5430 Theory Notes

Bookmark:

Introduction

Probability and Statistical Inference

- **Probability** is a branch of mathematics concerned with the study of *random* phenomena (e.g., experiments, models of populations).
- **Statistical inference** is the science of drawing inferences about populations based on only a part of the population (i.e., a sample).
  (Inference is based on probability.)

Random Samples

Definition.
Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with common cdf $F(x)$ and pdf/pmf $f(x)$. Then we say:

1. $X_1, \ldots, X_n$ is a **random sample (r.s.)**.
   $F(x)$ is the population cdf and $f(x)$ is the population pdf/pmf.
2. $X_1, \ldots, X_n$ is a random sample from $F(x)$ or from $f(x)$.
   (Both are equivalent ways of describing the population distribution.)

Statistical Inference

- Statistical inference is about **making statements about population distributions based on samples**.
- For a collection $\mathcal{F}$ of cdf's, let $F(x) \in \mathcal{F}$ be the underlying population cdf.
  Given $X_1, \ldots, X_n$, our objective is to draw inferences about $F(x)$.

Parametric Considerations

Parametric vs. Nonparametric Models

Definition.

If

$$\mathcal{F} = \{F(x \mid \theta) : \theta \in \Theta\}, \qquad \Theta \subset \mathbb{R}^k,\ 1 \le k < \infty,$$

then the inference problem is called **parametric**; otherwise, it is **nonparametric**.

- θ is called the **parameter**
- Θ is called the **parameter space**

Statistics and Estimators

Definition.
Let $X_1, \dots, X_n$ be a random sample. A (Borel measurable) function of the random sample,

$$T = h(X_1, \dots, X_n),$$

is called a **statistic** (or an **estimator**).
(That is, T is computable from the data.)

Sampling Distributions

Definition.
The probability distribution of a statistic T is called the **sampling distribution** of T.

Parametric Functions and Estimation

Definitions.

1. A (Borel measurable) function

$$\gamma : \Theta \to \mathbb{R}^d, \qquad 1 \le d < \infty,$$

is called a **parametric function**.

2. If a statistic $T = h(X_1, \dots, X_n)$ is used to estimate $\gamma(\theta)$, then:
   - T is called an **estimator** of $\gamma(\theta)$
   - The observed value $t = h(x_1, \dots, x_n)$ is called an **estimate** of $\gamma(\theta)$

Method of Moments Estimation (MME)

Introduction

Definition.
Let $X_1, \dots, X_n$ be a random sample from pdf/pmf $f(x \mid \theta_1, \dots, \theta_k)$.

Population Moments

$$E(X_1^j) \equiv \mu_j(\theta_1, \dots, \theta_k)$$

is the jth **population moment**, for $j = 1, 2, \dots$

Example:
If $X_1 \sim N(\mu, \sigma^2)$, then

$$E(X_1) = \mu, \qquad E(X_1^2) = \mathrm{Var}(X_1) + [E(X_1)]^2 = \sigma^2 + \mu^2.$$

Sample Moments

$$\mu_j' = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

is the jth **sample moment**, for $j = 1, 2, \ldots$

Method of Moments Estimators

The **method of moments estimators (MMEs)** $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$ are defined as the solution to the system:

$$\begin{aligned} \mu_1(\tilde{\theta}_1, \ldots, \tilde{\theta}_k) &= \mu_1', \\ \vdots \qquad\qquad &\ \ \vdots \\ \mu_k(\tilde{\theta}_1, \ldots, \tilde{\theta}_k) &= \mu_k'. \end{aligned} \tag{*}$$

(Choose $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$ so that the population moments match the sample moments.)

Moment Equations

The system of equations $(*)$ is called the **method of moments equations (MME equations)**.

Method of Moments Estimation for Parametric Functions

Definition.
For a parametric function $\gamma(\theta_1, \ldots, \theta_k)$, we define the **method of moments estimator (MME)**

$$\tilde{\gamma}(\theta_1, \ldots, \theta_k)$$

of $\gamma(\theta_1, \ldots, \theta_k)$ as

$$\tilde{\gamma}(\theta_1, \ldots, \theta_k) = \gamma(\tilde{\theta}_1, \ldots, \tilde{\theta}_k),$$

where $\tilde{\theta}_1, \ldots, \tilde{\theta}_k$ are the MMEs of $\theta_1, \ldots, \theta_k$.

Maximum Likelihood Estimation (MLE)

Introduction

Definition.
Let $f(x_1, \ldots, x_n \mid \theta)$ be the joint pdf/pmf of $(X_1, \ldots, X_n)$. Then

$$L(\theta) = f(x_1, \ldots, x_n \mid \theta), \qquad \theta \in \Theta,$$

viewed as a function of θ for fixed data $(x_1, \ldots, x_n)$, is called the **likelihood function**.

Notes

1. If $X_1, \ldots, X_n$ are i.i.d. with common pdf/pmf $f(x \mid \theta)$, then

$$L(\theta) = f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

2. If $X_1, \ldots, X_n$ are discrete random variables, then

$$L(\theta) = P(X_1 = x_1, \ldots, X_n = x_n \mid \theta).$$

Definition of the MLE

Definition.
Let $(X_1, \ldots, X_n)$ have joint pdf/pmf $f(x_1, \ldots, x_n \mid \theta)$, $\theta \in \Theta$.
For observed data $(x_1, \ldots, x_n)$, the **maximum likelihood estimate (MLE)** of θ is a point

$$\hat{\theta} = h(x_1, \ldots, x_n) \in \Theta$$

such that

$$f(x_1, \ldots, x_n \mid \hat{\theta}) = \max_{\theta \in \Theta} f(x_1, \ldots, x_n \mid \theta) = \max_{\theta \in \Theta} L(\theta).$$

The **maximum likelihood estimator** is defined as

$$\hat{\theta} = h(X_1, \ldots, X_n).$$

Finding Maximum Likelihood Estimators

Finding the MLE $\hat{\theta}$ requires maximizing the likelihood function $L(\theta)$ over Θ.

1. If $L(\theta)$ is smooth (differentiable) in θ, use calculus.
2. If $L(\theta)$ is not smooth, maximization requires more care.
3. In practice, $L(\theta)$ is often maximized numerically.
4. Maximizing $\log L(\theta)$ is equivalent to maximizing $L(\theta)$ and is often easier.
5. If the support $\{x : f(x \mid \theta) > 0\}$ depends on θ, indicator functions can be useful.

Using Calculus to Determine the MLE

Assume $\Theta \subset \mathbb{R}$ is open and $L(\theta)$ is twice differentiable on Θ. Then

$$\hat{\theta} \text{ maximizes } L(\theta) \iff \left. \frac{dL(\theta)}{d\theta} \right|_{\hat{\theta}} = 0 \quad \text{and} \quad \left. \frac{d^2 L(\theta)}{d\theta^2} \right|_{\hat{\theta}} < 0.$$

Since $\log(\cdot)$ is increasing,

$$\hat{\theta} \text{ maximizes } L(\theta) \iff \hat{\theta} \text{ maximizes } \log L(\theta).$$

Hence, $\hat{\theta}$ is an MLE if

$$\left. \frac{d \log L(\theta)}{d\theta} \right|_{\hat{\theta}} = 0 \quad \text{and} \quad \left. \frac{d^2 \log L(\theta)}{d\theta^2} \right|_{\hat{\theta}} < 0.$$

Multiparameter Case

Suppose $(X_1, \dots, X_n)$ have joint pdf/pmf $f(x_1, \dots, x_n \mid \theta)$ where

$$\theta = (\theta_1, \theta_2, \dots, \theta_k)' \in \Theta \subset \mathbb{R}^k.$$

We seek MLEs

$$\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)'$$

that satisfy

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta).$$

Result

If $\Theta \subset \mathbb{R}^k$ is open and $L(\theta)$ has second-order partial derivatives, then $\hat{\theta}_1, \dots, \hat{\theta}_k$ are MLEs provided:

1. For each $i = 1, \dots, k$,

$$\left. \frac{\partial \log L(\theta)}{\partial \theta_i} \right|_{\hat{\theta}} = 0.$$

2. Let H be the Hessian matrix at $\hat{\theta}$:

$$H = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1k} \\ h_{21} & h_{22} & \cdots & h_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ h_{k1} & h_{k2} & \cdots & h_{kk} \end{pmatrix}, \qquad h_{ij} = \left. \frac{\partial^2 \log L(\theta)}{\partial \theta_i \partial \theta_j} \right|_{\hat{\theta}}.$$

Let

$$\Delta_i = \det(\text{leading } i \times i \text{ submatrix of } H), \qquad i = 1, \dots, k.$$

Then we require

$$\Delta_1 < 0,\ \Delta_2 > 0,\ \Delta_3 < 0,\ \dots$$

(i.e., alternating signs).

MLEs of Parametric Functions

Definition.
For a parametric function $\gamma(\theta_1, \theta_2, \dots, \theta_k)$, we define

$$\gamma(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$$

to be the **MLE of** $\gamma(\theta_1, \theta_2, \dots, \theta_k)$, where $\hat{\theta}_1, \dots, \hat{\theta}_k$ are the MLEs of $\theta_1, \dots, \theta_k$.

Estimator Evaluation (for Point Estimators)

Bias

Definition.
An estimator $T = h(X_1, \ldots, X_n)$ of a parametric function $\gamma(\theta)$ is called **unbiased** if

$$E_\theta(T) = E(T) = \gamma(\theta), \qquad \forall \theta \in \Theta.$$

Definition.
T is **biased** if it is not unbiased.

Definition.
The **bias** of T is

$$b_\theta(T) = E(T) - \gamma(\theta).$$

If T is unbiased, then

$$b_\theta(T) = 0 \quad \forall \theta \in \Theta.$$

Notes

0. "U.E." denotes *unbiased estimator*.
1. If T is a U.E. of θ, then $\gamma(T)$ need **not** be a U.E. of $\gamma(\theta)$.
2. It is **not always possible** to find a U.E. of $\gamma(\theta)$.

Variance

Uniform Minimum Variance Unbiased Estimator (UMVUE)

Definition.
Let $f(x_1, \ldots, x_n \mid \theta)$ be the joint pdf/pmf of $X_1, \ldots, X_n$.
An estimator T of a real-valued parametric function $\gamma(\theta)$ is called the
Uniform Minimum Variance Unbiased Estimator (UMVUE) of $\gamma(\theta)$ if:

1. T is an unbiased estimator (U.E.) of $\gamma(\theta)$, i.e.,

$$E_\theta(T) = \gamma(\theta), \qquad \forall \theta \in \Theta.$$

2. $\mathrm{Var}_\theta(T) < \infty$, for all $\theta \in \Theta$.

3. For any other unbiased estimator T_1 of $\gamma(\theta)$,

$$\mathrm{Var}_\theta(T) \leq \mathrm{Var}_\theta(T_1), \qquad \forall \theta \in \Theta.$$

(That is, T has the smallest variance among all unbiased estimators of $\gamma(\theta)$.)

Finding a UMVUE

There are two general strategies for finding a UMVUE:

- Use the **Cramér–Rao Lower Bound (CRLB)** (does not always work).
- Use **sufficiency + completeness** (introduced later).

Cramér–Rao Lower Bound (CRLB)

Motivation

- Suppose T is an unbiased estimator of a real-valued parametric function $\gamma(\theta)$, and we wish to know whether T is the UMVUE of $\gamma(\theta)$.
- Suppose there exists a function $c(\theta)$ such that, for any unbiased estimator T_1 of $\gamma(\theta)$,

$$\mathrm{Var}_\theta(T_1) \geq c(\theta), \qquad \forall \theta \in \Theta.$$

- If we find that

$$\mathrm{Var}_\theta(T) = c(\theta), \qquad \forall \theta \in \Theta,$$

then T must be the UMVUE of $\gamma(\theta)$. - Sometimes such a lower bound $c(\theta)$ can be obtained via the **Cramér–Rao inequality**, also called the **Cramér–Rao Lower Bound (CRLB)**.

Theorem (Cramér–Rao Inequality)

Let $f(x_1, x_2, \ldots, x_n \mid \theta)$ be the joint pdf/pmf of $X_1, X_2, \ldots, X_n$, with $\theta \in \Theta$.
Assume regularity conditions hold, specifically:

1. Θ is an open subset of $\mathbb{R}$.
2. $A \equiv \{(x_1, \ldots, x_n) : f(x_1, \ldots, x_n \mid \theta) > 0\}$ does **not** depend on θ.
3. $\dfrac{d}{d\theta} f(x_1, \ldots, x_n \mid \theta)$ exists on Θ, for all $(x_1, \ldots, x_n) \in A$.
4. For any estimator $T^* = T^*(X_1, \ldots, X_n)$ with $E_\theta[(T^*)^2] < \infty$,

$$\frac{d}{d\theta} E_\theta(T^*) = \begin{cases} \displaystyle\int_A T^*(x_1, \ldots, x_n) \frac{d}{d\theta} f(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n, & \text{if } X_i \text{ are continuous,} \\ \displaystyle\sum_{(x_1, \ldots, x_n) \in A} T^*(x_1, \ldots, x_n) \frac{d}{d\theta} f(x_1, \ldots, x_n \mid \theta), & \text{if } X_i \text{ are discrete.} \end{cases}$$

5. For all $\theta \in \Theta$,

$$0 < I_n(\theta) \equiv E_\theta\left[\left(\frac{d}{d\theta} \log f(X_1, X_2, \ldots, X_n \mid \theta)\right)^2\right] < \infty.$$

Then, for any unbiased estimator T of $\gamma(\theta)$,

$$\mathrm{Var}_\theta(T) \geq \frac{[\gamma'(\theta)]^2}{I_n(\theta)}, \qquad \forall \theta \in \Theta. \tag{CRLB}$$

Here $\gamma'(\theta) = \dfrac{d}{d\theta}\gamma(\theta)$ is assumed to exist on Θ.

Fisher Information

- $I_n(\theta)$ is called the **Fisher information number** for a sample of size n.
- If $X_1, X_2, \ldots, X_n$ are i.i.d. with common pdf/pmf $f(x \mid \theta)$, then

$$I_n(\theta) = nI_1(\theta), \qquad I_1(\theta) = E_\theta\left[\left(\frac{d}{d\theta}\log f(X_1 \mid \theta)\right)^2\right].$$

- If $\dfrac{d^2}{d\theta^2}f(x_1, \ldots, x_n \mid \theta)$ exists on Θ, then

$$I_n(\theta) = E_\theta\left[\left(\frac{d}{d\theta}\log f(X_1, \ldots, X_n \mid \theta)\right)^2\right] = -E_\theta\left[\frac{d^2}{d\theta^2}\log f(X_1, \ldots, X_n \mid \theta)\right].$$

- If, in addition, $X_1, \ldots, X_n$ are i.i.d. with common $f(x \mid \theta)$, then

$$I_n(\theta) = nI_1(\theta), \quad \text{where} \quad I_1(\theta) = E_\theta\left[\left(\frac{d}{d\theta}\log f(X_1 \mid \theta)\right)^2\right] = -E_\theta\left[\frac{d^2}{d\theta^2}\log f(X_1 \mid \theta)\right].$$

Relative Efficiency

We compare unbiased estimators (U.E.'s) in terms of variance; **smaller variance is preferred**.

Definitions.
Let T, T_1, and T_2 be unbiased estimators of $\gamma(\theta)$.

1. The **relative efficiency** of T_1 with respect to T_2 is

$$\text{r.e.}(T_1, T_2, \theta) \equiv \frac{\text{Var}_\theta(T_2)}{\text{Var}_\theta(T_1)}.$$

2. T is called **efficient** if

$$\text{r.e.}(T_1, T, \theta) \leq 1, \qquad \forall \theta \in \Theta$$

for every other unbiased estimator T_1 of $\gamma(\theta)$. *(Equivalently, T is the UMVUE.)*

3. If T is an efficient estimator and T_1 is any unbiased estimator of $\gamma(\theta)$, the **efficiency** of T_1 is

$$e_{T_1}(\theta) = \text{r.e.}(T_1, T, \theta) = \frac{\text{Var}_\theta(T)}{\text{Var}_\theta(T_1)} \leq 1.$$

Comparing Biased and Unbiased Estimators: Mean Squared Error (MSE)

Previously, we compared unbiased estimators using variance.
When estimators may be biased, we use **mean squared error (MSE)**.

Definition.
For an estimator T of $\gamma(\theta)$, the **mean squared error** is

$$\mathrm{MSE}_\theta(T) \equiv E_\theta\left[(T - \gamma(\theta))^2\right].$$

Facts about MSE

1. The MSE decomposes as

$$\mathrm{MSE}_\theta(T) = \mathrm{Var}_\theta(T) + [b_\theta(T)]^2,$$

where

$$b_\theta(T) = E_\theta(T) - \gamma(\theta)$$

is the bias of T.

2. If T is an unbiased estimator of $\gamma(\theta)$, then

$$b_\theta(T) = 0 \quad \Rightarrow \quad \mathrm{MSE}_\theta(T) = \mathrm{Var}_\theta(T).$$

Decision Theory

Introduction

Loss Function

Definition.
A real-valued function $L(t, \theta)$ is called a **loss function** for estimating $\gamma(\theta)$ if:

1. $L(t, \theta) \geq 0$ for all t and θ,
2. $L(t, \theta) = 0$ if $t = \gamma(\theta)$.

That is, think of $L(t, \theta)$ as a **penalty** for guessing $\gamma(\theta)$ by the value t.

Risk Function

Definition.
For an estimator T of $\gamma(\theta)$, the **risk function** of T is

$$R_T(\theta) \equiv E_\theta[L(T, \theta)], \qquad \theta \in \Theta.$$

Comparing Estimators via Risk

1. An estimator T_1 is **at least as good as** T_2 if

$$R_{T_1}(\theta) \le R_{T_2}(\theta) \quad \text{for all } \theta \in \Theta.$$

2. An estimator T_1 is **better than** T_2 if
   (a) $R_{T_1}(\theta) \le R_{T_2}(\theta)$ for all $\theta \in \Theta$, and
   (b) $R_{T_1}(\theta_0) < R_{T_2}(\theta_0)$ for some $\theta_0 \in \Theta$.
3. An estimator T is called **admissible** if there does not exist another estimator that is better than T. Otherwise, T is called **inadmissible**.

Remarks on Admissibility

- If T_1 is inadmissible, then there exists an estimator T that is better than T_1. Hence, it suffices to consider only **admissible estimators**.
- In general, a single "best" estimator does **not** exist. Instead, one may:
  1. Restrict the class of estimators (e.g., consider only unbiased estimators) and find the best estimator within that class (e.g., the UMVUE), or
  2. Define a different optimality criterion for ordering the risk function, such as:
     - the **Bayes principle**, or
     - the **minimax principle**.

Minimax Principle & Estimator

Rationale

- If the statistician chooses estimator T_1, nature will choose θ_1 such that

$$R_{T_1}(\theta_1) = \max_{\theta \in \Theta} R_{T_1}(\theta).$$

- If the statistician chooses estimator T_2, nature will choose θ_2 such that

$$R_{T_2}(\theta_2) = \max_{\theta \in \Theta} R_{T_2}(\theta).$$

- Thus, the statistician should choose an estimator that **minimizes the worst-case risk**.

Minimax Estimator

Definition.
An estimator T is called **minimax** if

$$\max_{\theta \in \Theta} R_T(\theta) = \min_{T_1} \max_{\theta \in \Theta} R_{T_1}(\theta).$$

Notes

1. If the maximum is not attained, replace "max" with "sup".
2. The minimax criterion is **conservative**, as it guards against the worst-case scenario.

Bayes

Principle and Terminology

Definitions.

1. Let $\pi(\theta)$ be a pdf/pmf on Θ.
   Then $\pi(\theta)$ is called a **prior distribution**.

2. The **Bayes risk** of an estimator T (with respect to $\pi(\theta)$ and loss function $L(t,\theta)$) is

$$\mathrm{BR}_T = \begin{cases} \displaystyle\int_\Theta R_T(\theta)\,\pi(\theta)\,d\theta, & \text{if } \pi(\cdot) \text{ is continuous,} \\ \displaystyle\sum_{\theta\in\Theta} R_T(\theta)\,\pi(\theta), & \text{if } \pi(\cdot) \text{ is discrete.} \end{cases}$$

3. An estimator T_0 is called a **Bayes estimator** (with respect to $\pi(\theta)$) if

$$\mathrm{BR}_{T_0} = \min_T \mathrm{BR}_T.$$

Posterior Distributions

Notation

Let $X = (X_1, X_2, \ldots, X_n)$ and let $x = (x_1, x_2, \ldots, x_n)$ denote an observed value of X.

Set-up

1. θ is treated as a random variable on Θ with marginal pdf/pmf $\pi(\theta)$.
2. $f(x \mid \theta)$ is the conditional pdf/pmf of X given θ.
3. $f(x,\theta) = f(x \mid \theta)\pi(\theta)$ is the joint pdf/pmf of (X,θ).
4.

$$m(x) = \int_\Theta f(x,\theta)\,d\theta$$

is the marginal pdf/pmf of X.

Definition.
The conditional pdf of θ given x is

$$f_{\theta|x}(\theta) = \frac{f(x \mid \theta)\pi(\theta)}{m(x)}, \qquad \theta \in \Theta,$$

and is called the **posterior distribution** of θ.

Finding Bayes Estimators

For an estimator $T = h(X)$ and loss function $L(t, \theta)$:

$$R_T(\theta) = E_\theta[L(T, \theta)] = E_{X|\theta}[L(h(X), \theta)].$$

The Bayes risk is

$$\text{BR}_T = E_\theta[R_T(\theta)] = E_{X,\theta}[L(T, \theta)] = E_X\left[E_{\theta|X}[L(h(X), \theta)]\right].$$

Main Idea

To minimize BR_T, it is sufficient that **for each fixed data value** x, we choose $h(x)$ to minimize the **posterior risk**

$$E_{\theta|x}[L(h(x), \theta)] = \int_\Theta L(h(x), \theta)\, f_{\theta|x}(\theta)\, d\theta.$$

Bayes Estimator Theorem

Theorem.
A Bayes estimator minimizes the posterior risk

$$E_{\theta|x}[L(h(x), \theta)]$$

over all estimators $T = h(X)$, for fixed observed data $x = (x_1, x_2, \ldots, x_n)$.

Corollary.
Let T_0 denote the Bayes estimator of $\gamma(\theta)$.

1. If $L(t, \theta) = (t - \gamma(\theta))^2$, then

$$T_0 = E[\gamma(\theta) \mid x],$$

the **posterior mean** of $\gamma(\theta)$.

2. If $L(t, \theta) = |t - \gamma(\theta)|$, then

$$T_0 = \text{median}(\gamma(\theta) \mid x),$$

the **posterior median** of $\gamma(\theta)$.

Conjugate Priors

Definition.
Let

$$\mathcal{F} = \{f(x \mid \theta) : \theta \in \Theta\}$$

denote the class of joint pdfs/pmfs for $X_1, \ldots, X_n$. A class Π of priors is called a **conjugate family** for $\mathcal{F}$ if the posterior distribution belongs to Π for all $\pi \in \Pi$ and all x.

In a nutshell: A prior is conjugate to a likelihood if the posterior distribution of θ belongs to the same parametric family as the prior, with updating occurring through changes in the parameter values.

Bayes and Minimax Estimators

Theorem.
For a given loss function $L(t, \theta)$, if T^* is a Bayes estimator with respect to some prior and the risk of T^* is constant,

$$R_{T^*}(\theta) = c \quad \text{for all } \theta \in \Theta,$$

then T^* is the **minimax estimator** under the same loss function.