

Lecture 24: Sample median and quantiles

Robustness against outliers

When the population is normal, the sample mean is the MLE and asymptotically optimal.

What happens if the population is not normal?

Suppose that $X_1, \dots, X_n$ are iid with pdf

$$(1 - \varepsilon) \frac{1}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right) + \varepsilon f(x)$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the standard normal pdf and f is another pdf with mean μ and variance τ^2 , $0 \leq \varepsilon \leq 1$.

Then,

$$E(\bar{X}) = \mu \quad \text{and} \quad \text{Var}(\bar{X}) = (1 - \varepsilon) \frac{\sigma^2}{n} + \varepsilon \frac{\tau^2}{n}$$

If τ^2 is very large, then the performance of $\bar{X}$ is not good.

A large variance means that there may be extremely large or small values (outliers).

The sample mean $\bar{X}$ is not robust against outliers: if we let $x_{(n)} \rightarrow \infty$, then $\bar{x} \rightarrow \infty$ and we say that the estimate $\bar{x}$ breaks down.

Definition 10.2.2.

Let $T_n = T_n(X_1, \dots, X_n)$ be an estimator. The breakdown percentage of T_n is the largest percentage b between 0 and 50% such that, if bn of the largest (or smallest) values of $X_1, \dots, X_n$ are driven to ∞ (or $-\infty$), then $|T_n| \rightarrow \infty$.

The sample mean $\bar{X}$ has breakdown percentage 0: if any fraction of X_i 's is driven to ∞ , so is $|\bar{X}|$.

Suppose that n is odd.

The middle order statistic $X_{(\frac{n+1}{2})}$ is called the sample median.

What is the breakdown percentage of the sample median?

If we let all values larger (or smaller) than $X_{(\frac{n+1}{2})}$ go to ∞ , it does not affect the sample median!

When n is large, $\frac{(n-1)/2}{n} \approx \frac{1}{2}$ and the breakdown percentage for sample median is 50%.

However, we still need to check the performance of the sample median, such as consistency, asymptotic normality, and asymptotic efficiency.

For this purpose, we now study the general class of sample quantiles which includes the sample median as a special case.

Population quantiles

For any cdf F , previously we have defined

$$F^{-1}(t) = \inf\{x : F(x) \geq t\}$$

Let $p \in (0, 1)$ be a fixed constant.

We call $Q_p = F^{-1}(p)$ to be the p th quantile of F , which is also called the p th population quantile if F is the cdf of a random sample.

If $F = F_\theta$ depends on a parameter, then Q_p may be estimated by $F_{\hat{\theta}_n}^{-1}(p)$ if θ is estimated by $\hat{\theta}_n$.

Otherwise, the most popular estimator of Q_p based on iid $X_1, \dots, X_n$ is the p th sample quantile defined as $\hat{Q}_p = \hat{F}_n^{-1}(p)$, where is the $\hat{F}_n(x)$ empirical cdf defined as $\hat{F}_n(x) = n^{-1} \sum_{i=1}^n I(X_i \leq x)$.

However, we still need to check the performance of the sample median, such as consistency, asymptotic normality, and asymptotic efficiency.

For this purpose, we now study the general class of sample quantiles which includes the sample median as a special case.

Population quantiles

For any cdf F , previously we have defined

$$F^{-1}(t) = \inf\{x : F(x) \geq t\}$$

Let $p \in (0, 1)$ be a fixed constant.

We call $Q_p = F^{-1}(p)$ to be the p th quantile of F , which is also called the p th population quantile if F is the cdf of a random sample.

If $F = F_\theta$ depends on a parameter, then Q_p may be estimated by $F_{\hat{\theta}_n}^{-1}(p)$ if θ is estimated by $\hat{\theta}_n$.

Otherwise, the most popular estimator of Q_p based on iid $X_1, \dots, X_n$ is the p th sample quantile defined as $\hat{Q}_p = \hat{F}_n^{-1}(p)$, where is the $\hat{F}_n(x)$ empirical cdf defined as $\hat{F}_n(x) = n^{-1} \sum_{i=1}^n I(X_i \leq x)$.

Define $l_p = np$ if np is an integer and $l_p = 1 +$ the integer part of np if np is not an integer.

Then

$$\hat{Q}_p = X_{(l_p)} \quad (\text{the } l_p\text{th order statistic})$$

For the median, $p = 0.5$ and the sample median is

$$\hat{Q}_{0.5} = \begin{cases} X_{(n/2)} & \text{if } n \text{ is even} \\ X_{((n+1)/2)} & \text{if } n \text{ is odd} \end{cases}$$

This is the same as the definition in many textbooks in which the sample median is define as

$$\tilde{Q}_{0.5} = \begin{cases} (X_{(n/2)} + X_{(n/2+1)})/2 & \text{if } n \text{ is even} \\ X_{((n+1)/2)} & \text{if } n \text{ is odd} \end{cases}$$

When n is large, $\hat{Q}_{0.5}$ and $\tilde{Q}_{0.5}$ are not very different.

We take $\hat{Q}_{0.5}$ since it is equal to $\hat{F}_n^{-1}(0.5)$ and mathematically more convenient to handle, but $\tilde{Q}_{0.5}$ is more suitable for practical uses when n is not very large.

Similar discussions can be made for the sample quantiles.

The distribution of a sample quantile

The exact distribution of $\hat{Q}_p$ can be obtained since it is simply the l_p th order statistic (see Chapter 5), but we can derive it again as follows.

Note that $n\hat{F}_n(t) = \sum_{i=1}^n I(X_i \leq t)$ has the *binomial*($n, F(t)$) distribution for any $t \in \mathcal{R}$:

$$\begin{aligned} P(\hat{Q}_p \leq t) &= P(\hat{F}_n(t) \geq p) && \hat{F}_n(t) \geq p \text{ iff } t \geq \hat{F}_n^{-1}(p) \\ &= P(\text{a binomial}(n, F(t)) \text{ variable} \geq np) \\ &= \sum_{i=l_p}^n \binom{n}{i} [F(t)]^i [1 - F(t)]^{n-i} \end{aligned}$$

If F has a pdf f , then $\hat{Q}_p$ has a pdf that can be derived by differentiating $P(\hat{Q}_p \leq t)$ term by term (see Chapter 5), i.e.,

$$\frac{d}{dt} P(\hat{Q}_p \leq t) = n \binom{n-1}{l_p-1} [F(t)]^{l_p-1} [1 - F(t)]^{n-l_p} f(t).$$

Asymptotic normality of sample quantiles

For large n , it is more convenient to use the following asymptotic approximation.

The asymptotic normality of sample quantiles is also important in the consideration of asymptotic efficiency and comparing estimators.

Theorem (an extension of Example 10.2.3).

Let $X_1, \dots, X_n$ be iid random variables from F . If F is differentiable at Q_p and $f(Q_p) = F'(Q_p) > 0$, then

$$\sqrt{n}(\hat{Q}_p - Q_p) \text{ converges in distribution to } N\left(0, \frac{p(1-p)}{[f(Q_p)]^2}\right)$$

Proof.

Let t be a fixed constant, $p_n = F(Q_p + tn^{-1/2})$, and

$$Y_i = I(X_i \leq Q_p + tn^{-1/2})$$

Then $Y_1, \dots, Y_n$ are iid Bernoulli random variables with $P(Y_i = 1) = p_n$.

Hence

$$\begin{aligned}P\left(\hat{Q}_p \leq Q_p + tn^{-1/2}\right) &= P\left(p \leq F_n(Q_p + tn^{-1/2})\right) \\&= P\left(\sum_{i=1}^n Y_i \geq np\right) \\&= P\left(\frac{\sum_{i=1}^n Y_i - np_n}{\sqrt{np_n(1-p_n)}} \geq \frac{np - np_n}{\sqrt{np_n(1-p_n)}}\right)\end{aligned}$$

Under the assumed conditions on F , as $n \rightarrow \infty$,

$$p_n = F(Q_p + tn^{-1/2}) \rightarrow F(Q_p) = p$$

From the definition of the derivative, as $n \rightarrow \infty$,

$$\frac{np_n - np}{\sqrt{n}} = \frac{t(p_n - p)}{t/\sqrt{n}} = \frac{t(F(Q_p + tn^{-1/2}) - p)}{tn^{-1/2}} \rightarrow tf(Q_p)$$

Hence,

$$\lim_{n \rightarrow \infty} \frac{np - np_n}{\sqrt{np_n(1-p_n)}} = \frac{-tf(Q_p)}{\sqrt{p(1-p)}}$$

By the CLT,

$$\frac{\sum_{i=1}^n Y_i - np_n}{\sqrt{np_n(1-p_n)}} \text{ converges in distribution to } Z \sim N(0, 1)$$

Hence,

$$\lim_{n \rightarrow \infty} P\left(\hat{Q}_p \leq Q_p + tn^{-1/2}\right) = P\left(Z \geq \frac{-t f(Q_p)}{\sqrt{p(1-p)}}\right)$$

i.e.,

$$\lim_{n \rightarrow \infty} P\left(\sqrt{n}(\hat{Q}_p - Q_p) \leq t\right) = P\left(Z \leq \frac{t f(Q_p)}{\sqrt{p(1-p)}}\right)$$

The result follows since $\sqrt{p(1-p)}Z/f(Q_p) \sim N(0, p(1-p)/[f(Q_p)]^2)$.

In the special case of $p = 0.5$, the sample median as an estimator of the population median is asymptotically normal:

$$\sqrt{n}(\hat{Q}_{0.5} - Q_{0.5}) \text{ converges in distribution to } N\left(0, \frac{1}{4[f(Q_{1/2})]^2}\right)$$

By the CLT,

$$\frac{\sum_{i=1}^n Y_i - np_n}{\sqrt{np_n(1-p_n)}} \text{ converges in distribution to } Z \sim N(0, 1)$$

Hence,

$$\lim_{n \rightarrow \infty} P\left(\hat{Q}_p \leq Q_p + tn^{-1/2}\right) = P\left(Z \geq \frac{-t f(Q_p)}{\sqrt{p(1-p)}}\right)$$

i.e.,

$$\lim_{n \rightarrow \infty} P\left(\sqrt{n}(\hat{Q}_p - Q_p) \leq t\right) = P\left(Z \leq \frac{t f(Q_p)}{\sqrt{p(1-p)}}\right)$$

The result follows since $\sqrt{p(1-p)}Z/f(Q_p) \sim N(0, p(1-p)/[f(Q_p)]^2)$.

In the special case of $p = 0.5$, the sample median as an estimator of the population median is asymptotically normal:

$$\sqrt{n}(\hat{Q}_{0.5} - Q_{0.5}) \text{ converges in distribution to } N\left(0, \frac{1}{4[f(Q_{1/2})]^2}\right)$$

Example 10.2.4.

If F is symmetric about $Q_{0.5}$ and the mean μ exists, then $\mu = Q_{0.5}$; the sample mean and the sample median estimate the same quantity and we compare their asymptotic relative efficiency.

If $\text{Var}(X_1) = \infty$ (e.g., F is Cauchy), then the sample median is clearly better.

Assuming that $\text{Var}(X_i) = \sigma^2$ is finite, we obtain that

$$\text{ARE}(\hat{Q}_p, \bar{X}) = 4[f(\mu)]^2 \sigma^2$$

Thus, the sample mean is asymptotically more efficient than the sample median iff $4[f(\mu)]^2 \sigma^2 < 1$.

If F is in a location-scale family, then this ARE is a constant.

The following table gives some examples.

	Normal	Logistic	t_5	t_4	t_3	Double exponential
ARE	0.637	0.822	0.96	1.12	1.62	2

t_v : t-distribution with v degrees of freedom

Example.

We compare the sample median with the parametric MLE under a non-symmetric distribution.

Let $X_1, \dots, X_n$ be iid from $\text{exponential}(0, \theta)$ with unknown $\theta > 0$.

Since the cdf is $F_\theta(t) = 1 - e^{-t/\theta}$ for $t > 0$ and 0 for $t \leq 0$, we solve

$$0.5 = F_\theta(Q_{0.5}) = 1 - e^{-Q_{0.5}/\theta} \quad \text{iff} \quad \log 2 = Q_{0.5}/\theta$$

i.e., $Q_{0.5} = \theta \log 2$.

Since the MLE of θ is $\bar{X}$, the MLE of $Q_{0.5}$ is $\bar{X} \log 2$.

By the CLT and Slutsky's theorem,

$$\sqrt{n}(\bar{X} \log 2 - Q_{0.5}) \text{ converges in distribution to } N(0, (\log 2)^2 \theta^2)$$

On the other hand, $4[F'_\theta(Q_{0.5})]^2 = 4[\theta^{-1} e^{-\log 2}]^2 = \theta^{-2}$ and hence

$$\sqrt{n}(\hat{Q}_{0.5} - Q_{0.5}) \text{ converges in distribution to } N(0, \theta^2)$$

Hence,

$$\text{ARE}(\hat{Q}_{0.5}, \bar{X} \log 2) = (\log 2)^2 = 0.480$$

Thus, the MLE under the parametric assumption is much better.