Statistical Inference and Multivariate Analysis (MA324)

LECTURE SLIDES

Lecture 09

Point Estimation: Random Sample, Statistic, Point Estimator and Estimate

Indian Institute of Technology Guwahati

Jan-May 2023

Random Sample:

In the standard framework of parametric inference, we start with a data, say $(x_1, x_2, \ldots, x_n)$. Each x_i is an observation on the numerical characteristic under study. There are n observations and n is **fixed, pre-assigned, and known positive integer**.

Our job is to identify (based on a data) the CDF (or equivalently PMF/PDF) of the RV X, which **denote the numerical characteristic** in the population.

Def: [**Random Sample**] The random variables $X_1, X_2, \ldots, X_n$ is said to be a random sample (RS) of size n from the population F if $X_1, X_2, \ldots, X_n$ are ***i.i.d.* random variables with marginal CDF** F. If F has a PMF/PDF f, we will write that $X_1, \ldots, X_n$ is a RS from a PMF/PDF f.

JCDF :

Let X_i denote the ith observation for $i = 1, 2, \ldots, n$, where n is the **sample size**. Then, a meaningful assumption is that each X_i has same CDF F, as X_i is a copy of X. Now, if we can ensure that the observation are taken such a way that the **value of one does not effect the others**, then we can assume that $X_1, X_2, \ldots, X_n$ are **independent**. Thus, a RS can be used to model the situation.

Note that JCDF of a RS $X_1, \ldots, X_n$ is

$$F_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F(x_i).$$

Similarly, JPMF/JPDF of a RS $X_1, \ldots, X_n$ from PMF/PDF f is

$$f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i).$$

- In a typical problem of parametric inference, we further assume that the **functional form** of the CDF/PMF/PDF of RV X is **known**, but the CDF/PMF/PDF **involves unknown but fixed real or vector valued parameter** $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$.
- Thus, if the value of $\boldsymbol{\theta}$ is known, the stochastic properties of the numerical characteristic is completely known. Therefore, our aim is to find the value of $\boldsymbol{\theta}$ or a function of $\boldsymbol{\theta}$.
- We also assume that the possible values of $\boldsymbol{\theta}$ belong to a set Θ, which is called ***parametric space***.
- Here, $\boldsymbol{\theta}$ is an indexing or a labelling parameter. We say that $\boldsymbol{\theta}$ is an ***indexing or a labelling parameter*** if the CDF/PMF/PDF is uniquely specified by $\boldsymbol{\theta}$.
- That means that $F(x, \boldsymbol{\theta}_1) = F(x, \boldsymbol{\theta}_2)$ for all $x \in \mathbb{R}$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, where $F(\cdot, \boldsymbol{\theta})$ is the CDF of X.

AIM:

- As discussed, our **main aim is to identify the CDF/PMF/PDF** of the RV X based on a RS.
- In other words, we want to **identify which member of the family** $\{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ can **represent** the CDF of X, which is equivalent to decide the value of $\boldsymbol{\theta}$ in Θ based on a realization of a RS.
- Note that, as we **know the functional form** of the CDF of X, the value of $\boldsymbol{\theta} \in \Theta$ **completely specifies** the member in $\{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$.
- Here, it is **assumed** implicitly that the data **has information** regarding the **unknown parameter**.

Def: [**Statistic**] Let $X_1, \ldots, X_n$ be a RS. Let $T(x_1, \ldots, x_n)$ be a real-valued function having domain that includes the sample space, χ^n, of $X_1, X_2, \ldots, X_n$. Then the RV $\boldsymbol{Y} = T(X_1, \ldots, X_n)$ is called a **statistic** if it is **not a function of unknown parameters**.

- Note that our aim is to **find a guess value of unknown parameters** based on a RS. Hence, we are considering a function of RS. If the function involve any unknown parameters, we will not be able to compute the value of the function given a realization of a RS.

- Hence, the function that involves unknown parameters is of no use in this respect.

- Therefore, we define a **statistic as a function of RS, but statistic should not involve** an unknown parameter. Note that the **distribution of a statistic may depend** on unknown parameters.

Example 1: Let $X_1, \ldots, X_n$ be a RS from a $N(\mu, \sigma^2)$ distribution, where $\mu \in \mathbb{R}$ and $\sigma > 0$ are both unknown. Then $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ are examples of statistic. However, $\frac{\overline{X}-\mu}{\sigma}$ is not a statistic. Note that $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$. Clearly, the distribution of $\overline{X}$ depends on the unknown parameters.

Point Estimator and Estimate:

Def: [**Point Estimator and Estimate**] In the context of estimation, **a statistic** is called a **point estimator** (or simply estimator). A **realization** of a point estimator is called an **estimate**.

- In the above definition of an estimator, we do not mention about the parameter that is to be estimated and its parametric space.
- However, in practice, we need to take care of the **parameter to be estimated** and its **parametric space**. For **example**, to estimate population variance, we should not use an estimator that can be negative.
- There are **several methods** to **find an estimator**. We will consider three of them in this course: 1) **method of moment estimator** (MME), 2) **maximum likelihood estimator** (MLE) and 3) **least square estimator** (LSE). We will study the first two methods in this estimation and the third method will be discussed when we will study regression.
- Before discussing the methods of estimation, we will study **sufficiency, information, ancillary, and completeness**. These are useful concepts for the theory of estimation.