Stat 710: Mathematical Statistics
Lecture 19

Jun Shao

Department of Statistics
University of Wisconsin
Madison, WI 53706, USA

Mean vs median

Let F be a c.d.f. on $\mathscr{R}$ symmetric about $\theta \in \mathscr{R}$ with $F'(\theta) > 0$.

Then $\theta = \theta_{0.5}$ and is called the *median* of F.

If F has a finite mean, then θ is also equal to the mean.

We consider the estimation of θ based on i.i.d. X_i's from F.

If F is normal, it has been shown in previous chapters that the sample mean $\bar{X}$ is the UMVUE and MLE of θ and is asymptotically efficient.

On the other hand, if F is the c.d.f. of the Cauchy distribution $C(\theta,1)$, it follows from Exercise 78 in §1.6 that $\bar{X}$ has the same distribution as X_1, i.e., $\bar{X}$ is as variable as X_1, and is inconsistent as an estimator of θ.

Why does $\bar{X}$ perform so differently?

An important difference between the normal and Cauchy p.d.f.'s is that the former tends to 0 at the rate $e^{-x^2/2}$ as $|x| \to \infty$, whereas the latter tends to 0 at the much slower rate x^{-2}, which results in $\int |x| dF(x) = \infty$.

The poor performance of $\bar{X}$ in the Cauchy case is due to the high probability of getting extreme observations and the fact that $\bar{X}$ is sensitive to large changes in a few of the X_i's.

Mean vs median

This suggests the use of a robust estimator that discards some extreme observations.

The *sample median*, which is defined to be the 50%th sample quantile $\widehat{\theta}_{0.5}$ described in §5.3.1, is insensitive to the behavior of F as $|x| \to \infty$.

Since both the sample mean and the sample median can be used to estimate θ, a natural question is when is one better than the other, using a criterion such as the amse (asymptotic efficiency).

Unfortunately, a general answer does not exist, since the asymptotic relative efficiency between these two estimators depends on the unknown distribution F.

If F does not have a finite variance, then $\mathrm{Var}(\bar{X}) = \infty$ and $\bar{X}$ may be inconsistent.

In such a case the sample median is certainly preferred, since $\widehat{\theta}_{0.5}$ is consistent and asymptotically normal as long as $F'(\theta) > 0$, and may have a finite variance (Exercise 60).

The following example, which compares the sample mean and median in some cases, shows that the sample median can be better even if $\mathrm{Var}(X_1) < \infty$.

Example 5.10 (asymptotic efficiency and robustness)

Suppose that $\mathrm{Var}(X_1) < \infty$.
Then, by the CLT,

$$\sqrt{n}(\bar{X} - \theta) \to_d N(0, \mathrm{Var}(X_1)).$$

By Theorem 5.10(iv),

$$\sqrt{n}(\widehat{\theta}_{0.5} - \theta) \to_d N(0, [2F'(\theta)]^{-2}).$$

Hence, the asymptotic relative efficiency of $\widehat{\theta}_{0.5}$ w.r.t. $\bar{X}$ is

$$e(F) = 4[F'(\theta)]^2 \,\mathrm{Var}(X_1).$$

- If F is the c.d.f. of $N(\theta,\sigma^2)$, then $\mathrm{Var}(X_1) = \sigma^2$, $F'(\theta) = (\sqrt{2\pi}\sigma)^{-1}$, and $e(F) = 2/\pi = 0.637$.
- If F is the c.d.f. of the logistic distribution $LG(\theta,\sigma)$, then $\mathrm{Var}(X_1) = \sigma^2\pi^2/3$, $F'(\theta) = (4\sigma)^{-1}$, and $e(F) = \pi^2/12 = 0.822$.
- If $F(x) = F_0(x-\theta)$ and F_0 is the c.d.f. of the t-distribution t_ν with $\nu \geq 3$, then $\mathrm{Var}(X_1) = \nu/(\nu-2)$, $F'(\theta) = \Gamma(\frac{\nu+1}{2})/[\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})]$, $e(F) = 1.62$ when $\nu = 3$, $e(F) = 1.12$ when $\nu = 4$, and $e(F) = 0.96$ when $\nu = 5$.

Example 5.10 (continued)

- If F is the c.d.f. of the double exponential distribution $DE(\theta,\sigma)$, then $F'(\theta) = (2\sigma)^{-1}$ and $e(F) = 2$.
- Consider the Tukey model

$$F(x) = (1-\varepsilon)\Phi\left(\frac{x-\theta}{\sigma}\right) + \varepsilon\Phi\left(\frac{x-\theta}{\tau\sigma}\right),$$

where $\sigma > 0$, $\tau > 0$, and $0 < \varepsilon < 1$. Then $\mathrm{Var}(X_1) = (1-\varepsilon)\sigma^2 + \varepsilon\tau^2\sigma^2$, $F'(\theta) = (1-\varepsilon+\varepsilon/\tau)/(\sqrt{2\pi}\sigma)$, and $e(F) = 2(1-\varepsilon+\varepsilon\tau^2)(1-\varepsilon+\varepsilon/\tau)^2/\pi$. Note that $\lim_{\varepsilon\to 0} e(F) = 2/\pi$ and $\lim_{\tau\to\infty} e(F) = \infty$.

Trimmed sample mean

Since the sample median uses at most two actual values of x_i's, it may go too far in discarding observations, which results in a possible loss of efficiency.

The trimmed sample mean is a natural compromise between the sample mean and median.

Example 5.10 (continued)

- If F is the c.d.f. of the double exponential distribution $DE(\theta, \sigma)$, then $F'(\theta) = (2\sigma)^{-1}$ and $e(F) = 2$.
- Consider the Tukey model

$$F(x) = (1-\varepsilon)\Phi\left(\tfrac{x-\theta}{\sigma}\right) + \varepsilon\Phi\left(\tfrac{x-\theta}{\tau\sigma}\right),$$

where $\sigma > 0$, $\tau > 0$, and $0 < \varepsilon < 1$. Then
$\mathrm{Var}(X_1) = (1-\varepsilon)\sigma^2 + \varepsilon\tau^2\sigma^2$, $F'(\theta) = (1-\varepsilon+\varepsilon/\tau)/(\sqrt{2\pi}\sigma)$, and
$e(F) = 2(1-\varepsilon+\varepsilon\tau^2)(1-\varepsilon+\varepsilon/\tau)^2/\pi$. Note that $\lim_{\varepsilon\to 0} e(F) = 2/\pi$
and $\lim_{\tau\to\infty} e(F) = \infty$.

Trimmed sample mean

Since the sample median uses at most two actual values of x_i's, it may go too far in discarding observations, which results in a possible loss of efficiency.
The trimmed sample mean is a natural compromise between the sample mean and median.

The α-trimmed sample mean and its properties

The α-trimmed sample mean is defined as

$$\bar{X}_\alpha = \frac{1}{(1-2\alpha)n} \sum_{j=m_\alpha+1}^{n-m_\alpha} X_{(j)},$$

where m_α is the integer part of $n\alpha$ and $\alpha \in (0, \frac{1}{2})$.
It discards the m_α smallest and m_α largest observations.
The sample mean and median can be viewed as two extreme cases of $\bar{X}_\alpha$ as $\alpha \to 0$ and $\frac{1}{2}$, respectively.
If $F(x) = F_0(x-\theta)$, where F_0 is symmetric about 0 and has a Lebesgue p.d.f. positive in the range of X_1, then

$$\sqrt{n}(\bar{X}_\alpha - \theta) \to_d N(0, \sigma_\alpha^2),$$

where

$$\sigma_\alpha^2 = \frac{2}{(1-2\alpha)^2} \left\{ \int_0^{F_0^{-1}(1-\alpha)} x^2 dF_0(x) + \alpha [F_0^{-1}(1-\alpha)]^2 \right\}.$$

(These will be further discussed in the next lecture.)

Comparisons

From the asymptotic normality of $\bar{X}_\alpha$, the asymptotic relative efficiency between $\bar{X}_\alpha$ and the sample mean $\bar{X}$ is

$$e_{\bar{X}_\alpha,\bar{X}}(F) = \mathrm{Var}(X_1)/\sigma_\alpha^2.$$

Lehmann (1983, §5.4) provides various values of the asymptotic relative efficiency $e_{\bar{X}_\alpha,\bar{X}}(F)$.
For instance, when $F(x) = F_0(x-\theta)$ and F_0 is the c.d.f. of the t-distribution t_3, $e_{\bar{X}_\alpha,\bar{X}}(F) = 1.70$, 1.91, and 1.97 for $\alpha = 0.05$, 0.125, and 0.25, respectively;
when

$$F(x) = (1-\varepsilon)\Phi\left(\tfrac{x-\theta}{\sigma}\right) + \varepsilon\Phi\left(\tfrac{x-\theta}{\tau\sigma}\right)$$

with $\tau = 3$ and $\varepsilon = 0.05$, $e_{\bar{X}_\alpha,\bar{X}}(F) = 1.20$, 1.19, and 1.09 for $\alpha = 0.05$, 0.125, and 0.25, respectively;
when $\tau = 3$ and $\varepsilon = 0.01$, $e_{\bar{X}_\alpha,\bar{X}}(F) = 1.04$, 0.98, and 0.89 for $\alpha = 0.05$, 0.125, and 0.25, respectively.

M-estimators

Note that the sample mean $\bar{X}$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2=\min_{t\in\Theta}\frac{1}{n}\sum_{i=1}^{n}(X_i-t)^2\min_{t\in\Theta}\int(x-t)^2dF_n$$

This idea can be generalized to get a class of estimators obtained by minimizing some functions.

Let $\rho(x,t)$ be a Borel function on $\mathscr{R}^d\times\mathscr{R}$ and $\Theta\subset\mathscr{R}$ be an open set. An *M-functional* is defined to be a solution of

$$\int\rho(x,T(G))dG(x)=\min_{t\in\Theta}\int\rho(x,t)dG(x),\qquad G\in\mathscr{F}$$

For $X_1,...,X_n$ i.i.d. from $F\in\mathscr{F}$, $T(F_n)$ is called an *M-estimator* of $T(F)$.

$$\int\rho(x,T(F_n))dF_n(x)=\min_{t\in\Theta}\int\rho(x,t)dF_n(x)$$

i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}\rho(X_i,T(F_n))=\min_{t\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\rho(X_i,t)$$

M-estimators

Assume that $\psi(x,t) = \partial\rho(x,t)/\partial t$ exists a.e. and

$$\lambda_G(t) = \int \psi(x,t)dG(x) = \frac{\partial}{\partial t}\int \rho(x,t)dG(x).$$

Then $\lambda_G(T(G)) = 0$ and $T(F_n)$ is a solution of

$$\sum_{i=1}^n \psi(X_i,t) = 0.$$

Example 5.7

The following are some examples of M-estimators.

(i) If $\rho(x,t) = (x-t)^2/2$, then $T(F_n) = \bar{X}$ is the sample mean.

(ii) If $\rho(x,t) = |x-t|^p/p$, where $p \in [1,2)$, then

$$\psi(x,t) = \begin{cases} |x-t|^{p-1} & x \le t \\ -|x-t|^{p-1} & x > t. \end{cases}$$

When $p = 1$, $T(F_n)$ is the sample median. When $1 < p < 2$, $T(F_n)$ is called the pth least absolute deviations estimator or the minimum L_p distance estimator.

M-estimators

Assume that $\psi(x,t) = \partial\rho(x,t)/\partial t$ exists a.e. and

$$\lambda_G(t) = \int \psi(x,t)dG(x) = \frac{\partial}{\partial t}\int \rho(x,t)dG(x).$$

Then $\lambda_G(T(G)) = 0$ and $T(F_n)$ is a solution of

$$\sum_{i=1}^{n} \psi(X_i,t) = 0.$$

Example 5.7

The following are some examples of M-estimators.
(i) If $\rho(x,t) = (x-t)^2/2$, then $T(F_n) = \bar{X}$ is the sample mean.
(ii) If $\rho(x,t) = |x-t|^p/p$, where $p \in [1,2)$, then

$$\psi(x,t) = \begin{cases} |x-t|^{p-1} & x \le t \\ -|x-t|^{p-1} & x > t. \end{cases}$$

When $p = 1$, $T(F_n)$ is the sample median. When $1 < p < 2$, $T(F_n)$ is called the pth least absolute deviations estimator or the minimum L_p distance estimator.

Example 5.7 (continued)

(iii) Let $\mathscr{F}_0 = \{f_\theta : \theta \in \Theta\}$ be a parametric family of p.d.f.'s with $\Theta \subset \mathscr{R}$ and $\rho(x,t) = -\log f_t(x)$.
Then $T(F_n)$ is an MLE.
Thus, M-estimators are extensions of MLE's in parametric models.
(iv) Let $C > 0$ be a constant.
Huber (1964) considers

$$\rho(x,t) = \begin{cases} \frac{1}{2}(x-t)^2 & |x-t| \le C \\ \frac{1}{2}C^2 & |x-t| > C \end{cases}$$

with

$$\psi(x,t) = \begin{cases} t-x & |x-t| \le C \\ 0 & |x-t| > C. \end{cases}$$

The corresponding $T(F_n)$ is a type of trimmed sample mean.
(v) Let $C > 0$ be a constant.
Huber (1964) considers

$$\rho(x,t) = \begin{cases} \frac{1}{2}(x-t)^2 & |x-t| \le C \\ C|x-t| - \frac{1}{2}C^2 & |x-t| > C \end{cases}$$

Example 5.7 (continued)

with

$$\psi(x,t) = \begin{cases} C & t-x > C \\ t-x & |x-t| \le C \\ -C & t-x < -C. \end{cases}$$

The corresponding $T(F_n)$ is a type of Winsorized sample mean.

(vi) Hampel (1974) considers $\psi(x,t) = \psi_0(t-x)$ with $\psi_0(s) = -\psi_0(-s)$ and

$$\psi_0(s) = \begin{cases} s & 0 \le s \le a \\ a & a < s \le b \\ \frac{a(c-s)}{c-b} & b < s \le c \\ 0 & s > c, \end{cases}$$

where $0 < a < b < c$ are constants.

A smoothed version of ψ_0 is

$$\psi_1(s) = \begin{cases} \sin(as) & 0 \le s < \pi/a \\ 0 & s > \pi/a. \end{cases}$$

Theorem 5.7

Let $X_1,...,X_n$ be i.i.d. from F and T be an M-functional.
Assume that ψ is a bounded and continuous function on $\mathscr{R}^d \times \mathscr{R}$ and that $\lambda_F(t)$ is continuously differentiable at $T(F)$ and $\lambda_F'(T(F)) \neq 0$.
Then

$$\sqrt{n}[T(F_n) - T(F)] \to_d N(0, \sigma_F^2)$$

with

$$\sigma_F^2 = \frac{\int [\psi(x, T(F))]^2 dF(x)}{[\lambda_F'(T(F))]^2}.$$

Example 5.13

Consider Huber's ψ given in Example 5.7(v).
Assume that F is continuous at $\theta - C$ and $\theta + C$.
Then

$$\sigma_F^2 = \frac{\int_{\theta-C}^{\theta+C} (\theta - x)^2 dF(x) + C^2 F(\theta - C) + C^2[1 - F(\theta + C)]}{[F(\theta + C) - F(\theta - C)]^2}$$

Asymptotic relative efficiency between Huber's M-estimator and the sample mean can be obtained.

Theorem 5.7

Let $X_1, ..., X_n$ be i.i.d. from F and T be an M-functional.
Assume that ψ is a bounded and continuous function on $\mathscr{R}^d \times \mathscr{R}$ and that $\lambda_F(t)$ is continuously differentiable at $T(F)$ and $\lambda_F'(T(F)) \neq 0$.
Then

$$\sqrt{n}[T(F_n) - T(F)] \to_d N(0, \sigma_F^2)$$

with

$$\sigma_F^2 = \frac{\int [\psi(x, T(F))]^2 dF(x)}{[\lambda_F'(T(F))]^2}.$$

Example 5.13

Consider Huber's ψ given in Example 5.7(v).
Assume that F is continuous at $\theta - C$ and $\theta + C$.
Then

$$\sigma_F^2 = \frac{\int_{\theta-C}^{\theta+C} (\theta - x)^2 dF(x) + C^2 F(\theta - C) + C^2 [1 - F(\theta + C)]}{[F(\theta + C) - F(\theta - C)]^2}$$

Asymptotic relative efficiency between Huber's M-estimator and the sample mean can be obtained.