Statistics **Lecture 9 - 03/20/2025**

Lecture 9: Parameter Inference

Lecturer:Xiangyu Chang *Scribe: Haige Zheng,Wang Miao*

Edited by: Zhihong Liu

1 Parametric Inference

Parametric inference is a statistical method used to estimate the parameters of a population distribution based on a sample. The process can be summarized as follows:

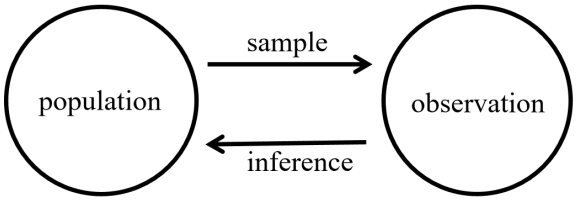

Figure 1: Parameter Inference

- **Population**: The entire set of data points, characterized by a distribution $F(\theta)$, where θ represents the parameters of the distribution.
- **Sample**: A subset of the population, from which observations are drawn.
- **Observation**: The actual data points collected from the sample.
- **Inference**: Using the observations to estimate the parameters θ of the population distribution.

2 Examples of Parametric Models

2.1 Example 1: Bernoulli distribution (Ber (P))

The Bernoulli distribution is a discrete distribution with a single parameter P, which represents the probability of success.

- The population follows Ber(P), and the sample is used to estimate P.

2.2 Example 2: Linear Regression Model

Given data points $\{\xi_i = (x_i, y_i)\}_{i=1}^n$ for $i = 1, 2, \ldots, n$, the relationship between x and y is modeled as:

$$r(x) = \theta^T x + \epsilon.$$

where θ is the parameter vector to be estimated.

2.3 Example 3: Large Language Models (Parametric Models)

In large language models, the goal is to predict the next token ω_{n+1} given a sequence of tokens $\omega_1, \omega_2, \ldots, \omega_n$.

The probability of the sequence is given by:

$$\mathbb{P}(w_1, w_2, \ldots, w_n) = \mathbb{P}(w_1)\mathbb{P}(w_2|w_1)\mathbb{P}(w_3|w_1, w_2)\cdots\mathbb{P}(w_n|w_1, w_2, \ldots, w_{n-1}).$$

$$\mathbb{P}(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} \mathbb{P}(w_i|w_1, w_2, \ldots, w_{i-1}).$$

This is an example of a parametric model where the parameters are learned from data.

3 Moment Estimation

Moment estimation is a method for estimating the parameters of a distribution by matching the sample moments to the theoretical moments.

Let $\theta = (\theta_1, \theta_2, \ldots, \theta_k)^T$ be the parameter vector. The steps for moment estimation are:

1. **Step 1**: Define the theoretical moments:

$$\alpha_k(\theta) = E_\theta[X^k] = \int X^k dF_\theta(X), \quad 1 \le k \le K.$$

2. **Step 2**: Compute the sample moments:

$$\hat{\alpha_k} = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

3. **Step 3**: Set the theoretical moments equal to the sample moments and solve for θ:

$$\alpha_k(\hat{\theta}) = \hat{\alpha_k}, \quad (k = 1, 2, \ldots, K).$$

3.1 Examples of Moment Estimation

3.1.1 Example 1: Bernoulli Distribution (Ber(P))

The first moment (mean) of the Bernoulli distribution is:

$$\alpha_1(p) = \mathbb{E}[X] = P.$$

The sample mean is:

$$\hat{\alpha}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}_n.$$

Therefore, the moment estimator for P is:

$$\hat{P} = \hat{\alpha}_1 = \overline{x}_n.$$

3.1.2 Example 2: Normal Distribution $N(\mu, \sigma^2)$

For a normal distribution with mean μ and variance σ^2, the first two moments are:

$$\alpha_1(\theta) = \mathbb{E}[X] = \mu,$$

$$\alpha_2(\theta) = \mathbb{E}[X^2] = \mu^2 + \sigma^2.$$

The sample moments are:

$$\hat{\alpha}_1 = \frac{1}{n}\sum x_i = \overline{x}_n,$$

$$\hat{\alpha}_2 = \frac{1}{n}\sum x_i^2.$$

The moment estimators for μ and σ^2 are:

$$\hat{\mu} = \hat{\alpha}_1 = \overline{x}_n,$$

$$\hat{\sigma}^2 = \hat{\alpha}_2 - (\hat{\alpha}_1)^2 = \frac{1}{n}\sum (x_i - \overline{x}_n)^2.$$

Note that this estimator for σ^2 is biased.

4 Maximum Likelihood Estimation (MLE)

Maximum likelihood estimation is a method for estimating the parameters of a statistical model by maximizing the likelihood function.

Given a sample $X_1, X_2, \ldots, X_n$ drawn independently and identically distributed (i.i.d.) from a distribution $F_0(x)$, the goal is to find the parameter θ that maximizes the likelihood function:

$$\max \prod_{i=1}^{n} f_\theta(x_i).$$

where $f_\theta(x_i)$ is the probability density function (PDF) or probability mass function (PMF) of the distribution.

This is equivalent to maximizing the log-likelihood function:

$$\max \sum_{i=1}^{n} \log f_\theta(x_i). \quad \Longleftrightarrow \quad \min - \sum_{i=1}^{n} \log f_\theta(x_i).$$

4.1 Examples of Maximum Likelihood Estimation

4.1.1 Example 1: Bernoulli Distribution (Ber(P))

1. Bernoulli Distribution (Ber(p))

The Bernoulli distribution models a random variable X that takes the value 1 with probability p and the value 0 with probability $1 - p$.

- **Probability Mass Function (PMF):**

$$f_\theta(x_i) = p^{x_i}(1-p)^{1-x_i}.$$

  where $x_i \in \{0, 1\}$.

- **Likelihood Function:** The likelihood function for a sample of n independent observations $x_1, x_2, \ldots, x_n$ is:

$$\prod_{i=1}^{n} f_\theta(x_i) = p^{\sum x_i}(1-p)^{\sum (1-x_i)}.$$

2. Log-Likelihood Function

To simplify the maximization, we take the natural logarithm of the likelihood function:

$$\log L(p) = \sum_{i=1}^{n} x_i \log p + \sum_{i=1}^{n} (1 - x_i) \log(1-p).$$

- **Objective:** Maximize the log-likelihood function with respect to p:

$$\max_p \left(\sum_{i=1}^{n} x_i \log p + \sum_{i=1}^{n} (1 - x_i) \log(1-p) \right),$$

  subject to $0 < p < 1$.

- **Equivalent Minimization Problem:**

$$\min_p \left(- \sum_{i=1}^{n} x_i \log p - \sum_{i=1}^{n} (1 - x_i) \log(1-p) \right).$$

3. Derivative and Optimization

To find the maximum likelihood estimate (MLE) of p, we take the derivative of the log-likelihood function with respect to p and set it to zero:

$$\frac{d}{dp}\log L(p) = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot \frac{1}{p} - \frac{1}{n}\sum_{i=1}^{n}(1-x_i)\cdot\frac{1}{1-p} = 0.$$

Solving for p:

$$\frac{1}{n}\sum_{i=1}^{n} x_i \cdot \frac{1}{p} = \frac{1}{n}\sum_{i=1}^{n}(1-x_i)\cdot\frac{1}{1-p}.$$

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x_n}.$$

4. Conclusion

The maximum likelihood estimate (MLE) of p is the sample mean of the observed data:

$$\hat{p} = \overline{x_n}.$$

4.1.2 Example 2: Normal Distribution $N(\mu, \sigma^2)$

1. Probability Density Function(PDF)

The normal distribution, also known as the Gaussian distribution, is characterized by its mean μ and variance b^2. The probability density function (PDF) for a normal distribution is given by:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

2. Likelihood Function

For a sample of n independent observations $x_1, x_2, \ldots, x_n$, the likelihood function is the product of the individual PDFs:

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} f(x_i;\mu,\sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

3. Log-Likelihood Function

To simplify the maximization, we take the natural logarithm of the likelihood function:

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

4. Maximization with Respect to μ and σ^2

To find the maximum likelihood estimates (MLE) of μ and σ^2, we take the partial derivatives of the log-likelihood function with respect to μ and σ^2 and set them to zero.

- **Partial Derivative with Respect to μ:**

$$\frac{\partial}{\partial \mu}\log L(\mu, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0.$$

  Solving for μ:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}.$$

- **Partial Derivative with Respect to σ^2:**

$$\frac{\partial}{\partial \sigma^2}\log L(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0.$$

  Solving for σ^2:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.$$

5. Conclusion

The maximum likelihood estimates (MLE) for the parameters of the normal distribution are:

$$\hat{\mu} = \overline{x},$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

4.1.3 Example 3: Linear Regression Model

In statistics, linear regression is a method for modeling the relationship between a dependent variable y_i and one or more independent variables x_i. The linear relationship between y_i and x_i can be expressed as:

$$y_i = \beta^T x_i + \xi_i, \quad \xi_i | x_i \sim N(0, 1).$$

where β is a vector of coefficients, x_i is a vector of independent variables, and ξ_i is a random error term with a normal distribution with mean 0 and variance 1.

The likelihood function for this model is given by:

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_i - \beta^T x_i)^2}{2}\right).$$

This function measures the probability of observing the data given the parameters β.

To simplify the maximization of the likelihood function, we often work with the log-likelihood function:

$$\log\left(\prod_{i=1}^{n} \frac{1}{sqrt2\pi} \exp\left(-\frac{(y_i - \beta^T x_i)^2}{2}\right)\right) = \log\left(2\pi\right)^{-\frac{n}{2}} - \sum_{i=1}^{n} \left(y_i - \beta^T x_i\right)^2.$$

Maximizing the log-likelihood function is equivalent to minimizing the sum of squared residuals:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2.$$

This can be written in matrix form as:

$$\min_{\beta} \|Y - X\beta\|^2.$$

where Y is an $n \times 1$ vector of dependent variables and X is an $n \times p$ matrix of independent variables.

Steps to Maximize the Likelihood Function

1. Write down the likelihood function $Ln(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$**.**

2. Take the natural logarithm of the likelihood function to obtain the log-likelihood function $ln(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i)$**.**

3. Maximize the log-likelihood function, which is equivalent to **minimizing** $-\ln(\theta)$**.**

4. Solve for the parameters θ that minimize $-\ln(\theta)$. For unconstrained optimization problems, this can be done using techniques such as gradient descent or Newton's method.

5 Properties of MLE

5.1 Likelihood Function

The log-likelihood for n observations is given by:

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i).$$

The MLE $\hat{\theta}_n$ is the value of θ that maximizes the log-likelihood function:

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \ell_n(\theta).$$

5.2 Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence is a measure of the difference between two probability distributions. The empirical KL divergence for a sample of size n is defined as:

$$D_n(\theta^*, \theta) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_{\theta^*}(X_i)}{f_\theta(X_i)}.$$

Properties of the KL divergence include:

- $D(\theta^*, \theta) \geq 0$ with equality if $\theta = \theta^*$.
- Asymmetric: $D(\theta^*, \theta) \neq D(\theta, \theta^*)$.

5.3 Key Assumptions

For the consistency of MLEs, the following assumptions are typically made:

A1. Uniform convergence:

$$\sup_{\theta \in \Theta} |\ell_n(\theta) - \mathbb{E}[\log f_\theta(X)]| \xrightarrow{p} 0.$$

A2. Identifiability: For any $\epsilon > 0$,

$$\sup_{\|\theta - \theta^*\| \geq \epsilon} \mathbb{E}[\log f_\theta(X)] < \mathbb{E}[\log f_{\theta^*}(X)].$$

5.4 Consistency Theorem

Under assumptions A1 and A2, the MLE $\hat{\theta}_n$ converges in probability to the true parameter θ^* as the sample size n tends to infinity:

$$\hat{\theta}_n \xrightarrow{p} \theta^* \quad \text{as } n \to \infty.$$

$\widehat{\text{By}}$ Law of Large Numbers:extbf

$$\frac{1}{n}\ell_n(\theta) \xrightarrow{p} \mathbb{E}[\log f_\theta(X)].$$

A2 implies θ^* is unique maximizer. Combined with A1, we get:

$$\|\hat{\theta}_n - \theta^*\| < \epsilon \quad \text{with probability } \to 1.$$

5.5 Definitions

Let $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta(x)$, where $\theta \in \Theta \subseteq \mathbb{R}^d$.

- **Score Function**:The score function $S_\theta(x)$ is the gradient of the log -likelihood function with respect to θ:

$$S_\theta(x) = log f_\theta(x) = \frac{\nabla_\theta f_\theta(x)}{f_\theta(x)}.$$

- **Fisher Information Matrix**:The Fisher Information Matrix $I(\theta)$ is the variance of the score function:

$$I(\theta) = \text{Var}[S_\theta(X)] = -\mathbb{E}\left[\nabla_\theta^2 \log f_\theta(X)\right].$$

5.6 Key Properties

1. **The mean of the score function is zero**: The score function $S_\theta(x) = \nabla_\theta \log f_\theta(x)$ measures the sensitivity of the log-likelihood to parameter changes. Remarkably, its expectation is always zero under the true distribution. This follows from:

$$\mathbb{E}[S_\theta(X)] = \int \frac{\nabla_\theta f_\theta(x)}{f_\theta(x)} f_\theta(x)\, dx = \int \nabla_\theta f_\theta(x)\, dx.$$

   Here we used the definition of expectation and canceled $f_\theta(x)$ terms. Crucially, we can interchange integration and differentiation:

$$\nabla_\theta \underbrace{\int f_\theta(x)\, dx}_{=1} = \nabla_\theta(1) = 0.$$

   This shows the score function is centered – positive and negative sensitivities balance out in expectation.

2. **Fisher Information Matrix (FIM) has dual representations**:

   The FIM $I(\theta)$ quantifies information about θ in the data. It has two equivalent expressions:

- **Variance of the score**:

$$I(\theta) = \mathrm{Var}(S_\theta(X)) = \mathbb{E}[S_\theta(X)S_\theta(X)^\top].$$

  This measures the dispersion of the score function. A "peakier" likelihood surface corresponds to higher information.

- **Negative expected Hessian**:

$$I(\theta) = -\mathbb{E}\left[\nabla_\theta^2 \log f_\theta(X)\right].$$

  This characterizes the curvature of the log-likelihood. The Hessian's eigenvalues indicate sensitivity along different parameter directions.

Connecting the representations: For n i.i.d. observations, the total log-likelihood $l_n(\theta) = \sum_{i=1}^n \log f_\theta(x_i)$ has variance:

$$\mathrm{Var}(l_n(\theta)) = nI(\theta) = -\mathbb{E}[\nabla_\theta^2 l_n(\theta)].$$

This shows information accumulates linearly with sample size.

Derivation of equivalence: Starting from the score function definition $S_\theta(x) = \nabla_\theta \log f_\theta(x)$, compute the second derivative:

$$\nabla_\theta^2 \log f_\theta(x) = \frac{\nabla_\theta^2 f_\theta(x)}{f_\theta(x)} - \frac{\nabla_\theta f_\theta(x) \nabla_\theta f_\theta(x)^\top}{f_\theta(x)^2}.$$

Taking expectation over $X \sim f_\theta$:

$$\begin{aligned}
\mathbb{E}\left[\nabla_\theta^2 \log f_\theta(X)\right] &= \int \left(\frac{\nabla_\theta^2 f_\theta(x)}{f_\theta(x)} - \frac{\nabla_\theta f_\theta(x) \nabla_\theta f_\theta(x)^\top}{f_\theta(x)^2} \right) f_\theta(x)\, dx \\
&= \int \nabla_\theta^2 f_\theta(x)\, dx - \mathbb{E}[S_\theta(X)S_\theta(X)^\top] \\
&= \nabla_\theta^2 \underbrace{\int f_\theta(x)\, dx}_{=1} - I(\theta) \\
&= 0 - I(\theta) = -I(\theta).
\end{aligned}$$

The critical step uses differentiation under the integral sign (requiring regularity conditions). This establishes $I(\theta) = -\mathbb{E}[\text{Hessian}]$, completing the duality.