IE605: Engineering Statistics

Lecture 11

Manjesh K. Hanawal

Previous Lecture:

- Data Reduction
- Sufficiency principle
- Sufficient Statistics
- Factorization theorem
- Minimum statistics
- Ancillary statistics

This Lecture:

- Likelihood functions

Likelihood function

Given that $\boldsymbol{X} = \boldsymbol{x}$ is observed, the function $\boldsymbol{\theta}$ defined by

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}|\boldsymbol{\theta}) & \text{if } \boldsymbol{X} \text{ is continuous} \\ P_\theta(\boldsymbol{X} = \boldsymbol{x}) & \text{if } \boldsymbol{X} \text{ is discrete} \end{cases}$$

When a sample x is observed. For a given parameters $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ if

$$P_{\boldsymbol{\theta}_1}(\boldsymbol{X} = \boldsymbol{x}) = L(\boldsymbol{\theta}_1|\boldsymbol{x}) > L(\boldsymbol{\theta}_2|\boldsymbol{x}) = P_\theta(\boldsymbol{X} = \boldsymbol{x})$$

Sample would have come more likely from parameter $\boldsymbol{\theta}_1$ than $\boldsymbol{\theta}_2$.
$\boldsymbol{\theta}_1$ is more **plausible** value of true parameter $\boldsymbol{\theta}$ than $\boldsymbol{\theta}_2$

Examples

- Binomial: Suppose X is Binomial with known n and unknown p. If $X = x$ is observed, then the likelihood function is

$$L(p|x) = P_p(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

- Exponential: if $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, where X_is are iid exponential with unknown λ

$$L(\lambda|x) = f(x|\lambda) = \prod_{i=1}^{n} \lambda \exp\{-\lambda x_i\}$$

- Gaussian: if $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, where X_is are iid Gaussian with unknown μ and known σ^2

$$L(\mu|x) = f(x|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(x_i - \mu)^2/2\sigma^2\}$$

Likelihood Principle

If samples $\boldsymbol{x}$ and $\boldsymbol{y}$ are such that $L(\boldsymbol{\theta}|\boldsymbol{x}) = C(\boldsymbol{x}, \boldsymbol{y})L(\boldsymbol{\theta}|\boldsymbol{y})$ for all $\boldsymbol{\theta}$, then conclusion drawn from $\boldsymbol{x}$ and $\boldsymbol{y}$ are identical. $C(\boldsymbol{x}, \boldsymbol{y})$ is independent of $\boldsymbol{\theta}$.

Example: Let $\boldsymbol{X} = (X_1, X_2, \ldots X_n)$ are iid Gaussian with unknown μ and known σ^2. For any samples $\boldsymbol{x}$ and $\boldsymbol{y}$

$$L(\mu|\boldsymbol{x}) = f(\boldsymbol{x}|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(x_i - \mu_i)^2/2\sigma^2\}$$

$$L(\mu|\boldsymbol{y}) = f(\boldsymbol{y}|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(y_i - \mu_i)^2/2\sigma^2\}$$

$$C(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{-\sum_{i=1}^{n}(x_i - \bar{x})/2\sigma^2 + \sum_{i=1}^{n}(y_i - \bar{y})/2\sigma^2\right\}$$

$L(\mu|x) = L(\mu|y)$ if and only if $\bar{x} = \bar{y}$. Same conclusion is drawn about μ from two samples $\boldsymbol{x}$ and $\boldsymbol{y}$ satisfying $\bar{x} = \bar{y}$

Maximum Likelihood Estimators

For a given $\boldsymbol{x}$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ be the parameter at which $L(\boldsymbol{\theta}|\boldsymbol{x})$ attains maximum value as a function of θ, i.e.,

$$\hat{\boldsymbol{\theta}}(\boldsymbol{x}) \in \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{x})$$

$\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ is the Maximum Likelihood Estimator (MLE) of parameter $\boldsymbol{\theta}$ at $\boldsymbol{X} = \boldsymbol{x}$.

- MLE is the **parameter point** for which the oberved point is most likely
- I general, MLE is a good **point estimator** and is by far the most popular technique for deriving estimators.

Solving MLE

- How to find global optima?
- How to verify the found solution is the global optima?
- When $L(\boldsymbol{\theta}|\boldsymbol{x})$ is differentiable in each component θ_i, simple differential calculus can be applied.

$$\frac{\partial}{\partial \theta_i} L(\boldsymbol{\theta}|\boldsymbol{x}) = 0, \quad i = 1, 2, \ldots, k.$$

- How sensitive is the solution to small changes in sample data

Examples:

Normal Likelihood: $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be i.i.d. $\sim \mathcal{N}(\theta, 1)$.

- Likelihood function is

$$L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{(x_i - \theta)^2}{2}\right\}$$

- $\frac{d}{d\theta} L(\theta|\boldsymbol{x}) = 0 \implies \sum_{i=1}^{n}(x_i - \theta) = 0.$ (first order condition)
- $\hat{\theta}(\boldsymbol{x}) = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{\boldsymbol{x}}$. Is $\bar{x}$ global optimal?
- Check $\frac{d^2}{d\theta^2} L(\theta|\boldsymbol{x})|_{\theta=\bar{x}} < 0$? (second order condition)

Bernoulli Likelihood: $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be i.i.d. $\sim Ber(p)$.

- Likelihood function is

$$L(p|\boldsymbol{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

- Solve $\frac{d}{d\theta} L(p|\boldsymbol{x}) = 0$ and find $\bar{p}$ (first order condition)
- Check $\frac{d^2}{dp^2} L(p|\boldsymbol{x})|_{p=\bar{p}} < 0$ (second order condition)

Log Likelihood Functions

- It is easier work with logarithms of Likelihood functions
- $\log L(\boldsymbol{\theta}|\boldsymbol{x})$ instead of $L(\boldsymbol{\theta}|\boldsymbol{x})$, called log likelihood function
- As $\log(\cdot)$ is monotone, solution of maximization problem does not change!
- Normal Log Likelihood:

$$\log L(\theta|\boldsymbol{x}) = n \log \frac{1}{\sqrt{2\pi}} - \frac{\sum_{i=1}^{n}(x_i - \theta)^2}{2}$$

- Binomial Log Likelihood:

$$\log L(p|\boldsymbol{x}) = \left(\sum_{i=1}^{n} x_i\right) \log p + \left(n - \sum_{i=1}^{n} x_i\right) \log(1 - p)$$

$$\frac{d}{dp} L(p|\boldsymbol{x}) = 0 \implies \hat{p} = \bar{x}$$

Other Point Estimators:

- Maximum Likelihood estimator
- Method of moments
- Bayes method
- Expectation Maximization (EM) method

Method of Moments (MM)

- Method of Moments(MM) is one the oldest method for finding point estimator (since 1800!)
- MM estimators are found by equating the first k sample moments to the corresponding population moments
- $\boldsymbol{X}$ be a sample from pmf/pdf $f(\boldsymbol{x}|(\theta_1, \theta_2, \ldots, \theta_k))$

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i^1, \qquad \mu_1' = \mathbb{E}(X^1)$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \qquad \mu_2' = \mathbb{E}(X^2)$$

$$\vdots$$

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \qquad \mu_k' = \mathbb{E}(X^k)$$

Usually, μ_j'S will be function of $(\theta_1, \theta_2, \ldots, \theta_k)$, say $\mu_j'(\theta_1, \theta_2, \ldots, \theta_k)$

Solving MMs

- Obtain k equations by equating

$$
\begin{aligned}
m_1 &= \mu_1'(\theta_1, \theta_2, \ldots, \theta_k) \\
m_2 &= \mu_2'(\theta_1, \theta_2, \ldots, \theta_k) \\
&\vdots \\
m_k &= \mu_k'(\theta_1, \theta_2, \ldots, \theta_k)
\end{aligned}
$$

Normal Method of Moments: $X = (X_1, X_2, \ldots, X_n)$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$. $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\mu, \sigma^2)$.

- $m_1 = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}$ (first moment), $m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$ (second moment)
- $\mu_1'(\theta_1, \theta_2) = \mu$ (mean), $\mu_2'(\theta_1, \theta_2) = \mu^2 + \sigma^2$ (second moment).
- $$\frac{\sum_{i=1}^n X_i}{n} = \theta_1 \quad \text{and} \quad \frac{\sum_{i=1}^n X_i^2}{n} = \theta_1^2 + \theta_2.$$