

Emory University

**MATH 362 Mathematical Statistics II**

Learning Notes

Jiuru Lyu

February 29, 2024

**Contents**

|          |  |          |
|----------|--|----------|
| <b>1</b> | <b>Estimation</b>  | <b>2</b> |
| 1.1      | Introduction . . . . .   | 2        |
| 1.2      | The Method of Maximum Likelihood and the Method of Moments . . . . . | 3        |
| 1.3      | The Method of Moment . . . . .                                       | 10       |
| 1.4      | Interval Estimation . . . . .  | 12       |
| 1.5      | Properties of Estimation . . . . .                                   | 15       |
| 1.6      | Best Unbiased Estimator . . . . .                                    | 18       |
| 1.7      | Sufficiency . . . . .  | 21       |
| 1.8      | Consistency . . . . .  | 24       |

# 1 Estimation

## 1.1 Introduction

**Definition 1.1.1 (Model).** A *model* is a distribution with certain parameters.

**Example 1.1.2** The normal distribution:  $N(\mu, \sigma^2)$ .

**Definition 1.1.3 (Population).** The *population* is all the objects in the experiment.

**Definition 1.1.4 (Data, Sample, and Random Sample).** *Data* refers to observed value from sample. The *sample* is a subset of the population. A *random sample* is a sequence of independent, identical (*i.i.d.*) random variables.

**Definition 1.1.5 (Statistics).** *Statistics* refers to a function of the random sample.

**Example 1.1.6** The sample mean is a function of the sample:

$$\bar{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n).$$

**Example 1.1.7 Central Limit Theorem**

We randomly toss  $n = 200$  fair coins on the table. Calculate, using the central limit theorem, the probability that at least 110 coins have turned on the same side.

$$\bar{X} = \frac{X_1 + \cdots + X_{200}}{200} \stackrel{\text{CLT}}{\sim} N(\mu, \sigma^2),$$

where

$$\mu = \mathbf{E}(\bar{X}) = \frac{\sum_{i=1}^{200} \mathbf{E}(X_i)}{200},$$

$$\sigma^2 = \mathbf{Var}(\bar{X}) = \mathbf{Var}\left(\frac{X_1 + \cdots + X_{200}}{200}\right) = \frac{\sum_{i=1}^{200} \mathbf{Var}(X_i)}{200^2}.$$

**Definition 1.1.8 (Statistical Inference).** The process of *statistical inference* is defined to be the process of using data from a sample to gain information about the population.

**Example 1.1.9** Goals in statistical inference

1. **Definition 1.1.10 (Estimation).** To obtain values of the parameters from the data.
2. **Definition 1.1.11 (Hypothesis Testing).** To test a conjecture about the parameters.
3. **Definition 1.1.12 (Goodness of Fit).** How well does the data fit a given distribution.
4. Linear Regression

**1.2 The Method of Maximum Likelihood and the Method of Moments**

**Example 1.2.1** Given an unfair coin, or  $p$ -coin, such that

$$X = \begin{cases} 1 & \text{head with probability } p, \\ 0 & \text{tail with probability } 1 - p. \end{cases}$$

How can we determine the value  $p$ ?

**Solution 1.**

1. Try to flip the coin several times, say, three times. Suppose we get HHT.
2. Draw a conclusion from the experiment.

**Key idea:** The choice of the parameter  $p$  should be the value that maximizes the probability of the sample.

$$\mathbf{P}(X_1 = 1, X_2 = 1, X_3 = 0) = \mathbf{P}(X_1 = 1)\mathbf{P}(X_2 = 1)\mathbf{P}(X_3 = 0) = p^2(1 - p) := f(p).$$

Solving the optimization problem  $\max_{p>0} f(p)$ , we find it is most likely that  $p = \frac{2}{3}$ . This method is called the *likelihood maximization method*. □

**Definition 1.2.2 (Likelihood Function).** For a random sample of size  $n$  from the discrete (or continuous) pdf  $p_X(k; \theta)$  (or  $f_Y(y; \theta)$ ), the *likelihood function*,  $L(\theta)$ , is the product of the pdf evaluated at  $X_i = k_i$  (or  $Y_i = y_i$ ). That is,

$$L(\theta) := \prod_{i=1}^n p_X(k_i; \theta) \quad \text{or} \quad L(\theta) := \prod_{i=1}^n f_Y(y_i; \theta).$$

**Definition 1.2.3 (Maximum Likelihood Estimate).** Let  $L(\theta)$  be as defined in Definition 1.2.2. If  $\theta_e$  is a value of the parameter such that  $L(\theta_e) \geq L(\theta)$  for all possible values of  $\theta$ , then we call  $\theta_e$  the *maximum likelihood estimate* for  $\theta$ .

**Theorem 1.2.4 The Method of Maximum Likelihood**

Given random samples  $X_1, \dots, X_N$  and a density function  $p_X(x)$  (or  $f_X(x)$ ), then we have the likelihood function defined as

$$\begin{aligned} L(\theta) &= p_X(X; \theta) = \mathbf{P}(X_1, X_2, \dots, X_N) \\ &= \mathbf{P}(X_1)\mathbf{P}(X_2) \cdots \mathbf{P}(X_N) && [\text{independent}] \\ &= \prod_{i=1}^N p_X(X_i; \theta) && [\text{identical}] \end{aligned}$$

Then, the maximum likelihood estimate for  $\theta$  is given by

$$\theta^* = \arg \max_{\theta} L(\theta),$$

where

$$L\left(\arg \max_{\theta} L(\theta)\right) = L^*(\theta) = \max_{\theta} L(\theta).$$

**Example 1.2.5** Consider the Poisson distribution  $X = 0, 1, \dots$ , with  $\lambda > 0$ . Then, the pdf is given by

$$p_X(k, \lambda) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, \dots$$

Given data  $k_1, \dots, k_n$ , we have the likelihood function

$$L(\lambda) = \prod_{i=1}^n p_X(X = k_i; \lambda) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{k_i}}{k_i!} = e^{-n\lambda} \frac{\lambda^{\sum k_i}}{k_1! \cdots k_n!}$$

Then, to find the maximum likelihood estimate of  $\lambda$ , we need to  $\max_{\lambda} L(\lambda)$ . That is to solve

$$\frac{\partial L(\lambda)}{\partial \lambda} = 0 \text{ and } \frac{\partial^2 L(\lambda)}{\partial \lambda^2} < 0.$$

**Example 1.2.6** Waiting Time.

Consider the exponential distribution  $f_Y(y) = \lambda e^{-\lambda y}$  for  $y \geq 0$ . Find the MLE  $\lambda_e$  of  $\lambda$ .

**Solution 2.**

The likelihood function of the exponential distribution is given by

$$\mathbf{L}(\lambda) = \prod_{i=1}^n \lambda e^{-\lambda y_i} = \lambda^n \exp \left( -\lambda \sum_{i=1}^n y_i \right).$$

Now, define

$$\ell(\lambda) = \ln \mathbf{L}(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^n y_i.$$

To optimize  $\ell(\lambda)$ , we compute

$$\frac{d}{d\lambda} \ell(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^n y_i \stackrel{set}{=} 0$$

So,

$$\frac{n}{\lambda} = \sum_{i=1}^n y_i \implies \lambda_e = \frac{n}{\sum_{i=1}^n y_i} =: \frac{1}{\bar{y}},$$

where  $\bar{y}$  is the sample mean. □

**Example 1.2.7** Given the exponential distribution  $f_Y(y) = \lambda e^{-\lambda y}$  for  $y \geq 0$ . Find the MLE of  $\lambda^2$ .

**Solution 3.**

Define  $\tau = \lambda^2$ . Then,  $\lambda = \sqrt{\tau}$ , and so

$$f_Y(y) = \sqrt{\tau} e^{-\sqrt{\tau} y}, \quad y \geq 0.$$

Then, the likelihood function becomes

$$\mathbf{L}(\tau) = \prod_{i=1}^n f_Y(y) = \tau^{\frac{n}{2}} \exp \left( -\sqrt{\tau} \sum_{i=1}^n y_i \right).$$

Similarly, after maximization, we find

$$\tau_e = \frac{1}{(\bar{y})^2}.$$

□

**Theorem 1.2.8 Invariant Property for MLE**

Suppose  $\lambda_e$  is the MLE of  $\lambda$ . Define  $\tau := h(\lambda)$ . Then,  $\tau_e = h(\lambda_e)$ .

**Proof 4.** In this proof, we will prove the case when  $h$  is a one-to-one function. The case of  $h$  being a many-to-one function is beyond the scope of this course.

Suppose  $h(\cdot)$  is a one-to-one function. Then,  $\lambda = h^{-1}(\tau)$  is well-defined. Then,

$$\max_{\lambda} \mathbf{L}(\lambda; y_1, \dots, y_n) = \max_{\tau} \mathbf{L}(h^{-1}(\tau); y_1, \dots, y_n) = \max_{\tau} \mathbf{L}(\tau; y_1, \dots, y_n).$$

■

**Example 1.2.9** Waiting Time with an unknown Threshold.

Let  $\lambda = 1$  in exponential but there is an unknown threshold  $\theta$ , that, is  $f_Y(y) = e^{-(y-\theta)}$  for  $y \geq \theta$ ,  $\theta > 0$ .

**Solution 5.**

Note that the likelihood function is given by

$$\begin{aligned} \mathbf{L}(\theta; y_1, \dots, y_n) &= \prod_{i=1}^n f_Y(y_i) = \exp \left( - \sum_{i=1}^n (y_i - \theta) \right), \quad y_i \geq \theta, \theta > 0 \\ &= \exp \left( - \sum_{i=1}^n (y_i - \theta) \right) \cdot \mathbb{1}_{[y_i \geq \theta, \theta > 0]}, \end{aligned}$$

where

$$\mathbb{1}_{x \in A} = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

Using order statistics,

$$\begin{aligned} \mathbf{L}(\theta) &= \exp \left( - \sum_{i=1}^n (y_i - \theta) \right) \cdot \mathbb{1}_{[y_{(n)} \geq y_{(n-1)} \geq \dots \geq y_{(1)} \geq \theta, \theta > 0]} \\ &= \exp \left( - \sum_{i=1}^n y_i + n\theta \right) \mathbb{1}_{[y_{(n)} \geq \dots \geq y_{(1)} \geq \theta, \theta > 0]}. \end{aligned}$$

So, we know  $\theta \leq y_{(1)} = y_{\min}$ .

To maximize the likelihood function, we want to maximize  $-\sum y_i + n\theta$ . That is, to maximize  $\theta$ , as  $\theta \leq y_{\min}$ , it must be that  $\theta_{\max} = y_{\min}$ . Therefore, the MLE is  $\theta^* = y_{\min}$ .  $\square$

**Example 1.2.10** Suppose  $Y_1, \dots, Y_n \sim \text{Uniform}[0, a]$ . That is,  $f_Y(y; a) = \frac{1}{a}$  for  $y \in [0, a]$ . Find MLE  $a_e$  of  $a$ .

**Solution 6.**

Note that

$$\begin{aligned} f_Y(y; a) &= \frac{1}{a} \cdot \mathbb{1}_{\{y \in [0, a]\}} \\ &= \frac{1}{a} \cdot \mathbb{1}_{\{0 \leq y_{(1)} \leq \dots \leq y_{(n)} \leq a\}} \end{aligned} \quad \text{where } y_{(1)} = \min y_i \text{ and } y_{(n)} = \max y_i$$

Then,

$$\mathbf{L}(a) = \frac{1}{a^n} \mathbb{1}_{\{0 \leq y_{(1)} \leq \dots \leq y_{(n)} \leq a\}}$$

To maximize  $\mathbf{L}(a)$ , we want to minimize  $a^n$ . Since  $a \geq y_{(n)}$ , it must be that  $a_e = y_{(n)}$ . Here, we call  $a_e = y_{(n)}$  an *estimate*, and  $\widehat{a_{\text{MLE}}} = Y_{(n)}$  an *estimator*.  $\square$

**Example 1.2.11 MLE that Does Not Exist**

Suppose  $f_Y(y; a) = \frac{1}{a}$ ,  $y \in [0, a)$ . Find the MLE.

**Solution 7.**

The likelihood function is the same:

$$\mathbf{L}(a) = \frac{1}{a^n} \mathbb{1}_{\{0 \leq y_{(1)} \leq \dots \leq y_{(n)} < a\}}.$$

However, since  $[0, a)$  is not a closed set, the optimization problem  $\max_{a \in [0, a)} \mathbf{L}(a)$  does not have a solution. Hence, the estimate does not exist.  $\square$

**Remark 1.1** MLE may not be unique all the time.

**Example 1.2.12 Multiple MLE Values**

Suppose  $X_1, \dots, X_n \sim \text{Uniform}\left[a - \frac{1}{2}, a + \frac{1}{2}\right]$ , where  $f_X(x; a) = 1$ ,  $x \in \left[a - \frac{1}{2}, a + \frac{1}{2}\right]$ . Find the MLE.

**Solution 8.**

In the indicator function notation, we can rewrite the pdf to be

$$f_X(x; a) = \mathbb{1}_{\{a - \frac{1}{2} \leq x \leq a + \frac{1}{2}\}} = \mathbb{1}_{\{a - \frac{1}{2} \leq x_{(1)} \leq \dots \leq x_{(n)} \leq a + \frac{1}{2}\}}.$$

So, the likelihood function will be

$$L(a) = \prod_{i=1}^n f_x(x_i; a) = \begin{cases} 1, & a \in \left[ x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2} \right] \\ 0, & \text{otherwise.} \end{cases}$$

So, the  $L(a)$  will be maximized whenever  $a \in \left[ x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2} \right]$ . Therefore, MLE can be any value in the range  $\left[ x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2} \right]$ . Say,

$$a_e = x_{(n)} - \frac{1}{2} \quad \text{or} \quad a_e = x_{(1)} + \frac{1}{2} \quad \text{or} \quad a_e = \frac{x_{(n)} - \frac{1}{2} + x_{(1)} + \frac{1}{2}}{2} = \frac{x_{(n)} + x_{(1)}}{2}.$$

□

### Theorem 1.2.13 MLE for Multiple Parameters

In general, we have the likelihood function  $L(\theta)$ , where  $\theta = (\theta_1, \dots, \theta_p)$ . To find the MLE, we need

$$\frac{\partial L(\theta)}{\partial \theta_i} = 0 \quad i = 1, \dots, p,$$

and the Hessian matrix

$$\left( \frac{\partial^2 L(\theta)}{\partial \theta_i \partial \theta_j} \right)_{i,j=1,\dots,p} := \begin{pmatrix} \frac{\partial^2 L(\theta)}{\partial \theta_1^2} & \cdots & \frac{\partial^2 L(\theta)}{\partial \theta_1 \partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 L(\theta)}{\partial \theta_p \partial \theta_1} & \cdots & \frac{\partial^2 L(\theta)}{\partial \theta_p^2} \end{pmatrix}$$

should be negative definite.

### Example 1.2.14 MLE for Multiple Parameters: Normal Distribution

Suppose  $Y_1, \dots, Y_n \sim N(\mu, \sigma)$ . Then,

$$f_{Y_i}(u; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i - \mu)^2 / (2\sigma^2)}.$$

Find the MLE for  $\mu$  and  $\sigma$ .

**Solution 9.**



The likelihood function will be

$$\mathbf{L}(\mu, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i - \mu)^2 / (2\sigma^2)}.$$

Then, we define

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

Set

$$\begin{cases} \frac{\partial \ell(\mu, \sigma)}{\partial \mu} = 0 & \textcircled{1} \\ \frac{\partial \ell(\mu, \sigma)}{\partial \sigma} = 0 & \textcircled{2} \end{cases}$$

From ①, we have

$$\begin{aligned} \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu) &= 0 \\ \sum_{i=1}^n y_i &= n\mu \implies \boxed{\mu_e = \frac{\sum y_i}{n} = \bar{y}} \end{aligned}$$

From ②, by the invariant property of MLE, we instead set

$$\begin{aligned} \frac{\partial \ell(\mu, \sigma)}{\partial \sigma^2} &= 0 \\ -\frac{n}{2} \cdot \frac{1}{\sigma^2} + \frac{1}{2} \left( \frac{1}{\sigma^2} \right)^2 \sum_{i=1}^n (y_i - \mu)^2 &= 0 \\ \frac{1}{2\sigma^2} \left( -n + \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right) &= 0 \\ -n\sigma^2 + \sum_{i=1}^n (y_i - \mu)^2 &= 0 \quad (\mu_e = \bar{y}) \\ \sum_{i=1}^n (y_i - \bar{y})^2 &= n\sigma^2 \\ \sigma_e^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 &\implies \boxed{\sigma_e = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} \end{aligned}$$

□

### 1.3 The Method of Moment

**Definition 1.3.1 (Moment Generating Function).** The *Moment Generating Function (MGF)* is defined as

$$\mathbf{M}_X(t) = \mathbf{E}[e^{tX}],$$

and it uniquely determines a probability distribution.

**Definition 1.3.2 (Moment).** The *k-th order moment* of  $X$  is  $\mathbf{E}[X^k]$ .

#### Example 1.3.3 Meaning of Different Moments

- $\mathbf{E}[X]$ : location of a distribution
- $\mathbf{E}[X^2] = \text{Var}(X) + \mathbf{E}[X]^2$ : width of a distribution
- $\mathbf{E}[X^3]$ : skewness – positively skewed / negatively skewed
- $\mathbf{E}[X^4]$ : kurtosis / tailedness – speed decaying to 0.

#### Example 1.3.4 Moment Estimate: Moments of Population and Sample

| Population                           | Sample, $X_1, \dots, X_n$  |
|--------------------------------------|--|
| $\mathbf{E}[X] = \mu$                | $\hat{\mu} = \overline{X} = \frac{X_1 + \dots + X_n}{n}$         |
| $\mathbf{E}[X^2] = \mu^2 + \sigma^2$ | $\hat{\mu}^2 + \hat{\sigma}^2 = \frac{X_1^2 + \dots + X_n^2}{n}$ |
| $\vdots$                             | $\vdots$   |
| $\mathbf{E}[X^k]$                    | $\frac{X_1^k + \dots + X_n^k}{n}$                                |

**Rationale:** The population moments should be close to the sample moments.

#### Example 1.3.5

- Consider  $N(\mu, \sigma^2)$ , where  $\sigma$  is given. Estimate  $\mu$ .

By the method of moment estimate, we have  $\mu_e = \overline{X}$ .

- Consider  $N(\mu, \sigma^2)$ . Estimate  $\mu$  and  $\sigma$ .

We have  $\mu_e = \overline{X}$  and  $\mu_e^2 + \sigma_e^2 = \frac{X_1^2 + \dots + X_n^2}{n}$ .

- Consider  $N(\theta, \sigma^2)$ . Given  $E(X^4) = 3\sigma^4$ , estimate  $\mu$  and  $\sigma$ .

We have  $\mu_e = \bar{X}$ ,  $\mu_e^2 + \sigma_e^2 = \frac{X_1^2 + \cdots + X_n^2}{n}$ , and  $3\sigma^4 = \frac{X_1^4 + \cdots + X_n^4}{n}$ . We have three equations but only two unknowns, then a solution is not guaranteed. So, we need some restrictions on this method (see Remark 1.2).

### Theorem 1.3.6 Method of Moments Estimates

For a random sample of size  $n$  from the discrete (or continuous) population/pdf  $p_X(k; \theta_1, \dots, \theta_s)$  (or  $f_Y(y; \theta_1, \dots, \theta_s)$ ), solutions to the system

$$\begin{cases} E(Y) = \frac{1}{n} \sum_{i=1}^n y_i \\ \vdots \\ E(Y^s) = \frac{1}{n} \sum_{i=1}^n y_i^s \end{cases}$$

which are denoted by  $\theta_{1e}, \dots, \theta_{se}$ , are called the **method of moments estimates** of  $\theta_1, \dots, \theta_s$ .

**Remark 1.2** To estimate  $k$  parameters with the method of moments estimates, we will only match the first  $k$  orders of moments.

**Example 1.3.7** Consider the Gamma distribution:

$$f_Y(y; r, \lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y} \quad \text{for } y \geq 0.$$

Given  $E(Y) = \frac{r}{\lambda}$  and  $E(Y^2) = \frac{r}{\lambda^2} + \frac{r^2}{\lambda^2}$ . Estimate  $r$  and  $\lambda$ .

**Solution 1.**

$$E(Y) = \frac{r}{\lambda} \implies \frac{r_e}{\lambda_e} = \frac{y_1 + \cdots + y_n}{n} = \bar{y} \quad \text{①}$$

$$E(Y^2) = \frac{r}{\lambda^2} + \frac{r^2}{\lambda^2} \implies \frac{r_e}{\lambda_e^2} + \frac{r_e^2}{\lambda_e^2} = \frac{y_1^2 + \cdots + y_n^2}{n} \quad \text{②}$$

Substitute ① into ②, we have

$$\frac{\bar{y}}{\lambda_e} + (\bar{y})^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 \implies \boxed{\lambda_e = \frac{\bar{y}}{\frac{1}{n} \sum y_i^2 - \bar{y}^2}} \quad \text{③}$$

Substitute ③ into ①, we have

$$r_e = \bar{y}\lambda_e = \frac{\bar{y}^2}{\frac{1}{n} \sum y_i^2 - \bar{y}^2}.$$

□

**Remark 1.3** *The sample variance is defined as*

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 &= \frac{1}{n} \sum_{i=1}^n (y_i^2 - 2y_i\bar{y} + \bar{y}^2) \\ &= \frac{1}{n} \sum_{i=1}^n y_i^2 - 2\bar{y} \cdot \frac{\sum y_i}{n} + \frac{1}{n} \cdot n\bar{y}^2 \\ &= \frac{1}{n} \sum_{i=1}^n y_i^2 - 2\bar{y}^2 + \bar{y}^2 \\ &= \frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{y}^2. \end{aligned} \quad \bar{y} = \frac{\sum y_i}{n}$$

So, in Example 1.3.7, if we define  $\hat{\sigma}^2$  to be the sample variance, we can further simplify our estimate as follows:

$$\lambda_e = \frac{\bar{y}}{\hat{\sigma}^2}, \quad r_e = \frac{\bar{y}^2}{\hat{\sigma}^2}.$$

## 1.4 Interval Estimation

**Example 1.4.1** Estimate  $\mu$ , where  $X \sim N(\mu, 1)$ .

We take some samples and compute their sample means:

$$\bar{X}^1 = \frac{x_1 + \cdots + x_n}{n}, \bar{X}^2 = \frac{\tilde{x}_1 + \cdots + \tilde{x}_n}{n}, \dots$$

Finding the distribution of  $\bar{X}$ , we can find an interval  $[\hat{\theta}_L, \hat{\theta}_U]$  such that

$$\mathbf{P}(\hat{\theta}_L \leq \bar{X} \leq \hat{\theta}_U) = 1 - \alpha.$$

**Remark 1.4** *By using the variance of the estimator, one can construct an interval such that with a high probability that the interval contains the unknown parameter.*

**Definition 1.4.2 (Confidence Interval).** The interval,  $[\hat{\theta}_L, \hat{\theta}_U]$  is called the *confidence interval*, and the high probability is  $1 - \alpha$ , where  $\alpha$  is given.

**Remark 1.5** Take  $\alpha = 5\%$ , then  $[\hat{\theta}_L, \hat{\theta}_U]$  is the 95% confidence interval of  $\mu$ . It does not mean that  $\mu$  has 95% chance to be in  $[\hat{\theta}_L, \hat{\theta}_U]$ . However, if we construct 1000 such intervals, 950 of them will contain  $\mu$ .

**Example 1.4.3** A random sample of size 4, ( $Y_1 = 6.5, Y_2 = 9.2, Y_3 = 9.9, Y_4 = 12.4$ ), from a normal population:

$$f_Y(y; \mu) = \frac{1}{\sqrt{2\pi}0.8} e^{-\frac{1}{2}\left(\frac{y-\mu}{0.8}\right)^2} \sim N(\mu, \sigma^2 = 0.64).$$

Both MLE and MME give  $\mu_e = \bar{y} = 9.5$ . The estimator  $\hat{\mu} = \bar{Y}$  follows normal distribution. Construct 95%-confidence interval for  $\mu$ .

**Solution 1.**

$E(\bar{Y}) = \mu$  and  $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n} = \frac{0.64}{4}$ . By the Central Limit Theorem,  $\bar{Y}$  approximately follow  $N\left(\mu, \frac{\sigma^2}{n}\right)$ . So,  $\frac{\bar{Y} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1)$ . Then,

$$\mathbf{P}\left(z_1 \leq \frac{\bar{Y} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq z_2\right) = 0.95 \implies \mathbf{P}\left(\bar{Y} - z_2 \sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{Y} - z_1 \sqrt{\frac{\sigma^2}{n}}\right) = 0.95$$

There are infinite many ways to construct a confidence interval by selecting different  $z_1$  and  $z_2$ . However, since we don't have any prior knowledge on  $\mu$ , it is good for us to choose  $z_1$  and  $z_2$  symmetrically. Moreover, symmetric  $z_1$  and  $z_2$  will yield a smaller interval. We know the symmetric  $z_1, z_2$  pair will be  $z_1 = -1.96$  and  $z_2 = 1.96$ . Therefore,

$$\mathbf{P}\left(\bar{Y} - 1.96 \sqrt{\frac{0.64}{4}} \leq \mu \leq \bar{Y} + 1.96 \sqrt{\frac{0.64}{4}}\right) = 0.95.$$

Then, 95% confidence interval is  $[9.5 - 1.96 \times 0.4, 9.5 + 1.96 \times 0.4]$ . □

#### Theorem 1.4.4 Confidence Interval

In general, for a normal population with  $\sigma$  known, the  $100(1 - \alpha)\%$  *two-sided confidence interval* for  $\mu$  is

$$\left(\bar{y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$$

**Theorem 1.4.5 Variation of Confidence Interval**

- One-sided interval:

$$\left( \bar{y} - z_{\alpha} \frac{\sigma}{\sqrt{n}}, \bar{y} \right) \text{ or } \left( \bar{y}, \bar{y} + z_{\alpha} \frac{\sigma}{\sqrt{n}} \right)$$

- $\sigma$  is unknown and sample size is small:  $z$ -score  $\rightarrow t$ -score.
- $\sigma$  is unknown and sample size is large:  $z$ -score by CLT.
- Non Gaussian population but sample size is large:  $z$ -score by CLT.

**Theorem 1.4.6**

Let  $k$  be the number of successes in  $n$  independent trials, where  $n$  is large and  $p = P(\text{success})$  is unknown. An approximate  $100(1 - \alpha)\%$  confidence interval for  $p$  is the set of numbers

$$\left( \frac{k}{n} - z_{\alpha/2} \sqrt{\frac{(k/n)(1 - k/n)}{n}}, \frac{k}{n} + z_{\alpha/2} \sqrt{\frac{(k/n)(1 - k/n)}{n}} \right).$$

**Definition 1.4.7 (Margin of Error).** The *margin of error*, denoted by  $d$ , is the quantity

$$d = z_{\alpha/2} \sqrt{\frac{(k/n)(1 - k/n)}{n}}.$$

**Remark 1.6** Stating the sample mean and the margin of error is equivalent to stating the confidence interval. Note that C.I. =  $\hat{p} \pm d$ .

**Theorem 1.4.8 Estimate Margin of Error**

When  $p$  is close to  $\frac{1}{2}$ , then  $d \approx d_m = \frac{z_{\alpha/2}}{2\sqrt{n}}$ , which is equivalent to  $\sigma_n \approx \frac{1}{2\sqrt{n}}$ . However, if  $p$  is away from  $\frac{1}{2}$ ,  $d$  and  $d_m$  are very different.

**Remark 1.7** Theorem 1.4.8 gives a conservative estimation of the margin of error, which is  $d_m$ .

**Proposition 1.9 :** Given  $d$ , we can estimate the sample size.

**Proof 2.**

$$d = z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \implies n \approx \hat{p}(1 - \hat{p}) / \left( \frac{d}{z_{\alpha/2}} \right)^2.$$

However, since  $n$  is unknown,  $\hat{p}$  is also unknown. We, therefore, need information on the actual  $p$  to conclude an estimation of the sample size.

- If  $p$  is known,

$$n = \frac{p(1-p)}{\left(\frac{d}{z_{\alpha/2}}\right)^2}.$$

- If  $p$  is unknown. Let  $f(p) = p(1-p)$ .  $f$  will be maximized when  $p = 0.5$ . So,  $f(p) = p(1-p) \leq 0.25$ . Then,

$$n \leq \frac{0.25}{\left(\frac{d}{z_{\alpha/2}}\right)^2}.$$

Since we are conservative, take  $n = \frac{\frac{1}{4}z_{\alpha/2}^2}{d^2} = \frac{z_{\alpha/2}^2}{4d^2}$ . This estimation is a conservative estimation of the sample size. ■

## 1.5 Properties of Estimation

The main question is that estimators are not unique in general. How do we choose a good estimator?

**Definition 1.5.1 (Unbiasedness).** Given a random sample of size  $n$  when whose population distribution depends on an unknown parameter  $\theta$ . Let  $\hat{\theta}$  be an estimator of  $\theta$ . Then,

- $\hat{\theta}$  is called *unbiased* if  $\mathbf{E}(\hat{\theta}) = \theta$ .
- $\hat{\theta}$  is called *asymptotically unbiased* if  $\lim_{n \rightarrow \infty} \mathbf{E}(\hat{\theta}) = \theta$ .
- If  $\theta$  is biased, then the *bias* is given by the quantity  $\mathbf{B}(\hat{\theta}) = \mathbf{E}(\hat{\theta}) - \theta$ .

**Example 1.5.2** Consider the exponential distribution:  $f_Y(y; \lambda) = \lambda e^{-\lambda y}$  for  $y \geq 0$ . Determine if the estimator  $\hat{\lambda} = \frac{1}{\bar{Y}}$  is biased or not.

*Hint:*  $n\bar{Y} = \sum_{i=1}^n Y_i \sim \text{Gamma}(n, \lambda)$ .

**Solution 1.**

Recall that  $\mathbf{E}[g(x)] = \int_x g(x) f_X(x) dx$ . Define  $X = \sum_{i=1}^n Y_i \sim \text{Gamma}(n, \lambda)$ . Also, recall the following facts:

$$\Gamma(n) = (n-1)! = (n-1)\Gamma(n-1)$$

and the integration over any probability density function will yield a result of 1 by definition.

Then,

$$\begin{aligned}
 \mathbf{E}(\hat{\lambda}) &= \mathbf{E}\left(\frac{1}{\bar{Y}}\right) = \mathbf{E}\left(\frac{n}{\sum Y_i}\right) = n\mathbf{E}\left(\frac{1}{\sum Y_i}\right) \\
 &= n\mathbf{E}\left(\frac{1}{X}\right) \\
 &= n \int_x \frac{1}{x} \cdot \frac{\lambda^n}{\Gamma(n)} x^{n-1} e^{-\lambda x} dx \\
 &= n \int_x \frac{\lambda^n}{(n-1)!} x^{n-2} e^{-\lambda x} dx \\
 &= \frac{n\lambda}{(n-1)} \underbrace{\int_x \frac{\lambda^{n-1}}{\Gamma(n-1)} x^{n-2} e^{-\lambda x} dx}_{=1} \\
 &= \frac{n}{n-1} \lambda.
 \end{aligned}$$

Therefore,  $\mathbf{E}(\hat{\lambda}) \neq \lambda$ , and so  $\hat{\lambda}$  is biased. However, note that

$$\lim_{n \rightarrow \infty} \mathbf{E}(\hat{\lambda}) = \lim_{n \rightarrow \infty} \frac{n}{n-1} \lambda = \lambda.$$

By definition, then  $\hat{\lambda}$  is asymptotically unbiased. □

**Example 1.5.3** Consider the exponential distribution  $f(y; \theta) = \frac{1}{\theta} e^{-y/\theta}$  for  $y \geq 0$ . Then,  $\hat{\theta} = \bar{Y}$  is unbiased.

**Remark 1.8** Suppose  $\{X_1, \dots, X_n\}$  are i.i.d. random variables, and  $\mathbf{E}(X_i) = \mu$  for  $i = 1, \dots, n$ . Then,  $\bar{X}$ , the sample mean, is always an unbiased estimator:

$$\mathbf{E}(\bar{X}) = \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n \mathbf{E}(X_i) = \frac{1}{n} \cdot n \cdot \mu = \mu.$$

#### Theorem 1.5.4 Sample Variance is Biased

Suppose  $\{X_1, \dots, X_n\}$  are i.i.d. random variables, and  $\mathbf{E}(X_i) = \mu$ ,  $\text{Var}(X_i) = \sigma^2$  for  $i = 1, \dots, n$ . Then, the sample variance  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$  is biased.



**Proof2.** Note that

$$\begin{aligned}
 \mathbf{E}(\hat{\sigma}^2) &= \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right) \\
 &= \mathbf{E}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu + \mu - \bar{X})^2\right) \\
 &= \frac{1}{n} \sum_{i=1}^n \mathbf{E}\left[(X_i - \mu)^2 + (\mu - \bar{X})^2 + 2(X_i - \mu)(\mu - \bar{X})\right] \\
 &= \frac{1}{n} \sum_{i=1}^n \left\{ \underbrace{\mathbf{E}(X_i - \mu)^2}_{\text{Var}(X_i)} + \mathbf{E}(\mu - \bar{X})^2 + 2\mathbf{E}[(\mu - \bar{X})(X_i - \mu)] \right\} \\
 &\quad \left| \begin{array}{l} \text{Hint: } \frac{1}{n} \sum_{i=1}^n (X_i - \mu) = \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n \mu = \bar{X} - \mu \end{array} \right. \\
 &= \frac{1}{n} \sum_{i=1}^n \text{Var}(X_i) + \frac{1}{n} \cdot n \mathbf{E}(\mu - \bar{X})^2 + 2\mathbf{E}\left[(\mu - \bar{X}) \frac{1}{n} \sum_{i=1}^n (X_i - \mu)\right] \\
 &= \frac{1}{n} \sum_{i=1}^n \sigma^2 + \mathbf{E}(\mu - \bar{X})^2 + 2\mathbf{E}[(\mu - \bar{X})(\bar{X} - \mu)] \\
 &= \frac{1}{n} \cdot n \cdot \sigma^2 + \mathbf{E}(\mu - \bar{X})^2 - 2\mathbf{E}[(\mu - \bar{X})^2] \\
 &= \sigma^2 - \mathbf{E}(\mu - \bar{X})^2 \\
 &= \sigma^2 - \underbrace{\mathbf{E}(\bar{X} - \mu)^2}_{=\text{Var}(\bar{X})} \\
 &= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2 \neq \sigma^2
 \end{aligned}$$

Therefore,  $\hat{\sigma}^2$  is not an unbiased estimator. ■

### Theorem 1.5.5 Adjusted Sample Variance is Unbiased

With the same set up in Theorem 1.5.4, define the adjusted sample variance to be

$$S^2 = \frac{n}{n-1} \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Then,  $S^2$  is an unbiased estimator of  $\sigma^2$ .

**Definition 1.5.6 (Decision Theory).** Minimize the error of an estimator (sample statistics) relative to the true parameter (population parameter) using a loss function.

**Definition 1.5.7 (Mean Squared Error).** The *mean squared error* (MSE) is defined by

$$\text{MSE}(\hat{\theta}) = \mathbf{E}[(\hat{\theta} - \theta)^2]$$

**Theorem 1.5.8 Decomposition of MSE**

Generally,

$$\text{MSE}(\theta) = \text{Var}(\hat{\theta}) + \mathbf{B}(\hat{\theta})^2$$

If  $\hat{\theta}$  is unbiased,  $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta})$ .  $\text{Var}(\theta)$  measures the precision of the estimator.

**Proof 3.** Note that we will the following:

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= \mathbf{E}[(\hat{\theta} - \theta)^2] \\ &= \mathbf{E}(\hat{\theta}^2 + \theta^2 - 2\hat{\theta}\theta) \\ &= \mathbf{E}(\hat{\theta}^2) - 2\theta\mathbf{E}(\hat{\theta}) + \theta^2 \\ &= \underbrace{\mathbf{E}(\hat{\theta}^2) - \mathbf{E}(\hat{\theta})^2}_{\text{Var}(\hat{\theta})} + \underbrace{\mathbf{E}(\hat{\theta})^2 - 2\theta\mathbf{E}(\hat{\theta}) + \theta^2}_{[\mathbf{E}(\hat{\theta}) - \theta]^2} \\ &= \text{Var}(\hat{\theta}) + [\mathbf{E}(\hat{\theta}) - \theta]^2 \\ &= \text{Var}(\theta) + \mathbf{B}(\hat{\theta})^2 \end{aligned}$$

If  $\hat{\theta}$  is unbiased,  $\mathbf{B}(\hat{\theta}) = 0$ , and so  $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta})$ . ■

**Definition 1.5.9 (Efficiency).** Let  $\hat{\theta}_1$  and  $\hat{\theta}_2$  be two unbiased estimators for a parameter  $\theta$ . If we have  $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$ , then we say that  $\hat{\theta}_1$  is *more efficient* than  $\hat{\theta}_2$ . The *relative efficiency* of  $\hat{\theta}_1$  with respect to  $\hat{\theta}_2$  is the ratio  $\frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)}$ .

## 1.6 Best Unbiased Estimator

**Definition 1.6.1 (Best/Minimum-Variance Estimator).** Let  $\Theta$  be the set of all estimators  $\hat{\theta}$  that are unbiased for the parameter  $\theta$ . We say that  $\hat{\theta}^*$  is a *best* or *minimum-variance estimator* (MVE) if  $\hat{\theta}^* \in \Theta$  and  $\text{Var}(\hat{\theta}^*) \leq \text{Var}(\hat{\theta}) \quad \forall \hat{\theta} \in \Theta$ .

**Definition 1.6.2 (Fisher's Information).** The *Fisher's information* of a continuous random variable  $Y$  with pdf  $f_Y(y; \theta)$  is defined as

$$\mathbf{I}(\theta) = \mathbf{E} \left[ \left( \frac{\partial \ln f_Y(y; \theta)}{\partial \theta} \right)^2 \right] = -\mathbf{E} \left[ \frac{\partial^2}{\partial \theta^2} \ln f_Y(y; \theta) \right].$$

**Remark 1.9** The Fisher's information measures the amount of information that a sample  $Y$  contains about the unknown parameter  $\theta$ . If  $\mathbf{I}(\theta)$  is big, then the curvature of  $f_Y(y; \theta)$  is big, and

thus it is more likely that we can find a region where  $\hat{\theta}$  is concentrated.

**Extension 1.1 (Joint Fisher's Information)** Suppose  $Y_1, \dots, Y_n$  are continuous i.i.d. random variables, each has a Fisher's information of  $\mathbf{I}(\theta)$ . Then,

$$\mathbf{E} \left[ \left( \frac{\partial}{\partial \theta} \ln f_{Y_1, \dots, Y_n}(y_1, \dots, y_n; \theta) \right)^2 \right] = n\mathbf{I}(\theta).$$

**Theorem 1.6.3 Properties of Fisher's Information**

Define the *Fisher's Score Function*  $\frac{\partial}{\partial \theta} \ln f_Y(y; \theta)$ . Then,

$$\mathbf{E}_Y \left[ \frac{\partial}{\partial \theta} \ln f_Y(y; \theta) \right] = 0.$$

**Proof 1.** Note that by chain rule, we have

$$\begin{aligned} \mathbf{E}_Y \left[ \frac{\partial}{\partial \theta} \ln f_Y(y; \theta) \right] &= \int_Y \left( \frac{\partial}{\partial \theta} \ln f_Y(y; \theta) \right) f_Y(y; \theta) \, dy \\ &= \int_Y \frac{1}{f_Y(y; \theta)} \left( \frac{\partial}{\partial \theta} f_Y(y; \theta) \right) f_Y(y; \theta) \, dy \\ &= \int_Y \frac{\partial}{\partial \theta} f_Y(y; \theta) \, dy \\ &= \frac{\partial}{\partial \theta} \int_Y f_Y(y; \theta) \, dy = \frac{\partial}{\partial \theta} (1) = 0. \end{aligned}$$

■

**Corollary 1.4 :**

$$\mathbf{I}(\theta) = \mathbf{Var} \left( \frac{\partial}{\partial \theta} \ln f_Y(y; \theta) \right).$$

**Proof 2.** By definition, we have

$$\begin{aligned} \mathbf{Var} \left( \frac{\partial}{\partial \theta} \ln f_Y(y; \theta) \right) &= \mathbf{E} \left[ \left( \frac{\partial}{\partial \theta} \ln f_Y(y; \theta) \right)^2 \right] - \underbrace{\left( \mathbf{E} \left( \frac{\partial}{\partial \theta} \ln f_Y(y; \theta) \right) \right)^2}_{=0, \text{ by Theorem 1.6.3.}} \\ &= \mathbf{E} \left[ \left( \frac{\partial}{\partial \theta} \ln f_Y(y; \theta) \right)^2 \right] \\ &= \mathbf{I}(\theta). \end{aligned}$$

■

**Theorem 1.6.5 Cramér-Rao Inequality**

Under regular condition, let  $Y_1, \dots, Y_n$  be a random sample of size  $n$  form the continuous population pdf  $f_Y(y; \theta)$ . Let  $\hat{\theta} = \hat{\theta}(Y_1, \dots, Y_n)$  be any unbiased estimator for  $\theta$ . Then,

$$\text{Var}(\hat{\theta}) \geq \frac{1}{n\mathbf{I}(\theta)}.$$

**Remark 1.10** A similar statement holds for the discrete case  $p_X(k; \theta)$ .

**Definition 1.6.6 (Efficiency of Unbiased Estimator).** An unbiased estimator  $\hat{\theta}$  is *efficient* if  $\text{Var}(\hat{\theta})$  is equal to the Cramér-Rao lower bound. That is,  $\text{Var}(\hat{\theta}) = (n\mathbf{I}(\theta))^{-1}$ . Such an estimator is the MVE defined in Definition 1.6.1. The *efficiency* of an unbiased estimator  $\hat{\theta}$  is defined to be the quantity

$$\left(n\mathbf{I}(\theta)\text{Var}(\hat{\theta})\right)^{-1}.$$

**Example 1.6.7** Suppose  $X \sim \text{Bernoulli}(p)$ . Is  $\hat{p} = \bar{X}$  efficient?

**Solution 3.**

Note that we have the following

$$\begin{aligned} f_X(x; p) &= p^x(1-p)^{1-x}, \quad x = 0, 1 \\ \ln f_X(x; p) &= x \ln p + (1-x) \ln(1-p) \\ \frac{\partial}{\partial p} \ln f_X(x; p) &= \frac{x}{p} - \frac{1-x}{1-p} \\ \frac{\partial^2}{\partial p^2} \ln f_X(x; p) &= -\frac{x}{p^2} - \frac{1-x}{(1-p)^2} \end{aligned}$$

Therefore, the Fisher's information can be computed by

$$\begin{aligned} \mathbf{I}(p) &= -\mathbf{E} \left[ \frac{\partial^2}{\partial p^2} \ln f_X(x; p) \right] = -\mathbf{E} \left[ -\frac{x}{p^2} - \frac{1-x}{(1-p)^2} \right] \\ &= \mathbf{E} \left[ \frac{x}{p^2} \right] + \mathbf{E} \left[ \frac{1-x}{(1-p)^2} \right] \\ &= \frac{\mathbf{E}(x)}{p^2} + \frac{1 - \mathbf{E}(x)}{(1-p)^2} \\ &= \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}. \end{aligned}$$

Note that

$$\text{Var}(\bar{X}) = \text{Var} \left( \frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n} \text{Var}(X_i) = \frac{1}{n} \cdot p(1-p).$$

So, we have

$$\text{Var}(\bar{X}) = \frac{p(1-p)}{n} = \frac{1}{n \left( \frac{1}{p(1-p)} \right)} = \frac{1}{n\mathbf{I}(p)}.$$

Therefore,  $\hat{p}$  is efficient. □

**Example 1.6.8** Suppose  $X \sim N(\mu, \sigma^2)$ , with  $\sigma^2$  is known. What is  $\mathbf{I}(\mu)$ ?

**Solution 4.**

Note that

$$\frac{d^2}{d\mu^2} \ln f_X(x; \mu) = -\frac{1}{\sigma^2}.$$

Then,

$$\mathbf{I}(\mu) = -\mathbf{E} \left[ \frac{d^2}{d\mu^2} \ln f_X(x; \mu) \right] = -\mathbf{E} \left[ -\frac{1}{\sigma^2} \right] = \frac{1}{\sigma^2}.$$

□

## 1.7 Sufficiency

**Remark 1.11** Use Likelihood Function to Define Fisher's Information

- We can define the score function as  $\frac{\partial \ln \mathbf{L}(Y_1, \dots, Y_n; \theta)}{\partial \theta} = 0 \implies \text{MLE}.$
- $\mathbf{E} \left[ \frac{\partial \ln \mathbf{L}(Y; \theta)}{\partial \theta} \right] = 0$
- $\mathbf{I}(\theta) = \mathbf{E} \left[ \left( \frac{\partial \ln \mathbf{L}(Y; \theta)}{\partial \theta} \right)^2 \right] = -\mathbf{E}_Y \left[ \frac{\partial^2 \ln \mathbf{L}(Y; \theta)}{\partial \theta^2} \right]$
- $-\mathbf{E}_Y \left[ \frac{\partial^2 \ln \mathbf{L}(Y_1, \dots, Y_n; \theta)}{\partial \theta^2} \right] = n\mathbf{I}(\theta).$

**Proof 1.**

$$\begin{aligned} -\mathbf{E}_Y \left[ \frac{\partial^2 \ln \mathbf{L}(Y_1, \dots, Y_n; \theta)}{\partial \theta^2} \right] &= -\mathbf{E}_Y \left[ \frac{\partial^2}{\partial \theta^2} \ln \mathbf{L}(Y_1, \dots, Y_n; \theta) \right] \\ &= -\mathbf{E}_Y \left[ \frac{\partial^2}{\partial \theta^2} \ln \left( \prod_{i=1}^n f_Y(Y_i; \theta) \right) \right] \\ &= -\mathbf{E}_Y \left[ \frac{\partial^2}{\partial \theta^2} \sum_{i=1}^n \ln f_Y(y_i; \theta) \right] = \sum_{i=1}^n \left( -\mathbf{E}_Y \left[ \frac{\partial^2}{\partial \theta^2} \ln f_Y(y_i; \theta) \right] \right) = n\mathbf{I}(\theta) \end{aligned}$$

■

- $\widehat{\theta}_{MLE} \xrightarrow{n \rightarrow \infty} N\left(\theta, \frac{1}{\mathbf{I}(\theta)}\right)$ . Note that  $\frac{1}{\mathbf{I}(\theta)}$  is the C-R lower bound. We see that  $\widehat{\theta}_{MLE}$  is asymptotically efficient.

**Remark 1.12 (Sufficiency Intuition)** Sufficiency tells us how much information can we get out of the data.

**Rationale** Let  $\widehat{\theta}$  be an estimator to the unknown parameter  $\theta$ . Does  $\widehat{\theta}$  contain all information about  $\theta$ ? e.g., The data itself is a sufficient estimator.

**Definition 1.7.1 (Sufficiency).** Let  $(X_1, \dots, X_n)$  be a random sample of size  $n$  from a continuous population with an unknown parameter  $\theta$ . We call  $\theta$  is *sufficient* if

$$f_{Y_1, \dots, Y_n | \widehat{\theta}}(Y_1, \dots, Y_n | \widehat{\theta} = \theta_e) = b(y_1, \dots, y_n),$$

where  $b(y_1, \dots, y_n)$  is independent of  $\theta$  ( $\perp \theta$ ). Also,  $\widehat{\theta} = h(Y_1, \dots, Y_n)$  and  $\theta_e = h(y_1, \dots, y_n)$ . In this case,  $\widehat{\theta}$  contains all the information about  $\theta$  from  $\{y_1, \dots, y_n\}$ .

### Example 1.7.2

- Toss a coin 5 times and get 3 heads. Estimate  $p =$  probability of  $H$ .

**Solution 2.**

$$\mathbf{P}\left(HHHTT \mid p_e = \frac{3}{5}\right) = \frac{1}{\binom{5}{3}} \perp p \implies \text{sufficient}$$

□

- A random sample of size  $n$  from Bernoulli( $p$ ). Check the sufficiency of  $p = \sum_{i=1}^n X_i$ .

**Solution 3.**

Suppose the random sample is  $\{X_1, \dots, X_n\}$ . Then, consider

$$\mathbf{P}(X_1 = x_1, \dots, X_n = x_n, \sum_{i=1}^n X_i = C \mid \widehat{p} = C) = \frac{\mathbf{P}(X_1 = x_1, \dots, X_n = x_n, \sum_{i=1}^n X_i = C)}{\mathbf{P}(\widehat{p} = C)}.$$

What new information can  $\sum_{i=1}^n X_i = C$  tell us?  $X_n = C - \sum_{i=1}^{n-1} X_i$ .

Note that  $\mathbf{P}(\hat{p} = C) = \mathbf{P}\left(\sum_{i=1}^n X_i = C\right)$ . Since the summation of Bernoulli( $p$ ) random variables is a Binomial( $n, p$ ) random variable, we have  $\mathbf{P}(\hat{p} = C) = \binom{n}{C} p^C (1-p)^{n-C}$ .

**Case I** Suppose  $\sum_{i=1}^n X_i = C$ . Then,

$$\begin{aligned} & \frac{\mathbf{P}(X_1 = x_1, \dots, X_n = x_n, \sum_{i=1}^n X_i = C)}{\mathbf{P}(\hat{p} = C)} \\ &= \frac{\left(\prod_{i=1}^{n-1} p^{X_i} (1-p)^{1-X_i} p^{C - \sum_{i=1}^{n-1} X_i} (1-p)^{\left(1-C + \sum_{i=1}^{n-1} X_i\right)}\right)}{\binom{n}{C} p^C (1-p)^{n-C}} \\ &= \frac{p^{\sum_{i=1}^{n-1} X_i + C - \sum_{i=1}^{n-1} X_i} (1-p)^{(n-1) - \sum_{i=1}^{n-1} X_i + 1 - C + \sum_{i=1}^{n-1} X_i}}{\binom{n}{C} p^C (1-p)^{n-C}} \\ &= \frac{p^C (1-p)^{n-C}}{\binom{n}{C} p^C (1-p)^{n-C}} = \frac{1}{\binom{n}{C}} \perp\!\!\!\perp p \implies \text{ sufficient} \end{aligned}$$

**Case II** Suppose  $\sum_{i=1}^n X_i \neq C$ . Then,

$$\frac{\mathbf{P}(X_1 = x_1, \dots, X_n = x_n, \sum_{i=1}^n X_i = C)}{\mathbf{P}(\hat{p} = C)} = \frac{0}{\mathbf{P}(\hat{p} = C)} = 0 \perp\!\!\!\perp p \implies \text{ sufficient}$$

□

### Theorem 1.7.3 Factorization Property

$\hat{\theta}$  is sufficient if and only if the likelihood can be factorized as

$$\mathbf{L}(\theta) = \underbrace{g(\theta_e; \theta)}_{\theta_e \text{ \& } \theta} \cdot \underbrace{u(y_1, \dots, y_n)}_{\perp\!\!\!\perp \theta}$$

## 1.8 Consistency

**Definition 1.8.1 (Consistency).** An estimator  $\hat{\theta}_n = h(W_1, \dots, W_n)$  is said to be *consistent* if it converges to  $\theta$  in probability; i.e., for any  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbf{P}\left(\left|\hat{\theta}_n - \theta\right| < \varepsilon\right) = 1.$$

**Remark 1.13** 1. Consistency is an asymptotical property (defined in a large sample limit).

2.  $n =$  sample size.  $\left|\hat{\theta}_n - \theta\right|$  is the distance between estimator and true  $\theta$ .

**Lemma 1.2 Markov Inequality:** Suppose  $X \geq 0$  is a random variable and  $a > 0$  is a constant. Then,

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}(X)}{a}.$$

**Remark 1.14** Markov inequality is good for determining extreme values. If  $\mathbf{E}(X)$  is small, then it is very unlikely that  $X$  will take some extremely large numbers.

### Theorem 1.8.3 Chebyshev Inequality

Let  $W$  be some random variable with finite mean  $\mu$  and variance  $\sigma^2$ . Then, for any  $\varepsilon > 0$ , we have

$$\mathbf{P}(|W - \mu| < \varepsilon) \leq 1 - \frac{\sigma^2}{\varepsilon^2}$$

or, equivalently,

$$\mathbf{P}(|W - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

**Proof 1.** Consider the random variable  $|W - \mu|$ . Then, by Markov Inequality,

$$\begin{aligned} \mathbf{P}(|W - \mu| \geq \varepsilon) &= \mathbf{P}(|W - \mu|^2 \geq \varepsilon^2) \\ &= \mathbf{P}((W - \mu)^2 \geq \varepsilon^2) \leq \frac{\mathbf{E}[(W - \mu)^2]}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2} \end{aligned}$$

■

**Corollary 1.4 :** The sample mean  $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n W_i$  is a consistent estimator for  $\mathbf{E}(W) = \mu$ , provided that the population  $W$  has finite mean  $\mu$  and variance  $\sigma^2$ .

**Proposition 1.5 :** If  $\hat{\theta}_n$  is an unbiased estimator of  $\theta$ , then  $\hat{\theta}_n$  is consistent if

$$\lim_{n \rightarrow \infty} \mathbf{Var}(\hat{\theta}_n) = 0.$$

**Proof 2.** Suppose  $\hat{\theta}_n$  is an unbiased estimator of  $\theta$ . Then,  $\mathbf{E}(\hat{\theta}_n) = \theta$ . So, by Chebyshev



Inequality, we have

$$\mathbf{P}\left(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right) = \mathbf{P}\left(\left|\hat{\theta}_n - \mathbf{E}\left(\hat{\theta}_n\right)\right| \geq \varepsilon\right) \leq \frac{\mathbf{E}\left[\left(\hat{\theta}_n - \mathbf{E}\left(\hat{\theta}_n\right)\right)^2\right]}{\varepsilon^2} = \frac{\mathbf{Var}\left(\hat{\theta}_n\right)}{\varepsilon^2}.$$

If we have  $\mathbf{Var}\left(\hat{\theta}_n\right) \rightarrow 0$  when  $n \rightarrow \infty$ , then

$$\lim_{n \rightarrow \infty} \mathbf{P}\left(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right) \leq \lim_{n \rightarrow \infty} \frac{\mathbf{Var}\left(\hat{\theta}_n\right)}{\varepsilon^2} = \frac{0}{\varepsilon} = 0.$$

Therefore, it must be that  $\lim_{n \rightarrow \infty} \mathbf{P}\left(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right) = 0$  as probability cannot take negative values. Hence,

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbf{P}\left(\left|\hat{\theta}_n - \theta\right| < \varepsilon\right) &= \lim_{n \rightarrow \infty} \left(1 - \mathbf{P}\left(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right)\right) \\ &= 1 - \lim_{n \rightarrow \infty} \mathbf{P}\left(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right) \\ &= 1 - 0 = 1. \end{aligned}$$

Then, by definition,  $\hat{\theta}_n$  is consistent. ■