Statistics for Data Science

MSc Data Science WiSe 2019/20

Prof. Dr. Dirk Ostwald

Frequentist inference

(7) Finite-sample estimator properties

Bibliographic remarks

The material presented in this section is based on Wasserman (2004, Chapter 9), Held and Sabanés Bové (2014, Sections 3.1,3.2,4.1,4.2), and DeGroot and Schervish (2012, Section 8.8).

The frequentist sampling intuition

- Statistical models often assume $X_1, ..., X_n \sim p_\theta$
- Real observed data is considered one possible realization of $X_1, ..., X_n \sim p_\theta$.
- From a sampling perspective, however, we could sample data and estimators

$$
\begin{aligned}
&x_1^{(1)}, x_2^{(1)}, ..., x_n^{(1)} \text{ and } \hat{\theta}_n\left(x_1^{(1)}, x_2^{(1)}, ..., x_n^{(1)}\right), \text{ e.g., } \hat{\theta}_n^{(1)} = \tfrac{1}{n}\textstyle\sum_{i=1}^n x_i^{(1)} \\
&x_1^{(2)}, x_2^{(2)}, ..., x_n^{(2)} \text{ and } \hat{\theta}_n\left(x_1^{(2)}, x_2^{(2)}, ..., x_n^{(2)}\right), \text{ e.g., } \hat{\theta}_n^{(2)} = \tfrac{1}{n}\textstyle\sum_{i=1}^n x_i^{(2)} \\
&x_1^{(3)}, x_2^{(3)}, ..., x_n^{(3)} \text{ and } \hat{\theta}_n\left(x_1^{(3)}, x_2^{(3)}, ..., x_n^{(3)}\right), \text{ e.g., } \hat{\theta}_n^{(3)} = \tfrac{1}{n}\textstyle\sum_{i=1}^n x_i^{(3)} \\
&x_1^{(4)}, x_2^{(4)}, ..., x_n^{(4)} \text{ and } \hat{\theta}_n\left(x_1^{(4)}, x_2^{(4)}, ..., x_n^{(4)}\right), \text{ e.g., } \hat{\theta}_n^{(4)} = \tfrac{1}{n}\textstyle\sum_{i=1}^n x_i^{(4)} \\
&x_1^{(5)}, x_2^{(5)}, ..., x_n^{(5)} \text{ and } \hat{\theta}_n\left(x_1^{(5)}, x_2^{(4)}, ..., x_n^{(5)}\right), \text{ e.g., } \hat{\theta}_n^{(5)} = \tfrac{1}{n}\textstyle\sum_{i=1}^n x_i^{(5)} \\
&\ldots
\end{aligned}
$$

- Frequentist inference is interested in the distributional properties of estimators.
- For example, what is the distribution of $\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}, \hat{\theta}_n^{(3)}, ...$?
- Finite-sample estimator properties concern the distribution of $\hat{\theta}_n$ for fixed n.
- Asymptotic estimator properties concern the distribution of $\hat{\theta}_n$ for $n \to \infty$.

Finite-sample estimator properties

- Error, bias, and unbiasedness
- Variance and standard error
- Cramér-Rao bound
- Mean squared error

Finite-sample estimator properties

- **Error, bias, and unbiasedness**
- Variance and standard error
- Cramér-Rao bound
- Mean squared error

Definition (Error, bias, and unbiasedness)

Let $\mathcal{P}$ denote a parametric statistical model with PMF/PDF p_θ, let $X_1, ..., X_n \sim p_\theta$, and let $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ denote an estimator for θ.

- The *error* of $\hat{\theta}_n$ is defined as
$$\hat{\theta}_n - \theta. \tag{1}$$
- The *bias* of $\hat{\theta}_n$ is defined as
$$\mathrm{B}(\hat{\theta}_n) := \mathbb{E}_\theta(\hat{\theta}_n) - \theta. \tag{2}$$
- $\hat{\theta}_n$ is called *unbiased*, if
$$\mathrm{B}(\hat{\theta}_n) = 0 \Leftrightarrow \mathbb{E}_\theta(\hat{\theta}_n) = \theta \text{ for all } \theta \in \Theta, n \in \mathbb{N}. \tag{3}$$
Otherwise, $\hat{\theta}_n$ is called *biased*.

Remarks

- The error depends on the realization of $X_1, ..., X_n$.
- The bias is the expected error over many realizations of $X_1, ..., X_n$.
- $\mathbb{E}_\theta$ means expectation with respect to p_θ.

Theorem (Unbiasedness of sample mean and sample variance)

Let $X_1, ..., X_n \sim p_\theta$ be a random sample of a parametric statistical model $\mathcal{P}$ with expectation $\mu := \mathbb{E}(X_i)$ and variance $\sigma^2 := \mathbb{V}(X_i)$ for $i = 1, ..., n$. Then

- The *sample mean*

$$\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i \tag{4}$$

  is an unbiased estimator of the expectation μ, and

- the *sample variance*

$$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \tag{5}$$

  is an unbiased estimator of the variance σ^2.

Proof

For ease of notation, we set $\mathbb{E} := \mathbb{E}_\theta$ and $\mathbb{V} := \mathbb{V}_\theta$. With the linearity of expectations, we then have

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

which proofs the unbiasedness of the sample mean as an estimator of the expectation.

To show the unbiasedness of the sample variance, we first note that we have

$$\mathbb{V}(\bar{X}) = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left(X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

We further note that basic algebraic manipulations yield

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

Proof (cont.)

We then have

$$\begin{aligned}
\mathbb{E}\left((n-1)S^2\right) &= \mathbb{E}\left(\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right) \\
&= \mathbb{E}\left(\sum_{i=1}^{n}(X_i-\mu)^2 - n(\bar{X}-\mu)^2\right) \\
&= \sum_{i=1}^{n}\mathbb{E}\left((X_i-\mu)^2\right) - n\mathbb{E}\left((\bar{X}-\mu)^2\right) \\
&= n\mathbb{V}(X_i) - n\mathbb{V}(\bar{X}) \\
&= n\sigma^2 - n\frac{\sigma^2}{n} \\
&= n\sigma^2 - \sigma^2 \\
&= (n-1)\sigma^2
\end{aligned}$$

Finally, we have

$$\mathbb{E}(S^2) = \mathbb{E}\left(\frac{1}{n-1}(n-1)S^2\right) = \frac{1}{n-1}\mathbb{E}\left((n-1)S^2\right) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2,$$

which shows the unbiasedness of the sample variance as an estimator of the variance.

□

Theorem (Biasedness of the sample standard deviation)

Let $X_1, ..., X_n$ be a random sample of a parametric statistical model $\mathcal{P}$ with variance $\sigma^2 := \mathbb{V}(X_i)$ and standard deviation $\sigma := \sqrt{\mathbb{V}(X_i)}$ for $i = 1, ..., n$. Then the *sample standard deviation*

$$S := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{6}$$

is a biased estimator of the standard deviation σ.

Proof

We first note that $\sqrt{\cdot}$ is a strictly concave function and $\sigma^2 > 0$. Then, with Jensen's inequality $\mathbb{E}(f(X)) < f(\mathbb{E}(X))$ for strictly concave functions, we have

$$\mathbb{E}(S) = \mathbb{E}\left(\sqrt{S^2}\right) < \sqrt{\mathbb{E}(S^2)} = \sqrt{\sigma^2} = \sigma. \tag{7}$$

□ Remark

- Nonlinear transformations of unbiased estimators are often biased.

Finite-sample estimator properties

- Error, bias, and unbiasedness
- **Variance and standard error**
- Cramér-Rao bound
- Mean squared error

Definition (Variance and standard error)

Let $\mathcal{P}$ denote a parametric statistical model with PMF/PDF p_θ, let $X_1, ..., X_n \sim p_\theta$, and let $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ denote an estimator for θ.

- The *variance* of $\hat{\theta}_n$ is defined as

$$\mathbb{V}_\theta(\hat{\theta}_n) := \mathbb{E}_\theta\left((\hat{\theta}_n - \mathbb{E}_\theta(\hat{\theta}_n))^2\right). \tag{8}$$

- The *standard error* of $\hat{\theta}_n$ is defined as

$$\mathrm{SE}(\hat{\theta}_n) := \sqrt{\mathbb{V}_\theta(\hat{\theta}_n)} \tag{9}$$

Remark

- The estimator variance is the variance of the random variable $\hat{\theta}_n$.
- The estimator standard error is the standard deviation of $\hat{\theta}_n$.
- All expectations, variances and the standard error are with respect to p_θ.

Theorem (Standard error of the sample mean)

Let $X_1, ..., X_n \sim p_\theta$ be a random sample of a parametric statistical model $\mathcal{P}$ with expectation $\mu := \mathbb{E}(X_i)$ and variance $\sigma^2 := \mathbb{V}(X_i)$ for $i = 1, ..., n$. Then the *standard error of the sample mean*, also referred to as *standard error of the mean* is given by

$$\mathrm{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}. \tag{10}$$

Proof

By definition and with $\mathbb{V}_\theta(\bar{X}) = \sigma^2/n$, we have

$$\mathrm{SE}(\bar{X}) = \sqrt{\mathbb{V}_\theta(\bar{X})} = \sqrt{\sigma^2/n} = \sigma/\sqrt{n}. \tag{11}$$

□

Remark

- A biased estimator for the standard error of the sample mean is given in terms of the sample standard deviation by $\hat{\mathrm{SE}}(\bar{X}) = S/\sqrt{n}$.

Example (Standard error of the Bernoulli parameter MLE)

Let $X_1, ..., X_n \sim \text{Bern}(\mu)$ and let $\hat{\mu}_{\text{ML}}$ denote the maximum likelihood estimator for μ. Then the standard error of $\hat{\mu}_{\text{ML}}$ is

$$\text{SE}\left(\hat{\mu}_{\text{ML}}\right) = \sqrt{\frac{\mu(1-\mu)}{n}}. \tag{12}$$

Proof
We have

$$\text{SE}\left(\hat{\mu}_{\text{ML}}\right) = \sqrt{\mathbb{V}\left(\hat{\mu}_{\text{ML}}\right)} = \sqrt{\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i)} = \sqrt{\frac{n\mu(1-\mu)}{n^2}} = \sqrt{\frac{\mu(1-\mu)}{n}}, \tag{13}$$

where the third equation follows with the independence of the $X_i, i = 1, ..., n$ and the fourth equation follows with the variance $\mathbb{V}(X) = \mu(1-\mu)$ of a Bernoulli distributed random variable (cf. Lecture (5)).

□

Remark

- An estimator for the standard error is given by $\hat{\text{SE}}\left(\hat{\mu}_{\text{ML}}\right) = \sqrt{\frac{\hat{\mu}_{\text{ML}}(1-\hat{\mu}_{\text{ML}})}{n}}$.

Finite-sample estimator properties

- Error, bias, and unbiasedness
- Variance and standard error
- **Cramér-Rao bound**
- Mean squared error

Cramér-Rao bound

- The smaller the variance of an estimator, the better.
- The Cramér-Rao bound is a lower variance bound for unbiased estimators.
- An unbiased estimator with variance equal to the Cramér-Rao bound has minimal variance among all unbiased estimators and is "optimal" in this sense.
- The Cramér-Rao bound rests on the notion of expected *Fisher information*.
- Most results in this respect hold only under the *Fisher regularity conditions*.

Fisher regularity conditions

1. Θ is an open interval, i.e., θ must not be at a parameter space boundary.
2. The support of p_θ does not depend on θ.
3. PMFs or PDFs indexed by θ are distinct.
4. The likelihood function is twice continuously differentiable.
5. Integration and differentiation can be exchanged.

Definition (Score function and Fisher information)

Let $\mathcal{P}$ denote a statistical model with PMF or PDF p_θ with scalar parameter $\theta \in \Theta$, let $X_1, ..., X_n \sim p_\theta$ denote a random sample from this model, and let ℓ_n denote the respective log likelihood function.

- The first derivative of the log likelihood function ℓ_n is referred to as the *score function*,

$$S_n(\theta) := \frac{d}{d\theta}\ell_n(\theta). \tag{14}$$

  For $n = 1$, we write $S(\theta) := S_1(\theta)$ and call $S(\theta)$ the *score function of a random variable*.

- The negative second derivative of the log likelihood function ℓ_n is referred to as *Fisher information* of the random sample $X_1, ..., X_n$,

$$I_n(\theta) := -\frac{d^2}{d\theta^2}\ell_n(\theta). \tag{15}$$

  For $n = 1$, we write $I(\theta) := I_1(\theta)$ and call $I(\theta)$ the *Fisher information of a random variable*.

Definition (Expected and observed Fisher information)

Let $\mathcal{P}$ denote a statistical model with PMF or PDF p_θ with scalar parameter $\theta \in \Theta$, let $X_1, ..., X_n \sim p_\theta$ denote a random sample from this model, let ℓ_n denote the respective log likelihood function, and let $\hat{\theta}_n^{ML}$ denote a maximum likelihood estimator of θ.

- The *expected Fisher information* of a random sample is defined as

$$J_n(\theta) := \mathbb{E}_\theta(I_n(\theta)). \tag{16}$$

  For $n = 1$, we write $J(\theta) := J_n(\theta)$ and refer to $J(\theta)$ as the *expected Fisher information of a random variable.*

- The *observed Fisher information* of a random sample is defined as

$$I_n\left(\theta_n^{ML}\right) = -\frac{d^2}{d\theta^2}\ell_n\left(\theta_n^{ML}\right), \tag{17}$$

  i.e., the observed Fisher information of a random sample is the Fisher information at the location of the maximum likelihood estimate θ_n^{ML}.

Theorem (Expectation and variance of the score function)

The expectation and the variance of the score function are given by

$$\mathbb{E}_\theta(S(\theta)) = 0 \text{ and } \mathbb{V}_\theta(S(\theta)) = J(\theta), \tag{18}$$

respectively.

Remarks

- The expectation of the derivative of the log likelihood function is zero.
- The expected Fisher information equals the variance of the score function.

Proof

We only consider the case that p_θ is a PDF and first show $\mathbb{E}_\theta(S(\theta)) = 0$:

$$\begin{aligned}
\mathbb{E}_\theta(S(\theta)) &= \int S(\theta) p_\theta(x)\, dx \\
&= \int \frac{d}{d\theta} \ell(\theta) p_\theta(x)\, dx \\
&= \int \frac{d}{d\theta} \ln L(\theta) p_\theta(x)\, dx \\
&= \int \frac{1}{L(\theta)} \frac{d}{d\theta} L(\theta) p_\theta(x)\, dx \\
&= \int \frac{1}{p_\theta(x)} \frac{d}{d\theta} L(\theta) p_\theta(x)\, dx \\
&= \int \frac{d}{d\theta} L(\theta)\, dx \\
&= \frac{d}{d\theta} \int p_\theta(x)\, dx \\
&= \frac{d}{d\theta} 1 \\
&= 0.
\end{aligned} \tag{19}$$

With the definition of the variance, it immediately follows that $\mathbb{V}_\theta(S(\theta)) = \mathbb{E}_\theta\left(S(\theta)^2\right)$. We next show that $J(\theta) = \mathbb{E}_\theta\left(S(\theta)^2\right)$ and thus $\mathbb{V}_\theta(S(\theta)) = J(\theta)$.

Proof (cont.)

$$
\begin{aligned}
J(\theta) &= \mathbb{E}_\theta \left(-\frac{d^2}{d\theta^2} \ln L(\theta) \right) \\
&= \mathbb{E}_\theta \left(-\frac{d}{d\theta} \frac{\frac{d}{d\theta} L(\theta)}{L(\theta)} \right) \\
&= \mathbb{E}_\theta \left(-\frac{\frac{d^2}{d\theta^2} L(\theta) L(\theta) - \frac{d}{d\theta} L(\theta) \frac{d}{d\theta} L(\theta)}{L(\theta) L(\theta)} \right) \\
&= -\mathbb{E}_\theta \left(\frac{\frac{d^2}{d\theta^2} L(\theta)}{L(\theta)} \right) + \mathbb{E}_\theta \left(\frac{\left(\frac{d}{d\theta} L(\theta) \right)^2}{(L(\theta))^2} \right) \qquad (20) \\
&= -\int \frac{\frac{d^2}{d\theta^2} L(\theta)}{L(\theta)} p_\theta(x)\, dx + \int \frac{\left(\frac{d}{d\theta} L(\theta) \right)^2}{(L(\theta))^2} p_\theta(x)\, dx \\
&= -\frac{d^2}{d\theta^2} \int L(\theta)\, dx + \int \left(\frac{\frac{d}{d\theta} L(\theta)}{L(\theta)} \right)^2 p_\theta(x)\, dx \\
&= -\frac{d^2}{d\theta^2} 1 + \int \left(\frac{d}{d\theta} \ln L(\theta) \right)^2 p_\theta(x)\, dx \\
&= \mathbb{E}_\theta \left(S(\theta)^2 \right).
\end{aligned}
$$

□

Theorem (Cramér-Rao bound)

Let $\mathcal{P}$ denote a parametric statistical model with PMF or PDF p_θ and let $\hat{\theta}$ denote an unbiased estimator for $g(\theta)$. Then

$$\mathbb{V}_\theta(\hat{\theta}) \geq \frac{\left(\frac{d}{d\theta} g(\theta)\right)^2}{J(\theta)}. \tag{21}$$

In particular, for $g(\theta) := \theta$ and thus $\left(\frac{d}{d\theta} g(\theta)\right)^2 = 1$.

$$\mathbb{V}_\theta(\hat{\theta}) \geq \frac{1}{J(\theta)}. \tag{22}$$

The right-hand sides of the above are referred to as *Cramér-Rao lower bounds*.

Remarks

- In words, the variance of an unbiased estimator $\hat{\theta}$ for θ is larger or equal to the reciprocal expected Fisher information $J(\theta)$.
- If $\mathbb{V}_\theta(\hat{\theta}) = \frac{1}{J(\theta)}$, the variance of the estimator is as low as possible.
- A highly Fisher-informative unbiased estimator has a low variance lower bound.

Proof

We first note that for the random variables $S(\theta)$ and $\hat{\theta}$ we have with the correlation inequality and the fact that $\mathbb{V}_\theta(S(\theta)) = J(\theta)$

$$\frac{\mathbb{C}_\theta(S(\theta),\hat{\theta})^2}{\mathbb{V}_\theta(S(\theta))\mathbb{V}_\theta(\hat{\theta})} \leq 1 \Rightarrow \mathbb{V}_\theta(\hat{\theta}) \geq \frac{\mathbb{C}_\theta(S(\theta),\hat{\theta})^2}{J(\theta)}. \tag{23}$$

With the translation theorem for covariances, $\mathbb{E}_\theta(S(\theta)) = 0$, and the unbiasedness of $\hat{\theta}$, we then have

$$\begin{aligned}
\mathbb{C}_\theta(S(\theta),\hat{\theta}) &= \mathbb{E}_\theta(S(\theta)\hat{\theta}) - \mathbb{E}_\theta(S(\theta))\mathbb{E}_\theta(\hat{\theta}) \\
&= \mathbb{E}_\theta(S(\theta)\hat{\theta}) \\
&= \int S(\theta)\,\hat{\theta}\,p_\theta(x)\,dx \\
&= \int \frac{d}{d\theta}\ln L(\theta)\,\hat{\theta}\,p_\theta(x)\,dx \\
&= \int \frac{\frac{d}{d\theta}L(\theta)}{L(\theta)}\,\hat{\theta}\,p_\theta(x)\,dx \\
&= \int \frac{\frac{d}{d\theta}L(\theta)}{p_\theta(x)}\,\hat{\theta}\,p_\theta(x)\,dx \\
&= \int \frac{d}{d\theta}L(\theta)\,\hat{\theta}\,dx \\
&= \frac{d}{d\theta}\int L(\theta)\,\hat{\theta}\,dx = \frac{d}{d\theta}\int \hat{\theta}\,p_\theta(x)\,dx = \frac{d}{d\theta}\mathbb{E}_\theta(\theta) = \frac{d}{d\theta}g(\theta)
\end{aligned} \tag{24}$$

□

Finite-sample estimator properties

- Error, bias, and unbiasedness
- Variance and standard error
- Cramér-Rao bound
- **Mean squared error**

Definition (Mean squared error)

Let $\mathcal{P}$ denote a parametric statistical model with PMF/PDF p_θ, let $X_1, ..., X_n \sim p_\theta$, and let $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ denote an estimator for θ. Then the *mean squared error* of $\hat{\theta}_n$ is defined as

$$\mathrm{MSE}(\hat{\theta}_n) := \mathbb{E}_\theta \left((\hat{\theta}_n - \theta)^2 \right). \tag{25}$$

Remarks

- The MSE is the expected squared deviation of $\hat{\theta}_n$ from θ.
- The variance is the expected squared deviation of $\hat{\theta}_n$ from $\mathbb{E}_\theta(\hat{\theta}_n)$.
- The expectation $\mathbb{E}_\theta(\hat{\theta}_n)$ may or may not coincide with θ.

Theorem (Mean squared error decomposition)

Let $\mathcal{P}$ denote a parametric statistical model with PMF/PDF p_θ, let $X_1, ..., X_n \sim p_\theta$, and let $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ denote an estimator for θ. Then

$$\mathrm{MSE}(\hat{\theta}_n) = \mathrm{B}(\hat{\theta}_n)^2 + \mathbb{V}_\theta(\hat{\theta}_n) \tag{26}$$

Remarks

- Mean squared error = Bias^2 + Variance.
- The MSE can be used as a bias-variance trade-off criterion.
- Small biases may be favoured over large variances.

Proof

Let $\bar{\theta}_n := \mathbb{E}_\theta(\hat{\theta}_n)$. Then

$$\begin{aligned}
\mathbb{E}_\theta\left((\hat{\theta}_n - \theta)^2\right) &= \mathbb{E}_\theta\left((\hat{\theta}_n - \bar{\theta}_n + \bar{\theta}_n - \theta)^2\right) \\
&= \mathbb{E}_\theta\left((\hat{\theta}_n - \bar{\theta}_n)^2 + 2(\hat{\theta}_n - \bar{\theta}_n)(\bar{\theta}_n - \theta) + (\bar{\theta}_n - \theta)^2\right) \\
&= \mathbb{E}_\theta\left((\hat{\theta}_n - \bar{\theta}_n)^2\right) + 2\mathbb{E}_\theta\left((\hat{\theta}_n - \bar{\theta}_n)(\bar{\theta}_n - \theta)\right) + \mathbb{E}_\theta\left((\bar{\theta}_n - \theta)^2\right) \\
&= \mathbb{E}_\theta\left((\hat{\theta}_n - \bar{\theta}_n)^2\right) + 2\mathbb{E}_\theta\left(\hat{\theta}_n\bar{\theta}_n - \hat{\theta}_n\theta - \bar{\theta}_n\bar{\theta}_n + \bar{\theta}_n\theta\right) + \mathbb{E}_\theta\left((\bar{\theta}_n - \theta)^2\right) \\
&= \mathbb{E}_\theta\left((\hat{\theta}_n - \bar{\theta}_n)^2\right) + 2\left(\bar{\theta}_n\bar{\theta}_n - \bar{\theta}_n\theta - \bar{\theta}_n\bar{\theta}_n + \bar{\theta}_n\theta\right) + \mathbb{E}_\theta\left((\bar{\theta}_n - \theta)^2\right) \\
&= \mathbb{E}_\theta\left((\hat{\theta}_n - \bar{\theta}_n)^2\right) + 0 + \mathbb{E}_\theta\left((\bar{\theta}_n - \theta)^2\right) \\
&= \mathbb{E}_\theta\left((\bar{\theta}_n - \theta)^2\right) + \mathbb{E}_\theta\left((\hat{\theta}_n - \bar{\theta}_n)^2\right) \\
&= \mathbb{E}_\theta\left((\mathbb{E}_\theta(\hat{\theta}_n) - \theta)^2\right) + \mathbb{E}_\theta\left((\hat{\theta}_n - \mathbb{E}_\theta(\hat{\theta}_n))^2\right) \\
&= (\mathbb{E}_\theta(\hat{\theta}_n) - \theta)^2 + \mathbb{V}_\theta(\hat{\theta}_n) \\
&= \mathrm{B}(\hat{\theta}_n)^2 + \mathbb{V}_\theta(\hat{\theta}_n).
\end{aligned}$$

□

References

DeGroot, M. H. and Schervish, M. J. (2012). *Probability and Statistics*. Pearson Education.

Held, L. and Sabanés Bové, D. (2014). *Applied statistical inference*, volume 10. Springer.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer.