Lesson 9

Statistics II

Todays topics

- **maximum likelihood (最尤推定)**

来嶋 秀治 (Shuji Kijima)
Dept. Informatics,
Graduate School of ISEE

Remind: Probability III

i.i.d. (multivariate distribution)

Distribution of random variables X and Y of $(\Omega, \mathcal{F}, P)$.

Ex1. two dice.

$\Omega = \{(1,1),(1,2),\ldots,(6,5),(6,6)\}$

X = sum of casts

Y = product of casts

例2. poker

choose five cards,

X = # of A's

Y = # of spades

Prop.

$$\Pr[(X_1 = x_1)\&(X_2 = x_2)\&\cdots\&(X_n = x_n)] = \Pr[X_1 = x_1]\Pr[X_2 = x_2]\cdots\Pr[X_n = x_n]$$

Prop.

Suppose ***discrete*** r.v. $X_1, X_2, \ldots X_n$ are i.i.d. w/pmf f.

Then $f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n)$.

Prop.

Suppose ***continuous*** r.v. $X_1, X_2, \ldots X_n$ are i.i.d. w/pdf f.

Then $f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n)$.

Prop.

$\Pr[(X \le x)\&(Y \le y)] = \Pr[X \le x]\Pr[Y \le y]$
i.e., $F_{XY}(x,y) = F_X(x)F_Y(y)$

Prop.

Suppose ***continuous*** r.v. X and Y are **independent**.

Then, $f_{XY}(x,y) = f_X(x)f_Y(y)$.

Prop.

$\Pr[(X \le x)\&(Y \le y)] = \Pr[X \le x]\Pr[Y \le y]$
i.e., $F_{XY}(x,y) = F_X(x)F_Y(y)$

Prop.

Suppose ***continuous*** r.v. X and Y are **independent**.

Then, $f_{XY}(x,y) = f_X(x)f_Y(y)$.

Proof.

$$f(x,y) \coloneqq \frac{\partial^2}{\partial x \partial y} F_{XY}(x,y)$$

$$= \frac{\partial^2}{\partial x \partial y}\left(F_X(x)F_Y(y)\right)$$

$$= \frac{\partial}{\partial x}\left(\left(\frac{\partial}{\partial y}F_X(x)\right)F_Y(y)\right) + \frac{\partial}{\partial x}\left(F_X(x)\left(\frac{\partial}{\partial y}F_Y(y)\right)\right)$$

$$= 0 + \frac{\partial}{\partial x}\left(F_X(x)f_Y(y)\right)$$

$$= \frac{\partial}{\partial x}F_X(x)f_Y(y) + F_X(x)\frac{\partial}{\partial x}f_Y(y)$$

$$= f_X(x)f_Y(y)$$

Joint distribution of i.i.d. r.v.

Prop.

Suppose $X_1, \ldots, X_n$ are i.i.d. with density function $f(x)$.

Then, the pdf of the joint distribution of $(X_1, \ldots, X_n)$ is

$$f(x_1, .., x_n) := f(x_1) \cdots f(x_n).$$

Estimating the population parameters

母パラメータの推定

Maximum likelihood (最尤推定)

maximum likelihood

Suppose samples $x_1, \ldots, x_n$ are i.i.d. with a prob. func. $f(x; \theta)$ where θ is an unknown parameter.

Find θ to maximize $\Pr[\boldsymbol{X} = \boldsymbol{x} \mid \theta]$.

Statistical inference: maximum likelihood

Example 1

The number of defective products per 10,000 products.

lot	1	2	3	4	5	6	7	8	9	10
#defective	0	2	0	0	1	1	0	3	1	0

How often do detectives appear?

Let X be a r.v. denoting #defectives,

Then $X \sim \mathrm{Po}(\lambda)$, i.e.,

$$\Pr[X = x] =: f(x) = \mathrm{e}^{-\lambda} \frac{\lambda^x}{x!}$$

for unknown parameter λ.

Maximum likelihood

θ: parameters
e.g. N(μ,σ), E(λ)

Preparation

Let $X_1, \dots, X_n$ be i.i.d. with density function $f(x;\theta)$, and

let $f(x_1,..,x_n;\theta) \coloneqq f(x_1;\theta)\cdots f(x_n;\theta)$,

i.e., density function of a joint distribution of $(X_1, \dots, X_n)$.

remark
$X_1,\dots,X_n$ are independent

Maximum likelihood estimation

Given sample values $X_1 = a_1,\ \dots,\ X_n = a_n$,

let $L(\theta \mid \boldsymbol{a}) \coloneqq f(a_1, \dots, a_n;\theta)$, called likelihood function, and

$\theta^* = \arg\max_{\theta} L(\theta)$ is called maximum likelihood estimator.

Ex. Maximum likelihood

Poisson distribution $\mathrm{Po}(\lambda)$

$$f(x) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}$$

Ex. Poisson distribution

Let $X_1, \ldots, X_n$ be independent r.v.s according to $\mathrm{Po}(\lambda)$, and $a_1, \ldots, a_n$ are sample values.

$$L(\lambda; a_1, \ldots, a_n) = \mathrm{e}^{-\lambda}\frac{\lambda^{a_1}}{a_1!}\mathrm{e}^{-\lambda}\frac{\lambda^{a_2}}{a_2!}\cdots\mathrm{e}^{-\lambda}\frac{\lambda^{a_n}}{a_n!}$$

max. likelihood estimator of λ = argmax_λ $L(\lambda)$

$$\frac{\partial}{\partial\lambda}L(\lambda; a_1, \ldots, a_n) = \cdots$$

Ex. Maximum likelihood

Poisson distribution $\mathrm{Po}(\lambda)$

$$f(x) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}$$

Ex. Poisson distribution

Let $X_1, \dots, X_n$ be independent r.v.s according to $\mathrm{Po}(\lambda)$, and $a_1, \dots, a_n$ are sample values.

$$\log\big(L(\lambda)\big) = \sum_{i=1}^{n} \log\big(f(a_i;\lambda)\big) = \sum_{i=1}^{n} (-\lambda + a_i \log\lambda - \log(a_i!))$$

$$= n(\bar{a}\log\lambda - \lambda) - \sum_{i=1}^{n} \log(a_i!)$$

$$\frac{\partial}{\partial\lambda}\log\big(L(\lambda)\big) = n\left(\bar{a}\frac{1}{\lambda} - 1\right)$$

$\bar{a}$ is the maximum likelihood estimator of λ

Statistical inference: maximum likelihood

Example 2

The scores of examnation.

student	1	2	3	4	5	6	7	8	9	10
score	72	89	64	52	96	64	70	83	56	70

How much ratio do they understand?

Let X be a r.v. denoting scores,

Then $X \sim \mathrm{N}(\mu, \sigma^2)$, i.e.,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

for unknown parameters μ and σ.

Ex. Maximum likelihood

Ex. Normal distribution

Let $X_1, \ldots, X_n$ be independent r.v.s according to $\mathrm{N}(\mu, \sigma^2)$, and $a_1, \ldots, a_n$ are sample values.

$$\ln\big(L(\mu,\sigma;a)\big) = \sum_{i=1}^{n} \ln\big(f(a_i;\mu,\sigma)\big)$$

$$= \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(a_i-\mu)^2}{2\sigma^2}\right)\right)$$

$$= \sum_{i=1}^{n}\left(-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{(a_i-\mu)^2}{2\sigma^2}\right)$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \sum_{i=1}^{n}\frac{(a_i-\mu)^2}{2\sigma^2}$$

Ex. Maximum likelihood

Ex. Normal distribution

Let $X_1, \ldots, X_n$ be independent r.v.s according to $\mathrm{N}(\mu, \sigma^2)$, and $a_1, \ldots, a_n$ are sample values.

$$\frac{\partial}{\partial \mu} \ln\big(L(\mu, \sigma; a)\big) = -\frac{\partial}{\partial \mu} \frac{\sum_{i=1}^{n} (a_i - \mu)^2}{2\sigma^2} = -\frac{\sum_{i=1}^{n} (a_i - \mu)}{\sigma^2}$$

$\bar{a}$ is the maximum likelihood estimator of μ

$$\frac{\partial}{\partial \sigma} \ln\big(L(\mu, \sigma; a)\big)$$

$$= -\frac{\partial}{\partial \sigma} \frac{n}{2} \ln \sigma^2 - \frac{\partial}{\partial \sigma} \frac{\sum_{i=1}^{n} (a_i - \mu)^2}{2\sigma^2}$$

$$= -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n} (a_i - \mu)^2}{\sigma^3}$$

Since $\mu = \bar{a}$ maximize $L(\mu, \sigma)$ (independent of σ),

$$(\sigma^*)^2 = \frac{\sum (a_i - \bar{a})^2}{n}$$

Ex. maximum likelihood

Example 2 (Hardy Weinberg)

Blood type	A	B	AB	O
samples	43	12	6	39
possible type	aa, ao	bb,bo	ab	oo

Estimate the ratio $\theta_a, \theta_b, \theta_o$ of a, b, o

Model

Hardy Weinberg equilibrium

$$p_A = \theta_a^2 + 2\theta_a\theta_o$$
$$p_B = \theta_b^2 + 2\theta_b\theta_o$$
$$p_{AB} = 2\theta_a\theta_b$$
$$p_O = \theta_o^2$$

where

$$p_A + p_B + p_{AB} + p_O = 1$$
$$\theta_a + \theta_b + \theta_o = 1$$

藤澤洋徳: 確率と統計より

Ex. maximum likelihood

Example 2 (Hardy Weinberg)

Blood type	A	B	AB	O
samples	43	12	6	39
possible type	aa, ao	bb,bo	ab	oo

Estimate the ratio $\theta_a, \theta_b, \theta_o$ of a, b, o

$$\log L(\theta)$$
$$= \log\left(\frac{n!}{x_A!\,x_B!\,x_{AB}!\,x_O!}\,p_A^{x_A}p_B^{x_B}p_{AB}^{x_{AB}}p_O^{x_O}\right)$$
$$= x_A \log p_A + x_B \log p_B + x_{AB} \log p_{AB} + x_O \log p_O + \log\left(\frac{n!}{x_A!\,x_B!\,x_{AB}!\,x_O!}\right)$$
$$= x_A \log(\theta_a^2 + 2\theta_a\theta_o) + x_B \log\left(\theta_b^2 + 2\theta_b\theta_o\right) + x_{AB} \log(2\theta_a\theta_b) + x_O \log(\theta_o^2)$$
$$+ \log\left(\frac{n!}{x_A!\,x_B!\,x_{AB}!\,x_O!}\right)$$

藤澤洋徳: 確率と統計より

Ex. maximum likelihood

Example 2 (Hardy Weinberg)

Blood type	A	B	AB	O
samples	43	12	6	39
possible type	aa, ao	bb,bo	ab	oo

Estimate the ratio $\theta_a, \theta_b, \theta_o$ of a, b, o

solution: $\widehat{\theta_a} = 0.285$, $\widehat{\theta_a} = 0.094$, $\widehat{\theta_a} = 0.621$

Blood type	A	B	AB	O
samples	43	12	6	39
	$\theta_a^2 + 2\theta_a\theta_o$	$\theta_b^2 + 2\theta_b\theta_o$	$2\theta_a\theta_b$	θ_o^2
estimation	43.5	12.6	5.4	38.5

藤澤洋徳: 確率と統計より

Topics Related to max. likelihood

Todays topics

- **KL divergence**
- **Consistency of max. likelihood**
- **Fisher Information (Fisher情報量)**
  - sharpness of max. likelihood
- **Cramer-Rao inequality**
  - a kind of central limit theorem

Kullback-Leibler divergence

Def. (KL divergence)

Let f, g be density function.

$$\mathrm{KL}(g, f) = \mathrm{E}_g\left[\log\left(\frac{g(X)}{f(X)}\right)\right] = \int g(x)\log\left(\frac{g(x)}{f(x)}\right)\mathrm{d}x$$

Prop.

- $\mathrm{KL}(g, f) \geq 0$
- $\mathrm{KL}(g, f) = 0$ iff $g \equiv f$

note that

- ✓ $\mathrm{KL}(g, f) \neq \mathrm{KL}(f, g)$, in general
- ✓ *triangle inequality* does not hold, in general

  i.e. $\mathrm{KL}(g, f) + \mathrm{KL}(f, h) > \mathrm{KL}(g, h)$ may hold.

Consistency of Maximum likelihood

Thm. (consistency of maximum likelihood estimator)

If f is strictly positive (i.e., $f(x) > 0$ for any x) and continuous, then the maximum likelihood estimator θ^* of θ_0 is *consistent*.

i.e., $\forall \varepsilon > 0, \lim_{n\to\infty} \Pr[|\theta^* - \theta_0| < \varepsilon] = 1$

Rem.

$\mathrm{KL}(g, f) = \mathrm{E}_g[\log g(X)] - \mathrm{E}_g[\log f(X)]$

$$\begin{aligned}\theta^* &= \arg\min_\theta \mathrm{KL}(f_{\theta^*}, f_\theta) \\ &= \arg\min_\theta \left(\mathrm{E}_{f_{\theta^*}}[\log f_{\theta^*}(X)] - \mathrm{E}_{f_{\theta^*}}[\log f_\theta(X)]\right) \\ &= \arg\max_\theta \mathrm{E}_{f_{\theta^*}}[\log f_\theta(X)]\end{aligned}$$

From the law of large number, intuitively

$$\arg\max_\theta \frac{1}{n}\sum_i^n \log f_\theta(x_i) \to \arg\max_\theta E_{f_{\theta^*}}[\log f_\theta(X)]$$

Fisher Information

Proposition.

For a density function $f(x;\theta)$,

$$\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]$$

Fisher Information

Proposition.

For a density function $f(x;\theta)$,

$$\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta\right.$$

$$\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right]$$

$$= \mathrm{E}\left[\frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)}\right] - \mathrm{E}\left[\frac{\left(\frac{\partial}{\partial\theta}f(x;\theta)\right)^2}{\left(f(x;\theta)\right)^2}\right]$$

$$= \int_{-\infty}^{\infty}\frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)}f(x;\theta)\mathrm{d}x - \mathrm{E}\left[\frac{\left(\frac{\partial}{\partial\theta}f(x;\theta)\right)^2}{\left(f(x;\theta)\right)^2}\right]$$

$$= \int_{-\infty}^{\infty}\frac{\partial^2}{\partial\theta^2}f(x;\theta)\mathrm{d}x - \mathrm{E}\left[\frac{\left(\frac{\partial}{\partial\theta}f(x;\theta)\right)^2}{\left(f(x;\theta)\right)^2}\right]$$

$$= -\mathrm{E}\left[\frac{\left(\frac{\partial}{\partial\theta}f(x;\theta)\right)^2}{\left(f(x;\theta)\right)^2}\right]$$

$$\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)$$

$$= \frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)$$

$$= \frac{\partial}{\partial\theta}\left(\frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}\right)$$

$$= \frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)} - \frac{\left(\frac{\partial}{\partial\theta}f(x;\theta)\right)^2}{\left(f(x;\theta)\right)^2}$$

$$\int_{-\infty}^{\infty} f(x;\theta)\mathrm{d}x = 1$$

$$\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x;\theta)\mathrm{d}x = 0$$

$$\int_{-\infty}^{\infty}\frac{\partial^2}{\partial\theta^2}f(x;\theta)\mathrm{d}x = 0$$

Fisher Information

Proposition.

For a density function $f(x;\theta)$,

$$\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]$$

Def.

For a density function $f(x;\theta)$,

the *Fisher Information* $I(\theta)$ is defined by

$$I(\theta) \coloneqq \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]$$

A type of central limit theorem

Thm.

Let f(x;θ) be positive and twice differentiable, and let θ^* be a max likelihood estimator of θ_o.

Then $\sqrt{n}(\theta^* - \theta_o) \sim N\left(0, \frac{1}{I(\theta)}\right)$ when n→+∞.

… meaning that

$$\mathrm{E}[\theta^* | \theta_o] \simeq \theta_o$$

$$\mathrm{Var}[\theta^* | \theta_o] \simeq \frac{1}{nI(\theta)}$$

for n→ +∞.

A type of central limit theorem

Thm. (Cramer-Rao's bound)

Let $X_1,\ldots,X_n$ be i.i.d. with density function $f(x;\theta)$, where $f(x;\theta)$ is positive and twice differentiable, and let $I(\theta)$ be the Fisher Information of $f(x;\theta)$.

If $\hat{\theta}$ is an unbiased estimator, then

$$\mathrm{Var}\left[\hat{\theta}\right] \geq \frac{1}{nI(\theta)}$$