Ann. Inst. Statist. Math.
38 (1986), Part A, 439-443

A NOTE ON BOOTSTRAPPING THE VARIANCE OF SAMPLE QUANTILE

GUTTI JOGESH BABU

(Received Apr. 9, 1985; July 18, 1985)

Summary

Let $m_{n,p}$ denote the p-th quantile based on n observations and let λ_p denote the population quantile. In this paper consistency of the bootstrap estimate of variance of $\sqrt{n}(m_{n,p}-\lambda_p)$ is established.

1. Introduction

Let $\{X_n\}$ be i.i.d. random variables with distribution function F having unique median μ. Suppose F has a continuous derivative f at μ and $f(\mu)>0$. It is well known that under suitable conditions $M_n=\sqrt{n}(m_n-\mu)$ is asymptotically normal with mean zero and variance $(4f^2(\mu))^{-1}$, where m_n is a sample median. As $f(\mu)$ is not known in general, this result is of not much use in estimating μ. So one should look for a suitable estimate of the variance of M_n. Recently, Efron [7] introduced a very general resampling procedure called the bootstrap. Babu [1], Babu and Singh [3]-[5], Singh [10] and Bickel and Freedman [6] studied the asymptotic properties of this method. In this paper we use the bootstrap method to estimate the variance of M_n. In this connection, the theorem gives much more than we require.

2. Bootstrap estimation of variance

To describe bootstrap, let $\{X_1, X_2, \cdots, X_n\}$ be i.i.d. random variables with distribution function F and let $T(X_1, \cdots, X_n; F)$ be the specified statistic of interest, possibly depending on the unknown distribution F. Let F_n denote the empirical distribution function of $X_1, \cdots, X_n$. The method consists of approximating the distribution of $T(X_1, \cdots, X_n; F)$ under F by that of $T(Y_1, \cdots, Y_n; F_n)$ under F_n, where $Y_1, \cdots, Y_n$ is a random sample from F_n.

Key words and phrases: Bootstrap, quantile, Borel-Cantelli lemma, Bahadur-Kiefer representation of quantiles.

Let G_n denote the empirical distribution function of $Y_1, \cdots, Y_n$. For any distribution function G, $0<u<1$, let $G^{-1}(u)=\inf\{x: G(x)\geqq u\}$. Note that with this notation, $F_n^{-1}(u)$ is a u-th sample quantile. Let $0<p<1$ and let F be continuously differentiable in a neighbourhood of $F^{-1}(p)$. Further we assume that $f(F^{-1}(p))>0$, where $f=F'$. From the proof of Theorem 2 of Singh [10] (see also proposition 5.1 of Bickel and Freedman [6]), it follows that a.e., $B_{n,p}=\sqrt{n}(G_n^{-1}(p)-F_n^{-1}(p))$ is asymptotically normal with mean zero and variance $\sigma^2=pq(f(F^{-1}(p)))^{-2}$, where $q=1-p$. So if for almost all samples, $\{B_{n,p}^2\}$ are uniformly integrable, then the bootstrap variance of $B_{n,p}$ converges a.e. to σ^2. In particular, this holds if for some $\delta>2$, the δ-th moment of $|B_{n,p}|$ are uniformly bounded a.e. The following theorem is useful in this connection.

THEOREM. *Let E_* denote the expectation under the bootstrap distribution. We have*

(1) *If $E(\log(1+|X_1|))<\infty$, then for all $k>0$, $E_*|B_{n,p}|^k \ll 1$ a.e.*
(2) *If $E\{[\log(1+|X_1|)]/\log\log(3+|X_1|)\}=\infty$, then for every $k>0$ $\limsup_{n\to\infty} E_*|B_{n,p}|^k=\infty$ a.e.*
(3) *If $E(X_1^2)<\infty$, then for every real t, $E_*(e^{t|B_{n,p}|})\ll 1$ a.e.*
(4) *If $E(X_1/\log(1+|X_1|))^2=\infty$, then for all real t, $\limsup_{n\to\infty} E_*(e^{t|B_{n,p}|})=\infty$ a.e.*

Often the notation "$\ll$" is used instead of the standard $O(\cdot)$. Ghosh, Parr, Singh and Babu [8] proved that if $E|X_1|^\alpha<\infty$ for some $\alpha>0$, then for some $\delta>2$, $E_*|B_{n,0.5}|^\delta \ll 1$ a.e.

To prove the Theorem, we need the following lemmas.

LEMMA 1. *Let $\{Z_i\}$ be i.i.d. random variables with $Z_i\geqq 0$ and $E(Z_1)<\infty$. Then there exists a sequence $0\leqq a_n\to 0$ such that $\max_{i\leqq n} Z_i\leqq na_n$ for all large n, a.e.*

PROOF. Let $10<s_1<s_2<\cdots$ be continuity points of the distribution G of Z_i such that $2^i<s_i$ and $\int_{s_i}^\infty x\,dG(x)<E(Z_1)4^{-i}$. Define $h(x)=1$ for $0<x\leqq s_1$ and $h(x)=2^j$ for $s_j<x\leqq s_{j+1}$, $j=1, 2, \cdots$. Clearly $h(x)\uparrow\infty$, $h(x)\leqq x$ for $x\geqq 1$ and $E(Z_1h(Z_1))<\infty$. So

$$\begin{aligned}[Z_1>n/h(\sqrt{n})]&=[(Z_1>n/h(\sqrt{n}))\cap(Z_1>\sqrt{n})]\\&\subset[(Z_1>n/h(\sqrt{n}))\cap(h(Z_1)>h(\sqrt{n}))]\\&\subset(Z_1h(Z_1)\geqq n)\,.\end{aligned}$$

As a consequence,

$$\sum_{n=s_1}^\infty P(Z_n\geqq n/h(\sqrt{n}))\leqq\sum_1^\infty P(Z_1h(Z_1)\geqq n)\leqq 1+E(Z_1h(Z_1))<\infty\,.$$

Hence by Borel-Cantelli lemma a.e., $Z_n \leqq n/h(\sqrt{n})$ for all large n. If we put $a_n=1$ for $1 \leqq n \leqq s_1$ and

$$a_n = \max\{(2/h(\sqrt{i})):\ \log n \leqq i \leqq n\}$$

for $n > s_1$, then for $n \geqq 10$, $a_n \geqq 2/\sqrt{\log n}$ and

$$\max_{i \leq n}(i/h(\sqrt{i})) \leqq (n/\sqrt{\log n}) + \max\{(i/h(\sqrt{i})):\ \sqrt{n} \leqq i \leqq n\}$$
$$\leqq (n/\sqrt{\log n}) + \frac{1}{2} n a_n \leqq n a_n .$$

So $\max_{i \leq n} Z_i \leqq n a_n$ for all large n, a.e.

LEMMA 2. *Let* $0<\delta<1$. *Let* $\{Z_{i,\delta}\}$ *be i.i.d. random variables with* $P(Z_{1,\delta}=1)=\delta=1-P(Z_{1,\delta}=0)$. *Let* $S_{n,\delta}=\sum_{i=1}^{n} Z_{i,\delta}$. *Then*

$$P(S_{n,\delta} \geqq 4 \max\{\sqrt{n}, n\delta\}) \ll n^{-3} .$$

PROOF. For any $a, b>0$, we have by Markov's inequality that

$$P(S_{n,\delta} \geqq 4b) \leqq e^{-4ab}(\delta e^a + (1-\delta))^n \leqq \exp(-4ab + n\delta(e^a - 1)) .$$

The result follows now by putting $b=\max\{\sqrt{n}, n\delta\}$ and $a=(\log n)/b$ in the above equation.

PROOF OF THE THEOREM. Let $T_1=F_n^{-1}(0)=\min_{i \leq n} X_i$ and $T_2=F_n^{-1}(1)=\max_{i \leq n} X_i$. First, we prove (2). By using Borel-Cantelli lemma, we get for any $\delta>0$ that $|X_n| \geqq \exp(\delta n \log n)$ infinitely often a.e. Consequentlly, $T=\max\{|T_1|, |T_2|\}=\max_{i \leq n}|X_i| \geqq n^{\delta n}$ infinitely often a.e. So for any $k>0$, by taking $\delta=2/k$, we obtain

$$T^k P_*(G_n^{-1}(p) \in \{T_1, T_2\}) \geqq n^{\delta k n} n^{-n} \geqq n^n$$

infinitely often a.e. This proves (2). A similar proof gives (4).

To prove (1) and (3), let $\{U_i\}$ be i.i.d. $U[0, 1]$ random variables. Let V_n denote the empirical distribution function of $\{U_1, \cdots, U_n\}$. Without loss of generality we can take $X_i=F^{-1}(U_i)$. As f is continuous and positive in a neighbourhood of $F^{-1}(p)$ and since

$$\sup_{0<t<1} \sqrt{n}\,|V_n^{-1}(t)-t|(\log\log n)^{-1/2} \ll 1 \text{ a.e.}, \tag{5}$$

there exists a $d \in (0, \min\{p, q\})$ such that, a.e. we have

$$\begin{aligned} |F_n^{-1}(t+p)-F_n^{-1}(p)| &= |F^{-1}(V_n^{-1}(t+p))-F^{-1}(V_n^{-1}(p))| \\ &\ll |V_n^{-1}(t+p)-V_n^{-1}(p)| \end{aligned} \tag{6}$$

uniformly for $|t| \leqq d$. By Bahadur-Kiefer representation of quantiles (see Kiefer [9]), we have uniformly for $|t|<d$,

$$\text{(7)} \qquad \text{the l.h.s. of (6)} \ll |V_n(t+p)-V_n(p)-2t|+n^{-3/4}\log n \ll \max\{|t|, n^{-1/2}\} \text{ a.e.}$$

The last inequality in (7) follows from Lemma 2 and the Borel-Cantelli lemma.

Let $r=np+1$ or $[np]$ depending on whether np is an integer or not, where $[x]$ denotes the integral part of x. Since for any $v\in(0,1)$,

$$\sum_{j=r}^{n}\binom{n}{j}v^j(1-v)^{n-j}=n\binom{n-1}{r-1}\int_0^v u^{r-1}(1-u)^{n-r}du\ ,$$

we have for $1\leqq i\leqq n$,

$$P_*(G_n^{-1}(p)=F_n^{-1}(i/n))=n\binom{n-1}{r-1}\int_{(i-1)/n}^{i/n} u^{r-1}(1-u)^{n-r}du\ .$$

By Stirling's formula $n\binom{n-1}{r-1}\ll\sqrt{n}\,p^{1-r}q^{r-n}$. So for any $D>1$, $n^{-1/2}\leqq \varepsilon_n\to 0$, by (7) a.e. there exists an $A>1$ such that

$$\begin{aligned}
\text{(8)} \quad &\sum_{|p-i/n|\leqq\varepsilon_n} P_*(G_n^{-1}(p)=F_n^{-1}(i/n))\exp\left(D\sqrt{n}\left|F_n^{-1}\left(\frac{i}{n}\right)-F_n^{-1}(p)\right|\right)\\
&\ll 1+\sqrt{n}\,p^{1-r}q^{r-n}\sum_{|p-i/n|\leqq\varepsilon_n}\exp\left(A\sqrt{n}\,|(i/n)-p|\right)\\
&\qquad\cdot\int_{(i-1)/n}^{i/n} u^{r-1}(1-u)^{n-r}du\\
&\ll 1+\sqrt{n}\int_{-2\varepsilon_n}^{2\varepsilon_n}\exp(A\sqrt{n}\,|v|)\left(1+\frac{v}{p}\right)^{r-1}\left(1-\frac{v}{q}\right)^{n-r}dv\\
&\ll 1+\sqrt{n}\int_0^{2\varepsilon_n}\left[\exp\left(A\sqrt{n}\,v-\frac{nv^2}{2pq}+O(nv^2\varepsilon_n)\right)\right]dv\\
&\ll 1+\int_0^{2\sqrt{n}\varepsilon_n}\left[\exp\left(Av-\frac{v^2}{2pq}+O(v^2\varepsilon_n)\right)\right]dv\ll 1\ .
\end{aligned}$$

Note that the function $u^{r-1}(1-u)^{n-r}$, $0<u<1$ is increasing in $(0,(r-1)/(n-1))$ and decreasing in $((r-1)/(n-1),1)$. Since $1+x\leqq\exp(x-x^2/4)$ for $|x|\leqq 1/2$, and since $pq\leqq 1/4$, we have, a.e.

$$\begin{aligned}
\text{(9)} \quad &\sum_{|p-i/n|>\varepsilon_n} P_*(G_n^{-1}(p)=F_n^{-1}(i/n))\\
&\leqq n\binom{n-1}{r-1}\int_{(|u-p|\geqq\varepsilon_n-1/n)} u^{r-1}(1-u)^{n-r}I_{(0,1)}(u)du\\
&\ll\sqrt{n}\,[(1+(\varepsilon_n-n^{-1})/p)^{r-1}(1-(\varepsilon_n-n^{-1})/q)^{n-r}\\
&\qquad+(1+(\varepsilon_n-n^{-1})/q)^{n-r}(1-(\varepsilon_n-n^{-1})/p)^{r-1}]\\
&\ll\sqrt{n}\,\exp(-n(4pq)^{-1}\varepsilon_n^2)\ll\sqrt{n}\,\exp(-n\varepsilon_n^2)\ .
\end{aligned}$$

If $E(\log(1+|X_1|))<\infty$, then by Lemma 1, there exists a sequence $0\leqq a_n\to 0$ such that $T\ll e^{na_n}$ for all large n a.e. So if we take $\varepsilon_n^2=\max\{2a_n, n^{-1/4}\}$ in (8) and (9) we get

$$\begin{aligned} E_*|B_n, p|^k &\ll 1+T^k n^{(k+1)/2}\exp(-n\varepsilon_n^2) \\ &\ll 1+n^{(k+1)/2}(\exp(na_n-n\varepsilon_n^2))\ll 1 \text{ a.e.} \end{aligned}$$

This proves (1). If $E(X_1^2)<\infty$, then by Lemma 1, there exists a sequence $0\leqq a_n\to 0$ such that $T\leqq\sqrt{na_n}$ for all large n a.e. So for any $t>0$, if we take $\varepsilon_n^2=\max\{2t\sqrt{a_n}, n^{-1/4}\}$ in (8) and (9) we get (3).

This completes the proof of the Theorem.

Remark. The condition that F is differentiable in a neighbourhood of $F^{-1}(p)$ can be relaxed. The only place in the proof where it is used is in (6). Without this assumption a weaker version of the theorem can be obtained using Theorem 5 of Babu and Singh [2].

INDIAN STATISTICAL INSTITUTE

REFERENCES

[1] Babu, G. J. (1984). Bootstrapping statistics with linear combinations of chi-squares as weak limit, *Sankhya*, Series A, **46**, 85-93.

[2] Babu, G. J. and Singh, K. (1978). On deviations between empirical and quantile processes for mixing random variables, *J. Multivar. Anal.*, 8, 532-549.

[3] Babu, G. J. and Singh, K. (1983). Inference on means using bootstrap, *Ann. Statist.*, **11**, 999-1003.

[4] Babu, G. J. and Singh, K. (1984). Asymptotic representations related to jackknifing and bootstrapping L-statistics, *Sankhya*, Series A, **46**, 195-206.

[5] Babu, G. J. and Singh, K. (1984). On one term Edgeworth correction by Efron's bootstrap, *Sankhya*, Series A, **46**, 219-232.

[6] Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap, *Ann. Statist.*, **9**, 1196-1217.

[7] Efron, B. (1979). Bootstrap methods: Another look at the Jackknife, *Ann. Statist.*, **7**, 1-26.

[8] Ghosh, M., Parr, W. C., Singh, K. and Babu, G. J. (1984). A note on bootstrapping the sample median, *Ann. Statist.*, **12**, 1130-1135.

[9] Kiefer, J. (1967). On Bahadur's representation of sample quantiles, *Ann. Math. Statist.*, **38**, 1323-1342.

[10] Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap, *Ann. Statist.*, **9**, 1187-1195.