Proof of (23): By the central limit theorem (11) on p.254,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(X_i;\theta)}{\partial\theta} \longrightarrow N\Big(0, \frac{1}{\sigma^2(\theta)}\Big), \qquad (26)$$

where $\sigma^2(\theta)$ is defined in (14), i.e., $\mathrm{V}\Big(\frac{\partial \log f(X_i;\theta)}{\partial\theta}\Big) = \frac{1}{\sigma^2(\theta)}$.

Note that $\mathrm{E}\Big(\frac{\partial \log f(X_i;\theta)}{\partial\theta}\Big) = 0$.

Apply the central limit theorem, taking $\frac{\partial \log f(X_i;\theta)}{\partial\theta}$ as the ith random variable.

By the Taylor series expansion around $\hat{\theta}_n = \theta$,

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(X_i;\hat{\theta}_n)}{\partial\theta} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(X_i;\theta)}{\partial\theta} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i;\theta)}{\partial\theta^2}(\hat{\theta}_n-\theta) \\ &\quad + \frac{1}{2!}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial^3 \log f(X_i;\theta)}{\partial\theta^3}(\hat{\theta}_n-\theta)^2 + \cdots \end{aligned}$$

The third and above terms in the right-hand side are:

$$\frac{1}{2!}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial^3 \log f(X_i;\theta)}{\partial\theta^3}(\hat{\theta}_n-\theta)^2 + \cdots \longrightarrow 0.$$

It can be shown that the sum of the above terms is equal to $O(n^{-1/2})$.

Note that $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3 \log f(X_i;\theta)}{\partial\theta^3} \longrightarrow \mathrm{E}\Big(\frac{\partial^3 \log f(X_i;\theta)}{\partial\theta^3}\Big)$ from Chebyshev's inequality.

In addition, for now, we consider $\sqrt{n}(\hat{\theta}_n-\theta)^2 \longrightarrow 0$ as $n \longrightarrow \infty$. Actually, we obtain $\sqrt{n}(\hat{\theta}_n-\theta)^2 = O(n^{-1/2})$ from $\hat{\theta}_n - \theta = O(n^{-1/2})$.

Therefore,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(X_i;\theta)}{\partial\theta} \approx -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i;\theta)}{\partial\theta^2}(\hat{\theta}_n-\theta)$$

which implies that the asy. dist. of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(X_i;\theta)}{\partial\theta}$ is equivalent to that of $-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i;\theta)}{\partial\theta^2}(\hat{\theta}_n-\theta)$.

From (26) and the above equations, we obtain:

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i;\theta)}{\partial\theta^2}\sqrt{n}(\hat{\theta}_n-\theta) \longrightarrow N\Big(0, \frac{1}{\sigma^2(\theta)}\Big).$$

The law of large numbers indicates as follows:

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i;\theta)}{\partial\theta^2} \longrightarrow -\mathrm{E}\Big(\frac{\partial^2 \log f(X_i;\theta)}{\partial\theta^2}\Big) = \frac{1}{\sigma^2(\theta)},$$

where the last equality comes from (14).

Thus, we have the following relationship:

$$\begin{aligned} -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i;\theta)}{\partial\theta^2}\sqrt{n}(\hat{\theta}_n-\theta) &\longrightarrow \frac{1}{\sigma^2(\theta)}\sqrt{n}(\hat{\theta}_n-\theta) \\ &\longrightarrow N\Big(0, \frac{1}{\sigma^2(\theta)}\Big) \end{aligned}$$

Therefore, the asymptotic normality of the maximum likelihood estimator is obtained as follows:

$$\sqrt{n}(\hat{\theta}_n-\theta) \longrightarrow N(0, \sigma^2(\theta)).$$

Thus, (23) is obtained.

7.5.2 Least Squares Estimation Method (最小二乗法)

$X_1, X_2, \cdots, X_n$ are mutually independently distributed with mean μ.

$x_1, x_2, \cdots, x_n$ are generated from $X_1, X_2, \cdots, X_n$, respectively.

Solve the following problem:

$$\min_{\mu} S(\mu), \qquad \text{where} \quad S(\mu) = \sum_{i=1}^{n} (x_i - \mu)^2.$$

Let $\hat{\mu}$ be the least squares estimate of μ.

$$\frac{\mathrm{d}S(\mu)}{\mathrm{d}\mu} = 0 \quad \Longrightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The least squares estimator is given by:

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

which is equivalent to MLE.

7.5.3 Method of Moment (積率法)

The distribution of X_i is $f(x;\theta)$.

Let μ'_k be the kth moment.

From the definition of the kth moment,

$$\mathrm{E}(X^k) = \mu'_k$$

where μ'_k depends on θ.

Let $\hat{\mu}'_k$ be the estimate of the kth moment.

$$\mathrm{E}(X^k) \approx \frac{1}{n}\sum_{i=1}^{n} x_i^k = \hat{\mu}'_k$$

The estimator of μ'_k is:

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

Example: $\theta = (\mu, \sigma^2)$: Because we have two parameters, we use the 1st and 2nd moments.

$\mu'_1 = \mathrm{E}(X) = \mu$

$\mu'_2 = \mathrm{E}(X^2) = \mathrm{V}(X) + (\mathrm{E}(X))^2 = \sigma^2 + \mu^2$

Estimates:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \qquad \hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

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

Estimators:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

7.6 Interval Estimation

In Sections 7.1 – 7.5.1, the point estimation is discussed.

It is important to know where the true parameter value of θ is likely to lie.

Suppose that the population distribution is given by $f(x;\theta)$.

Using the random sample $X_1, X_2, \cdots, X_n$ drawn from the population distribution, we construct the two statistics, say, $\theta_U(X_1, X_2, \cdots, X_n)$ and $\theta_L(X_1, X_2, \cdots, X_n)$, where

$$P(\theta_L(X_1, X_2, \cdots, X_n) < \theta < \theta_U(X_1, X_2, \cdots, X_n)) = 1 - \alpha. \tag{27}$$

(27) implies that θ lies on the interval $\big(\theta_L(X_1, X_2, \cdots, X_n), \theta_U(X_1, X_2, \cdots, X_n)\big)$ with probability $1 - \alpha$.

Now, we replace the random variables $X_1, X_2, \cdots, X_n$ by the experimental values $x_1, x_2, \cdots, x_n$.

Then, we say that the interval:

$$\big(\theta_L(x_1, x_2, \cdots, x_n),\ \theta_U(x_1, x_2, \cdots, x_n)\big)$$

is called the $100 \times (1 - \alpha)\%$ **confidence interval (信頼区間)** of θ.

Thus, estimating the interval is known as the **interval estimation (区間推定)**, which is distinguished from the point estimation.

In the interval, $\theta_L(x_1, x_2, \cdots, x_n)$ is known as the **lower bound** of the confidence interval, while $\theta_U(x_1, x_2, \cdots, x_n)$ is the **upper bound** of the confidence interval.

Given probability α, the $\theta_L(X_1, X_2, \cdots, X_n)$ and $\theta_U(X_1, X_2, \cdots, X_n)$ which satisfies equation (27) are not unique.

For estimation of the unknown parameter θ, it is more optimal to minimize the width of the confidence interval.

Therefore, we should choose θ_L and θ_U which minimizes the width $\theta_U(X_1, X_2, \cdots, X_n) - \theta_L(X_1, X_2, \cdots, X_n)$.

Interval Estimation of $\overline{X}$: Let $X_1, X_2, \cdots, X_n$ be mutually independently and identically distributed random variables.

X_i has a distribution with mean μ and variance σ^2.

From the central limit theorem,

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \longrightarrow N(0, 1).$$

Replacing σ^2 by its estimator S^2 (or S^{**2}),

$$\frac{\overline{X} - \mu}{S / \sqrt{n}} \longrightarrow N(0, 1).$$

Therefore, when n is large enough,

$$P(z^* < \frac{\overline{X} - \mu}{S / \sqrt{n}} < z^{**}) = 1 - \alpha,$$

where z^* and z^{**} ($z^* < z^{**}$) are percent points from the standard normal density function.

Solving the inequality above with respect to μ, the following expression is obtained.

$$P\Big(\overline{X} - z^{**}\frac{S}{\sqrt{n}} < \mu < \overline{X} - z^{*}\frac{S}{\sqrt{n}}\Big) = 1 - \alpha,$$

where $\hat{\theta}_L$ and $\hat{\theta}_U$ correspond to $\overline{X} - z^{**}\dfrac{S}{\sqrt{n}}$ and $\overline{X} - z^{*}\dfrac{S}{\sqrt{n}}$,

respectively.

The length of the confidence interval is given by:

$$\hat{\theta}_U - \hat{\theta}_L = \frac{S}{\sqrt{n}}(z^{**} - z^*),$$

which should be minimized subject to:

$$\int_{z^*}^{z^{**}} f(x)\,\mathrm{d}x = 1 - \alpha,$$

i.e.,

$$F(z^{**}) - F(z^*) = 1 - \alpha,$$

where $F(\cdot)$ denotes the standard normal cumulative distribution function.

Solving the minimization problem above, we can obtain the conditions that $f(z^*) = f(z^{**})$ for $z^* < z^{**}$ and that $f(x)$ is symmetric.

Therefore, we have:

$$-z^* = z^{**} = z_{\alpha/2},$$

where $z_{\alpha/2}$ denotes the $100 \times \alpha/2$ percent point from the standard normal density function.

Accordingly, replacing the estimators $\overline{X}$ and S^2 by their estimates $\overline{x}$ and s^2, the $100 \times (1 - \alpha)\%$ confidence interval of μ is approximately represented as:

$$\Big(\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}},\ \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\Big),$$

for large n.

For now, we do not impose any assumptions on the distribution of X_i.

If we assume that X_i is normal, $\dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ has a t distribution with $n - 1$ degrees of freedom for any n.

Therefore, $100 \times (1 - \alpha)\%$ confidence interval of μ is given by:

$$\Big(\overline{x} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}},\ \overline{x} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\Big),$$

where $t_{\alpha/2}(n-1)$ denotes the $100 \times \alpha/2$ percent point of the t distribution with $n - 1$ degrees of freedom.

Interval Estimation of $\hat{\theta}_n$: Let $X_1, X_2, \cdots, X_n$ be mutually independently and identically distributed random variables.

X_i has the probability density function $f(x_i; \theta)$.

Suppose that $\hat{\theta}_n$ represents the maximum likelihood estimator of θ.

From (25), we can approximate the $100\times(1-\alpha)\%$ confidence interval of θ as follows:

$$\Big(\hat{\theta}_n - z_{\alpha/2}\frac{\sigma(\hat{\theta}_n)}{\sqrt{n}},\ \hat{\theta}_n + z_{\alpha/2}\frac{\sigma(\hat{\theta}_n)}{\sqrt{n}}\Big).$$

8 Testing Hypothesis (仮説検定)

8.1 Basic Concepts in Testing Hypothesis

Given the population distribution $f(x; \theta)$, we want to judge from the observed values $x_1, x_2, \cdots, x_n$ whether the hypothesis on the parameter θ, e.g. $\theta = \theta_0$, is correct or not.

The hypothesis that we want to test is called the **null hypothesis (帰無仮説)**, which is denoted by $H_0 : \theta = \theta_0$.

The hypothesis against the null hypothesis, e.g. $\theta \neq \theta_0$, is called the **alternative hypothesis (対立仮説)**, which is denoted by $H_1 : \theta \neq \theta_0$.

Type I and Type II Errors (第一種の誤り，第二種の誤り):

When we test the null hypothesis H_0, as shown in Table 1 we have four cases, i.e.,

(i) we accept H_0 when H_0 is true,

(ii) we reject H_0 when H_0 is true,

(iii) we accept H_0 when H_0 is false, and

(iv) we reject H_0 when H_0 is false.

(i) and (iv) are correct judgments, while (ii) and (iii) are not correct.

(ii) is called a **type I error (第一種の誤り)** and (iii) is called a **type II error (第二種の誤り)**.

The probability which a type I error occurs is called the **significance level (有意水準)**, which is denoted by α, and the probability of committing a type II error is denoted by β.

Probability of (iv) is called the **power (検出力)** or the **power function (検出力関数)**, because it is a function of the parameter θ.

Table 1: Type I and Type II Errors

	H_0 is true.	H_0 is false.
Acceptance of H_0	Correct judgment	**Type II Error** 第二種の誤り (Probability β)
Rejection of H_0	**Type I Error** 第一種の誤り (Probability α = **Significance Level**) 有意水準	Correct judgment ($1-\beta$ = **Power**) 検出力

Testing Procedures: The testing procedure is summarized as follows.

1. Construct the null hypothesis (H_0) on the parameter.
2. Consider an appropriate statistic, which is called a **test statistic (検定等計量)**.

   Derive a distribution function of the test statistic when H_0 is true.
3. From the observed data, compute the observed value of the test statistic.

4. Compare the distribution and the observed value of the test statistic.

   When the observed value of the test statistic is in the tails of the distribution, we consider that H_0 is not likely to occur and we reject H_0.