STAT 423/523 Statistical Methods for Engineers and Scientists

Lecture 3: Point Estimation II

Chencheng Cai

Washington State University

Methods of Point Estimation

We have discussed the definitions and properties of the estimators.

Methods of Point Estimation

We have discussed the definitions and properties of the estimators.
Now we introduce some methods to construct point estimators:

- Method of Moments (MoM)
- Maximum Likelihood Estimation (MLE)

Method of Moments (MoM)

Definition (Moments)

Let $X_1, \ldots, X_n$ be a random sample from a population with pmf or pdf $f(x)$. For $k = 1, 2, \ldots$, the **kth population moment** or **kth moment of the distribution** $f(x)$, is $E(X^k)$. The **kth sample moment** is

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Method of Moments (MoM)

Definition (Moments)

Let $X_1, \ldots, X_n$ be a random sample from a population with pmf or pdf $f(x)$. For $k = 1, 2, \ldots$, the **kth population moment** or **kth moment of the distribution** $f(x)$, is $E(X^k)$. The **kth sample moment** is

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

- kth moment of a distribution is the expected value of X^k.
- kth sample moment is the sample average of X^k.
- When $n \to \infty$, the two moments are equal (by Law of Large Numbers).

Method of Moments (MoM)

Let $X_1, \ldots, X_n$ be a random sample from a population with pmf or pdf $f(x; \theta_1, \ldots, \theta_m)$, where $\theta_1, \ldots, \theta_m$ are the unknown parameters we want to estimate.

Method of Moments (MoM)

Let $X_1, \ldots, X_n$ be a random sample from a population with pmf or pdf $f(x; \theta_1, \ldots, \theta_m)$, where $\theta_1, \ldots, \theta_m$ are the unknown parameters we want to estimate.

The **method of moments estimator** of $\theta_1, \ldots, \theta_m$ is the solution to the following system of equations:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = E(X) = g_1(\theta_1, \ldots, \theta_m)$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = E(X^2) = g_2(\theta_1, \ldots, \theta_m)$$

$$\vdots$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^m = E(X^k) = g_m(\theta_1, \ldots, \theta_m)$$

Method of Moments (MoM)

Let $X_1, \ldots, X_n$ be a random sample from a population with pmf or pdf $f(x; \theta_1, \ldots, \theta_m)$, where $\theta_1, \ldots, \theta_m$ are the unknown parameters we want to estimate.

The **method of moments estimator** of $\theta_1, \ldots, \theta_m$ is the solution to the following system of equations:

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} X_i &= E(X) = g_1(\theta_1, \ldots, \theta_m) \\
\frac{1}{n}\sum_{i=1}^{n} X_i^2 &= E(X^2) = g_2(\theta_1, \ldots, \theta_m) \\
&\vdots \\
\frac{1}{n}\sum_{i=1}^{n} X_i^m &= E(X^k) = g_m(\theta_1, \ldots, \theta_m)
\end{aligned}
$$

In short: **MoM matches the sample moments with the population moments.**

Example

Let $X_1, \ldots, X_n$ be a random sample from a population with unknown mean μ and unknown variance σ^2.

Example

Let $X_1, \ldots, X_n$ be a random sample from a population with unknown mean μ and unknown variance σ^2.

MoM matches the first two moments:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = E(X) = \mu$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = E(X^2) = E(X)^2 + \mathrm{Var}(X) = \mu^2 + \sigma^2$$

Example

Let $X_1, \ldots, X_n$ be a random sample from a population with unknown mean μ and unknown variance σ^2.

MoM matches the first two moments:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = E(X) = \mu$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = E(X^2) = E(X)^2 + \mathrm{Var}(X) = \mu^2 + \sigma^2$$

The solution, the MoM estimator, is

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

Example (Textbook 6.13)

Let $X_1, \ldots, X_n$ be a random sample from a Gamma distribution with parameters α and β. The pdf is

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

Example (Textbook 6.13)

Let $X_1, \ldots, X_n$ be a random sample from a Gamma distribution with parameters α and β. The pdf is

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

The first two moments of the Gamma distribution are

$$E(X) = \alpha\beta, \quad E(X^2) = \alpha(\alpha+1)\beta^2.$$

Example (Textbook 6.13)

Let $X_1, \ldots, X_n$ be a random sample from a Gamma distribution with parameters α and β. The pdf is

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

The first two moments of the Gamma distribution are

$$E(X) = \alpha\beta, \quad E(X^2) = \alpha(\alpha+1)\beta^2.$$

The MoM estimator of α and β are the solutions to the following equations:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = E(X) = \alpha\beta$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = E(X^2) = \alpha(\alpha+1)\beta^2$$

Example (Textbook 6.13)

$$\frac{1}{n}\sum_{i=1}^{n} X_i = E(X) = \alpha\beta$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = E(X^2) = \alpha(\alpha+1)\beta^2$$

The solutions are

$$\hat{\alpha} = \frac{\bar{X}^2}{\overline{X^2} - \bar{X}^2}$$
$$\hat{\beta} = \frac{\overline{X^2} - \bar{X}^2}{\bar{X}}$$

Method of Moments

- MoM only requires the first few moments of the distribution. (Not the explicit pmf or pdf)
- If the first m moments do not give a unique solution, we can use more moments.
- MoM estimator is approximately normal if the sample size is large enough (by CLT).

Maximum Likelihood Estimation (MLE)

Definition (Likelihood Function)

Let $X_1, \ldots, X_n$ be a random sample from a population with pmf or pdf $f(x; \theta_1, \ldots, \theta_m)$. The **likelihood function** is

$$L(\theta_1, \ldots, \theta_m) = \prod_{i=1}^{n} f(x_i; \theta_1, \ldots, \theta_m).$$

Maximum Likelihood Estimation (MLE)

Definition (Likelihood Function)

Let $X_1, \ldots, X_n$ be a random sample from a population with pmf or pdf $f(x; \theta_1, \ldots, \theta_m)$. The **likelihood function** is

$$L(\theta_1, \ldots, \theta_m) = \prod_{i=1}^{n} f(x_i; \theta_1, \ldots, \theta_m).$$

- The likelihood function is a function of the parameters $\theta_1, \ldots, \theta_m$.
- though it has exactly the same formula as the joint pmf or pdf of the sample.
- In order to compute the likelihood, we need to know the pmf or pdf explicitly. (compare it to MoM)

Important Clarificaitons on Likelihood

- The likelihood function is **the probability of observing the sample** given the parameters.
- **NOT** the probability of the parameters given the sample.

Important Clarificaitons on Likelihood

- The likelihood function is **the probability of observing the sample** given the parameters.
- **NOT** the probability of the parameters given the sample.
- That is

$$L(\theta_1, \ldots, \theta_m) \neq p(\theta_1, \ldots, \theta_m \mid X_1, \ldots, X_n)$$

- The r.h.s. of above is

$$p(\theta_1, \ldots, \theta_m \mid X_1, \ldots, X_n) = \frac{p(X_1, \ldots, X_n \mid \theta_1, \ldots, \theta_m) p(\theta_1, \ldots, \theta_m)}{p(X_1, \ldots, X_n)}$$

but we assume $\theta_1, \ldots, \theta_m$ to be fixed.

Example (Textbook 6.15)

Suppose 10 email accounts are randomly sampled and the 1st, 3rd and 10th accounts are found to have strong passwords. We want to estiamte p: the proportion of email accounts with strong passwords.

Example (Textbook 6.15)

Suppose 10 email accounts are randomly sampled and the 1st, 3rd and 10th accounts are found to have strong passwords. We want to estiamte p: the proportion of email accounts with strong passwords.

Let the random variables $X_1, \ldots, X_{10}$ be the indicator variables for the 10 accounts to have strong passwords.

Example (Textbook 6.15)

Suppose 10 email accounts are randomly sampled and the 1st, 3rd and 10th accounts are found to have strong passwords. We want to estiamte p: the proportion of email accounts with strong passwords.

Let the random variables $X_1, \ldots, X_{10}$ be the indicator variables for the 10 accounts to have strong passwords.

The likelihood function is

$$L(p) = f(x_1, \ldots, x_{10}; p) = p(1-p)p(1-p)\cdots p = p^3(1-p)^7.$$

Example (Textbook 6.15)

Suppose 10 email accounts are randomly sampled and the 1st, 3rd and 10th accounts are found to have strong passwords. We want to estiamte p: the proportion of email accounts with strong passwords.

Let the random variables $X_1, \ldots, X_{10}$ be the indicator variables for the 10 accounts to have strong passwords.

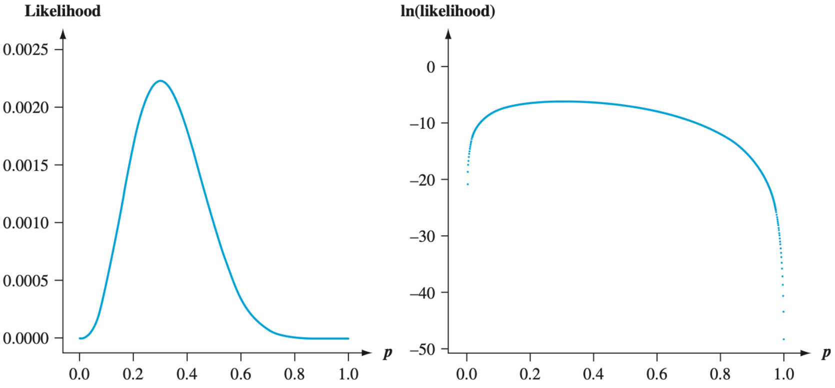
The likelihood function is

$$L(p) = f(x_1, \ldots, x_{10}; p) = p(1-p)p(1-p)\cdots p = p^3(1-p)^7.$$

The logarithm of the likelihood function is called the **log-likelihood function**:

$$\ell(p) := \log L(p) = 3\log p + 7\log(1-p).$$

Example (Textbook 6.15)



The intuitively best guess of p is the value that maximizes the likelihood function.

Maximum Likelihood Estimation (MLE)

The **maximum likelihood estimator** of $\theta_1, \ldots, \theta_m$ is the value of $\theta_1, \ldots, \theta_m$ that maximizes the likelihood function $L(\theta_1, \ldots, \theta_m)$.

The **log-likelihood function** is

$$\ell(\theta_1, \ldots, \theta_m) = \log L(\theta_1, \ldots, \theta_m).$$

Maximum Likelihood Estimation (MLE)

The **maximum likelihood estimator** of $\theta_1, \ldots, \theta_m$ is the value of $\theta_1, \ldots, \theta_m$ that maximizes the likelihood function $L(\theta_1, \ldots, \theta_m)$.

The **log-likelihood function** is

$$\ell(\theta_1, \ldots, \theta_m) = \log L(\theta_1, \ldots, \theta_m).$$

In many cases, the MLE of $\theta_1, \ldots, \theta_m$ is the solution to the following system of equations: The MLE of $\theta_1, \ldots, \theta_m$ is the solution to the following system of equations:

$$\frac{\partial \ell}{\partial \theta_1} = 0, \ldots, \frac{\partial \ell}{\partial \theta_m} = 0.$$

The first order derivatives are called the **score functions**. The MLE is a zero of the score functions.

Example (Textbook 6.15) Cont.

Continue the example of passwords. The score function is

$$\frac{d\ell(p)}{dp} = \frac{d(3\log p + 7\log(1-p))}{dp} = \frac{3}{p} - \frac{7}{1-p}.$$

Example (Textbook 6.15) Cont.

Continue the example of passwords. The score function is

$$\frac{d\ell(p)}{dp} = \frac{d(3\log p + 7\log(1-p))}{dp} = \frac{3}{p} - \frac{7}{1-p}.$$

The MLE is the solution to

$$\frac{3}{p} - \frac{7}{1-p} = 0.$$

The solution is $\hat{p} = 3/10$.

Example

Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with parameter $\lambda > 0$. The pdf is

$$f(x; \lambda) = \lambda e^{-\lambda x}.$$

Example

Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with parameter $\lambda > 0$. The pdf is

$$f(x; \lambda) = \lambda e^{-\lambda x}.$$

The likelihood function is

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

The log-likelihood function is

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} X_i.$$

Example

Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with parameter $\lambda > 0$. The pdf is

$$f(x; \lambda) = \lambda e^{-\lambda x}.$$

The likelihood function is

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

The log-likelihood function is

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} X_i.$$

The score function is

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i.$$

Example

The MLE is the solution to

$$\frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0.$$

The solution is $\hat{\lambda} = n / \sum_{i=1}^{n} X_i = 1/\bar{X}$.

Example

The MLE is the solution to

$$\frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0.$$

The solution is $\hat{\lambda} = n / \sum_{i=1}^{n} X_i = 1/\bar{X}$.

- $\hat{\lambda}$ is biased because (Jensen's inequality)

$$E\left(\frac{1}{\bar{X}}\right) > \frac{1}{E(\bar{X})} = \lambda.$$

- The MoM estimator for λ is $\hat{\lambda} = 1/\bar{X}$ as well (based on the first moment).

Example

Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. The likelihood function is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left(-\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{2\sigma^2}\right)$$

Example

Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. The likelihood function is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left(-\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{2\sigma^2}\right)$$

The log-likelihood function is

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2.$$

The score functions are

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu), \qquad \frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Example

The MLE is the solution to

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)=0, \quad -\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(X_i-\mu)^2=0.$$

The solution is

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

The MLE of μ is the sample mean, and the MLE of σ^2 is the sample variance.

Example

The MLE is the solution to

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)=0, \quad -\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(X_i-\mu)^2=0.$$

The solution is

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

The MLE of μ is the sample mean, and the MLE of σ^2 is the sample variance.

- $\hat{\mu}$ is unbiased.
- $\hat{\sigma}^2$ is biased.

Maximum Likelihood Estimation

- MLE is not unique. (The likelihood function may have multiple maxima.)
- MLE is not always the solution to the score functions. (The score functions may not have zeros.)
- When the sample size is large enough and the MLE is a zero of the score functions, the MLE is approximately normal (by CLT).
- When the sample size is large enough, the MLE is approximately unbiased (by Law of Large Numbers).
- When the sample size is large enough, the MLE is approximately efficient (with smallest variance).
- MLE is transformation invariant. (If $\hat{\theta}$ is the MLE of θ, then $\phi(\hat{\theta})$ is the MLE of $\phi(\theta)$ for any function ϕ).

Example: Domain-related Distribution

Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on the interval $[0, \theta]$. The pdf is

$$f(x; \theta) = \frac{1}{\theta} I(0 \leq x \leq \theta).$$

Example: Domain-related Distribution

Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on the interval $[0, \theta]$. The pdf is

$$f(x; \theta) = \frac{1}{\theta} I(0 \leq x \leq \theta).$$

The likelihood function is

$$L(\theta) = \frac{1}{\theta^n} I(\max(X_1, \ldots, X_n) \leq \theta).$$

Example: Domain-related Distribution

Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on the interval $[0, \theta]$. The pdf is

$$f(x; \theta) = \frac{1}{\theta} I(0 \leq x \leq \theta).$$

The likelihood function is

$$L(\theta) = \frac{1}{\theta^n} I(\max(X_1, \ldots, X_n) \leq \theta).$$

The likelhood function is monotone decreasing in $\theta > X_{max}$. The MLE is $\hat{\theta} = X_{max}$.

Example: Domain-related Distribution

Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on the interval $[0, \theta]$. The pdf is

$$f(x; \theta) = \frac{1}{\theta} I(0 \leq x \leq \theta).$$

The likelihood function is

$$L(\theta) = \frac{1}{\theta^n} I(\max(X_1, \ldots, X_n) \leq \theta).$$

The likelhood function is monotone decreasing in $\theta > X_{max}$. The MLE is $\hat{\theta} = X_{max}$.

- $\hat{\theta}$ is biased. (because $E(X_{max}) < \theta$)
- When the sample size is large enough, $\hat{\theta}$ is **not** approximately normal.

Example: Domain-related Distribution

If we consider MoM for the same problem. The MoM estimator is the solution to

$$\frac{1}{n}\sum_{i=1}^{n} X_i = E(X) = \frac{\theta}{2}.$$

The MoM estimator is $\hat{\theta} = 2\bar{X}$.

Example: Domain-related Distribution

If we consider MoM for the same problem. The MoM estimator is the solution to

$$\frac{1}{n}\sum_{i=1}^{n} X_i = E(X) = \frac{\theta}{2}.$$

The MoM estimator is $\hat{\theta} = 2\bar{X}$.

- $\hat{\theta}$ is unbiased.
- When the sample size is large enough, $\hat{\theta}$ is approximately normal.
- However, it could happen that $\hat{\theta} < X_{max}$.