5430 Theory Notes

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

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

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

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

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

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

is called a **parametric function**.

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

Method of Moments Estimation (MME)

Introduction

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

Population Moments

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

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

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

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

We seek MLEs

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

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 $\hat{\theta}_1, \ldots, \hat{\theta}_k$ are MLEs provided:

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

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

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

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

Then we require

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

(i.e., alternating signs).

MLEs of Parametric Functions

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

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

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

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

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

2. An estimator T_1 is **better than** T_2 if
   (a) $R_{T_1}(\theta) \leq R_{T_2}(\theta)$ 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.

Large Sample Properties of Estimators

Let $\{T_n\}$ be a sequence of estimators of a parametric function $\gamma(\theta)$.

1. Consistency

$\{T_n\}$ is called **consistent** for $\gamma(\theta)$ if, for any $\varepsilon > 0$,

(a)

$$\lim_{n\to\infty} P_\theta \left(|T_n - \gamma(\theta)| < \varepsilon\right) = 1, \quad \forall \theta \in \Theta,$$

(b) equivalently,

$$\lim_{n\to\infty} P_\theta \left(|T_n - \gamma(\theta)| \geq \varepsilon\right) = 0, \quad \forall \theta \in \Theta,$$

(c) equivalently, T_n converges in probability to $\gamma(\theta)$ as $n \to \infty$, i.e.,

$$T_n \xrightarrow{p} \gamma(\theta).$$

Two Useful Results for Establishing Consistency

1. Weak Law of Large Numbers (WLLN) If $W_1, W_2, \ldots$ are i.i.d. k-dimensional random vectors with

$$E|W_1| \equiv E|W_{1,1}| + \cdots + E|W_{1,k}| < \infty,$$

where $W_1 = (W_{1,1}, \ldots, W_{1,k})$, then

$$\bar{W}_n = \frac{1}{n}\sum_{i=1}^{n} W_i \xrightarrow{p} E(W_1) = \begin{pmatrix} E(W_{1,1}) \\ \vdots \\ E(W_{1,k}) \end{pmatrix}.$$

2. Continuous Mapping Theorem Suppose $Y_n \xrightarrow{p} Y$, where $\{Y_n\}$ and Y are k-dimensional random vectors, $k \geq 1$.

(a) For any continuous function $g : \mathbb{R}^k \to \mathbb{R}^p$ $(p \geq 1)$,

$$g(Y_n) \xrightarrow{p} g(Y).$$

(b) If $P(Y = c) = 1$ for some $c \in \mathbb{R}^k$ and g is continuous at c, then

$$g(Y_n) \xrightarrow{p} g(c).$$

2. Mean Squared Error Consistency (MSEC)

$\{T_n\}$ is called **mean squared error consistent (MSEC)** for $\gamma(\theta)$ if

$$\lim_{n\to\infty} \mathrm{MSE}_\theta(T_n) \equiv \lim_{n\to\infty} E_\theta\big[(T_n - \gamma(\theta))^2\big] = 0, \quad \forall \theta \in \Theta.$$

3. Asymptotic Unbiasedness

$\{T_n\}$ is called **asymptotically unbiased** for $\gamma(\theta)$ if

$$\lim_{n\to\infty} E_\theta(T_n) = \gamma(\theta), \quad \forall \theta \in \Theta.$$

Remarks and Relationships on Consistency

1. MSEC $\Rightarrow$ Consistency Pick $\varepsilon > 0$ and use Chebyshev's inequality:

$$P_\theta(|T_n - \gamma(\theta)| > \varepsilon) = P_\theta\big((T_n - \gamma(\theta))^2 > \varepsilon^2\big) \leq \frac{E_\theta[(T_n - \gamma(\theta))^2]}{\varepsilon^2} = \frac{\mathrm{MSE}_\theta(T_n)}{\varepsilon^2} \to 0.$$

Hence, T_n is consistent for $\gamma(\theta)$.

2. Characterization of MSEC $\{T_n\}$ is MSEC if and only if

$$\lim_{n\to\infty} \mathrm{Var}_\theta(T_n) = 0 \quad \text{and} \quad \{T_n\} \text{ is asymptotically unbiased.}$$

This follows from

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

3. Logical Relationships

- MSEC $\Rightarrow$ asymptotically unbiased
- MSEC $\Rightarrow$ consistency
- Asymptotically unbiased $\not\Rightarrow$ consistency
- Consistency $\not\Rightarrow$ asymptotically unbiased

4. Asymptotic Efficiency

Recall: For two unbiased estimators T and T^* of $\gamma(\theta)$, relative efficiency is

$$\mathrm{RE}(T, T^*; \theta) = \frac{\mathrm{Var}_\theta(T^*)}{\mathrm{Var}_\theta(T)}.$$

Similarly, we compare large-sample variances of asymptotically unbiased estimators.
Let $\{T_n\}$ and $\{T_n^*\}$ be asymptotically unbiased for $\gamma(\theta)$.

1. Asymptotic Relative Efficiency (ARE)

$$\mathrm{ARE}(T_n, T_n^*; \theta) = \lim_{n\to\infty} \frac{\mathrm{Var}_\theta(T_n^*)}{\mathrm{Var}_\theta(T_n)}, \quad \theta \in \Theta.$$

2. Asymptotically Efficient Estimator $\{T_n^*\}$ is called **asymptotically efficient** if

$$\mathrm{ARE}(T_n, T_n^*; \theta) \leq 1$$

for all $\theta \in \Theta$ and any other $\{T_n\}$ that is asymptotically unbiased for $\gamma(\theta)$.

3. Asymptotic Efficiency If $\{T_n^*\}$ is asymptotically efficient, define

$$\mathrm{AE}(T_n, \theta) \equiv \mathrm{ARE}(T_n, T_n^*; \theta).$$

Asymptotic Properties of MLEs

Let $X_1, X_2, \ldots, X_n$ be i.i.d. with common pdf/pmf $f(x \mid \theta)$, $\theta \in \Theta \subset \mathbb{R}$.

Let $\hat{\theta}_n$ be the MLE of θ.

Under regularity conditions, as $n \to \infty$:

1. **Consistency**

$$\hat{\theta}_n \xrightarrow{p} \theta.$$

2. **Asymptotic Normality**

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, \frac{1}{I_1(\theta)}\right),$$

where

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

3. **Asymptotic Efficiency**

The sequence $\{\hat{\theta}_n\}$ is asymptotically efficient.

5. Asymptotic Normality: Delta Method

Suppose

$$\sqrt{n}(T_n - a) \xrightarrow{d} N(0, \sigma^2(a))$$

for some $a \in \mathbb{R}$.

If $g : \mathbb{R} \to \mathbb{R}$ is continuously differentiable at a with $g'(a) \neq 0$, then

$$\sqrt{n}(g(T_n) - g(a)) \xrightarrow{d} N\big(0, [g'(a)]^2\sigma^2(a)\big).$$

Sufficiency and Completeness (In Point Estimation)

Sufficiency as Data Reduction

Let $X_1, \ldots, X_n$ be random variables with joint pdf/pmf $f(x \mid \theta)$, $\theta \in \Theta \subset \mathbb{R}^p$.

Let $S = (S_1, \ldots, S_k)$ be a vector of estimators.

S is called **(jointly) sufficient** for θ if the conditional distribution of $(X_1, \ldots, X_n)$ given S does **not** depend on θ.

Factorization Theorem

S is sufficient for θ **if and only if** there exist functions $g(S, \theta)$ and $h(x)$ such that $h(x)$ does not depend on θ and

$$f(x \mid \theta) = g(S, \theta)h(x), \quad \text{for all } x \text{ and all } \theta.$$

Remarks:

1. The choice of $g(\xi, \theta)$ and $h(x)$ is not unique.
2. Any 1-to-1 function of a sufficient statistic is also sufficient.

Example: In last example, suppose $A = I_{n\times n}$.

Minimal Sufficiency

Definition. A vector of statistics S is called **minimally sufficient** if

1. S is sufficient for θ, and
2. for any other vector T of sufficient statistics for θ, S is a function of T.

Remarks on Sufficiency

1. If $X_1, \ldots, X_n$ is a random sample (i.i.d.) from pdf/pmf $f(x \mid \theta)$, $\theta \in \Theta$, then the order statistics

$$X_{(1)}, \ldots, X_{(n)}$$

are sufficient for θ.

Proof. By the factorization theorem,

$$f(x \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = g(x_{(1)}, \ldots, x_{(n)}, \theta)\, h(x),$$

for all x, θ, where

$$g(x_{(1)}, \ldots, x_{(n)}, \theta) = \prod_{i=1}^{n} f(x_{(i)} \mid \theta), \quad h(x) = 1.$$

2. If $S = (S_1, \ldots, S_k)$ is sufficient for real-valued $\theta \in \Theta \subset \mathbb{R}$, then any Bayes estimator is a function of S.

Example. Let $X_1, \ldots, X_n \sim$ iid Bernoulli(θ), $0 < \theta < 1$, with loss

$$L(t, \theta) = \frac{(t-\theta)^2}{\theta(1-\theta)},$$

and prior $\pi(\theta) = \text{Uniform}(0, 1)$.

Then the Bayes estimator is

$$T_0 = \bar{X}_n,$$

which is sufficient for θ by the factorization theorem.

3. If $S = (S_1, \ldots, S_k)$ is sufficient for $\theta \in \Theta \subset \mathbb{R}^p$ and $\hat{\theta}$ is the unique MLE of θ, then $\hat{\theta}$ is a function of S.

Rao–Blackwell Theorem and Sufficiency

Rao–Blackwell Theorem.
Let $f(x \mid \theta) = f(x_1, \ldots, x_n \mid \theta)$ be the joint pdf/pmf of $(X_1, \ldots, X_n)$ and let

$$S = (S_1, \ldots, S_k)$$

be sufficient for $\theta = (\theta_1, \ldots, \theta_p) \in \Theta \subset \mathbb{R}^p$.

Let T be any unbiased estimator (UE) of a real-valued $\gamma(\theta)$ and define

$$T^* = E(T \mid S).$$

Then:

1. T^* is a function of S and an unbiased estimator of $\gamma(\theta)$.
2.

$$\mathrm{Var}_\theta(T^*) \le \mathrm{Var}_\theta(T), \quad \forall \theta \in \Theta.$$

3. If $\mathrm{Var}_{\theta_0}(T^*) = \mathrm{Var}_{\theta_0}(T)$ for some $\theta_0 \in \Theta$, then

$$P_{\theta_0}(T = T^*) = 1.$$

Remark. Any UE T can be improved (or left unchanged) by conditioning on a sufficient statistic S. This process is called **Rao–Blackwellization**.

Completeness

Definition. A statistic T is called **complete** if the only function $u(T)$ of T that is an unbiased estimator of zero is $u(T) = 0$ with probability 1.

Equivalently:

Let $f(x \mid \theta)$ be the joint pdf/pmf of $(X_1, \ldots, X_n)$ and let $f_T(t \mid \theta)$ denote the sampling distribution of T.

T (or the family $\{f_T(t \mid \theta) : \theta \in \Theta\}$) is **complete** if, for any real-valued function $u(T)$,

$$E_\theta[u(T)] = 0 \ \forall \theta \in \Theta \quad \Longrightarrow \quad P_\theta(u(T) = 0) = 1 \ \forall \theta \in \Theta.$$

T is called **boundedly complete** if the above implication holds for all bounded functions $u(\cdot)$.

Remarks on Completeness

1. If T is complete, then T is boundedly complete (the converse is false).
2. If T is sufficient and boundedly complete, then T is minimal sufficient.
   Hence, if T is sufficient and complete, then T is minimal sufficient.
3. Suppose T is complete and $h_1(T)$, $h_2(T)$ are two estimators of $\gamma(\theta)$ such that

$$E_\theta[h_1(T)] = \gamma(\theta) = E_\theta[h_2(T)], \quad \forall \theta \in \Theta.$$

Then defining $u(T) = h_1(T) - h_2(T)$,

$$E_\theta[u(T)] = 0 \ \forall \theta \Rightarrow P_\theta(u(T) = 0) = 1 \ \forall \theta,$$

so

$$P_\theta(h_1(T) = h_2(T)) = 1 \ \forall \theta.$$

Hence, there is at most one (unique) unbiased estimator of $\gamma(\theta)$ that is a function of a complete statistic.

Lehmann–Scheffé Theorem

Lehmann–Scheffé Theorem.

Let $f(x \mid \theta) = f(x_1, \ldots, x_n \mid \theta)$ be the joint pdf/pmf of $(X_1, \ldots, X_n)$, $\theta \in \Theta \subset \mathbb{R}^p$.

Let

$$S = (S_1, \ldots, S_k)$$

be a complete and sufficient statistic.

If $T^* = T(S)$ is an unbiased estimator of $\gamma(\theta)$ and is a function of S, then T^* is the **UMVUE** of $\gamma(\theta)$.

Proof (sketch).
Let T be any UE of $\gamma(\theta)$ and define

$$T_1 = E(T \mid S).$$

By the Rao–Blackwell theorem, T_1 is a function of S, is unbiased, and

$$\mathrm{Var}_\theta(T_1) \leq \mathrm{Var}_\theta(T).$$

Since S is complete, $T_1 = T^*$ with probability 1, so

$$\mathrm{Var}_\theta(T^*) \leq \mathrm{Var}_\theta(T), \quad \forall \theta.$$

Exponential Families

Definition. A family of pdf/pmf $\{f(x \mid \theta) : \theta \in \Theta\}$ is called an **exponential family** if it can be written as

$$f(x \mid \theta) = \begin{cases} c(\theta)h(x)\exp\left(\sum_{i=1}^{k} q_i(\theta)t_i(x)\right), & x \in A, \\ 0, & \text{otherwise,} \end{cases}$$

where

- $A = \{x : f(x \mid \theta) > 0\}$ does not depend on θ,
- $c(\theta) > 0$, $h(x) > 0$,
- $q_i(\theta)$ and $t_i(x)$ are real-valued functions.

Theorem.
Let $X_1, \ldots, X_n$ be a (possibly vector-valued) random sample from an exponential family.

If

$$\{(q_1(\theta), \ldots, q_k(\theta)) : \theta \in \Theta\} \supset (a_1, b_1) \times \cdots \times (a_k, b_k),$$

then

$$S = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j)\right)$$

is complete and sufficient.

Ancillarity

Definition. A statistic T is called **ancillary** if its distribution does **not** depend on any unknown parameters.

Basu's Theorem

Basu's Theorem.

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

Suppose $S = (S_1, \ldots, S_k)$ is complete and sufficient, and $T = (T_1, \ldots, T_m)$ is ancillary.

Then S and T are independent for all θ.

Hypothesis Testing I

Definition.

1. A (statistical) hypothesis is a statement about a population parameter.
2. The two complementary hypotheses in a testing problem are called the **null hypothesis** H_0 and the **alternative hypothesis** H_1 (or H_a).

Example.
Let $\theta = (\theta_1, \theta_2)$ denote average blood sugar levels before and after taking a drug.

- H_0: no effect $\iff \theta_1 = \theta_2$
- H_1: effective $\iff \theta_1 \neq \theta_2$ or $\theta_1 > \theta_2$

Definition.
If a statistical hypothesis H completely specifies the distribution of $(X_1, \ldots, X_n)$, then H is called **simple**; otherwise, H is called **composite**.

Test Functions

Definition.
Let $\mathcal{X}$ be the set of all possible values of $X = (X_1, \ldots, X_n)$.

A **test function** (or test rule) $\phi(X_1, \ldots, X_n)$ is a function

$$\phi : \mathcal{X} \to [0, 1],$$

with the interpretation that if $X = (X_1, \ldots, X_n)$ is observed, then H_0 is rejected with probability

$$\phi(X) \equiv \phi(X_1, \ldots, X_n).$$

Definition.
If $\phi(X) \in \{0, 1\}$ for all $X \in \mathcal{X}$, then:

1. $\phi(X)$ is called a **simple test function (rule)**.
2. $$R_\phi = \{X : \phi(X) = 1\}$$
is called the **rejection region** of $\phi(X)$.
3. $$A_\phi = \{X : \phi(X) = 0\}$$
is called the **acceptance region** of $\phi(X)$.

Error Probabilities

Suppose $\phi(X) \in \{0, 1\}$ is a simple test rule for testing

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \notin \Theta_0,$$

where $\Theta_0 \subset \Theta$.

action based on $\phi(\cdot)$

	fail to reject H_0	reject H_0
the state of nature		
H_0 is true	OK	Type I error
H_1 is true (or H_0 false)	Type II error	OK

1. For any $\theta \in \Theta_0$, the probability of a **Type I error** is

$$P_\theta(\text{reject } H_0) = P_\theta(X \in R_\phi) = E_\theta[\phi(X)].$$

2. For any $\theta \notin \Theta_0$, the probability of a **Type II error** is

$$P_\theta(\text{fail to reject } H_0) = P_\theta(X \in A_\phi) = 1 - P_\theta(X \in R_\phi) = 1 - E_\theta[\phi(X)].$$

Remark.
For any general (possibly randomized) test function $\phi(X) \in [0, 1]$, the same expressions hold:

1. For $\theta \in \Theta_0$,

$$P_\theta(\text{Type I error}) = E_\theta[\phi(X)].$$

2. For $\theta \notin \Theta_0$,

$$P_\theta(\text{Type II error}) = 1 - E_\theta[\phi(X)].$$

Size and Power

Definition.
Let $\phi(X)$ be a test rule for testing $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \notin \Theta_0$.

1.

$$\max_{\theta \in \Theta_0} E_\theta[\phi(X)]$$

is called the **size** (or **level**) of $\phi(X)$.

2.

$$\Pi_\phi(\theta) = E_\theta[\phi(X)]$$

is called the **power function** of $\phi(X)$.

Note.

- For $\theta \in \Theta_0$, $\Pi_\phi(\theta)$ is the probability of a Type I error.
- For $\theta \notin \Theta_0$, $1 - \Pi_\phi(\theta)$ is the probability of a Type II error.

Error Probabilities, Size, and Power

In general, it is not possible to minimize both Type I and Type II error probabilities simultaneously (for a fixed sample size).

One can at best minimize the probability of Type II error while fixing the probability of Type I error at a given level α.

Equivalently, maximizing power

$$\Pi_\phi(\theta) = E_\theta[\phi(X)]$$

for $\theta \notin \Theta_0$ subject to

$$\max_{\theta \in \Theta_0} E_\theta[\phi(X)] \leq \alpha$$

is the goal of hypothesis testing.

Most Powerful Tests (Simple vs. Simple)

Let $f(x \mid \theta)$ be the joint pdf/pmf of $X = (X_1, \ldots, X_n)$.

We wish to test

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1, \qquad \theta_0 \neq \theta_1.$$

Definition.
A test function $\phi(x)$ is called a **most powerful (MP)** test of size α if:

1.

$$E_{\theta_0}[\phi(X)] = \alpha,$$

2.

$$E_{\theta_1}[\phi(X)] \geq E_{\theta_1}[\tilde{\phi}(X)]$$

for any other test rule $\tilde{\phi}(x)$ satisfying

$$E_{\theta_0}[\tilde{\phi}(X)] \leq \alpha.$$

Uniformly Most Powerful (UMP) Tests

Let $f(x \mid \theta)$ be the joint pdf/pmf of $X = (X_1, \ldots, X_n)$, $\theta \in \Theta \subset \mathbb{R}^p$, and let Θ_0 be a nonempty proper subset of Θ.

A test rule $\phi(x)$ for testing

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \notin \Theta_0$$

is called a **uniformly most powerful (UMP)** test of size α if:

1.

$$\max_{\theta \in \Theta_0} E_\theta[\phi(X)] = \alpha,$$

2.

$$E_\theta[\phi(X)] \geq E_\theta[\tilde{\phi}(X)]$$

for all $\theta \notin \Theta_0$ and any other test rule $\tilde{\phi}(x)$ with

$$\max_{\theta \in \Theta_0} E_\theta[\tilde{\phi}(X)] \leq \alpha.$$

Finding UMP Tests

1. **Method I:** Based on the Neyman–Pearson Lemma
2. **Method II:** Using the Monotone Likelihood Ratio (MLR) property

Method I: Neyman–Pearson Lemma

Let $f(x \mid \theta)$ be the joint pdf/pmf of $X_1, \ldots, X_n$.
For testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, a MP test of size α exists for all $\alpha \in [0, 1]$ and is given by

$$\phi(x) = \begin{cases} 1, & f(x \mid \theta_1) > k f(x \mid \theta_0), \\ \gamma, & f(x \mid \theta_1) = k f(x \mid \theta_0), \\ 0, & f(x \mid \theta_1) < k f(x \mid \theta_0), \end{cases}$$

where $\gamma \in [0, 1]$ and $0 \leq k \leq \infty$ are constants chosen to satisfy

$$E_{\theta_0}[\phi(X)] = \alpha.$$

Remarks on Neyman–Pearson Lemma

- Let

$$L(\theta) = f(x \mid \theta)$$

denote the likelihood function. The MP test rejects H_0 when $L(\theta_1)$ is large relative to $L(\theta_0)$.

- The constants (γ, k) satisfying the size condition are not unique in general.
- If the distribution of $f(X \mid \theta_1)/f(X \mid \theta_0)$ under θ_0 is continuous, then we may take $\gamma = 0$.

Method II: Monotone Likelihood Ratio (MLR)

Let $f(x \mid \theta)$, $\theta \in \Theta \subset \mathbb{R}$, be the joint pdf/pmf of $X = (X_1, \ldots, X_n)$.

The family $\{f(x \mid \theta) : \theta \in \Theta\}$ is said to have a **monotone likelihood ratio (MLR)** in a real-valued statistic $T = t(X)$ if, for any $\theta_1 < \theta_2$, there exists a nondecreasing function

$$g_{\theta_1,\theta_2} : \mathbb{R} \to [0, \infty)$$

such that

$$\frac{f(x \mid \theta_2)}{f(x \mid \theta_1)} = g_{\theta_1,\theta_2}(t(x)),$$

for all x such that $f(x \mid \theta_1) + f(x \mid \theta_2) > 0$.

Using MLR Context.

Method II applies only when $\Theta \subset \mathbb{R}$ and the testing problem is one-sided:

$$H_0 : \theta \leq \theta_0 \text{ vs. } H_1 : \theta > \theta_0 \quad \text{or} \quad H_0 : \theta \geq \theta_0 \text{ vs. } H_1 : \theta < \theta_0.$$

Theorem (UMP Tests via MLR).
Assume $\{f(x \mid \theta) : \theta \in \Theta\}$ has MLR in $T = t(X)$.

1. A size α UMP test for

$$H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0$$

is given by

$$\phi(x) = \begin{cases} 1, & t(x) > k, \\ \gamma, & t(x) = k, \\ 0, & t(x) < k, \end{cases}$$

where $\gamma \in [0, 1]$ and $k \in [-\infty, \infty]$ satisfy

$$E_{\theta_0}[\phi(X)] = \alpha.$$

2. A size α UMP test for

$$H_0 : \theta \geq \theta_0 \quad \text{vs.} \quad H_1 : \theta < \theta_0$$

is given by

$$\phi(x) = \begin{cases} 1, & t(x) < k, \\ \gamma, & t(x) = k, \\ 0, & t(x) > k, \end{cases}$$

where $\gamma \in [0, 1]$ and $k \in [-\infty, \infty]$ satisfy

$$E_{\theta_0}[\phi(X)] = \alpha.$$

Hypothesis Testing II

Likelihood Ratio Tests (LRT)

Definition.

Let $f(x \mid \theta)$, $\theta \in \Theta \subset \mathbb{R}^p$, be the joint pdf/pmf of

$$X = (X_1, \ldots, X_n),$$

and let Θ_0 be a nonempty proper subset of Θ.

The **likelihood ratio statistic (LRS)** for testing

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \notin \Theta_0$$

is defined as

$$\lambda(x) = \frac{\max_{\theta \in \Theta_0} f(x \mid \theta)}{\max_{\theta \in \Theta} f(x \mid \theta)}.$$

Interpreting $\lambda(x)$

1. If $\hat{\theta}$ is the MLE over Θ and $\tilde{\theta}$ is the maximizer of $f(x \mid \theta)$ over Θ_0, then

$$\lambda(x) = \frac{f(x \mid \tilde{\theta})}{f(x \mid \hat{\theta})} \in [0, 1].$$

2. If H_0 is true, $\lambda(x)$ is expected to be **large** (close to 1).
3. If H_0 is false, $\lambda(x)$ is expected to be **small** (close to 0).

Likelihood Ratio Test (LRT)

Definition.
A size α **likelihood ratio test (LRT)** for testing

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \notin \Theta_0$$

is defined by

$$\phi(x) = \begin{cases} 1, & \lambda(x) < k, \\ \gamma, & \lambda(x) = k, \\ 0, & \lambda(x) > k, \end{cases}$$

where $\gamma \in [0, 1]$ and $0 \leq k \leq 1$ are chosen so that

$$\max_{\theta \in \Theta_0} E_\theta[\phi(X)] = \alpha.$$

Large-Sample Calibration of LRTs

Theorem (Wilks).
Let $X_1, \ldots, X_n$ be i.i.d. with joint pdf/pmf $f(x \mid \theta)$, $\theta \in \Theta \subset \mathbb{R}^p$.

Suppose

$$\Theta_0 = \{\theta = (\theta_1, \ldots, \theta_p) \in \Theta : \theta_1 = \theta_1^0, \ldots, \theta_r = \theta_r^0\},$$

for some $1 \leq r \leq p$.

Under regularity conditions, if H_0 is true,

$$-2 \log \lambda_n(X_1, \ldots, X_n) \xrightarrow{d} \chi^2_r \quad \text{as } n \to \infty.$$

Remark.
An approximate size α LRT rejects H_0 if

$$-2 \log \lambda_n(X_1, \ldots, X_n) > \chi^2_{1-\alpha}(r),$$

where $\chi^2_{1-\alpha}(r)$ is the $(1 - \alpha)$ quantile of a χ^2_r distribution.

Equivalently,

$$\lambda_n(X_1, \ldots, X_n) < \exp\left(-\frac{1}{2}\chi^2_{1-\alpha}(r)\right).$$

Bayes Tests

Let $X_1, \ldots, X_n$ have joint pdf/pmf $f(x \mid \theta)$, $\theta \in \Theta \subset \mathbb{R}^p$, and consider testing

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \notin \Theta_0.$$

Assume:

- $\pi(\theta)$ is a prior pdf,
- the posterior density is

$$f_{\theta|x}(\theta) \propto \pi(\theta) f(x \mid \theta).$$

Define posterior probabilities:

$$P(\theta \in \Theta_0 \mid x) = \int_{\Theta_0} f_{\theta|x}(\theta)\, d\theta,$$
$$P(\theta \notin \Theta_0 \mid x) = \int_{\Theta \setminus \Theta_0} f_{\theta|x}(\theta)\, d\theta.$$

Bayes Test Rule

A Bayes test for H_0 vs. H_1 is

$$\phi(x) = \begin{cases} 1, & P(\theta \notin \Theta_0 \mid x) \geq P(\theta \in \Theta_0 \mid x), \\ 0, & \text{otherwise}, \end{cases}$$

equivalently,

$$\phi(x) = \begin{cases} 1, & P(\theta \notin \Theta_0 \mid x) \geq \frac{1}{2}, \\ 0, & \text{otherwise}. \end{cases}$$

Decision-Theoretic Justification

Consider a simple test $\phi_1(x) \in \{0, 1\}$ with actions:

- reject H_0 if $\phi_1(x) = 1$,
- do not reject H_0 if $\phi_1(x) = 0$.

Define the 0–1 loss:

$$L(\theta, \phi_1(x)) = I_{\{\theta \in \Theta_0\}} I_{\{\phi_1(x)=1\}} + I_{\{\theta \notin \Theta_0\}} I_{\{\phi_1(x)=0\}}.$$

The posterior risk is

$$E_{\theta|x}[L(\theta, \phi_1(x))] = I_{\{\phi_1(x)=1\}} P(\theta \in \Theta_0 \mid x) + I_{\{\phi_1(x)=0\}} P(\theta \notin \Theta_0 \mid x).$$

Minimizing posterior risk yields the Bayes test above.

Interval Estimation

Introduction - Terminology

Let $X_1, \ldots, X_n$ have joint pdf/pmf $f(x \mid \theta)$, $\theta \in \Theta \subset \mathbb{R}$.
Let $L(X)$ and $U(X)$ be statistics such that $L(X) \leq U(X)$.

1. The random interval

$$I(X) = [L(X), U(X)]$$

is called an **interval estimator** for θ.

2.

$$I(X) = (-\infty, U(X)]$$

is a **one-sided upper interval estimator**.

3.

$$I(X) = [L(X), \infty)$$

is a **one-sided lower interval estimator**.

4. The **coverage probability** at θ is

$$P_\theta(\theta \in I(X)).$$

5. The **confidence coefficient (CC)** is

$$\min_{\theta \in \Theta} P_\theta(\theta \in I(X)).$$

Remarks.

1. An interval estimator $I(X) = [L(X), U(X)]$ together with its confidence coefficient is called a **confidence interval**.
2. If $\theta \in \Theta \subset \mathbb{R}^p$ is vector-valued, confidence intervals are replaced by **confidence regions** (or confidence sets).

General Methods of Interval Estimation

1. Inverting a Test
2. Pivotal Quantities and Asymptotically Pivotal Quantities
3. Bayes "Credible" Intervals

Inverting a Test

Theorem.
Let $X_1, \ldots, X_n$ have joint pdf/pmf $f(x \mid \theta)$, $\theta \in \Theta \subset \mathbb{R}^p$.

Let $A(\theta_0)$ denote the acceptance region of a size α test for

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0.$$

Define, for observed x,

$$C_x = \{\theta_0 : x \in A(\theta_0)\}.$$

Then C_X is a **confidence set** for θ with confidence coefficient $1 - \alpha$.

Pivotal Quantities

Definition.
Let $X_1, \ldots, X_n$ have joint pdf/pmf $f(x \mid \theta)$, where $\theta \in \Theta \subset \mathbb{R}^p$. A random variable $Q(X, \theta)$ is called a **pivot** (or **pivotal quantity**) if the distribution of

$$Q(X, \theta)$$

under θ does **not depend on** θ.

Note.

- $Q(X, \theta)$ is **not necessarily a statistic** (it may depend on θ).
- The defining property is
$$P_\theta(Q(X, \theta) \in A) = P(Q \in A),$$
independent of θ.

Examples

1. If $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

and

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

is a pivot.

If σ^2 is unknown, then

$$\frac{\bar{X}_n - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

is also a pivot.

2. Let f_0 be a pdf on $\mathbb{R}$ and suppose

$$f(x \mid \theta) = \frac{1}{\theta_2} f_0\left(\frac{x - \theta_1}{\theta_2}\right), \quad x \in \mathbb{R},$$

where $\theta = (\theta_1, \theta_2)$, $\theta_1 \in \mathbb{R}$ (location) and $\theta_2 > 0$ (scale).

Define

$$Q(X, \theta) = \frac{\bar{X}_n - \theta_1}{\theta_2}.$$

Let

$$Y_i = \frac{X_i - \theta_1}{\theta_2} \sim f_0,$$

so that

$$Q(X, \theta) = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Hence the distribution of $Q(X, \theta)$ does not depend on θ, so Q is a pivot.

Interval Estimation via Pivotal Quantities

Remark.
Let $Q(X, \theta)$ be a pivot and let $0 < \alpha < 1$. Suppose $a \leq b$ satisfy

$$P_\theta(a \leq Q(X, \theta) \leq b) = 1 - \alpha.$$

Then

$$C_X = \{\theta \in \Theta : a \leq Q(X, \theta) \leq b\}$$

is a **confidence region** for θ with confidence coefficient $1 - \alpha$.

That is,

$$\min_{\theta \in \Theta} P_\theta(\theta \in C_X) = 1 - \alpha.$$

Remark. If $\Theta \subset \mathbb{R}$ and $Q(X, \theta)$ is monotone in θ, then C_X is an **interval**.

Asymptotically Pivotal Quantities

Definition.
A sequence $Q_n = Q_n(X_1, \ldots, X_n, \theta)$ is called **asymptotically pivotal** if

$$Q_n \xrightarrow{d} Q \quad \text{as } n \to \infty,$$

where the limiting distribution of Q does not depend on θ.

Example

If $X_1, \ldots, X_n$ are iid with

$$E(X_i) = \mu, \qquad \text{Var}(X_i) = \sigma^2,$$

then by the CLT,

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2),$$

and hence

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1).$$

Remark.
If Q_n is asymptotically pivotal and

$$P(a \leq Q \leq b) = 1 - \alpha,$$

then for large n,

$$C_X = \{\theta : a \leq Q_n(X, \theta) \leq b\}$$

is an **approximate** confidence region with confidence coefficient approximately $1 - \alpha$.

Variance Stabilizing Transformations (VST)

Definition.

Let $\hat{\theta}_n$ be an estimator such that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta)).$$

A function $g : \mathbb{R} \to \mathbb{R}$ is a **variance stabilizing transformation (VST)** if

$$g'(\theta)\sigma(\theta) = 1.$$

Remark.
By the Delta Method,

$$\sqrt{n}\big(g(\hat{\theta}_n) - g(\theta)\big) \xrightarrow{d} N(0, 1).$$

Thus,

$$Q_n(X, \theta) = \sqrt{n}\big(g(\hat{\theta}_n) - g(\theta)\big)$$

is asymptotically pivotal.

An approximate $(1 - \alpha)$ CI for θ is

$$C_X = \left\{\theta : z_{\alpha/2} \leq \sqrt{n}\big(g(\hat{\theta}_n) - g(\theta)\big) \leq z_{1-\alpha/2}\right\}.$$

Mood–Graybill–Boes (MGB) Interval Method

Let T be a statistic with cdf

$$F(t \mid \theta) = P(T \leq t \mid \theta),$$

where $\theta \in \Theta \subset \mathbb{R}$.

By the probability integral transform,

$$Q(T, \theta) = F(T \mid \theta) \sim \text{Uniform}(0, 1),$$

so $Q(T, \theta)$ is a pivot.

For $0 < \alpha < 1$,

$$P\Big(\frac{\alpha}{2} \leq Q \leq 1 - \frac{\alpha}{2}\Big) = 1 - \alpha.$$

Thus,

$$C_T = \Big\{\theta : \frac{\alpha}{2} \leq F(T \mid \theta) \leq 1 - \frac{\alpha}{2}\Big\}$$

is a confidence region with C.C. $1 - \alpha$.

Inversion

For observed $T = t$, if $F(t \mid \theta)$ is monotone in θ, then

$$C_T = [\theta_L(t), \theta_U(t)],$$

where

$$F(t \mid \theta_L(t)) = \frac{\alpha}{2}, \qquad F(t \mid \theta_U(t)) = 1 - \frac{\alpha}{2}.$$

Discrete vs Continuous T

- If T is **continuous**, then

$$\min_{\theta \in \Theta} P_\theta(\theta \in C_T) = 1 - \alpha.$$

- If T is **discrete**, then

$$\min_{\theta \in \Theta} P_\theta(\theta \in C_T) \geq 1 - \alpha,$$

so the interval is **conservative**.

Conservative Confidence Intervals

A CI I is called **conservative** if

$$\min_{\theta \in \Theta} P_\theta(\theta \in I) \geq 1 - \alpha.$$

Discrete-data intervals obtained via inversion are typically conservative.

Bayes Intervals

Let $X = (X_1, \ldots, X_n)$ with joint pdf/pmf $f(x \mid \theta)$, prior $\pi(\theta)$, and posterior density $f_{\theta|x}(\theta)$.
A set C_x is a $(1 - \alpha)$ **credible set** if

$$P(\theta \in C_x \mid X = x) = 1 - \alpha,$$

or equivalently,

$$\int_{C_x} f_{\theta|x}(\theta)\, d\theta = 1 - \alpha.$$

Note. A credible set is **not** a confidence region; it does not satisfy frequentist coverage guarantees.

Highest Posterior Density (HPD) Sets

A $(1 - \alpha)$ **HPD credible set** has the form

$$C_x = \{\theta : f_{\theta|x}(\theta) \geq c\},$$

for some cutoff $c > 0$, chosen so that

$$P(\theta \in C_x \mid X = x) = 1 - \alpha.$$

Evaluating Interval Estimators

Remark 1.

For two CIs $I_C = [L_C(X), U_C(X)]$ and $I_D = [L_D(X), U_D(X)]$ with the same confidence coefficient $1 - \alpha$, I_D is preferred if

$$E_\theta[\text{length}(I_D)] \leq E_\theta[\text{length}(I_C)], \quad \forall \theta \in \Theta.$$

Remark 2.
For confidence regions C_X and D_X in $\mathbb{R}^p$, D_X is preferred if

$$E_\theta[\text{volume}(D_X)] \leq E_\theta[\text{volume}(C_X)], \quad \forall \theta.$$

Theorem on Shortest Intervals

Let $f(x)$ be a unimodal pdf with mode x^*.

If $[a, b]$ satisfies:

1. $\int_a^b f(x)\,dx = 1 - \alpha$
2. $f(a) = f(b) > 0$
3. $a \leq x^* \leq b$

then $[a, b]$ has **minimum length** among all intervals with probability $1 - \alpha$.