Example 2.28

Let $X_1, \ldots, X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathbb{R}$ and a known $\sigma^2$.

Consider the hypotheses $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$, where $\mu_0$ is a fixed constant.

Since the sample mean $\bar{X}$ is sufficient for $\mu \in \mathbb{R}$, it is reasonable to consider the following class of tests: $T_c(X) = I_{(c, \infty)}(\bar{X})$, i.e., $H_0$ is rejected (accepted) if $\bar{X} > c$ ($\bar{X} \leq c$), where $c \in \mathbb{R}$ is a fixed constant.

Let $\Phi$ be the c.d.f. of $N(0, 1)$.

By the property of the normal distributions,

$$\alpha_{T_c}(\mu) = P(T_c(X) = 1) = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right).$$

Figure 2.2 provides an example of a graph of two types of error probabilities, with $\mu_0 = 0$.

Since $\Phi(t)$ is an increasing function of $t$, 

\[
\sup_{P \in \mathcal{P}_0} \alpha_{T_c}(\mu) = 1 - \Phi \left( \frac{\sqrt{n}(c - \mu_0)}{\sigma} \right).
\]

In fact, it is also true that

\[
\sup_{P \in \mathcal{P}_1} [1 - \alpha_{T_c}(\mu)] = \Phi \left( \frac{\sqrt{n}(c - \mu_0)}{\sigma} \right).
\]

If we would like to use an \( \alpha \) as the level of significance, then the most effective way is to choose a \( c_\alpha \) (a test \( T_{c_\alpha}(X) \)) such that

\[
\alpha = \sup_{P \in \mathcal{P}_0} \alpha_{T_{c_\alpha}}(\mu),
\]

in which case \( c_\alpha \) must satisfy

\[
1 - \Phi \left( \frac{\sqrt{n}(c_\alpha - \mu_0)}{\sigma} \right) = \alpha,
\]

i.e., \( c_\alpha = \sigma z_{1-\alpha} / \sqrt{n} + \mu_0 \), where \( z_a = \Phi^{-1}(a) \).

In Chapter 6, it is shown that for any test \( T(X) \) satisfying

\[
\sup_{P \in \mathcal{P}_0} \alpha_T(P) \leq \alpha,
\]

\[
1 - \alpha_T(\mu) \geq 1 - \alpha_{T_{c_\alpha}}(\mu), \quad \mu > \mu_0.
\]
Choice of significance level

- The choice of a level of significance $\alpha$ is usually somewhat subjective.
- In most applications there is no precise limit to the size of $T$ that can be tolerated.
- Standard values, 0.10, 0.05, and 0.01, are often used for convenience.
- For most tests satisfying $\sup_{P \in \mathcal{P}_0} \alpha_T(P) \leq \alpha$, a small $\alpha$ leads to a “small” rejection region.

$p$-value

It is good practice to determine not only whether $H_0$ is rejected for a given $\alpha$ and a chosen test $T_\alpha$, but also the smallest possible level of significance at which $H_0$ would be rejected for the computed $T_\alpha(x)$, i.e.,

$$\hat{\alpha} = \inf\{\alpha \in (0, 1) : T_\alpha(x) = 1\}.$$

Such an $\hat{\alpha}$, which depends on $x$ and the chosen test and is a statistic, is called the $p$-value for the test $T_\alpha$. 
Example 2.29

Let us calculate the $p$-value for $T_{c_\alpha}$ in Example 2.28. Note that

$$\alpha = 1 - \Phi \left( \frac{\sqrt{n}(c_\alpha - \mu_0)}{\sigma} \right) > 1 - \Phi \left( \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} \right)$$

if and only if $\bar{x} > c_\alpha$ (or $T_{c_\alpha}(x) = 1$).

Hence

$$1 - \Phi \left( \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} \right) = \inf\{\alpha \in (0, 1) : T_{c_\alpha}(x) = 1\} = \hat{\alpha}(x)$$

is the $p$-value for $T_{c_\alpha}$.

It turns out that $T_{c_\alpha}(x) = I_{(0, \alpha)}(\hat{\alpha}(x))$.

Remarks

- With the additional information provided by $p$-values, using $p$-values is typically more appropriate than using fixed-level tests in a scientific problem.

- In some cases, a fixed level of significance is unavoidable when acceptance or rejection of $H_0$ is a required decision.
Randomized tests

In Example 2.28, \( \sup_{P \in \mathcal{P}_0} \alpha_T(P) = \alpha \) can always be achieved by a suitable choice of \( c \).
This is, however, not true in general.
We need to consider *randomized tests*.
Recall that a randomized decision rule is a probability measure \( \delta(x, \cdot) \) on the action space for any fixed \( x \).
Since the action space contains only two points, 0 and 1, for a hypothesis testing problem, any randomized test \( \delta(X, A) \) is equivalent to a statistic \( T(X) \in [0, 1] \) with \( T(x) = \delta(x, \{1\}) \) and \( 1 - T(x) = \delta(x, \{0\}) \).
A nonrandomized test is obviously a special case where \( T(x) \) does not take any value in \((0, 1)\).

For any randomized test \( T(X) \), we define the type I error probability to be \( \alpha_T(P) = E[T(X)], P \in \mathcal{P}_0 \), and the type II error probability to be \( 1 - \alpha_T(P) = E[1 - T(X)], P \in \mathcal{P}_1 \).
For a class of randomized tests, we would like to minimize \( 1 - \alpha_T(P) \) subject to \( \sup_{P \in \mathcal{P}_0} \alpha_T(P) = \alpha \).
Example 2.30

Assume that the sample $X$ has the binomial distribution $Bi(\theta, n)$ with an unknown $\theta \in (0, 1)$ and a fixed integer $n > 1$. Consider the hypotheses $H_0 : \theta \in (0, \theta_0]$ versus $H_1 : \theta \in (\theta_0, 1)$, where $\theta_0 \in (0, 1)$ is a fixed value.

Consider the following class of randomized tests:

$$T_{j,q}(X) = \begin{cases} 
1 & X > j \\
q & X = j \\
0 & X < j,
\end{cases}$$

where $j = 0, 1, \ldots, n - 1$ and $q \in [0, 1]$.

$$\alpha_{T_{j,q}}(\theta) = P(X > j) + qP(X = j) \quad 0 < \theta \leq \theta_0$$

$$1 - \alpha_{T_{j,q}}(\theta) = P(X < j) + (1 - q)P(X = j) \quad \theta_0 < \theta < 1.$$  

It can be shown that for any $\alpha \in (0, 1)$, there exist an integer $j$ and $q \in (0, 1)$ such that the size of $T_{j,q}$ is $\alpha$. 
Asymptotic approach

- In decision theory and inference, a key is to find moments and/or distributions of various statistics, which is difficult in general.
- When the sample size \( n \) is large, we may approximate the moments and distributions of statistics by those of the limiting distributions using the asymptotic tools discussed in §1.5, which leads to some asymptotic statistical procedures and asymptotic criteria for assessing performances.
- The asymptotic approach also provides a simpler solution (e.g., in computation) and requires less stringent model/loss assumption that itself is an approximation, as for a large sample, the statistical properties is less dependent on the loss functions and models.
- A major weakness of the asymptotic approach is that typically we don’t know whether a particular \( n \) in a problem is large enough.
- To overcome this difficulty, asymptotic results are often used with some numerical/empirical studies for selected values of \( n \) to examine the finite sample performance of asymptotic procedures.
Definition 2.10 (Consistency of point estimators)

Let $X = (X_1, ..., X_n)$ be a sample from $P \in \mathcal{P}$, $T_n(X)$ be an estimator of $\vartheta$ for every $n$, and $\{a_n\}$ be a sequence of positive constants, $a_n \to \infty$.

(i) $T_n(X)$ is consistent for $\vartheta$ iff $T_n(X) \to_p \vartheta$ w.r.t. any $P$.

(ii) $T_n(X)$ is $a_n$-consistent for $\vartheta$ iff $a_n[T_n(X) - \vartheta] = O_p(1)$ w.r.t. any $P$.

(iii) $T_n(X)$ is strongly consistent for $\vartheta$ iff $T_n(X) \to_{a.s.} \vartheta$ w.r.t. any $P$.

(iv) $T_n(X)$ is $L_r$-consistent for $\vartheta$ iff $T_n(X) \to_{L_r} \vartheta$ w.r.t. any $P$ for some fixed $r > 0$; if $r = 2$, $L_2$-consistency is called consistency in mse.

- Consistency is actually a concept relating to a sequence of estimators, $\{T_n\}$, but we just say “consistency of $T_n$” for simplicity.
- Each of the four types of consistency in Definition 2.10 describes the convergence of $T_n(X)$ to $\vartheta$ in some sense, as $n \to \infty$.
- A reasonable point estimator is expected to perform better, at least on the average, if more data (larger $n$) are available.
- Although the estimation error of $T_n$ for a fixed $n$ may never be 0, it is distasteful to use $T_n$ which, if sampling were to continue indefinitely, could still have a nonzero estimation error.
Methods of proving consistency

One or a combination of the WLLN, the CLT, Slutsky’s theorem, and the continuous mapping theorem (Theorems 1.10 and 1.12) can typically be applied to establish consistency of point estimators. For example, $\bar{X}$ is consistent for population mean $\mu$ (SLLN), and $g(\bar{X}^2)$ is consistent for $g(\mu)$ for any continuous function $g$.

Example 2.34

Let $X_1, \ldots, X_n$ be i.i.d. from an unknown $P$ with a continuous c.d.f. $F$ satisfying $F(\theta) = 1$ for some $\theta \in \mathbb{R}$ and $F(x) < 1$ for any $x < \theta$. Consider the largest order statistic $X_{(n)}$ as an estimator of $\theta$. For any $\varepsilon > 0$, $F(\theta - \varepsilon) < 1$ and

$$P(|X_{(n)} - \theta| \geq \varepsilon) = P(X_{(n)} \leq \theta - \varepsilon) = [F(\theta - \varepsilon)]^n,$$

which imply (according to Theorem 1.8(v)) $X_{(n)} \rightarrow a.s. \theta$, i.e., $X_{(n)}$ is strongly consistent for $\theta$.

If we assume that $F^{(i)}(\theta -)$, the $i$th-order left-hand derivative of $F$ at $\theta$, exists and vanishes for any $i \leq m$ and that $F^{(m+1)}(\theta -)$ exists and is nonzero, where $m$ is a nonnegative integer, then
Example 2.34 (continued)

\[ 1 - F(X_n) = \frac{(-1)^m F(m+1)(\theta - \cdot)}{(m+1)!} (\theta - X_n)^{m+1} + o\left(|\theta - X_n|^{m+1}\right) \text{ a.s.} \]

This result and the fact that \( P\left(n[1 - F(X_n)] \geq s\right) = (1 - s/n)^n \) imply that \((\theta - X_n)^{m+1} = O_P(n^{-1})\), i.e., \(X_n\) is \(n^{(m+1)^{-1}}\)-consistent.

If \(m = 0\), then \(X_n\) is \(n\)-consistent; if \(m = 1\), then \(X_n\) is \(\sqrt{n}\)-consistent.

The limiting distribution of \(n^{(m+1)^{-1}}(X_n - \theta)\) can be derived as follows. Let

\[
h_n(\theta) = \left[ \frac{(-1)^m (m+1)!}{nF(m+1)(\theta - \cdot)} \right]^{(m+1)^{-1}}.
\]

For \(t \leq 0\), by Slutsky's theorem,

\[
\lim_{n \to \infty} P\left(\frac{X_n - \theta}{h_n(\theta)} \leq t\right) = \lim_{n \to \infty} P\left(\left[\frac{\theta - X_n}{h_n(\theta)}\right]^{m+1} \geq (-t)^{m+1}\right)
\]

\[
= \lim_{n \to \infty} P\left(n[1 - F(X_n)] \geq (-t)^{m+1}\right)
\]

\[
= \lim_{n \to \infty} \left[1 - (-t)^{m+1}/n\right]^n = e^{-(t)^{m+1}}.
\]
Consistency is an essential requirement

- Like the admissibility, consistency is an essential requirement: any inconsistent estimators should not be used, but there are many consistent estimators and some may not be good.
- Thus, consistency should be used together with other criteria.

Approximate and asymptotic bias

- Unbiasedness is a criterion for point estimators (§2.3.2).
- In some cases, however, there is no unbiased estimator.
- Furthermore, having a “slight” bias in some cases may not be a bad idea.
- For a point estimator \( T_n(X) \) of \( \vartheta \) if \( E(T_n) \) 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 \textit{approximately unbiased}.
- There are many reasonable point estimators whose expectations are not well defined.
- It is desirable to define a concept of \textit{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) For a point estimator $T_n$ of $\vartheta$, 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 $P$, then $T_n$ is asymptotically unbiased.

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 $\{\tilde{b}_{T_n}(P)\}$).

Proposition 2.3 (asymptotic expectation is essentially unique)

For a sequence of random variables $\{\xi_n\}$, suppose both $E\xi/a_n$ and $E\eta/b_n$ are asymptotic expectations of $\xi_n$ defined by 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 $b_n/a_n \to 0$;
(c) $E\xi \neq 0$, $E\eta \neq 0$, and $(E\xi/a_n)/(E\eta/b_n) \to 1$. 
If $T_n$ is consistent for $\vartheta$, then $T_n = \vartheta + o_p(1)$ and $T_n$ is asymptotically unbiased, although $T_n$ may not be approximately unbiased.

### Precise order of asymptotic bias

When $a_n(T_n - \vartheta) \xrightarrow{d} Y$ 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.

A more precise order of the asymptotic bias of $T_n$ may be obtained (for comparing different estimators in terms of their asymptotic biases). In Example 2.34, $X(n)$ has the asymptotic bias $\tilde{b}_{X(n)}(P) = h_n(\theta)EY$, which is of order $n^{-(m+1)^{-1}}$.

Suppose that there is a sequence of random variables $\{\eta_n\}$ such that

$$a_n \eta_n \xrightarrow{d} Y \quad \text{and} \quad a_n^2 (T_n - \vartheta - \eta_n) \xrightarrow{d} W,$$

where $Y$ and $W$ are random variables with $EY = 0$ 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 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 $\Sigma = \text{Var}(X_1)$, $T_n = g(\bar{X})$, where $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ and $g$ is a function on $\mathbb{R}^k$ that is second-order differentiable at $\mu = E X_1 \in \mathbb{R}^k$. Consider $T_n$ as an estimator of $\vartheta = g(\mu)$.

By Taylor’s expansion,

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

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),

$$2^{-1} n(\bar{X} - \mu)^\tau \nabla^2 g(\mu)(\bar{X} - \mu) \to_d 2^{-1} Z_\Sigma^\tau \nabla^2 g(\mu) Z_\Sigma,$$

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

Thus,

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

is the $n^{-1}$ order asymptotic bias of $T_n = g(\bar{X})$, where tr($A$) denotes the trace of the matrix $A$. 
Example 2.35

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

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

$$E[\bar{X}(1 - \bar{X})] = E\bar{X} - E\bar{X}^2 = p - (E\bar{X})^2 - \text{Var}(\bar{X}) = p(1 - p) - p(1 - p)/n.$$ 

The exact bias of $T_n$ is the same as the $n^{-1}$ order asymptotic bias.

Consider next the estimation of $\vartheta = p^{-1}$.

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 $n$. 
Like the bias, the mse of an estimator $T_n$ of $\vartheta$, $\text{mse}_{T_n}(P) = E(T_n - \vartheta)^2$, 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 estimators of a parameter.

**Definition 2.12 (asymptotic variance and amse)**

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 as the asymptotic expectation of $(T_n - \vartheta)^2$, $\text{amse}_{T_n}(P) = EY^2/a_n^2$. The asymptotic variance of $T_n$ is defined as $\sigma^2_{T_n}(P) = \text{Var}(Y)/a_n^2$.

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

(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 $P$. 
The amse and asymptotic variance are the same iff $EY = 0$. In Example 2.33, $\text{amse}_X(P) = \sigma^2_X(P) = 4\mu^2\sigma^2/n$. In Example 2.34, $\sigma^2_X(n) = [h_n(\theta)]^2 \text{Var}(Y)$, $\text{amse}_X(n)(P) = [h_n(\theta)]^2 EY^2$.

When both $\text{mse}_X(T_n)(P)$ 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) (Exercise 115).

The following result shows that when the exact mse of $T_n$ exists, it is no smaller than the amse of $T_n$, and when they 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 $a_n \to a > 0$. If $a_n(T_n - \vartheta) \to_d Y$ 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.
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, $\sqrt{n}(T_{1n} - \vartheta) \rightarrow_d N(0, e^{-\theta}(1 - e^{-\theta}))$ 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 $E T_{2n} = e^{n\theta(e^{-1/n} - 1)}$.
Hence $\text{nb}_{T_{2n}}(\theta) \rightarrow \theta e^{-\theta}/2$.
Using Theorem 1.12 and the CLT, we can show that
$\sqrt{n}(T_{2n} - \vartheta) \rightarrow_d N(0, e^{-2\theta})$.
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 $T_{1n}$. 