Lecture 15: Sample quantiles and their asymptotic properties

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 $p$th quantile for any c.d.f. $G$ on $\mathbb{R}$.
Quantiles of $F$ are often the parameters of interest.
$$\theta_p = F^{-1}(p) = p\text{th quantile of } F$$
$$F_n = \text{empirical c.d.f. based on } X_1, ..., X_n$$
$$\hat{\theta}_p = F_n^{-1}(p) = \text{the } p\text{th sample quantile.}$$
$$\hat{\theta}_p = c_{np}X_{(m_p)} + (1 - c_{np})X_{(m_p+1)},$$
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.
\[ F(\theta_p-) = \lim_{x\to\theta_p, x<\theta_p} F(x) \]
\[ F(\theta_p) = \lim_{x\to\theta_p, x>\theta_p} F(x) \]
\[ F(\theta_p-) \leq \rho \leq F(\theta_p) \]

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

**Theorem 5.9**

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

\[ P(|\hat{\theta}_p - \theta_p| > \varepsilon) \leq 2Ce^{-2n\delta^2_\varepsilon}, \]

where \( \delta_\varepsilon \) is the smaller of \( F(\theta_p + \varepsilon) - \rho \) and \( \rho - F(\theta_p - \varepsilon) \) 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).
\[ F(\theta_p-) = \lim_{x \to \theta_p, x < \theta_p} F(x) \]
\[ F(\theta_p) = \lim_{x \to \theta_p, x > \theta_p} F(x) \]
\[ F(\theta_p-) \leq p \leq F(\theta_p) \]

\( F \) 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 \( X_1, \ldots, X_n \) 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 \( n = 1, 2, \ldots, \)

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

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).
\[ F(\theta_p^-) = \lim_{x \to \theta_p, x < \theta_p} F(x) \]
\[ F(\theta_p) = \lim_{x \to \theta_p, x > \theta_p} F(x) \]
\[ F(\theta_p^-) \leq p \leq F(\theta_p) \]

\( F \) 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 \( X_1, \ldots, X_n \) 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 \( n = 1, 2, \ldots, \)

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

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 $\mathbb{R}$,

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

(exercise).

Hence

$$P(\hat{\theta}_p > \theta_p + \varepsilon) = P(p > F_n(\theta_p + \varepsilon))$$

$$= P(F(\theta_p + \varepsilon) - F_n(\theta_p + \varepsilon) > F(\theta_p + \varepsilon) - p)$$

$$\leq P(\rho_\infty(F_n, F) > \delta_\varepsilon)$$

$$\leq Ce^{-2n\delta_\varepsilon^2},$$

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 \( t \in \mathbb{R} \),

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

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.

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

This can be shown by differentiating \( P(F_n(t) \geq p) \) term by term, which leads to
$$\varphi_n(t) = \sum_{i=l_p}^{n} \binom{n}{i} i[F(t)]^{i-1} [1 - F(t)]^{n-i} f(t)$$

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

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

$$+ n \sum_{i=l_p+1}^{n} \binom{n-1}{i-1} [F(t)]^{i-1} [1 - F(t)]^{n-i} f(t)$$

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

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

The following result provides an asymptotic distribution for $\sqrt{n}(\hat{\theta}_p - \theta_p)$. 
Theorem 5.10

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

(i) If $F(\theta_p) = p$, then $P(\sqrt{n}(\hat{\theta}_p - \theta_p) \leq 0) \to \Phi(0) = \frac{1}{2}$, 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

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

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

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

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

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

$$\sqrt{n}(\hat{\theta}_p - \theta_p) \to_d N(0, \sigma^2_F),$$

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

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

Then

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

Under the assumed conditions on \( F, \rho_{nt} \to \rho \) and \( c_{nt} \to t \).

Hence, the result follows from

\[
P(Z_{nt} < -c_{nt}) - \Phi(-c_{nt}) \to 0.
\]

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 \( \sqrt{n}(\hat{\theta}_p - \theta_p) \) has the c.d.f.

\[
\Phi\left(\frac{t}{\sigma_F^-}\right)I_{(-\infty,0)}(t) + \Phi\left(\frac{t}{\sigma_F^+}\right)I_{[0,\infty)}(t),
\]
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 \( X_1, \ldots, X_n \) 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.

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

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 \( X_1, \ldots, X_n \) 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.

\[
\Phi\left(\frac{t}{\sigma^-}\right)I_{(-\infty,0)}(t) + \Phi\left(\frac{t}{\sigma^+}\right)I_{[0,\infty)}(t),
\]

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 \( X_1, \ldots, X_n \) be i.i.d. random variables from \( F \).

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

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

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

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

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

Then, for any \( t \in \mathbb{R} \) and \( \varepsilon > 0 \),

\[
P(\xi_n \leq t, Z_n(0) \geq t + \varepsilon) = P(Z_n(t) \leq U_n(t), Z_n(0) \geq t + \varepsilon) \\
\leq P(|Z_n(t) - Z_n(0)| \geq \varepsilon/2) \\
+ P(|U_n(t) - t| \geq \varepsilon/2) \\
\rightarrow 0
\]
Proof (continued)

Similarly,
\[ P(\xi_n \geq t + \varepsilon, Z_n(0) \leq t) \to 0. \]

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 \( X_1, \ldots, X_n \) be i.i.d. random variables from \( F \) having positive derivatives at \( \theta_{p_j} \), where \( 0 < p_1 < \cdots < p_m < 1 \) are fixed constants. Then
\[ \sqrt{n}[(\hat{\theta}_{p_1}, \ldots, \hat{\theta}_{p_m}) - (\theta_{p_1}, \ldots, \theta_{p_m})] \to_d N_m(0, D), \]
where \( D \) is the \( m \times m \) symmetric matrix whose \((i, j)\)th element is
\[ p_i(1 - p_j)/[F'(\theta_{p_i})F'(\theta_{p_j})], \quad i \leq j. \]
Proof (continued)

Similarly,

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

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

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

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

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