Approximate and asymptotic bias

- Unbiasedness is a criterion for point estimators. However, in some cases, there is no unbiased estimator.
- Furthermore, having a “slight” bias in some cases may not be a bad idea.
- Let $T_n(X)$ be a point estimator of $\vartheta$ for every $n$. If $ET_n$ exists for every $n$ and $\lim_{n \to \infty} E(T_n - \vartheta) = 0$ for any $P \in \mathcal{P}$, then $T_n$ is said to be approximately unbiased.
- But, there are many reasonable point estimators whose expectations are not well defined. It is desirable to define a concept of asymptotic bias for point estimators whose expectations are not well defined.
Definition 2.11

(i) Let \( \xi, \xi_1, \xi_2, \ldots \) be random variables and \( \{a_n\} \) be a sequence of positive numbers satisfying \( a_n \to \infty \) or \( a_n \to a > 0 \). If \( a_n \xi_n \to_d \xi \) and \( E|\xi| < \infty \), then \( E\xi/a_n \) is called an asymptotic expectation of \( \xi_n \).

(ii) Let \( T_n \) be a point estimator of \( \vartheta \) for every \( n \). An asymptotic expectation of \( T_n - \vartheta \), if it exists, is called an asymptotic bias of \( T_n \) and denoted by \( \tilde{b}_{T_n}(P) \) (or \( \tilde{b}_{T_n}(\theta) \) if \( P \) is in a parametric family).

If \( \lim_{n \to \infty} \tilde{b}_{T_n}(P) = 0 \) for any \( P \in \mathcal{P} \), then \( T_n \) is said to be asymptotically unbiased.

Remarks

- Like the consistency, the asymptotic expectation (or bias) is a concept relating to sequences \( \{\xi_n\} \) and \( \{E\xi/a_n\} \) (or \( \{T_n\} \) and \( \{\tilde{b}_{T_n}(P)\}\)).

- The exact bias \( b_{T_n}(P) \) is not necessarily the same as \( \tilde{b}_{T_n}(P) \) when both of them exist.
Proposition 2.3
Let $\{\xi_n\}$ be a sequence of random variables. Suppose that both $E\xi/a_n$ and $E\eta/b_n$ are asymptotic expectations of $\xi_n$ defined according to Definition 2.11(i). Then, one of the following three must hold:

(a) $E\xi = E\eta = 0$;
(b) $E\xi \neq 0$, $E\eta = 0$, and $b_n/a_n \to 0$; or $E\xi = 0$, $E\eta \neq 0$, and $a_n/b_n \to 0$;
(c) $E\xi \neq 0$, $E\eta \neq 0$, and $(E\xi/a_n)/(E\eta/b_n) \to 1$.

Remark
If $T_n$ is a consistent estimator of $\vartheta$, then $T_n = \vartheta + o_p(1)$ and, by Definition 2.11(ii), $T_n$ is asymptotically unbiased, although $T_n$ may not be approximately unbiased.
Precise order of asymptotic bias

When \( a_n(T_n - \vartheta) \to_d Y \) with \( EY = 0 \) (e.g., \( T_n = \bar{X}^2 \) and \( \vartheta = \mu^2 \) in Example 2.33), the asymptotic bias of \( T_n \) is 0 (\( T_n \) is asymptotically unbiased).

A more precise order of the asymptotic bias of \( T_n \) may be obtained (for comparing different estimators in terms of their asymptotic biases).

Suppose there is a sequence of random \( \{\eta_n\} \) such that

\[
\begin{align*}
    a_n\eta_n & \to_d Y \\
    a_n^2(T_n - \vartheta - \eta_n) & \to_d W,
\end{align*}
\]

where \( Y \) and \( W \) are random variables with finite means, \( EY = 0 \) and \( EW \neq 0 \).

Then we may define \( a_n^{-2} \) to be the order of \( \tilde{b}_{T_n}(P) \) or define \( EW/a_n^2 \) to be the \( a_n^{-2} \) order asymptotic bias of \( T_n \).

However, \( \eta_n \) may not be unique.

Some regularity conditions have to be imposed so that the order of asymptotic bias of \( T_n \) can be uniquely defined.
Functions of sample means

We consider the case where $X_1, \ldots, X_n$ are i.i.d. random $k$-vectors with finite $\Sigma = \text{Var}(X_1)$.

Let $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$, and $T_n = g(\bar{X})$, where $g$ is a function on $\mathcal{R}^k$ that is second-order differentiable at $\mu = EX_1 \in \mathcal{R}^k$.

Consider $T_n$ as an estimator of $\vartheta = g(\mu)$.

By Taylor's expansion,

$$T_n - \vartheta = [\nabla g(\mu)]^\top (\bar{X} - \mu) + \frac{1}{2} (\bar{X} - \mu)^\top \nabla^2 g(\mu) (\bar{X} - \mu) + o_p \left( \frac{1}{n} \right),$$

where $\nabla g$ is the $k$-vector of partial derivatives of $g$ and $\nabla^2 g$ is the $k \times k$ matrix of second-order partial derivatives of $g$. 
By the CLT and Theorem 1.10(iii),

\[
\frac{n}{2}(\bar{X} - \mu)^\top \nabla^2 g(\mu)(\bar{X} - \mu) \xrightarrow{d} \frac{Z_\Sigma^\top \nabla^2 g(\mu) Z_\Sigma}{2},
\]

where \( Z_\Sigma = N_k(0, \Sigma) \).

Thus,

\[
E[Z_\Sigma^\top \nabla^2 g(\mu) Z_\Sigma] = \frac{\text{tr} (\nabla^2 g(\mu) \Sigma)}{2n}
\]

is the \( n^{-1} \) order asymptotic bias of \( T_n = g(\bar{X}) \), where \( \text{tr}(A) \) denotes the trace of the matrix \( A \).

**Example 2.35**

Let \( X_1, ..., X_n \) be i.i.d. binary random variables with \( P(X_i = 1) = p \), where \( p \in (0, 1) \) is unknown.
Example 2.35 (continued)

Consider first the estimation of \( \vartheta = p(1 - p) \).
Since \( \text{Var}(\bar{X}) = p(1 - p)/n \), the \( n^{-1} \) order asymptotic bias of \( T_n = \bar{X}(1 - \bar{X}) \) according to the formula \( \text{tr} \left( \nabla^2 g(\mu) \Sigma \right) / 2n \) with \( g(x) = x(1 - x) \) is \( -p(1 - p)/n \).
On the other hand, a direct computation shows \( E[\bar{X}(1 - \bar{X})] = E\bar{X} - E\bar{X}^2 = p - (E\bar{X})^2 - \text{Var}(\bar{X}) = p(1 - p) - p(1 - p)/n \).
Hence, the exact bias of \( T_n \) is the same as the \( n^{-1} \) order asymptotic bias.

Consider next the estimation of \( \vartheta = p^{-1} \).
In this case, there is no unbiased estimator of \( p^{-1} \) (Exercise 84 in §2.6).
Let \( T_n = \bar{X}^{-1} \).
Then, an \( n^{-1} \) order asymptotic bias of \( T_n \) according to the formula \( \text{tr} \left( \nabla^2 g(\mu) \Sigma \right) / 2n \) with \( g(x) = x^{-1} \) is \( (1 - p)/(p^2 n) \).
On the other hand, \( ET_n = \infty \) for every \( n \).
Asymptotic variance and mse

Like the bias, $\text{MSE}_{T_n}(P) = E(T_n - \vartheta)^2$, is not well defined if the second moment of $T_n$ does not exist. We now define a version of asymptotic mean squared error (amse) and a measure of assessing different point estimators of a common parameter.

**Definition 2.12**

Let $T_n$ be an estimator of $\vartheta$ for every $n$ and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$. Assume that $a_n(T_n - \vartheta) \to_d Y$ with $0 < EY^2 < \infty$.

(i) The asymptotic mean squared error of $T_n$, denoted by $\text{AMSE}_{T_n}(P)$ or $\text{AMSE}_{T_n}(\theta)$ if $P$ is in a parametric family indexed by $\theta$, is defined to be the asymptotic expectation of $(T_n - \vartheta)^2$, i.e., $\text{AMSE}_{T_n}(P) = EY^2/a_n^2$. The asymptotic variance of $T_n$ is defined to be $\sigma_{T_n}^2(P) = \text{Var}(Y)/a_n^2$. 
Definition 2.12 (continued)

(ii) Let $T_n'$ be another estimator of $\vartheta$. The *asymptotic relative efficiency* of $T_n'$ w.r.t. $T_n$ is defined to be $e_{T_n',T_n}(P) = \text{AMSE}_{T_n}(P)/\text{AMSE}_{T_n'}(P)$.

(iii) $T_n$ is said to be *asymptotically more efficient* than $T_n'$ iff $\limsup_n e_{T_n',T_n}(P) \leq 1$ for any $P$ and $< 1$ for some $P$.

Remarks

▶ The amse and asymptotic variance are the same iff $EY = 0$.
▶ By Proposition 2.3, the amse or the asymptotic variance of $T_n$ is essentially unique and, therefore, the concept of asymptotic relative efficiency in Definition 2.12(ii)-(iii) is well defined.
▶ In Example 2.33, $\text{AMSE}_{\bar{X}_2}(P) = \sigma^2_{\bar{X}_2}(P) = 4\mu^2\sigma^2/n$. 
When both $\text{MSE}_{T_n}(P)$ and $\text{MSE}_{T'_n}(P)$ exist, one may compare $T_n$ and $T'_n$ by evaluating the relative efficiency $\text{MSE}_{T_n}(P)/\text{MSE}_{T'_n}(P)$. However, this comparison may be different from the one using the asymptotic relative efficiency in Definition 2.12(ii), since the mse and amse of an estimator may be different (Exercise 115 in §2.6).

The following result shows that when the exact mse of $T_n$ exists, it is no smaller than the amse of $T_n$. It also provides a condition under which the exact mse and the amse are the same.

**Proposition 2.4**

Let $T_n$ be an estimator of $\vartheta$ for every $n$ and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$. Suppose that $a_n(T_n - \vartheta) \to_d Y$ with $0 < EY^2 < \infty$.

Then

(i) $EY^2 \leq \liminf_n E[a_n^2(T_n - \vartheta)^2]$ and

(ii) $EY^2 = \lim_{n \to \infty} E[a_n^2(T_n - \vartheta)^2]$ if and only if $\{a_n^2(T_n - \vartheta)^2\}$ is uniformly integrable.
Proof

(i) By Theorem 1.10(iii),
\[
\min\{a_n^2(T_n - \vartheta)^2, t\} \xrightarrow{d} \min\{Y^2, t\} \quad \text{for any } t > 0.
\]
Since \(\min\{a_n^2(T_n - \vartheta)^2, t\}\) is bounded by \(t\), by Theorem 1.8(viii),
\[
\lim_{n \to \infty} \mathbb{E}(\min\{a_n^2(T_n - \vartheta)^2, t\}) = \mathbb{E}(\min\{Y^2, t\})
\]
Then
\[
\mathbb{E}Y^2 = \lim_{t \to \infty} \mathbb{E}(\min\{Y^2, t\})
= \lim_{t \to \infty} \lim_{n \to \infty} \mathbb{E}(\min\{a_n^2(T_n - \vartheta)^2, t\})
= \liminf_{t, n} \mathbb{E}(\min\{a_n^2(T_n - \vartheta)^2, t\})
\leq \liminf_n \mathbb{E}[a_n^2(T_n - \vartheta)^2],
\]
where the third equality follows from the fact that 
\(\mathbb{E}(\min\{a_n^2(T_n - \vartheta)^2, t\})\) is nondecreasing in \(t\) for any fixed \(n\).

(ii) The result follows from Theorem 1.8(viii).
Example 2.36

Let $X_1, \ldots, X_n$ be i.i.d. from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$. Consider the estimation of $\vartheta = P(X_i = 0) = e^{-\theta}$.

Let $T_{1n} = F_n(0)$, where $F_n$ is the empirical c.d.f. Then $T_{1n}$ is unbiased and has $\text{MSE}_{T_{1n}}(\theta) = e^{-\theta} (1 - e^{-\theta})/n$.

Also, $\sqrt{n}(T_{1n} - \vartheta) \to_d N(0, e^{-\theta} (1 - e^{-\theta}))$ by the CLT.

Thus, in this case $\text{AMSE}_{T_{1n}}(\theta) = \text{MSE}_{T_{1n}}(\theta)$.

Consider $T_{2n} = e^{-\bar{X}}$. Note that $ET_{2n} = e^{n\theta (e^{-1/n} - 1)}$.
Hence $b_{T_{2n}}(\theta) \to 0$.

Using Theorem 1.12 and the CLT, we can show that $\sqrt{n}(T_{2n} - \vartheta) \to_d N(0, e^{-2\theta} \theta)$.
By Definition 2.12(i), $\text{AMSE}_{T_{2n}}(\theta) = e^{-2\theta} \theta/n$.

Thus, the asymptotic relative efficiency of $T_{1n}$ w.r.t. $T_{2n}$ is

$$e_{T_{1n}, T_{2n}}(\theta) = \theta/(e^{\theta} - 1) < 1$$

This shows that $T_{2n}$ is asymptotically more efficient than $T_{1n}$.
The result for $T_{2n}$ in Example 2.36 is a special case (with $U_n = \bar{X}$) of the following general result.

**Theorem 2.6**

Let $g$ be a function on $\mathcal{R}^k$ that is differentiable at $\theta \in \mathcal{R}^k$ and let $U_n$ be a $k$-vector of statistics satisfying $a_n(U_n - \theta) \xrightarrow{d} Y$ for a random $k$-vector $Y$ with $0 < E\|Y\|^2 < \infty$ and a sequence of positive numbers $\{a_n\}$ satisfying $a_n \to \infty$.

Let $T_n = g(U_n)$ be an estimator of $\vartheta = g(\theta)$.

Then, the amse and asymptotic variance of $T_n$ are, respectively,

$$\text{AMSE}_{T_n}(P) = E\{[\nabla g(\theta)]^\top Y\}^2 / a_n^2$$

and

$$\sigma^2_{T_n}(P) = [\nabla g(\theta)]^\top \text{Var}(Y) \nabla g(\theta) / a_n^2.$$
Method of Moments — An approach to derive asymptotically unbiased estimators

An exactly unbiased estimator may not exist, or is hard to obtain. We often derive asymptotically unbiased estimators. The method of moments is the oldest method of deriving asymptotically unbiased estimators, although they may not be the best estimators.

Consider a parametric problem where $X_1, \ldots, X_n$ are i.i.d. random variables from $P_{\theta}$, $\theta \in \Theta \subset \mathbb{R}^k$, and $E|X_1|^k < \infty$. Let $\mu_j = E X_1^j$ be the $j$th moment of $P$ and let

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

be the $j$th sample moment, which is an unbiased estimator of $\mu_j$, $j = 1, \ldots, k$. 

Typically, 
\[ \mu_j = h_j(\theta), \quad j = 1, \ldots, k, \] 
(1)

for some functions \( h_j \) on \( \mathbb{R}^k \).

By substituting \( \mu_j \)’s on the left-hand side of (1) by the sample moments \( \hat{\mu}_j \), we obtain a moment estimator \( \hat{\theta} \), i.e., \( \hat{\theta} \) satisfies 
\[ \hat{\mu}_j = h_j(\hat{\theta}), \quad j = 1, \ldots, k, \]

which is a sample analogue of (1). This method of deriving estimators is called the method of moments, which is an application of the important statistical principle, the substitution principle.

Let \( \hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_k) \) and \( h = (h_1, \ldots, h_k) \). Then \( \hat{\mu} = h(\hat{\theta}) \). If the inverse function \( h^{-1} \) exists, then the unique moment estimator of \( \theta \) is \( \hat{\theta} = h^{-1}(\hat{\mu}) \).
When $h^{-1}$ does not exist (i.e., $h$ is not one-to-one), any solution of
\[ \hat{\mu} = h(\hat{\theta}) \]
is a moment estimator of $\theta$. If possible, we always choose a solution $\hat{\theta}$ in the parameter space $\Theta$. In some cases, however, a moment estimator does not exist (see Exercise 111).

Assume that $\hat{\theta} = g(\hat{\mu})$ for a function $g$. If $h^{-1}$ exists, then $g = h^{-1}$. If $g$ is continuous at $\mu = (\mu_1, \ldots, \mu_k)$, then $\hat{\theta}$ is strongly consistent for $\theta$, since $\hat{\mu}_j \to_{a.s.} \mu_j$ by the SLLN. If $g$ is differentiable at $\mu$ and $E|X_1|^{2k} < \infty$, then $\hat{\theta}$ is asymptotically normal, by the CLT and Theorem 1.12, and

\[ \text{AMSE}_\hat{\theta}(\theta) = n^{-1}[\nabla g(\mu)]^\top V_\mu \nabla g(\mu), \]

where $V_\mu$ is a $k \times k$ matrix whose $(i,j)$th element is $\mu_{i+j} - \mu_i \mu_j$. Furthermore, the $n^{-1}$ order asymptotic bias of $\hat{\theta}$ is

\[ (2n)^{-1} \text{tr} \left( \nabla^2 g(\mu)V_\mu \right). \]
Example 3.24

Let $X_1, \ldots, X_n$ be i.i.d. from a population $P_{\theta}$ indexed by the parameter $\theta = (\mu, \sigma^2)$, where $\mu = EX_1 \in \mathcal{R}$ and $\sigma^2 = \text{Var}(X_1) \in (0, \infty)$. This includes cases such as the family of normal distributions, double exponential distributions, or logistic distributions (Table 1.2, page 20).

Since $EX_1 = \mu$ and $EX_1^2 = \text{Var}(X_1) + (EX_1)^2 = \sigma^2 + \mu^2$, setting $\hat{\mu}_1 = \mu$ and $\hat{\mu}_2 = \sigma^2 + \mu^2$ we obtain the moment estimator

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

Note that $\bar{X}$ is unbiased, but $\frac{n-1}{n} S^2$ is not.

If $X_i$ is normal, then $\hat{\theta}$ is sufficient and is nearly the same as an optimal estimator such as the UMVUE.

On the other hand, if $X_i$ is from a double exponential or logistic distribution, then $\hat{\theta}$ is not sufficient and can often be improved.
Consider now the estimation of $\sigma^2$ when we know that $\mu = 0$. Obviously we cannot use the equation $\hat{\mu}_1 = \mu$ to solve the problem. Using $\hat{\mu}_2 = \mu_2 = \sigma^2$, we obtain the moment estimator

$$\hat{\sigma}^2 = \hat{\mu}_2 = n^{-1} \sum_{i=1}^{n} X_i^2.$$ 

This is still a good estimator when $X_i$ is normal, but is not a function of sufficient statistic when $X_i$ is from a double exponential distribution.

For the double exponential case one can argue that we should first make a transformation $Y_i = |X_i|$ and then obtain the moment estimator based on the transformed data. The moment estimator of $\sigma^2$ based on the transformed data is

$$\bar{Y}^2 = \left( \frac{1}{n} \sum_{i=1}^{n} |X_i| \right)^2,$$

which is sufficient for $\sigma^2$. Note that this estimator can also be obtained based on absolute moment equations.
Example 3.25
Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on $(\theta_1, \theta_2)$, $-\infty < \theta_1 < \theta_2 < \infty$.

Note that

$$EX_1 = (\theta_1 + \theta_2)/2 \quad \text{and} \quad EX_1^2 = (\theta_1^2 + \theta_2^2 + \theta_1 \theta_2)/3.$$ 

Setting $\hat{\mu}_1 = EX_1$ and $\hat{\mu}_2 = EX_1^2$ and substituting $\theta_1$ in the second equation by $2\hat{\mu}_1 - \theta_2$ (the first equation), we obtain that

$$(2\hat{\mu}_1 - \theta_2)^2 + \theta_2^2 + (2\hat{\mu}_1 - \theta_2)\theta_2 = 3\hat{\mu}_2,$$

which is the same as

$$(\theta_2 - \hat{\mu}_1)^2 = 3(\hat{\mu}_2 - \hat{\mu}_1^2).$$

Since $\theta_2 > EX_1$, we obtain that

$$\hat{\theta}_2 = \hat{\mu}_1 + \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1^2)} = \bar{X} + \sqrt{\frac{3(n-1)}{n}} S^2$$

$$\hat{\theta}_1 = \hat{\mu}_1 - \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1^2)} = \bar{X} - \sqrt{\frac{3(n-1)}{n}} S^2.$$ 

These estimators are not functions of the sufficient and complete statistic $(X_{(1)}, X_{(n)})$. 
Example 3.26

Let $X_1, ..., X_n$ be i.i.d. from the binomial distribution $Bi(p, k)$ with unknown parameters $k \in \{1, 2, ...\}$ and $p \in (0, 1)$.

Since

$$EX_1 = kp$$

and

$$EX^2_1 = kp(1 - p) + k^2 p^2,$$

we obtain the moment estimators

$$\hat{p} = (\hat{\mu}_1 + \hat{\mu}^2_1 - \hat{\mu}_2)/\hat{\mu}_1 = 1 - \frac{n-1}{n} S^2 / \bar{X}$$

and

$$\hat{k} = \hat{\mu}^2_1/(\hat{\mu}_1 + \hat{\mu}^2_1 - \hat{\mu}_2) = \bar{X} / (1 - \frac{n-1}{n} S^2 / \bar{X}).$$

The estimator $\hat{p}$ is in the range of $(0, 1)$.
But $\hat{k}$ may not be an integer.
It can be improved by an estimator that is $\hat{k}$ rounded to the nearest positive integer.
The method of moments can also be applied to nonparametric problems. Consider, for example, the estimation of the central moments

\[ c_j = E(X_1 - \mu_1)^j, \quad j = 2, \ldots, k. \]

Since

\[ c_j = \sum_{t=0}^{j} \binom{j}{t} (-\mu_1)^t \mu_{j-t}, \]

the moment estimator of \( c_j \) is

\[ \hat{c}_j = \sum_{t=0}^{j} \binom{j}{t} (-\bar{X})^t \hat{\mu}_{j-t}, \]

where \( \hat{\mu}_0 = 1. \)
It can be shown (exercise) that

\[ \hat{c}_j = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^j, \quad j = 2, \ldots, k, \]

which are sample central moments.

From the SLLN, \( \hat{c}_j \)'s are strongly consistent.

If \( E|X_1|^{2k} < \infty \), then

\[ \sqrt{n} (\hat{c}_2 - c_2, \ldots, \hat{c}_k - c_k) \to_d N_{k-1}(0, D) \]

where the \((i, j)\)th element of the \((k - 1) \times (k - 1)\) matrix \( D \) is

\[ c_{i+j+2} - c_{i+1}c_{j+1} - (i+1)c_i c_{j+2} - (j+1)c_{i+2}c_j + (i+1)(j+1)c_i c_j c_2. \]