Confidence Intervals

Some ~~parameter~~ Terms

1. Population
2. Sample
3. Parameter :-

$$\text{Population} \begin{cases} \rightarrow \text{mean}(\mu) \\ \rightarrow \text{Var} \\ \rightarrow \text{Std}(\sigma) \end{cases} \Big\} \text{ parameter}$$

↑ sigma

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.

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

5. Inferential Statistics

Sample → predict about ~~prop~~ ~~of~~ population.

Inferential statistics → use hypothesis testing, confidence interval and regression analysis, among others.

These methods help researchers answer question like:

a) is there a significant difference betn two group?
b) Can we predict the outcome of a variable based on the value of other values?

c. What is the relationship betn two or more ~~variable~~ variable!

Point Estimate

Subs → 77 k → any age?

or

Live → 100 → $\bar{x}$ → sample mean (point estimate)

↳ viewer

10 live classes

↳ 100 → 10 mean ($\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots \bar{x}_{10}$)

↳ viewer

$\boxed{\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots \bar{x}_{10}}$ → $\bar{x}_{mean}$ (point estimate)

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 inference about a population based on a sample.

Confidence Interval

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.

μ, σ

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

[25, 32] → confidence Interval

95% of data (confidence level) betn confidence Interval.

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

25 ± 4

$25+4 = 29$
$25-4 = 21$

Confidence Interval = [21, 29]

Ways to calculate CI:

- z procedure
  pop → std (available) (σ)
- t procedure
  pop → std (not available) (σ)

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

Example of CI usage

Confidence Interval (Sigma known)
↳ Z procedure

Assumption

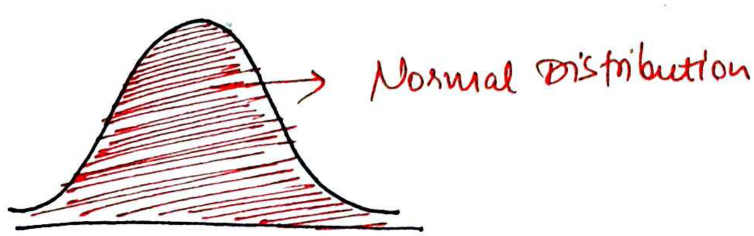
1. Random Sampling

pop ⟶ Sample
↳ Sample must be random.

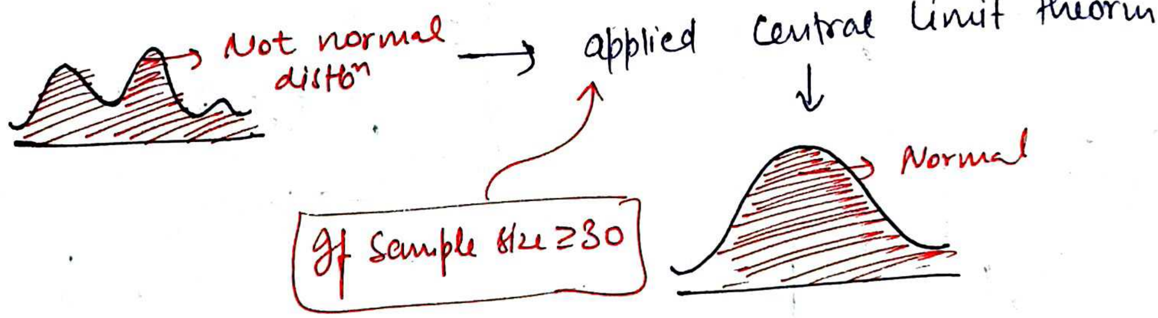
e.g → Collect salary data of all India. So, we have to collect data from Random state not only single state.

2. Known population standard deviation

3. Normal distribution or large sample size



However, if the population distbⁿ is not normal, the Central limit theorm can be applied when the sample size is large (usually, sample size $n \geq 30$ is considered large enough). According to the Central limit Theorm, the sampling distbⁿ of the sample mean will approach a normal distbn as the sample size increase, regardless of the shape of the population distbⁿ.



A (1- alpha) * 100 % confidence interval for mu:

$YT \rightarrow$ Campusx $\rightarrow$ $\boxed{77k} \rightarrow 28 \pm 14$

$[16, 42] \leftarrow$ Confidence interval

$\boxed{\sigma = 15}$

↳ 95% confidence level

formula CI using Z procedure

$$\underset{(\bar{X})}{\text{Point estimate}} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

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

$\sigma \rightarrow$ std pop

$n \rightarrow$ Sample size $\rightarrow 100$

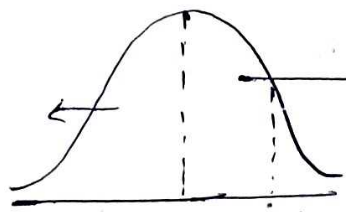
Intuition

Point estimate ($\bar{x}$) → CLT

10 live class → 50 Student → Avg age

$\boxed{\bar{x}_{1\,class}, \bar{x}_{2\,class}, \text{----} \bar{x}_{10\,class}}$ → Sampling dist of Sample mean.

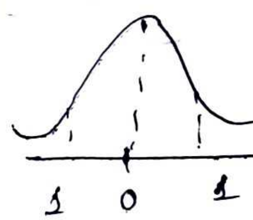
normal dist



→ Convert into Standard Normal variate (Z) $\mu = 0$, $\sigma = 1$

$$\boxed{Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}}$$

→ graph



95% = (1-α)

sure

$1-\alpha = 95\%$ $\quad -Z_{0.025} \longleftrightarrow Z_{0.025}$

95%

$\boxed{\alpha = 5}$

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

95%

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

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

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

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

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

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

$\bar{x} \rightarrow$ Sample

* μ is a fixed (Age is fixed) so, prob is also fixed. ($\bar{x}$) is sample and it is not fixed bcz every time ($\bar{x}$) sample always change. We cannot say " ~~sample~~ μ lie betn $\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $\bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$" prob is 95%".

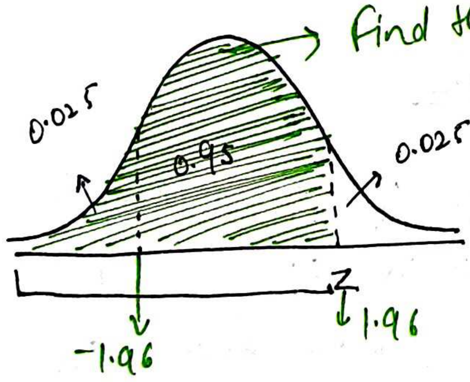
We can say " μ lie between $\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $\bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ confidence level is 95%".

CI $\mu = \bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ $\quad (1-\alpha) \rightarrow 95\%$

$\frac{\alpha}{2} = \frac{0.50}{2}$

$\alpha/2 = 0.250$

$$\boxed{\mu = \bar{x} \pm Z_{0.250}\frac{\sigma}{\sqrt{n}}}$$



Find this area

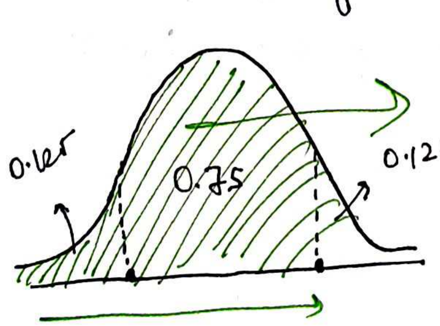
950
025
$\overline{975}$ Find this value on Z table.

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

$\boxed{Z_{\alpha/2} \rightarrow 1.96}$

* If I want to to $Z_{\alpha/2}$ of 75% then what we have to do is.

Area using Z score.



* What is confidence Interval?

→ Confidence Interval is not a Probability to defined interval [14-42] or not a ~~age~~ avg lie beth these [14-42].

real defination

↳ 95% → [18, 42]
  ↓
Channel

77k → 100 times (Random Samples)(50)

avg age of 50 people → 1 time
⋮
avg age of 50 people → 100times

→ 100 time mai se 95 time of avg age [18, 42) ke beech mai ayega

Interpreting Confidence Interval

A confidence interval is a range of values within which a population parameter, such as the pop 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 ~~Interval~~ level**: The confidence level (commonly set at 90%, 95% or 99%) represents the prob that the confidence interval will contain the true pop 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 pop and calculate the confidence interval for each, approximately 95 of those interval would contain the true pop parameter.

2. **Interval range**: The width of the confidence interval gives an indication of the precision of the estimation. A narrower confidence interval suggests a more precise estimate of the pop 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 pop 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 choose when constructing the interval.

Factor Affecting Margin Error

1. Confidence level (1 - alpha). Upr - lower / 2 = margin of error

CI = point estimate ± margin of error

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

Sample mean

Critical ① Value

Pop std ②

Sample size ③.

① confidence level ↓ Margin of error ↓

ex: 50% → [49.09, 50]

95% → [39.01, 55.01]

② Standard deviation ↓ Margin of error ↓

③ Sample size ↑ Margin of error ↓

Confidence Interval (Sigma not known)

Using the t procedure

Assumption

1. Random Sampling: The data must be collected using a random sampling method to ensure that the sample is representative of the pop. This helps to minimize biases and ensures that the results can be generalized to the entire pop.

2. Sample Standard deviation: The pop standard deviation (σ) is unknown, and the sample standard deviation (S) is an estimate. The t-dist is ~~specially~~ specifically designed to account for the additional uncertainty introduced by using the sample standard deviation insted of the pop standard deviation.

3. Approximately normal Distribution: The t-procedure assumes that the underlying pop is approximately normally distributed, or the sample size is larger enough for the central limit Theorm to apply. If the pop distb^n is heavily skewed or has extreme outliers, the t-procedure may not be accurate and non-parametric methods should be considered.

4. Independent Observation: The observations in the sample should independent of each other. In other words, the value of one observation should not influence the value of another observation. This is particularly imp when working with time

Series data or data with inherent dependencies.

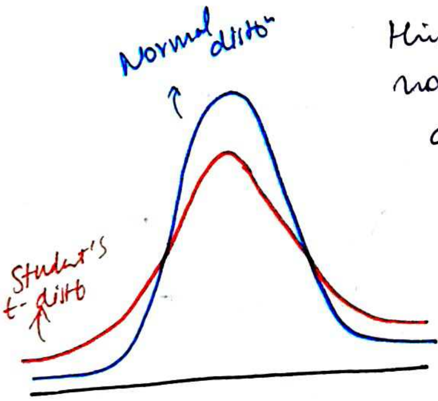
In CI (with sigma) $= \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ $\sigma \xrightarrow{X} S$

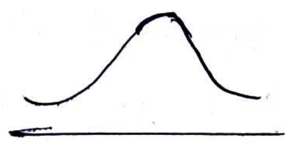
$$CI = \bar{X} + Z_{\alpha/2} \frac{S}{\sqrt{n}}$$ → create complexity

$\bar{X} \rightarrow \boxed{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}} \rightarrow$ std normal dist

$Z' \rightarrow \boxed{\frac{\bar{X}-\mu}{S/\sqrt{n}}} \rightarrow$ Student's t distribution (look like normal distbn but not normal distb)

this is not normal distbn





theoritical distbn (Not exist in nature like Normal distbn)

Only one parameter → degree of freedom (n−1)

degree of freedom − ∞

→ Student's t distbn = Normal distribution

degree of freedom ↑↑ (Student's t distbn ≃ Normal distbn)

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

↓ use t table

95% of $Z_{\alpha/2} \rightarrow 1.96$

95% of $t_{\alpha/2} = 2.0008$

} always $t_{\alpha/2} > Z_{\alpha/2}$ bcz we are not confident so increase the interval for more confidence.