Confidence Intervals – (Interval Estimation)

Understand Central Limit Theorem

$X_1, X_2, \dots, X_n$ are i.i.d. random variables with mean μ and variance σ^2

$$\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} N(0,1)$$

$x \in \mathbb{R}$,

$$\mathbb{P}\left(\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \leq x\right) \xrightarrow[n \to \infty]{} \mathbb{P}(Z \leq x)$$

where $Z \overset{d}{=} N(0,1)$ random variable.

Proof: OMIT – for this class.

Ex:- Encourage to work out a proof for the case :-

$X_i \overset{d}{=} \text{Bernoulli}(p)$; $0 < p < 1$

& independent

$$\Rightarrow \quad S_n = \sum_{i=1}^{n} X_i \overset{d}{=} \text{Binomial}(n, p)$$

observe :- $$\frac{S_n - n\mu}{\sqrt{n}\,\sigma} = \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right)$$

Write out explicitly : $$\mathbb{P}\left(\frac{S_n - n\mu}{\sqrt{n}\,\sigma} \leq x\right) = \dots$$

and compute its limit $\square$

Confidence Intervals :

Estimate : $-\ \theta$ from a distribution X

— Sample $X_1, X_2, \dots, X_n$ i.i.d X from the population

— Construct :- $g(X_1, X_2, \dots, X_n)$ as an estimate of θ.

Unbiased : $E[g(X_1, X_2, \dots, X_n)] = \theta$

Consistent :- $Var[g(X_1, X_2, X_3, \dots, X_n)] \to 0$ as $n \to \infty$.

Example: $\theta = E[X]$ $\quad \frac{1}{n}\sum_{i=1}^{n} X_i := g(X_1, X_2, \dots, X_n)$

If $Var[X] < \infty$ then $g(\cdot)$ was unbiased and consistent.

Question: Can we understand " $g(X_1, X_2, \dots, X_n) - \theta$ " better ?

Resolve this for $g(X_1, X_2, \dots, X_n) := \overline{X}_n$

Using the Central limit Theorem.

Question: Can we understand " $g(X_1, X_2, \ldots, X_n) - \theta$ " better?

Resolve this for $g(X_1, X_2, \ldots, X_n) := \bar{X}_n$

Using the Central limit Theorem.

Normal Random Variable $\quad Z \stackrel{d}{=} N(0,1)$

(1) — $\mathbb{P}(|Z| \le 1.96) \cong 0.95$

approximation

Central limit Theorem :-

$X_1, X_2, \ldots, X_n$ i.i.d. sample from X

with $E[X] = \mu$ and $\operatorname{var}[X] = \sigma^2$

(2) - $\mathbb{P}\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \le x \right) \cong \mathbb{P}(Z \le x)$

for large n.

From (1) and (2)

(3) — $\mathbb{P}\left(\left| \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \right| \le 1.96 \right) \cong 0.95$

for large n.

(3) $\Longleftrightarrow$

$$\mathbb{P}\left(-\frac{\sigma}{\sqrt{n}} 1.96 \leq \bar{X}_n - \mu \leq 1.96 \frac{\sigma}{\sqrt{n}}\right) \cong 0.95$$

for large n

i.e.

$$\mathbb{P}\left(-\frac{\sigma}{\sqrt{n}} 1.96 + \bar{X}_n \leq \mu \leq \frac{\sigma}{\sqrt{n}} 1.96 + \bar{X}_n\right) \cong 0.95$$

for large n

i.e

$$\mathbb{P}\left(\mu \in \left(-\frac{\sigma}{\sqrt{n}} 1.96 + \bar{X}_n , \frac{\sigma}{\sqrt{n}} 1.96 + \bar{X}_n\right)\right) \cong 0.95$$

for large n

(4) ✓

$$\text{Interval}_n \equiv \left(-\frac{\sigma}{\sqrt{n}} 1.96 + \bar{X}_n , \frac{\sigma}{\sqrt{n}} 1.96 + \bar{X}_n\right)$$

True mean $\equiv \mu$

i.e $\mathbb{P}(\text{True mean} \in \text{Interval}_n) \cong 0.95$

for large n

Instead of $\bar{X}_n \equiv$ Point estimator for μ

we can provide a 95%-Confidence Interval.

Implications of 4 :-

A How large should n be for $\cong$ to take effect ?

[check Earlier worksheet for Berry Essen bounds]

B $$\boxed{\mathbb{P}\left(\mu \in \left(-\frac{\sigma}{\sqrt{n}}1.96+\bar{X}_n, \frac{\sigma}{\sqrt{n}}1.96+\bar{X}_n\right)\right) \cong 0.95 \quad \text{for large } n}$$

4

Does 4 $\Longrightarrow$

- Sample $X_1, X_2, \dots, X_n$ from X

- Compute $\left(-\frac{\sigma}{\sqrt{n}}1.96+\bar{X}_n, \frac{\sigma}{\sqrt{n}}1.96+\bar{X}_n\right) \equiv I_n$

Then there is a 95% chance that $\mu \in I_n$?

NO ! - Be careful in interpretation of 4

Example: $I_n = (-3, 1)$ from one sample.

$I_n = (-5, -4)$ from another sample.

Clearly (4) interpretation above is not correct.

Correct interpretation of (4)

- Experiment:

  - Sample $X_1, X_2, \ldots, X_n$ from X
  - Compute $\left(-\frac{\sigma}{\sqrt{n}} 1.96 + \bar{X}_n, \frac{\sigma}{\sqrt{n}} 1.96 + \bar{X}_n\right) \equiv I_n$

- Repeat Experiment 100 times

let $I_n^{(1)}, I_n^{(2)}, \ldots, I_n^{(100)}$ be the intervals computed in each trial

(4) $\Longrightarrow$ The true mean μ will belong to approximately 95 of the intervals.

Because :- n - large fixed

Set $A = \{\mu \in (-\frac{\sigma}{\sqrt{n}} 1.96 + \bar{X}_n , \frac{\sigma}{\sqrt{n}} 1.96 + \bar{X}_n)\}$

(4) $\Rightarrow$ $\mathbb{P}(A) = 0.95$

$\therefore$ If we perform "m" trials of the experiment and Construct $I^{(1)}, \ldots, I^{(m)}$ then A will occur approximately 95% of the time