

Confidence Interval: In simple words, It 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 percentage like 95%, indicates how sure we are that, the true values lies within the interval.

Confidence interval = Point estimate \pm margin of error

There are two ways to calculate Confidence Interval \rightarrow

① Z procedure

When you have population std deviation (σ)

② T procedure

When you ^{don't} have population std deviation (σ)

Confidence Intervals always created for population perspective, Not for sample statistics. But we took the help of the sample statistics to create Confidence Interval.

Some conditions to apply Z procedure

- ① Data must be collected using Random sampling method to ensure that the sample is the representative of the population.
- ② Population std deviation must be known.
- ③ Z procedure assumes that 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 (≥ 30).

Problem: CampusX channel has 77K youtube subscribers. Predict the average age of the subscribers? Confidence level should be 95%. Population $\sigma = 50$ and it is normally distributed.

Formula to calculate C.I using Z procedure \rightarrow

\rightarrow Find the point estimate

$$\begin{aligned}\text{Formula: point estimate} &\pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \\ &= \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\end{aligned}$$

Similar problem is noted down on PW-skills section.

Point estimate $\rightarrow \bar{x}$

$$1 - \alpha = 95\%$$

$$\sigma = 50$$

n = sample size

$$= 100$$

What confidence intervals interpret?

confidence level 95%, and interval $[18, 42]$. What it means?

of age from the 77K subscribers (population) if we take 100 random samples, where every sample size (≥ 30). Then from that 100 samples 95 times the average age would be found in this $[18, 42]$ range.

The more the confidence level is, the bigger the confidence interval would be. For example,

Suppose, Dhoni's run prediction for a match in 20.

Confidence level	Run Range (Interval)
100%	$[0, 150]$
95%	$[15, 120]$
80%	$[35, 100]$
1%	$[25]$

What are the various aspects that affect Confidence Interval?

The formula was,

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

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

\nearrow sample mean \downarrow critical value

$\sigma \rightarrow$ std
 $\sqrt{n} \rightarrow$ sample size

→ First of all, confidence level affect confidence interval.

(Generally, 95% confidence level followed by industry standard)

→ The more the std deviation will increase, the ~~more~~^{less} the confidence interval will increase. (Range will be smaller)

→ The more the sample size (n), the more the confidence interval becomes. (Range will be bigger)

T-Procedure (Which actually is more usable than z procedure)

The conditions to apply T-Procedure →

- ① Data must be collected using random sampling
- ② The population std (σ) is unknown and the sample std (s) is used as an estimate.
- ③ T-procedure assumes it's dataset to be normally distributed on the sample sizes are large enough to apply central limit theorem. If population distribution is heavily skewed or has extreme outliers, the t-procedure may not be accurate.
- ④ Every samples should be independent to each other.

For T-procedure formula would be →

$$C.I = \bar{X} + t_{\alpha/2} \frac{s}{\sqrt{n}} \quad [we \text{ will use } t \text{ table instead of } z]$$

More info is noted on pw-skills part. (Please watch the session 44 couple of more time to understand + stats more)

Review → session 44