COLUMBIA UNIVERSITY
Statistical Inference and Modeling
Lecture Notes 2: Estimation

Prof. Marco Avella
Scribe: Monica Yan
Spring 2020

This note introduces key ideas about the estimation of model parameters based on observed data. The basic framework is summarized in the following quote by Ronald A. Fisher [1]

> Briefly, and in its most concrete form, the object of statistical methods is the reduction of data ... This object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample. The law of distribution of this hypothetical population is specified by relatively few parameters, which are sufficient to describe it exhaustively in respect of all qualities under discussion.

1 Statistical Models and Identification

Throughout this course we will be interested in an statistical experiments where the observed outcome is a sample $X_1, \ldots, X_n$ of n i.i.d random variables in some measurable space E (e.g. $E \subset \mathbb{R}^d$) and denote by F their common distribution.

Definition (Statistical Model)**.** Let $X_1, \ldots, X_n$ be n independent copies of a random variable X. A *statistical model* for a sample from X is any family of frequency or density functions $\{f(x;\theta)\}_{\theta\in\Theta}$ for the law of X. The index set Θ is called the *parameter set*.

We will assume that the statistical model is *well specified*, i.e. such that $F = F_{\theta_0}$ for some $\theta_0 \in \Theta$. This particular θ_0 is called the *true parameter*. In words, we assume that the true generating probability law F belongs to the family of distributions postulated by the statistical model.

Example. Think about the paper "Adult persistence of head-turning asymmetry" [2] discussed in the first lecture, where $x_i = 1$ if the couple turned their head to the right, and $x_i = 0$ if they turned it to the left. In this context, our model F is a Bernoulli distribution $\text{Ber}(p)$, and it is well specified if the true generating process is indeed a Bernoulli with a particular value of $p \in (0, 1)$, say $p = 0.5$ if there is no asymmetry. In this case, the model is well-specified, and $p = 0.5$ is called the true parameter.

[1] R.A Fisher, M."On the mathematical foundations of theoretical statistics" Phil. Trans. R. Soc. Lond. A 222.594-604 (1922): 309-368. Link to source file. This breaktrough paper laid out the foundations of modern statistics by introducing the maximum likelihood estimator and the key notions of consistency, efficiency and sufficiency!

[2] Güntürkün, O., 2003. Human behaviour: adult persistence of head-turning asymmetry. Nature, 421(6924), p.711. Link to source file

Given a hypothesized statistical model, the goal of parameter estimation is to figure out the value of the underlying parameter that generated the observed data. Our statistical models will need the following regularity condition if one hopes to successfully estimate the unknown true parameter.

Definition (Identification)**.** The parameter θ is called identified if and only if the map $\theta \in \Theta \mapsto F_\theta$ is injective, i.e.

$$\theta \neq \theta' \Rightarrow F_\theta \neq F_{\theta'}$$

Example. In the head-turning example, the statistical model was $Ber(p)$ and the parameter p is identified. Indeed, trivially a different parameter $p' \neq p$ will lead to a the model $Ber(p')$ that will generate data with a different distrinbution from that of $Ber(p)$.

Example. If $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, but *we only observe* $Y_i = \mathbb{1}_{X_i \geq 0}$ for $i = 1, \ldots, n$. In this case the parameters μ and σ^2 are not identified. To see this first note that

$$\mathbb{P}(Y_i = 1) = \mathbb{P}(X_i \geq 0) = 1 - \mathbb{P}(X_i \leq 0) = 1 - \Phi(-\mu/\sigma).$$

Since the ratio $\theta = \mu/\sigma$ completely determines the distribution of the *observed* random sample $Y_1 \ldots, Y_n$, we can easily see that the pairs $(c\mu, c\sigma)$ and (μ, σ) lead to the same distribution of Y_i for any $c > 0$. In this case only $\theta = \mu/\sigma$ is identified.

Typically, we want to know the true parameter. There are a few ways general ways to construct estimators based on an observed random sample. In the following we will discuss the *Maximum Likelihood Estimator* and the *Method of Moments*, but first let's introduce some important concepts.

2 Parameter estimation: key concepts

Definition (Estimator/Estimate)**.** Suppose that the observable random variables of interest are $X_1, \ldots, X_n$. We define a *statistic* $T_n = T(\mathbf{X})$ to be a function of $\mathbf{X} = X_1, \ldots, X_n$ that does not depend on unknown parameters.
An *estimator* of $\theta_0 \in \Theta$ is a statistic whose primary goal is to estimate θ_0. If $\{X_1 = x_1, \ldots, X_n = x_n\}$ are observed, then $T(x_1, \ldots, x_n)$ is called an *estimate* of θ_0.

Here are some central notions for assessing the performance of an estimator:

- An estimator $\hat{\theta}_n$ of θ_0 is unbiased if $\mathbb{E}[\hat{\theta}_n] = \theta_0$.
- An estimator $\hat{\theta}_n$ is (weakly) consistent for the parameter θ_0 if

$$\hat{\theta}_n \xrightarrow[n\to\infty]{\mathcal{P}} \theta_0,$$

that is, if and only if for every $\epsilon > 0$ we have that

$$\mathbb{P}(\|\widehat{\theta} - \theta\| > \epsilon) \to 0, \qquad \text{as } n \to \infty.$$

- The *Mean Square Error* (MSE) is a measure of accuracy of an estimator.
  - Mean Square Error

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}_n) =& \mathbb{E}[\|\hat{\theta}_n - \theta_0\|_2^2] = \mathbb{E}[(\hat{\theta}_n - \theta_0)^\top(\hat{\theta}_n - \theta_0)] \\ =& \mathbb{E}[\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n] + \mathbb{E}[\hat{\theta}_n] - \theta_0)^\top(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n] + \mathbb{E}[\hat{\theta}_n] - \theta_0)] \\ =& \mathbb{E}[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|_2^2] + \mathbb{E}[\|\mathbb{E}[\hat{\theta}_n] - \theta_0\|_2^2] + 2\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^\top(\mathbb{E}[\hat{\theta}_n] - \theta_0)] \\ =& tr[var(\hat{\theta}_n)] + \|bias(\hat{\theta}_n, \theta_0)\|_2^2\end{aligned}$$

    since $2\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^\top(\mathbb{E}[\hat{\theta}_n] - \theta_0)] = 0$.

  - One way to compare two estimator is to consider their *relative efficiency* by comparing the ratio of their MSE i.e. given two estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ of θ_0 we have that

$$\mathrm{Eff}(\hat{\theta}_n, \tilde{\theta}_n) = \frac{\mathrm{MSE}(\hat{\theta}_n)}{\mathrm{MSE}(\tilde{\theta}_n)}$$

3 Method of Moments

Let $X_1, \ldots, X_n$ be an i.i.d. sample of univariate random variables associated with a statistical model $(F_\theta)_{\theta\in\Theta}$ and assume that $\Theta \subseteq \mathbb{R}^d$, for some $d \geq 1$.

- Population moments:

$$\mu_k(\theta_0) = \mathbb{E}[X_1^k], \quad 1 \leq k \leq d$$

- Empirical moments:

$$\hat{\mu}_k = \overline{X_n^k} = \frac{1}{n}\sum_{i=1}^n X_i^k, \quad 1 \leq k \leq d$$

The key intuition behind the method of moments is that when the sample size gets larger, the empirical moments will converge to the population moments. We can therefore try to use empirical moments to estimate their population counterparts. Furthermore, one can hope to use empirical moments to estimate the true parameter θ_0 provided it can be written as a function of population parameters.

Definition. Assumming that the function $\psi(\theta) = (\mu_1(\theta), \ldots, \mu_d(\theta))$ is bijective we have that $\theta = \psi^{-1}(\mu_1(\theta), \ldots, \mu_d(\theta))$. The method of moments estimator of θ_0 is

$$\hat{\theta}_n^{MM} = \psi^{-1}(\hat{\mu}_1, \ldots, \hat{\mu}_d)$$

provided it exists.

The above definition assumes the existence of a bijective map ψ for the computation of the estimator. In practice, this function will be derived on a case by case basis depending on the moments that we use. Typically we will choose the number of moments to match the dimension of the parameter θ that we want to estimate.

Example. Let $X_1, \ldots, X_n$ be an iid random sample with distribution Ber(p). The first moment of a Bernoulli distribution is

$$\mathbb{E}[X^1] = p$$

the empirical moment is

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

therefore, we use the empirical moment to estimate the population moment

$$\hat{p}^{MM} = \overline{X}$$

Example. Let $X_1, \ldots, X_n$ be i.i.d. Exponential random variables with density $f(x) = \lambda e^{-\lambda x}$. The first moment of X is

$$\mu_1 = \mathbb{E}[X^1] = \frac{1}{\lambda} = \psi(\lambda)$$

and the empirical first moment is

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{\mu}_1.$$

The resulting MM estimator of λ is

$$\hat{\lambda}^{MM} = \psi^{-1}(\overline{X}) = \frac{1}{\overline{X}}$$

Example. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d Gamma(α, β) with density function

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{\beta x},$$

where $\alpha, \beta > 0$ and $\Gamma(\cdot)$ is the gamma function. In this case we want to estimate the two dimensional parameter $\theta = (\alpha, \beta)$. The first two moments of this distribution are:

$$\mathbb{E}[X_1] = \mu_1(\theta) = \frac{\alpha}{\beta}, \qquad \mathbb{E}[X_1^2] = \mu_2(\theta) = \frac{\alpha(\alpha+1)}{\beta^2},$$

which implies that

$$\alpha = \frac{\mu_1^2}{\mu_2 - \mu_1^2}, \qquad \beta = \frac{\mu_1}{\mu_2 - \mu_1^2}.$$

The MOM says that we replace the right-hand sides of these equations by the *sample moments* and then solve for α and β. In this case, we get

$$\hat{\alpha} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2} = \frac{(\overline{X})^2}{\overline{X^2} - (\overline{X})^2}, \qquad \hat{\beta} = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2} = \frac{\overline{X}}{\overline{X^2} - (\overline{X})^2}.$$

The next theorem states that the method of moments estimator is a consistent and asymptotically normally distributed. In order to state the result we need some additional notation. Let $\mathbf{X}_1 = (X_1^1, X_1^2, \ldots, X_1^d)$ and $M(\theta_0) = (\mu_1(\theta_0), \ldots, \mu_d(\theta_0))$ denote its expectation and $\Sigma(\theta_0) = \text{var}[(X_1^1, X_1^2, \ldots, X_1^d)]$ its variance

Theorem. *If ψ^{-1} is continuously differentiable at $M(\theta_0)$ then*

$$\sqrt{n}(\hat{\theta}_n^{MM} - \theta_0) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, V(\theta_0)),$$

where $V(\theta_0) = [\nabla\psi^{-1}(M(\theta_0))]\Sigma(\theta_0)[\nabla\psi^{-1}(M(\theta_0))]^\top$.

The CLT tells us that if we standardize $\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$, it will converge to a multivariate normal distribution. Therefore the proof of the above theorem follows from the delta method.

4 Maximum Likelihood Estimation

Definition. Let $X_1, \ldots, X_n$ be an i.i.d. sample of random variables with density or frequency function $f(x;\theta_0)$ and assume that $\Theta \subseteq \mathbb{R}^d$, for some $d \geq 1$, the likelihood function is

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^n f(x_i;\theta)$$

and the maximum likelihood estimator of θ_0 is

$$\hat{\theta}_n^{MLE} = \operatorname*{argmax}_{\theta\in\Theta} L(\theta; X_1, \ldots, X_n) = \operatorname*{argmax}_{\theta\in\Theta} \log L(\theta; X_1, \ldots, X_n)$$

Example. Let $X_1, \ldots, X_n \overset{iid}{\sim} Ber(p)$. The joint density function is

$$f(x_1, \ldots, x_n; p) = \prod_{i=1}^n P(X_i = x_i) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}$$

and the log-likelihood is

$$\ell(p; X_1, \ldots, X_n) = \log(L(p; X_1, \ldots, X_n)) = \log p \sum_{i=1}^n X_i + \log(1-p)\sum_{i=1}^n (1 - X_i)$$

to get the argmax log-likelihood

$$\frac{\partial(\ell(p; X_1, \ldots, X_n)}{\partial p} = \frac{\sum_{i=1}^n X_i}{p} - \frac{\sum_{i=1}^n (1-X_i)}{(1-p)} \overset{\text{set}}{=} 0$$

$$\hat{p}^{MLE} = \frac{1}{n}\sum_{i=1}^n X_i$$

Example. If $X_1, \ldots, X_n$ are i.i.d. Exponential random variables with mean $1/\lambda$, the likelihood function is

$$L(\lambda; X_1, \ldots, X_n) = \prod_{i=1}^n f(X_i;\lambda) = \prod_{i=1}^n \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda\sum_{i=1}^n X_i}.$$

$$\ell(\lambda; X_1, \ldots, X_n) = n\log(\lambda) - \lambda \sum_{i=1}^{n} X_i$$

$$\frac{\partial(\ell(\lambda; X_1, \ldots, X_n)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i \overset{\text{set}}{=} 0$$

$$\hat{\lambda}^{MLE} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}}$$

Example. Suppose that $X_1, \ldots, X_n$ are i.i.d from a Gamma distribution for which the p.d.f is as follows:

$$f(x; \alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, \qquad \text{for } x > 0.$$

The likelihood function is

$$L(\alpha; X_1, \ldots, X_n = \frac{1}{\Gamma(\alpha)^n} \left(\prod_{i=1}^{n} X_i \right)^{\alpha-1} e^{-\sum_{i=1}^{n} X_i},$$

and thus the log-likelihood is

$$\log L(\alpha; X_1, \ldots, X_n) = -n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log(X_i) - \sum_{i=1}^{n} X_i,$$

The MLE of α will be the value of α that satisfies the equation

$$\begin{aligned} \frac{\partial}{\partial \alpha} L(\alpha) \quad &= \quad -n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} \log(X_i) = 0 \\ \text{i.e.,} \quad & \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \frac{1}{n} \sum_{i=1}^{n} \log(X_i). \end{aligned}$$

In this case we do not have an analytical solution for the estimator. Instead, we would have to rely on numerical methods (e.g. Newton's method) in order to compute $\hat{\alpha}^{MLE}$.

The maximum likelihood estimator can also be shown to be consistent and asymptotically normally distributed. Before presenting a more detailed statement, let us introduce the important notion of *Fisher Information*[3] defined as

$$I(\theta) = \mathbb{E}\Big[\frac{\partial}{\partial \theta} \log L(\theta; X_1) \frac{\partial}{\partial \theta^\top} \log L(\theta; X_1) \Big] = -\mathbb{E}\Big[\frac{\partial^2}{\partial \theta \partial \theta^\top} \log L(\theta; X_1) \Big].$$

where X_1 means one single observation and $L(\theta; X_1)$ is its likelihood function. When it comes to the whole sample, as X_i are i.i.d. random variables, the Fisher information of the whole sample $I_n(\theta)$ is

$$I_n(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta \partial \theta^\top} \sum_{i=1}^{n} \log L(\theta; X_i) \right] = -\sum_{i=1}^{n} \mathbb{E}\Big[\frac{\partial^2}{\partial \theta \partial \theta^\top} \log L(\theta; X_i) \Big] = nI(\theta)$$

[3] see appendix for more details

Theorem. *Under regularity conditions* [a]*, we have*

$$\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{\mathcal{P}} \theta_0$$

$$\sqrt{n}(\hat{\theta}_n^{MLE} - \theta_0) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, I(\theta_0)^{-1})$$

where $I(\theta_0)$ is Fisher Information.

[a] Theorem 5.4 in K. Knight (2000) "Mathematical Statistics"

Remark. (Invariance Property) For any function $\tau(\theta)$, its MLE is $\tau(\hat{\theta}^{MLE})$ [4]. The invariance property tells us that the MLE of known function of the unknown parameter can be found by plugging-in the MLE of the unknown parameter. Since the MLE is consistent and asymptotically normally distributed, $\tau(\hat{\theta}^{MLE})$ will also be consistent and asymptotically normally distributed provided τ is continuous and differentiable at θ_0.

5 Optimality in Estimation

When we compute an estimator, one would intuitively like them to be (nearly) unbiased and as efficiency as possible. The following theorem establishes a strong connection between the Fisher information and the precision one can hope to achieve with an estimator.

Theorem (Cramér-Rao Lower Bound)**.** *Let $X_1, \ldots, X_n$ be an i.i.d. sample of random variables with density or frequency function $f(x;\theta_0)$ and assume: (1) the support of $f(x;\theta)$ does not depend on θ; (2) $f(x;\theta)$ is differentiable with respect to θ for all x. Then for an unbiased estimator $T(\mathbf{X})$ of θ_0 we have that*

$$var(T(\mathbf{X})) \geq \frac{1}{nI(\theta_0)},$$

where $I(\theta_0)$ is the Fisher Information.

Remark. The Cramér-Rao lower bound shows that the MLE enjoys certain optimality properties. Indeed, the asymptotic normality of the MLE gives us the approximation

$$\hat{\theta}_n^{MLE} \approx \mathcal{N}(\theta_0, \frac{1}{nI(\theta_0)})$$

which tells us that the MLE is approximately unbiased and its variance attains the minimum variance given by the Cramér-Rao lower bound asymptotically. The following example shows that in certain cases this lower bound can be also be attained for a fixed sample size n.

[4] Theorem 7.2.10 in G. Casella & R.L. Berger (2002). "Statistical Inference"

Example. $X_1, \ldots, X_n$ are i.i.d. Poisson random variables with mean λ, the log-likelihood function is

$$\ell(\lambda; x_1, \ldots, x_n) = \sum_{i=1}^{n} [x_i \log(\lambda) - \lambda - log(x_i!)]$$

the MLE of λ is

$$\hat{\lambda}^{MLE} = \overline{X}$$

and

$$\mathbb{E}[\hat{\lambda}^{MLE}] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \sum_{i=1}^{n} \lambda = \lambda.$$

Furthermore the Fisher Information is

$$\begin{aligned} I(\lambda) &= -\mathbb{E}[\frac{\partial^2 \ell(\lambda; X_1)}{(\partial \lambda)^2}] \\ &= -\mathbb{E}[-\frac{X_1}{\lambda^2}] = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}. \end{aligned}$$

Therefore any unbiased estimator will have a variance of at least λ/n. This lower bound is attained by $\hat{\lambda}^{MLE}$, which shows that us is a minimum variance unbiased estimator.

Appendix A Fisher Information

Fisher information measures the amount of information one observation X_1 carries out about the unknown parameter θ of the distribution models X. Let $f(X;\theta)$ be the probability density function for random variable X, then it's also the likelihood function $L(\theta; X_1)$. The partial derivative with respect to θ of the natural logarithm of the likelihood function is called the "scoreâĂİ, $s(\theta; X_1)$.

$$s(\theta; X_1) = \frac{\partial}{\partial\theta} \log L(\theta; X_1)$$

It measures the sensitivity of log-likelihood to small changes to the parameter values. Under certain regularity conditions, given θ is the "true" parameter, it can be shown that the expected value of score function over the whole sample space $\mathcal{X}$ is 0

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial \log L(\theta; X_1)}{\partial\theta}\Big|\theta\right] &= \int_{\mathcal{X}} \frac{\partial \log f(x;\theta)}{\partial\theta} f(x;\theta)\mathrm{d}x \quad \text{(rewrite } L(\theta; X_1) \text{ as } f(x;\theta)\text{)} \\
&= \int_{\mathcal{X}} \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} f(x;\theta)\mathrm{d}x \quad \text{(chain rule)} \\
&= \frac{\partial}{\partial\theta} \int_{\mathcal{X}} f(x;\theta)\mathrm{d}x \quad \text{(under assumed regularity conditions)} \\
&= \frac{\partial}{\partial\theta} 1 = 0
\end{aligned}$$

Fisher information is defined as the variance of the score function, as the expectation of score function is 0, thus

$$I(\theta) = Var(s(\theta; X_1)) = \mathbb{E}\Big[\Big(\frac{\partial}{\partial\theta} \log L(\theta; X_1)\Big)^2\Big]$$

If the log-likelihood function is twice differentiable with respect to θ and under certain regularity conditions, the Fisher information can also written as

$$I(\theta) = -\mathbb{E}\Big[\frac{\partial^2}{\partial\theta\partial\theta^\top} \log L(\theta; X_1)\Big]$$

This shows relationship between the Fisher information and the curvature of log-likelihood function, which are the two important quantities when describing $\hat{\theta}_{MLE}$. Near the maximum likelihood estimate, a low Fisher information means that maximum value of appear in flat area, in other words, there are many similar log-likelihood values nearby. While a high Fisher information indicates the maximum of log-likelihood is sharp.

When $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is n i.i.d. random variables, the Fisher information of the whole data $I_n(\theta)$ is

$$\begin{aligned}
I_n(\theta) &= -\mathbb{E}\Big[\frac{\partial^2}{\partial\theta\partial\theta^\top} \log L(\theta; X_1, X_2, \ldots, X_n)\Big] \\
&= \sum_{i=1}^{n} -\mathbb{E}\Big[\frac{\partial^2}{\partial\theta\partial\theta^\top} \log L(\theta; X_i)\Big] \\
&= nI(\theta)
\end{aligned}$$