In this lecture, we present a fairly general method to construct estimators, for parameters of interest which may be more complicated to estimate than the parameters we considered thus far, which had natural sample analogs.

1 Maximum likelihood estimation

1.1 Maximum likelihood principle for discrete random variables

Suppose that we want to model a data set $a_1, a_2, \dots, a_n$ as the realization of random variables $X_1, X_2, \dots, X_n$ which are a random sample, understood to therefore be i.i.d., from a probability mass function p_θ which is characterized by θ, our parameter of interest (which is often a number, but could also be a vector of several quantities).

We define the joint probability mass function

$$p_{\theta,X_1,X_2,\dots,X_n}(x_1,\dots,x_n) = p_\theta(x_1)p_\theta(x_2)\dots p_\theta(x_n)$$

For a given, fixed data set $a_1, a_2, \dots, a_n$, we construct the **likelihood function**

$$L(\theta) = p_{\theta,X_1,X_2,\dots,X_n}(a_1,\dots,a_n) = p_\theta(a_1)p_\theta(a_2)\dots p_\theta(a_n) \qquad \text{Likelihood function, discrete random variable} \tag{1}$$

The idea behind the maximum likelihood principle is to say that we should choose θ in such a way that the probability of observing the actual data set $a_1, \dots, a_n$ is highest. In other words **we should choose θ in such a way that $L(\theta)$ is maximized**.

This idea was first proposed by the English statistician Ronald Fisher in 1912, and may be the most widely used method of estimation in statistics nowadays.

Definition: **The maximum likelihood estimate** of the parameter of interest θ is the value $t = h(a_1, a_2, \dots, a_n)$ that **maximizes the likelihood function** $L(\theta)$.

The corresponding random variable, $T = h(X_1, X_2, \dots, X_n)$ is called the **maximum likelihood estimator of θ**.

1.2 Maximum likelihood principle for continuous random variables

We will accept here, without proving it, that the usual analogy

$$\begin{array}{ccc} \text{discrete random variable} & \leftrightarrow & \text{continuous random variable} \\ \text{probability mass function} & \leftrightarrow & \text{probability density function} \end{array}$$

also applies to the maximum likelihood principle, which can therefore be stated in the same terms for continuous random variables as for discrete random variables, with the likelihood function defined by:

$$L(\theta) = f_{\theta,X_1,X_2,\dots,X_n}(a_1,\dots,a_n) = f_\theta(a_1)f_\theta(a_2)\dots f_\theta(a_n) \qquad \text{Likelihood function, continuous random variable} \tag{2}$$

where $f_{\theta,X_1,X_2,\dots,X_n}$ may now be interpreted as the joint probability density function, and f_θ as a probability density function.

1.3 Likelihood and loglikelihood

$L(\theta)$ as introduced in the previous sections is the product of n terms. Maximizing $L(\theta)$ can therefore be complicated and tedious, since the product rule will lead to n complicated terms when differentiating $L(\theta)$.

To address this difficulty, let us define the **loglikelihood function** according to:

$$l(\theta) = \ln\left[L(\theta)\right] \qquad \text{Log-likelihood function} \tag{3}$$

We have the following equivalences:

$$\frac{dl}{d\theta} = 0 \qquad \Leftrightarrow \qquad \frac{1}{L(\theta)}\frac{dL}{d\theta} = 0 \qquad \Leftrightarrow \qquad \frac{dL}{d\theta} = 0$$

We therefore conclude that **extremizing $l(\theta)$ is equivalent to extremizing $L(\theta)$**. And since ln is an increasing function, we can go further and say that maximizing $l(\theta)$ is equivalent to maximizing $L(\theta)$: **$l(\theta)$ and $L(\theta)$ will have the same maximum $\theta_{\max}$.**

Why is that interesting? Because finding the maximum of $l(\theta)$ is often much easier than finding the maximum of $L(\theta)$ since $l(\theta) = \ln[L(\theta)]$ is the *sum of n terms*:

$$l(\theta) = \ln[p_\theta(a_1)] + \ldots + \ln[p_\theta(a_n)] \qquad \text{(Discrete random variable)}$$
$$l(\theta) = \ln[f_\theta(a_1)] + \ldots + \ln[f_\theta(a_n)] \qquad \text{(Continuous random variable)}$$

The derivative of a sum of n terms is much more convenient to evaluate and use to find a maximum than the derivative of a product of n terms.

2 Examples

2.1 Example 1

Consider a discrete random variable whose probability mass function is

$$\begin{cases} p_X(0) = \frac{2\theta}{3} \\ p_X(1) = \frac{\theta}{3} \\ p_X(2) = \frac{2(1-\theta)}{3} \\ p_X(3) = \frac{1-\theta}{3} \\ p_X(x) = 0 \ \text{ if } x \notin \{0,1,2,3\} \end{cases}$$

where θ is the parameter of interest, with $0 \leq \theta \leq 1$.

The following 10 independent observations were taken from this distribution: $(3, 0, 2, 1, 3, 2, 1, 0, 2, 1)$.
What is the maximum likelihood estimate of θ?

We construct the likelihood function

$$L(\theta) = \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{1-\theta}{3}\right)^2$$

$L(\theta)$ clearly does not seem pleasant to maximize. Let us therefore consider the loglikelihood function $l(\theta)$.

$$\begin{aligned} l(\theta) = \ln[L(\theta)] &= 2\left(\ln\left(\frac{2}{3}\right) + \ln\theta\right) + 3\left(-\ln 3 + \ln\theta\right) + 3\left(\ln\left(\frac{2}{3}\right) + \ln(1-\theta)\right) + 2\left(-\ln 3 + \ln(1-\theta)\right) \\ &= 5\ln\theta + 5\ln(1-\theta) + K \end{aligned}$$

where K is a constant which we do not need to spend time computing, since we are looking for the maximum of l.

$$\begin{aligned} \frac{dl}{d\theta} = 0 \quad &\Leftrightarrow \quad \frac{5}{\theta} - \frac{5}{1-\theta} = 0 \\ &\Leftrightarrow \quad \theta = 1 - \theta \\ &\Leftrightarrow \quad \theta = \frac{1}{2} \end{aligned}$$

Let us verify that $\theta = \frac{1}{2}$ is indeed a maximum.

$$\frac{d^2l}{d\theta^2} = -\frac{5}{\theta^2} - \frac{5}{(1-\theta)^2} < 0 \quad \forall\, \theta$$

$\theta = \frac{1}{2}$ is the maximum likelihood estimate of θ.

2.2 Example 2

Let us return once more to the polling problem we first considered in Lectures 10 and 11.

$X_1, X_2, \ldots, X_n$ are independent Bernoulli random variables with unknown parameter of interest μ. Their probability mass function may be conveninently written as

$$p_{X_i}(x) = \mu^x(1-\mu)^{1-x} \quad \text{for } x = 0 \text{ or } x = 1 \qquad , \qquad p_{X_i}(x) = 0 \quad \text{if } x \notin \{0,1\}$$

Let $(x_1, x_2, \ldots, x_n)$ be the responses from the people polled about the upcoming election. The likelihood function is

$$L(\mu) = \mu^{x_1}(1-\mu)^{1-x_1} \ldots \mu^{x_n}(1-\mu)^{1-x_n} = \mu^{\sum_{i=1}^n x_i}(1-\mu)^{\sum_{i=1}^n (1-x_i)} = \mu^{\sum_{i=1}^n x_i}(1-\mu)^{n-\sum_{i=1}^n x_i}$$

Once again, it does not seem particularly pleasant to find the maximum of L. We therefore consider the log-likelihood function

$$l(\mu) = \ln[L(\mu)] = \sum_{i=1}^n x_i \ln \mu + \left(n - \sum_{i=1}^n x_i \right) \ln(1-\mu)$$

Let us look for the maximum of l.

$$\begin{aligned} \frac{dl}{d\mu} = 0 \quad &\Leftrightarrow \quad \frac{\sum_{i=1}^n x_i}{\mu} = \frac{n - \sum_{i=1}^n x_i}{1-\mu} \\ &\Leftrightarrow \quad \sum_{i=1}^n x_i - \mu \sum_{i=1}^n x_i = n\mu - \mu \sum_{i=1}^n x_i \\ &\Leftrightarrow \quad \mu = \frac{\sum_{i=1}^n x_i}{n} \end{aligned}$$

It is straightforward to verify, in much the same way as we did for the previous example, that this value of μ indeed corresponds to a maximum of l.

We conclude that the maximum likelihood estimate we obtain is the sample mean

$$\mu = \frac{\sum_{i=1}^n x_i}{n} = \overline{x_n}$$

The maximum likelihood estimator therefore is

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n}$$

2.3 Example 3: Estimating the parameters of the normal distribution

Suppose the data set $x_1, x_2, \ldots, x_n$ is a realization of a random sample from a normal distribution with mean μ and variance σ^2.

You will derive in recitation, and can also read in our textbook, that the maximum likelihood estimate for μ is the sample mean:

$$\overline{x_n} = \frac{x_1 + \ldots + x_n}{n} \qquad \text{Maximum likelihood estimate for } \mu$$

and the maximum likelihood estimate for σ^2 is:

$$\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x_n})^2 \qquad \text{Maximum likelihood estimate for } \sigma^2$$

Remember that in Lecture 13, we had found that

$$\hat{S}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$$

was an estimator for σ^2 which had a smaller MSE than

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$$

This is one reason for the popularity of $\hat{S}_n^2$. Now we see another reason for this popularity: $\hat{S}_n^2$ is the maximum likelihood estimator for σ^2.