

# Stat 710: Mathematical Statistics

## Lecture 18

Jun Shao

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

# Lecture 18: Sample quantiles and their asymptotic properties

## Estimation of quantiles (percentiles)

Suppose that  $X_1, \dots, X_n$  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  $p$ th quantile for any c.d.f.  $G$  on  $\mathcal{R}$ .

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  $X_1, \dots, X_n$

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

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

where  $X_{(j)}$  is the  $j$ th 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,  $\hat{\theta}_p$  is a linear function of order statistics.

$$F(\theta_p-) = \lim_{x \rightarrow \theta_p, x < \theta_p} F(x)$$

$$F(\theta_p) = \lim_{x \rightarrow \theta_p, x > \theta_p} F(x)$$

$$F(\theta_p-) \leq p \leq F(\theta_p)$$

$F$  is not flat in a neighborhood of  $\theta_p$  if and only if  $p < F(\theta_p + \varepsilon)$  for any  $\varepsilon > 0$ .

## Theorem 5.9

Let  $X_1, \dots, X_n$  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  $n = 1, 2, \dots$ ,

$$P(|\hat{\theta}_p - \theta_p| > \varepsilon) \leq 2Ce^{-2n\delta_\varepsilon^2},$$

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

$$F(\theta_p-) = \lim_{x \rightarrow \theta_p, x < \theta_p} F(x)$$

$$F(\theta_p) = \lim_{x \rightarrow \theta_p, x > \theta_p} F(x)$$

$$F(\theta_p-) \leq p \leq F(\theta_p)$$

$F$  is not flat in a neighborhood of  $\theta_p$  if and only if  $p < F(\theta_p + \varepsilon)$  for any  $\varepsilon > 0$ .

## Theorem 5.9

Let  $X_1, \dots, X_n$  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  $n = 1, 2, \dots$ ,

$$P(|\hat{\theta}_p - \theta_p| > \varepsilon) \leq 2Ce^{-2n\delta_\varepsilon^2},$$

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

$$F(\theta_p-) = \lim_{x \rightarrow \theta_p, x < \theta_p} F(x)$$

$$F(\theta_p) = \lim_{x \rightarrow \theta_p, x > \theta_p} F(x)$$

$$F(\theta_p-) \leq p \leq F(\theta_p)$$

$F$  is not flat in a neighborhood of  $\theta_p$  if and only if  $p < F(\theta_p + \varepsilon)$  for any  $\varepsilon > 0$ .

## Theorem 5.9

Let  $X_1, \dots, X_n$  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  $n = 1, 2, \dots$ ,

$$P(|\hat{\theta}_p - \theta_p| > \varepsilon) \leq 2Ce^{-2n\delta_\varepsilon^2},$$

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

## Proof of Theorem 5.9

Let  $\varepsilon > 0$  be fixed.

Note that, for any c.d.f.  $G$  on  $\mathcal{R}$ ,

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

(exercise).

Hence

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

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

Similarly,

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

This completes the proof.

## The distribution of a sample quantile

The exact distribution of  $\hat{\theta}_p$  can be obtained as follows.

Since  $nF_n(t)$  has the binomial distribution  $Bi(F(t), n)$  for any  $t \in \mathcal{R}$ ,

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

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  $\hat{\theta}_p$  has the Lebesgue p.d.f.

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

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

## Theorem 5.10

Let  $X_1, \dots, X_n$  be i.i.d. random variables from  $F$ .

(i) If  $F(\theta_p) = p$ , then  $P(\sqrt{n}(\hat{\theta}_p - \theta_p) \leq 0) \rightarrow \Phi(0) = \frac{1}{2}$ , where  $\Phi$  is the c.d.f. of the standard normal.

(ii) If  $F$  is continuous at  $\theta_p$  and there exists  $F'(\theta_p-) > 0$ , then

$$P(\sqrt{n}(\hat{\theta}_p - \theta_p) \leq t) \rightarrow \Phi(t/\sigma_F^-), \quad t < 0,$$

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

(iii) If  $F$  is continuous at  $\theta_p$  and there exists  $F'(\theta_p+) > 0$ , then

$$P(\sqrt{n}(\hat{\theta}_p - \theta_p) \leq t) \rightarrow \Phi(t/\sigma_F^+), \quad t > 0,$$

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

(iv) If  $F'(\theta_p)$  exists and is positive, then

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

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

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

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

Hence, the result follows from

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

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  $\sqrt{n}(\hat{\theta}_p - \theta_p)$  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

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

## Bahadur's representation

When  $F'(\theta_p^-) = F'(\theta_p^+) = F'(\theta_p) > 0$ , Theorem 5.9 shows that the asymptotic distribution of  $\sqrt{n}(\hat{\theta}_p - \theta_p)$  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  $X_1, \dots, X_n$  be i.i.d. random variables from  $F$ .

If  $F'(\theta_p)$  exists and is positive, then

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

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  $\sqrt{n}(\hat{\theta}_p - \theta_p)$  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

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

## Bahadur's representation

When  $F'(\theta_p^-) = F'(\theta_p^+) = F'(\theta_p) > 0$ , Theorem 5.9 shows that the asymptotic distribution of  $\sqrt{n}(\hat{\theta}_p - \theta_p)$  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  $X_1, \dots, X_n$  be i.i.d. random variables from  $F$ .

If  $F'(\theta_p)$  exists and is positive, then

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

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  $\sqrt{n}(\hat{\theta}_p - \theta_p)$  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

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

## Bahadur's representation

When  $F'(\theta_p^-) = F'(\theta_p^+) = F'(\theta_p) > 0$ , Theorem 5.9 shows that the asymptotic distribution of  $\sqrt{n}(\hat{\theta}_p - \theta_p)$  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  $X_1, \dots, X_n$  be i.i.d. random variables from  $F$ .

If  $F'(\theta_p)$  exists and is positive, then

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

Let  $t \in \mathcal{R}$ ,  $\theta_{nt} = \theta_p + tn^{-1/2}$ ,  $Z_n(t) = \sqrt{n}[F(\theta_{nt}) - F_n(\theta_{nt})]/F'(\theta_p)$ , and  $U_n(t) = \sqrt{n}[F(\theta_{nt}) - F_n(\hat{\theta}_p)]/F'(\theta_p)$ .

It can be shown (exercise) that

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

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

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

Let  $\xi_n = \sqrt{n}(\hat{\theta}_p - \theta_p)$ .

Then, for any  $t \in \mathcal{R}$  and  $\varepsilon > 0$ ,

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

## Proof (continued)

Similarly,

$$P(\xi_n \geq t + \varepsilon, Z_n(0) \leq t) \rightarrow 0.$$

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

$$\sqrt{n}[(\hat{\theta}_{p_1}, \dots, \hat{\theta}_{p_m}) - (\theta_{p_1}, \dots, \theta_{p_m})] \rightarrow_d N_m(0, D),$$

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

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

## Proof (continued)

Similarly,

$$P(\xi_n \geq t + \varepsilon, Z_n(0) \leq t) \rightarrow 0.$$

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

$$\sqrt{n}[(\hat{\theta}_{p_1}, \dots, \hat{\theta}_{p_m}) - (\theta_{p_1}, \dots, \theta_{p_m})] \rightarrow_d N_m(0, D),$$

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

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