Lecture 17: Point Estimation

March 10, 2016

In this lecture , we will start our discussion on methods for finding estimators, and evaluating these estimators.

1 Introduction:

In previous lectures, we studied the problem of detection, which is nothing but deciding between two (or more) different hypothesis. In estimation, we estimate or guess the value of an unknown (point). This unknown need not to be a real number value, it can also be a vector or may take a range of interval. In this lecture, we will focus on *point estimation*. Our purpose is to estimate a point, which will yield to the knowledge of entire population. Following is the definition of a *point estimator*.

Definition 1.1. A *point estimator* is any function $W(X_1, X_2, ..., X_n)$ of samples. That means, any statistic(like sample mean, sample variance) is a point estimator.

2 Methods of Constructing Estimators:

In some cases our intuition lead us to a very good estimator. For example, estimating a parameter with it's sample analogue is usually reasonable. Like, the sample mean is a good estimator for the population mean, but this is not the case always. Sometimes our intuition lead us to a very bad estimator that seem to be correct. So we need a more methodical way of estimating parameters. Following are some methods of finding estimators.

2.1 Method of Moments

The method of moments is , perhaps, the oldest method of finding point estimators. It is quite simple to use and almost always yields some sort of estimate.

Let $X_1, X_2, \ldots, X_n$ be a sample from a population with pdf or pmf $f(x|\theta_1, \ldots, \theta_k)$. Methods of moment estimators are found by equating the first k sample moments to the corresponding k population moments, and solving the resulting system of simultaneous equations. More precisely, define

$$m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j, \quad \mu_j'. \qquad \text{for } j = 1, 2, ..., k \tag{1}$$

The population moment μ_j' will typically be a function of $\theta_1, ..., \theta_k$, say $\mu_j'(\theta_1, \ldots, \theta_k)$. The method of moments estimator $(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$ of $\theta_1, \ldots, \theta_k$ is obtained by solving the following system of equations for $(\theta_1, \ldots, \theta_k)$ in terms of $m_1, \ldots, m_k$:

$$m_j = \mu_j'(\theta_1, \ldots, \theta_k). \qquad \text{for } j = 1, 2, ..., k \tag{2}$$

Definitely, there will be some error in our estimate of the parameter but as we will increase the number of samples, the error will reduce.

Example 2.1. (Normal method of moments)

Let samples $X_1, X_2, \ldots, X_n$ are independent and Gaussian distributed with mean θ and variance σ^2. So our parameters for estimation are, $\theta_1 = \theta$ and $\theta_2 = \sigma^2$. We have $m_1 = \bar{X}, m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \mu_1' = \theta, \mu_2' = \theta^2 + \sigma^2$, and hence we must solve

$$\bar{X} = \theta, \ \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \theta^2 + \sigma^2. \tag{3}$$

Solving for θ and σ^2 yields the method of moments estimators

$$\tilde{\theta} = \bar{X}, \quad \tilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2. \tag{4}$$

Example 2.2. (Binomial method of moments)

Let samples $X_1, X_2, \ldots, X_n$ are independent and binomial distributed with parameters(k, p), that is,

$$P(X_i = x|k, p) = \binom{k}{x} p^x (1-p)^{k-x}, \quad x = 0, 1, ..., k. \tag{5}$$

Here we want the point estimator for both unknown parameters k and p. Equating the first two sample moments to their corresponding population moments yields the system of equations

$$\bar{X} = kp, \tag{6}$$

$$\frac{1}{n}\sum_{i=1}^{n} {X_i}^2 = kp(1-p) + k^2p^2. \tag{7}$$

Now we can solve it for k and p. Substituting value of kp from eqn. (6) in eqn. (7), we get,

$$\frac{1}{n}\sum_{i=1}^{n} {X_i}^2 = \bar{X}(1-p) + \bar{X}^2, \tag{8}$$

and the estimates of p and k as,

$$\tilde{p} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{\bar{X}}, \quad \tilde{k} = \frac{\bar{X}^2}{\bar{X} - \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}. \tag{9}$$

By observing $\tilde{p}$ and $\tilde{k}$, we can say that it is possible to get the negative estimates of p and k which, of course, must be positive numbers, Which implies that it is not necessary to coincide the range of estimator to the range of parameter it is estimating. However, we may reduce the probability of occurance of such event by taking large number of observables.

2.2 Maximum Likelihood Estimators

Let $X_1, X_2, \ldots, X_n$ are an iid sample from a population with pdf or pmf $f(x|\theta_1, \ldots, \theta_k)$, the likelihood function is defined by

$$L(\mathrm{X}|\theta) = L(x_1, \ldots, x_n|\theta_1, ..., \theta_k) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k), \quad \theta \in \Theta \subseteq \mathbb{R}^k \tag{10}$$

Definition 2.3. Given observations $x_1, x_2, ..., x_n$, a maximum likelihood estimate of θ is an element of $\arg\max_{\theta\in\Theta} L_\theta(x)$.

Likelihood of $x_1, x_2, ..., x_n$ under $p(x_1, ..., x_n|\theta)$ is,

$$L_\theta(x) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{11}$$

Remark 1. By it's construction, the range of the MLE (maximum likelihood estimate) coincides with the range of the parameter.

Remark 2. MLE is the solution to an optimizing problem, i.e. an optimizer.

Example 2.4. Normal likelihood

Let samples $X_1, X_2, \ldots, X_n$ are independent and Gaussian distributed with mean θ and variance σ^2. We want to get an estimate of θ and σ^2. For that, we

need to solve:

$$\arg\max_{(\theta,\sigma^2)} \prod_{i=1}^{n} f(x_i|\theta,\sigma^2) = \arg\max \sum_{i=1}^{n} \log f(x_i|\theta,\sigma^2), \tag{12}$$

$$= \arg\max \sum_{i=1}^{n} \{(-1/2)\log(2\pi\sigma^2) - (1/2\sigma^2)(x_i-\theta)^2\},$$

$$= \arg\max \{-(n/2)\log(\sigma^2) - (1/2\sigma^2)\sum_{i=1}^{n}(x_i-\theta)^2\},$$

$$= \arg\max \ g(\theta,\sigma^2). \tag{13}$$

At optimality:

$$\nabla g(\theta,\sigma^2) = 0. \tag{14}$$

To get the estimate of θ, we need to compute the partial differentiation of $g(\theta,\sigma^2)$ w.r.t. θ, and set it equal to 0.

$$\frac{\partial g}{\partial \theta} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta) \tag{15}$$

Equating it to zero, gives,

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i, \text{ and } \frac{\partial^2 g}{\partial \theta^2} = -\frac{n}{\sigma^2}. \tag{16}$$

Since,the second derivative is negative at θ=$\tilde{\theta}$, so we can say that $\tilde{\theta}$ is the maximum of $g(\theta)$ and hence MLE of θ.

Similarly, to get the estimate of σ^2 ,

$$\frac{\partial g}{\partial \sigma^2} = 0 \ \text{ gives } \ \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\tilde{\theta})^2 = 0, \tag{17}$$

and we get the estimate,

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\tilde{\theta})^2. \tag{18}$$

Let $\sigma^2 = t$, $\tilde{\sigma}^2 = \tilde{t}$,

$$\left(\frac{\partial^2 g}{\partial t^2}\right)_{t=\tilde{t}} = \frac{n}{2\tilde{t}^2} - \frac{1}{\tilde{t}^3}\sum_{i=1}^{n}(x_i-\tilde{\theta})^2 \tag{19}$$

$$= -\frac{n^3}{2\left[\sum_{i=1}^{n}(x_i-\tilde{\theta})^2\right]^2} \tag{20}$$

Hence, $\tilde{\sigma}^2$ is the ML estimate of the σ^2.

Example 2.5. Bernoulli MLE Let $X_1, X_2, \ldots, X_n$ be iid Bernoulli(p). Then the likelihood function is

$$L(X|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^y(1-p)^{n-y}, \tag{21}$$

where $y = \sum_{i=1}^{n} x_i$. Sometimes, it is much easier to differentiate the log likelihood,

$$\log L(\mathbf{x}|p) = y \log p + (n-y)\log(1-p) \tag{22}$$

Since log is a monotonic function, maximizing likelihood and its log value yield the same value for the maxima.

If $0 < y < n$, differentiating $\log L(\mathbf{x}|p)$ and setting the result equal to 0 give the solution, $\tilde{p} = y/n$. It is also straightforward to verify that y/n is the global maximum in this case. If $y = 0$ or $y = n$, then

$$\log L(\mathbf{x}|p) = \begin{cases} n\log(1-p) & \text{if } y = 0 \\ n \log p & \text{if } y = n. \end{cases} \tag{23}$$

In either case $\log L(\mathbf{x}|p)$ is a monotone function of p, and it is again straightforward to verify that $\tilde{p} = y/n$ in each case. Thus, we have shown that $\sum_{i=1}^{n} X_i/n$ is the MLE of p.

2.3 Bayes Estimators

In the Bayesian approach parameter θ is considered to be a quantity whose variation can be described by a probability distribution (called the prior distribution). This is a subjective distribution, based on the experimenter's belief, and is formulated before the data are seen (hence the name prior distribution). A sample is then taken from a population indexed by θ and the prior distribution is updated with this sample information. The updated prior is called the posterior distribution. This updating is done with the use of Bayes Rule, hence the name Bayesian statistics.

Given $\{f(x|\theta) : \theta \in \Theta\}$, underlying assumption is that $\theta \in \Theta$ is randomly chosen from a prior distribution π over Θ, and therefore $X_1, X_2, ..., X_n$ are iid over distribution $f(\mathbf{x}|\theta)$.

Note 1. Choice of prior is subjective i.e. up to designer.

If $x_1, x_2, \ldots, x_n$ are the observed samples, construct the posterior distribution for $\theta \in \Theta$ using Bayes rule.

$$\pi(\theta|(x_1, x_2, \ldots, x_n)) = \frac{\pi(\theta) f(x_1, x_2, \ldots, x_3|\theta)}{m(x_1, x_2, \ldots, x_n)} \tag{24}$$

where $m(\mathbf{x})$ is the marginal distribution of $\mathbf{X}$, that is,

$$m(x_1, x_2, \ldots, x_n) = \int_\Theta \pi(\theta') f(x_1, x_2, \ldots, x_3|\theta') d\theta' \tag{25}$$

One can write down possible estimators depending on the posterior distribution: (i) Mode, (ii) Mean and (iii) Median.

Example 2.6. Bernoulli Bayes Estimators

Let $X_1, X_2, \ldots, X_n$ be iid Bernoulli(θ), $\theta \in [0, 1]$. Let prior has Beta(a, b) distribution then,

$$\pi(\theta) = \begin{cases} \dfrac{\theta^{a-1}(1-\theta)^{b-1}}{\beta(a,b)}, & \text{if } \theta \in [0,1] \\ 0, & \text{otherwise} \end{cases} \tag{26}$$

where,

$$\beta(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx. \tag{27}$$

Let consider posterior distribution,

$$\begin{aligned} \pi(\theta|(x_1, x_2, \ldots, x_n)) &= \frac{\pi(\theta) f(x_1, x_2, \ldots, x_n|\theta)}{f(x_1, x_2, \ldots, x_n)}, \\ &= \frac{1}{f(\mathbf{x})} \frac{\theta^{a-1}(1-\theta)^{b-1}}{\beta(a,b)} \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}, \\ &= \frac{\theta^{\sum_i x_i + a - 1} (1-\theta)^{n - \sum_i x_i + b - 1}}{\beta(a,b) f(\mathbf{x})}, \end{aligned} \tag{28}$$

which is a Beta $\left(\sum_i x_i + a, n - \sum_i x_i + b\right)$ distribution. Given the posterior distribution, the possible estimators are,

1. Mode $= \dfrac{\sum_i x_i + a - 1}{a + b + n - 2}$
2. Expectation $= \dfrac{\sum_i x_i + a}{a + b + n}$

Definition 2.7. Let F denote the class of pdfs or pmfs $\{f(x|\theta) : \theta \in \Theta\}$. A class $\prod$ of prior distributions on Θ is a *conjugate* family for F, if the posterior distribution is in the class $\prod$ for all $f \in F$, for all priors in $\prod$.