MATH 450: Mathematical statistics

Oct 3rd, 2019

Lecture 12: Information

Topics

Week 2	*Chapter 6: Statistics and Sampling Distributions*
Week 4	Chapter 7: Point Estimation
Week 7	*Chapter 8: Confidence Intervals*
Week 10	*Chapter 9: Test of Hypothesis*
Week 11	Chapter 10: Two-sample inference
Week 13	Regression

Chapter 7: Overview

7.1 Point estimate
- unbiased estimator
- mean squared error

7.2 Methods of point estimation
- method of moments
- method of maximum likelihood.

7.3 Sufficient statistic

7.4 Information and Efficiency

Sufficient statistic

Some observations

- Basic estimation problem:
  - Given a density function $f(x,\theta)$ and a sample $X_1, X_2, \ldots, X_n$
  - Construct a statistic $\hat{\theta} = T(X_1, X_2, \ldots, X_n)$
  - Different methods lead to different estimates with different accuracies
- If, however, the distribution of $t(X_1, X_2, \ldots, X_n)$ does not depend on θ, then it is no good
- Similarly, if the conditional probability

$$P(X_1, X_2, \ldots, X_n | T)$$

  does not depend on θ, then this means that $T(X_1, X_2, \ldots, X_n)$ contained all the information to estimate θ

Sufficient statistic

Definition

A statistic $T = t(X_1, \ldots, X_n)$ is said to be sufficient for making inferences about a parameter θ if the joint distribution of $X_1, X_2, \ldots, X_n$ given that $T = t$ does not depend upon θ for every possible value t of the statistic T.

Fisher-Neyman factorization theorem

Theorem

T is sufficient for θ if and only if nonnegative functions g and h can be found such that

$$f(x_1, x_2, \ldots, x_n; \theta) = g(t(x_1, x_2, \ldots, x_n), \theta) \cdot h(x_1, x_2, \ldots, x_n)$$

i.e. the joint density can be factored into a product such that one factor, h does not depend on θ; and the other factor, which does depend on θ, depends on x only through $t(x)$.

Example 1

Problem

Let $X_1, X_2, ..., X_n$ be a random sample of from a Poisson distribution with parameter λ

$$f(x, \lambda) = \frac{1}{x!}e^{-\lambda x} \qquad x = 0, 1, 2, \ldots,$$

where λ is unknown.
Find a sufficient statistic of λ.

Example 2

Problem

Let $X_1, X_2, ..., X_n$ be a random sample of from a Poisson distribution with parameter λ

$$f(x) = \begin{cases} \frac{\beta}{x^{\beta+1}} & \textit{if } x > 1 \\ 0 & \textit{otherwise} \end{cases}$$

where β is unknown.
Find a sufficient statistic of β.

Jointly sufficient statistic

Definition

The m statistics $T_1 = t_1(X_1, \dots, X_n)$, $T_2 = t_2(X_1, \dots, X_n)$, ..., $T_m = t_m(X_1, \dots, X_n)$ are said to be jointly sufficient for the parameters $\theta_1, \theta_2, \dots, \theta_k$ if the joint distribution of $X_1, \dots, X_n$ given that

$$T_1 = t_1, T_2 = t = 2, \dots, T_m = t_m$$

does not depend upon $\theta_1, \theta_2, \dots, \theta_k$ for every possible value $t_1, t_2, \dots, t_m$ of the statistics.

Fisher-Neyman factorization theorem

Theorem

$T_1, T_2, \ldots, T_m$ are sufficient for $\theta_1, \theta_2, \ldots, \theta_k$ if and only if nonnegative functions g and h can be found such that

$$f(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_k) = g(t_1, t_2, \ldots, t_m, \theta_1, \theta_2, \ldots, \theta_k) \cdot h(x_1, x_2, \ldots, x_n)$$

Example 3

- Let $X_1, X_2, ..., X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Prove that

$$T_1 = X_1 + \ldots + X_n, \qquad T_2 = X_1^2 + X_2^2 + \ldots + X_n^2$$

are jointly sufficient for the two parameters μ and σ.

Example 4

- Let $X_1, X_2, ..., X_n$ be a random sample from a Gamma distribution

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

where α, β is unknown.

- Prove that

$$T_1 = X_1 + \ldots + X_n, \qquad T_2 = \prod_{i=1}^{n} X_i$$

are jointly sufficient for the two parameters α and β.

Information

Fisher information

Definition

The Fisher information $I(\theta)$ in a single observation from a pmf or pdf $f(x;\theta)$ is the variance of the random variable $U = \frac{\partial \log f(X,\theta)}{\partial \theta}$, which is

$$I(\theta) = Var\left[\frac{\partial \log f(X,\theta)}{\partial \theta}\right]$$

Note: We always have $E[U] = 0$

Fisher information

We have

$$\sum_x f(x,\theta) = 1 \quad \forall\theta$$

Thus

$$\begin{aligned} E[U] &= E\left[\frac{\partial \log f(X,\theta)}{\partial\theta}\right] \\ &= \sum_x \frac{\partial \log f(x,\theta)}{\partial\theta} f(x,\theta) \\ &= \sum_x \frac{\partial f(x,\theta)}{\partial\theta} = 0 \end{aligned}$$

Example

Problem

Let X be distributed by

x	0	1
$f(x,\theta)$	$1-\theta$	θ

Compute $I(X,\theta)$.

Hint:

- If $x = 1$, then $f(x,\theta) = \theta$. Thus

$$u(x) = \frac{\partial \log f(x,\theta)}{\partial \theta} = \frac{1}{\theta}$$

- How about $x = 0$?

Example

Problem

Let X be distributed by

x	0	1
$f(x,\theta)$	$1-\theta$	θ

Compute $I(X,\theta)$.

We have

$$\begin{aligned} Var[U] &= E[U^2] - (E[U])^2 = E[U^2] \\ &= \sum_{x=0,1} U^2(x) f(x,\theta) \\ &= \frac{1}{(1-\theta)^2} \cdot (1-\theta) + \frac{1}{\theta^2} \cdot \theta \end{aligned}$$

The Cramer-Rao Inequality

Theorem

Assume a random sample $X_1, X_2, ..., X_n$ from the distribution with pmf or pdf $f(x, \theta)$ such that the set of possible values does not depend on θ. If the statistic $T = t(X_1, X_2, ..., X_n)$ is an unbiased estimator for the parameter θ, then

$$Var(T) \geq \frac{1}{n \cdot I(\theta)}$$

Proof for $n = 1$

Recall that $E[U] = 0$ and $E[T] = \theta$ (since T is an unbiased estimator of θ) we have

$$\begin{aligned} Cov(T, U) &= E[TU] - E[U] \cdot E[T] \\ &= \sum_x t(x) \frac{\partial \log f(x, \theta)}{\partial \theta} f(x, \theta) \\ &= \sum_x t(x) \frac{\partial f(x, \theta)}{\partial \theta} \frac{1}{f(x, \theta)} f(x, \theta) \\ &= \frac{\partial}{\partial \theta} \left(\sum_x t(x) f(x, \theta) \right) = 1 \end{aligned}$$

Proof for $n = 1$

The Cauchy–Schwarz inequality shows that

$$Cov(T, U) \leq \sqrt{Var(T) \cdot Var(U)}$$

which implies

$$Var(T) \geq \frac{1}{I(\theta)}.$$

Heisenberg's Uncertainty Principle

uncertainty in momentum

$$\Delta x \Delta p \geq \frac{h}{4\pi} = \frac{\hbar}{2}$$

uncertainty in position

The more accurately you know the position (i.e., the smaller Δx is), the less accurately you know the momentum (i.e., the larger Δp is); and vice versa

Efficiency

Theorem

Let $T = t(X_1, X_2, ..., X_n)$ is an unbiased estimator for the parameter θ, the ratio of the lower bound to the variance of T is its efficiency

$$\mathit{Efficiency} = \frac{1}{nI(\theta)V(T)} \leq 1$$

T is said to be an efficient estimator if T achieves the Cramer–Rao lower bound (i.e., the efficiency is 1).

Note: An efficient estimator is a minimum variance unbiased (MVUE) estimator.

Large Sample Properties of the MLE

Theorem

Given a random sample $X_1, X_2, ..., X_n$ from the distribution with pmf or pdf $f(x, \theta)$ such that the set of possible values does not depend on θ. Then for large n the maximum likelihood estimator $\hat{\theta}$ has approximately a normal distribution with mean θ and variance $\frac{1}{n \cdot I(\theta)}$.
More precisely, the limiting distribution of $\sqrt{n}(\hat{\theta} - \theta)$ is normal with mean 0 and variance $1/I(\theta)$.