Lecture 11: Sample quantiles, robustness, and asymptotic efficiency

Estimation of quantiles (percentiles)

Suppose that $X_1,...,X_n$ are i.i.d. random variables from an unknown nonparametric F

For $p \in (0,1)$,

$$G^{-1}(p) = \inf\{x : G(x) \geq p\}$$

is the pth quantile for any c.d.f. G on $\mathscr{R}$.

Quantiles of F are often the parameters of interest.

$\theta_p = F^{-1}(p) = p$th quantile of F

$F_n =$ empirical c.d.f. based on $X_1,...,X_n$

$\widehat{\theta}_p = F_n^{-1}(p) =$ the pth sample quantile.

$$\widehat{\theta}_p = c_{np} X_{(m_p)} + (1 - c_{np}) X_{(m_p+1)},$$

where $X_{(j)}$ is the jth order statistic, m_p is the integer part of np, $c_{np} = 1$ if np is an integer, and $c_{np} = 0$ if np is not an integer.

Thus, $\widehat{\theta}_p$ is a linear function of order statistics.

$F(\theta_p-) = \lim_{x\to\theta_p, x<\theta_p} F(x)$
$F(\theta_p) = \lim_{x\to\theta_p, x>\theta_p} F(x)$
$F(\theta_p-) \le p \le F(\theta_p)$
F is not flat in a neighborhood of θ_p if and only if $p < F(\theta_p + \varepsilon)$ for any $\varepsilon > 0$.

Theorem 5.9

Let $X_1, ..., X_n$ be i.i.d. random variables from a c.d.f. F satisfying $p < F(\theta_p + \varepsilon)$ for any $\varepsilon > 0$. Then, for every $\varepsilon > 0$ and $n = 1, 2, ...$,

$$P\left(|\widehat{\theta}_p - \theta_p| > \varepsilon\right) \le 2Ce^{-2n\delta_\varepsilon^2},$$

where δ_ε is the smaller of $F(\theta_p + \varepsilon) - p$ and $p - F(\theta_p - \varepsilon)$ and C is the same constant in Lemma 5.1(i).

Remarks

- Theorem 5.9 implies that $\widehat{\theta}_p$ is strongly consistent for θ_p (exercise)
- Theorem 5.9 implies that $\widehat{\theta}_p$ is $\sqrt{n}$-consistent for θ_p if $F'(\theta_p-)$ and $F'(\theta_p+)$ (the left and right derivatives of F at θ_p) exist (exercise).

Proof of Theorem 5.9

Let $\varepsilon > 0$ be fixed.
Note that, for any c.d.f. G on $\mathscr{R}$,

$$G(x) \geq t \text{ if and only if } x \geq G^{-1}(t)$$

(exercise).
Hence

$$\begin{aligned}
P\big(\widehat{\theta}_p > \theta_p + \varepsilon\big) &= P\big(p > F_n(\theta_p + \varepsilon)\big) \\
&= P\big(F(\theta_p + \varepsilon) - F_n(\theta_p + \varepsilon) > F(\theta_p + \varepsilon) - p\big) \\
&\leq P\big(\rho_\infty(F_n, F) > \delta_\varepsilon\big) \\
&\leq Ce^{-2n\delta_\varepsilon^2},
\end{aligned}$$

where the last inequality follows from DKW's inequality (Lemma 5.1(i)).
Similarly,

$$P\big(\widehat{\theta}_p < \theta_p - \varepsilon\big) \leq Ce^{-2n\delta_\varepsilon^2}.$$

This completes the proof.

The distribution of a sample quantile

The exact distribution of $\widehat{\theta}_p$ can be obtained as follows.
Since $nF_n(t)$ has the binomial distribution $Bi(F(t), n)$ for any $t \in \mathscr{R}$,

$$\begin{aligned} P\big(\widehat{\theta}_p \le t\big) &= P\big(F_n(t) \ge p\big) \\ &= \sum_{i=l_p}^{n} \binom{n}{i} [F(t)]^i [1 - F(t)]^{n-i}, \end{aligned}$$

where $l_p = np$ if np is an integer and $l_p = 1+$ the integer part of np if np is not an integer.
If F has a Lebesgue p.d.f. f, then $\widehat{\theta}_p$ has the Lebesgue p.d.f.

$$\varphi_n(t) = n \binom{n-1}{l_p - 1} [F(t)]^{l_p - 1} [1 - F(t)]^{n-l_p} f(t).$$

This can be shown by differentiating $P\big(F_n(t) \ge p\big)$ term by term, which leads to

$$
\begin{aligned}
\varphi_n(t) &= \sum_{i=l_p}^{n} \binom{n}{i} i[F(t)]^{i-1}[1-F(t)]^{n-i} f(t) \\
&\quad - \sum_{i=l_p}^{n} \binom{n}{i} (n-i)[F(t)]^{i}[1-F(t)]^{n-i-1} f(t) \\
&= \binom{n}{l_p} l_p [F(t)]^{l_p-1}[1-F(t)]^{n-l_p} f(t) \\
&\quad + n \sum_{i=l_p+1}^{n} \binom{n-1}{i-1} [F(t)]^{i-1}[1-F(t)]^{n-i} f(t) \\
&\quad - n \sum_{i=l_p}^{n-1} \binom{n-1}{i} [F(t)]^{i}[1-F(t)]^{n-i-1} f(t) \\
&= n \binom{n-1}{l_p-1} [F(t)]^{l_p-1}[1-F(t)]^{n-l_p} f(t).
\end{aligned}
$$

The following result provides an asymptotic distribution for $\sqrt{n}(\widehat{\theta}_p - \theta_p)$.

Theorem 5.10

Let $X_1, ..., X_n$ be i.i.d. random variables from F.
(i) If $F(\theta_p) = p$, then $P(\sqrt{n}(\widehat{\theta}_p - \theta_p) \leq 0) \to \Phi(0) = \frac{1}{2}$, where Φ is the c.d.f. of the standard normal.
(ii) If F is continuous at θ_p and there exists $F'(\theta_p-) > 0$, then

$$P\left(\sqrt{n}(\widehat{\theta}_p - \theta_p) \leq t\right) \to \Phi(t/\sigma_F^-), \qquad t < 0,$$

where $\sigma_F^- = \sqrt{p(1-p)}/F'(\theta_p-)$.
(iii) If F is continuous at θ_p and there exists $F'(\theta_p+) > 0$, then

$$P\left(\sqrt{n}(\widehat{\theta}_p - \theta_p) \leq t\right) \to \Phi(t/\sigma_F^+), \qquad t > 0,$$

where $\sigma_F^+ = \sqrt{p(1-p)}/F'(\theta_p+)$.
(iv) If $F'(\theta_p)$ exists and is positive, then

$$\sqrt{n}(\widehat{\theta}_p - \theta_p) \to_d N(0, \sigma_F^2),$$

where $\sigma_F = \sqrt{p(1-p)}/F'(\theta_p)$.

Proof

The proof of (i) is left as an exercise.

Part (iv) is a direct consequence of (i)-(iii) and the proofs of (ii) and (iii) are similar.

Thus, we only give a proof for (iii).

Let $t > 0$, $p_{nt} = F(\theta_p + t\sigma_F^+ n^{-1/2})$, $c_{nt} = \sqrt{n}(p_{nt} - p)/\sqrt{p_{nt}(1-p_{nt})}$, and $Z_{nt} = [B_n(p_{nt}) - np_{nt}]/\sqrt{np_{nt}(1-p_{nt})}$, where $B_n(q)$ denotes a random variable having the binomial distribution $Bi(q,n)$.

Then

$$\begin{aligned} P\big(\widehat{\theta}_p \le \theta_p + t\sigma_F^+ n^{-1/2}\big) &= P\big(p \le F_n(\theta_p + t\sigma_F^+ n^{-1/2})\big) \\ &= P\big(Z_{nt} \ge -c_{nt}\big). \end{aligned}$$

Under the assumed conditions on F, $p_{nt} \to p$ and $c_{nt} \to t$.

Hence, the result follows from

$$P\big(Z_{nt} < -c_{nt}\big) - \Phi(-c_{nt}) \to 0.$$

But this follows from the CLT (Example 1.33) and Pólya's theorem (Proposition 1.16).

If $F'(\theta_p-)$ and $F'(\theta_p+)$ exist and are positive, but $F'(\theta_p-) \neq F'(\theta_p+)$, then the asymptotic distribution of $\sqrt{n}(\widehat{\theta}_p - \theta_p)$ has the c.d.f.

$$\Phi(t/\sigma_F^-) I_{(-\infty,0)}(t) + \Phi(t/\sigma_F^+) I_{[0,\infty)}(t),$$

a mixture of two normal distributions.
An example of such a case when $p = 1/2$ is

$$F(x) = x I_{[0,\frac{1}{2})}(x) + (2x - \tfrac{1}{2}) I_{[\frac{1}{2},\frac{3}{4})}(x) + I_{[\frac{3}{4},\infty)}(x).$$

Bahadur's representation

When $F'(\theta_p-) = F'(\theta_p+) = F'(\theta_p) > 0$, Theorem 5.9 shows that the asymptotic distribution of $\sqrt{n}(\widehat{\theta}_p - \theta_p)$ is the same as that of $\sqrt{n}[F_n(\theta_p) - F(\theta_p)]/F'(\theta_p)$.
The next result reveals a stronger relationship between sample quantiles and the empirical c.d.f.

Theorem 5.11 (Bahadur's representation)

Let $X_1, ..., X_n$ be i.i.d. random variables from F.
If $F'(\theta_p)$ exists and is positive, then

$$\sqrt{n}(\widehat{\theta}_p - \theta_p) = \sqrt{n}[F_n(\theta_p) - F(\theta_p)]/F'(\theta_p) + o_p(1).$$

Proof

Let $t \in \mathscr{R}$, $\theta_{nt} = \theta_p + tn^{-1/2}$, $Z_n(t) = \sqrt{n}[F(\theta_{nt}) - F_n(\theta_{nt})]/F'(\theta_p)$, and $U_n(t) = \sqrt{n}[F(\theta_{nt}) - F_n(\widehat{\theta}_p)]/F'(\theta_p)$.
It can be shown (exercise) that

$$Z_n(t) - Z_n(0) = o_p(1).$$

Since $|p - F_n(\widehat{\theta}_p)| \leq n^{-1}$,

$$\begin{aligned} U_n(t) &= \sqrt{n}[F(\theta_{nt}) - p + p - F_n(\widehat{\theta}_p)]/F'(\theta_p) \\ &= \sqrt{n}[F(\theta_{nt}) - p]/F'(\theta_p) + O(n^{-1/2}) \\ &\to t. \end{aligned}$$

Let $\xi_n = \sqrt{n}(\widehat{\theta}_p - \theta_p)$.
Then, for any $t \in \mathscr{R}$ and $\varepsilon > 0$,

$$\begin{aligned} P\big(\xi_n \leq t, Z_n(0) \geq t + \varepsilon\big) &= P\big(Z_n(t) \leq U_n(t), Z_n(0) \geq t + \varepsilon\big) \\ &\leq P\big(|Z_n(t) - Z_n(0)| \geq \varepsilon/2\big) \\ &\quad + P\big(|U_n(t) - t| \geq \varepsilon/2\big) \\ &\to 0 \end{aligned}$$

because, if $Z_n(t) \le U_n(t)$, $Z_n(0) \ge t+\varepsilon$, and $|Z_n(t) - Z_n(0)| < \varepsilon/2$, then

$$-\varepsilon/2 < Z_n(t) - Z_n(0) \le U_n(t) - Z_n(0) \le U_n(t) - (t+\varepsilon)$$

i.e., $U_n(t) - t > \varepsilon/2$.
Similarly,

$$P\big(\xi_n \ge t+\varepsilon, Z_n(0) \le t\big) \to 0.$$

It follows from the result in Exercise 128 of §1.6 that

$$\xi_n - Z_n(0) = o_p(1),$$

which is what we need to prove.

Corollary 5.1

Let $X_1,...,X_n$ be i.i.d. random variables from F having positive derivatives at θ_{p_j}, where $0 < p_1 < \cdots < p_m < 1$ are fixed constants. Then

$$\sqrt{n}[(\widehat{\theta}_{p_1},...,\widehat{\theta}_{p_m}) - (\theta_{p_1},...,\theta_{p_m})] \to_d N_m(0,D),$$

where D is the $m \times m$ symmetric matrix whose (i,j)th element is

$$p_i(1-p_j)/[F'(\theta_{p_i})F'(\theta_{p_j})], \qquad i \le j.$$

Robustness and efficiency: median vs mean

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.

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 $\nu = 5$.

- 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

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

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.