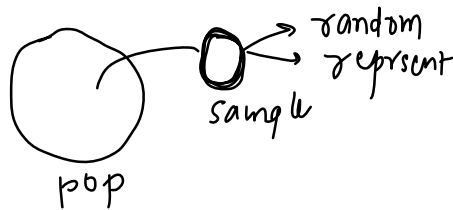


Some Terms

30 March 2023 07:09



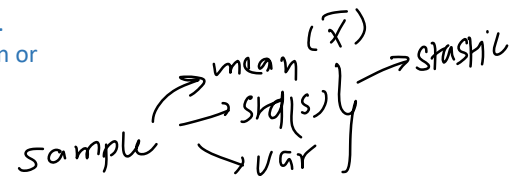
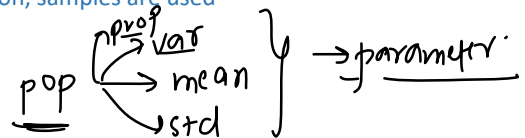
Population Vs Sample

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.

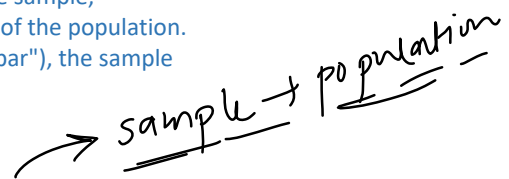
Parameter Vs Estimate

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).



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:

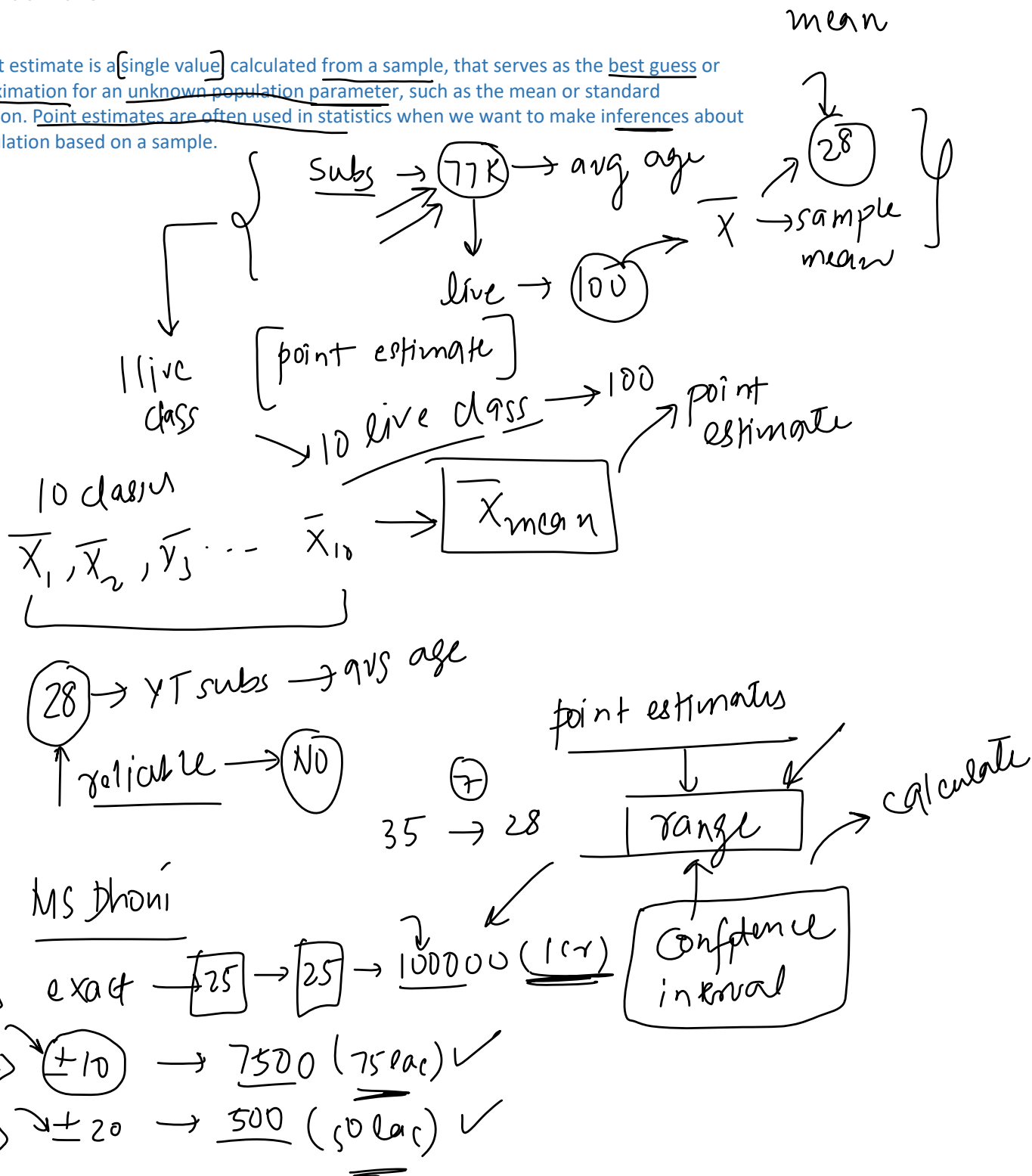
- Is there a significant difference between two groups?
- Can we predict the outcome of a variable based on the values of other variables?
- What is the relationship between two or more variables?

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.

Point Estimate

30 March 2023 07:19

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

$$\mu, \sigma, \bar{X} \rightarrow [25, 32]$$

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.

95% confident

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

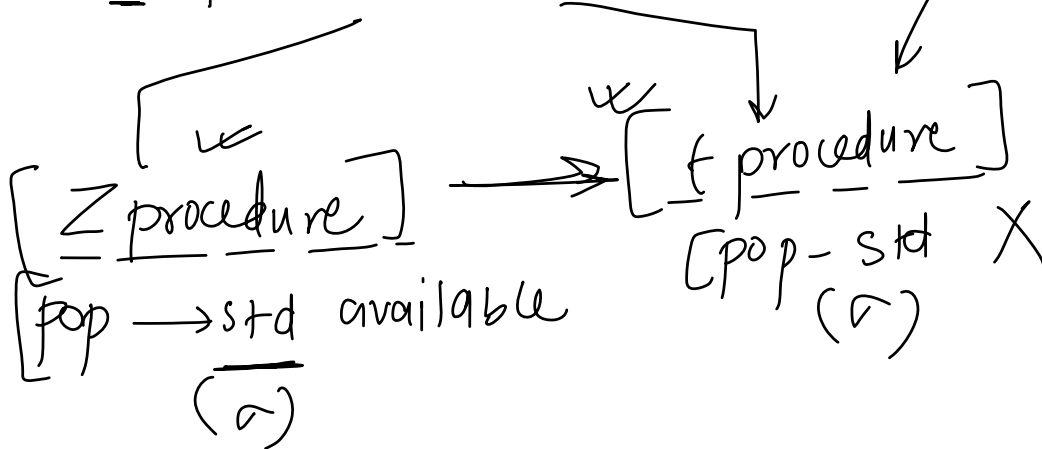
95%

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

$$25 \pm 4$$

$$[21, 29]$$

Ways to calculate CI:

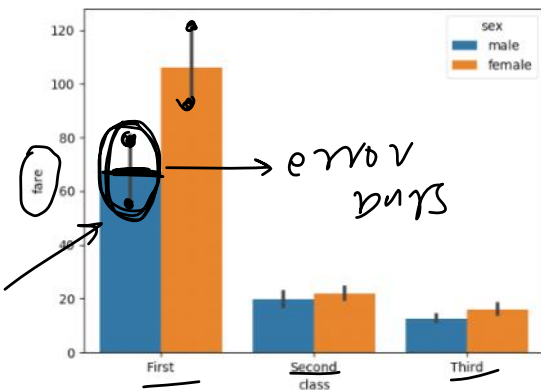


Confidence Interval is created for Parameters and not statistics. Statistics help us get the confidence interval for a parameter.

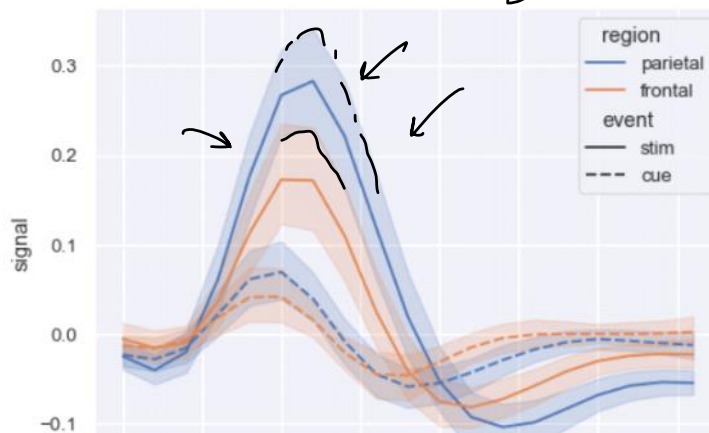
Examples of CI usage

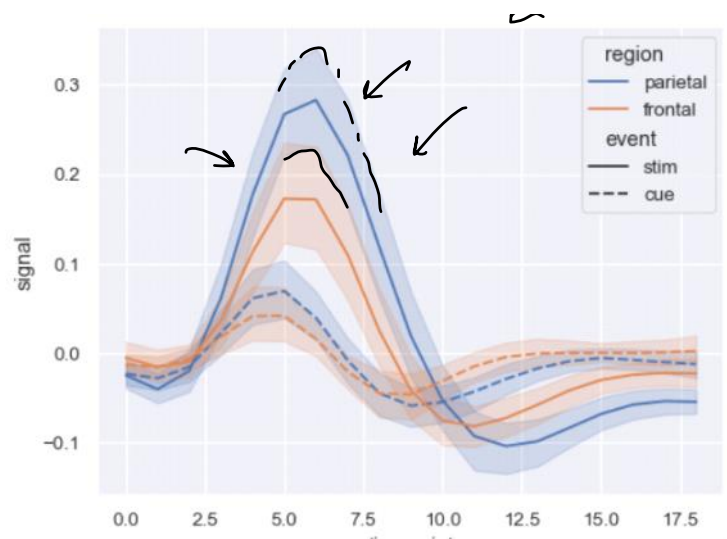
seaborn

bar plot



(CI)





Confidence Interval (Sigma Known) Z procedure

30 March 2023 07:13

(σ) pop std available \rightarrow pop \rightarrow sample random

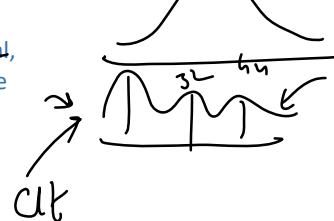
Assumptions

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 $n \geq 30$ 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.

Sample size $n \geq 30 \rightarrow (28) \rightarrow$ Z procedure

YT \rightarrow pop \rightarrow std (σ)
distribution \rightarrow normal

77k subs \rightarrow age
normal dist



A $(1 - \alpha) * 100\%$ Confidence Interval for μ :

$$\sigma = 15$$

YT \rightarrow campus \rightarrow 77k \rightarrow 28 ± 14

$[16, 42]$ \leftarrow confidence interval
Confidence level \rightarrow 95%

formula
CI using
Z procedure

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

1) Intuition
2) $Z_{\alpha/2}$

Z \rightarrow ?

$(1 - \alpha) \rightarrow$ confidence level

$(1 - \alpha) \rightarrow 95\%$

$\sigma \rightarrow$ std pop

$n \rightarrow$ sample size $\rightarrow 100$

Intuition

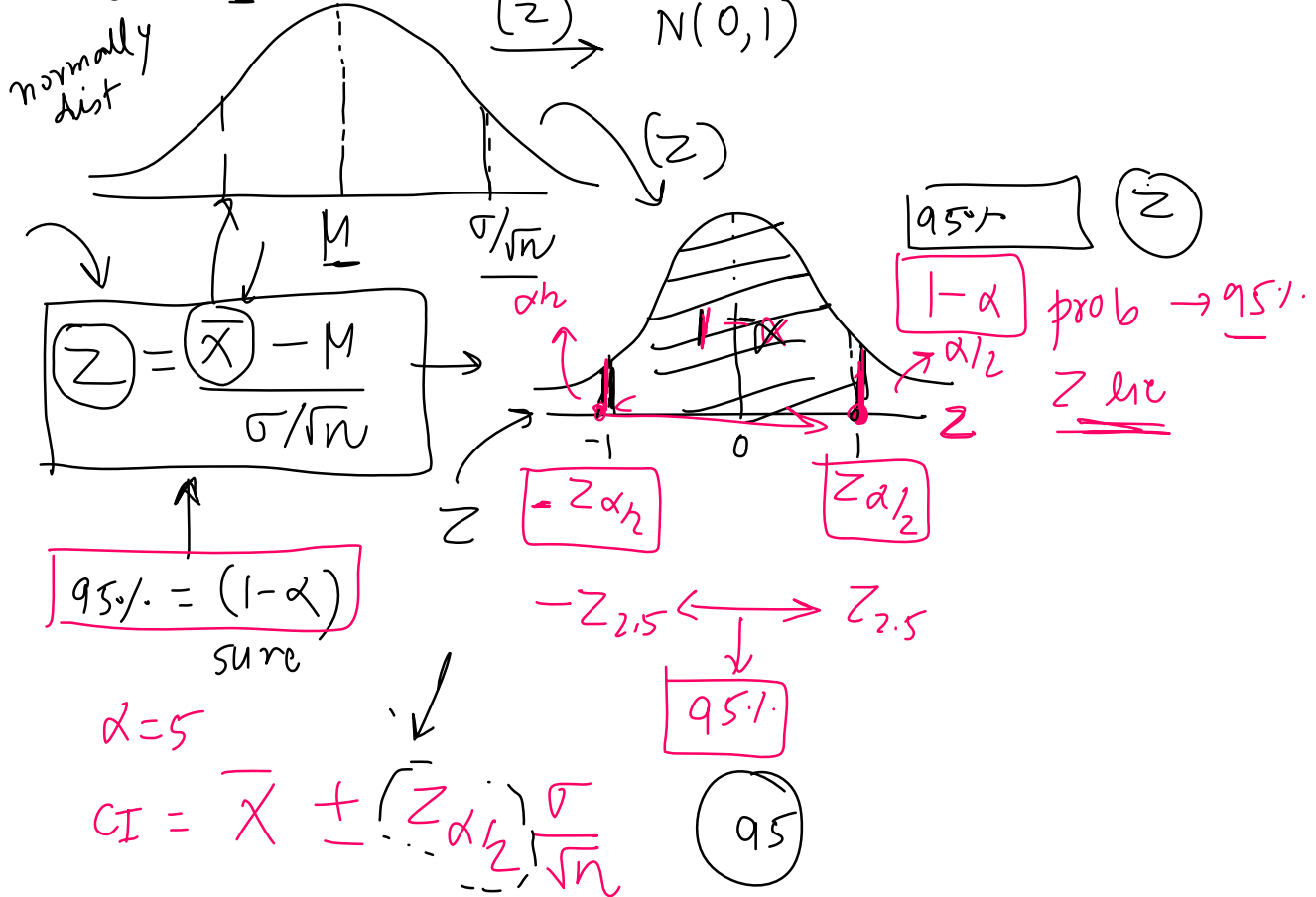
Point estimate (\bar{x}) \rightarrow CLT

point estimate $(\bar{x}) \rightarrow CL1$

10 live class \rightarrow (50) \rightarrow avg class

$$\left[\overline{X_1}, \overline{X_2}, \dots, \overline{X_{10}} \right] \rightarrow \text{sampling dist of sample mean}$$

normally
list

$$\xrightarrow{L(z)} N(0,1)$$


$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$$

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha \quad \textcircled{\mu} \rightarrow \textcircled{CI}$$

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

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

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

$$P\left(\underbrace{\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}_{\bar{X} \rightarrow \text{sample}} < \underbrace{\mu}_{\text{fixed}} < \underbrace{\bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}_{\bar{X} \rightarrow \text{sample}}\right) = \boxed{1-\alpha}$$

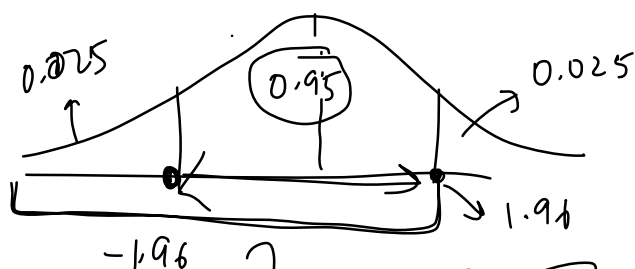
→ 95%
1-α 0.95 α = 0.50
0.250

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

confidence
(1-α) → 95%

$$\mu = \bar{X} \pm z_{0.250} \frac{\sigma}{\sqrt{n}} \quad \leftarrow \textcircled{Z}$$

$$\begin{array}{r} 0.50 \\ 0.25 \\ \hline 0.75 \end{array}$$



range

→ confidence interval

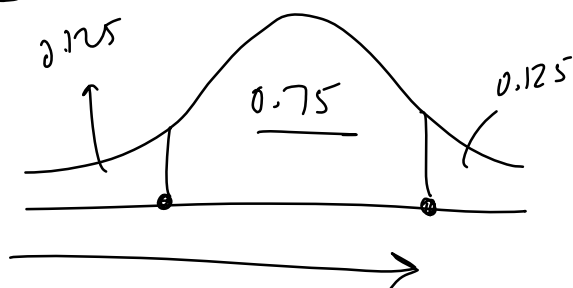
$$\mu = \bar{X} \pm \boxed{1.96} \frac{\sigma}{\sqrt{n}}$$

CI with 95% confidence level

$$z_{\alpha/2} \rightarrow 1.96 \quad 50\% \quad 75\%$$

99%

$$\textcircled{Z_{\alpha/2}} \quad 0.87$$



$$Z \rightarrow 1.13$$

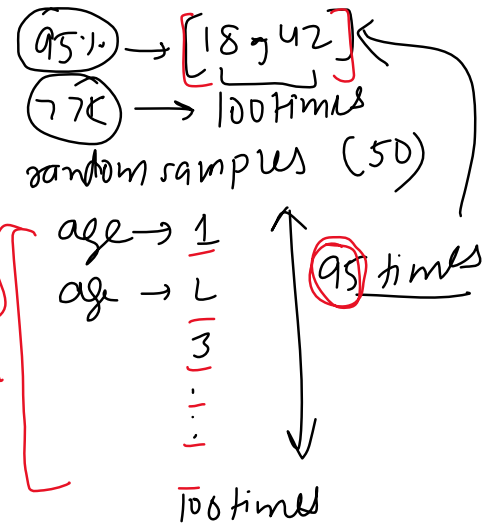
Interpreting Confidence Interval

30 March 2023 08:33

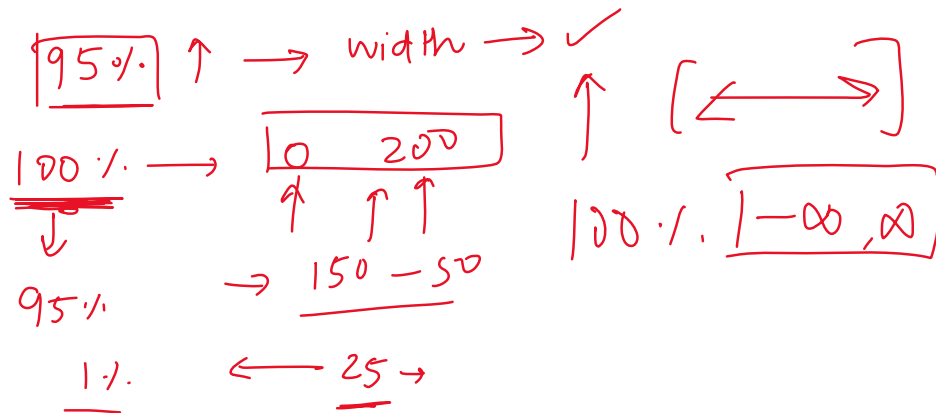
pop fixed $95 \rightarrow [14 - 42] \leftarrow$

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:

- 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.
- 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.
- 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



Factors Affecting Margin of Error

30 March 2023 07:15

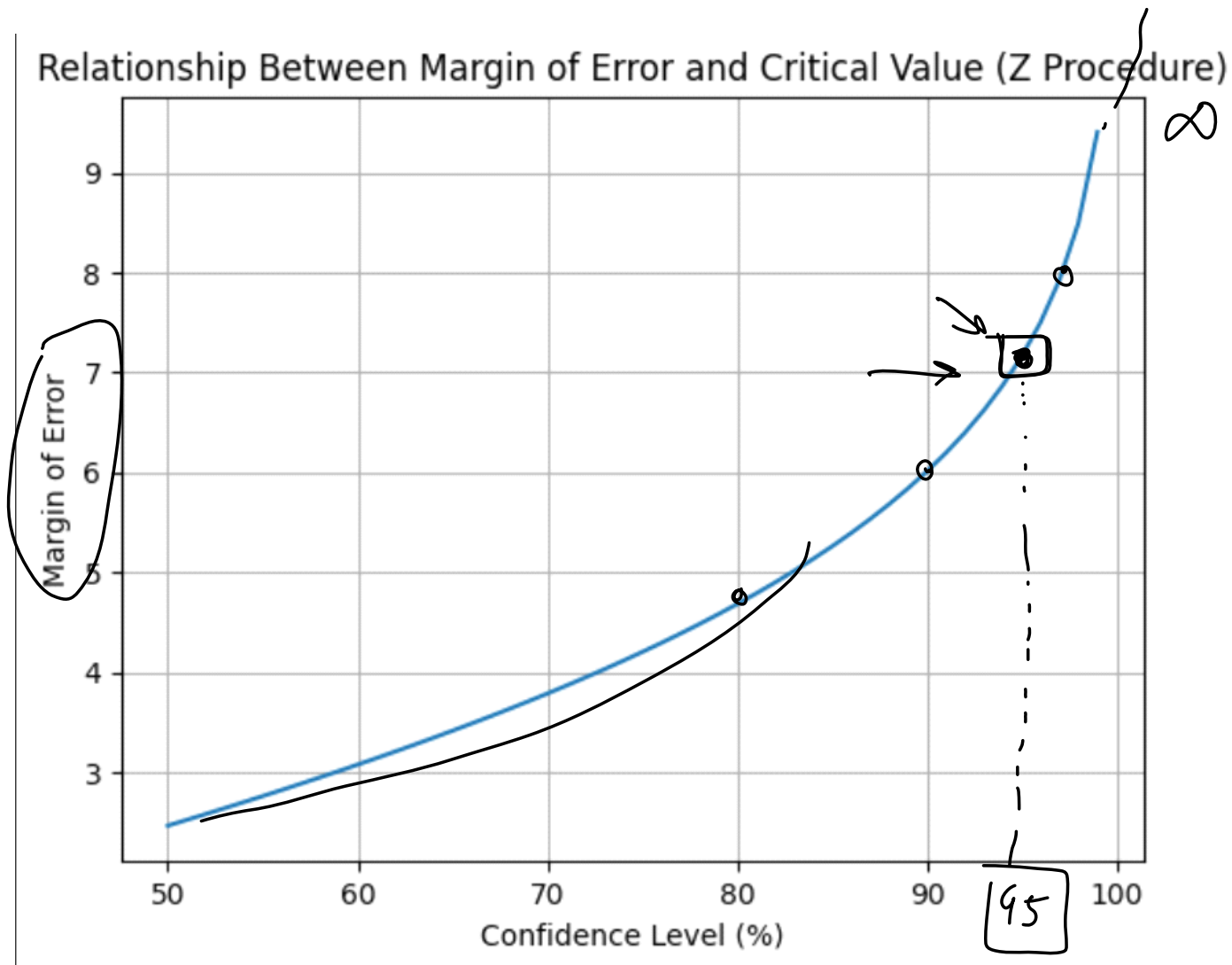
1. Confidence Level (1-alpha)
2. Sample Size
3. Population Standard Deviation

upper - lower \rightarrow margin
 z

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

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

Sample mean $\rightarrow \bar{x}$
 (1) critical value $\rightarrow z_{\alpha/2}$
 (2) pop std $\rightarrow \sigma$
 (3) sample size $\rightarrow n$



30 March 2023 07:15

$\boxed{\tau} \rightarrow \underline{t\text{-product}}$

Assumptions

- and the
ally
ole
- Sample $> 30 \rightarrow$ if it's not
normal
- \rightarrow normal \rightarrow small sample
size \rightarrow CT
- ough for
ved or has
ethods
- ent of
the value of
es data or
- $$\frac{t_{\alpha/2}}{z_{\alpha/2}} \gg z_{\alpha/2}$$

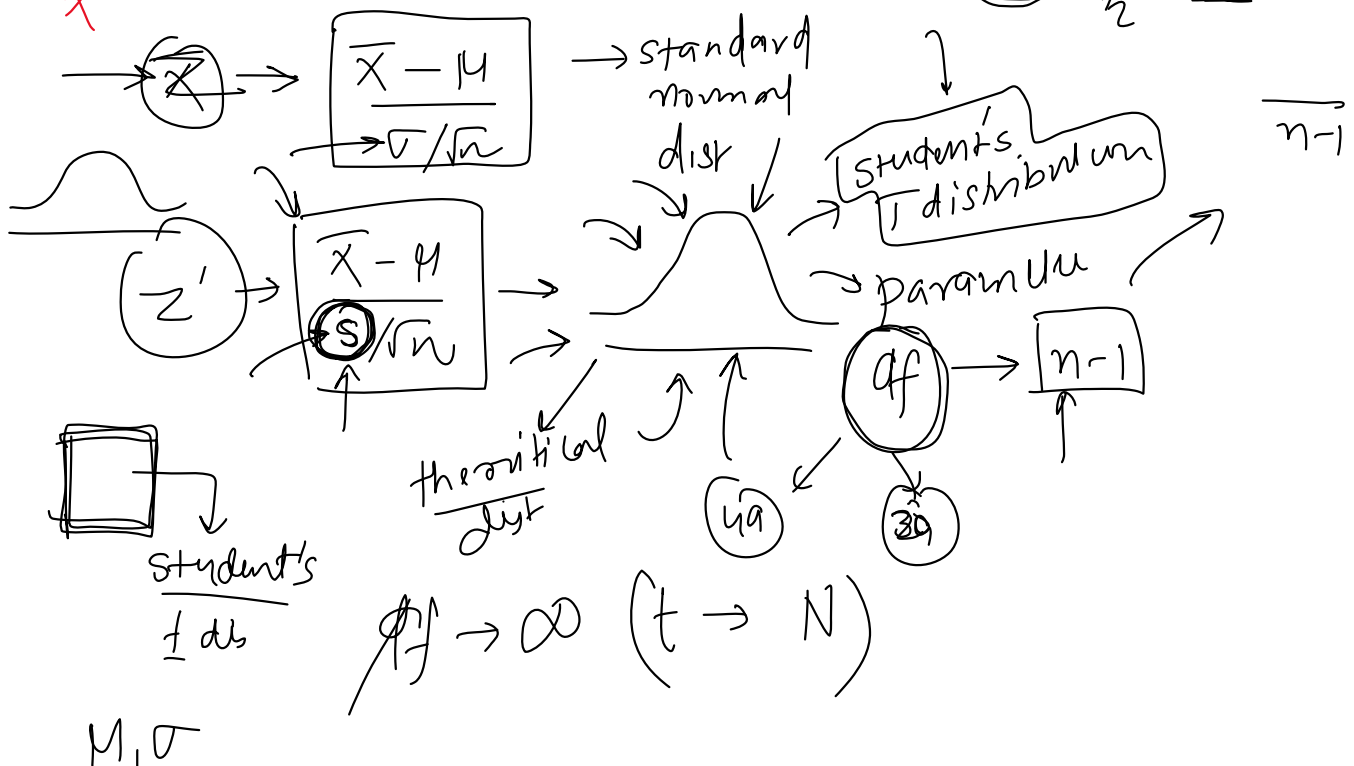
$$\frac{t_{\alpha/2}}{z_{\alpha/2}} \approx z_{\alpha/2}$$
- Z-table

$$CI = \bar{X} \pm z_{\alpha/2} \frac{\boxed{\sqrt{s^2}}}{\sqrt{n}} \rightarrow S$$

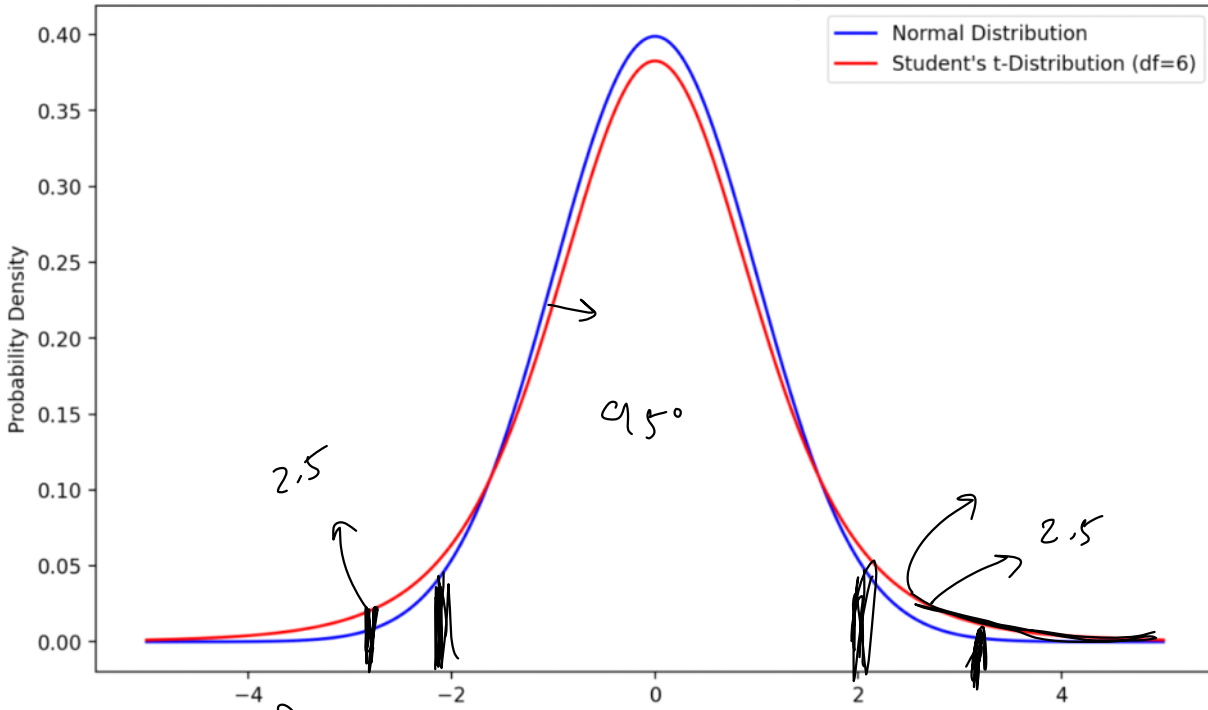
$$CI = \bar{X} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}}$$

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

$$95\% \rightarrow z_{\alpha/2} \rightarrow 1.96 \quad t_{\alpha/2} = 2.0009$$



Normal and t-Distribution Comparison



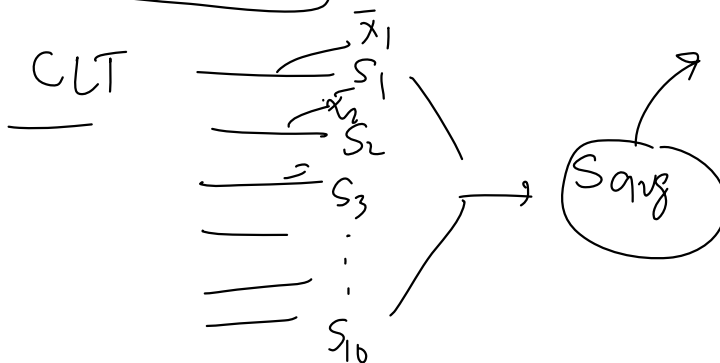
$CI = \bar{X} \pm \frac{Z_{\alpha/2} \cdot \sigma}{\sqrt{n}}$

$= \bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$

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

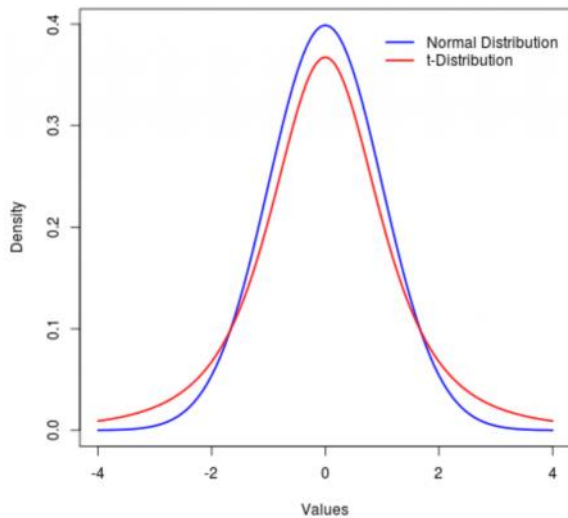
↔

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



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

pop \rightarrow 1360

$\mu \rightarrow x$

$\sigma \rightarrow x$

CLT \rightarrow 10 times \rightarrow 30 size

95% confidence level
inference