Lecture 3

3.1 Method of moments.

Consider a family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$ and and consider a sample $X = (X_1, \ldots, X_n)$ of i.i.d. random variables with distribution $\mathbb{P}_{\theta_0}$, where $\theta_0 \in \Theta$. We assume that θ_0 is unknown and we want to construct an estimate $\hat{\theta} = \hat{\theta}_n(X_1, \cdots, X_n)$ of θ_0 based on the sample X.

Let us recall some standard facts from probability that we be often used throughout this course.

- **Law of Large Numbers (LLN):**

  If the distribution of the i.i.d. sample $X_1, \ldots, X_n$ is such that X_1 has finite expectation, i.e. $|\mathbb{E}X_1| < \infty$, then the sample average

  $$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n} \to \mathbb{E}X_1$$

  converges to the expectation in some sense, for example, for any arbitrarily small $\varepsilon > 0$,

  $$\mathbb{P}(|\bar{X}_n - \mathbb{E}X_1| > \epsilon) \to 0 \text{ as } n \to \infty.$$

  Convergence in the above sense is called convergence in probability.

  Note. Whenever we will use the LLN below we will simply say that the average converges to the expectation and will not mention in what sense. More mathematically inclined students are welcome to carry out these steps more rigorously, especially when we use LLN in combination with the Central Limit Theorem.

- **Central Limit Theorem (CLT):**

  If the distribution of the i.i.d. sample $X_1, \ldots, X_n$ is such that X_1 has finite expectation and variance, i.e. $|\mathbb{E}X_1| < \infty$ and $\mathrm{Var}(X) < \infty$, then

  $$\sqrt{n}(\bar{X}_n - \mathbb{E}X_1) \to^d N(0, \sigma^2)$$

converges in distribution to normal distribution with zero mean and variance σ^2, which means that for any interval $[a, b]$,

$$\mathbb{P}\Big(\sqrt{n}(\bar{X}_n - \mathbb{E}X_1) \in [a, b]\Big) \to \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx.$$

Motivating example. Consider a family of normal distributions

$$\{N(\alpha, \sigma^2) : \alpha \in \mathbb{R}, \sigma^2 \geq 0\}.$$

Consider a sample $X_1, \ldots, X_n \sim N(\alpha_0, \sigma_0^2)$ with distribution from this family and suppose that the parameters α_0, σ_0 are unknown. If we want to estimate these parameters based on the sample then the law of large numbers above provides a natural way to do this. Namely, LLN tells us that

$$\hat{\alpha} = \bar{X}_n \to \mathbb{E}X_1 = \alpha_0 \text{ as } n \to \infty$$

and, similarly,

$$\frac{X_1^2 + \ldots + X_n^2}{n} \to \mathbb{E}X_1^2 = \mathrm{Var}(X) + \mathbb{E}X^2 = \sigma_0^2 + \alpha_0^2.$$

These two facts imply that

$$\hat{\sigma}^2 = \frac{X_1^2 + \cdots + X_n^2}{n} - \Big(\frac{X_1 + \cdots + X_n}{n}\Big)^2 \to \mathbb{E}X^2 - (\mathbb{E}X)^2 = \sigma_0^2.$$

It, therefore, makes sense to take $\hat{\alpha}$ and $\hat{\sigma}^2$ as the estimates of unknown α_0, σ_0^2 since by the LLN for large sample size n these estimates will approach the unknown parameters.

We can generalize this example as follows.

Suppose that the parameter set $\Theta \subseteq \mathbb{R}$ and suppose that we can find a function $g : \mathcal{X} \to \mathbb{R}$ such that a function

$$m(\theta) = \mathbb{E}_\theta g(X) : \Theta \to \mathrm{Im}(\Theta)$$

has a continuous inverse m^{-1}. Here $\mathbb{E}_\theta$ denotes the expectation with respect to the distribution $\mathbb{P}_\theta$. Take

$$\hat{\theta} = m^{-1}(\bar{g}) = m^{-1}\Big(\frac{g(X_1 + \cdots + g(X_n)}{n}\Big)$$

as the estimate of θ_0. (Here we implicitely assumed that $\bar{g}$ is always in the set $\mathrm{Im}(m)$.) Since the sample comes from distribution with parameter θ_0, by LLN we have

$$\bar{g} \to \mathbb{E}_{\theta_0} g(X_1) = m(\theta_0).$$

Since the inverse m^{-1} is continuous, this implies that our estimate

$$\hat{\theta} = m^{-1}(\bar{g}) \to m^{-1}(m(\theta_0)) = \theta_0$$

converges to the unkown parameter θ_0.

Typical choices of the function g are $g(x) = x$ or x^2. The quantity $\mathbb{E}X^k$ is called the k^{th} moment of X and, hence, the name - *method of moments*.

Example: Family of exponential distributions $E(\alpha)$ with p.d.f.

$$p(x) = \begin{cases} \alpha e^{-\alpha x}, & x \geq 0, \\ 0, & x < 0 \end{cases}$$

Take $g(x) = x$. Then

$$m(\alpha) = \mathbb{E}_\alpha g(X) = \mathbb{E}_\alpha X = \frac{1}{\alpha}.$$

($\frac{1}{\alpha}$ is the expectation of exponential distribution, see Pset 1.) Let us recall that we can find inverse by solving for α the equation

$$m(\alpha) = \beta, \text{ i.e. in our case } \frac{1}{\alpha} = \beta.$$

We have,

$$\alpha = m^{-1}(\beta) = \frac{1}{\beta}.$$

Therefore, we take

$$\hat{\alpha} = m^{-1}(\bar{g}) = m^{-1}(\bar{X}) = \frac{1}{\bar{X}}$$

as the estimate of unkown α_0.

Take $g(x) = x^2$. Then

$$m(\alpha) = \mathbb{E}_\alpha g(X^2) = \mathbb{E}_\alpha X^2 = \frac{2}{\alpha^2}.$$

The inverse is

$$\alpha = m^{-1}(\beta) = \sqrt{\frac{2}{\beta}}$$

and we take

$$\hat{\alpha} = m^{-1}(\bar{g}) = m^{-1}(\bar{X^2}) = \sqrt{\frac{2}{\bar{X^2}}}$$

as another estimate of α_0.

The question is, which estimate is better?

1. **Consistency.** We say that an estimate $\hat{\theta}$ is consistent if $\hat{\theta} \to \theta_0$ in probability as $n \to \infty$.. We have shown above that by construction the estimate by method of moments is always consistent.

2. **Asymptotic Normality.** We say that $\hat{\theta}$ is asymptotically normal if

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N(0, \sigma^2_{\theta_0})$$

where $\sigma^2_{\theta_0} \equiv$ is called the asymptotic variance of the estimate $\hat{\theta}$.

Theorem. *The estimate $\hat{\theta} = m^{-1}(\bar{g})$ by the method of moments is asymptotically normal with asymptotic variance*

$$\sigma^2_{\theta_0} = \frac{Var_{\theta_0}(g)}{(m'(\theta_0))^2}.$$

Proof. Writing Taylor expansion of the function m^{-1} at point $m(\theta_0)$ we have

$$m^{-1}(\bar{g}) = m^{-1}(m(\theta_0)) + (m^{-1})'(m(\theta_0))(\bar{g} - m(\theta_0)) + \frac{(m^{-1})''(c)}{2!}(\bar{g} - m(\theta_0))^2$$

where $c \in [m(\theta_0), \bar{g}]$. Since $m^{-1}(m(\theta_0)) = \theta_0$, we get

$$m^{-1}(\bar{g}) - \theta_0 = (m^{-1})'(m(\theta_0))(\bar{g} - m(\theta_0) + \frac{(m^{-1})''(c)}{2!})(\bar{g} - m(\theta_0)^2$$

Let us prove that the left hand side multiplied by $\sqrt{n}$ converges in distribution to normal distribution.

$$\sqrt{n}(m^{-1}(\bar{g}) - \theta_0) = (m^{-1})'(m(\theta_0))\underbrace{\sqrt{n}(\bar{g} - m(\theta_0))}+\frac{(m^{-1})''(c)}{2!}\frac{1}{\sqrt{n}}(\underbrace{\sqrt{n}(\bar{g} - m(\theta_0))})^2 \tag{3.1}$$

Let us recall that

$$\bar{g} = \frac{g(X_1) + \cdots + g(X_n)}{n}, \mathbb{E}g(X_1) = m(\theta_0).$$

Central limit theorem tells us that

$$\sqrt{n}(\bar{g} - m(\theta_0) \to N(0, \mathrm{Var}_{\theta_0}(g(X_1)))$$

where convergence is in distribution. First of all, this means that the last term in (3.1) converges to 0 (in probability), since it has another factor of $1/\sqrt{n}$. Also, since from calculus the derivative of the inverse

$$(m^{-1})'(m(\theta_0)) = \frac{1}{m'(m^{-1}(m(\theta_0)))} = \frac{1}{m'(\theta_0)},$$

the first term in (3.1) converges in distribution to

$$(m^{-1})'(m(\theta_0))\sqrt{n}(m^{-1}(\bar{g}) - \theta_0) \to \frac{1}{m'(\theta_0)} N(0, \mathrm{Var}_{\theta_0}(g(X_1))) = N\Big(0, \frac{Var_{\theta_0}(g(X_1))}{(m'(\theta_0))^2}\Big)$$

□

What this result tells us is that the smaller $\frac{Var_{\theta_0}(g)}{m'(\theta_0)}$ is the better is the estimate $\hat{\theta}$ in the sense that it has smaller deviations from the unknown parameter θ_0 asymptotically.