36-705: Intermediate Statistics **Fall 2017**

Lecture 14: October 2

Lecturer: Siva Balakrishnan

We continue our discussion of point estimation. We will follow Larry's notes very closely for the next few lectures. We focus on the parametric setting where we observe $X_1, \ldots, X_n \sim p(X; \theta)$ and would like to estimate θ.

Roughly, the questions we are trying to answer are:

1. Are there general purpose methods to come up with estimators of θ?
2. Given two (or more) estimators is there a general framework in which we can compare estimators?
3. Finally, are there general purpose ways to analyze complex estimators (say estimators that are not simple averages)?

$X_1, \ldots, X_n \sim p(x; \theta)$. Want to estimate $\theta = (\theta_1, \ldots, \theta_k)$. An *estimator*

$$\widehat{\theta} = \widehat{\theta}_n = w(X_1, \ldots, X_n)$$

is a function of the data. Keep in mind that the parameter is a fixed, unknown constant. The estimator is a random variable.

For now, we will discuss three methods of constructing estimators:

1. The Method of Moments (MOM)
2. Maximum likelihood (MLE)
3. Bayesian estimators.

Later we will discuss some other methods. We will also discuss several methods for evaluating estimators including:

1. Bias and Variance
2. Mean squared error (MSE)
3. Minimax Theory
4. Large sample theory.

Some Terminology. Throughout these notes, we will use the following terminology:

1. $\mathbb{E}_\theta(\widehat{\theta}) = \int \cdots \int \widehat{\theta}(x_1, \ldots, x_n) p(x_1; \theta) \cdots p(x_n; \theta) dx_1 \cdots dx_n$.
2. Bias: $\mathbb{E}_\theta(\widehat{\theta}) - \theta$.
3. The distribution of $\widehat{\theta}_n$ is called its *sampling distribution.*
4. The standard deviation of $\widehat{\theta}_n$ is called the *standard error* denoted by $\mathrm{se}(\widehat{\theta}_n)$.
5. $\widehat{\theta}_n$ is *consistent* if $\widehat{\theta}_n \xrightarrow{p} \theta$.
6. Later we will see that if bias → 0 and $\mathrm{Var}(\widehat{\theta}_n) \to 0$ as $n \to \infty$ then $\widehat{\theta}_n$ is consistent.

14.1 The Method of Moments

Suppose that $\theta = (\theta_1, \ldots, \theta_k)$. Define

$$\begin{aligned} m_1 &= \frac{1}{n}\sum_{i=1}^n X_i, & \mu_1(\theta) &= \mathbb{E}(X_i) \\ m_2 &= \frac{1}{n}\sum_{i=1}^n X_i^2, & \mu_2(\theta) &= \mathbb{E}(X_i^2) \\ &\vdots & &\vdots \\ m_k &= \frac{1}{n}\sum_{i=1}^n X_i^k, & \mu_k(\theta) &= \mathbb{E}(X_i^k). \end{aligned}$$

Let $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ solve:

$$m_j = \mu_j(\widehat{\theta}), \quad j = 1, \ldots, k.$$

In other words, we equate the first k sample moments with the first k theoretical moments. This defines k equations with k unknowns.

Example 14.1 *$N(\beta, \sigma^2)$ with $\theta = (\beta, \sigma^2)$. Then $\mu_1 = \beta$ and $\mu_2 = \sigma^2 + \beta^2$. Equate:*

$$\frac{1}{n}\sum_{i=1}^n X_i = \widehat{\beta}, \quad \frac{1}{n}\sum_{i=1}^n X_i^2 = \widehat{\sigma}^2 + \widehat{\beta}^2$$

to get

$$\widehat{\beta} = \overline{X}_n, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

Example 14.2 *Suppose*

$$X_1, \ldots, X_n \sim \text{Binomial}(k, p)$$

where both k and p are unknown. We get

$$kp = \overline{X}_n, \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = kp(1-p) + k^2p^2$$

giving

$$\widehat{p} = \frac{\overline{X}_n}{k}, \quad \widehat{k} = \frac{\overline{X}_n^2}{\overline{X}_n - \frac{1}{n}\sum_i (X_i - \overline{X}_n)^2}.$$

The method of moments was popular many years ago because it is often easy to compute. Lately, it has attracted attention again. For example, there is a large linterature on estimating "mixtures of Gaussians" using the method of moments.

14.2 Maximum Likelihood

The most popular method for estimating parameters is maximum likelihood. The reason is that, under certain conditions, the maximum likelihood estimator is optimal. This result was established by Sir Ronald Fisher and Lucian LeCam. We'll discuss optimality later.

The maximum likelihood estimator (mle) $\widehat{\theta}$ is defined as the maximizer of

$$\mathcal{L}(\theta) = p(X_1, \ldots, X_n; \theta) \overset{iid}{=} \prod_i p(X_i; \theta).$$

This is the same as maximizing the log-likelihood

$$\mathcal{LL}(\theta) = \log \mathcal{L}(\theta).$$

Often it suffices to solve

$$\frac{\partial \mathcal{LL}(\theta)}{\partial \theta_j} = 0, \quad j = 1, \ldots, k.$$

Example 14.3 *Binomial.* $\mathcal{L}(p) = \prod_i p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$ *where* $S = \sum_i X_i$. *So*

$$\mathcal{LL}(p) = S\log p + (n-S)\log(1-p)$$

and $\widehat{p} = \overline{X}_n$.

Example 14.4 $X_1, \ldots, X_n \sim N(\mu, 1)$.

$$\mathcal{L}(\mu) \propto \prod_i e^{-(X_i-\mu)^2/2} \propto e^{-n(\overline{X}_n-\mu)^2}, \quad \mathcal{LL}(\mu) = -\frac{n}{2}(\overline{X}_n - \mu)^2$$

and $\widehat{\mu} = \overline{X}_n$. *For* $N(\mu, \sigma^2)$ *we have*

$$\mathcal{L}(\mu, \sigma^2) \propto \prod_i \frac{1}{\sigma} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2\right\}$$

and

$$\mathcal{LL}(\mu, \sigma^2) = -n \log \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2.$$

Set

$$\frac{\partial \mathcal{LL}}{\partial \mu} = 0, \quad \frac{\partial \mathcal{LL}}{\partial \sigma^2} = 0$$

to get

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2.$$

Example 14.5 *Let* $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$. *Then*

$$\mathcal{L}(\theta) = \frac{1}{\theta^n} I(\theta > X_{(n)})$$

and so $\widehat{\theta} = X_{(n)}$.

14.2.1 MLE and MoM for exponential families

The log-likelihood in an exponential family is concave and given by

$$\mathcal{LL}(\theta; x_1, \ldots, x_n) \propto \left[\sum_{i=1}^s \theta_i \sum_{j=1}^n T_i(x_j) - nA(\theta)\right],$$

so we can simply take the derivative with respect to θ and set this equal to 0. Using the facts we have seen in the last lecture about the derivative of A, we can see that this amounts to solving the following system of equations for θ:

$$\mathbb{E}_{p(X;\theta)}[T_i(X)] = \frac{1}{n}\sum_{j=1}^n T_i(x_j) \quad \text{for} \quad i \in \{1, \ldots, s\}.$$

So the maximum likelihood estimator simply picks the parameters θ to match the empirical expectations of the sufficient statistics to the expected value of the sufficient statistics under the distribution.

Usually we cannot compute this estimator in closed form so we use an iterative algorithm (like gradient ascent) to maximize the likelihood. However, you should remember that exponential families have concave likelihoods so this is usually a tractable endeavour (at least for simple enough families).

For exponential families as we can see above the method of moments coincides with the MLE (if we chose the sufficient statistics to direct which moments to compute).

14.2.2 Equivariance and the profile likelihood

Suppose that $\theta = (\eta, \xi)$. The *profile likelihood* for η is defined by

$$\mathcal{L}(\eta) = \sup_{\xi} \mathcal{L}(\eta, \xi).$$

To find the mle of η we can proceed in two ways. We could find the overall mle $\widehat{\theta} = (\widehat{\eta}, \widehat{\xi})$. The mle for η is just the first coordinate of $(\widehat{\eta}, \widehat{\xi})$. Alternatively, we could find the maximizer of the profile likelihood. These give the same answer. Do you see why?

The mle is *equivariant.* if $\eta = g(\theta)$ then $\widehat{\eta} = g(\widehat{\theta})$. Suppose g is invertible so $\eta = g(\theta)$ and $\theta = g^{-1}(\eta)$. Define $\mathcal{L}^*(\eta) = \mathcal{L}(\theta)$ where $\theta = g^{-1}(\eta)$. So, for any η,

$$\mathcal{L}^*(\widehat{\eta}) = \mathcal{L}(\widehat{\theta}) \geq \mathcal{L}(\theta) = \mathcal{L}^*(\eta)$$

and hence $\widehat{\eta} = g(\widehat{\theta})$ maximizes $\mathcal{L}^*(\eta)$. For non invertible functions this is still true if we define

$$\mathcal{L}^*(\eta) = \sup_{\theta : g(\theta) = \eta} \mathcal{L}(\theta).$$

(In other words, the profile likelihood.)

Example 14.6 *Binomial. The mle is $\widehat{p} = \overline{X}_n$. Let $\psi = \log(p/(1-p))$. Then $\widehat{\psi} = \log(\widehat{p}/(1-\widehat{p}))$.*

14.3 Bayes Estimator

To define the Bayes estimator, we begin by treating θ as a random variable. This point requires much discussion (which we will have later). For now, just tentatively think of θ as random. We start with a *prior distribution* $p(\theta)$ on θ. Note that

$$p(x_1, \ldots, x_n|\theta)p(\theta) = p(x_1, \ldots, x_n, \theta).$$

Now compute the *posterior distribution* by Bayes' theorem:

$$p(\theta|x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n|\theta)p(\theta)}{p(x_1, \ldots, x_n)}$$

where

$$p(x_1, \ldots, x_n) = \int p(x_1, \ldots, x_n|\theta)p(\theta)d\theta.$$

This can be written as

$$p(\theta|x_1, \ldots, x_n) \propto \mathcal{L}(\theta)p(\theta) = \text{Likelihood } \times \text{ prior}.$$

Now compute a point estimator from the posterior. For example:

$$\widehat{\theta} = \mathbb{E}(\theta|x_1, \ldots, x_n) = \int \theta p(\theta|x_1, \ldots, x_n)d\theta = \frac{\int \theta p(x_1, \ldots, x_n|\theta)p(\theta)d\theta}{\int p(x_1, \ldots, x_n|\theta)p(\theta)d\theta}.$$

Example 14.7 *Let* $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$. *Let the prior be* $\theta \sim \text{Beta}(\alpha, \beta)$. *Hence*

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

and

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt.$$

Set $Y = \sum_i X_i$. *Then*

$$p(\theta|X) \propto \underbrace{\theta^Y 1 - \theta^{n-Y}}_{\text{likelihood}} \times \underbrace{\theta^{\alpha-1} 1 - \theta^{\beta-1}}_{\text{prior}} \propto \theta^{Y+\alpha-1} 1 - \theta^{n-Y+\beta-1}.$$

Therefore, $\theta|X \sim \text{Beta}(Y+\alpha, n-Y+\beta)$. *The Bayes estimator is*

$$\widetilde{\theta} = \frac{Y+\alpha}{(Y+\alpha)+(n-Y+\beta)} = \frac{Y+\alpha}{\alpha+\beta+n} = (1-\lambda)\widehat{\theta}_{mle} + \lambda\,\overline{\theta}$$

where

$$\overline{\theta} = \frac{\alpha}{\alpha+\beta}, \quad \lambda = \frac{\alpha+\beta}{\alpha+\beta+n}.$$

This is an example of a conjugate prior.

Example 14.8 *Let* $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ *with* σ^2 *known. Let* $\mu \sim N(m, \tau^2)$. *Then*

$$\mathbb{E}(\mu|X) = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}}\overline{X}_n + \frac{\frac{\sigma^2}{n}}{\tau^2 + \frac{\sigma^2}{n}}m$$

and

$$\text{Var}(\mu|X) = \frac{\sigma^2\tau^2/n}{\tau^2 + \frac{\sigma^2}{n}}.$$

14.4 MSE

Now we discuss the evaluation of estimators. The mean squared error (MSE) is

$$\mathbb{E}_\theta(\widehat{\theta}-\theta)^2 = \int \cdots \int (\widehat{\theta}(x_1,\ldots,x_n)-\theta)^2 p(x_1;\theta)\cdots p(x_n;\theta)dx_1\ldots dx_n.$$

The bias is

$$B = \mathbb{E}_\theta(\widehat{\theta})-\theta$$

and the variance is

$$V = \mathrm{Var}_\theta(\widehat{\theta}).$$

Theorem 14.9 *We have*

$$MSE = B^2+V.$$

Proof: Let $m=\mathbb{E}_\theta(\widehat{\theta})$. Then

$$\begin{aligned} MSE &= \mathbb{E}_\theta(\widehat{\theta}-\theta)^2 = \mathbb{E}_\theta(\widehat{\theta}-m+m-\theta)^2 \\ &= \mathbb{E}_\theta(\widehat{\theta}-m)^2 + (m-\theta)^2 + 2\mathbb{E}_\theta(\widehat{\theta}-m)(m-\theta) \\ &= \mathbb{E}_\theta(\widehat{\theta}-m)^2 + (m-\theta)^2 = V+B^2. \end{aligned}$$

■

An estimator is *unbiased* if the bias is 0. In that case, the MSE = Variance. There is often a tradeoff between bias and variance. So low bias can imply high variance and vice versa.

Example 14.10 *Let $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$. Then*

$$\mathbb{E}(\overline{X})=\mu, \quad \mathbb{E}(S^2)=\sigma^2.$$

The MSE's are

$$\mathbb{E}(\overline{X}-\mu)^2 = \frac{\sigma^2}{n}, \quad \mathbb{E}(S^2-\sigma^2)^2 = \frac{2\sigma^4}{n-1}.$$

We would like to choose an estimator with small MSE. However, the MSE is a function of θ. Later, we shall discuss minimax estimators, that use the maximum of the MSE over θ as a way to compare estimators.