• method of finding estimators II --- Maximum Likelihood Estimator (MLE)

Questions:

parameter

• If assign a distribution on parameter space ⇒ Bayesian approach
• If not (i.e., θ fixed & unknown) ⇒ Frequentist approach

- Toss a coin 10 times. Let θ be the probability of getting a head. Suppose that we know

$$\theta \in \{0.1, 0.5, 0.9\}. \leftarrow \text{parameter space}$$

- When we get 7 heads out of the 10 tosses, which θ is more plausible to generate the output?

Q: Why sum $\neq 1$?
Hint. total probability $= 1$

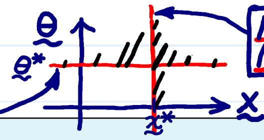
Hint.

$X = \#$ of heads
$X \sim B(10, \theta)$

$$P(7 \text{ heads}|\theta = 0.1) \approx 0.000,$$
$$P(7 \text{ heads}|\theta = 0.5) \approx 0.117,$$
$$P(7 \text{ heads}|\theta = 0.9) \approx 0.057.$$

distribution (pmf/pdf)

like-lihood



Definition 6.8 (likelihood, log likelihood, TBp. 267, 268)

Suppose random variables $X_1, \ldots, X_n$ have a joint pdf or pmf

varying fixed

$$f(x_1, \ldots, x_n|\Theta).$$

not prob., but proportional to prob.

$\sum_{\underset{\sim}{x}} f(\underset{\sim}{x}|\underset{\sim}{\theta}) = 1 = \int_{\underset{\sim}{x}} f(\underset{\sim}{x}|\underset{\sim}{\theta})d\underset{\sim}{x}$

Given the observed values $X_1 = x_1^*, \ldots, X_n = x_n^*$, the **likelihood function** of Θ is defined as

pdf/pmf c.f.

fixed varying

$$\mathcal{L}(\Theta) = f(x_1^*, x_2^*, \ldots, x_n^*|\Theta),$$

$\ell(\underset{\sim}{\theta})$
$\|$

which is a function of Θ. The **log likelihood function** is defined as $\log \mathcal{L}(\Theta)$.

Notes.

1. We consider likelihood function as a function of θ while joint pdf/pmf as a function of x_i's.
2. For discrete case, likelihood function gives the probability of observing the data as a function of θ.

How about continuous case?

$\sum_{\underset{\sim}{\theta}} f(\underset{\sim}{x}|\underset{\sim}{\theta}) \neq 1 \neq \int_{\underset{\sim}{\theta}} f(\underset{\sim}{x}|\underset{\sim}{\theta})d\underset{\sim}{\theta}$ fixed

Definition 6.9 (maximum likelihood estimator, TBp. 267)

The **maximum likelihood estimator** (**MLE**) of θ is the value of θ that maximizes the likelihood. → Is it an estimator? i.e., a function of $X_1, \cdots, X_n$: $\hat{\theta}(X_1, \cdots, X_n)$?

Interpretation. MLE makes the observed data "most probable" or "most likely," i.e., MLE gives the most "plausible" model given the observed data.
in terms of probability

Note.

1. For i.i.d. case, the likelihood function and the log likelihood function are, respectively,

function of $\underset{\sim}{\theta}$

marginal pdf/pmf

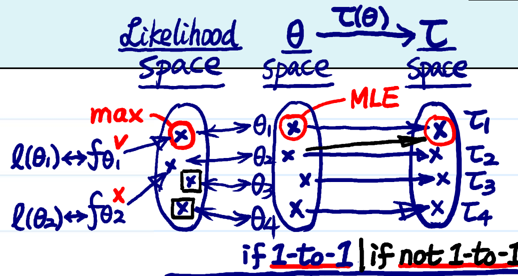
$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i^*|\theta), \quad \text{and} \quad l(\theta) = \sum_{i=1}^{n} \log f(x_i^*|\theta). \equiv \log(\mathcal{L}(\theta))$$

fixed

2. Maximizing the likelihood function, $\mathcal{L}(\theta)$, is equivalent to maximizing its natural logarithm, $l(\theta)$, since the logarithm is a monotonic function.

Theorem 6.1 (invariance property of MLE) e.g. Gamma(α,λ), $\underline{\theta}=(\alpha,\lambda)$, $\tau(\underline{\theta})=\alpha/\lambda=$ mean

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



Proof. MLE of $\tau(\theta)$ is a solution of the maximization problem

$$\max_{\tau^*}\ \max_{\theta:\tau(\theta)=\tau^*} l(\theta) = \max_{\tau(\theta)} l(\theta).$$

(FYI) profile likelihood $\mathcal{L}(\tau^*) = \sup_{\theta:\tau(\theta)=\tau^*} \mathcal{L}(\theta)$

Since $\hat{\theta}$ is the MLE of θ, the maximum is attained when $\theta=\hat{\theta}$, which implies the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Example 6.10 (i.i.d Poisson distribution, TBp. 268)

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $P(\lambda)$. The log likelihood is

statistical modeling

$$l(\lambda) = \sum_{i=1}^{n} \log \frac{e^{-\lambda}\lambda^{X_i}}{X_i!} = -n\lambda + \log\lambda \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log X_i!.$$

Setting $l'(\lambda)=0$ gives $\frac{1}{\lambda}\sum_{i=1}^{n} X_i - n = 0.$

The MLE is then

a function of data → $\hat{\lambda} = \overline{X}$. ← same as the moment estimator (LNp 9) → Sampling distribution discussed in LNp.11

Check that this is a maximum:

$$l''(\lambda) = -\frac{n\overline{X}}{\lambda^2} < 0 \quad \Rightarrow \quad l(\lambda) \text{ is concave.}$$

次 單位時間 / 單位時間 次

- **Example for Thm 6.1, LNp.19** $\Rightarrow$ the MLE of $\frac{1}{\lambda}$ is $\frac{1}{\overline{X}}$.

Example 6.11 (i.i.d normal distribution, TBp. 269)

Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu,\sigma^2)$ random variables. The joint density is

statistical modeling

$$f(x_1, x_2, \ldots, x_n|\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right].$$

note: not σ^2

The log likelihood is

$$l(\mu,\sigma) = \sum_{i=1}^{n}\left[-\log\sigma - \frac{1}{2}\log(2\pi) - \frac{1}{2}\left(\frac{X_i-\mu}{\sigma}\right)^2\right].$$

Setting

$$\begin{cases} 0 = \frac{\partial l}{\partial \mu} = \sigma^{-2}\sum_{i=1}^{n}(X_i-\mu) \\ 0 = \frac{\partial l}{\partial \sigma} = -n\sigma^{-1} + \sigma^{-3}\sum_{i=1}^{n}(X_i-\mu)^2 \end{cases}$$

The MLE is then

sampling distribution discussed in LNp.13

$$\begin{cases} \hat{\mu} &= \overline{X} \quad \leftarrow \text{sample mean} \\ \hat{\sigma} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2} \quad \leftarrow \text{sample variance} = \hat{\sigma}^2 \end{cases}$$

which is the same as the method of moments estimators. ← LNp.13

Check maximum ⇒

$$\begin{pmatrix} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \sigma \partial \mu} \\ \frac{\partial^2 l}{\partial \mu \partial \sigma} & \frac{\partial^2 l}{\partial \sigma^2} \end{pmatrix} = -\begin{pmatrix} \frac{n}{\sigma^2} & \frac{2}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu) \\ \frac{2}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu) & \frac{3}{\sigma^4}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{n}{\sigma^2} \end{pmatrix}$$

which is negative definite when $\mu=\hat{\mu}$ and $\sigma=\hat{\sigma}$ and $\mathcal{L} \to 0$ as (μ,σ) tends to boundary.

→ local maximum.

→ It's global maximum.

- **Example for Thm 6.1, LNp.19,**
  - MLE of μ^2, the square of a normal mean, is $\overline{X}^2$
  - MLE of σ^2, the variance, is $\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$

    ← same as the moment estimator

Example 6.12 (i.i.d restricted normal distribution)

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. from $N(\mu, 1)$ with $0 \le \mu < \infty$. The log likelihood is ← statistical modeling

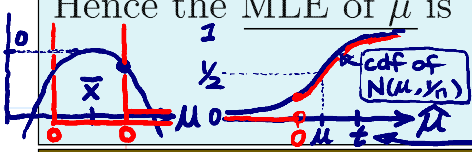
from the $\ell(\mu,\sigma)$ in LNp.20, $\sigma = 1$

sampling distribution=? Note. $\overline{X} \sim N(\mu, 1/n)$

$$\begin{aligned} l(\mu) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(\underbrace{X_i - \mu}_{-\overline{X}+\overline{X}})^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(X_i-\overline{X})^2 - \frac{n}{2}(\overline{X}-\mu)^2 \cancel{-\sum_{i=1}^{n}(X_i-\overline{X})(\overline{X}-\mu)} \end{aligned}$$

Hence the MLE of μ is

always falls in $(0,\infty)$ (Why?)



$$\hat{\mu} = \begin{cases} \overline{X}, & \text{if } \overline{X} \ge 0 \\ 0, & \text{if } \overline{X} < 0 \end{cases}$$

cf. moment estimator $\hat{\mu} = \overline{X}$ (Ec)

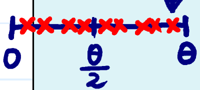
reasonable?

$P(\hat{\mu} \le t) = P(\overline{X} \le t)$ if $t > 0$. 3/19

Example 6.13 (i.i.d uniform(0, θ) distribution)

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $U(0,\theta)$, where $\theta > 0$. Then the likelihood of θ is ← $X_{(n)}$

← statistical modeling



c.f. moment estimator $\hat{\theta} = 2\overline{X}$ (Ec) ⇒ unbiased ⇒ but $2\overline{X}$ might be $< X_{(n)}$

$$\mathcal{L}(\theta) = \begin{cases} \theta^{-n}, & \text{if } 0 \le X_i \le \theta,\ i=1,\ldots,n \\ 0, & \text{otherwise} \end{cases}$$

Q: Are $X_{(n)}$, $2\overline{X}$ reasonable estimators?

$$= \begin{cases} \theta^{-n}, & \text{if } \theta \ge \max_{1 \le i \le n} X_i = X_{(n)} \ge 0 \\ 0, & \text{otherwise} \end{cases}$$

sampling distribution=?

alternative estimator $\frac{n+1}{n}X_{(n)} = (1+\frac{1}{n})X_{(n)}$

Because $\mathcal{L}(\theta)$ decreases when θ increases, the MLE of θ is $X_{(n)}$.

① $E(X_{(n)}) \underset{(Ec)}{=} \frac{n}{n+1}\theta$ ⇐ biased ② always underestimate