Statistics II: Introduction to Inference

Week 4: Methods of Point Estimation

In the previous modules, we have talked about properties of *good* estimators. Next, we will discuss some common methods of finding point estimators of θ.

Example 1. Let $X_1, \ldots, X_n$ be a random sample from `Gamma`(α, β) distribution with the common pdf of the form

$$f_{\boldsymbol{\theta}}(x)\frac{\beta^{\alpha}}{\Gamma\alpha}x^{\alpha-1}\exp\{-\beta x\}, \qquad x > 0, \quad \alpha, \beta > 0.$$

We are interested in estimating the parameter $\boldsymbol{\theta} = (\alpha, \beta)$.

The two most popular frequentist methods of finding estimators are:
(A) method of moments (MoM), and
(B) maximum likelihood (ML) estimation.
We will now discuss these methods in details.

1 Method of Moments (MoM)

One of the simplest and oldest methods of finding estimators is the method of moments or substitution principle.

Definition 1 (Population and sample raw moments)**.** *Let $X_1, \ldots, X_n$ be a random sample from a population with distribution function $\{F_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$. Then the r-th order population moment is defined as*

$$\mu'_r = E\left(X_1^r\right); \qquad r = 0, 1, 2, \ldots.$$

Further, the r-th order sample raw moment is

$$M'_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r; \qquad r = 0, 1, 2, \ldots.$$

Method of moments (MoM) estimators are found by equating the sample moments to the population moments, and solving the resulting system of equations.

MoM Method. Let $X_1, \cdots, X_n$ be random sample of size n from $F_{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ be a k-dimensional parameter. Then usually we require a system of k-equations to get an estimate of $\boldsymbol{\theta}$. We obtain estimates of $\boldsymbol{\theta}$ by solving the following equations:

$$M'_r = \mu'_r = \mu'_r(\boldsymbol{\theta}), \qquad r = 1, \ldots, k. \tag{1}$$

Note here that, for each r, the r-th order population raw moment of $F_{\boldsymbol{\theta}}$, μ'_r, is a function of $\boldsymbol{\theta}$. The solution of (1) w.r.t. $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}_{\text{MoM}}$, is the MoM estimator of $\boldsymbol{\theta}$.

Further, if the parameter of interest is $\psi(\boldsymbol{\theta})$, then the MoM estimator of the same is

$$\widehat{\psi}_{\text{MoM}} = \psi(\widehat{\boldsymbol{\theta}}_{\text{MoM}}).$$

Recall the above example. Here $\boldsymbol{\theta} = (\alpha, \beta)$, and so $k = 2$. To find the MoM estimator of $\boldsymbol{\theta}$ can be obtained by solving the equations

$$M'_1 = \bar{X}_n = \frac{\alpha}{\beta}, \qquad \text{and} \qquad M'_2 = \frac{\alpha}{\beta^2}(1+\alpha).$$

Thus, the MoM solutions are

$$\widehat{\alpha}_{\text{MoM}} = \frac{\bar{X}_n^2}{S_n^2}; \qquad \text{and} \qquad \widehat{\beta}_{\text{MoM}} = \frac{\bar{X}_n}{S_n^2},$$

where S_n^2 is the sample variance.

Example 2. Let $X_1, \ldots, X_n$ be a random sample from `Normal`(μ, σ^2). The MoM estimators of μ and σ^2 are $\bar{X}$ and S^2 respectively.

Remark 1. *The r-th order sample raw moment, M_r' is a* good *estimator of the r-th order population raw moment μ_r'. It is unbiased (WHY?) and consistent (consistency will be discussed later).*

Remark 2. *However, MoM may lead to estimators having sub-optimal sampling properties, and may lead to absurd estimators in some cases.*

Example 3. Let $X_1, \ldots, X_n$ be a random sample from `Uniform`$(0, \theta)$. The MoM estimator of θ is $\widehat{\theta}_{\text{MoM}} = 2\bar{X}_n$. For a particular realization with $n = 5$ can be

0.8850 4.5775 3.2539 9.6740 5.6156

Here $\widehat{\theta}_{\text{MoM}} = 9.6024$. But the maximum realization $\max\{x_1, \ldots, x_5\} = 9.6740 > \widehat{\theta}_{\text{MoM}}$. Thus, MoM estimate is absurd.

2 Maximum Likelihood (ML) Estimation

Example 4. Suppose in a shop there are three types of mobile phones with battery lives 2 years (*Low*), 3 years (*Medium*) and 6 years (*Hi*), respectively. A buys a mobile from the shop and it lasts for 4.5 years. What type of mobile phone did A buy?

To answer this question one has to model the data first. Let X be the battery life of a mobile phone with average battery life θ. We model $X \sim$ `exponential`(θ) with pdf

$$f_\theta(x) = \frac{1}{\theta} e^{-x/\theta}; \qquad x > 0,\ \theta > 0.$$

By the problem θ can be either 2, 3 or 6 for the mobile types *Low*, *Medium* and *Hi*; i.e., $\theta \in \Theta = \{2, 3, 6\}$, and the observed $x = 4.5$.

Given this model, one way to estimate θ would be to see under which value of θ the observation $x = 4.5$ is most likely. This can be obtained by inserting the value of θ in the pdf evaluated at 4.5. Observe that

$$f_2(4.5) = 0.0527, \quad f_3(4.5) = 0.0744, \quad \text{and} \quad f_6(4.5) = 0.0787.$$

Therefore, 6 years (*Hi*) is the most likely value of $\theta \in \Theta$.

Example 5. Next, we generalize the problem. Suppose A buys 5 mobile phones of the same type, and they last for 1 year, 2.5 years, 6 years, 3.2 years and 4 years. If we assume that the distribution of life-time of the mobile phones independently follow an `exponential` distribution with pdf f_θ given by

$$f_\theta(x) = \theta \exp\{-\theta x\}; \qquad x > 0, \quad \theta > 0,$$

then how do we estimate θ? If we use the procedure followed in Example 4, then we need to find the $\theta \in \Theta = [0, \infty)$ for which the joint likelihood evaluated that the observations is maximized, i.e., we estimate θ by

$$\widehat{\theta} = \text{argmax}_\theta \theta^5 \exp\left\{-\theta \sum_{i=1}^{5} x_i\right\},$$

where $\sum_{i=1}^5 x_i = 16.7$. One can easily verify that $\widehat{\theta} = 5/\sum_{i=1}^5 x_i = 0.2994$.

This method of estimating θ is called maximum likelihood estimator. We describe the method in details below.

Definition 2 (Likelihood Function)**.** *Let* $\mathbf{X} = (X_1, \ldots, X_n)'$ *be a random sample from a population with distribution function* $\{F_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$*. Suppose the distribution* $\{F_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$ *is characterized by a pdf (or, pmf)* $f_{\boldsymbol{\theta}}(\cdot)$*. Further, suppose* $\mathbf{x}$ *is a realization of* $\mathbf{X}$*. Then the function of* $\boldsymbol{\theta}$ *defined as* $L(\boldsymbol{\theta}; \mathbf{x}) = f_{\boldsymbol{\theta}}(\mathbf{x})$ *is called the likelihood function.*

Definition 3 (Maximum Likelihood Estimate, MLE)**.** *Given a realization* $\mathbf{x}$*, let* $\widehat{\boldsymbol{\theta}}$ *be the value in* Θ *that maximizes the likelihood function* $L(\boldsymbol{\theta}; \mathbf{x})$ *with respect to* $\boldsymbol{\theta}$*, then* $\widehat{\boldsymbol{\theta}}$ *is called the MLE of the parameter* $\boldsymbol{\theta}$*.*

Note that, the maximizer $\widehat{\boldsymbol{\theta}}$ is nothing but a function of the realization $\mathbf{x}$. Thus we can treat the maximizer of the likelihood function as a statistic or estimator of $\boldsymbol{\theta}$. This estimator is called Maximum Likelihood (ML) estimator. Notationally, we write $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_{\mathrm{ML}}(\mathbf{X})$.

Example 6. Suppose there are n tosses of a coin, and we do not know the value of n, or the probability of head (p). However, we know that n is between 3 to 5 and one of the sides of the coin is twice as heavy as the other (i.e., either $p = 2(1-p)$ or $(1-p) = 2p$). Then what is the MLE of $\boldsymbol{\theta} = (n, p)$ given we observe x heads, $x = 1, \ldots, 5$?

Remark 3. *If the likelihood function is differentiable with respect to* $\boldsymbol{\theta}$*, then one may take the differentiation approach for finding the MLE. In case of a several value function* $L(\boldsymbol{\theta}; \mathbf{X})$*, if the function is twice continuously differentiable with respect to each* θ_j*, then a critical point of* $L(\boldsymbol{\theta}; \mathbf{X})$ *can be obtained by equating* $\frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{X}) = \mathbf{0}$*. Then to verify, if the critical point is a maximizer, one can check if the Hessian matrix* $\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} L(\boldsymbol{\theta}; \mathbf{X})$ *is negative definite at the critical point.*

Remark 4. *Often it is convenient to work with the log likelihood function, instead of the likelihood function. As logarithm is a monotone function, the maximizer of likelihood and the log likelihood are the same. The log likelihood is generally denoted by* $l(\boldsymbol{\theta}; \mathbf{X})$*.*

Example 7. Let $X_1, \ldots, X_n$ be a random sample from `normal`(μ, σ^2). Then the MLE of μ and σ^2 are $\bar{X}_n$ and S_n^2, respectively.

Example 8. Let $X_1, \ldots, X_n$ be a random sample from `uniform`(α, β). Then the MLE of α and β are $X_{(1)}$ and $X_{(n)}$, respectively.

Remark 5. *Maximum likelihood estimate may not exist. See the following example.*

Example 9. Let X_1, X_2 be a random sample from `Bernoulli`(θ), $\theta \in (0, 1)$. Suppose the realization $(0, 0)$ is observed. Then the MLE does not exist.

Remark 6. *Even if MLE exists, it may not be unique. See the following example.*

Example 10. Let $X_1, \ldots, X_n$ be a random sample from `Double Exponential`(θ, σ) distribution, $\theta \in \mathbb{R}$. Then the $\widehat{\theta}_{\mathrm{ML}}$ is the median $X_1, \ldots, X_n$, which is not unique.

Remark 7. *The method of maximum likelihood estimation may produce an absurd (not meaningful) estimator.*

In spite of all the above shortcomings, MLE is by far the most popular and reasonable frequentist method of estimation. The reason is that MLE possesses a list of desirable properties. We will discuss some of them below.

Theorem 1 (Properties of MLE: 1)**.** *Suppose the regularity conditions of CRLB are satisfied, the log-likelihood is twice differentiable, and there exists an unbiased estimator* $\widehat{\theta}^{\star}$ *of* θ*, the variance of which attains the CRLB. Suppose further that the likelihood equation has a unique maximizer* $\widehat{\theta}_{\mathrm{ML}}(\mathbf{X})$*, then* $\widehat{\theta}^{\star} = \widehat{\theta}_{\mathrm{ML}}(\mathbf{X})$*.*

Corollary 1. *Theorem 1 implies that if the CRLB is attained by any estimator, then it must be an MLE. However, the converse is not true, i.e., the variance of an MLE may not attain the CRLB.*

2.1 Invariance Property

Let $\boldsymbol{\eta} := \Psi(\boldsymbol{\theta})$ be any function of $\boldsymbol{\theta}$, and we are interested in finding the MLE of $\boldsymbol{\eta}$ given a sample $X_1, \cdots, X_n$. The following theorem states that $\widehat{\boldsymbol{\eta}}_{\mathrm{ML}} = \Psi(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})$ for any function Ψ.

Theorem 2 (Properties of MLE: 2, Invariance Property)**.** *Let $\{f_{\boldsymbol{\theta}}(\cdot) : \boldsymbol{\theta} \in \Theta\}$ be a family of PDFs (PMFs), and let $L(\boldsymbol{\theta}; \mathbf{X})$ be the likelihood function. Suppose $\Theta \subseteq \mathbb{R}^k, k \geq 1$. Let $\Psi : \Theta \to \Lambda$ be a mapping of Θ onto Λ, where $\Lambda \subseteq \mathbb{R}^p$ $(1 \leq p \leq k)$. If $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}(\mathbf{X})$ is an MLE of $\boldsymbol{\theta}$, then $\Psi\left(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}(\mathbf{X})\right)$ is an MLE of $\Psi(\boldsymbol{\theta})$.*

Example 11. Let $X_1, \ldots, X_n$ be a random sample from `Gamma`$(1, \theta)$ distribution, $\theta > 0$. Find an MLE of θ.

Example 12. Let $X_1, \ldots, X_n$ be a random sample from `Poisson`(θ) distribution, $\theta > 0$. Find an MLE of $P(X = 1) = \exp\{-\theta\}$.