Lecture 4

Let us go back to the example of exponential distribution $E(\alpha)$ from the last lecture and recall that we obtained two estimates of unknown parameter α_0 using the first and second moment in the method of moments. We had:

1. Estimate of α_0 using first moment:

$$g(X) = X, \ \ m(\alpha) = \mathbb{E}_\alpha g(X) = \frac{1}{\alpha}, \ \ \hat{\alpha}_1 = m^{-1}(\bar{g}) = \frac{1}{\bar{X}}.$$

2. Estimate of α using second moment:

$$g(X) = X^2, \ \ m(\alpha) = \mathbb{E}_\alpha g(X^2) = \frac{2}{\alpha^2}, \ \ \hat{\alpha}_2 = m^{-1}(\bar{g}) = \sqrt{\frac{2}{\overline{X^2}}}.$$

How to decide which method is better? The asymptotic normality result states:

$$\sqrt{n}(m^{-1}(\bar{g}) - \theta_0) \to N\left(0, \frac{\operatorname{Var}_{\theta_0}(g(X))}{(m'(\theta_0))^2}\right).$$

It makes sense to compare two estimates by comparing their asymptotic variance. Let us compute it in both cases:

1. In the first case:

$$\frac{\operatorname{Var}_{\alpha_0}(g(X))}{(m'(\alpha_0))^2} = \frac{\operatorname{Var}_{\alpha_0}(X)}{(-\frac{1}{\alpha_0^2})^2} = \frac{\frac{1}{\alpha_0^2}}{(-\frac{1}{\alpha_0^2})^2} = \alpha_0^2.$$

In the second case we will need to compute the fourth moment of the exponential distribution. This can be easily done by integration by parts but we will show a different way to do this.

The moment generating function of the distribution $E(\alpha)$ is:

$$\varphi(t) = \mathbb{E}_\alpha e^{tX} = \int_0^\infty e^{tx} \alpha e^{-\alpha x} dx = \frac{\alpha}{\alpha - t} = \sum_{k=0}^\infty \frac{t^k}{\alpha^k},$$

where in the last step we wrote the usual Taylor series. On the other hand, writing the Taylor series for e^{tX} we can write,

$$\varphi(t) = \mathbb{E}_\alpha e^{tX} = \mathbb{E}_\alpha \sum_{k=0}^{\infty} \frac{(tX)^k}{k!} = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbb{E}_\alpha X^k.$$

Comparing the two series above we get that the kth moment of exponential distribution is

$$\mathbb{E}_\alpha X^k = \frac{k!}{\alpha^k}.$$

2. In the second case:

$$\frac{\mathrm{Var}_{\alpha_0}(g(X))}{(m'(\alpha_0))^2} = \frac{\mathrm{Var}_{\alpha_0}(X^2)}{(-\frac{4}{\alpha_0^3})^2} = \frac{\mathbb{E}_{\alpha_0} X^4 - (\mathbb{E}_{\alpha_0} X^2)^2}{(-\frac{4}{\alpha_0^3})^2} = \frac{\frac{4!}{\alpha_0^4} - (\frac{2}{\alpha_0^2})^2}{(-\frac{4}{\alpha_0^3})^2} = \frac{5}{4} \alpha_0^2$$

Since the asymptotic variance in the first case is less than the asymptotic variance in the second case, the first estimator seems to be better.

4.1 Maximum likelihood estimators.

(Textbook, Section 6.5)

As always we consider a parametric family of distributions $\{\mathbb{P}_\theta, \theta \in \Theta\}$. Let $f(X|\theta)$ be either a probability function or a probability density function of the distribution $\mathbb{P}_\theta$. Suppose we are given a sample $X_1, \ldots, X_n$ with unknown distribution $\mathbb{P}_\theta$, i.e. θ is unknown. Let us consider a *likelihood function*

$$\varphi(\theta) = f(X_1|\theta) \times \ldots \times f(X_n|\theta)$$

seen as a function of the parameter θ only. It has a clear interpretation. For example, if our distributions are discrete then the probability function

$$f(x|\theta) = \mathbb{P}_\theta(X = x)$$

is a probability to observe a point x and the likelihood function

$$\varphi(\theta) = f(X_1|\theta) \times \ldots \times f(X_n|\theta) = \mathbb{P}_\theta(X_1) \times \ldots \times \mathbb{P}_\theta(X_n) = \mathbb{P}_\theta(X_1, \cdots, X_n)$$

is the probability to observe the sample $X_1, \ldots, X_n$.

In the continuous case the likelihood function $\varphi(\theta)$ is the probability density function of the vector $(X_1, \cdots, X_n)$.

Definition: (Maximum Likelihood Estimator.) Let $\hat{\theta}$ be the parameter that maximizes $\varphi(\theta)$, i.e.

$$\varphi(\hat{\theta}) = \max_{\theta} \varphi(\theta).$$

Then $\hat{\theta}$ is called the maximum likelihood estimator (MLE).

To make our discussion as simple as possible, let us assume that the likelihood function behaves like shown on the figure 4.1, i.e. the maximum is achieved at the unique point $\hat{\theta}$.
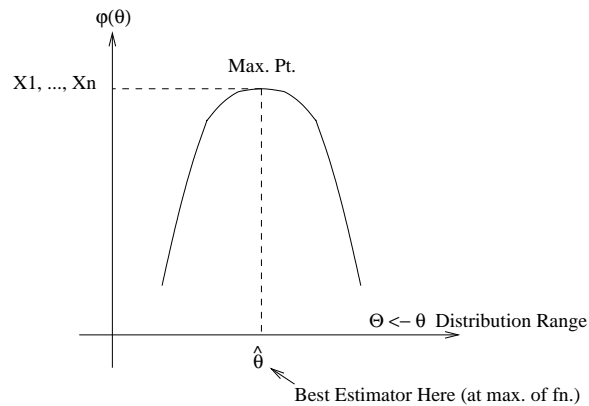

Figure 4.1: Maximum Likelihood Estimator (MLE)

When finding the MLE it sometimes easier to maximize the log-likelihood function since

$$\varphi(\theta) \rightarrow \text{ maximize } \Leftrightarrow \log \varphi(\theta) \rightarrow \text{ maximize}$$

maximizing φ is equivalent to maximizing $\log \varphi$. Log-likelihood function can be written as

$$\log \varphi(\theta) = \sum_{i=1}^{n} \log f(X_i | \theta).$$

Let us give several examples of MLE.

Example 1. Bernoulli distribution $B(p)$.

$$\mathcal{X} = \{0, 1\}, \ \mathbb{P}(X = 1) = p, \ \mathbb{P}(X = 0) = 1 - p, \ p \in [0, 1].$$

Probability function in this case is given by

$$f(x|p) = \begin{cases} p, & x = 1, \\ 1 - p, & x = 0. \end{cases}$$

Likelihood function

$$\begin{aligned} \varphi(p) &= f(X_1|p)f(X_2|p)\dots f(X_n|p) \\ &= p^{\# \ of \ 1's}(1-p)^{\# \ of \ 0's} = p^{X_1+\dots+X_n}(1-p)^{n-(X_1+\dots+X_n)} \end{aligned}$$

and the log-likelihood function

$$\log \varphi(p) = (X_1 + \cdots + X_n) \log p + (n - (X_1 + \cdots + X_n)) \log(1 - p).$$

To maximize this over p let us find the critical point $\frac{d \log \varphi(p)}{dp} = 0$,

$$(X_1 + \cdots + X_n)\frac{1}{p} - (n - (X_1 + \cdots + X_n))\frac{1}{1 - p} = 0.$$

Solving this for p gives,

$$p = \frac{X_1 + \cdots + X_n}{n} = \bar{X}$$

and therefore $\hat{p} = \bar{X}$ is the MLE.

Example 2. Normal distribution $N(\alpha, \sigma^2)$ has p.d.f.

$$f(X|(\alpha, \sigma^2)) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-\alpha)^2}{2\sigma^2}}.$$

likelihood function

$$\varphi(\alpha, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i - \alpha)^2}{2\sigma^2}}.$$

and log-likelihood function

$$\begin{aligned}
\log \varphi(\alpha, \sigma^2) &= \sum_{i=1}^{n} \left(\log \frac{1}{\sqrt{2\pi}} - \log \sigma - \frac{(X - \alpha)^2}{2\sigma^2} \right) \\
&= n \log \frac{1}{\sqrt{2\pi}} - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \alpha)^2.
\end{aligned}$$

We want to maximize the log-likelihood with respect to α and σ^2. First, obviously, for any σ we need to minimize $\sum(X_i - \alpha)^2$ over α. The critical point condition is

$$\frac{d}{d\alpha} \sum (X_i - \alpha)^2 = -2 \sum (X_i - \alpha) = 0.$$

Solving this for α gives that $\hat{\alpha} = \bar{X}$. Next, we need to maximize

$$n \log \frac{1}{\sqrt{2\pi}} - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

over σ. The critical point condition reads,

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum (X_i - \bar{X})^2 = 0$$

and solving this for σ we get

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

is the MLE of σ^2.