

Stochastic Processes



Week 06 (version 1.0)
Estimation Theory 01

Hamid R. Rabiee

Fall 2024

Outline of Week 06 Lectures

- Introduction to Estimation Theory
- Sufficient Statistic
- Minimal Sufficient Statistic
- Complete Sufficient Statistic
- Likelihood Principle
- Frequentist's Estimators: MLE, MM

Introduction to Estimation Theory

- **Estimation Theory:** is a branch of statistics that deals with estimating the values of parameters based on observed data that has a random component.
- In this course we focus on point estimation:
Given $X = \{x_1, x_2, \dots, x_n\}$ where x_i s are independent and identically distributed **(i.i.d) observations** with $f(x_i|\theta)$, we want to find an statistics $T(X) = \hat{\theta}$ that is a **good estimator** for θ .

Introduction to Estimation Theory

- Three basic Questions:
 - 1) Do we need all the i.i.d observations to estimate θ ?
 - 2) What do we mean by “good estimator”?
 - 3) Do we need prior information on θ (i.e. $f(\theta)$) to estimate it?
- Answers:
 - 1) Not necessarily! We may use **Sufficient Statistic (SS)**; a function or statistic of observations, instead.
 - 2) The goodness of an estimator is measured by three properties: unbiasedness, efficiency (minimum variance) and consistency.

Introduction to Estimation Theory

- **Unbiasedness:**

An estimator $\hat{\theta}$ is said to be unbiased if its expected value is identical to θ ; $E(\hat{\theta}) = \theta$.

- **Efficiency:**

If two competing estimators are both unbiased, the one with the smaller variance is said to be relatively more efficient.

- **Consistency:**

If an estimator $\hat{\theta}$ approaches the parameter θ closer and closer as the sample size n increases, $\hat{\theta}$ is said to be a consistent estimator of θ (not a rigorous definition).

Introduction to Estimation Theory

- 3) The frequentist believe we do not need prior information on θ (i.e. $f(\theta)$) to estimate it. However, the Bayesian believe we do need prior information on θ .

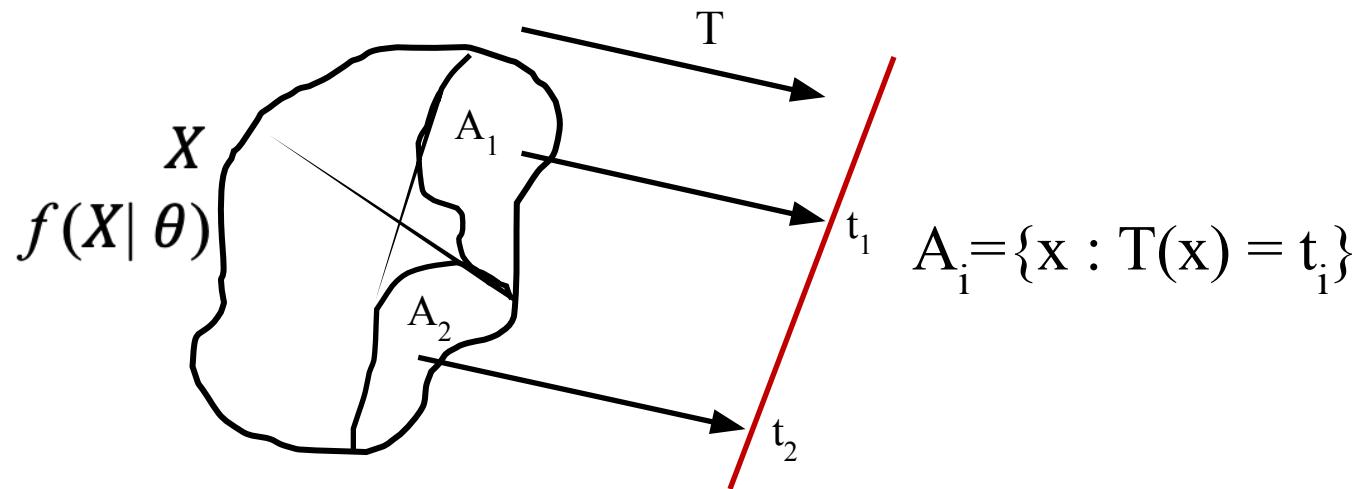
In the following we focus on Sufficient Statistic.

Outline of Week 06 Lectures

- Introduction to Estimation Theory
- **Sufficient Statistic**
- Minimal Sufficient Statistic
- Complete Sufficient Statistic
- Likelihood Principle
- Frequentist's Estimators: MLE, MM

Sufficient Statistic (SS)

Assume the statistic T partitions the sample space into sets.



Goal of SS: Data reduction without discarding information about θ . Examples of statistics:

$$T(X) = 2$$

$$T(X) = X$$

Sufficient Statistic

- A statistic $T(X)$ is a sufficient statistic for θ if the conditional density of X given the value of $T(X)$ does not depend on θ .
- In other words, if $T(X)$ is a sufficient statistic for θ then any inference about θ should depend on the sample X only through $T(X)$; meaning $\hat{\theta}$ is a function of $T(X)$.
- How to find sufficient statistics for θ ?

Sufficient Statistic

Factorization Theorem:

Let $f(x|\theta)$ be the pdf of X .

$T(X)$ is a sufficient stat for θ iff \exists functions g and h such that:

$$f(x|\theta) = g(T(x)|\theta) h(x) \quad \forall x \in \chi, \quad \theta \in \Theta$$

proof: (discrete case)

\Rightarrow : Assume T is a sufficient statistic:

$$\begin{aligned} f(x|\theta) &= P_\theta(X = x, T(X) = T(x)) \\ &= \underbrace{P_\theta(T(X) = T(x))}_{g(T(x)|\theta)} \underbrace{P_\theta(X = x | T(X) = T(x))}_{h(x)} \end{aligned}$$

Sufficient Statistic

\Leftarrow : Assume factorization holds, let $q(t|\theta)$ be the pmf of $T(X)$

Let $A_t = \{y: T(y) = t\}$

$$q(t|\theta) = P_\theta(T(X) = t) = \sum_{x \in A_t} f(x|\theta) = \sum_{x \in A_t} g(T(x)|\theta)h(x)$$

$$P_\theta(X = x | T(X) = T(x)) = \frac{P_\theta(X=x, T(X)=T(x))}{P_\theta(T(X)=T(x))} = \frac{P_\theta(X=x)}{q(t|\theta)}$$

$$= \frac{g(T(x)|\theta)h(x)}{g(T(x)|\theta) \sum_{x \in A_t} h(x)} = \frac{h(x)}{\sum_{x \in A_t} h(x)} \text{ does not depend on } \theta.$$

Sufficient Statistic

Example: x_1, \dots, x_n be i.i.d Bernouli(θ), $0 < \theta < 1$.

Then $T(x) = \sum_{i=1}^n x_i$ is a sufficient statistic for θ .

$$f(x|\theta) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\sum x_i} (1-\theta)^{n-\sum x_i}$$

$$g(t|\theta) := \theta^t (1-\theta)^{n-t}$$

$$h(x) := 1$$

Sufficient Statistic

Example: x_1, \dots, x_n be i.i.d $\text{U}(0, \theta)$.

$$f(x_1, \dots, x_n | \theta) = \begin{cases} \frac{1}{\theta^n} & \text{all } x_i \text{ in } [0, \theta] \\ 0 & \text{o.w.} \end{cases}$$

$$\text{Recall: } I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{o.w.} \end{cases}$$

$$\text{Let: } T(x) = \max_i x_i$$

$$\text{Define: } g(t|\theta) := \frac{1}{\theta^n} I_{(-\infty, \theta]}(t) \quad h(x) = I_{[0, +\infty)} \left(\min_i x_i \right)$$

$$\Rightarrow g(T(x)|\theta)h(x) = \frac{1}{\theta^n} I_{(-\infty, \theta]} \left(\max_i x_i \right) \cdot I_{[0, +\infty)} \left(\min_i x_i \right) = f(x_1, \dots, x_n | \theta)$$

$\Rightarrow T(X)$ is sufficient statistic.

Sufficient Statistic

Example: x_1, \dots, x_n be i.i.d $\text{Normal}(\mu, \delta^2)$.

$$f(x|\mu, \delta^2) = (2\pi\delta^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\delta^2}\right)$$

We show that following t_1 and t_2 together is a sufficient statistic.

$$t_1 = \sum_{i=1}^n (x_i - \bar{x})^2, \quad t_2 = \bar{x}$$

need: $g(t_1, t_2 | \theta)$

$$g(t|\theta) = g(t_1, t_2 | \mu, \delta^2) = (2\pi\delta^2)^{-\frac{n}{2}} \exp\left(-\frac{(t_2 + n(t_1 - \mu))}{2\delta^2}\right)$$

$$h(x) = 1$$

$\Rightarrow T(X)$ is sufficient statistic.

Sufficient Statistic

Exponential Family:

Family of pdfs or pmfs is called a k-parameter exponential family if:

$$f(x|\theta) = h(x) c(\theta) \exp\left(\sum_{i=1}^k w_i(\theta) t_i(x)\right)$$

Example: x_1, \dots, x_n be i.i.d Bernouli(θ), $0 < \theta < 1$.

$$\begin{aligned} f(x|\theta) &= \theta^{\sum x_i} (1-\theta)^{n-\sum x_i} = \exp\left(\ln \theta \sum_{i=1}^n x_i + \ln(1-\theta) \left(n - \sum_{i=1}^n x_i\right)\right) \\ &= \exp\left(\ln \frac{\theta}{1-\theta} \sum_{i=1}^n x_i + n \ln(1-\theta)\right) = \exp(n \ln(1-\theta)) \cdot \exp\left(\ln \frac{\theta}{1-\theta} \sum_{i=1}^n x_i\right) \end{aligned}$$

$$k = 1, \quad h(x) = 1, \quad c(\theta) = \exp(n \ln(1-\theta)), \quad t_1 = \sum_{i=1}^n x_i, \quad w_1(\theta) = \ln \frac{\theta}{1-\theta}$$

Sufficient Statistic

Example: x_1, \dots, x_n be i.i.d $\text{Normal}(\mu, \delta^2)$.

$$f(x|\mu, \delta^2) = \frac{1}{\sqrt{2\pi\delta^2}} \exp\left(-\frac{(x-\bar{\mu})^2}{2\delta^2}\right) = \frac{1}{\sqrt{2\pi}} \frac{1}{\delta} \exp\left(-\frac{\mu^2}{2\delta^2}\right) \exp\left(-\frac{x^2}{2\delta^2} + \frac{\mu x}{\delta^2}\right)$$

Exponential family:

$$f(x|\theta) = h(x) c(\theta) \exp\left(\sum_{i=1}^k w_i(\theta) t_i(x)\right)$$

\Rightarrow

$$k = 2, \quad h(x) = 1, \quad c(\mu, \delta^2) = \frac{1}{\sqrt{2\pi}} \frac{1}{\delta} \exp\left(-\frac{\mu^2}{2\delta^2}\right),$$

$$t_1(x) = \frac{x^2}{2}, \quad w_1(\mu, \delta^2) = \frac{1}{\delta^2}$$

$$t_2(x) = x, \quad w_2(\mu, \delta^2) = \frac{\mu}{\delta^2}$$

Sufficient Statistic

Sufficient statistic for exponential family:

Let x_1, \dots, x_n be i.i.d observations from a pdf or pmf $f(x|\theta)$. Suppose $f(x|\theta)$ belongs to the exponential family:

$$f(x|\theta) = h(x) c(\theta) \exp\left(\sum_{i=1}^k w_i(\theta) t_i(x)\right)$$

Then

$T(X) = (\sum_{i=1}^n t_1(x_i), \sum_{i=1}^n t_2(x_i), \dots, \sum_{i=1}^n t_k(x_i))$ is a sufficient statistic for θ .

Example: x_1, \dots, x_n be i.i.d $\text{Normal}(\mu, \delta^2)$.

$$t_1(x) = -\frac{x^2}{2} \quad t_2(x) = x$$

Sufficient Statistic

$\Rightarrow T(X) = \left(-\frac{1}{2} \sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i \right)$ is sufficient statistic for (μ, δ^2)

$$T'(X) = \left(\sum_{i=1}^n (x_i - \bar{x})^2, \bar{x} \right)$$

$$T(X) = T(Y) \quad iff \quad T'(X) = T'(Y)$$

Results:

- 1) $T(X) = X$ is a sufficient statistic.

Proof:

$$f(x|\theta) = f(T(x)|\theta)h(x)$$

$$T(x) = x, \quad h(x) = 1$$

Sufficient Statistic

2) Any **one-to-one function of a sufficient statistic** is also a sufficient statistic.

Proof: Suppose T is a sufficient statistic.

Define $T^*(x) = r(T(x))$ where r is one-to-one and has inverse r^{-1}

$$f(x|\theta) = g(T(x)|\theta)h(x) = g(r^{-1}(T^*(x))|\theta)h(x)$$

Define $g^*(t|\theta) = g(r^{-1}(t)|\theta)h(x)$

$\Rightarrow f(x|\theta) = g^*(T^*(x)|\theta) h(x)$ so T^* is a sufficient static for θ .

Example: x_1, \dots, x_n be i.i.d Bernouli(θ), $0 < \theta < 1$.

All of the following are sufficient statics for θ :

$$T_1(X) = \sum_{i=1}^n x_i, \quad T_2(X) = (x_{(1)}, x_{(2)}, \dots, x_{(n)}), \quad T_3(X) = (x_1, x_2, \dots, x_n)$$

Outline of Week 06 Lectures

- Introduction to Estimation Theory
- Sufficient Statistic
- Minimal Sufficient Statistic
- Complete Sufficient Statistic
- Likelihood Principle
- Frequentist's Estimators

Minimal Sufficient Statistic

Minimal sufficient statistic:

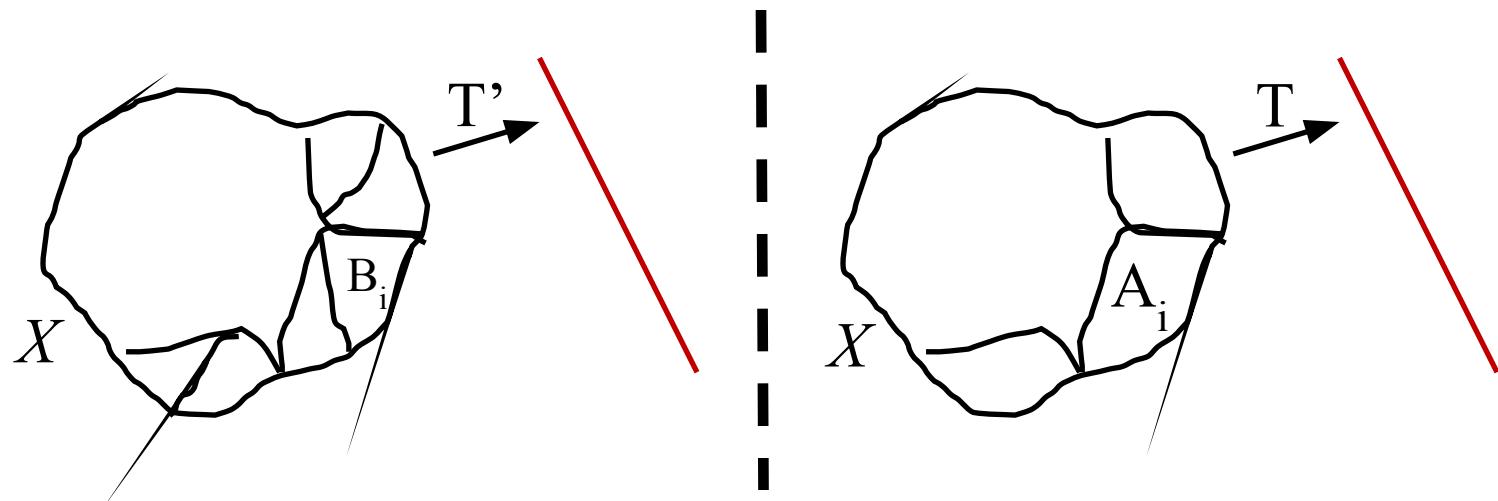
A sufficient statistic $T(X)$ is called minimal sufficient statistic, if for any other sufficient statistic $T'(X)$, $T(X)$ is a function of $T'(X)$.

It achieve maximum possible data reduction without losing info about θ .

T partitions χ into sets; $A_t = \{\underline{X} : T(\underline{X}) = t\}$

T' partitions χ into sets; $B_{t'} = \{\underline{X} : T'(\underline{X}) = t'\}$

Each set $B_{t'} \subset$ some set A_t



Minimal Sufficient Statistic

Theorem:

Let $f(x|\theta)$ be pdf or pmf. Suppose that for any 2 sample points \underline{X} and \underline{Y} the ratio:

$$\frac{f(\underline{X}|\theta)}{f(\underline{Y}|\theta)}$$

is constant as a function of θ iff $T(\underline{X}) = T(\underline{Y})$:

$T(\underline{X})$ is a minimal sufficient statistic for θ , iff the above holds.

Proof: assume $f(x|\theta) > 0$

Let $I = \{t: t = T(x) \text{ for some } x \in \chi\}$

$$A_t = \{\underline{X} : T(\underline{X}) = t\}$$

Minimal Sufficient Statistic

for each A_t , choose a fix element $X_t \in A_t$. For any \underline{X} , let $X_{T(x)}$ be the fixed element that is in the same A_t as \underline{X} , Hence:

$$T(\underline{X}) = T(X_{T(x)})$$

$\Rightarrow \frac{f(\underline{X}|\theta)}{f(X_{T(x)}|\theta)}$ is constant as a function of θ .

$$g(t|\theta) := f(X_{T(x)}|\theta)$$

$$f(x|\theta) = \frac{f(X_{T(x)}|\theta) f(\underline{x}|\theta)}{f(X_{T(x)}|\theta)} = g(T(x)|\theta) h(x)$$

$\Rightarrow T(x)$ is sufficient.

Minimal Sufficient Statistic

\Leftarrow Let T' be an arbitrary sufficient statistic. Then from factorization theorem:

\exists functions g', h' s.t. $f(x|\theta) = g'(T'(x)|\theta) h'(x)$

For any 2 sample points like $\underline{x}, \underline{y}$ with $T'(\underline{x}) = T'(\underline{y})$:

$$\frac{f(\underline{x}|\theta)}{f(\underline{y}|\theta)} = \frac{g'(T'(\underline{x})|\theta) h'(\underline{x})}{g'(T'(\underline{y})|\theta) h'(\underline{y})} = \frac{h'(\underline{x})}{h'(\underline{y})} \text{ which is a constant as a function of } \theta.$$

So by the assumption about $T(x)$ we have: $T(\underline{x}) = T(\underline{y})$.

Therefore, T is a function of T' .

Hence T is minimal.

Minimal Sufficient Statistic

Example: x_1, \dots, x_n be i.i.d Bernoulli(θ), $0 < \theta < 1$

$$f(x|\theta) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\sum x_i} (1-\theta)^{n-\sum x_i}$$

$$\Rightarrow \frac{f(x|\theta)}{f(y|\theta)} = \theta^{\sum x_i - \sum y_i} (1-\theta)^{\sum y_i - \sum x_i}$$

need: $\sum x_i - \sum y_i = 0$

So $T(X) = \sum_{i=1}^n x_i$ is minimal sufficient for θ .

Minimal Sufficient Statistic

Example: x_1, \dots, x_n be i.i.d $\text{Normal}(\mu, \delta^2)$.

$$f(x|\mu, \delta^2) = (2\pi\delta^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\delta^2}\right)$$

$$\frac{f(x|\mu, \delta^2)}{f(y|\mu, \delta^2)} = \exp\left(\frac{-n(\bar{x}^2 - \bar{y}^2) + 2n\mu(\bar{x} - \bar{y}) - (n-1)(\sum_{i=1}^n (x_i - \bar{x})^2 - \sum_{i=1}^n (y_i - \bar{y})^2)}{2\delta^2}\right)$$

Need:

$$\bar{x} = \bar{y}$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (y_i - \bar{y})^2$$

So $(\bar{x}, \sum_{i=1}^n (x_i - \bar{x})^2)$ is a minimal sufficient statistic for θ .

But it is not unique. E.g. (\bar{x}, s^2) is also a minimal sufficient statistic for θ .

Minimal Sufficient Statistic

Any 1-1 function of a minimal sufficient statistic is a minimal sufficient statistic.

Example: x_1, \dots, x_n be i.i.d $U(\theta, \theta + 1)$

$$f(x|\theta) = \begin{cases} 1 & \text{all } x_i \text{ in } (\theta, \theta + 1) \\ 0 & \text{o.w.} \end{cases} = \begin{cases} 1 & \max(x_i) - 1 < \theta < \min(x_i) \\ 0 & \text{o.w.} \end{cases}$$

$\frac{f(x|\theta)}{f(y|\theta)}$ is constant as a function of θ iff $\begin{cases} \max(x_i) = \max(y_i) \\ \min(x_i) = \min(y_i) \end{cases}$

Hence, $T(X) = (x_{(1)}, x_{(n)})$ is a minimal sufficient statistic for θ .

Note: $T'(x) = \left(x_{(n)} - x_{(1)}, \frac{x_{(1)} + x_{(n)}}{2}\right)$ is also minimal sufficient statistic.

Outline of Week 06 Lectures

- Introduction to Estimation Theory
- Sufficient Statistic
- Minimal Sufficient Statistic
- Complete Sufficient Statistic
- Likelihood Principle
- Frequentist's Estimators

Complete Sufficient Statistic

Def: let $f(t|\theta)$ be family of pdfs (pmfs) for a statistic $T(x)$, the family of probability distributions is called complete if:

$$E_{\theta} g(T) = 0 \quad \forall \theta$$

$$\Rightarrow p_{\theta}(g(T) = 0) = 1 \quad \forall \theta$$

or $T(x)$ is a **complete statistic**.

Note: completeness is a property of the family of distributions not a particular distribution.

Complete Sufficient Statistic

Example: Let X be a random sample of size n such that each X_i has the same Bernoulli distribution with parameter p . Let T be the number of 1s observed in the sample, i.e.

$$T = \sum_{i=1}^n X_i$$

T is a statistic of X which has a binomial distribution with parameters (n, p) . If the parameter space for p is $(0, 1)$, then T is a complete statistic:

$$\mathbb{E}_p(g(T)) = \sum_{t=0}^n g(t) \binom{n}{t} p^t (1-p)^{n-t} = (1-p)^n \sum_{t=0}^n g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t$$

neither p nor $1 - p$ can be 0.

Complete Sufficient Statistic

Hence: $E_p(g(T)) = 0$ iff:

$$\sum_{t=0}^n g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t = 0$$

Replacing $p/(1 - p)$ by r :

$$\sum_{t=0}^n g(t) \binom{n}{t} r^t = 0$$

The range of r is the positive reals. Also, $E(g(T))$ is a polynomial in r and, therefore, can only be identical to 0 if all coefficients are 0, that is, $g(t) = 0$ for all t .

Complete Sufficient Statistic

- It is important to notice that the result that all coefficients must be 0 was obtained because of the range of r .
- For example, for a single observation and a single parameter value; if $n = 1$ and the parameter space is $\{0.5\}$, T is not complete: $g(t) = 2(t - 0.5)$ and then, $E(g(T)) = 0$ although $g(t)$ is not 0 for $t = 0$ nor for $t = 1$.

Theorem: (exponential family)

Let x_1, \dots, x_n iid $F(x|\theta)$ $f(x|\theta) = h(x) c(\theta) \exp(\sum w_i(\theta)t_i(x))$

Suppose that the range of $(w_1(\theta), \dots, w_k(\theta))$ contains an n dimensional rectangle.

Then: $T(\underline{x}) = (\sum_{j=1}^n t_1(x_j), \dots, \sum_{j=1}^n t_k(x_j))$ is complete.

Outline of Week 06 Lectures

- Introduction to Estimation Theory
- Sufficient Statistic
- Minimal Sufficient Statistic
- Complete Sufficient Statistic
- Likelihood Principle
- Frequentist's Estimators

The Likelihood Principle

The likelihood principle:

Def: $\underline{X} \sim f(x|\theta)$

Then given $\underline{X} = \underline{x}$ observed, the function of θ defined by:

$$L(\theta|\underline{X}) = f(\underline{X}|\theta)$$

Is called the **likelihood function**.

Interpretation:

1) X discrete

$$L(\theta|\underline{X}) = p_\theta(\underline{X} = \underline{x})$$

$$L_1(\theta_1|\underline{X}) > L_2(\theta_2|\underline{X})$$

Sample had a higher likelihood of occurring if $\theta = \theta_1$ than $\theta = \theta_2$.

The Likelihood Principle

2) X continuous (real valued pdf)

for small ε :

$$2\varepsilon L(\theta | X) = 2\varepsilon f(X | \theta) \cong p_\theta(X - \varepsilon < X < X + \varepsilon)$$

$$\frac{L(\theta_1 | X)}{L(\theta_0 | X)} = \frac{p_{\theta_1}(X - \varepsilon < X < X + \varepsilon)}{p_{\theta_0}(X - \varepsilon < X < X + \varepsilon)} > 1 \ ?$$

approx. the same interpretation as discrete.

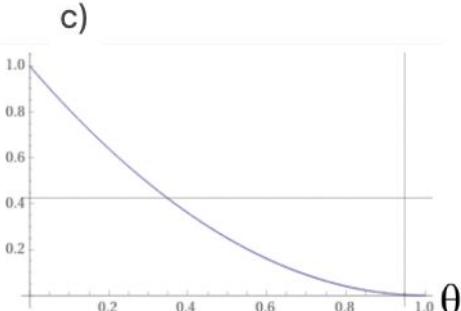
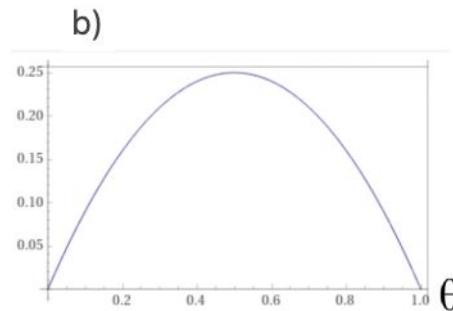
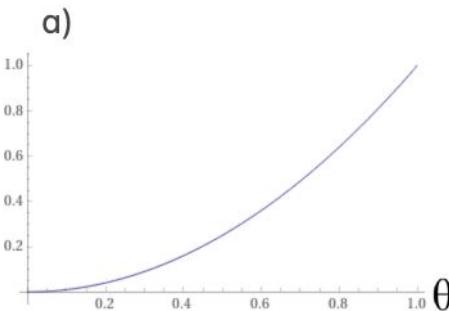
Example: x_1, \dots, x_n iid Bernoulli(θ)

$$L(\theta | x) = f(x | \theta) = \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i}$$

Let $n = 2$

The Likelihood Principle

- (a) $\sum x_i = 2 \Rightarrow L(\theta | x) = \theta^2$
- (b) $\sum x_i = 1 \Rightarrow L(\theta | x) = \theta(1 - \theta)$
- (c) $\sum x_i = 0 \Rightarrow L(\theta | x) = (1 - \theta)^2$



consider $L\left(\frac{3}{4} | x\right) / L\left(\frac{1}{4} | x\right)$

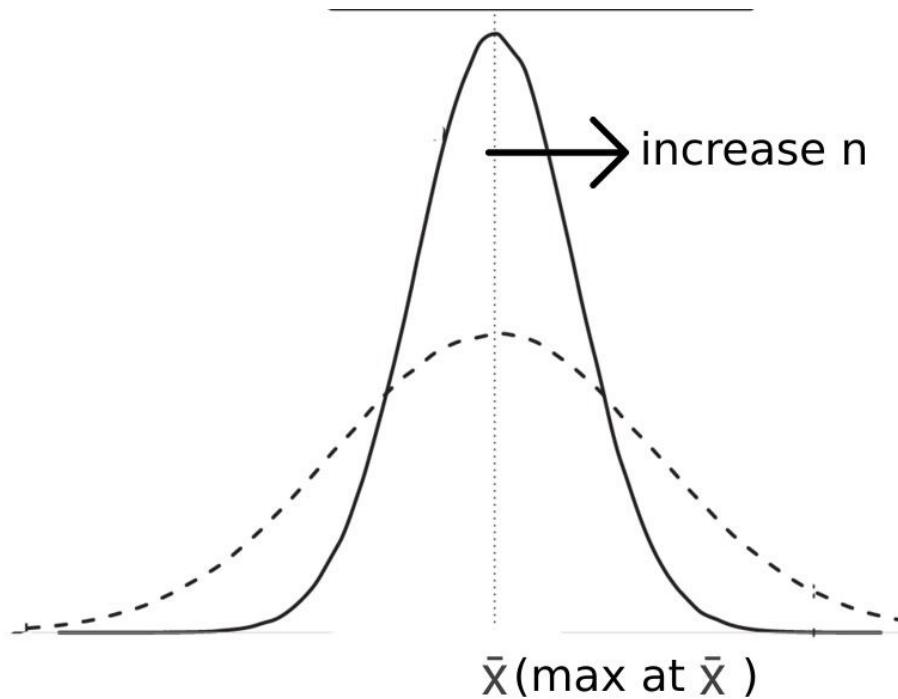
(a) $\frac{L(3/4|x)}{L(1/4|x)}$

$$\begin{cases} \frac{1}{9} & \text{when } \sum x_i = 0 \\ & \end{cases}$$

The Likelihood Principle

Example: x_1, \dots, x_n iid $N(\mu, \delta^2)$. Assume δ^2 is fixed.

$$L(\mu | x) = f(x | \mu) = (2\pi\delta^2)^{-n/2} e^{-\frac{1}{2\delta^2} [\sum_i (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2]}$$
$$= k(x) e^{-n(\bar{x} - \mu)^2 / 2\delta^2}$$



The Likelihood Principle

Likelihood principle:

If X and Y are two sample points (with the same parameter θ) s.t. $L(\theta|X)$ is proportional to $L(\theta|Y)$:

$$L(\theta|X) = C(X, Y) L(\theta|Y) \quad \forall \theta$$

Then the conclusions drawn from X and Y about θ should be identical.

Idea: use the likelihood function to compare the “probability” of various parameter values.

if $L(\theta_2|X) = 2L(\theta_1|X)$ θ_2 is twice as likely as θ_1 and:

Outline of Week 06 Lectures

- Introduction to Estimation Theory
- Sufficient Statistic
- Minimal Sufficient Statistic
- Complete Sufficient Statistic
- Likelihood Principle
- Frequentist's Estimators: MLE, MM

Frequentist's Estimators

Def: A point estimator is any statistic $T(x)$.

Estimator: function of samples (function of SS).

Estimate: actual value of the estimator.

Methods of finding estimators for this course:

- (1) Maximum Likelihood Estimator (MLE) ~ (frequentist)
- (2) Method of Moments (MM) ~ (frequentist)
- (3) UMVUE ~ (frequentist)
- (4) Maximum APosteriori (MAP) ~ (Bayes)
- (5) Bayes Minimum Risk ~ (Bayes)

Maximum Likelihood Estimator: MLE

Maximum likelihood estimator (MLE):

$$L(\theta|X) = L(\theta_1, \dots, \theta_k | X_1, \dots, X_n) = \prod_{i=1}^n f(X_i|\theta)$$

Def:

for each \underline{X} , let $\hat{\theta}(X)$ be the value which maximizes $L(\theta|X)$

then, $\hat{\theta}(X)$ is the maximum likelihood estimator (MLE) of θ .

$$\hat{\theta}_{ML}(X) = \text{Arg Max } (L(\theta|X))$$

Log likelihood:

use $\log L(\theta|X)$.

Maximum Likelihood Estimator: MLE

How to find MLE's:

(1) Differentiation

if $L(\theta|X)$ is differentiable in θ_i , possible θ_i 's are solutions to:

$$\frac{\partial}{\partial \theta_i} L(\theta|X) = 0 , \quad i = 1, \dots, k$$

a) 1-dimension

solve $\frac{\partial}{\partial \theta} L(\theta|X) = 0$ for $\hat{\theta}$

check $\frac{\partial^2}{\partial \theta^2} L(\theta|X) < 0$ for $\theta = \hat{\theta}$

(check boundaries)

Maximum Likelihood Estimator: MLE

Example: x_1, \dots, x_n iid $Bern(\theta)$

$$L(\theta | x) = \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i}$$

$$\log L(\theta | x) = \sum x_i \log \theta + (n - \sum x_i) \log(1 - \theta)$$

$$\frac{\partial \log L(\theta | x)}{\partial \theta} = \frac{\sum x_i}{\theta} - \frac{n - \sum x_i}{1 - \theta} = 0 \Rightarrow \hat{\theta} = \bar{x}$$

$$\frac{\partial^2 \log L}{\partial \theta^2} = -\frac{\sum x_i}{\theta^2} - \frac{n - \sum x_i}{(1 - \theta)^2} < 0 @ \theta = \hat{\theta}$$

check boundaries; $\sum x_i = 0, \sum x_i = n$

$$n \log(1 - \theta) \text{ if } \sum x_i = 0$$

$$\begin{aligned} \log L(\theta | x) = \\ n \log(\theta) \text{ if } \sum x_i = n \end{aligned}$$

Maximum Likelihood Estimator: MLE

b) 2-dimensions

solve $\frac{\partial}{\partial \theta_1} L(\theta_1, \theta_2 | X) = 0$

, $\frac{\partial}{\partial \theta_2} L(\theta_1, \theta_2 | X) = 0 \quad \text{for } \theta_1, \theta_2$

check that $\frac{\partial^2}{\partial \theta_1^2} L(\theta_1, \theta_2 | X) < 0 \quad \text{for } \theta_1 = \hat{\theta}_1, \theta_2 = \hat{\theta}_2$

or: $\frac{\partial^2}{\partial \theta_2^2} L(\theta_1, \theta_2 | X) < 0 \quad \text{for } \theta_1 = \hat{\theta}_1, \theta_2 = \hat{\theta}_2$

and: $\frac{\partial^2}{\partial \theta_1^2} L(\theta_1, \theta_2 | X) \frac{\partial^2}{\partial \theta_2^2} L(\theta_1, \theta_2 | X) - \left[\frac{\partial^2}{\partial \theta_1 \partial \theta_2} L(\theta_1, \theta_2 | X) \right]^2 > 0$

for $\theta_1 = \hat{\theta}_1, \theta_2 = \hat{\theta}_2$.

Maximum Likelihood Estimator: MLE

Example: x_1, \dots, x_n iid $N(\mu, \delta^2)$

$$\log L(\mu, \delta^2 | x) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log s^2 - \frac{1}{2\delta^2} \sum (x_i - \mu)^2$$

$$\frac{\partial}{\partial \mu} \log L = \frac{1}{\delta^2} \sum (x_i - \mu) = 0 \Rightarrow \hat{\mu} = \bar{x}$$

$$\frac{\partial}{\partial \delta^2} \log L = -\frac{n}{2\delta^3} + \frac{1}{2\delta^4} \sum (x_i - \mu)^2 = 0 \Rightarrow \hat{\delta}^2 = \sum (x_i - \bar{x})^2$$

$$(i) \frac{\partial^2}{\partial \mu^2} \log L = -\frac{n}{\delta^2}$$

$$(ii) \frac{\partial^2}{\partial (s^2)^2} \log L = \frac{n}{2\delta^4} - \frac{1}{\delta^6} \sum (x_i - \mu)^2$$

$$(iii) \frac{\partial^2}{\partial \mu \partial \delta^2} \log L = -\frac{1}{\delta^4} \sum (x_i - \mu)$$

Maximum Likelihood Estimator: MLE

$$\frac{1}{\delta^6} \left[-\frac{n^2}{2} + \frac{n}{\delta^2} \sum (x_i - \mu)^2 - \frac{1}{\delta^2} \left(\sum (x_i - \mu) \right)^2 \right] \Big| \begin{array}{l} \mu = \hat{\mu}_i \\ \delta^2 = \hat{\delta}^2 \end{array}$$

$$= \frac{1}{\hat{\delta}^6} \left[-\frac{n^2}{2} + \frac{n}{\hat{\delta}^2} n \hat{\delta}^2 - \frac{1}{\hat{\delta}^2} (0) \right] = \frac{n^2}{2 \hat{\delta}^2} > 0$$

Maximum Likelihood Estimator: MLE

How to find MLE's:

(2) Direct maximization

- find global upper bound on likelihood function
 - show bound is attained
-

Example: x_1, \dots, x_n iid $N(\mu, 1)$

$$L(\mu | \underline{x}) = \left(\frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{1}{2} \sum (x_i - \mu)^2}$$

Recall for any number a: $\sum (x_i - \bar{x})^2 \leq \sum (x_i - a)^2$
 $\Rightarrow L(\mu | \underline{x}) \leq L(\bar{x} | \underline{x}) \Rightarrow \hat{\mu} = \bar{x}$

Maximum Likelihood Estimator: MLE

(3) Numerically (by computer)

With or without (1) and (2)

Example: x_1, \dots, x_n iid truncated poisson:

$$p[x_i = r] = \frac{e^{-m} m^r}{(1 - e^{-m})r!}, m \leq 0, 1, \dots$$

$$L(m | x) = \prod_{i=1}^n \frac{e^{-m} m^{x_i}}{(1 - e^{-m})x_i!} = \left(\frac{e^{-m}}{1 - e^{-m}} \right)^n m^{\sum x_i} \prod_{i=1}^n \frac{1}{x_i!}$$

$$\log L = -mn - n \log(1 - e^{-m}) + \sum x_i \lg m - \sum \log(x_i!)$$

$$\frac{\partial \log L}{\partial m} s + n - \frac{ne^{-m}}{1 - e^{-m}} + \frac{\sum x_i}{m} = 0 \Rightarrow \hat{m} = ?$$

Define: $\phi(m) = \frac{\partial \log L}{\partial m}$, need \hat{m} s/t $\phi(\hat{m}) = 0$

Maximum Likelihood Estimator: MLE

Let m_0 be an initial estimate for \hat{m} .

$$0 \approx \phi(\hat{m}) \approx \phi(m_0) + (\hat{m} - m_0)\phi'(m_0)$$

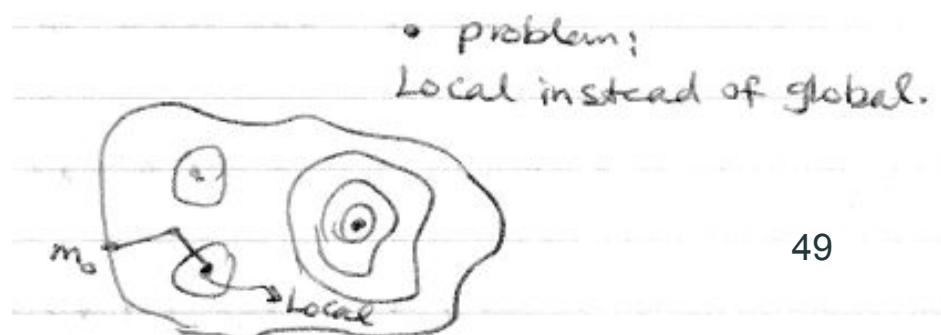
$$\hat{m} \approx m_0 - \frac{\phi(m_0)}{\phi'(m_0)}$$

- (1) Choose an initial estimate m_0
- (2) Define a sequence $\{m_k\}$ of estimates by:

$$m_{k+1} = m_k - \frac{\phi(m_k)}{\phi'(m_k)}, k = 0, 1, 2, \dots$$

- (3) Stop when $|\hat{m}_{k+1} - \hat{m}_k| < \varepsilon$

Let $\hat{m} = m_k$



Maximum Likelihood Estimator: MLE

Note: maximization takes place only over the range of parameter values.

Example: x_1, \dots, x_n iid $N(\mu, 1)$ but $\mu \geq 0$

$\hat{\mu} = \bar{x}$ what if $\bar{x} < 0$?

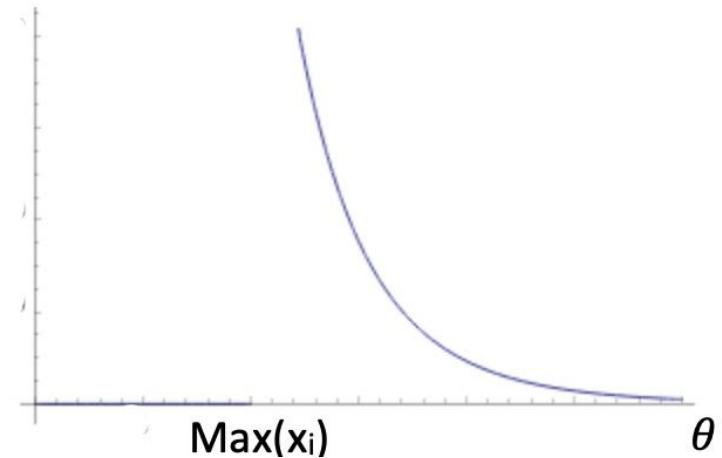
$$\hat{\mu} = 0 \text{ if } \bar{x} < 0 \Rightarrow \hat{\mu} = \begin{cases} \bar{x}, & \bar{x} \geq 0 \\ 0, & \bar{x} < 0 \end{cases}$$

Maximum Likelihood Estimator: MLE

Note: maximization can occur on boundaries.

Example: x_1, \dots, x_n iid $U(0, \theta)$

$$L(\theta | X) = \begin{cases} \frac{1}{\theta^n} & \text{if } \theta \geq \\ 0 & \end{cases}$$
$$\therefore \hat{\theta}_{mLE} = \max(:$$



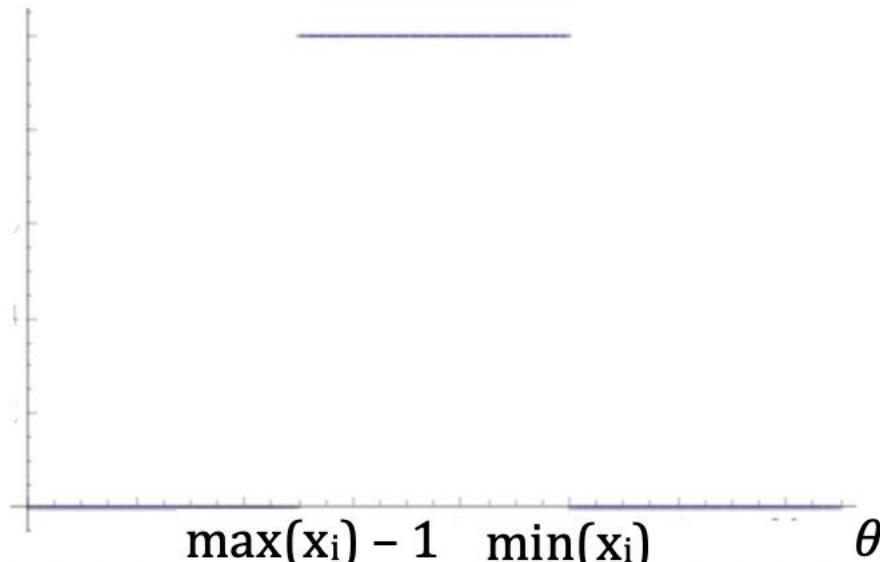
Note: maximum likelihood estimate may not be unique.

Maximum Likelihood Estimator: MLE

Note: maximum likelihood estimate may not be unique.

Example: x_1, \dots, x_n iid $U(\theta, \theta + 1)$

$$L(\theta | \underline{x}) = \begin{cases} 1 & \max x_i - 1 < \theta < \min x_i \\ 0 & 0, \omega. \end{cases}$$



$\therefore \hat{\theta} = \text{any value in the interval}$
 $(\max(x_i) - 1, \min(x_i))$

Maximum Likelihood Estimator: MLE

Note: MLE's can be numerically unstable.

Example: x_1, \dots, x_n iid $\text{Bin}(k, p)$; k, p unknowns

Can show:

$$\begin{aligned} &\text{if } \underline{x} = (16, 18, 22, 25, 27) \Rightarrow \hat{k} = 99 \\ &\text{if } \underline{\underline{x}} = (16, 18, 22, 25, 28) \Rightarrow \hat{k} = 190 \end{aligned}$$

Maximum Likelihood Estimator: MLE

Theorem: (invariance property)

If $\hat{\theta}$ is the MLE of θ , then for any function $r(\theta)$, $r(\hat{\theta})$ is the MLE of $r(\theta)$.

Example: x_1, \dots, x_n iid $N(\mu, 1)$

\bar{X} is the MLE of μ , then \bar{X}^2 is the MLE of μ^2 .

Method of Moments

Method of moments:

$$x_1, \dots, x_n \quad iid \quad f(x|\theta_1, \dots, \theta_k)$$

Equate the first k sample moments to the k first population moments.

$$\text{Let } m_1 = \frac{1}{n} \sum X_i \quad \mu_1 = E(X)$$

$$m_2 = \frac{1}{n} \sum X_i^2 \quad \mu_2 = E(X^2)$$

⋮

⋮

$$m_k = \frac{1}{n} \sum X_i^k \quad \mu_k = E(X^k)$$

$$m_j = \mu_j(\theta_1, \dots, \theta_k)$$

$$\text{Let } m_1 = \mu_1(\theta_1, \dots, \theta_k)$$

⋮

$$m_k = \mu_k(\theta_1, \dots, \theta_k) \quad \text{solve for } \theta_1, \dots, \theta_k$$

Method of moments

Example: x_1, \dots, x_n iid $N(\mu, \delta^2)$

$$m_1 = \frac{1}{n} \sum x_i \quad \mu_1 = \mu$$

$$m_2 = \frac{1}{n} \sum x_i^2 \quad \hat{\mu}_2 = \delta^2 + \mu^2$$

$$\bar{x} = \mu, \frac{1}{n} \sum x_i^2 = s^2 + \mu^2 \Rightarrow \hat{\mu} = \bar{x} + \hat{\delta}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Example: x_1, \dots, x_n iid binomial(k, p) both unknown

$$\bar{x} = kp$$

$$\frac{1}{n} \sum x_i^2 = kp(1-p) + k^2 p^2$$

Solving to get: $\hat{k} = \frac{\bar{x}^2}{\left[\bar{x} - \frac{1}{n} \sum (x_i - \bar{x})^2 \right]}$

$$\hat{p} = \frac{\bar{x}}{\hat{k}}$$

Method of moments

Note: this method can also be used for moment matching.

-match moments of distributions of statistics to obtain approximation to distributions.

Example: x_1, \dots, x_n iid $p(\lambda)$

$$(1) E(x_1) = \lambda$$

$$(2) E(x_1^2) = \lambda + \lambda^2$$

$$m_1 = \frac{1}{n} \sum x_i$$

$$m_2 = \frac{1}{n} \sum x_i^2$$

$$(1) \hat{\lambda} = \bar{x}$$

$$(2) \hat{\lambda}^2 + \hat{\lambda} - \frac{1}{n} \sum x_{\hat{\lambda}}^2 = 0 \Rightarrow \hat{\lambda} = -\frac{1}{2} + \left[\frac{1}{4} + \frac{1}{2} \sum x_i^2 \right]^{1/2}$$

$\hat{\lambda}$ is not unique, using method of moments.

Next Week:

**Estimation Theory 02:
UMVE & Bayes**

Have a good day!