INIAD statistics and probability A

Week 7
“Statistical estimation”

2018.7th Nov.

Hirotada Honda

Lecture contents

Week 1: Mathematical preliminaries
(1) Introduction
(2) Mathematical preliminaries needed throughout this course (set theory, algebraic equation, differential and integral calculus, etc).
Week 2: Frequency Distributions
(3) What is a frequency distribution?
(4) Representation and understanding of frequency distributions To study the methods of understanding the characteristics of data distributions with graphs.
Week 3: Descriptive Statistics
(5) What are descriptive statistics?
(6) Various measures of descriptive statistics To study the methods of describing the characteristics of data with quantitative measures.
Week 4: Basics of Probability and Probability Distributions
(7) Introduction to probability
(8) Introduction to probability distributions To acquire a basic understanding of probability theory and probability distributions.
Week 5: Populations and Samples
(9) Relationship between populations and samples
(10) Relationship between parameters and statistics To understand the relationship between an entire set of cases of interest (a population) and a subset of the cases extracted from the population (a sample).
Week 6: Introduction to Statistical Inference
(11) What is statistical inference?
(12) Basic ideas of statistical inference To study the methods of estimating population values (parameters) from observed values (sample statistics).
Week 7: Statistical Estimation
(13) Various methods of statistical estimation
(14) Application of statistical estimation To study the basics of statistical estimation (point estimation and interval estimation).

Week 8: Summary (15) Summary of basic ideas of statistical data analysis

1-1. Statistical estimation

Statistical estimation

Find the unknown population parameters on the basis of the observed data.

Ex) After tossing a coin n times, its specific side appeared X times. Then, estimate the actual probability p that this specific side appears.

⇒ It's natural to estimate as: $\hat{p} = \frac{X}{n}$

Since X is a random variable, the quantity above is also a r.v..
Called as the *estimator*.

Suitable estimator?

Assumption: The observed value X (it’s a r.v.) follows a certain probability density with a parameter θ.

Question: Make an estimator $\tilde{\theta}(X)$

What kind of the method is “suitable”?

-If you know the actual value θ_0, then $\tilde{\theta}(X) = \theta_0$

-But it’s unknown.

Suitable characteristics of estimator

Unbiased:

-The expected value of the estimator matches with the actual value θ_0. (Unbiased estimator)

Minimized mean squared error (MSE)

- MSE (the sum of the variance and bias) is minimized.

Maximum likelihood estimate (MLE)

- Can be applied in case unbiased estimator cannot be obtained.

Examples of unbiased estimator

Consider again the estimator $\hat{p}$ of the former example.

-Since binomial distribution, E[X] = np and so

$$E\left[\hat{p}\right] = \frac{E[X]}{n} = p$$

Thus, $\hat{p}$ is the unbiased estimator of *p*.

Examples of unbiased estimator

Given a sequence $X_1, X_2, \ldots, X_n$, estimate the population mean and Variance.

The expected mean of the sample mean matches with the Population mean.
→ The sample mean is the unbiased estimator of the population mean.

How about the variance?

The variance of

$$\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

is <u>not</u> an unbiased estimator.

The unbiased variance

$$\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

Is the unbiased estimator of the variance.

Bias of estimator

The difference between the expected value of the estimator
And the actual value:

$$Bias = E[\hat{\theta}(X)] - \theta_0$$

Usually unknown since the actual value is.

Bias of estimator

In the former example, We can consider other estimators:

$$\hat{p}^* = \frac{X+1}{n+2}$$

Bias

$$E[\hat{p}^*] = \frac{np+1}{n+2} = p + \boxed{\frac{1-2p}{n+2}}$$

Accuracy of estimator

The estimator is a r.v. Usually, the accuracy is measured by MSE (mean squared error):

$$E[(\hat{p}^* - p)^2] = \underbrace{V(\hat{p}^*)}_{\text{Variance of estimator}} + \underbrace{\{E[\hat{p}^*] - p\}^2}_{\text{Squared bias}}$$

Smaller the MSE is, better the estimator is.
- Smaller variance, smaller bias.

Accuracy of estimator

$X_1, X_2, \ldots, X_n$ (n>=3) iid, that follow N(μ、 σ²)
#μ is unknown, σ² is known / unknown
-All the followings are the unbiased estimator of μ.

- $\hat{\mu}_1 = X_1;$
- $\hat{\mu}_2 = (X_1 + X_2)/2;$
- $\hat{\mu}_3 = (4 - \pi)X_1 + (\pi - 3)X_n;$
- $\hat{\mu}_4 = \sum_{i=1}^{n} X_i/n$

But then, which is the most desirable?

$\hat{\mu}_1$ is simple

$\hat{\mu}_4$ is popular…

Accuracy of estimator

The estimator with the smallest variance is best!
(Efficiency)

$$V[\mu_1] = V[X_1] = \sigma^2$$

$$V[\hat{\mu}_2] = (V[X_1] + V[X_2])/4 = (\sigma^2 + \sigma^2)/4 = \frac{\sigma^2}{2}$$

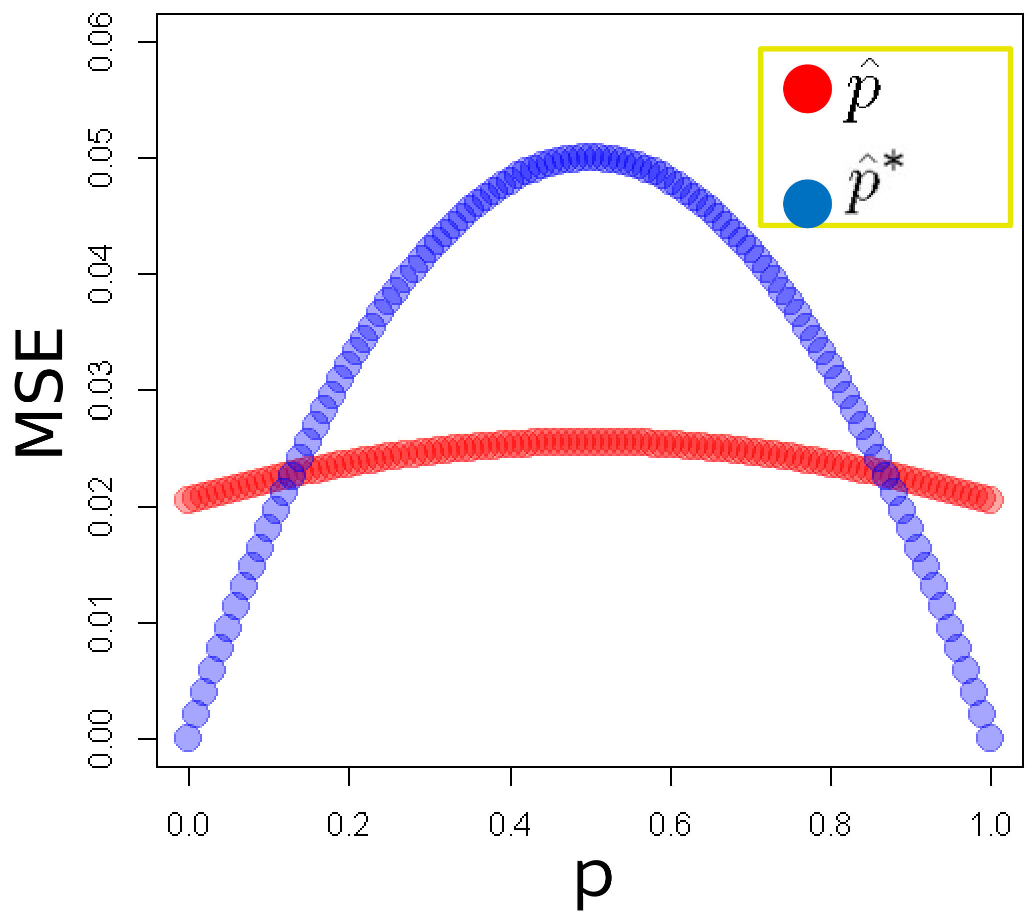
$$V[\hat{\mu}_3] = (4-\pi)^2 V[X_1] + (\pi - 3)^2 V[X_n] = 0.76\sigma^2$$

$$\boxed{V[\hat{\mu}_4] = \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}}$$

$\hat{\mu}_4$ has the highest efficiency among these.

Example of coin tossing

Compare $\hat{p}$ and $\hat{p}^*$ from the viewpoint of the MSE.
$\hat{p}^*$ has smaller MSE around p=0 and p=1.



Maximum Likelihood Estimate

Example of lottery

In a certain lottery, they can get the winning piece with the probability of p.

Bernoille trial

Estimate the actual value of p in the following situations:

i) At first, you got a piece, which was a winning piece.
ii) Second, you got a piece, which was a losing piece.
iii) After that, you get a piece 3 times, whose results were win / win / lose, respectively.

Idea of maximum likelihood estimate (MLE)

Estimate p as a maximizer of the probability that the observed Situation happens.

Idea of maximum likelihood estimate (MLE)

Estimate *p* as a maximizer of the probability that the observed Situation happens.

i) At first, you got a piece, which was a winning piece.

→ Obviously, $\hat{p} = 1$ at this moment,

Idea of maximum likelihood estimate (MLE)

Estimate *p* as a maximizer of the probability that the observed Situation happens.

i) Second, you got a piece, which was a losing piece.

→ Probability of such a situation is... denoted as L(p)!

$L(p) = p*(1-p) = p-p^2$ Convex upward!

What is the maximizer of L(p) above on $0 \leqq p \leqq 1$?
p=0.5.
Thus, $\hat{p} = 0.5$ at this moment.

Idea of maximum likelihood estimate (MLE)

Estimate *p* as a maximizer of the probability that the observed Situation happens.

After you tried 5 times, 3 wins and 2 loses.

→ Probability of such a situation is... denoted as L(p)!

$$L(p) = p^3(1-p)^2$$

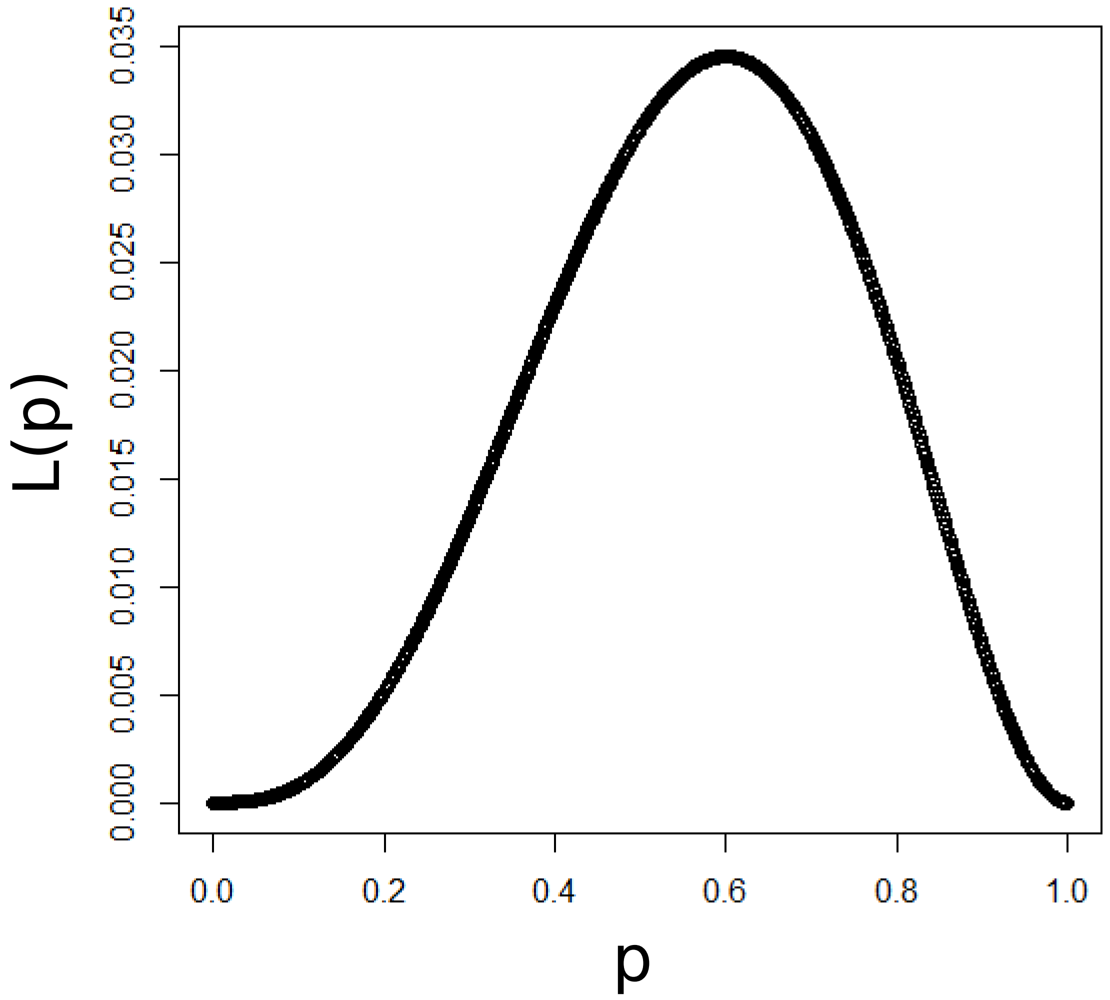
What is the maximizer of L(p) above on $0 \leqq p \leqq 1$?

Idea of maximum likelihood estimate (MLE)

$L(p) = p^3(1-p)^2$

What is the maximizer of L(p) above on $0 \leqq p \leqq 1$?

$$\frac{\partial L(p)}{\partial p} = 3p^2(1-p)^2 - p^3 \times 2(1-p) = p^2(1-p)(3-5p)$$



$$\hat{p} = 0.6$$

Likelihood

Estimate p as a maximizer of the probability that the observed Situation happens.

The colored term above is called as the *likelihood*.

Thus, MLE estimates parameters of population as the maximizer of the likelihood. The estimated parameters in MLE are called as *Maximum likelihood estimator*.

Formulation

Let $f(x;\Theta)$ be a pdf (or distribution function for a discrete r.v.), and θ is a parameter to be estimated.

Θ may be a vector!
i.e., multiple arameters are ok.

Ex)In the former example, let X be the number of winning piece out of N trials.

Then,

$$f(x;\theta)=\theta^{x}(1-\theta)^{N-x}$$

Parametric distribution and *Kullback–Leibler divergence*

Distance between two continuous models

Let f(x) and g(x) be continuous pdfs.

Then, the Kulback-Leibler (KL) divergence is:

$$I(g;f) \equiv E_G\left[\log \frac{g(X)}{f(X)}\right] = \int \log\left\{\frac{g(x)}{f(x)}\right\} g(x)\,\mathrm{d}x$$

Here, E_G means the expected value with respect to the distribution of g(x).
Distance between two probability distributions!

Distance between two discrete models

Let f(x) and g(x) be discrete probability functions.

Then, the Kulback-Leibler (KL) divergence is:

$$I(g;f) \equiv E_G\left[\log \frac{g(X)}{f(X)}\right] = \sum g(x_i) \log \frac{g(x_i)}{f(x_i)}$$

Here, E_G means the expected value with respect to the distribution of g(x).
Distance between two probability distributions!

Features of KL divergence

(i) $I(g; f) \geq 0;$

(ii) $I(g; f) = 0 \Leftrightarrow g(x) = f(x).$

Example of KL divergence

Two dice A and B, with probabilities of pips (1,2,3,4,5,6):

$$f_A = \{0.2, 0.12, 0.18, 0.12, 0.2, 0.18\},$$

$$f_B = \{0.18, 0.12, 0.14, 0.19, 0.22, 0.15\},$$

Which is closer to the ideal one: g={1/6,...,1/6}?

Example of KL divergence

Two dice A and B, with probabilities of pips (1,2,3,4,5,6):
Which is closer to the ideal one: g={1/6,...,1/6}?

$$I(g; f_A) = \sum_{i=1}^{6} g_i \log \frac{g_i}{f_{Ai}} = 0.023$$

$$I(g; f_B) = \sum_{i=1}^{6} g_i \log \frac{g_i}{f_{Bi}} = 0.020$$

So B is closer.

KL divergence as a measure of estimator

Mean log likelihood

Suppose now we want to estimate an unknown pdf g(x). Let f(x) be its estimator. Then,

$$I(g;f) = E_G\left[\log\frac{g(X)}{f(X)}\right] = E_G[\log g(X)] - \boxed{E_G[\log f(X)]}$$

But the 1st term is a constant (unknown, but constant).

Larger the 2nd term, closer these two models are.

Find f(x) that maximizes the 2nd term!

Mean log likelihood

The 2nd term is called as the mean log likelihood:

$$E_G[\log f(X)] = \begin{cases} \int g(x) \log f(x) \, dx; & \text{(if continuous)} \\ \sum g(x_i) \log f(x_i) \, dx; & \text{(if discrete)} \end{cases}$$

But still contains unknown g(x)...
Replace it with the empirical distribution:

Empirical distribution

Let the observed sample be: $\{x_1, x_2, \ldots, x_n\}$

Then, the empirical distribution is:

$$\hat{g}(x) = \begin{cases} \dfrac{1}{n} & (x = x_i \text{ for each i}) \\ 0 & \text{(otherwise).} \end{cases}$$

By using this…

Estimator of mean log likelihood

$$E_{\hat{G}}[\log f(X)] \equiv \int \hat{g}(x) \log f(x) \, \mathrm{d}x = \sum_{i=1}^{n} \hat{g}(x_i) \log f(x_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \log f(x_i)$$

Law of large number states:

$$\frac{1}{n} \sum_{i=1}^{n} \log f(x_i) \to E_G[\log f(X)] \quad n \to +\infty$$

Estimator of mean log likelihood

Therefore,

$$\frac{1}{n}\sum_{i=1}^{n}\log f(x_i)$$

is a natural estimator of the mean log likelihood.
Now, by multiplying n, the following quantity is called as the log likelihood.

$$\sum_{i=1}^{n}\log f(x_i)$$

Maximum likelihood estimate (MLE)

Now, assume that the unknown pdf f(x) contains a parameter (a vector in general)

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^{\mathrm{T}}$$

It can be denoted as

$$f(x|\boldsymbol{\theta})$$

Log-likelihood

Consider the log-likelihood:

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(X_i; \theta)$$

Maximize this!

The maximizer remains the same.

Example of log-likelihood

The normal dist.

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

- Two parameters μ and σ^2.
- ⇒We denote the likelihood as $L(\mu,\sigma^2)$.

Example of log-likelihood

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

If we have the observed data X_1, X_2, …, X_n:

$$\ln L(\mu, \sigma^2) = \sum_{i=1}^{n} \ln \left[\frac{1}{\sqrt{2\pi\sigma}} \exp\left\{ -\frac{(X_i - \mu)^2}{2\sigma^2} \right\} \right]$$

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

Example of log-likelihood

Then, we should find the pair of (μ, σ^2) that maximizes
The log-likelihood $\ln L(\mu, \sigma^2)$.

$$\frac{\partial}{\partial \mu} \ln L(\mu, \sigma^2)\Big|_{\mu=\hat{\mu}} = \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(X_i - \hat{\mu})$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} X_i - n\mu\right) = 0$$

In this case, we luckily have such a value of μ by just considering the above equality (it’s a special case).

Example of log-likelihood

The maximum likelihood estimator of μ is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The maximum likelihood estimator of the population mean of the normal distribution is equal to the sample mean.

On the other hand,

$$\left.\frac{\partial}{\partial \sigma^2} \ln L(\mu, \sigma^2)\right|_{\mu=\hat{\mu}, \sigma^2=\hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 = 0$$

$$\Downarrow$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Cautions in MLE

MLE is applicable to the cases in which we cannot find the Unbiased estimator.
→ In general, it has some bias.

Bias is unknown, but some statistical approaches tries to estimate it through the observed data, and correct the bias Included in MLE.
→ AIC, BIC, EIC

Bias and Information Criteria (IC)

A pdf g(x): unknown

We estimate g(x) within a parametric pdfs {f(x|θ)}.

Minimize the KL divergence between g and f by taking
The appropriate θ.

That is, maximize the mean log-likelihood:

$$E_G[\log f(X)]$$

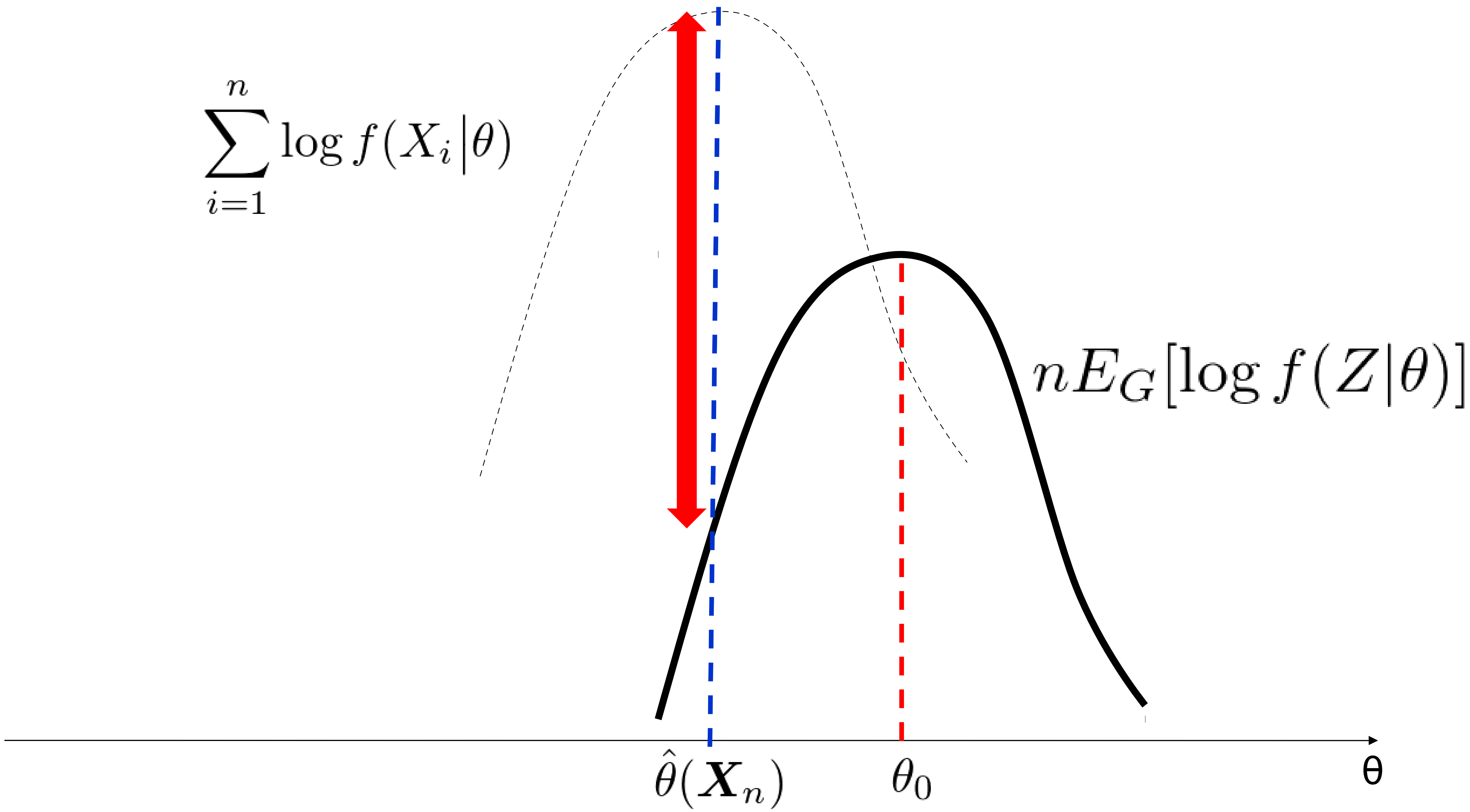
Estimator

But it's unknown. So approximate it with

$$\frac{1}{n}\sum_{i=1}^{n}\log f(x_i)$$

Actually, the log-likelihood is $\sum_{i=1}^{n}\log f(x_i)$. This is an estimator of

$$nE_G\left[\log f(X)\right]$$


$\sum_{i=1}^{n} \log f(X_i|\theta)$
$nE_G[\log f(Z|\theta)]$
$\hat{\theta}(\boldsymbol{X}_n)$
θ_0
θ

Definition of bias

Then, the bias of this estimator is:

$$b(G) \equiv E_{G(\boldsymbol{x}_n)}\left[\sum_{i=1}^{n} \log f(x_i|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n)) - nE_{G(z)}[\log f(Z|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n))]\right]$$

$E_{G(\boldsymbol{x}_n)}$: The expected value with respect to the simultaneous distribution of samples $\boldsymbol{X}_n$

$E_{G(z)}$: The expected value with respect to g(x)

IC

The information criteria is defined as:

$$IC(\boldsymbol{X}_n; \hat{G}) \equiv -2\sum_{i=1}^{n} \log f(X_i | \hat{\boldsymbol{\theta}}) + 2\widetilde{b(G)}$$

Here, $\widetilde{b(G)}$ is the estimator of the bias b(G).

The smaller this value is, the better the estimator is.

AIC (Akaike Information Criteria)

Approximately estimate the bias by the dimension p
Of the model’s parameter.

$$AIC \equiv -2\sum_{i=1}^{n} \log f(X_i | \hat{\boldsymbol{\theta}}) + 2p$$

The smaller this value is, the better the estimator is.

Example of AIC

In the polynomial regression:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

$$AIC_p = n(\log 2\pi + 1) + n \log \sigma^2 + 2(p + 2)$$

Example of AIC (2)

If we model a certain pdf by the normal distribution

$N(\mu, \sigma^2)$ with μ and σ^2 as parameters, take p=2.

So the bias correction term is 4.

Summary(checklist)

-How MLE works?

-What is bias?

-What is likelihood / log-likelihood?

-You can apply MLE to the normal distribution?

【Ref.】 Intuitive understanding of log-likelihood for continuous distributions

Likelihood

Let X_1, X_2, …,X_n be elements of a sample. Then, the likelihood for the continuous distribution is

$$L(\theta) = f(X_1;\theta)f(X_2;\theta)\ldots f(X_n;\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

Can be regarded as a function of θ!

Maximum likelihood estimator

The maximum likelihood estimator of a r.v. that follows a pdf *f* is, a maximizer of

$$L(\theta) = f(X_1;\theta)f(X_2;\theta)\dots f(X_n;\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

Log-likelihood

$$L(\theta) = f(X_1;\theta)f(X_2;\theta)\ldots f(X_n;\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

is hard to deal with.

We often consider the log-likelihood:

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(X_i;\theta)$$

The maximizer remains the same.

Q. 1

Recall that the pdf of the exponential distribution is as follows.

$$f(x;\lambda) = \lambda e^{-\lambda x}$$

i) Under the observed data of X_1, X_2, ...,X_n, find the log likelihood ln L(λ).

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(X_i;\theta)$$

ii) Find the maximum-likelihood estimator λ.

A.1

i)

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(X_i;\theta)$$

$$\log f(x|\theta) = \log \lambda - \lambda x$$

$$\log L(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

A.1

ii)

$$\frac{\partial}{\partial \lambda}\left(\log L(\lambda)\right) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

Thus, $\frac{\partial}{\partial \lambda}\left(\log L(\lambda)\right) = 0$ yields

$$\boxed{\lambda = \frac{n}{\sum_{i=1}^{n} x_i}}$$

Q.2

Recall that the pdf of the exponential distribution is as follows.

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

Now, given the observed data of 0.30, 0.06, 0.05, 0.08, 0.12
That follow the exponential distribution, find the ML estimator λ.

Q.2

Recall that the pdf of the exponential distribution is as follows.

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

Now, given the observed data of 0.30, 0.06, 0.05, 0.08, 0.12
That follow the exponential distribution, find the ML estimator
λ.

Hint; You can apply the
result of Q.1

A.2

0.30, 0.06, 0.05, 0.08, 0.12

Since

$$\lambda = \frac{n}{\sum_{i=1}^{n} x_i}$$

and n=5,

$$\sum_{i=1}^{n} x_i = 0.30 + 0.06 + 0.05 + 0.08 + 0.12 = 0.61$$

$$\lambda = 5/0.61 = \boxed{8.2}$$

Q.3

As we observed the customer arrival interals in a certain amusement park, the observed data were:

1.51, 0.13, 0.21, 2.29, 0.11, 0.79, 0.65, 1.10, 1.08, 2.11 [sec].

Given that they follow the exponential distribution,

i) Find the ML estimator λ;
ii) Find the probability that an interval is 1 sec or less.

A.3

i) Find the ML estimator λ;

$$\sum_{i=1}^{n} x_i = 1.51 + 0.13 + 0.21 + 2.29 + 0.11 + 0.79 + 0.65 + 1.10 + 1.08 + 2.11 = 0.998$$

$$\lambda = 10/0.998 = \boxed{1.0}$$

ii) Find the probability that an interval is 1 sec or less.

$$P(X \leqq x) = 1 - e^{-\lambda x}$$

$$P(X \leqq 1) = 1 - e^{-1.0\times 1} = = \underline{0.63}$$

Q.4

The Weibull distribution is used to model the interval of system failures. A specific form of its pdf is:

$$f(x\ ;\ \lambda) = 2\lambda^{-2} x e^{-\frac{x^2}{\lambda^2}} \quad (x \geq 0, \lambda > 0)$$

Hint:

$$\ln(f(x)) = \ln(2\lambda^{-2}) + \ln x - \frac{x^2}{\lambda^2}$$

i) Given $X_1, X_2, \ldots, X_n$,
Find the log-likelihood;
ii) Find the ML estimator λ in general;
iii) Given the following data, find the ML estimator λ:
7.01, 7.72, 3.57, 2.56, 3.53

$$\ln(f(x)) = \ln(2\lambda^{-2}) + \ln x - \frac{x^2}{\lambda^2}$$

$$\Downarrow$$

$$\ln L(\lambda) = n\ln(2\lambda^{-2}) + \sum \ln(x_i) - \frac{1}{\lambda^2}\sum x_i^2$$

A.4

$$L(\lambda) = f(x_1;\lambda) \times f(x_2;\lambda) \times \dots f(x_n;\lambda) = \frac{2}{\lambda^2} x_1 \mathrm{e}^{-\frac{x_1^2}{\lambda^2}} \times \frac{2}{\lambda^2} x_2 \mathrm{e}^{-\frac{x_2^2}{\lambda^2}} \times \dots \frac{2}{\lambda^2} x_n \mathrm{e}^{-\frac{x_n^2}{\lambda^2}}$$

$$\boxed{= \frac{2^n}{\lambda^{2n}} \left(\prod_{j=1}^{n} x_j\right) \mathrm{e}^{-\frac{\sum_{k=1}^{n} x_k^2}{\lambda^2}}}$$

$$= \frac{2^n}{\lambda^{2n}} \left(\prod_{j=1}^{n} x_j\right) \mathrm{e}^{-\frac{\sum_{k=1}^{n} x_k^2}{\lambda^2}}$$

①
$$\boxed{\log L(\lambda) = n \log 2 - 2n \log \lambda + \sum_{j=1}^{n} \log x_j - \frac{1}{\lambda^2} \sum_{k=1}^{n} x_k^2}$$

A.4

$$\log L(\lambda) = n \log 2 - 2n \log \lambda + \sum_{j=1}^{n} \log x_j - \frac{1}{\lambda^2} \sum_{k=1}^{n} x_k^2$$

$$\downarrow$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}(\log L(\lambda)) = -\frac{2n}{\lambda} + \frac{2}{\lambda^3} \sum_{k=1}^{n} x_k^2 = 0$$

②

$$\boxed{\lambda = \sqrt{\frac{\sum_{k=1}^{n} x_k^2}{n}}}$$

A.4

$$\lambda = \sqrt{\frac{\sum_{k=1}^{n} x_k^2}{n}}$$

③

$$\hat{\lambda} = \sqrt{\frac{(7.01^2 + 7.72^2 + 3.57^2 + 2.56^2 + 3.53^2)}{5}} = 5.30$$

Q.5

There are two machines A and B, whose lifetime, denoted as X1 and X2, follow the same exponential distribution:

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

i) We have observed X1=a and X2=b. Then, find the MLE of λ.

ii) At time t, we have observed X1 = a but the machine B was still running. Then, find the MLE of λ.

A.5

(i)

$$f(x_1, x_2) = f(x_1)f(x_2) = \lambda^2 \mathrm{e}^{-\lambda(x_1+x_2)}$$

$$\ln L(\lambda) = 2\ln\lambda - \lambda(a+b)$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}L(\lambda) = \frac{2}{\lambda} - (a+b) = 0 \iff \lambda = \frac{a+b}{2}$$

A.5

(ii) Consider the probability of lifetime of A and the situation “B is still running”.

P(Lifetime of A is x or less and B is working at t)

$$= P(X_1 \le x)P(X_2 \ge t)$$

So, the pdf of x in this situation is

$$\frac{\mathrm{d}}{\mathrm{d}x} P(X_1 \le x)P(X_2 \ge t) = f(x)P(X_2 \ge t).$$

A.5

But

$$P(X_2 \geq t) = 1 - P(X_2 < t) = 1 - (1 - \mathrm{e}^{-\lambda t}) = \mathrm{e}^{-\lambda t}.$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}x} P(X_1 \leq x) P(X_2 \geq t) = \lambda \mathrm{e}^{-\lambda x} \mathrm{e}^{-\lambda t} = \lambda \mathrm{e}^{-\lambda(x+t)}.$$

$$\Downarrow$$

$$\ln L(\lambda) = \ln \lambda - \lambda(a+t)$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} L(\lambda) = \frac{1}{\lambda} - (a+t) = 0$$

$$\boxed{\lambda = \frac{1}{a+t}}$$

Q.6

We want to estimate the number of fish in a certain large pond.
Now, we marked m fish with a red marker, and then released them into the pond. Then, we caught n fish from the lake,
and found that k out of n were marked.
Now, estimate the number of fish N (including m fish released) in the lake by MLE.

A.6

Total N, caught n, k were marked.

$$L(N) = f(k; N) = \frac{{}_{m}C_{k}\ {}_{N-m}C_{n-k}}{{}_{N}C_{n}}$$

Find the maximizer *N* of this quantity (likelihood).
Because *N* is an integer, it's simpler to maximize this likelihood directly (not the log-likelihood).

Noting that

$$\frac{L(N+1)}{L(N)} = \frac{(N+1-m)(N+1-n)}{(N+1)(N+1-m-n+k)}$$

A.6

Find an N that satisfies

$$\frac{L(N+1)}{L(N)} < 1, \quad \frac{L(N1)}{L(N-1)} > 1$$

$\Leftrightarrow$

$$\frac{mn}{k} - 1 < N < \frac{mn}{k}$$

$$N = [\frac{mn}{k}]$$

Q.7

There is a coin that shows a specific side with the probability of θ in each trial. Now, after tossing this coin 100 times, the specific Side appeared 70 times. Find the MLE of θ.

A.7

The likelihood is $L(\theta) = {}_{100}\mathrm{C}_{70}\theta^{70}(1-\theta)^{30}$

$$\log L(\theta) = 70\log\theta + 30\log(1-\theta) + \log{}_{100}\mathrm{C}_{70}$$

$$\Downarrow$$

$$\frac{d}{d\theta}\log L(\theta) = \frac{70}{\theta} - \frac{30}{1-\theta}$$

$$\frac{70}{\theta} - \frac{30}{1-\theta} = 0 \iff \theta = 0.7$$

Q.8

Data $X_1, X_2, \ldots, X_n$, are observed, which are known to follow the Normal distribution $N(\mu, \sigma^2)$.

i) Find the MLE of μ。
ii) Find the MLE of σ^2

A.8

Data $X_1, X_2, \ldots, X_n$, are observed, which are known to follow the Normal distribution $N(\mu, \sigma^2)$.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Q.9

Data $X_1, X_2, \ldots, X_n$, are observed, which are known to follow the normal distribution $N(\mu, 1^2)$. It is also known that $0 \leq \mu \leq 1.$

Find the MLE of μ.

A.9

By using $\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$

with $\sigma^2 = 1,$ we should just focus on minimizing

$$\sum_{i=1}^{n}(X_i - \mu)^2 = n(\mu - \bar{X})^2 + n.$$

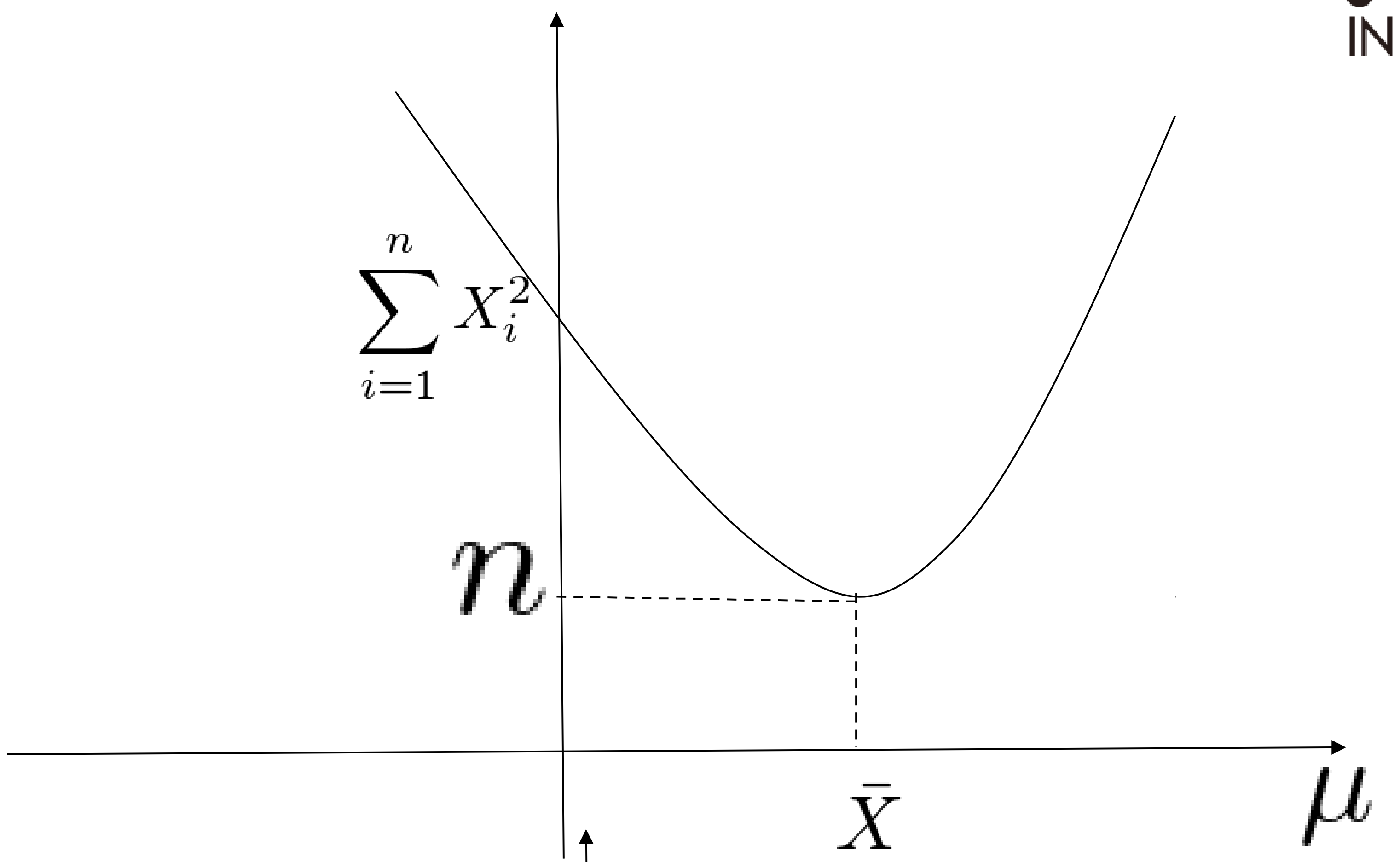
Here, $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$

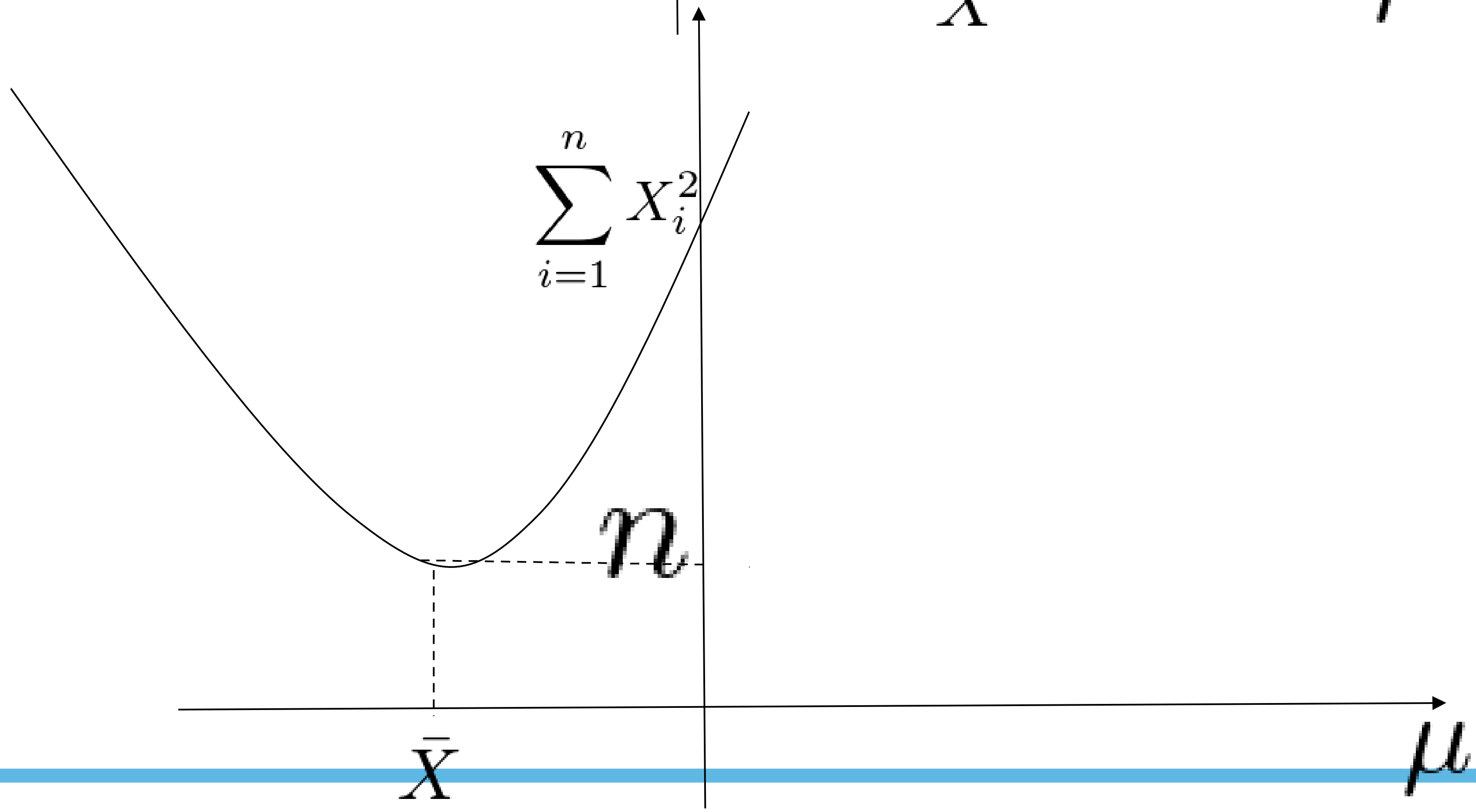
A.9

i) $\bar{X} \in [0,1]$ ➡ $\hat{\mu} = \bar{X}$

ii) $\bar{X} \geq 1$ ➡ $\hat{\mu} = 1$

iii) $\bar{X} \leq 0$ ➡ $\hat{\mu} = 0$





Q.10

Data $X_1, X_2, \ldots, X_n$, are observed, which are known to follow the normal distribution $N(\mu, \sigma^2)$. It is also known that $\mu = 0, \sigma^2 \geq 1.$

Find the MLE of σ^2.

A.10

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

Since μ=0 now,

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

For simplicity, we set

$$y = \sigma^2, \quad f(y) \equiv -\frac{n}{2}\ln(2\pi y) - \frac{\sum_{i=1}^{n} X_i^2}{2y}.$$

We should maximize this under: y≧1.

A.10

Equivalently, we should minimize:

$$g(y) \equiv n \ln(2\pi y) + \frac{\sum_{i=1}^{n} X_i^2}{y}$$

$$= n\left\{\ln(2\pi y) + \frac{V}{y}\right\},$$

Where

$$V = \frac{\sum_{i=1}^{n} X_i^2}{n}$$

A.10

Since the variance of sample is V, y=V is a candidate (global minimizer).

But how about y=1?
We should compare g(1) and g(V).

But it’ seen that g(1)≧g(V) holds for all V≧0.

Actually, $\ln(2\pi) + V \geq \ln(2\pi V) + 1 \quad V \geq 0$

So the answer is $\hat{\sigma}^2 = V = \dfrac{\sum_{i=1}^{n} X_i^2}{n}$