Statistical Inference and Multivariate Analysis (MA324)

LECTURE SLIDES
Lecture 22

Interval Estimation

Indian Institute of Technology Guwahati

Jan-May 2023

Interval Estimation

- Now, we assume that the parameter under consideration is a real valued parameter. We are **interested** to **find an interval** in $\Theta \subseteq \mathbb{R}$ such that the interval **covers** the unknown parameter with a **specified probability**.

- Of course, the **interval will be based on a RS**. Interval estimation is quite useful in practice. For example, one may be interested to find a upper limit of mean of toxic level of some drug or food.

Interval Estimation

- Note that for a RV X and two real constants $a > 0$ and $b > 0$,

$$P\left(a < X < b\right) = P\left(X < b < \frac{bX}{a}\right).$$

  Though, these two probabilities are same, there is a basic difference in these two probability statements.

- For the LHS, we are taking about **probability that a random quantity** X belongs to a fixed interval $(a,\, b)$. For the RHS, we are taking about **probability that a random interval** $(X,\, \frac{bX}{a})$ contains a fixed point b.

- For example, let $X \sim U(0,\, 1)$, $a = 0.5$, and $b = 1$. In this case, $P\left(X < 1 < 2X\right) = P(0.5 < X < 1) = 0.5$.

Def: An **interval estimate of a real valued parameter** θ is any pair of functions $L(\boldsymbol{x})$ and $U(\boldsymbol{x})$ of random sample only (do not involve any unknown parameters) that satisfy $L(\boldsymbol{x}) \leq U(\boldsymbol{x})$ for all $\boldsymbol{x}$ in the support of the RS. The **random interval** $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is called an **interval estimator** of θ.

Remark: Though in the definition, the closed interval $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is written, the interval may be **closed, open or semi-open** based on the problem. If $L(\boldsymbol{x}) = -\infty$, then $U(\boldsymbol{x})$ provides an upper limit and $(-\infty, U(\boldsymbol{X}))$ is called upper interval estimator. Similarly, if $U(\boldsymbol{x}) = \infty$, then $L(\boldsymbol{x})$ provides a lower limit, and $(L(\boldsymbol{X}), \infty)$ is called lower interval estimator.

Example 1: Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$. Consider $L_1(\boldsymbol{x}) = x_1 - 1$, $U_1(\boldsymbol{x}) = x_1 + 1$, $L_2(\boldsymbol{x}) = \bar{x} - 1$, and $U_2(\boldsymbol{x}) = \bar{x} + 1$. Then both $[L_1(\boldsymbol{X}), U_1(\boldsymbol{X})]$ and $[L_2(\boldsymbol{X}), U_2(\boldsymbol{X})]$ are interval estimator of μ. Which one should we use? Note that here the lengths of both intervals are same, hence, one should use the interval estimator which has higher probability that the random interval includes μ.

$$P\left(X_1 - 1 \leq \mu \leq X_1 + 1\right) = P\left(-1 \leq X_1 - \mu \leq 1\right) = 2\Phi(1) - 1,$$
$$P\left(\bar{X} - 1 \leq \mu \leq \bar{X} + 1\right) = P\left(-\sqrt{n} \leq \sqrt{n}\left(\bar{X} - \mu\right) \leq \sqrt{n}\right) = 2\Phi(\sqrt{n}) - 1.$$

Now, as $\Phi(\cdot)$ is an increasing function, we should prefer $[L_2(\boldsymbol{X}) = \overline{X} - 1, U_2(\boldsymbol{X}) = \overline{X} + 1]$ over $[L_1(\boldsymbol{X}) = X_1 - 1, U_1(\boldsymbol{X}) = X_1 + 1]$.

Remark: In the previous example, as the **length of the intervals are same**, we **prefer** an interval for which the **probability** that the random interval covers the parameter μ is **highest**. In other cases, we may have interval estimators that have equal probability of covering the parameter. In such cases, we should prefer an interval which has **minimum length**. We will not study such optimality issues in this course.

Def: [**Coverage Probability**] Coverage probability associated with an interval estimator $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ for θ is measured by

$$P_\theta\left(L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})\right).$$

Def: [**Confidence Coefficient**] The confidence coefficient associated with an interval estimator $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is defined by

$$\inf_{\theta \in \Theta} P_\theta\left(L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})\right).$$

Def: [**Confidence Interval**] Let $\alpha \in (0, 1)$. An interval estimator $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is said to be a confidence interval (CI) of level $1-\alpha$ (or a $100(1-\alpha)\%$ confidence interval) if

$$P_\theta\left(L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})\right) \geq 1-\alpha \text{ for all } \theta \in \Theta.$$

Remark: Typical values of α are 0.1, 0.05, 0.01.

Remark: Clearly, we are **loosing precision in interval estimation** compared to point estimation. Do we have **any gain**? Consider the previous example. A reasonable point estimator of μ is $\bar{X}$. However, $P\left(\bar{X} = \mu\right) = 0$ as $\bar{X}$ is a CRV. On the other hand,

$$P\left(\bar{X} - 1 \leq \mu \leq \bar{X} + 1\right) > 0.$$

Hence, in interval estimation we have **some confidence** which we gain by loosing precision.

Interpretation of Confidence Interval

- Let us try to interpret a CI. Let $[L(\boldsymbol{X}),\, U(\boldsymbol{X})]$ be an interval estimator of the parameter θ. Once we **observe** $\boldsymbol{X} = \boldsymbol{x}$, an interval estimate $[L(\boldsymbol{x}),\, U(\boldsymbol{x})]$ is a **fixed interval**.
- Also, recall that the parameter θ is an unknown but fixed entity. Therefore, **no probability is attached** to these observed interval estimate.
- The **interpretation** of the phrase "$(1-\alpha)$ confidence" can be discussed as follows. Suppose that the RS is drawn repeatedly. For the first observation $\boldsymbol{X} = \boldsymbol{x}_1$, the interval estimate is $[L(\boldsymbol{x}_1),\, U(\boldsymbol{x}_1)]$. For the second observation $\boldsymbol{X} = \boldsymbol{x}_2$, the interval estimate is $[L(\boldsymbol{x}_2),\, U(\boldsymbol{x}_2)]$, and so on. If we keep on repeating this procedure, we will have interval estimates

$$[L(\boldsymbol{x}_1),\, U(\boldsymbol{x}_1)]\,,\, [L(\boldsymbol{x}_2),\, U(\boldsymbol{x}_2)]\,,\, [L(\boldsymbol{x}_3),\, U(\boldsymbol{x}_3)]\,,\, [L(\boldsymbol{x}_4),\, U(\boldsymbol{x}_4)]\,,\, \ldots.$$

- In a **long haul**, out of these conceptual interval estimates found, **approximately** $100(1-\alpha)\%$ **would include the unknown value of the parameter** θ. This interpretation goes hand in hand with the relative frequency definition of probability.