Mathematical statistics

October 4^{th}, 2018

Lecture 12: Information

Where are we?

Week 1	Probability reviews
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 14	Regression

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**
- Large sample properties of the maximum likelihood estimator

Sufficient statistic

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 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

- 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.

Jointly sufficient statistic

Definition

The m statistics $T_1 = t_1(X_1, \ldots, X_n)$, $T_2 = t_2(X_1, \ldots, X_n)$, ..., $T_m = t_m(X_1, \ldots, X_n)$ are said to be jointly sufficient for the parameters $\theta_1, \theta_2, \ldots, \theta_k$ if the joint distribution of $X_1, X_2, \ldots, X_n$ given that

$$T_1 = t_1, T_2 = t_2, \ldots, T_m = t_m$$

does not depend upon $\theta_1, \theta_2, \ldots, \theta_k$ for every possible value $t_1, t_2, \ldots, t_m$ 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

Problem

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

Problem

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

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

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*

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

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

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

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

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

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 $1/I(\theta)$.

The Central Limit Theorem

Theorem

Let $X_1, X_2, \dots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then, in the limit when $n \to \infty$, the standardized version of $\bar{X}$ have the standard normal distribution

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z\right) = \mathbb{P}[Z \le z] = \Phi(z)$$

Chapter 7: Summary

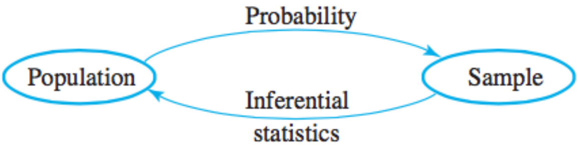
Overview

7.1 Point estimate
- unbiased estimator
- mean squared error
- bootstrap

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

7.3 Sufficient statistic

7.4 Information and Efficiency
- Large sample properties of the maximum likelihood estimator



Definition

A point estimate $\hat{\theta}$ of a parameter θ is a single number that can be regarded as a sensible value for θ.

$$\begin{array}{ccccc} \text{population parameter} & \Longrightarrow & \textit{sample} & \Longrightarrow & \textit{estimate} \\ \theta & \Longrightarrow & X_1, X_2, \ldots, X_n & \Longrightarrow & \hat{\theta} \end{array}$$

Mean Squared Error

- Measuring error of estimation

$$|\hat{\theta} - \theta| \qquad \text{or} \qquad (\hat{\theta} - \theta)^2$$

- The error of estimation is random

Definition

The mean squared error of an estimator $\hat{\theta}$ is

$$E[(\hat{\theta} - \theta)^2]$$

Bias-variance decomposition

Theorem

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + \left(E(\hat{\theta}) - \theta\right)^2$$

Bias-variance decomposition

Mean squared error = variance of estimator + $(bias)^2$

Unbiased estimators

Definition

A point estimator $\hat{\theta}$ is said to be an unbiased estimator of θ if

$$E(\hat{\theta}) = \theta$$

for every possible value of θ.

Unbiased estimator

$\Leftrightarrow$ Bias $= 0$

$\Leftrightarrow$ Mean squared error $=$ variance of estimator

Example 1

Problem

Consider a random sample $X_1, \ldots, X_n$ from the pdf

$$f(x) = \frac{1 + \theta x}{2} \qquad -1 \leq x \leq 1$$

Show that $\hat{\theta} = 3\bar{X}$ is an unbiased estimator of θ.

Overview

7.1 Point estimate

- unbiased estimator
- mean squared error
- bootstrap

7.2 Methods of point estimation

- **method of moments**
- **method of maximum likelihood.**

7.3 Sufficient statistic

7.4 Information and Efficiency

- Large sample properties of the maximum likelihood estimator

Method of moments: ideas

- Let $X_1, \dots, X_n$ be a random sample from a distribution with pmf or pdf
$$f(x; \theta_1, \theta_2, \dots, \theta_m)$$
- Assume that for $k = 1, \dots, m$
$$\frac{X_1^k + X_2^k + \dots + X_n^k}{n} = E(X^k)$$
- Solve the system of equations for $\theta_1, \theta_2, \dots, \theta_m$

Method of moments: Example 4

Problem

Suppose that for a parameter $0 \leq \theta \leq 1$, X is the outcome of the roll of a four-sided tetrahedral die

x	1	2	3	4
$p(x)$	$\frac{3\theta}{4}$	$\frac{\theta}{4}$	$\frac{3(1-\theta)}{4}$	$\frac{(1-\theta)}{4}$

Suppose the die is rolled 10 times with outcomes

$$4, 1, 2, 3, 1, 2, 3, 4, 2, 3$$

Use the method of moments to obtain an estimator of θ.

Maximum likelihood estimator

- Let $X_1, X_2, ..., X_n$ have joint pmf or pdf

$$f_{joint}(x_1, x_2, \ldots, x_n; \theta)$$

where θ is unknown.

- When $x_1, \ldots, x_n$ are the observed sample values and this expression is regarded as a function of θ, it is called the likelihood function.
- The maximum likelihood estimates θ_{ML} are the value for θ that maximize the likelihood function:

$$f_{joint}(x_1, x_2, \ldots, x_n; \theta_{ML}) \geq f_{joint}(x_1, x_2, \ldots, x_n; \theta) \qquad \forall \theta$$

How to find the MLE?

- Step 1: Write down the likelihood function.
- Step 2: Can you find the maximum of this function?
- Step 3: Try taking the logarithm of this function.
- Step 4: Find the maximum of this new function.

To find the maximum of a function of θ:

- compute the derivative of the function with respect to θ
- set this expression of the derivative to 0
- solve the equation

Example 3

- Let $X_1, \ldots, X_{10}$ be a random sample of size $n = 10$ from a distribution with pdf

$$f(x) = \begin{cases} (\theta + 1)x^{\theta} & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- The observed x_i's are

$$0.92, 0.79, 0.90, 0.65, 0.86, 0.47, 0.73, 0.97, 0.94, 0.77$$

- Question: Use the method of maximum likelihood to obtain an estimator of θ.

Overview

7.1 Point estimate
- unbiased estimator
- mean squared error
- bootstrap

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

7.3 **Sufficient statistic**

7.4 **Information and Efficiency**
- Large sample properties of the maximum likelihood estimator

Fisher-Neyman factorization theorem

Theorem

T is sufficient for 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 $t(x)$.

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 $E[U] = 0$.

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

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

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 $1/I(\theta)$.