



# A note on the bootstrapped empirical process<sup>☆</sup>

Gutti Jogesh Babu\*

*Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA*

Received 3 June 2003; accepted 3 September 2003

---

## Abstract

Let  $F_n$  denote the empirical distribution of a sample of size  $n$  from a distribution function  $F$ , and let  $H_{n,t}$  denote the distribution of  $\sqrt{n}(F_n(t) - F(t))$ . Contrary to the intuition,  $H_{n,t}$  is not approximated uniformly over  $t$  by its bootstrapped counterpart. The main problem is at the values of  $t$  near the tails of  $F$ . Two results exploring this phenomenon are presented here.

© 2003 Elsevier B.V. All rights reserved.

*MSC:* primary 60F05; secondary 62F40

*Keywords:* Bootstrap; Empirical distribution; Law of the iterated logarithm

---

Let  $X_1, \dots, X_n$  be i.i.d. random variables from a distribution  $F$ . Let  $F_n$  and  $F_n^*$  denote the empirical distribution and its bootstrap version. Further, let  $P^*$  denote the measure induced by the bootstrap sampling, that is, the conditional measure given the sample  $X_1, \dots, X_n$ . Note that  $F_n^*$  is the empirical distribution of a sample of size  $n$  from  $F_n$ . Define

$$A_n(x) = \sup_u |P(\sqrt{n}(F_n(x) - F(x)) \leq u) - P^*(\sqrt{n}(F_n^*(x) - F_n(x)) \leq u)|,$$

$$A_n = \sup_x A_n(x)$$

$$t_n = \inf \left\{ x : F(x) \geq \frac{1}{n+1} \right\}$$

$$D_n = \{F_n(t_n) = 0\} = \{X_1 > t_n, \dots, X_n > t_n\}.$$

---

<sup>☆</sup> Supported in part by National Science Foundation Grant DMS-0101360.

\* Tel.: +1-814-8651348; fax: +1-814-8637114.

*E-mail address:* [babu@stat.psu.edu](mailto:babu@stat.psu.edu) (G.J. Babu).

In this paper it is shown that  $A_n$  is bounded away from zero with positive probability, for all  $n \geq 1$ .

It is clear from the central limit theorem and the strong law of large numbers that  $A_n(x) \rightarrow 0$  a.e. for all  $x$ . The problem of asymptotic behavior of  $A_n$  arose in connection with the study of ‘False Discovery Rates’ by [Genovese and Wasserman \(2001\)](#). In view of the tightness of the empirical process and its bootstrap version, it may not be unrealistic to expect  $A_n \rightarrow_p 0$  or  $A_n \rightarrow 0$  a.e., as  $n \rightarrow \infty$ . However, an attempt to prove this quickly leads to doubt its validity. In this paper, it is shown that  $A_n \not\rightarrow_p 0$  for *any* continuous  $F$ . But if the range of  $x$  near the tails is restricted, then  $A_n(x) \rightarrow 0$  uniformly in certain intervals a.e. These results are established in the next two theorems.

**Theorem 1.** *For any continuous  $F$ ,  $P(A_n \geq 1 - 2e^{-1}) \geq \frac{1}{4}$  for all  $n \geq 1$ .*

**Proof.** As  $(1 - 1/(n + 1))^{n+1}$  is increasing in  $n$ , it follows that for all  $n \geq 1$ ,

$$\frac{1}{4} \leq \left(1 - \frac{1}{n+1}\right)^{n+1} \leq P(D_n) = \left(1 - \frac{1}{n+1}\right)^n \leq e^{-1} \left(1 + \frac{1}{n}\right) \leq 2e^{-1}. \tag{1}$$

Clearly,

$$P(\sqrt{n}(F_n(t_n) - F(t_n)) \leq 0) = P\left(F_n(t_n) \leq \frac{1}{n+1}\right) = P(F_n(t_n) = 0) = P(D_n),$$

and on  $D_n$ ,  $F_n^*(t_n) = F_n(t_n) = 0$ . So  $A_n \geq 1 - P(D_n)$  on  $D_n$ , and hence

$$P(A_n \geq 1 - P(D_n)) \geq P(D_n)$$

for all  $n \geq 1$ . In view of (1), this leads to  $P(A_n \geq 1 - 2e^{-1}) \geq \frac{1}{4}$  for all  $n \geq 1$ .  $\square$

To prove the next theorem, we need a special case of Berry–Esseen theorem (see for example, Theorem 12.4 and Eq. (12.16) of [Bhattacharya and Rao, 1986](#)), which is stated as the following Lemma.

**Lemma 1.** *Let  $Y_1, \dots, Y_n$  be i.i.d. Bernoulli random variables with  $P(Y_i = 1) = p = 1 - P(Y_i = 0)$ , where  $0 < p < 1$ . Then*

$$\left| P\left(\sum_{i=1}^n (Y_i - p) \leq x\sqrt{np(1-p)}\right) - \Phi(x) \right| \leq \frac{2}{\sqrt{np(1-p)}}.$$

Here  $\Phi$  and  $\phi$  denote the cumulative distribution function and the density function of the standard normal variable.

**Theorem 2.** *For any distribution function  $F$  and any sequence  $a_n \rightarrow \infty$ , we have*

$$B_n = \sup\{A_n(x) : F(x)(1 - F(x)) \geq a_n(\log \log n)^{1/2} n^{-1/2}\} \rightarrow 0 \quad \text{a.e.}$$

**Proof.** Let  $\|F_n - F\| = \sup_x |F_n(x) - F(x)|$ . If

$$\sigma^2(x, n) = F_n(x)(1 - F_n(x)) \quad \text{and} \quad \sigma^2(x) = F(x)(1 - F(x)),$$

then

$$|\sigma^2(x, n) - \sigma^2(x)| \leq \|F_n - F\|. \quad (2)$$

By Lemma 1, we have

$$\begin{aligned} & |P(\sqrt{n}(F_n(x) - F(x)) \leq u) - P^*(\sqrt{n}(F_n^*(x) - F_n(x)) \leq u)| \\ & \leq |\Phi(x/\sigma(x)) - \Phi(x/\sigma(x, n))| + \frac{2}{\sqrt{n}} \left( \frac{1}{\sigma(x)} + \frac{1}{\sigma(x, n)} \right). \end{aligned} \quad (3)$$

Since  $\sup_x |x\phi(x)| < 1$ , it follows by (2), that

$$\begin{aligned} |\Phi(x/\sigma(x)) - \Phi(x/\sigma(x, n))| & \leq \left| \frac{1}{\sigma(x)} - \frac{1}{\sigma(x, n)} \right| \max(\sigma(x), \sigma(x, n)) \\ & \leq |\sigma^2(x) - \sigma^2(x, n)| (\sigma(x)\sigma(x, n))^{-1} \\ & \leq \|F_n - F\| (\min(\sigma(x), \sigma(x, n)))^{-2}. \end{aligned} \quad (4)$$

By the well-known law of the iterated logarithm for empirical distribution functions (see for example, Theorem 5.1.1 of Csörgő and Révész (1981)), we have

$$\limsup_{n \rightarrow \infty} \sqrt{2n/\log \log n} \|F_n - F\| \leq 1 \quad \text{a.e.} \quad (5)$$

The theorem now follows from (2)–(5).  $\square$

**Remark.** The random variable  $nF_n(1/n)$  converges to the Poisson distribution with mean 1, while the bootstrap distribution of  $nF_n^*(1/n)$  given  $(nF_n(1/n) = k)$  converges weakly to the Poisson distribution with mean  $k$ . So the bootstrap distribution of  $nF_n^*(1/n)$  converges weakly to a random measure. Consequently,

$$P(nF_n(1/n) = j) - P^*(nF_n^*(1/n) = j) \rightarrow 0.$$

## Acknowledgements

I want to thank Larry Wasserman who posed the problem.

## References

- Bhattacharya, R.N., Ranga Rao, R., 1986. Normal Approximations and Asymptotic Expansions. Krieger, Malabar, FL.
- Csörgő, M., Révész, P., 1981. Strong Approximations in Probability and Statistics. Academic Press, New York.
- Genovese, C., Wasserman, L., 2001. False Discovery Rates. Preprint.