5 Methods for Constructing Estimators

In this section, we will consider different methods for constructing point estimators.

5.1 Method of Moments

This method was proposed by the British statistician Karl Pearson in 1894. Suppose we have a population with p.d.f. $f(x,\theta)$, where θ is a scalar. For any function $g(X,\theta)$, we can define its expectation (provided it is finite) as

$$E\left[g(X,\theta)\right] = \int_{-\infty}^{\infty} g\left(x,\theta\right) f(x;\theta)dx$$

This expectation is called a population moment.

For example, the population mean is the first-order moment

$$\mu = \mu_1(\theta) = E[X] \equiv \int_{-\infty}^{\infty} x f(x;\theta)dx$$

with $g(X,\theta) = X$.

Similarly, we can define moments of any order k:

$$\mu_k(\theta) = E[X^k] \equiv \int_{-\infty}^{\infty} x^k f(x,\theta)dx$$

The population variance is also a moment, since it is an expectation of function $g(X,\mu) = (X-\mu)^2$:

$$\sigma^2 = E\left[(X-\mu)^2\right]$$

Suppose that for some known function $g(X,\theta)$

$$E\left[g(X,\theta)\right] = 0 \tag{4}$$

and we are interested in estimated the unknown parameter θ. If we knew the p.d.f. $f(x;\theta)$, we could find the functional form of $E\left[g(X,\theta)\right]$ as a function of θ and equate it to zero, i.e., get rid of the expectation sign. Then, we could find θ by simply solving the resulting equation. However, we don't know $f(x;\theta)$.

Yet, there is an alternative. Suppose we have a random sample $\{X_1,...,X_n\}$. Since X_i are i.i.d., $g(X_i,\theta)$ are also i.i.d. Then, by the law of large numbers (discussed in class), the sample average $\frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta) \xrightarrow{p} E\left[g(X,\theta)\right]$. This suggests approximating $E\left[g(X,\theta)\right]$ by $\frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta)$. In other words, in equation (4) we can replace the population moment $E\left[g(X,\theta)\right]$ by its sample analogue $\frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta)$:

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i,\widehat{\theta}) = 0 \tag{5}$$

and then solve the last equation for $\widehat{\theta}$. Then, $\widehat{\theta}$ is called a *method of moments (MM)* estimator of θ.

We assumed that θ is a scalar, so that (5) is one equation in one unknown. In general, if θ is a m-dimensional vector and $g(x,\theta)$ is an m-dimensional vector-function that depends on the data x and the parameter, then a Mo estimator is defined as the solution to the system of m equations in m unknowns:

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i, \hat{\theta}) = \mathbf{0}_{m\times 1}$$

Because of sampling uncertainty, there is in general no guarantee that there is always a solution for the *sample* moment conditions, in particular if $g(x,\theta)$ is nonlinear in θ or the number of moment condition exceeds the dimension of the parameter vector. In that case, we may instead define an estimator as the minimizer of a quadratic form of the sample moment vector

$$Q_n(\theta) := \left[\frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta)\right]' W \left[\frac{1}{n}\sum_{i=1}^{n} g(X_i,\theta)\right]$$

where W is a known positive semi-definite matrix. An estimator of the form $\hat{\theta} = \arg\min_{\theta\in\Theta} Q(\theta)$ is called a *generalized method of moments* (GMM) estimator which plays an important role in econometrics.

Example 1. Poisson distribution.

Suppose $X_1,\ldots,X_n$ is an i.i.d. sample from a Poisson distribution with unknown parameter λ, i.e. $X_i \sim Poisson(\lambda)$. The distribution has only one unknown parameter, and the first population moment (mean) is given by

$$\mu = E[X] = \lambda$$

Therefore, the MM estimator of λ is simple the sample mean, i.e., we replace the population mean by the sample mean

$$\hat{\lambda} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Thus, the MM estimator of λ coincides with the sample mean.

Example 2. Uniform distribution

Let $X \sim U[0,\theta]$ be a uniformly distributed random variable over an interval depending on the unknown parameter θ. The p.d.f. is

$$f(x) = \begin{cases} \frac{1}{\theta} & if\ 0 \le x \le \theta \\ 0 & otherwise \end{cases}$$

How could we estimate θ? For uniform distribution $E[X] = \frac{\theta}{2}$. Now, in the left-hand side of the last equation, replace the population mean by the sample mean to get

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\hat{\theta}}{2}$$

Hence, the a method-of-moments estimator for θ is $\hat{\theta}_{MM} = 2\bar{X}_n$, where $\bar{X}_n$ is the sample mean.

5.2 Maximum Likelihood Estimation

While the method of moments only tries to match a selected number of moments of the population to their sample counterparts, we may alternatively construct an estimator which makes the population distribution as a whole match the sample distribution as closely as possible. This is what the *maximum likelihood estimator* of a parameter θ does, which is loosely speaking, the value which "most likely" would have generated the observed sample.

Suppose we have an i.i.d. sample $\{X_1, ..., X_n\}$ from the population with p.d.f. $f(x, \theta)$, which is known up to the parameter θ. That is, $X_1, ..., X_n$ are independent and identically distributed with the common p.d.f. $f(x, \theta)$. Since $\{X_1, ..., X_n\}$ are independent, their joint p.d.f. or *joint likelihood function* is

$$L(\theta) \equiv f(X_1, \theta) f(X_2, \theta) ... f(X_n, \theta) = \prod_{i=1}^{n} f(X_i, \theta) \tag{6}$$

The Maximum Likelihood estimator (MLE) $\widehat{\theta}_{MLE}$, is the value of θ that maximizes the likelihood function. Intuitively, $\widehat{\theta}_{MLE}$ maximizes the likelihood (or probability) that the data comes from the specified distribution. Note that we haven't said anything about whether the random variables X_i are continuous or discrete, so that the p.d.f. entering the likelihood can be either a density or a probability mass function, or a hybrid between the two if the distribution is mixed continuous-discrete.

It is usually much easier to work with the logarithm of the likelihood function:

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(X_i, \theta)$$

Maximization of likelihood function (6) is equivalent to maximization of the logarithm of the likelihood function since the log transformation is strictly increasing. That is, the value of θ that maximizes any increasing function of $L(\theta; X_1, ..., X_n)$ will also maximize $L(\theta; X_1, ..., X_n)$. Thus, $\widehat{\theta}_{MLE}$ solves the problem:

$$\max_{\theta} \ln L(\theta) = \sum_{i=1}^{n} \ln f(X_i, \theta). \tag{7}$$

Assuming that $\ln [f(X_i, \theta)]$ is differentiable, the necessary condition for maximum is given by:

$$\frac{\partial \ln L(\widehat{\theta}_{MLE})}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln \left[f(X_i, \widehat{\theta}_{MLE}) \right]}{\partial \theta} = 0. \tag{8}$$

This necessary condition will often be also sufficient for maximum, and therefore, $\widehat{\theta}_{MLE}$ could be found by setting the first condition (8) to zero and solving for θ.

Example 1. Bernoulli Distribution

Let $X_1, ..., X_n$ be a random sample from the Bernoulli distribution with a probability distribution:

$$P(X = x) = \theta^x (1-\theta)^{1-x}, 0 < \theta < 1.$$

The joint likelihood function is then given by

$$L(\theta) = \prod_{i=1}^{n} \theta^{X_i} (1-\theta)^{1-X_i} = \theta^y (1-\theta)^{n-y}$$

where $y = \sum_{i=1}^{n} X_i$ is the number of times X takes on the value 1. Taking the natural logs gives

$$\ln L(\theta) = y \ln \theta + (n-y) \ln(1-\theta).$$

First, consider the case when $0 < y < n$, the differentiating and setting the derivative to zero yields

$$\frac{\partial \ln L}{\partial \theta} = \frac{y}{\widehat{\theta}} - \frac{n-y}{1-\widehat{\theta}} = 0 \Longrightarrow \widehat{\theta}_{MLE} = n^{-1} \sum_{i=1}^{n} X_i.$$

Example 2. Poisson Distribution

Let $X_1, ..., X_n$ be a random sample from the Poisson distribution:

$$\begin{aligned} f(x, \lambda) &= \lambda^x e^{-\lambda}/x!, x = 0, 1, 2, ...; 0 < \lambda < \infty \\ E(X) &= Var(X) = \lambda \end{aligned}$$

The likelihood and log likelihood functions are

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda} \lambda^{X_i}/X_i! = \frac{e^{-n\lambda} \lambda^{\Sigma X_i}}{\prod_{i=1}^{n} X_i!}$$

and

$$\ln L(\lambda) = -n\lambda + \sum_{i=1}^{n} X_i \ln \lambda - \ln \left[\prod_{i=1}^{n} X_i! \right].$$

Differentiating the log likelihood, we have

$$\frac{\partial \ln L}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i.$$

Setting the derivative to zero gives

$$-n + \frac{1}{\widehat{\lambda}} \sum_{i=1}^{n} X_i = 0 \Longrightarrow \widehat{\lambda}_{MLE} = n^{-1} \sum_{i=1}^{n} X_i = \overline{X}.$$

That is the MLE estimator of the mean of Poisson distribution is the same as the MM estimator and equals the sample mean, which, as we know, is unbiased.

Example 3. Normal distribution

Suppose $X \sim N(\mu, \sigma^2)$, and we want to estimate the parameters μ and σ^2 from an i.i.d. sample $X_1, \ldots, X_n$. The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

It turns out that it's much easier to maximize the log-likelihood,

$$\begin{aligned} \ln L(\theta) &= \sum_{i=1}^{n} \ln \left\{ \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} \right\} \\ &= \sum_{i=1}^{n} \left\{ \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(X_i-\mu)^2}{2\sigma^2} \right\} \\ &= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \end{aligned}$$

To find the maximum, we take the derivatives with respect to μ and σ^2, and set them equal to zero:

$$0 = \frac{1}{2\widehat{\sigma}^2} \sum_{i=1}^{n} 2(X_i - \hat{\mu}) \Leftrightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Thus, a MLE of μ is the sample mean, which was shown to be unbiased.

Similarly,

$$0 = -\frac{n}{2} \frac{2\pi}{2\pi\widehat{\sigma}^2} + \frac{1}{2\left(\widehat{\sigma}^2\right)^2} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 \Leftrightarrow \widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

As shown earlier, $\widehat{\sigma}^2$ is a biased estimator for σ^2. So, in general, MLE may be biased.