Properties of $\overline{X}$: From Theorem on p.138, mean and variance of $\overline{X}$ are obtained as follows:

$$\mathrm{E}(\overline{X}) = \mu, \qquad \mathrm{V}(\overline{X}) = \frac{\sigma^2}{n}.$$

Properties of S^{*2}, S^2 and S^{**2}: The expectation of S^{*2} is:

$$\mathrm{E}(S^{*2}) = \mathrm{E}\Big(\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2\Big) = \frac{1}{n}\sum_{i=1}^n \mathrm{E}\Big((X_i-\mu)^2\Big)$$
$$= \frac{1}{n}\sum_{i=1}^n \mathrm{V}(X_i) = \frac{1}{n}\sum_{i=1}^n \sigma^2 = \sigma^2.$$

Next, the expectation of S^2 is given by:

$$\mathrm{E}(S^2)$$
$$= \mathrm{E}\Big(\frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X})^2\Big) = \frac{1}{n-1}\mathrm{E}\Big(\sum_{i=1}^n (X_i-\overline{X})^2\Big)$$
$$= \frac{1}{n-1}\mathrm{E}\Big(\sum_{i=1}^n ((X_i-\mu)-(\overline{X}-\mu))^2\Big)$$
$$= \frac{1}{n-1}\mathrm{E}\Big(\sum_{i=1}^n ((X_i-\mu)^2 - 2(X_i-\mu)(\overline{X}-\mu) + (\overline{X}-\mu)^2)\Big)$$
$$= \frac{1}{n-1}\mathrm{E}\Big(\sum_{i=1}^n (X_i-\mu)^2 - 2(\overline{X}-\mu)\sum_{i=1}^n (X_i-\mu) + n(\overline{X}-\mu)^2\Big)$$

$$= \frac{1}{n-1}\mathrm{E}\Big(\sum_{i=1}^n (X_i-\mu)^2 - n(\overline{X}-\mu)^2\Big)$$
$$= \frac{n}{n-1}\mathrm{E}\Big(\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2\Big) - \frac{n}{n-1}\mathrm{E}((\overline{X}-\mu)^2)$$
$$= \frac{n}{n-1}\sigma^2 - \frac{n}{n-1}\frac{\sigma^2}{n} = \sigma^2.$$

$\sum_{i=1}^n (X_i-\mu) = n(\overline{X}-\mu)$ is used in the sixth equality.

$\mathrm{E}\Big(\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2\Big) = \mathrm{E}(S^{*2}) = \sigma^2$ and

$\mathrm{E}((\overline{X}-\mu)^2) = \mathrm{V}(\overline{X}) = \dfrac{\sigma^2}{n}$ are required in the eighth equality.

Finally, the expectation of S^{**2} is represented by:

$$\mathrm{E}(S^{**2}) = \mathrm{E}\Big(\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2\Big) = \mathrm{E}\Big(\frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X})^2\Big)$$
$$= \mathrm{E}\Big(\frac{n-1}{n}S^2\Big) = \frac{n-1}{n}\mathrm{E}(S^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

Summarizing the above results, we obtain as follows:

$$\mathrm{E}(S^{*2}) = \sigma^2, \qquad \mathrm{E}(S^2) = \sigma^2, \qquad \mathrm{E}(S^{**2}) = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

7.4 Point Estimation: Optimality

θ denotes the parameter to be estimated.

$\hat{\theta}_n(X_1, X_2, \cdots, X_n)$ represents the estimator of θ, while $\hat{\theta}_n(x_1, x_2, \cdots, x_n)$ indicates the estimate of θ.

Hereafter, in the case of no confusion, $\hat{\theta}_n(X_1, X_2, \cdots, X_n)$ is simply written as $\hat{\theta}_n$.

As discussed above, there are numerous candidates of the estimator $\hat{\theta}_n$.

The desired properties of $\hat{\theta}_n$ are:

- **unbiasedness (不偏性)**,
- **efficiency (有効性)**.
- **consistency (一致性)** and
- **sufficiency (十分性)**. ⟵ Not discussed in this class.

Unbiasedness (不偏性): One of the desirable features that the estimator of the parameter should have is given by:

$$E(\hat{\theta}_n) = \theta, \tag{12}$$

which implies that $\hat{\theta}_n$ is distributed around θ.

When (12) holds, $\hat{\theta}_n$ is called the **unbiased estimator (不偏推定量)** of θ.

$E(\hat{\theta}_n) - \theta$ is defined as **bias (偏り)**.

As an example of unbiasedness, consider the case of $\theta = (\mu, \sigma^2)$.

Suppose that $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed with mean μ and variance σ^2.

Consider the following estimators of μ and σ^2.

1. The estimator of μ is:
   - $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

2. The estimators of σ^2 are:

- $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$,
- $S^{**2} = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Since we have obtained $E(\overline{X}) = \mu$ and $E(S^2) = \sigma^2$, $\overline{X}$ and S^2 are unbiased estimators of μ and σ^2.
We have obtained the result $E(S^{**2}) \neq \sigma^2$ and therefore S^{**2} is not an unbiased estimator of σ^2.
According to the criterion of unbiasedness, S^2 is preferred to S^{**2} for estimation of σ^2.

Efficiency (有効性): Consider two estimators, $\hat{\theta}_n$ and $\widetilde{\theta}_n$.
Both are assumed to be unbiased.
That is, $E(\hat{\theta}_n) = \theta$ and $E(\widetilde{\theta}_n) = \theta$.
When $V(\hat{\theta}_n) < V(\widetilde{\theta}_n)$, we say that $\hat{\theta}_n$ is more efficient than $\widetilde{\theta}_n$.
The unbiased estimator with the least variance is known as the **efficient estimator (有効推定量)**.
We have the case where an efficient estimator does not exist.
In order to find the efficient estimator, we utilize **Cramer-Rao inequality (クラメール・ラオの不等式)**.

Suppose that $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed and the distribution of X_i is $f(x_i;\theta)$.
For any unbiased estimator of θ, denoted by $\hat{\theta}_n$, it is known that we have the following inequality:

$$V(\hat{\theta}_n) \geq \frac{\sigma^2(\theta)}{n}, \tag{13}$$

where
$$\sigma^2(\theta) = \frac{1}{E\left(\left(\dfrac{\partial \log f(X;\theta)}{\partial\theta}\right)^2\right)} = \frac{1}{V\left(\left(\dfrac{\partial \log f(X;\theta)}{\partial\theta}\right)\right)} = -\frac{1}{E\left(\dfrac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right)}, \tag{14}$$

which is known as the **Cramer-Rao inequality (クラメール・ラオの不等式)**.
When there exists the unbiased estimator $\hat{\theta}_n$ such that the equality in (13) holds, $\hat{\theta}_n$ becomes the unbiased estimator with minimum variance, which is the **efficient estimator (有効推定量)**.
$\dfrac{\sigma^2(\theta)}{n}$ is called the **Cramer-Rao lower bound (クラメール・ラオの下限)**.

Proof of the Cramer-Rao inequality: We prove the above inequality and the equalities in $\sigma^2(\theta)$.

The **likelihood function (尤度関数)** $l(\theta; x) = l(\theta; x_1, x_2, \cdots, x_n)$ is a joint density of $X_1, X_2, \cdots, X_n$.

That is, $l(\theta; x) = l(\theta; x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i; \theta)$

See Section 7.5 for the **likelihood function (尤度関数)**.

The integration of $l(\theta; x_1, x_2, \cdots, x_n)$ with respect to $x_1, x_2, \cdots, x_n$ is equal to one.

That is, we have the following equation:

$$1 = \int l(\theta; x)\, \mathrm{d}x, \tag{15}$$

where the likelihood function $l(\theta; x)$ is given by $l(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta)$ and $\int \cdots \mathrm{d}x$ implies n-tuple integral.

Differentiating both sides of equation (15) with respect to θ, we obtain the following equation:

$$\begin{aligned} 0 &= \int \frac{\partial l(\theta; x)}{\partial \theta}\, \mathrm{d}x = \int \frac{1}{l(\theta; x)} \frac{\partial l(\theta; x)}{\partial \theta} l(\theta; x)\, \mathrm{d}x \\ &= \int \frac{\partial \log l(\theta; x)}{\partial \theta} l(\theta; x)\, \mathrm{d}x = \mathrm{E}\Big(\frac{\partial \log l(\theta; X)}{\partial \theta}\Big), \end{aligned} \tag{16}$$

which implies that the expectation of $\dfrac{\partial \log l(\theta; X)}{\partial \theta}$ is equal to zero.

In the third equality, note that $\dfrac{\mathrm{d} \log x}{\mathrm{d}x} = \dfrac{1}{x}$.

Now, let $\hat{\theta}_n$ be an estimator of θ. The definition of the mathematical expectation of the estimator $\hat{\theta}_n$ is represented as:

$$\mathrm{E}(\hat{\theta}_n) = \int \hat{\theta}_n l(\theta; x)\, \mathrm{d}x. \tag{17}$$

Differentiating equation (17) with respect to θ on both sides, we can rewrite as follows:

$$\begin{aligned} \frac{\partial \mathrm{E}(\hat{\theta}_n)}{\partial \theta} &= \int \hat{\theta}_n \frac{\partial l(\theta; x)}{\partial \theta}\, \mathrm{d}x = \int \hat{\theta}_n \frac{\partial \log l(\theta; x)}{\partial \theta} l(\theta; x)\, \mathrm{d}x \\ &= \int \Big(\hat{\theta}_n - \mathrm{E}(\hat{\theta}_n)\Big)\Big(\frac{\partial \log l(\theta; x)}{\partial \theta} - \mathrm{E}(\frac{\partial \log l(\theta; x)}{\partial \theta})\Big) l(\theta; x)\, \mathrm{d}x \\ &= \mathrm{Cov}\Big(\hat{\theta}_n, \frac{\partial \log l(\theta; X)}{\partial \theta}\Big). \end{aligned} \tag{18}$$

In the second equality, $\dfrac{\mathrm{d} \log x}{\mathrm{d}x} = \dfrac{1}{x}$ is utilized.

The third equality holds because of $\mathrm{E}(\dfrac{\partial \log l(\theta; X)}{\partial \theta}) = 0$ from equation (16).

For simplicity of discussion, suppose that θ is a scalar.

Taking the square on both sides of equation (18), we obtain the following expression:

$$\begin{aligned} \Big(\frac{\partial \mathrm{E}(\hat{\theta}_n)}{\partial \theta}\Big)^2 &= \Big(\mathrm{Cov}\Big(\hat{\theta}_n, \frac{\partial \log l(\theta; X)}{\partial \theta}\Big)\Big)^2 = \rho^2 \mathrm{V}(\hat{\theta}_n) \mathrm{V}\Big(\frac{\partial \log l(\theta; X)}{\partial \theta}\Big) \\ &\le \mathrm{V}(\hat{\theta}_n) \mathrm{V}\Big(\frac{\partial \log l(\theta; X)}{\partial \theta}\Big), \end{aligned} \tag{19}$$

where ρ denotes the correlation coefficient between $\hat{\theta}_n$ and $\dfrac{\partial \log l(\theta; X)}{\partial \theta}$.

Note that we have the definition of ρ is given by:

$$\rho = \frac{\mathrm{Cov}\Big(\hat{\theta}_n, \dfrac{\partial \log l(\theta; X)}{\partial \theta}\Big)}{\sqrt{\mathrm{V}(\hat{\theta}_n)}\sqrt{\mathrm{V}\Big(\dfrac{\partial \log l(\theta; X)}{\partial \theta}\Big)}}.$$

Moreover, we have $-1 \le \rho \le 1$ (i.e., $\rho^2 \le 1$).

Then, the inequality (19) is obtained, which is rewritten as:

$$\mathrm{V}(\hat{\theta}_n) \ge \frac{\Big(\dfrac{\partial \mathrm{E}(\hat{\theta}_n)}{\partial \theta}\Big)^2}{\mathrm{V}\Big(\dfrac{\partial \log l(\theta; X)}{\partial \theta}\Big)}. \qquad (20)$$

When $\mathrm{E}(\hat{\theta}_n) = \theta$, i.e., when $\hat{\theta}_n$ is an unbiased estimator of θ, the numerator in the right-hand side of equation (20) is equal to one.

Therefore, we have the following result:

$$\mathrm{V}(\hat{\theta}_n) \ge \frac{1}{\mathrm{V}\Big(\dfrac{\partial \log l(\theta; X)}{\partial \theta}\Big)} = \frac{1}{\mathrm{E}\Big(\Big(\dfrac{\partial \log l(\theta; X)}{\partial \theta}\Big)^2\Big)}.$$

Note that we have $\mathrm{V}\Big(\dfrac{\partial \log l(\theta; X)}{\partial \theta}\Big) = \mathrm{E}\Big(\Big(\dfrac{\partial \log l(\theta; X)}{\partial \theta}\Big)^2\Big)$ in the equality above, because of $\mathrm{E}\Big(\dfrac{\partial \log l(\theta; X)}{\partial \theta}\Big) = 0$.

Moreover, the denominator in the right-hand side of the above inequality is rewritten as follows:

$$\begin{aligned}
&\mathrm{E}\Big(\Big(\frac{\partial \log l(\theta; X)}{\partial \theta}\Big)^2\Big) \\
&= \mathrm{E}\Big(\Big(\sum_{i=1}^{n} \frac{\partial \log f(X_i; \theta)}{\partial \theta}\Big)^2\Big) = \sum_{i=1}^{n} \mathrm{E}\Big(\Big(\frac{\partial \log f(X_i; \theta)}{\partial \theta}\Big)^2\Big) \\
&= n\mathrm{E}\Big(\Big(\frac{\partial \log f(X; \theta)}{\partial \theta}\Big)^2\Big) = n \int_{-\infty}^{\infty} \Big(\frac{\partial \log f(x; \theta)}{\partial \theta}\Big)^2 f(x; \theta)\,\mathrm{d}x.
\end{aligned}$$

In the first equality, $\log l(\theta; X) = \sum_{i=1}^{n} \log f(X_i; \theta)$ is utilized.

Since X_i, $i = 1, 2, \cdots, n$, are mutually independent, the second equality holds.

The third equality holds because X_1, X_2, $\cdots$, X_n are identically distributed.

Therefore, we obtain the following inequality:

$$\mathrm{V}(\hat{\theta}_n) \ge \frac{1}{\mathrm{E}\Big(\Big(\dfrac{\partial \log l(\theta; X)}{\partial \theta}\Big)^2\Big)} = \frac{1}{n\mathrm{E}\Big(\Big(\dfrac{\partial \log f(X; \theta)}{\partial \theta}\Big)^2\Big)} = \frac{\sigma^2(\theta)}{n},$$

which is equivalent to (13).

Next, we prove the equalities in (14), i.e.,

$$\begin{aligned}
-\mathrm{E}\Big(\frac{\partial^2 \log f(X; \theta)}{\partial \theta^2}\Big) &= \mathrm{E}\Big(\Big(\frac{\partial \log f(X; \theta)}{\partial \theta}\Big)^2\Big) \\
&= \mathrm{V}\Big(\frac{\partial \log f(X; \theta)}{\partial \theta}\Big).
\end{aligned}$$

Differentiating $\int f(x; \theta)\,\mathrm{d}x = 1$ with respect to θ, we obtain as follows:

$$\int \frac{\partial f(x; \theta)}{\partial \theta}\,\mathrm{d}x = 0.$$

We assume that the range of x does not depend on the parameter θ and that $\dfrac{\partial f(x; \theta)}{\partial \theta}$ exists.

The above equation is rewritten as:

$$\int \frac{\partial \log f(x; \theta)}{\partial \theta} f(x; \theta)\,\mathrm{d}x = 0, \qquad (21)$$

or equivalently,

$$\mathrm{E}\Big(\frac{\partial \log f(X; \theta)}{\partial \theta}\Big) = 0. \qquad (22)$$

Again, differentiating equation (21) with respect to θ,

$$\int \frac{\partial^2 \log f(x; \theta)}{\partial \theta^2} f(x; \theta)\,\mathrm{d}x + \int \frac{\partial \log f(x; \theta)}{\partial \theta}\frac{\partial f(x; \theta)}{\partial \theta}\,\mathrm{d}x = 0,$$

i.e.,

$$\int \frac{\partial^2 \log f(x;\theta)}{\partial\theta^2} f(x;\theta)\,\mathrm{d}x + \int \Big(\frac{\partial \log f(x;\theta)}{\partial\theta}\Big)^2 f(x;\theta)\,\mathrm{d}x = 0,$$

i.e.,

$$\mathrm{E}\Big(\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\Big) + \mathrm{E}\bigg(\Big(\frac{\partial \log f(x;\theta)}{\partial\theta}\Big)^2\bigg) = 0.$$

Thus, we obtain:

$$-\mathrm{E}\Big(\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\Big) = \mathrm{E}\bigg(\Big(\frac{\partial \log f(x;\theta)}{\partial\theta}\Big)^2\bigg).$$

Moreover, from equation (22), the following equation is derived.

$$\mathrm{E}\bigg(\Big(\frac{\partial \log f(x;\theta)}{\partial\theta}\Big)^2\bigg) = \mathrm{V}\Big(\frac{\partial \log f(x;\theta)}{\partial\theta}\Big).$$

Therefore, we have:

$$-\mathrm{E}\Big(\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\Big) = \mathrm{E}\bigg(\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\Big)^2\bigg) = \mathrm{V}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\Big).$$

Thus, the Cramer-Rao inequality is derived as:

$$\mathrm{V}(\hat{\theta}_n) \geq \frac{\sigma^2(\theta)}{n},$$

where

$$\begin{aligned}\sigma^2(\theta) &= \frac{1}{\mathrm{E}\bigg(\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\Big)^2\bigg)} = \frac{1}{\mathrm{V}\bigg(\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\Big)\bigg)} \\ &= -\frac{1}{\mathrm{E}\Big(\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\Big)}.\end{aligned}$$