Random Sampling: Confidence Interval

Anis REZGUI
Mathematics Department & Computer Science
INSAT - Carthage University

April 19, 2022

Plan

Next Section

Limit Theorems

Random Sample

Let X be a given random variable. A random vector $(X_1, \cdots, X_n)$ is a random sample of X if all X_i's are mutually independent and identically distributed following all the same distribution as X.

1. A given statistical series $\{x_1 \cdots, x_n\}$ can be seen as a realization of the random sample of X, $(X_1, \cdots, X_n)$, this means that:
   - we have realized each X_i apart and we have obtained x_i, and this for all $i = 1, \cdots, n$.
2. Agiven function of $(X_1, \cdots, X_n)$

$$T = \phi(X_1, \cdots, X_n)$$

is called a statistic of X.

Examples of statistics of X

1. The sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

2. The sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2,$$

3. The minimum and the maximum

$$X_{min} = \min_i X_i \quad X_{max} = \max_i X_i.$$

Limit Theorems

SLLN: Strong Law of Large Numbers

Theorem

Suppose X a given random variable such that $\mathbb{E}(|X|) < \infty$. If $\{X_n\}_{n\in\mathbb{N}}$ is a sequence of independent and identically distribution "iid" random variable that follow all the same distribution as X, then

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to+\infty]{a.e} \mathbb{E}(X).$$

Notes

1. The SLLN justify the well known and intuitive point estimate for the expectation of X, $\mathbb{E}(X) = \mu$ by the sample mean $\overline{x}$ of any n independent realizations of X.
2. Unfortunately this estimate is not always enough since it depends too much on sample's variation "the n independent realizations of X"

The Central Limit Theorem (CLT)

Theorem

Suppose X *a given random variable with a finite variance* $\mathbb{V}(X) = \sigma^2 < \infty$*. If* $\{X_n\}_{n\in\mathbb{N}}$ *is a sequence of iid random variables that follow all the same distribution as* X*, then*

$$\frac{\overline{X} - \mathbb{E}(\overline{X})}{\sqrt{\mathbb{V}(\overline{X})}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n\to+\infty]{\mathcal{L}} N(0,1). \tag{1}$$

Remarks

1. The CLT ensure that asymptotically speaking the statistic $\overline{X}$ is normally distributed

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \tag{2}$$

and this, whatever is the initial distribution's nature of X (incredible but true!)

Limit Theorems

Remarks (continued)

2. If $X \sim N(\mu, \sigma)$ we have equality in both equations (3) and (2) and this is without letting n goes to infinity.
3. The CLT gives an idea about how the sample mean statistic $\overline{X}$ approach the expectation μ, actually on can say that the order of this approximation is "most likely" of order $\frac{1}{2}$ i.e

$$\mathbb{P}\left\{|\overline{X} - \mu| \leq \frac{\sigma}{\sqrt{n}}\right\} = \mathbb{P}\left\{-2.57 \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 2.57\right\} = 99\%. \tag{3}$$

Limit Theorems

What to do with this? Is it so important?

- Suppose σ known and we have a number of realizations of X, for instance, $\{x_1, x_2, \cdots, x_{100}\}$ and so a realization of $\overline{X}_{100} = \overline{x}_{100}$
- Equation (3) can be read: in 99% of cases one can assume that

$$\mu \in [\overline{x}_{100} - \frac{\sigma}{\sqrt{100}}, \overline{x}_{100} + \frac{\sigma}{\sqrt{100}}] = [\overline{x}_{100} - \frac{\sigma}{10}, \overline{x}_{100} + \frac{\sigma}{10}] \qquad (4)$$

- Yes indeed, this is actually very important!

Let $(X_1, \cdots, X_n)$ be a random sample of a given X, we define the following statistics

$$S^{*2} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{5}$$

Proposition

1. $\mathbb{E}(S^{*2}) = \sigma^2$
   the statistic S^{*2} is unbiased for σ^2,
2. $\mathbb{V}(S^{*2}) = \frac{\mu_4}{n} - \frac{n-3}{n(n-1)}\mu_2^2$ where μ_k sets for the k^{th} moment of X
   i.e $\mu_k = \mathbb{E}(X^k)$.

Limit Theorems

We apply the SLLN and the CLT to obtain

Theorem

1. $S^{*2} \xrightarrow[n\to+\infty]{a.e} \sigma^2$
2. $\dfrac{(n-1)S^{*2}}{\sigma^2} \xrightarrow[n\to+\infty]{\mathcal{D}} \chi^2_{n-1}$
3. $\dfrac{\overline{X}-\mu}{s^*/\sqrt{n-1}} \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{T}_{n-1}$

Theorem 3 justify the $\dfrac{1}{n-1}$ in formula (5) and especially in implemented formula of the sample variance in all statistical software

$$s^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

Next Section

Confidence Intervals for μ

When σ is known

We use Theorem 2, let $\alpha \in [0,1]$, a confidence interval of level $1-\alpha$ (or of risk α) of the population mean μ is

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

or equivalently

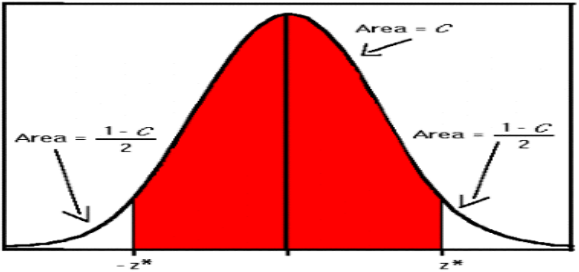
$$\mu \in [\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

where $z_{\alpha/2}$ satisfies

$$\phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$$

and ϕ the normal CDF.

Confidence Intervals for μ



Confidence Intervals for μ

Example

The number of defects in a sample of electric bulbs produced in a given factory is $x = 1\ |0\ |1\ |3\ |2\ |0\ |1\ |2\ |0$.
If we suppose that the population variance is given by $\sigma = 0.1$. An estimation of the defect rate of this factory with a risk level of 5% is given by
$\mu \in [\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}] = [\bar{x} - z_{2.5\%} \frac{0.1}{\sqrt{9}}, \bar{x} + z_{2.5\%} \frac{0.1}{\sqrt{9}}]$
where the sample mean is computed $\overline{x} = 1.11$ and $z_{2.5\%} = 1.96$, and so we get

$$\mu \in [1.04, 1.17].$$

Confidence Intervals for μ

When σ is unknown

We use 3. of Theorem 3, let $\alpha \in [0,1]$, a confidence interval of level $1-\alpha$ (or of risk α) of the population mean μ is

$$\bar{x} - t_{\alpha/2}\frac{s^*}{\sqrt{n-1}} \leq \mu \leq \bar{x} + t_{\alpha/2}\frac{s^*}{\sqrt{n-1}}$$

or equivalently

$$\mu \in [\bar{x} - t_{\alpha/2}\frac{s^*}{\sqrt{n-1}}, \bar{x} + t_{\alpha/2}\frac{s^*}{\sqrt{n-1}}]$$

where $t_{\alpha/2}$ satisfies

$$\phi_t(t_{\alpha/2}) = 1 - \frac{\alpha}{2}$$

and ϕ_t the CDF of t-student distribution.

Confidence Intervals for μ
Example

We reconsider the same example as for the case when σ was known. We need to compute the sample variance that is $s_x = 1.05$. We get the following confidence interval

$$\mu \in [\bar{x} - t_{\alpha/2}\frac{s_x}{\sqrt{n-1}}\,,\,\bar{x} + t_{\alpha/2}\frac{s_x}{\sqrt{n-1}}] \;=\; [1.11 - 2.26\frac{1.05}{\sqrt{8}}\,,\,1.11 + 2.26\frac{1.05}{\sqrt{8}}] \;=\; [0.27, 1.95].$$

Note that when we don't know σ we loose accuracy and this is expected since we do approximate σ by s^*.