Lecture 14: Inference and asymptotic approach

Example 2.28

Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathscr{R}$ and a known σ^2.
Consider the hypotheses $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$, where μ_0 is a fixed constant.
Since the sample mean $\bar{X}$ is sufficient for $\mu \in \mathscr{R}$, 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 $c \in \mathscr{R}$ is a fixed constant.

Let Φ 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,

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

In fact, it is also true that

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

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

$$\alpha = \sup_{P \in \mathscr{P}_0} \alpha_{T_{c_\alpha}}(\mu),$$

in which case c_α 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 $\sup_{P \in \mathscr{P}_0} \alpha_T(P) \le \alpha$,

$$1 - \alpha_T(\mu) \ge 1 - \alpha_{T_{c_\alpha}}(\mu), \qquad \mu > \mu_0.$$

Choice of significance level

- The choice of a level of significance α 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 $\sup_{P\in\mathscr{P}_0}\alpha_T(P)\leq\alpha$, a small α leads to a "small" rejection region.

p-value

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

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

Such an $\widehat{\alpha}$, which depends on x and the chosen test and is a statistic, is called the *p-value* for the test T_α.

Example 2.29

Let us calculate the p-value for T_{c_α} 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

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

is the p-value for T_{c_α}.
It turns out that $T_{c_\alpha}(x) = I_{(0,\alpha)}(\widehat{\alpha}(x))$.

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, $\sup_{P \in \mathscr{P}_0} \alpha_T(P) = \alpha$ 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 $\alpha_T(P) = E[T(X)]$, $P \in \mathscr{P}_0$, and the type II error probability to be $1 - \alpha_T(P) = E[1 - T(X)]$, $P \in \mathscr{P}_1$.

For a class of randomized tests, we would like to minimize $1 - \alpha_T(P)$ subject to $\sup_{P \in \mathscr{P}_0} \alpha_T(P) = \alpha$.

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

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

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

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 α.

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 $P \in \mathscr{P}$, $T_n(X)$ be an estimator of ϑ for every n, and $\{a_n\}$ be a sequence of positive constants, $a_n \to \infty$.

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

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

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

(iv) $T_n(X)$ is *L_r-consistent* for ϑ 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 ϑ 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 μ (SLLN), and $g(\bar{X}^2)$ is consistent for $g(\mu)$ for any continuous function g.

Example 2.34

Let $X_1,...,X_n$ be i.i.d. from an unknown P with a continuous c.d.f. F satisfying $F(\theta)=1$ for some $\theta \in \mathscr{R}$ and $F(x)<1$ for any $x<\theta$.
Consider the largest order statistic $X_{(n)}$ as an estimator of θ.
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)) $X_{(n)}\to_{a.s.}\theta$, i.e., $X_{(n)}$ is strongly consistent for θ.

If we assume that $F^{(i)}(\theta-)$, the ith-order left-hand derivative of F at θ, 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)

$$1-F(X_{(n)}) = \frac{(-1)^m F^{(m+1)}(\theta-)}{(m+1)!}(\theta - X_{(n)})^{m+1} + o\left(|\theta - X_{(n)}|^{m+1}\right) \quad \text{a.s.}$$

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

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

For $t \leq 0$, by Slutsky's theorem,

$$\begin{aligned}
\lim_{n\to\infty} P\left(\frac{X_{(n)}-\theta}{h_n(\theta)} \leq t\right) &= \lim_{n\to\infty} P\left(\left[\frac{\theta - X_{(n)}}{h_n(\theta)}\right]^{m+1} \geq (-t)^{m+1}\right) \\
&= \lim_{n\to\infty} P\left(n[1-F(X_{(n)})] \geq (-t)^{m+1}\right) \\
&= \lim_{n\to\infty} \left[1-(-t)^{m+1}/n\right]^n = e^{-(-t)^{m+1}}.
\end{aligned}$$

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 ϑ if $E(T_n)$ exists for every n and $\lim_{n\to\infty} E(T_n - \vartheta) = 0$ for any $P \in \mathscr{P}$, then T_n is said to be *approximately unbiased*.
- 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 $\xi, \xi_1, \xi_2, ...$ 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 ξ_n.

(ii) For a point estimator T_n of ϑ, 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 ξ_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 ϑ, 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 $a_n(T_n - \vartheta) \to_d Y$ 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.

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

$$a_n\eta_n \to_d Y \qquad \text{and} \qquad a_n^2(T_n - \vartheta - \eta_n) \to_d W,$$

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, η_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 $X_1,...,X_n$ are i.i.d. random k-vectors with finite $\Sigma = \mathrm{Var}(X_1)$, $T_n = g(\bar{X})$, where $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and g is a function on $\mathscr{R}^k$ that is second-order differentiable at $\mu = EX_1 \in \mathscr{R}^k$.
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 ∇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,

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

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

Consider first the estimation of $\vartheta = p(1-p)$.

Since $\mathrm{Var}(\bar{X}) = p(1-p)/n$, the n^{-1} order asymptotic bias of $T_n = \bar{X}(1-\bar{X})$ according to the formula $\mathrm{tr}\left(\nabla^2 g(\mu)\Sigma\right)/2n$ with $g(x) = x(1-x)$ is $-p(1-p)/n$.

On the other hand, a direct computation shows

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

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 $\mathrm{tr}\left(\nabla^2 g(\mu)\Sigma\right)/2n$ 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 ϑ, $\mathrm{mse}_{T_n}(P) = E(T_n - \vartheta)^2$, 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 ϑ 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 $\mathrm{amse}_{T_n}(P)$ or $\mathrm{amse}_{T_n}(\theta)$ if P is in a parametric family indexed by θ, is defined as the asymptotic expectation of $(T_n - \vartheta)^2$, $\mathrm{amse}_{T_n}(P) = EY^2/a_n^2$.
The asymptotic variance of T_n is defined as $\sigma^2_{T_n}(P) = \mathrm{Var}(Y)/a_n^2$.

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

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

The amse and asymptotic variance are the same iff $EY = 0$.
In Example 2.33, $\mathrm{amse}_{\bar{X}^2}(P) = \sigma^2_{\bar{X}^2}(P) = 4\mu^2\sigma^2/n$.
In Example 2.34, $\sigma^2_{X_{(n)}}(P) = [h_n(\theta)]^2\,\mathrm{Var}(Y)$,
$\mathrm{amse}_{X_{(n)}}(P) = [h_n(\theta)]^2 EY^2$.
When both $\mathrm{mse}_{T_n}(P)$ and $\mathrm{mse}_{T'_n}(P)$ exist, one may compare T_n and T'_n by evaluating the relative efficiency $\mathrm{mse}_{T_n}(P)/\mathrm{mse}_{T'_n}(P)$.
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 ϑ 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) $EY^2 \le \liminf_n E[a_n^2(T_n - \vartheta)^2]$ 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 $X_1, ..., X_n$ 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 $\mathrm{mse}_{T_{1n}}(\theta) = e^{-\theta}(1 - e^{-\theta})/n$.
Also, $\sqrt{n}(T_{1n} - \vartheta) \to_d N(0, e^{-\theta}(1 - e^{-\theta}))$ by the CLT.
Thus, in this case $\mathrm{amse}_{T_{1n}}(\theta) = \mathrm{mse}_{T_{1n}}(\theta)$.

Consider $T_{2n} = e^{-\bar{X}}$.
Note that $ET_{2n} = e^{n\theta(e^{-1/n} - 1)}$.
Hence $nb_{T_{2n}}(\theta) \to \theta e^{-\theta}/2$.
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), $\mathrm{amse}_{T_{2n}}(\theta) = e^{-2\theta}\theta/n$.
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}.