

**STATISTICS  
FOR THE  
21<sup>ST</sup> CENTURY**

**Methodologies  
for Applications  
of the Future**

edited by

**C. R. Rao**

*Pennsylvania State University  
University Park, Pennsylvania*

**Gábor J. Székely**

*Bowling Green State University  
Bowling Green, Ohio*

*and*

*Alfréd Rényi Institute  
of Mathematics  
Hungarian Academy of Sciences  
Budapest, Hungary*



MARCEL DEKKER, INC.

NEW YORK • BASEL

# Consistency and Accuracy of the Sequential Bootstrap

G. Jogesh Babu<sup>1</sup> and C. R. Rao<sup>2</sup>

Department of Statistics  
Pennsylvania State University  
University Park, Pennsylvania

P. K. Pathak

Department of Statistics and Probability  
Michigan State University  
East Lansing, Michigan

**Abstract.** The object of this paper is to present a brief account of the sequential bootstrap from a survey sampling point of view. This sequential resampling scheme entails resampling from the observed sample sequentially (with replacement) until a preassigned number of distinct original observations appear. This approach stems from the observation made by Efron in 1983 that the usual bootstrap samples are supported on approximately  $.632n$  of the original data points. We outline a number of approaches that can be employed to study the theoretical as well as the empirical properties of the sequential bootstrap. Our investigation shows that there is a great potential for sequential bootstrap in applications often encountered in practice.

---

<sup>1</sup> Supported by NSA grant MDA 904-97-1-0023 and NSA grant DMS-9626189.

<sup>2</sup> Supported by the Army Research Office under Grant DAAH04-96-1-0082.

## 1. Introduction

In this study we have examined the resampling procedure for bootstrap from a survey sampling point of view. Given an observed sample, resampling for the bootstrap involves  $n$  repeated trials of simple random sampling with replacement (SRSWR). It is well known that SRSWR does not yield samples that are equally informative ([10]) as it results in different number of distinct observations in different bootstrap samples. Our investigation shows that this randomness in the information content of the bootstrap samples is unnecessary. Stemming from the observation made by Efron ([6]) that the usual bootstrap samples are supported on approximately  $k = (1 - e^{-1})n \sim .632n$  of the original data points, we have proposed a sequential resampling procedure which keeps the information content of each sample at a constant level without affecting its correctness properties.

Let  $S = (x_1, \dots, x_n)$  denote a random sample from a distribution  $F$ , and let  $\theta(F)$  be a given parameter of interest. As an illustrative example we take  $\theta(F)$  to be the population mean:

$$(1.1) \quad \mu(F) = \int x dF.$$

Let  $F_n$  denote the empirical distribution function based on  $S$  so that  $F_n(x) = n^{-1} \sum \delta_{x_i}(x)$  in which  $\delta_{x_i}$  denotes the delta function at the  $i$ th sample observation  $x_i$ ,  $1 \leq i \leq n$ . We also assume for simplicity in the exposition that the population variance  $\sigma^2 = \int (x - \mu)^2 dF = 1$ . Consider now the plug in estimator  $\mu_n := \mu(F_n)$  of  $\mu(F)$  so that

$$(1.2) \quad \mu_n = \frac{1}{n} \sum_{i=1}^n x_i$$

with the corresponding pivot:

$$\pi_n = \sqrt{n}(\mu_n - \mu) = \sqrt{n}(\bar{x} - \mu)$$

in which  $\bar{x}$  is the sample mean and  $\mu$  is the population mean.

The central limit theorem entails that the sampling distribution of  $\pi_n$  can be approximated by the standard normal distribution. On the other hand, the bootstrap furnishes an alternative approach based on resampling to estimating the sampling distribution of  $\pi_n$  from  $S$  more precisely. Generally speaking the central limit approximation is accurate to  $o(1)$ , while resampling approximation is accurate to order  $o(\frac{1}{\sqrt{n}})$ . For example, let  $G_n$  denote the distribution of  $\pi_n$ . Then the central limit theorem yields that

$$(1.3) \quad \begin{aligned} \|G_n - \Phi\|_\infty &:= \sup_x |G_n(x) - \Phi(x)| \\ &= o(1) \end{aligned}$$

where  $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$ , while the bootstrap approximation captures the skewness of the distribution  $G_n$  in the following sense:

$$(1.4) \quad \|G_n - H_n\|_\alpha = o(1/\sqrt{n})$$

in which  $H_n$  represents a one-term Edgeworth expansion for  $G_n$ .

The usual bootstrap procedure to approximating the distribution  $G_n$  is based on resampling. Given the observed sample  $S$ , one selects a simple random sample with replacement (SRSWR) of size  $n$  from the observed sample  $S$ . Let  $S_n^* = (x_1^*, \dots, x_n^*)$  be an SRSWR sample drawn from  $S$  in this manner. Let  $F_n^*$  be the empirical distribution based on  $S_n^*$ . Let

$$(1.5) \quad \begin{aligned} \pi_n^* &= \sqrt{n}(\mu(F_n^*) - \mu(F_n))/\sigma(F_n) \\ &= \sqrt{n}(\mu_n^* - \mu_n)/s_n, \quad \text{say,} \end{aligned}$$

denote the pivot based on  $S_n^*$ , where  $s_n^2 = \frac{1}{n} \sum_1^n (x_i - \bar{x})^2$ . Then for large  $n$ , the conditional distribution of  $\pi_n^*$  given  $S$  is close to that of  $\pi_n$ . In practice, this conditional distribution of  $\pi_n^*$  is approximated by the observed frequency distribution (ensemble) of  $\pi_n^*$  obtained by repeated resamplings from  $S$  by SRSWR of size  $n$  a large number of times of order  $O(1000)$ . This observed frequency distribution is referred to as the bootstrap distribution of the original pivot  $\pi_n$ .

We can now examine what may be viewed as a certain drawback of this method of resampling. It is well known that owing to the with replacement nature of SRSWR, not all of the observations in an observed bootstrap sample  $S_n^*$  will be based on distinct observations from  $S$ . In fact, the information content of  $S_n^*$ , the set of distinct observations from  $S$ , is a random variable. Let  $\nu_n$  denote the number of distinct observations in  $S_n^*$ . Then

$$(1.6) \quad \begin{aligned} E(\nu_n) &= n[1 - (1 - \frac{1}{n})^n] \\ &= n(1 - e^{-1}) + O(1) \end{aligned}$$

$$(1.7) \quad \simeq n(.632)$$

$$(1.8) \quad V(\nu_n) = ne^{-1}(1 - e^{-1}) + O(1)$$

$$(1.9) \quad \simeq n(.32)n$$

so that

$$(1.10) \quad SD(\nu_n) \simeq (.48)\sqrt{n}.$$

Thus on the average, the usual bootstrap utilizes approximately 63% of the information from the original sample, the rest of the 37% of the data in it represents repetitious information. The  $2\sigma$ -limits for  $\nu_n$  are approximately  $(.63)n \pm 2(.48)\sqrt{n} \simeq (.63)n \pm \sqrt{n}$ . For example, if  $n = 100$ ,  $\nu_n$  ranges from a lower limit of 53 to an upper limit of 73 in approximately 95% of the bootstrap samples. This sort of randomness in the information content of bootstrap samples can be eliminated by adopting a sequential approach of the kind described in the following sections.

## 2. A Sequential Resampling Approach

To select a bootstrap sample, draw sample units (observations) from  $S$  sequentially by SRSWR until there are  $(m+1) \approx n(1 - e^{-1}) + 1$  distinct sample units (original observations from  $S$ ) in the observed bootstrap sample; discard the last observation. The recorded sequential bootstrap sample has the form:

$$(2.1) \quad S_N^* = (x_1^*, x_2^*, \dots, x_n^*)$$

in which the number of distinct observations is precisely  $[n(1 - e^{-1})]$ . The size  $N$  of the bootstrap sample is now a random variable with  $E(N) \approx n$ . The pivot based on  $S_N^*$  is

$$(2.2) \quad \begin{aligned} \pi_N^* &= \sqrt{N}(\mu(F_N^*) - \mu(F_n)) / \sigma(F_n) \\ &= \sqrt{N}(\mu_N^* - \mu_n) / s_n, \quad \text{say} \end{aligned}$$

in which  $F_N^*$  denotes the empirical distribution based on the sequential bootstrap sample  $S_N^*$ .

The simplest approach to studying the sequential bootstrap is to compare the pivot  $\pi_N^*$  (based on the sequential approach) with the pivot  $\pi_n^*$  (based on the bootstrap of constant size  $n$ ). It is easily seen that the random sample size  $N$  admits the following representation in terms of independent (geometric) random variables:

$$(2.3) \quad N = I_1 + I_2 + \dots + I_m$$

in which  $m = [n(1 - e^{-1})]$ ;  $I_1 = 1$ , and for each  $k$ ,  $2 \leq k \leq m$ ,

$$(2.4) \quad P(I_k = j) = \left(1 - \frac{k-1}{n}\right) \left(\frac{k-1}{n}\right)^{j-1}$$

for  $j = 1, 2, \dots$ . Therefore

$$(2.5) \quad \begin{aligned} E(N) &= n \left[ \frac{1}{n} + \frac{1}{(n-1)} + \dots + \frac{1}{(n-m+1)} \right] \\ &= n + O(1). \end{aligned}$$

Similarly

$$(2.6) \quad V(N) = \sum_{k=1}^m \frac{n(k-1)}{(n-k+1)^2} = n(e-2) + O(1).$$

Thus

$$(2.7) \quad \frac{E(N-n)^2}{n^2} = \frac{(e-2)}{n} + O\left(\frac{1}{n^2}\right)$$

so that  $(N/n)$  tends to 1 in probability. Further an analysis similar to that of [9] shows that Hajek's disparity between  $\pi_n^*$  and  $\pi_N^*$  satisfies:

$$(2.8) \quad \begin{aligned} \frac{E(\pi_N^* - \pi_n^*)^2}{\text{Var}(\pi_n)} &\leq k \sqrt{\frac{\text{Var}(N)}{n^2}} \\ &= O\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

This inequality implies that  $\pi_N^*$  and  $\pi_n^*$  are asymptotically equivalent, thus providing a simple justification that the sequential bootstrap provides its asymptotic correctness (consistency).

A second approach for consistency of the sequential bootstrap can be based on the so-called Mallows' metric (see, for example, [5]). Suppose that we define a sequence of distribution functions  $\{G_n : n \geq 1\}$  to converge in  $M$ -sense to a distribution function  $G$  iff (a) the sequence  $\{G_n : n \geq 1\}$  converges weakly to  $G$ , and (b)  $\int x^2 dG_n$  converges to  $\int x^2 dG$ . Then it is easily seen that this convergence is induced by the following metric ([8]):

$$(2.9) \quad d^2(F, G) = \inf_{x \sim F, y \sim G} E(x-y)^2$$

where the infimum is with respect to all random variables  $x$  and  $y$  such that  $x$  has the distribution  $F$  and  $y$  has the distribution  $G$ .

Under the metric (2.9), it is easily shown that

$$(2.10) \quad d(\pi_n^*, \pi_n) \leq d(F_n, F)$$

so that

$$(2.11) \quad \begin{aligned} d(\pi_n^*, \Phi) &\leq d(\pi_n^*, \pi_n) + d(\pi_n, \Phi) \\ &\leq d(F_n, F) + d(\pi_n, \Phi). \end{aligned}$$

The first term on the right converges to zero by the law of large numbers, while the second term does so by the central limit theorem. This establishes the consistency of the usual constant size bootstrap. The added complication in the sequential case is that the sample size  $N$  is now a random variable and for consistency to go through we need to show that  $d(\pi_N^*, \Phi)$  can be made arbitrarily small. Based on techniques similar to those of [9] and [10], it can be shown that

$$(2.12) \quad d(\pi_N^*, \pi_n^*) = O(n^{-1/4}).$$

Consistency of the pivot  $\pi_N^*$  now follows as a consequence of (2.11), (2.12) and the triangle inequality for metric spaces.

A limitation of the preceding two approaches is that they apply only to linear statistics and cannot be easily extended to more general statistics such as the sample quantiles. A third and more general approach is to treat pivots  $\pi_n^*$  and  $\pi_N^*$  as random signed measures and study their convergence in the functional sense. In this functional setting, we have established the following key results ([11]).

**Theorem 2.1** *Let  $F$  be a distribution function in  $R'$ . Let  $F_n^*$  and  $F_N^*$  denote empirical distribution functions based respectively on the usual and the sequential bootstrap samples. Then*

$$(2.13) \quad \|F_N^* - F_n^*\|_\infty = O_p(n^{-\frac{3}{4}})$$

$$(2.14) \quad \|\sqrt{N}(F_N^* - F_n^*) - \sqrt{n}(F_n^* - F_n)\|_\infty = O_p(n^{-\frac{1}{4}})$$

Moreover if  $F$  is uniform on  $[0,1]$ , the sequence of stochastic processes

$$\{\sqrt{N}(F_N^*(t) - F_n(t)) : 0 \leq t \leq 1\}$$

converge weakly to the standard Brownian bridge  $B(t) := w(t) - tw(1)$ , where  $w(t)$  is the standard Wiener process. More generally, if  $F$  is any continuous distribution function then this limit is  $(B \circ F)(t) = B(F(t))$ ,  $-\infty < t < \infty$ . Results such as these imply consistency of the sequential bootstrap for statistics of the form  $\theta_n = T(F_n)$  in which  $T$  is a compactly differentiable functional and includes functionals such as the quantiles.

The three different approaches of this section show that the distance between the sequential and the usual bootstrap is at most of the order of  $O(n^{-\frac{1}{4}})$ . Although this entails the consistency of the sequential bootstrap, it does not guarantee its second order correctness. To do so, one needs to capture the skewness of the pivot  $\pi_N^*$ . There are two approaches for accomplishing it.

### 3. Second Order Correctness of the Sequential Bootstrap

The proof of the second order correctness of the sequential bootstrap requires Edgeworth type expansions for dependent random variables. A simple rigorous justification of such expansion is unavailable in the literature. Along the lines of [7], we first outline an approach based on the computation of cumulants. This approach assumes that a formal Edgeworth expansion is valid for pivot under sequential bootstrap.

Let  $N_i$  denote the number of times the  $i$ th observation  $x_i$  from the original sample appears in the sequential bootstrap sample,  $1 \leq i \leq n$ . Then

$$(3.1) \quad N = N_1 + N_2 + \dots + N_n$$

in which  $N_1, N_2, \dots$  are exchangeable random variables.

The second order correctness of the sequential bootstrap for statistics such as the sample sum is closely related to the behavior of the moments of the random variables  $N_i$ 's,  $1 \leq i \leq n$ . Among other things the asymptotic distribution of each  $N_i$  should be Poisson with mean 1. In fact it can be shown that

$$(3.2) \quad E(N_1 - 1)^{k_1} \dots (N_l - 1)^{k_l} = \prod_{i=1}^l (e^\Delta - 1 - \Delta)(x - 1)^{k_i} + O\left(\frac{1}{n}\right)$$

where  $\Delta$  is the difference operator with unit increment.

It follows from (3.2) that to the order of  $O(1/n)$ , the  $N_i$ 's are asymptotically independent. This implies that the Hall-Mammen ([7]) type conditions for the second order correctness of the sequential bootstrap hold. This approach is based on the tacit assumption that formal Edgeworth type expansions go through in the sequential bootstrap.

A second approach without such assumption entails the following modification of the sequential bootstrap. It is based on the Poisson distribution.

### Poisson Resampling Scheme:

The original sample  $(x_1, \dots, x_n)$  is assumed to be from  $\mathbb{R}^k$  for greater flexibility. Let  $\alpha_1, \dots, \alpha_n$  denote  $n$  independent observations from  $P(1)$ , the Poisson distribution with unit mean. If there are exactly  $m = [n(1 - e^{-1})]$  non-zero values among  $\alpha_1, \dots, \alpha_n$ , take

$$(3.3) \quad S_N^* = \{(x_1, \alpha_1), \dots, (x_n, \alpha_n)\},$$

otherwise reject the  $\alpha$ 's and repeat the procedure. This is the conceptual definition. The sample size  $N$  of the Poisson resampling scheme admits the representation:

$$(3.4) \quad N = \alpha_1 + \alpha_2 + \dots + \alpha_n$$

in which  $\alpha_1, \dots, \alpha_n$  are iid Poisson random variable with mean  $\lambda = 1$  and with the added restriction that exactly  $m$  of the  $\alpha$ 's are non-zero, i.e.  $I_{\{\alpha_1 > 0\}} + \dots + I_{\{\alpha_n > 0\}} = m$ . Further it can be shown that the moment generating function  $M_N(t)$  of  $N$  is:

$$(3.5) \quad M_N(t) = \left[ \frac{(e^{(e^t-1)} - e^{-1})}{(1 - e^{-1})} \right]^m$$

so that the distribution of  $N$  can be viewed as that of  $m$  iid random variables with a common moment generating function:

$$(3.6) \quad m(t) = \frac{(e^{(e^t-1)} - e^{-1})}{(1 - e^{-1})}.$$

It is clear that  $m(t)$  is the moment generating function of the Poisson distribution with mean  $\lambda = 1$  and censored at  $x = 0$ . This representation of  $M_N(t)$  provides a practical way of implementing this scheme. To implement Poisson resampling scheme, first assign at random  $(n - m)$   $\alpha$ 's to zero and then to the remaining  $m$   $\alpha$ 's assign values independently chosen from the Poisson distribution with mean  $\lambda = 1$  and censored at  $x = 0$ .

This modification of the sequential bootstrap enables us to develop a rigorous proof of the second order correctness in the sequential case. The techniques we use are similar to those of [1], [3] and [4].

Let  $\{Y_j : j \geq 1\}$  be a sequence of iid Poisson random variables with mean  $\lambda = 1$ . Let

$$(3.7) \quad V_n^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})(x_j - \bar{x})'$$

$$(3.8) \quad c_j = V_n^{-1}(x_j - \bar{x})$$

$$(3.9) \quad P_n(x) = \frac{1}{6n} \sum_{j=1}^n ((c_j x)^3 - 3(c_j^2)(c_j x))$$

$$(3.10) \quad N = \sum_{j=1}^n y_j$$

$$(3.11) \quad T_n = \sum_{j=1}^n I_{\{y_j > 0\}}$$

$$(3.12) \quad \bar{Y} = \frac{N}{n}$$

$$(3.13) \quad U_n = \frac{1}{\sqrt{n}} \sum_{j=1}^n y_j V_n^{-1}(x_j - \bar{x})$$

$$(3.14) \quad \bar{\Psi}_n(x) = \left(1 + \frac{1}{\sqrt{n}} P_n(x)\right) \frac{e^{-\|x\|^2/2}}{(\sqrt{2\pi})^k}$$

For a real-valued function  $h$ , let

$$(3.15) \quad M_h = \sup |h(x)|(1 + \|x\|)^{-3}$$

and for any  $\delta > 0$ , let

$$(3.16) \quad \omega(h, \delta; x) = \sup_{\|x - z\| < \delta} |h(x) - h(z)|$$

$$(3.17) \quad \omega(h, \delta) = \int \omega(h, \delta; x) \phi(x) dx$$

where  $\phi$  denotes the standard normal density.

In terms of the foregoing terminology, the following results furnish a rigorous justification for the second order correctness of the sequential bootstrap. These results are consequences of a technical result on conditional Edgeworth expansions for weighted means of multivariate random vectors presented in [2].

**Theorem 3.1** *Let  $x_1, x_2, \dots$  be i.i.d. random vectors with mean  $\mu$  and covariance matrix  $\Sigma$ . Let  $H$  be a 3-times continuously differentiable function in a neighborhood of  $\mu$ . Suppose that  $x_1$  has a strongly non-lattice*

distribution and  $E(\|x_1\|^3) < \infty$ . Let  $l(y)$  denote the vector of first order partial derivatives at  $y$ , and  $l(\mu) \neq 0$ . If  $m - n(1 - e^{-1})$  is bounded, then for almost all sample sequences  $x_1, x_2, \dots$ , we have

$$\sup_z \sqrt{n} \left| P \left( \frac{\sqrt{N} (H(\frac{1}{N} \sum_{i=1}^n x_i y_i) - H(\bar{x}))}{\sqrt{l'(\bar{x}) V_n^2 l(\bar{x})}} \leq z \mid T_n = m; x_1, \dots, x_n \right) - P \left( \sqrt{n} (H(\bar{x}) - H(\mu)) \leq z \sqrt{l'(\mu) \Sigma l(\mu)} \right) \right| \rightarrow 0,$$

as  $n \rightarrow \infty$ .

The next result is more suitable for applications to studentized statistics.

**Theorem 3.2** Let  $\{x_n\}$  be as in Theorem 3.1. Suppose the function  $H$  is three times continuously differentiable in a neighborhood of the origin and  $H(0) = 0$ . If  $m - n(1 - e^{-1})$  is bounded, then for almost all sample sequences  $x_1, x_2, \dots$ , we have

$$\sup_z \sqrt{n} \left| P \left( \frac{\sqrt{N} H(\frac{1}{N} \sum_{i=1}^n (x_i - \bar{x}) y_i)}{\sqrt{l'(0) V_n^2 l(0)}} \leq z \mid T_n = m; x_1, \dots, x_n \right) - P \left( \sqrt{n} H(\bar{x} - \mu) \leq z \sqrt{l'(0) \Sigma l(0)} \right) \right| \rightarrow 0,$$

as  $n \rightarrow \infty$ .

It is easily seen that the second order correctness of the sequential bootstrap pivot such as

$$\pi_N^* = \sqrt{N} \left( \sum_{i=1}^n (x_j - \bar{x}) y_j \right) / s_n$$

given that  $T_n := \left( \sum_{j=1}^n I_{\{y_j > 0\}} \right) = m$  follows from Theorem 3.2. The one-term correction captures the skewness of the underlying distribution. For further details the reader is referred to [2].

## References

- [1] Babu, G.J. and Bai, Z.D. (1996), Mixtures of global and local Edgeworth expansions and their applications. *J. Multivariate Analysis*, 59, 282-307.
- [2] Babu, G.J., Pathak, P.K. and Rao, C.R. (1998) *Second order corrections of the sequential bootstrap*, submitted for publication.

- [3] Babu, G.J. and Singh, K. (1989), On Edgeworth expansions in the mixture cases. *Annals of Statistics*, **17**, 443–447.
- [4] Bai, Z.D. and Rao, C.R. (1992), A note on Edgeworth expansion for ratio of sample means. *Sankhya*, **54 A**, 309–322.
- [5] Bickel, P.J. and Freedman, D.A. (1981), Some asymptotic theory for the bootstrap. *Annals of Statistics*, **9**, 1196–1217.
- [6] Efron, B. (1983), Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.*, **78**, 316–331.
- [7] Hall, P. and Mammen, E. (1994), On general resampling algorithms and their performance in distribution estimation. *Annals of Statistics*, **24**, 2011–2030.
- [8] Mallows, C.L. (1972), A note on asymptotic joint normality. *Annals of Statistics*, **43**, 508–515.
- [9] Mitra, S.K. and Pathak, P.K. (1984), The nature of simple random sampling. *Annals of Statistics*, **12**, 1536–1542.
- [10] Pathak, P.K. (1964), Sufficiency in sampling theory. *Annals of Mathematical Statistics*, **35**, 795–808.
- [11] Rao, C.R., Pathak, P.K., Koltchinskii, V.I. (1997), Bootstrap by sequential resampling. *Journal of Statistical Planning and Inference*, **64**, 257–281.