*J. Austral. Math. Soc.* (*Series A*) **58** (1995), 47–53

# THE LIMITING DISTRIBUTION OF A RECURSIVE RESAMPLING PROCEDURE

**ZHENG ZUKANG and WU LIPENG**

(Received 19 October 1990; revised 18 March 1992)

Communicated by A. G. Pakes

**Abstract**

A recursive resampling method is discussed in this paper. Let $X_1, X_2, \cdots, X_n$ be i.i.d. random variables with distribution function $F$ and construct the empirical distribution function $F_n$. A new sample $X_{n+1}$ is drawn from $F_n$ and the new empirical distribution function $\tilde{F}_{n+1}$ in the wide sense, is computed from $X_1, X_2, \cdots, X_n, X_{n+1}$. Then $X_{n+2}$ is drawn from $\tilde{F}_{n+1}$ and $\tilde{F}_{n+2}$ is obtained. In this way, $X_{n+m}$ and $\tilde{F}_{n+m}$ are found. It will be proved that $\tilde{F}_{n+m}$ converges to a random variable almost surely as $m$ goes to infinity and the limiting distribution is a compound beta distribution. In comparison with the usual non-recursive bootstrap, the main advantage of this procedure is a reduction in unconditional variance.

1991 *Mathematics subject classification* (*Amer. Math. Soc.*): 60G50, 60F15.

## 1. Introduction

In recent years the jackknife, the bootstrap and other resampling methods have been discussed by many authors. Efron [2] gave a review. The goal of his study is to understand a collection of ideas concerning the non-parametric estimators of bias, variance and more general measures of error. Perhaps the bootstrap method is more interesting. Let $X_1, X_2, \cdots, X_n$ be i.i.d. random variables with distribution function $F$. Let $F_n$ be the empirical distribution function of the data, putting probability mass $1/n$ on each point, that is, we observe $\{X_j,\ j = 1, \cdots, n\}$ and construct

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \le x\}}. \tag{1.1}$$

Now let $X_1^*, X_2^*, \cdots, X_m^*$ be an i.i.d. sample from $F_n$. This means that $\{X_i^*\}$ is a random sample drawn with replacement from the observed $X_1, X_2, \cdots, X_n$. A spe-





https://doi.org/10.1017/S1446788700038088 Published online by Cambridge University Press

cified random variable $R(X, F)$ is given, possibly depending on both the unknown distribution function $F$ and $X = (X_1, X_2, \cdots, X_n)$. We use $R^* = R(X^*, F_n) = R((X_1^*, \cdots, X_n^*), F_n)$ to approximate $R(X, F)$ and call it the bootstrap estimator. Some asymptotic theorems for the bootstrap are investigated for interesting cases, and some counter-examples show that the approximations do not always succeed.

Instead of the above replacement, that is, equal mass on each observed value, Rubin [4] first suggested the Bayesian Bootstrap method. Some authors developed this to the random weighting method, in which case the Dirichlet distribution is introduced as a prior distribution.

In the other direction, we will suggest a recursive method of resampling. Suppose that $X_1, X_2, \cdots, X_n$ are i.i.d. random variables with distribution function $F$. We use the original data to get the empirical distribution function $F_n$ of (1.1). Now we draw one sample from the observed data with equal probability to obtain $X_{n+1}$. Then we have $X_1, X_2, \ldots, X_{n+1}$ and draw another sample from these $n + 1$ data with equal probability, to get $X_{n+2}$ with distribution $\tilde{F}_{n+1}$ the empirical distribution function of $X_1, X_2, \cdots, X_n, X_{n+1}$ in the wise sense, that is,

$$\tilde{F}_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} I_{\{X_i \le x\}}.$$

In this way we get $X_{n+1}, X_{n+2}, \cdots, X_{n+m}$ and $\tilde{F}_n + 1, \tilde{F}_{n+2}, \cdots, \tilde{F}_{n+m}$. We shall see that our recursive method, compared with the bootstrap, has the advantage of reduction in unconditional variance. For theoretical interest, the questions are:

(i) Does $\tilde{F}_{n+m}$ converges when $m$ tends to infinity?
(ii) What is the form of the limit of $\tilde{F}_{n+m}$, if it exists?

## 2. Main Lemmas

We invoke the following lemmas to solve the problems mentioned in Section 1. We always suppose $X_1, X_2, \cdots, X_n$ $(n > 1)$ are i.i.d. random variables with distribution function $F$. Construct the empirical distribution function $F_n$. Furthermore, let $X_{n+1}$ be a sample from $F_n$ and use $X_1, X_2, \cdots, X_n, X_{n+1}$ to construct the empirical distribution function $\tilde{F}_{n+1}$. In this way, we get $X_{n+m}$ and $\tilde{F}_{n+m}$. Denote $S_{n+m}(x) = \tilde{F}_{n+m}(x) - F(x)$, $\mathscr{F}_n = \sigma(X_1, \cdots, X_n)$, $\mathscr{F}_{n+m} = \sigma(X_1, \cdots, X_n, X_{n+1}, \cdots, X_{n+m})$. It is clear that $\mathscr{F}_{n+m}$ is increasing in $m$.

LEMMA 1. *For any $x$ and $m$,*

(2.1) (i) $$ES_{n+m}(x) = 0; \quad \textit{and}$$

(2.2) (ii) $$ES^2_{n+m} = \left[\frac{n+2m+1}{(n+m)(n+1)}\right] F(x)(1 - F(x)).$$

https://doi.org/10.1017/S1446788700038088 Published online by Cambridge University Press

PROOF. (2.1) is obvious and we only need to prove (2.2). Since

$$\begin{aligned}
&\left[\tilde{F}_{n+m}(x) - F(x)\right]^2 \\
&\quad = \frac{1}{(n+m)^2}\Big\{(n+m-1)^2\left(\tilde{F}_{n+m-1}(x) - F(x)\right)^2 \\
&\qquad\qquad + 2(n+m-1)\left(\tilde{F}_{n+m-1}(x) - F(x)\right)\left(I_{\{X_{n+m}\le x\}} - F(x)\right) \\
&\qquad\qquad\qquad + \left[I_{\{X_{n+m}\le x\}} - F(x)\right]^2\Big\},
\end{aligned}$$

we obtain

$$\begin{aligned}
&E\left[\left(\tilde{F}_{n+m}(x) - F(x)\right)^2 |\mathscr{F}_{n+m-1}\right] \\
&\quad = \frac{1}{(n+m)^2}\left[(n+m)^2\left(\tilde{F}_{n+m-1}(x) - F(x)\right)^2 - \left(\tilde{F}_{n+m-1}(x) - F(x)\right)^2\right] \\
&\qquad + \frac{1}{(n+m)^2}E\left[I_{\{X_{n+m}\le x\}} - 2I_{\{X_{n+m}\le x\}}F(x) + F^2(x)|\mathscr{F}_{n+m-1}\right] \\
&\quad = \left[1 - \frac{1}{(n+m)^2}\right]\left[\tilde{F}_{n+m-1}(x) - F(x)\right]^2 \\
&\qquad + \frac{1}{(n+m)^2}\left[\tilde{F}_{n+m-1}(x) - 2F(x)\tilde{F}_{n+m-1}(x) + F^2(x)\right].
\end{aligned}$$

Recursively,

$$\begin{aligned}
ES^2_{n+m}(x) &= \left[1 - \frac{1}{(n+m)^2}\right]ES^2_{n+m-1}(x) + \frac{1}{(n+m)^2}F(x)(1-F(x)) \\
&= \left[1 - \frac{1}{(n+m)^2}\right]\left[1 - \frac{1}{(n+m-1)^2}\right]ES^2_{n+m-2}(x) \\
&\quad + \left\{\frac{1}{(n+m)^2} + \frac{1}{(n+m-1)^2}\left[1 - \frac{1}{(n+m)^2}\right]\right\}F(x)(1-F(x)) \\
&= \left[1 - \frac{1}{(n+m)^2}\right]\left[1 - \frac{1}{(n+m-1)^2}\right]\cdots\left[1 - \frac{1}{(n+1)^2}\right]\frac{F(x)(1-F(x))}{n} \\
&\quad - \left[1 - \frac{1}{(n+m)^2}\right]\left[1 - \frac{1}{(n+m-1)^2}\right]\cdots\left[1 - \frac{1}{(n+1)^2}\right]F(x)(1-F(x)) \\
&\quad + F(x)(1-F(x)) \\
&= F(x)(1-F(x))\left\{-\frac{(n+m+1)(n+m-1)}{(n+m)^2}\cdot\frac{(n+m)(n+m-2)}{(n+m-1)^2}\right. \\
&\qquad\qquad \left.\cdots\frac{(n+2)n}{(n+1)^2}\cdot\frac{n-1}{n} + 1\right\}
\end{aligned}$$

https://doi.org/10.1017/S1446788700038088 Published online by Cambridge University Press

$$= F(x)\,(1-F(x))\left[\frac{n+2m+1}{(n+m)(n+1)}\right].$$

REMARK 1. For any $x$ and $n, m$, we have

$$E\left(\tilde{F}_{n+m}(x)|\mathscr{F}_n\right) = F_n(x)$$

and

$$\lim_{m\to\infty} ES^2_{n+m}(x) = \frac{2}{n+1}F(x)(1-F(x)).$$

REMARK 2. Let $X_1^*, X_2^*, \cdots, X_m^*$ be an i.i.d. sample with distribution $F_n$,

$$F^*_{n,m}(x) = \frac{1}{m}\sum_{j=1}^{m} I_{\{X_j^*\le x\}} \tag{2.3}$$

and $S^*_{n,m}(x) = F^*_{n,m}(x) - F(x)$. To compare $ES^2_{n+m}(x)$ with $ES^{*2}_{n+m}(x)$, we calculate

$$\begin{aligned} ES^{*2}_{n,m} &= E\left[(F^*_{n,m}(x) - F_n(x)) + (F_n(x) - F(x))\right]^2 \\ &= F(x)(1-F(x))\left[\frac{1}{m}\left(1-\frac{1}{n}\right)+\frac{1}{n}\right]. \end{aligned}$$

It is easy to check that if $m < (n+1+\sqrt{(n+1)^2+4n(n+1)})/2$ then

$$ES^{*2}_{n,m}(x) > ES^2_{n+m}(x). \tag{2.4}$$

This means that $\tilde{F}_{n+m}(x)$ is preferable and our procedure is better than the bootstrap in the sense of small unconditional variance.

Now we consider the problems of some particles distributed on the two points 0 and 1. Suppose that initially there are $v$ particles at 0 and $u$ particles at 1. Denote by $s_0 = u/(u+v)$ the initial proportion of particles at 1. The $(u+v+1)$-th particle is allocated to 1 with probability $u/(u+v)$ and to 0 with probability $v/(u+v)$. Furthermore, if the $u+v+(k-1)$ particles are allocated with $A$ particles at 0 and $B$ particles at 1 $(A+B = u+v+k-1)$, then the $(u+v+k)$-th particle will be put at 1 with probability $B/(A+B)$ and at 0 with probability $A/(A+B)$. Denote by $s_k$ the proportion of particles at 1 at the $k$-th stage. It is clear that $s_k$ equals $(u+j)/(u+v+k)$ for some $j = 0, 1, \cdots, k$. What is the probability of this equality? We may think of $(A, B)$ as a two dimensional vector with the above 'transition probabilities'. Consequently,

$$P\left(s_k = \frac{u}{u+v+k}\right) = \frac{v(v+1)\cdots(v+k-1)}{(u+v)\cdots(u+v+k-1)} \cdot \binom{k}{0},$$
$$P\left(s_k = \frac{u+1}{u+v+k}\right) = \frac{u}{u+v}\left(\frac{v}{u+v+1}\right)\left(\frac{v+1}{u+v+2}\right)\cdots\left(\frac{v+k-2}{u+v+k-1}\right) \cdot \binom{k}{1},$$
$$P\left(s_k = \frac{u+j}{u+v+k}\right) = \frac{u(u+1)\cdots(u+j-1)}{(u+v)(u+v+1)\cdots(u+v+j-1)}$$
$$\cdot\left(\frac{v}{u+v+j}\right)\left(\frac{v+1}{u+v+j+1}\right)\cdots\left(\frac{v+k-1-j}{u+v+k-1}\right) \cdot \binom{k}{j},$$
$$P\left(s_k = \frac{u+k}{u+v+k}\right) = \frac{u(u+1)\cdots(u+k-1)}{(u+v)(u+v+1)\cdots(u+v+k-1)} \cdot \binom{k}{k}.$$

LEMMA 2. *The limiting distribution of $s_k$ as $k \to \infty$ is the beta distribution with parameters $u$ and $v$.*

PROOF. Consider the general form of the Polya-Eggenberger distribution, that is, a mixed binomial distribution where the success probability has the beta $(u, v)$ distribution. That is, the random variable $\alpha_k$ with

$$P\,(\alpha_k = j) = \binom{k}{j}\frac{B(u+j, v+k-j)}{B(u, v)}$$

for $j = 0, 1, \cdots, k$. By the property of this distribution, $\alpha_k / k$ converges in law to beta $(u, v)$. See Johnson and Kotz [3, pp. 177–181]. Now let $s_k = (u + \alpha_k)/(u + v + k)$ in our problem and Lemma 2 follows immediately.

## 3. Convergence of the resampling procedure

In this section, we discuss the almost surely convergence of $S_{n+m}(x)$ as $m \to \infty$, and we find its limiting distribution.

THEOREM 3.1. *Let $x$ and $n$ be fixed. Then,*

(i) *$S_{n+m}(x)$ is a $\{\mathscr{F}_{n+m}\}$ martingale;*

(ii) *$S_{n+m}(x) \to S(x)$ as $m \to \infty$,* a.s. *where $S(x)$ is a random variable depending on $x$ and $n$; and*

(iii) *$S(x) + F(x)$ has the distribution function*

$$(3.1)\quad H(x) = D(0)(1-p)^n + \sum_{j=1}^{n-1}\binom{n}{j}\frac{(1-p)^{n-j}p^j}{B(j, n-j)} \cdot \int_0^x (1-y)^{n-j-1}y^{j-1}\,dy + D(1)p^n,$$

https://doi.org/10.1017/S1446788700038088 Published online by Cambridge University Press

*where $p = F(x)$, $D(0)$ and $D(1)$ are the degenerate distribution functions at 0 and 1, respectively.*

PROOF. (i) It is obvious from Lemma 1 that $E[\tilde{F}_{n+m}(x)|\mathscr{F}_{n+m-1}] = \tilde{F}_{n+m-1}(x)$ which leads directly to our assertion.

(ii) Since $|S_{n+m}(x)| \leq 1$, the martingale convergence theorem (cf. Chow and Teicher [1]) shows that $S_{n+m}(x) \to S(x)$(a.s.), an $L_1$ random variable.

(iii) Let $\xi_n(x) = \lim_{m\to\infty} \tilde{F}_{n+m}(x) = S(x) + F(x)$ and $p = F(x)$. Suppose that $u$ of the $X_1, X_2, \cdots, X_n$ are located to the left of $x$. This means that $u$ of $I_{\{X_i \leq x\}}$ equal 1 and $v = n - u$ of $I_{\{X_i \leq x\}}$ equal 0. This sets our problem into the context of the above particle scheme. Hence, from Lemma 2, the conditional limiting distribution is a beta distribution with parameters $u$ and $v$. In particular, $u = n$ leads to the degenerate limiting distribution at 1, while $u = 0$ leads to the degenerate limiting distribution at 0. Finally the distribution of $u$ is binomial with parameter $p = F(x)$, completing the proof.

We can calculate the moments of $\xi_n(x)$ using (3.1):

$$E\xi_n(x) = \sum_{j=1}^{n-1} \binom{n}{j} (1-p)^{n-j} p^j \int_0^1 \frac{(1-y)^{n-j-1} y^{j-1} y}{B(j, n-j)}\, dy + p^n, \tag{3.2}$$

$$\begin{aligned} E\xi_n^2(x) &= \sum_{j=1}^{n-1} \binom{n}{j} (1-p)^{n-j} p^j \int_0^1 \frac{(1-y)^{n-j-1} y^{j-1} y^2}{B(j, n-j)}\, dy + p^n \\ &= \frac{n-1}{n+1} p^2 + \frac{2}{n+1} p, \end{aligned} \tag{3.3}$$

$$\begin{aligned} \operatorname{Var} \xi_n(x) &= \frac{n-1}{n+1} p^2 + \frac{2}{n+1} p - p^2 \\ &= \frac{2}{n+1} F(x)(1 - F(x)). \end{aligned} \tag{3.4}$$

Note that $\xi_n(x)$ is a increasing function of $x$. In the same fashion as the proof of the Glivenko-Cantelli theorem, we obtain

THEOREM 3.2. *For fixed $n$*

$$\lim_{m\to\infty} \sup_x \left| \tilde{F}_{n+m}(x) - \xi_n(x) \right| = 0 \qquad \text{a.s.} \tag{3.5}$$

We omit the proof.

## Acknowledgement

We are very grateful to Professor Tony Pakes and the referee for helpful suggestions and discussion.

## References

[1] Y. S. Chow and H. Teicher, *Probability theory* (Springer-Verlag, New York, 1978).
[2] B. Efron, *The jackknife, the bootstrap, and other resampling schemes* (SIAM, Philadelphia, 1982).
[3] N. L. Johnson and S. Kotz, *Urn models and their application* (Wiley, New York, 1977).
[4] D. P. Rubin, 'The Bayesian bootstrap', *Ann. Statist.* **9** (1981), 130–134.

Department of Statistics and Operations Research
Fudan University
Shanghai, 200433
China

https://doi.org/10.1017/S1446788700038088 Published online by Cambridge University Press