QÜESTIIÓ, Vol 14, n° 1,2,3 pp. 43-88, 1990

EFFICIENT BOOTSTRAP SIMULATION AN OVERVIEW*

ALEX SANCHEZ

Dept. d'Estadística
Universitat de Barcelona

Two basic sources of error are associated to the use of bootstrap methods: one is derived from the fact that the true distribution is substituted by a suitable estimate, and the other is simulation errors. Some techniques to reduce or quantify these errors are discussed in this work. Some of them such as importance sampling or antithetic variates are adapted from classical Monte Carlo swindles, whereas others such as the centered and the balanced bootstrap are more specific. The existence of common methodological trends, such as the use of influence functions and Von Mises expansions to estimate the variance of the method is emphasized.

Keywords: Efficient bootstrap, Centered bootstrap, Linear bootstrap, Balanced bootstrap, Importance sampling, Antithetic sampling, Delta method, Influence Functions.

1. INTRODUCTION

Bootstrap methods (Efron (1979) [6], Efron (1982) [7] or Efron and Tibishirani (1986) [10]) allow us to assess the bias or the variability of a statistic (let us say T, an estimator for θ), or to estimate probabilities such as $Pr(T - \theta \leq d)$, most often in order to calculate confidence limits or to test hypotheses about θ.

*This work was done while the author was visiting the Department of Statistics at Stanford University. The author is grateful to his advisor, Dr. Jordi Ocaña, for introducing him to the topic and to Dr. Bradley Efron for a good number of helpful conversations. The work is partially supported by CGYCIT grant, PS89-0043.

-Article rebut el novembre de 1990.

Given a statistic T, which estimates a parameter $\theta(F)$, and a data sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from F we are interested in quantities such as the bias, the variance and the quantiles of $T(\mathbf{X})$, namely

$$(1.1) \qquad bias(T(\mathbf{X})) = E_F T(\mathbf{X}) - \theta(F)$$

$$(1.2) \qquad var(T(\mathbf{X})) = E_F T^2(\mathbf{X}) - (E_F T(\mathbf{X}))^2$$

$$(1.3) \qquad T^\alpha \quad \text{s.t.} \quad \text{prob}_F(T(\mathbf{X}) \leq T^\alpha) = \alpha.$$

The bootstrap procedure consists of replacing F with some estimate $\hat{F}$, in general the empirical distribution function, so that the estimates under the *bootstrap distribution* are:

$$(1.4) \qquad bias_B(T(\mathbf{X})) = E_{\hat{F}} T(\mathbf{X}^*) - \theta(\hat{F})$$

$$(1.5) \qquad var_B(T(\mathbf{X})) = E_{\hat{F}} T^2(\mathbf{X}^*) - (E_{\hat{F}} T(\mathbf{X}^*))^2$$

$$(1.6) \qquad T_B^\alpha \quad \text{s.t.} \quad \text{prob}_{\hat{F}}(T(\mathbf{X}^*) \leq T_B^\alpha) = \alpha$$

where $\mathbf{X}^* = (X_1^*, X_2^*, \ldots, X_n^*)$ is a *bootstrap sample*, i.e. a sample of size n from $\hat{F}$. If $\hat{F}$ is the empirical distribution and $T = \theta(F)$ then $\theta(\hat{F})$ is the observed value of T on the original sample. We shall consider it so from now on and call $\theta(\hat{F})$, T^o for brevity.

Expectations under the bootstrap distribution can only be obtained analytically in a few cases . However we can evaluate them using the following Monte Carlo algorithm:

1. Draw B samples, $\mathbf{X^*}_1, \mathbf{X^*}_2, \ldots, \mathbf{X^*}_B$ from $\hat{F}$
2. Form $T(\mathbf{X^*}_1), \ldots, T(\mathbf{X^*}_B)$
3. Let

$$(1.7) \qquad \overline{T} \equiv B^{-1} \sum_{b=1}^{B} T(\mathbf{X^*}_i)$$

then the approximate bootstrap estimates are:

$$(1.8) \qquad \overline{bias}_B(T(\mathbf{X})) = \overline{T} - T^o$$

$$(1.9) \qquad \overline{var}_B(T(\mathbf{X})) \quad = \quad 1/(B-1)\sum_{b=1}^{B}(T(\mathbf{X}^*{}_i) - \overline{T})^2$$

$$(1.10) \qquad \overline{T_B^\alpha} \quad \text{s.t.} \quad [\#(T(\mathbf{X}^*{}_i) \leq \overline{T_B^\alpha}]/\, B = \alpha$$

Let R be any of the preceding characteristics, the bias, the variance or the percentiles, and let R_F, $R_B(= R_{\hat{F}})$ and $\overline{R}_B(= \overline{R}_{\hat{F}})$ be the estimators under F, under the bootstrap distribution and the Monte Carlo approximation to R_B respectively.

All along the "bootstrapping" process there are different sources of error:

I . Those deriving from the fact that

$$R_B \neq R_F, \quad \text{because} \quad \hat{F} \neq F$$

Some questions arise here:

- Can we use another estimator of F, $\hat{F} = G$, instead of the empirical* c.d.f., $\hat{F} = F_n$ in order to improve the estimation of R_F by means of $R_{\hat{F}}$?.
- Can we find a better estimator than R_B $(R'_{\hat{F}})$
- How can both approaches be combined efficiently?.

II. Those due to the "finiteness" of the Monte Carlo approach, i.e. to the fact that:

$$\overline{R}_B \neq R_B, \quad \text{because} \ , \quad B < \infty$$

again some obvious questions are:

- Do any of the changes in (I.) (i.e. using R_G or $R'_{\hat{F}}$ instead of $R_{\hat{F}}$) affect the convergence rate of $\overline{R}_B$ to R_B as $B \to \infty$?
- Can we apply any of the classical ideas of Monte Carlo swindles (variance reduction techniques or VRT) to obtain more *efficient* estimators? (this generally meaning with minor variance).

  (Again we find ourselves in a similar situation to that in (I.) as some VRT use different estimators, whereas others change the sampling distribution).

We can, therefore find a series of *efficient bootstrap techniques* intended to reduce or quantify some of these errors (there is no necessary one-to-one relation between the methods and the errors mentioned above). From a wide perspective we can find methods that:

*whenever we need to differentiate between $\hat{F}$ and the empirical we shall use this notation.

- modify the estimates to obtain more efficient ones:

  The *centering method* of Efron (1988) [8].

  The *linear bootstrap* introduced by Davison, Hinkley and Schechtmann (1986) [4].

  Control function estimates, discussed by Therneau (1983) [34].

- use more sophisticated resampling schemes that hasten the convergence of $\hat{R}_B$ to $R_B = R_\infty$:

  The *balanced bootstrap* by Davison, Hinkley and Schechtmann (1986) [4] and Graham, Hinkley, John and Shi (1987) [14].

  The *accelerated simulation* procedures outlined by Ogbonwman and Wynn [32].

- use the Monte Carlo device of *importance sampling.*

  Introduced by Therneau (1983) [34] in the context of bootstrap estimation, it has been used by Johns (1988) [31] in a quantile estimation problem, by Hinkley and Shi (1989) [28] in a double bootstrap problem, and has been widely reviewed by Hesterberg (1988) [26].

- use computationally cheap methods to estimate the errors in (I.) or (II.):

  The *jackknife-after-bootstrap* measures of error introduced by Efron (1990) [9].

2. CENTERED AND LINEARIZED BOOTSTRAP ESTIMATES

2.1 The Centering Method. (Efron (1988))

The work by Efron* (1988) [8], in contrast with others we shall coment on later, obtains improved bootstrap estimates by means of sampling from the empirical c.d.f. in the *ordinary bootstrap* way, and modifying the calculations to estimate R_B, using say $\hat{R}_B$, instead of $\overline{R}_B$.

The estimators introduced, $\widehat{bias}_B$, $\widehat{var}_B$, and $\widehat{T^\alpha_B}$, are obtained with no additional cost using information contained in the bootstrap samples.

For each of them some kind of "diagnostic", is given to assess the gain in efficiency in using these estimates instead of the "straightforward ones".

*All along this section, when omitting the reference, we are referring to this work.

The improved bias estimate. Let $\mathbf{P}$ be the *resampling vector* associated with a given bootstrap sample, $(X_1^*, X_2^*, \ldots, X_n^*)$, i.e. the vector:

$$(2.1) \qquad (P_1, ..., P_n) \text{ s.t. } \quad P_i = \frac{\#\{X_j^* = x_i\}}{n}, \quad j = 1, \ldots, n$$

$$(2.2) \qquad \text{and} \quad \sum_{i=1}^{n} P_i = 1.$$

so that the statistic evaluated on a bootstrap sample, $T(\mathbf{X}^*)$ is $T(\mathbf{P})$. The resampling vector corresponding to the original sample is $P^o = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$.

Let $\overline{\mathbf{P}}$ be the average of the resampling vectors:

$$(2.3) \qquad \overline{\mathbf{P}} = B^{-1} \sum_{b=1}^{B} \mathbf{P}^b.$$

and let

$$(2.4) \qquad T(\overline{\mathbf{P}})$$

be the statistic evaluated on it.

The bias estimate introduced here

$$(2.5) \qquad \widehat{bias}_B = \overline{T} - T(\overline{\mathbf{P}})$$

is a substantial improvement on the straightforward bias estimate (1.8).

The superiority of $\widehat{bias}_B$ over $\overline{bias}_B$ may be exactly shown in the case where T is a *quadratic, functional* statistic (Efron (1982) [8], sections 2.6, 4.4. and 6.1.), i.e. a statistic that only depends on the data $\mathbf{X}$ through $\hat{F}$ (that is what functional means) and is of the form:

$$(2.6) \qquad T(\mathbf{P}) = T^o + (\mathbf{P} - \mathbf{P}^o)'\mathbf{U} + \frac{1}{2}(\mathbf{P} - \mathbf{P}^o)'\mathbf{V}(\mathbf{P} - \mathbf{P}^o)$$

where $\mathbf{U}$ is a $n \times 1$ vector s.t. $\sum_{j=1}^{n} U_j = 0$, and $\mathbf{V}$ a $n \times n$ symmetric matrix s.t. $\sum_{j'=1}^{n} V_{jj'} = \sum_{j'=1}^{n} V_{j'j} = 0$. $\mathbf{U}$ is the empirical influence function

for T and $\mathbf{V}$ is the second order empirical influence function for T (see Efron (1982) [7], Huber (1977) [30] or Hampel et al. (1986) [25]).

The justification of why $\widehat{bias}_B$ performs better than $\overline{bias}_B$ in this case comes from seeing that $\widehat{bias}_B$ matches $bias_B$ closer whereas $\overline{bias}_B$ adds extra linear terms that inflate the variance of the estimate.

A more generally centering argument (that gives its name to the method) is the following: The theoretical expectation of the resampling vectors $\mathbf{P}^1,\ldots,\mathbf{P}^n$ is $\mathbf{P}^o$, but their mean value is $\overline{\mathbf{P}}$, so using $\overline{T}-T(\overline{\mathbf{P}})$ instead of $\overline{T}-T(\mathbf{P}^o)$ corrects the bias estimate for values of $\overline{\mathbf{P}} \neq \mathbf{P}^o$.

Some simulation experiments in a problem of ratio estimation show $\widehat{bias}_B$ to be up to 50 times less variable than $\overline{bias}_B$, although, as the author comments, this gain is much greater than the usual one.

An orthogonal decomposition for the bootstrap replications of the statistic $T(\mathbf{P})$, based in its ANOVA decomposition (Efron and Stein (1981) [11]) is considered:

$$T(\mathbf{P}) = \mu + \alpha(\mathbf{P}) + \beta(\mathbf{P}) \tag{2.7}$$

where $\mu = E\{T(\mathbf{P})\}$, $\alpha(\mathbf{P}) = \mathbf{P}'\alpha$ ($\beta(\mathbf{P})$ is the "residual"), and α has components

$$\alpha_j = n[E\{T(\mathbf{P})|X_1^* = x_j\} - \mu]. \tag{2.8}$$

$\alpha(\mathbf{P})$ is the *linear part of* $T(\mathbf{P})$. It may be estimated by ordinary least squares, or by different versions of the empirical influence function* (which allows it to be called the *bootstrap influence vector*).

From the definition of

$$R^2 = var\{\alpha(\mathbf{P})\} \,/\, var\{T(\mathbf{P})\} \tag{2.9}$$

the relative efficiency of $\overline{bias}_B$ to $\widehat{bias}_B$ is established in *Theorem 1* in Efron (1988):

$$\frac{var(\overline{bias}_B)}{var(\widehat{bias}_B)} = \frac{1}{1-R^2}[1 + O_p(1/n) + O_p(1/\sqrt{B})] \tag{2.10}$$

*Although we do not develop the relation between α and $\mathbf{U}$ it's worth mentioning that they are closely related, especially in order to consider the relation between this method and others outlined below such as the linear bootstrap.

Thus, by estimating R^2, we can predict the gain in efficiency from using $\widehat{bias}_B$ instead of $\overline{bias}_B$. For instance in the ratio estimation example referred to above with $B = 20$ the average value of $\dfrac{var(\overline{bias}_B)}{var(\widehat{bias}_B)}$ over 10 simulated samples is 65.5 and the predicted ratio is 65.4

The improved variance estimate. We consider a decomposition of the straightforward variance estimate,

$$\overline{var}_B = \hat{\alpha}'\widehat{\mathbf{\Sigma}}\hat{\alpha} + \overline{var}_B(\hat{\beta}) \tag{2.11}$$

where $\widehat{\mathbf{\Sigma}}$ is the usual unbiased estimate of covariance for $\mathbf{P}$:

$$\widehat{\mathbf{\Sigma}} = (B-1)^{-1}\sum_{b=1}^{B}(\mathbf{P}^b - \overline{b}P)(\mathbf{P}^b - \overline{b}P)', \tag{2.12}$$

and

(2.13)

$$\overline{var}_B(\hat{\beta}) = (B-1)^{-1}\sum_{i=1}^{B}[\hat{\beta}]^2 = (B-1)^{-1}\sum_{i=1}^{B}[T^b - \overline{T} - (\mathbf{P}^b - \overline{\mathbf{P}})'\hat{\alpha}]^2$$

(Note that here we are working with centered versions of T and $\mathbf{P}$, $T^b - \overline{T}$ and $(\mathbf{P}^b - \overline{\mathbf{P}})$).

From the preceding decomposition an improved estimate of $var\{T(\mathbf{P})\}$ is obtained:

$$\widehat{var}_B = \| \hat{\alpha} \|^2 + d_{n,B} \cdot \overline{var}_B(\beta) \tag{2.14}$$

where $d_{n,B} = 1 - n(n-1)/(B-1)(B-n-1)$. Essentially, the improvement comes from the substitution of the true covariance matrix $\mathbf{\Sigma}$ in place of $\widehat{\mathbf{\Sigma}}$ in (1.9).

In some simulation experiments (see tables 4 and 5 of Efron (1988)) the observed ratio $\overline{var}_B/\widehat{var}_B$ gave average values ranging from 8.7 to 1.4, so that the gain in efficiency, though clear, is not as great as in the bias case.

There is no result available, such as theorem 1, relating $\overline{var}_B$ to $\widehat{var}_B$. A normal theory analog is suggested:

$$\frac{var(\overline{var}_B)}{var(\widehat{var}_B)} \simeq \frac{1-r}{(1-R^4)-c(1-R^2)^2} \tag{2.15}$$

Though in the simulation experiments mentioned above this relation happens to be too large, ranging from 28.6 to 1.0 it seems to be roughly proportional to the real relationship, so that it may give at least an idea of the gain in efficiency.

Efron (1988) [8] relates the estimator $\widehat{var}_B$ to the control function estimates of Therneau considered below.

The improved percentile estimates. Again, a decomposition for the bootstrap replications of the statistic $T(\mathbf{P})$, in a linear and a residual part is considered:

$$T^b = T(\mathbf{P}^b) = \hat{L}^b + \hat{M}^b \tag{2.16}$$

where $\hat{L}^b = \mathbf{P}^{b'}\hat{\alpha}$ and $\hat{M}^b = T^b - \hat{L}^b$

From there the improved percentile estimate $\tilde{T}_B^\alpha$ is obtained by means of a *cumulant adjustment formula.* It consists in applying a correction to the values $\overline{T}^b \rightarrow \tilde{T}^b, b = 1, \ldots B$ such that the first four empirical cumulants of the corrected values $\tilde{\mathbf{L}}^b$ match the first four theoretical cumulants of $(\hat{\mathbf{L}}^b|\hat{\alpha})$ (i.e. with $\hat{\alpha}$ fixed). The improved percentile estimate is then the observed percentile, $\tilde{T}_B^\alpha$ for the $\tilde{T}^b, (\quad b = 1...B)$ values

A result for the asymptotic relative efficiency of both estimates is available (as *Theorem 2*):

$$\frac{var(\overline{T}_B^\alpha)}{var(\tilde{T}_B^\alpha)} \leq \frac{\alpha(1-\alpha)}{E[\pi(L)(1-\pi(L)]} \tag{2.17}$$

where

$$\pi(L) \equiv P\left[T(\mathbf{P}) < T_\infty(\alpha)|L\right]. \tag{2.18}$$

The use of this result as a method for diagnosing the increase of the efficiency is not immediate, in as much as $E[\pi(L)(1-\pi(L)]$ is difficult to evaluate. (It seems however that it might be properly estimated through logistic regression).

Two simulation experiments are referred to in Efron (1988). In the ratio estimation problem, there is always a much greater gain in efficiency when central quantiles are estimated that in the extreme quantiles' case. In another example (with the "bootstrap stars", the law school data) the gains are not so great, and even a loss may be observed for 95 and 97.5 percentiles.

Finally two other estimates of T^α are introduced. A regression estimate $\check{T}_B^\alpha$ based on applying the cumulant adjustment process, starting with

$$T^b = T(\mathbf{P}^b) = g(\hat{L}^b) + \hat{M}^b \tag{2.19}$$

instead of starting with (1.10), where $g(.)$ is some smooth function fitted to the scatterplot of $(\hat{L}^b, \hat{T}^b)$. This estimate may represent a modest improvement on $\tilde{T}_B^\alpha$.

Another estimate $\hat{T}_B^\alpha$ based on a different estimation of $\pi(\mathbf{P})$ in (2.18) is introduced in the last section. It has the (theoretical) advantage of achieving equality in (2.17) but it is also more difficult to evaluate and, up to now, experimental results only show that it performs as well as, but no better than $\tilde{T}_B^\alpha$.

Open problems. Some problems related with or arising from this work are the following:

- Is there any advantage in combining this approach with different resampling schemes, such as, for instance, those of Davison et al [3], or Johns [31]?.
- How do these methods apply to more complicated situations?.
- Can $E[\pi(L)(1-\pi(L)]$ be easily and accurately estimated (for instance through logistic regression as suggested), in order to make the predictions of theorem 2 available?.
- Would the use of a smoothed version of the bootstrap help to sidestep the difficulties in the estimation of π?.
- Will the answers to the former problems clarify the theoretical advantage of $\hat{S}_B^\alpha$ over $\tilde{S}_B^\alpha$?.
- Can the percentile estimation be modified for the (very common) case where we are interested in extreme probabilities?.

2.2 The Linear Bootstrap. (Davison et al (1986))

One of the ideas behind this approximation, made by Davison, Hinkley and Schechtmann (1986) [4], is one of the main variance reduction strategies, namely, integrate as much as possible analytically leaving only a small variance remainder to be simulated.

Let us consider (whenever possible) a von Mises expansion (see Efron (1982) [7], Hinkley and Wei (1984) [29] or Fernholz (1983) [12]) for the statistic of interest, $T(F)$:

(2.20)

$$T(\hat{F}) = T(F) + n^{-1}\sum_{j=1}^{n} L(X_j, F) + \frac{1}{2}n^{-2}\sum_{j=1}^{n}\sum_{k=1}^{n} Q(X_j, X_k, F) + \dots$$

where $L(X;F)$ is the first-order influence function of $T(F)$, $Q(X_j, X_k; F)$ the second order influence function and so on (The ANOVA decomposition of Efron and Stein (1981) [11] approaches this expansion as $n \to \infty$ see, e.g. Efron (1982) [7] sec. 4.3.).

Let the frequency of X_i in a (bootstrap) sample, $(X_1^*, X_2^*, \dots, X_n^*)$ from $\hat{F}$ be:

(2.21) $$f_i^* = card\ \{j : X_j^* = X_i\}.$$

The bootstrap approximation to (2.20) will be:

(2.22)

$$T^* = T(\hat{F}) + n^{-1}\sum_{j=1}^{n} f_j^* \underbrace{L(X_j, \hat{F})}_{\hat{L}_j} + \frac{1}{2}n^{-2}\sum_{j=1}^{n}\sum_{k=1}^{n} f_j^* f_k^* \underbrace{Q(X_j, X_k, \hat{F})}_{\hat{Q}_{jk}} + \dots$$

The use of this approximation would require the calculation of the empirical influence functions $\hat{L}_j = L(X_j; \hat{F})$, $\hat{Q}_{jk} = Q(X_j, X_k; \hat{F})$ up to the desired degree of accuracy (taking the first two terms would give us a *linear* approximation, and the first three terms a *quadratic* approximation).

Now suppose that we know how to calculate the first-order empirical influence functions $\hat{L}_j$ and that we wish to estimate the bias and the variance of T, that is we want to calculate $E_{\hat{F}}(T) - T^o$ and $var_{\hat{F}}(T)$. Standard results about influence

functions make it possible to prove that $E^*(L) = 0$ and $var^*(L) = n^{-1}\sum \hat{L}_j^2$, so that if we write:

$$(2.23)\quad T(\hat{F}) = \underbrace{T(F) + n^{-1}\sum_{j=1}^{n} L(X_j, F)}_{T_L} + D_L, \qquad (D_L = T(\hat{F}) - T_L)$$

$$(2.24)\quad T^* = \underbrace{T(\hat{F}) + n^{-1}\sum_{j=1}^{n} f_j^* \hat{L}_j}_{T_L^*} + D_L^* \qquad (D_L^* = T^* - T_L^*)$$

we have:

$$(2.25)\quad E_{\hat{F}}(T) = T^0 + E_{\hat{F}}(D_L)$$

$$(2.26)\quad var_{\hat{F}}(T) = n^{-2}\sum_{j=1}^{n} \hat{L}_j^2 + 2cov_{\hat{F}}(D_L, T_L) + var\hat{F}(D_L).$$

where only the terms involving D_L need to be estimated by simulation, e.g. with the following algorithm:

1. Compute T^o and $\hat{L}_j$, $(j = 1, \ldots, n)$, on the original sample
2. Let $b = 1, \ldots, B$ and compute, for each b:

$$(2.27)\quad T^{*b}$$

$$(2.28)\quad T_L^{*b} = T^o + n^{-1}\sum_{j=1}^{n} f_j^{*b}\hat{L}_j$$

$$(2.29)\quad D_L^{*b} = T^{*b} - T_L^{*b}$$

1. Calculate $\overline{D}_L^* = B^{-1}\sum_{b=1}^{B} D_L^{b*}$.

The *linear estimates* of the bias and the variance are:

$$(2.30)\quad bias_L = \overline{D}_L^* = B^{-1}\sum_{b=1}^{B} D_L^{b*}$$

$$(2.31) \qquad var_L = n^{-2}\sum_{j=1}^{n}\hat{L}_j^2 + 2B^{-1}\sum_{b=1}^{B}D_L^{*b}(T^* - T_L^o)$$

$$(2.32) \qquad +(B-1)^{-1}\sum_{b=1}^{B}(D_L^{*b} - \overline{D_L^*})^2$$

It's worth noting that the leading term in the variance estimate contains no simulation terms. This method may also be applied when the influence function is not known, by substitution of jackknife estimates for influence values.

Davison et al.(1986) [4] suggest that these results may easily be derived for any moment or cumulant of T, directly, as done here, or exploiting some properties of the moment generating function $M(\lambda)$ that enable us to write the bootstrap estimate of $M(\lambda)$ as:

$$(2.33) \qquad M^\dagger(\lambda) = \{n^{-1}\sum_j \exp(\lambda\hat{L}_j/n)\}^n E_\lambda^\dagger\{\exp(\lambda D_L)\}$$

where $E_\lambda^\dagger$ denotes expectation with respect to the probability distribution

$$(2.34) \qquad Pr\,(X^* = X_j) \propto \exp(\lambda\hat{L}_j/n) \quad (j = 1...n)$$

and only $E_\lambda^\dagger$ and its derivatives in (2.33) require simulation.

Davison et al (1986) [4] give some examples where they examine numerically the variability of the bootstrap estimates, in three cases: the sample mean, the correlation coefficient and the eigenvalues of a covariance matrix. They focus on percentile estimates of $T - \theta$ (instead of T) and distinguish between the *raw percentiles*, i.e. the straightforward percentile estimate (1.10) put as $\overline{T}_B^\alpha - T^o$ and the *Normal percentiles* (coming from the somehow strong assumption that T is approximately normal) $\widehat{bias} + \widehat{var}^{\frac{1}{2}} + \Phi^{-1}(\alpha)$.

In the cases where the normal approximation is appropriate it always improves the percentile estimation (i.e. diminishes its variability). Analogously, whenever the normal approximation is appropriate the linear bootstrap increases the efficiency of both the ordinary and the balanced bootstrap (also with normal approximation). In any situation, simply switching from ordinary to balanced resampling only represents a slight improvement, mainly for the central percentiles.

The author of this work gives some hints on how to improve these results such as controlling the quadratic component of (2.20) for the balanced resampling case or making higher moment corrections for the normal approximations.

Open Problems.

- Can the use of formula (2.33) for the m.g.f. of T^* provide computationally efficient methods for estimating the exact distribution of T^*?.
- As seen before the thinnest point in percentile estimation is tail probabilities estimation. It seems to deserve specially differentiated attention.

2.3 Relation between centered and linear bootstrap estimates. (Hall (1989)

Introduction The equivalence of the two former methods (and to a third one, *the balanced bootstrap* to be discussed below) is shown by Hall (1989) [17], in the case where the estimates are smooth functions of means*, by proving that the variances (and MSE) of each bootstrap estimate are asymptotically equivalent up to the same constant $(Bn^2)^{-1}$, what means an improvement over the ordinary bootstrap for wich the rate of convergenece of the variance is $(Bn)^{-1}$.

This improvement is explained showing that the application of each algorithm is equivalent to removing the linear term in a Taylor expansion of the ordinary bootstrap estimator.

As far as the results apply to smooth functions of means we shall adapt our notation to it, writing:

$$T^o = T(\overline{X}) \quad (\text{for } T(\mathbf{X})) \tag{2.35}$$

$$\overline{T}_{OB} = E_{\hat{F}}(T(\overline{X}^*)) \quad (\text{for } E_{\hat{F}}(T(\mathbf{X}^*))) \tag{2.36}$$

$$= B^{-1}\sum_{b=1}^{B} T(\overline{X}_b^*) \tag{2.37}$$

Methodology The Taylor expansion of $T(\overline{X}_b^*)$ is:

$$T(\overline{X}_b^*) = T(\overline{X}) + (\overline{X}_b^* - \overline{X})T'(\overline{X}) + \frac{1}{2}(\overline{X}_b^* - \overline{X})^2 T''(\overline{X}) - \ldots \tag{2.38}$$

If we take expectations on it we obtain the ordinary (unbalanced) bootstrap estimate $\overline{T}_{OB}$, and, taking variances on this expansion we have:

*This enables the proofs to be based on Taylor expansions for the statistics. It seems that, given that the results to be related have been established using the influence function tools, it would be nicer to use them here as well.

$$(2.39)\ var_{\hat{F}}(\overline{T}_{OB}) = 0 + var_{\hat{F}}(B^{-1}\sum_{b=1}^{B}(\overline{X}_b^* - \overline{X})T'(\overline{X})) + O(B^{-1}n^{-2})$$

$$(2.40)\qquad (Bn)^{-1}\hat{\sigma}^2 T'(\overline{X})^2 + O(B^{-1}n^{-2})$$

where $\hat{\sigma}^2 = n^{-1}\sum_i (X_i - \overline{X})^2$.

The core of this work lies in proving that Efron's centered estimate, $T(\overline{\mathbf{P}})$ (2.4), ($\equiv T_{CB}$), and Davison et al's linear and balanced estimates T_{BL}, T_{BB} are built in such a way that the variance corresponding to the linear part is eliminated.

- In the *linear estimate*, T_{LB} the linear term is explicitely removed, so that:

$$(2.41)\quad T_L = B^{-1}\sum_{b=1}^{B}\left\{T(\overline{X}_b^*) - (\overline{X}_b^* - \overline{X})T'(\overline{X})\right\}$$

  and therefore, from (2.40) its variance is of order $O(B^{-1}n^{-2})$.

- In Efron's *centered estimate*, the variance is not explicitely removed, but writing the *grand mean* (i.e. $T(\overline{bP})$) as $\overline{X}_{\cdot}^* = B^{-1}\sum_{b=1}^{B}\overline{X}_b^*$ we can write T_{LB} as

$$(2.42)\quad T_{LB} = B^{-1}\sum_{b=1}^{B}T(\overline{X}_b^*) - (\overline{X}_{\cdot}^* = B^{-1}\sum_{b=1}^{B}\overline{X}_b^* - \overline{X})T'(\overline{X})$$

  and now using

$$(2.43)\qquad (\overline{X}_{\cdot}^* - \overline{X})T'(\overline{X}) \simeq T(\overline{X}_{\cdot}^*) - T(\overline{X})$$

  we obtain the centered estimate:

$$(2.44)\qquad T_{CB} = B^{-1}\sum_{b=1}^{B}T(\overline{X}_b^*) - T(\overline{X}_{\cdot}^*) + T(\overline{X})$$

- by noticing that if $\overline{X}_{\cdot}^* = \overline{X}$ then the linear term correction is 0 we come to the *balanced method estimate*, T_{BB} where this equality is achieved making each observation appear exactly B times in the B resamples, so that if we call $X_b^\dagger = n^{-1}\sum_i X_{bi}^\dagger$ the mean under balanced resampling then the balanced estimate is:

$$(2.45)\qquad T_{BB} = B^{-1}\sum_{i=1}^{B}T(X_b^\dagger)$$

  T_{BB} is examinated in the next section.

It is worth noticing that these three estimates are equivalent in their asymptotic variance, but not in their unbiasedness, kept only by T_{LB}, (and by the ordinary estimate T_{OB})

Some generalizations of minor applicability are also provided in this work.

3. THE BALANCED BOOTSTRAP AND RELATED TECHNIQUES

3.1 The Balanced Bootstrap. (Davison et al. (1986))

The key idea of this method arises from the fact that, when we resample in the ordinary way from the empirical distribution (i.e. drawing n values with replacement from the original sample, $(X_1, X_2, \ldots, X_n)$) the frequency of each value in all B samples is not necessarily (probably never) the same as it has in the original sample ($1/n$ in the case where all values are different; if B were infinite, by the laws of large numbers the observed frequencies would be the same as those in the sample).

In order to eliminate this source of error Davison et al. (1986) [4] propose a *balanced resampling method* consisting of:

1. Concatenate B copies of $(X_1, X_2, \ldots, X_n)$ in a string of length nB
2. Randomly permute this string
3. Read off the B bootstrap samples as successive blocks of length n in the permuted string.

Gleason(1988) [13] suggests different algorithms to perform the above efficiently.

It can be seen that this algorithm arises if we obtain the frequencies of the element X_i, $i = 1...n$, in the bth resample, $b = 1...B$, f^*_{bi}, from a hypergeometric distribution with row sums $f_{b.}* \equiv n$ and column sums $f_{.i}* \equiv B$ instead of taking them from a multinomial distribution with expectations (1,...,1) as does the ordinary bootstrap.

Arguments for the superiority of the balanced bootstrap over the ordinary way of resampling are also given by Davison et al. (1986) [4], applying some calculations for the hypergeometric and multinomial distributions derived from Haldane (1940) [15]. To do this, they calculate the expectation and the variance of the bias estimate and the expectation for the variance estimate for the estimates of both ordinary and balanced methods. In every case the balanced method estimate turns out to be better.

The balanced bootstrap, will be obtained by a straightforward application of formulas (1.8), (1.9), (1.10) to the b balanced resamples obtained from the preceding algorithm.

3.2 The Accelerated Simulation Method. (Ogbonmwan and Wynn (1986))

In a contribution to the discussion of a 1986 paper by C.F.J. Wu [32], Wynn, H.P. and Ogbonmwan S.M. introduce the idea of *accelerated simulation* establi-

shing a frame in which the efficiency of a given resampling plan may be quantified.

If S is the set of all possible configurations that may be obtained sampling with replacement from $(X_1, X_2, \ldots, X_n)$ the ordinary bootstrap takes B samples from S by simple random sampling (s.r.s.). As happens in sampling theory it may be expected that a more sophisticated sampling scheme would give more accurate estimates.

An alternative to s.r.s. given by J. Tukey (1978) [35] is considered. It consists in making a complete enumeration from $S' \subset S$ where S' is much smaller than S and it fills out S in a certain dense way such that any inference based in S' is valid for S.

The authors give some hints on how to measure the discrepancy between the true and the estimated statistic and argue through them why the balanced bootstrap performs better.

The problem, of course, here (as in any problem in sampling theory) is how to choose S' (e.g. the balanced method, of Davison et al, is a good choice). They suggest that subsets chosen by minimizing the discrepancy or by a judicious "space-filling", together with balanced resampling, would do better than the ordinary bootstrap.

4. IMPORTANCE SAMPLING AND THE BOOTSTRAP

4.1 Introduction. Importance Sampling

In this work – as in the following – a classical Monte Carlo swindle, importance sampling, is used as a way to increase the accuracy of bootstrap estimates.

The traditional importance sampling method is concerned with the estimation of expectations (Hammersley and Handscomb (1964) [21]). Very briefly stated, the method goes as follows: Suppose we want to estimate the expectation of a response function $\phi(\cdot)$ from an input random vector $\mathbf{U} = [U_1, \ldots, U_m]'$ uniformly distributed over the m-dimensional unit cube with probability density $f_0(\mathbf{u})$, so that, in terms of the random variable $Y(\mathbf{U})$, the estimand of interest is:

$$\theta = E(Y) = \int_{R^m} \phi(\mathbf{u}) f_0(\mathbf{u}) d\mathbf{u} = \int_{I^m} \phi(\mathbf{u}) d\mathbf{u}. \tag{4.1}$$

This may, of course, be estimated by the sample mean $\overline{Y}_n$. For a given sample of size n, variance reduction techniques (VRT) usually yield an alternative estimator $\hat{\theta}_n$, with

(4.2) $$E(\hat{\theta}_n) = \theta, \quad \text{and} \quad var(\hat{\theta}_n) < var(\overline{Y}_n).$$

Importance sampling requires the input vector **U** to be sampled from an alternative density $f(\cdot)$ instead of the uniform density $f_0(\cdot)$. To compensate for this distortion of the input so as to achieve condition (4.2) the original response $Y = \phi(\mathbf{U})$ is replaced by the variate $Z = \phi(\mathbf{U})/f(\mathbf{U})$. The *importance estimator* $\hat{\theta}_n$ is then taken to be the sample mean Z_n computed over n independent replications of the new response Z. When the *importance density* $f(\cdot)$ closely mimics $\phi(\cdot)$ the ratio $\phi(\mathbf{U})/f(\mathbf{U})$ is nearly constant, and a substantial variance reduction is achieved. The trouble in this technique lies in the appropriate selection of $f(\cdot)$ up to the point that not only is variance reduction not guaranteed but also, with a poorly chosen importance density, large variance increases can occur (Bratley, Fox and Schrage (1987) [2]).

4.2 Importance Sampling for Bootstrap Confidence Intervals. (M. Vernon Johns (1988))

Quantile estimation Importance sampling is here applied to quantile estimation, a different problem from mean estimation and therefore the procedures must be adapted. Standard results for order statistics guarantee that, as m becomes large, if $f = F'$ exists and is continuous at ξ_p (the p-quantile of an r.v. ξ) then:

(4.3) $$\sqrt{m}(\hat{\xi}_p - \xi_p) \overset{dist}{\rightarrow} N(0, r^2)$$

where $r^2 = p(1-p)/f^2(\xi_p)$ and $\hat{\xi}_p$ is the sample p-quantile. The importance sampling approach is as follows:

1. Generate $Y_1, \ldots, Y_m \simeq G; \quad (G' = g)$
2. Let the ordered values be: $Y_{(1)} \leq \ldots \leq Y_{(m)}$
3. Let

(4.4)

$$S_r = \frac{1}{m} \sum_{i=1}^{r} \frac{f(Y_{(i)})}{g(Y_{(i)})} \quad 1 \leq r \leq m$$

where, by means of the order statistics theory, $S_r = F(G^{-1}(\frac{r}{m}))$, and given that $S_r = p \rightarrow Y_{(r)} = G^{-1}(\frac{r}{m}) = F^{-1}(p) = p$ the following importance sampling estimator is suggested:

$$(4.5) \quad \xi^*_{p,m} = Y_{(R)}, \quad \text{R s.t.} \quad \begin{cases} S_{(R)} \leq p \\ S_{(R+1)} > p \end{cases}$$

The problem, of course, is how to appropriately choose G in order to obtain an accurate estimator $\xi^*_{p,m}$ with moderate sizes of m.

By studying the asymptotic distribution of $\xi^*_{p,m}$ the author obtains the following result:

$$(4.6) \qquad \sqrt{m}(\hat{\xi}^*_{p,m} - \xi_p) \overset{dist}{\rightarrow} N(0, r^2)$$

where, letting $IC(Y;G)$ represent the influence curve for Y under the distribution G we have (again):

$$(4.7) \qquad r^2 = E_G(IC^2(Y;G)) = \frac{1}{f^2(\xi_p)} \left\{ \int_{-\infty}^{\xi_p} \frac{f^2(y)}{g(y)} dy - p^2 \right\}.$$

Alhough this should , in principle, allow us to reduce r choosing the appropriate $g(y)$ (close to $f(y)1(y \leq \xi_p)/p$), in practice it is rarely feasible for f and ξ_p are not fully known and only in particular cases may a substantial reduction be obtained.

Bootstrap confidence limits for location estimates The preceding importance sampling scheme may be combined with the *percentile method* for building bootstrap confidence intervals (see Efron and Tibishirani (1986) [10]) to obtain an importance sampling estimate of the pth quantile, t_p, of a location estimate $T(\hat{F})$. This is accomplished by associating a resampling probability $\hat{G}(\mathbf{X}) = (g_1, g_2, ..., g_n)$, $\sum_{i=1}^n g_i = 1$ to each value of $\mathbf{X} = (X_1, ..., X_n)$ in the original sample. So the likelihood of a bootstrap sample $\mathbf{X}^* = (X^*_1, X^*_2, \ldots, X^*_n)$, $\hat{G}$ is given by:

$$(4.8) \qquad l_G(\mathbf{X}^*) = \prod_{i=1}^n \prod_{j=1}^n g^*_{ij}, \quad \begin{cases} g^*_{ij} & = g_j & \text{if } X^*_i = X_j \\ & = 1 & \text{otherwise.} \end{cases}$$

The corresponding likelihood under ordinary bootstrap is $l_F(\mathbf{X}^*) = n^{-n}$, and so we may obtain the bootstrap analogue of (4.4):

$$S_r = \frac{1}{m}\sum_{i=1}^{r}\frac{l_F(\mathbf{X}^*)}{l_G(\mathbf{X}^*)} \quad 1 \le r \le m \tag{4.9}$$

so that, as in (4.4), the estimate of the pth bootstrap quantile of $T(\hat{F})$ is $T_{(R)}$ where R is such that $S_R \le p$ and $S_{R+1} > p$.

To choose T appropriately , a representation for $T(\hat{F}) - T(F)$ is introduced: Let $Z_i = (X_i - T(F))/S(F)$, where $S(\cdot)$ is some location-and-scale equivariant scale functional, then:

$$\sqrt{n}(T(\hat{F}) - T(F)) = 1/\sqrt{n}\sum_{i=1}^{n} h(Z_i) + o_P(1) \tag{4.10}$$

This enables us to generate the distribution $\hat{G}$ by *exponential tilting*, which consists in letting:

$$h_i = f\left[\frac{X_i - T(\hat{F})}{S(\hat{F})}\right], \quad i = 1, 2...n \tag{4.11}$$

for the original $X_i's$ and setting:

$$g_i = \exp\{a(h_i/\sqrt{n}\hat{\sigma}) + b\} \tag{4.12}$$

where $\hat{\sigma}^2 = (1/n)\sum_{i=1}^{n} h_i^2$ and b is chosen so that $\sum_{i=1}^{n} g_i = 1$. Now resampling under $\hat{G}$ as before we obtain the following likelihood for the bootstrap sample:

$$l_G(\mathbf{X}^*) = \exp\{aV^* + nb\} \tag{4.13}$$

where $V^* = 1/\sqrt{n}\hat{\sigma}\sum_{i=1}^{n} h_i$ and $h_i^* = h_j \longleftrightarrow X_i^* = X_j$. Under *ordinary bootstrap* the quantity V^* is approximately $N(0,1)$, which in turn is the approximate distribution of $(\sqrt{n}/\sigma)(T(\hat{F}) - T(F))$. Under *tilted bootstrap*, V^* is approximately $N(a,1)$ which is then the approximate distribution of $(\sqrt{n}/\sigma)(T(\hat{F}^*) - T(F))$, so that the approximate distribution of $T(\hat{F}^*)$ is centered at $T(F) +$

$a\sigma/\sqrt{n}$ and has variance σ^2/n, which is the same as the approximate variance of $T(\hat{F})$.

Now, the importance sampling estimate, $T_{(R)}$ given by (4.9) has asymptotic variance given by (4.7), which may be minimized choosing g to minimize $I = \int_{-\infty}^{t_p} f_n^2(x)/g_n(x)dx$ where, in the present context, f_n and g_n are the approximate densities of $T(\hat{F})$ and $T(\hat{F}^*)$ respectively. With an adequate change of variables we can now substitute f_n and g_n in I by $N(0,1)$ and $N(a,1)$ so that the upper limit t_p now becomes the pth quantile of the $N(0,1)$, z_p and a that must be chosen to minimize I. The minimizing values of a may be obtained numerically for given values of p.

Johns (1988) [31] applies this method to build confidence intervals for a robust location estimate. The method is compared with ordinary bootstrap resampling. Both resampling procedures are also applied to a studentized version of the statistic. In every case the estimate obtained by importance sampling has as much as 10-fold less variability than the one obtained by ordinary bootstrap. Though this method works fairly well when the sample size is small, a slight increase in the variability of the lengths of the intervals is observed for a small number of replications (10). In general the studentized statistic produces coverage probabilities closer to the nominal value of the confidence level in both cases but also with an increase in the variablity of the interval lengths.

Open Problems

- The method of exponential tilting leaves the variance of the statistic unchanged. Some other transformations performing in the same way, but directly reducing the variance, might be tried.
- The performance of the method should be *numerically and extensively* contrasted with other methods to increase the efficiency such as saddlepoint approximations or the use of estimates of the moment-generating function of a linearized version of the statistic (as cited in Davison et al (1986) [4]).
- The method might be generalized to many other bootstrap confidence intervals methods, especially those in which the statistic could admit a representation of the form (4.10).

4.3 Importance sampling and the nested bootstrap. (D.V. Hinkley and S. Shi (1989))

Introduction Iterative bootstrap methods have been proposed by different authors (Beran (1987) [1], Hall and Martin (1988) [20], Di Ciccio and Romano (1988) [5] and in the work by Hinkley and Shi (1989)[28]) as a way to achieve

higher order accuracy when building confidence intervals. A comparative study of these methods and others that do not use iteration, such as Efron's percentile methods, has been done by Hall (1988) [16]. Except for the case of the ***automatic percentile method*** (Di Ciccio and Romano (1988) [5]) -which may be iterated but in which the number of required calculations increases linearly with the number of iterations- all other iterative methods have the drawback that the number of calculations increases exponentially with the number of iterations, so that higher accuracy is only obtainable at a higher computational cost. This is then an appropriate situation to apply some kind of variance reduction technique, and in this work the authors apply importance sampling to a new iterative bootstrap method, the *nested bootstrap*.

The nested bootstrap This method may be viewed as making a correction to the percentile limit (see Efron (1982) [7] or Efron and Tibishirani (1986) [10]), where the upper P confidence limit for an estimate $T(F)$ is simply $T^*_{(R)}$ where $T^*_{(\cdot)}$ represents the ordered values of the B bootstrap replications of $T(\hat{F})$ and $R = \lfloor BP \rfloor$ is the integer part of BP. The implicit idea is to estimate the deviation from P in coverage of the percentile limit $T'_{(R)}$ and hence to define an adjusted value R' of R such that $T^*_{(R')}$ will give coverage closer to P. It turns out that this *modified percentile limit* is the Q quantile of the bootstrap distribution T^*, that is $T^*_{(S)}$ with $S = \lfloor BQ \rfloor$ where Q satisfies the double bootstrap equation:

$$Pr\left[Pr(T^{**} \leq T|\hat{F}^*) \leq Q|\hat{F}\right] = P, \tag{4.14}$$

and writing $U^* = Pr(T^{**} \leq T|\hat{F}^*)$ (4.14) may be expressed as:

$$P = Pr(U^* \leq Q|\hat{F}) \tag{4.15}$$

where U^*, as an c.d.f it is, will be approximately uniformly distributed for large values of n.

Equation (4.14) may be solved using simulated samples to estimate the probabilities in the folllowing way:

1. Let b run from 1 to B.
2. Generate a sample $(X_1^*, X_2^*, \ldots, X_n^*)^b$ from $\hat{F}$ and calculate T_b^*
   (a) Let m run from 1 to M.
   (b) Generate a sample $(X_1^{**}, \ldots X_n^{**})^{bm}$ from $\hat{F}^*$ and calculate T_{bm}^* (this will give M estimates $T_{b1}^{**}, \ldots, T_{bM}^{**}$)

(this will give B estimates, $T_1^*, \ldots, T_B^*$ and a total of BM estimates T_{bm}^{**}).

Now if we let $I_A(y)$ denote the indicator function for a set A then the empirical version of (4.15) is:

$$P = B^{-1} \sum_{b=1}^{B} I_{[0,Q]}(\hat{U}_b^*) \tag{4.16}$$

where $\hat{U}_b^*$ $b = 1...B$ analogously estimates the innermost probabilities of (4.15). The solution $\hat{Q}$ to (4.16) can be found ordering the $\hat{U}^{*'}s$ and then solving

$$\sum_{i=1}^{R'} \hat{U}_{(i)}^* \leq BP \leq \sum_{i=1}^{R'+1} \hat{U}_{(i)}^* \tag{4.17}$$

for $R' = \lfloor BQ \rfloor$. The double bootstrap confidence limit is $T_{(R')}^*$.

Two-level Importance sampling The key computation of the method has a simple analogue in terms of normal distributions. An importance sampling scheme for Monte Carlo implementation of this analogue is first derived and then translated into the bootstrap context.

The main idea for the normal analogue consists of considering that, "as often happens" T is approximately normally distributed $T \sim N(\theta, n^{-1}V)$, with $V = var\{L(X,F)\}$, and $L(x,F)$ is the influence function of T at x under F. From this it follows that $T^*|\hat{F}$ is approximately $N(T, n^{-1}\hat{V})$ and T^{**} is approximately $N(T^*, n^{-1}\hat{V}^*)$ so that (4.14) becomes:

$$P = Pr\{Pr(Z \leq \mu|Y) \leq Q\} \tag{4.18}$$

where $Y \sim N(\mu, \sigma^2)$ and $(Z|Y = y) \sim N(y, \sigma^2)$.
The approach is done in two stages.

- First we suppose that $U = u(y)$ is known, so that we can estimate P as:

$$\hat{P}_0 = B^{-1} \textstyle\sum_{b=1}^{B} I_{[0,Q]}(U_i). \tag{4.19}$$

Following thue usual importance sampling approach we write:

$$P = \int I_{[0,Q]}(u(y)) \left[\frac{\sigma^{-1}\phi\{(y-\mu)/\sigma\}}{g(y)} \right] g(y) d(y) \tag{4.20}$$

which leads us to the estimate

$$B^{-1} \textstyle\sum_{b=1}^{B} I_{[0,Q]}(U_b') \left[\frac{\sigma^{-1}\phi\{(Y_b'-\mu)/\sigma\}}{g(Y_b')} \right] \tag{4.21}$$

where the Y_b' are sampled from $g(.)$ and $U_b' = u(Y_i')$. Now taking $g(y) = \sigma^{-1}\phi\{(y-\mu-\eta\sigma)/\sigma\}$ which makes $Y \sim N(\mu+\eta\sigma, \sigma^2)$, we arrive at the estimate, depending on a parameter, η:

$$\hat{P}_\eta = \begin{cases} \exp(\frac{1}{2}\eta^2)B^{-1}\sum_{b=1}^{B} I_{[0,Q]}(U_b')e_b & (P < \frac{1}{2}), \\ 1 - \exp(\frac{1}{2}\eta^2)B^{-1}\sum_{b=1}^{B} I_{(Q,1]}(U_b')e_b & (P \geq \frac{1}{2}) \end{cases} \tag{4.22}$$

where $E_b = \exp\{-\eta(Y_b' - \eta)/\sigma\}$ $(b = 1, \ldots, B)$. and η is chosen so to maximize in some sense $var(\hat{P}_0)/var(\hat{P}_\eta)$ (called EEF_η by the authors of the work). The optimal value of η, call it η_P, may be approximated by:

$$\eta_P = (|\kappa_p| + \tfrac{2}{3} - \tfrac{1}{3}\sqrt{|\kappa_P|})sgn(\tfrac{1}{2} - P). \tag{4.23}$$

where $\kappa_P = \Phi^{-1}(1-P)$. It is quite robust to minor changes in η.

- A second level sampling involves the estimation of $u(y)$. Proceeding in an analogous way to in the preceding step we obtain:

$$\hat{u}(y) = \begin{cases} \exp(\frac{1}{2}\beta^2)M^{-1}\sum_{m=1}^{M} I_{(-\infty,\mu]}(Z_m')f_m & (y \geq \mu), \\ 1 - \exp(\frac{1}{2}\beta^2)M^{-1}\sum_{m=1}^{M} I_{(\mu,\infty)}(Z_m')f_m & (y < \mu), \end{cases} \tag{4.24}$$

where $f_m = \exp\{-\beta(Z_m' - y)/\sigma\}$ $(m = 1, \ldots, M)$, the Z_m' are independent $N(y+\beta\sigma, \sigma^2)$, and β is also chosen optimally. Writing $r(y) = (\mu - y)/\sigma$ the optimal value for β is approximately:

$$\beta(y) = (|r(y)| + \tfrac{2}{3} - \tfrac{1}{3}\sqrt{|r(y)|})sgn\{r(y)\}. \tag{4.25}$$

Now setting $y = Y_b'$ in (4.24) and substituting $\hat{U}_b' = \hat{u}(Y_b')$ for $u(Y_b')$ in (4.21) we obtain the estimate:

$$\hat{P}_{\eta,\beta} = \begin{cases} \exp(\frac{1}{2}\eta^2)B^{-1}\sum_{b=1}^{B} I_{[0,Q]}(U_b')e_b & (P < \frac{1}{2}), \\ 1 - \exp(\frac{1}{2}\eta^2)B^{-1}\sum_{b=1}^{B} I_{(Q,1]}(U_b')e_b & (P \geq \frac{1}{2}) \end{cases} \tag{4.26}$$

where E_b is as before.

There is a problem in that, to choose optimal values of β and η, we must know P and $u(y)$. However, due to their robustness, the values obtained with the normal approximation continue to be right even when we don't know P and $u(y)$.

Two level importance sampling for the nested bootstrap In order to apply the foregoing to the nested bootstrap we need T^* and T^{**} to be approximately normal (so that they can play the role of Y_B and Z_M in (4.21) and (4.24)). In the work by Johns (1988) [31] it is established that approximate normality

for T^* may be achieved by sampling from an ***exponentially tilted*** distribution, i.e. if we sample $(\mathbf{X}^*)$ from a multinomial distribution $\hat{F}_\eta$ with probabilities

$$Pr(X^* = x_i) \propto \exp\{\eta \hat{L}_i/(n\hat{\sigma})\} \quad (i = 1, \ldots, n) \tag{4.27}$$

then T^* is approximately $N(T + \eta\hat{\sigma}, \hat{\sigma}^2)$, with $\hat{\sigma}^2 = n^{-2}\sum\{L(x_i, \hat{F})\}^2$.

A parallel is feasible for T^{**} so that, if we sample X^{**} from $\hat{F}_\beta$ a multinomial with probabilities:

$$Pr(X^{**} = x_i^* | \hat{F}^*) \propto \exp\{\beta \hat{L}_i^*/(n\hat{\sigma}^*)\} \quad (i = 1, \ldots, n) \tag{4.28}$$

with $\hat{L}_i^* = L(x_i^*, \hat{F})$, then T^{**} is approximately $N(T^* + \beta\hat{\sigma}, \hat{\sigma}^{*2})$. The optimal choice of β depending on T^{**} in a similar way as in (4.25):

$$\beta(T^*) = (|r(T^*)| + \frac{2}{3} - \frac{1}{3}\sqrt{|r(T^*)|})sgn\{r(T^*)\}. \tag{4.29}$$

where $r(T^*) = (T - T^*)/\hat{\sigma}^*$. Similarly the analogue to $\exp(\frac{1}{2}\eta^2)e_b$ in (4.26) will be:

$$LR_i^* = \prod_{j=1}^{n} d\hat{F}_\eta(x_{ij}^*)/d\hat{F}_0(x_{ij}^*). \tag{4.30}$$

With all the analogues completed, the bootstrap algorithm is as follows:

1. Calculate T, $\hat{L}_i = L(x_i, \hat{F})$, $i = 1, \ldots, n$ and $\hat{\sigma}^2 = n^{-2}\sum \hat{L}_i^2$
2. Calculate the optimal value of η, following (4.3) and the tilted probabilities

$$p_i = \hat{p}(x_i) \propto \exp\{\eta \hat{L}_i/(n\sigma)\}, \quad (i = 1, \ldots, n) \tag{4.31}$$

3. Repeat for b=1,...,B
4. Sample $x_{b1}, \ldots, x_{bn}$ randomly from $p_1, \ldots, p_n$. Then calculate T_b^* and $LR_i^* = \prod_{j=1}^{n}\{n\hat{p}(x_{bj}^*)\}$
5. Calculate $\hat{L}_{bk}^* = \hat{L}(x_{bk}^*, \hat{F}^*)$, $(k = 1, \ldots n)$ $\hat{\sigma}_b^{*2} = n^{-2}\sum_{k=1}^{n} \hat{L}_{bk}^{*2}$ $r_b = (T - T^*)/\hat{\sigma}_i^*$, optimal values of β_b following (4.29) and the probabilities:

$$p_{bk}^* = \hat{p}^*(x_{bk}^*) \propto \exp\{\beta_b \hat{L}_{bk}^*/(n\hat{\sigma}_i^*)\} \tag{4.32}$$

6. Repeat for j=1,..., M:

   Sample $x^{**}_{bj,1}, \ldots, x^{**}_{bj,n}$ randomly from $(p^*_{b1}, \ldots, p^*_{bn})$, calculate T^{**}_{bj}, and $\prod_{k=1}^{n}\{n\hat{p}^*(x^{**}_{bj,k})\}$

7. Calculate

$$\hat{U}^*_i = \begin{cases} M^{-1}\sum_{j=1}^{M} I_{(-\infty,T]}(T^{**}_{bj})/LR^{**}_{bj} & (T^*_i \geq T) \\ 1 - M^{-1}\sum_{j=1}^{M} I_{(T,\infty)}(T^{**}_{bj})/LR^{**}_{bj} & (T^*_b < T) \end{cases} \tag{4.33}$$

8. Calculate

$$\hat{P} = \begin{cases} B^{-1}\sum_{b=1}^{B} I_{(0,Q]}(\hat{U}^*_b)/LR^*_b & (Q^*_b < \tfrac{1}{2}), \\ 1 - B^{-1}\sum_{b=1}^{B} I_{(Q,1]}(\hat{U}^*_i)/LR^*_i & (Q^*_b \geq \tfrac{1}{2}) \end{cases} \tag{4.34}$$

Open Problems

- How can this approach be applied when T is not approximately normal?.
- Can we obtain theoretical approximations playing the same role as the double bootstrap iteration?
- To what extent could theoretical approximations achieve the same efficiency as the second bootstrap?.
- How is M to be optimized in the second level resampling?.
- How would these techniques apply to more complicated situations?.
- An improved second-level sampling is introduced in the paper as a kind of "adaptive correction", but it is not analyzed in much depth.
- As mentioned in other places more experimental evidence is required to draw reliable conclusions. It seems that a wide simulation study would be of general interest.

5. CONTROL FUNCTION ESTIMATES FOR THE BOOTSTRAP (THERNEAU (1982))

Introduction. This method arises from the following identity:

$$E\{f(X)\} = E\{f(X) - h(X)\} + E\{h(X)\} \tag{5.1}$$

and consists, when we try to estimate the expectation of f, in taking another function h, whose expectation is a known quantity $E\{h(X)\} = \mu_h$, so that $E\{f(X)\}$ may be estimated by the following *control function estimator*:

$$\hat{E}_{CONT.} \equiv B^{-1} \sum_{b=1}^{B} [f(x_b) - h(x_b))] + \mu_h \tag{5.2}$$

which has variance:

$$var(\hat{E}_{CONT.}) = B^{-2} var(f - h). \tag{5.3}$$

The *appropriate choice* of h will make the variance of $(f - h)$,

$$var(f - h) = var(f) + var(h) - 2cov(f, h) \tag{5.4}$$

much smaller than $var(f)$. A reasonable election of h will take it close to f so that $(f - h)$ is small, and correlated, so that $cov(f, h)$ is positive and as large as possible.

Therneau (1983) [34] in an unpublished Stanford Ph.D. thesis applies this method to the bootstrap and discusses a wide set of control functions. For brevity we shall only consider the bias and variance estimates.

If T is the statistic whose expectation we are trying to estimate, the bootstrap version of this method will consist in finding a function g whose expectation under the bootstrap distribution is known, so that we may write:

$$E_{\hat{F}} T(\mathbf{P}) = E_{\hat{F}}\{(T(\mathbf{P}) - g(\mathbf{P})\} + E_{\hat{F}} g(\mathbf{P}) \tag{5.5}$$

$$= E_{\hat{F}}(T(\mathbf{P}) - g(\mathbf{P})\} + \mu_g \tag{5.6}$$

Bias Estimation. Two natural choices for $g(\mathbf{P})$, when we want to estimate the bias seem to be the quadratic functional(see Efron (1982) [7]):

$$(\mathbf{P}) \equiv c + \mathbf{PU}' + \mathbf{PVP}'. \tag{5.7}$$

or simply the linear functional:

(5.8) $$lin(\mathbf{P}) = c + \mathbf{PU}'$$

These terms are closely related to the first two and first term respectively of the von Mises expansion (2.20) of T so that the requirement of being close to this expansion is in principle fulfilled.

Some simulation experiments prove that choosing g to be the linear functional of $\mathbf{P}$ allows the number of resamples needed to achieve a given variance to be reduced by a factor of 5 to 10 with respect to the ordinary bootstrap. Though the use of quadratic functionals would probably achieve higher reductions, they are not so recommendable as far as the number of coefficients to determine is also much bigger ($n(n+1)/2$ vs n for the case of linear functionals), so that linear functionals seem to be the right compromise.

Bootstrap Estimate of Variance. Given a linear approximation, $lin(\mathbf{P})$ to $T(\mathbf{P})$ as in the preceeding section, the control function suggested by Therneau to estimate the variance of T, $E_{\hat{F}} = \{T(\mathbf{P}) - \overline{T}\}^2$ is

(5.9) $$g_{var}(\mathbf{P}) = (lin(\mathbf{P}) - \mu_{lin})^2.$$

As before, this expression closely matches $\{T(\mathbf{P}) - \overline{T}\}$ and has known expectation. Moreover it works better than the more straightforward approximation $lin(\mathbf{P})^2$ to $T(\mathbf{P})^2$, seemingly due to the fact that the curvature of the functions increases when they are squared. The author leaves open the possibility of correcting this square so that it might work better.

6. ANTITHETIC RESAMPLING (P. HALL (1989))

6.1 Introduction

Hall (1989)[18] has introduced an antithetic resampling method for the bootstrap based on the well known VRT of *antithetic variates* introduced by Hammersley and Morton (1956)[22]. (See the classical book of Hammersley and Handscomb (1964)[21] for a description.) The basic idea of the method is as follows. Let $\hat{\theta}$ be a statistic. To estimate $E(\hat{\theta})$ by simulation we calculate $\hat{\theta}$ on n independent runs and then use

(6.1) $$\overline{\hat{\theta}} = n^{-1} \sum_{i=1}^{n} \hat{\theta}_i$$

to estimate $E(\hat{\theta})$. The variance of this estimate is:

(6.2) $$var(\overline{\hat{\theta}}) = n^{-1} var(\hat{\theta}).$$

If we can generate our simulated samples so that the estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$ are negatively correlated by pairs ($\rho(\hat{\theta}_{2i-1}, \hat{\theta}_{2i}) < 0$) the pairs mutually being independent among them we may define:

(6.3) $$\hat{\theta}_i^{\dagger} = \frac{\hat{\theta}_{2i-1} + \hat{\theta}_{2i}}{2}$$

that will give the same estimate as 6.1

(6.4) $$\overline{\hat{\theta}^{\dagger}} = (n/2)^{-1} \sum_{i=1}^{n/2} \hat{\theta}_i^{\dagger} = \overline{\hat{\theta}}$$

If the negative correlation is adequately induced this estimate may have a lower variance than 6.1 by the relation:

(6.5) $$var(\hat{\theta}_i^{\dagger}) = \frac{1}{4} var(\hat{\theta}_{2i-1}) + \frac{1}{4} var(\hat{\theta}_{2i}) + \underbrace{\frac{1}{2} cov(\hat{\theta}_{2i-1}, \hat{\theta}_{2i})}_{<0}$$

As in other VRT there is a subtle point that lies in *how to induce* the negative correlation between $\hat{\theta}_{2i-1}$ and $\hat{\theta}_{2i}$.

6.2 Antithetic resampling

Suppose we are interested in the bootstrap estimation of the expected value, Ψ of an order invariant statistic $\theta(X_1, \ldots, X_n)$. The bootstrap estimate of Ψ is:

(6.6) $$\hat{\Psi} = E\{\theta(X_1^*, X_2^*, \ldots, X_n^*)|\hat{F}\}$$

where "$|\hat{F}$" denotes the ordinary bootstrap sampling. As usual this expectation may very rarely be evaluated analytically and we approximate it by:

(6.7) $$\hat{\Psi}^* = B^{-1} \sum_{b=1}^{B} [\theta(X_1^*, X_2^*, \ldots, X_n^*)]^b.$$

To introduce antithetic resampling we use the following equivalent formula for 6.7:

(6.8) $$\hat{\Psi}^* = B^{-1} \sum_{b=1}^{B} [\theta(X_{I(b,1)}, \ldots, X_{I(b,n)})]^b.$$

where $I(b,i)$ $(1 \leq b \leq B, 1 \leq i \leq n)$ are independently and uniformly distributed on $1, 2, \ldots, n$.

The idea of antithetic resampling is to choose appropriately an *antithetic permutation* π of the integers $1, \ldots, n$ so that if we define

(6.9) $$\hat{\Psi}^{**} = B^{-1} \sum_{b=1}^{B} [\theta(X_{\pi\{I(b,1)\}}, \ldots, X_{\pi\{I(b,n)\}})]^b$$

then $\hat{\Psi}^*$ and $\hat{\Psi}^{**}$ should be negatively correlated and, as a consequence, the estimate:

(6.10) $$\hat{\Psi}^\dagger = \frac{1}{2}(\hat{\Psi}^* + \hat{\Psi}^{**})$$

should enjoy the traditional advantages of antithetic sampling.

Whe are going to see that this permutation, call it π, is of the form: $\pi(i) = n - 1 + 1$.

Let us first suppose, for simplicity, that the estimate $\theta(X_1, X_2, \ldots, X_n)$ is a function of the mean, (and henceforth it admits a Taylor series expansion),

$\theta(X_1, X_2, \ldots, X_n) = \theta(\overline{X})$. We assume X_i d-variate and θ smooth, and define $\theta_j(x) = (\partial/\partial x^j)\theta(x)$ and

$$Y_i = \sum_{j=1}^{d} \theta_j(\overline{X})(X_i - \overline{X})^j. \tag{6.11}$$

It may be easily seen, by a 1st order Taylor series expansion that

$$var(\hat{\Psi}^*|\hat{F}) = var(\hat{\Psi}^{**}|\hat{F}) \simeq (Bn)^{-1}n^{-1}\sum_{i=1}^{n} Y_i^2 \tag{6.12}$$

and

$$cov(\hat{\Psi}^*, \hat{\Psi}^{**}|\hat{F}) \simeq (Bn)^{-1}n^{-1}\sum_{i=1}^{n} Y_i Y_{\pi(i)} \tag{6.13}$$

and therefore

(6.14)

$$var(\hat{\Psi}^\dagger|\hat{F}) = (2Bn)^{-1}\left(n^{-1}\sum_{i=1}^{n} Y_i^2 + n^{-1}\sum_{i=1}^{n} Y_i Y_{\pi(i)}\right) + O(B^{-1}n^{-2})$$

It is known that, given a set of ordered real numbers $a_1 \leq a_2 \leq \ldots \leq a_n$, if $b_1, \ldots, b_n$ is a rearrangement of these numbers then $\sum a_i b_i$ is minimized by taking the smallest a_i with the largest b_i, the second smallest with the second largest and so on. Therefore, under the notational convention that $Y_1 \leq Y_2 \ldots \leq Y_n$ the series $\sum Y_i Y_{\pi(i)}$ is minimized over all permutations π by taking $\pi(i) = n - 1 + 1$.

In the more general case where θ is not a function of the mean, other generalizations may be made for the former idea to be valid. One possibility is to use the regression version of antithetic sampling replacing the X_j's by estimated residuals in 6.11. An alternative is considering the von Mises expansion of θ instead of the Taylor expansion, i.e. if we put: $\hat{\theta} = \theta(X_1, X_2, \ldots, X_n) = A(\hat{F})$ and let IC represent the influence function of A (see appendix I) we can write:

$$A(\hat{F}) = A(F) + n^{-1}\sum_{i=1}^{n} IC(X_i) + o_p(n^{-\frac{1}{2}}) \tag{6.15}$$

and replace definition 6.11 by

$$Y_i = IC(X_i). \tag{6.16}$$

If we do so, then the ensuing definition of the permutation π continues to be valid.

6.3 Estimation of distribution functions and quantiles

The former ideas apply not only to bootstrap expectation estimation but also to other situations that are typically analyzed by bootstrap methods. Suppose we want to estimate the distribution function of $\hat{\theta}$ or of a studentized version of $\hat{\theta}$ such as:

$$T = n^{\frac{1}{2}}\{\theta(\overline{X}) - \theta(\mu)\}/\hat{\sigma} \tag{6.17}$$

where $\mu = E(X)$.
As before, we define, the bth resample:

$$\{X_i^*,\ 1 \leq i \leq n\}^b = \{X^*_{I(b,i)},\ 1 \leq i \leq n\} \tag{6.18}$$

and the antithetic resample

$$\{X_i^{**},\ 1 \leq i \leq n\}^b = \{X^*_{\pi\{I(b,i)\}},\ 1 \leq i \leq n\} \tag{6.19}$$

The bootstrap estimate of G is

$$\hat{G}(x) = \mathrm{pr}[T^* \leq x|\hat{F}] \tag{6.20}$$

which we usually approximate by G^*, the empirical c.d.f. of the sample. If we now define:

$$\hat{G}^\dagger = \frac{1}{2}(\hat{G}^* + \hat{G}^{**}) \tag{6.21}$$

it may be proved that $\hat{G}^*$ and $\hat{G}^{**}$ are negatively correlated and that the antithetic permutation π defined before gives the greatest degree of negative correlation and so maximizes the performance of $\hat{G}^\dagger$.

It is possible from this idea to define an antithetic quantile estimate $\nu^\dagger_{p,B}$ and Hall (1989) indicates that under certain regularity conditions the asymptotic relative efficiency of $\nu^\dagger_{p,B}$ with respect to $\nu^*_{p,B}$ is the same as that of $\hat{G}_B(z_p) - G(z_p)$ relative to $\hat{G}^*_B(z_p) - G(z_p)$.

6.4 Antithetic resampling and other "efficient bootstrap methods"

Hall (1989) [18] compares the performance of antithetic resampling to importance resampling in problems of distribution function estimation and quantile estimation as introduced by Johns (1988) [31].

He develops importance resampling schemes for distribution function estimation and shows that antithetic resampling often gives greater efficiency towards the center of the distribution but not at the tails.

It must be kept in mind however that importance resampling may be much harder to carry out than uniform or even antithetic resampling, so that the theoretical increase in efficiency may be partially lost by the increase in the simulation costs. Hall (1989b) [18] gives some expressions to quantify this increase in both, antithetic and importance resampling.

The combination of antithetic resampling with other "efficient bootstrap techniques" is not a good strategy. Not only does this not improve the efficiency but it even decreases it (this is, in fact common in variance reduction). For instance Hall (1989b) [18] proves that combining antithetic resampling with Efron's centering method means that the covariance between the antithetic pairs cannot be made negative (this implies that we shouldn't either combine antithetic resampling with the balanced or with the linear bootstrap, asymptotically equivalent to the centered bootstrap). He also proves that the combination of antithetic and importance resampling gives a poorer performance than either of the methods alone.

7. MEASURING THE ACCURACY OF BOOTSTRAP ESTIMATES

7.1 Jackknife-After-Bootstrap Standard Errors and Influence Functions. (Efron (1990))

Introduction. One of the main uses of bootstrap is to obtain measures of accuracy for a given estimate, such as the bias or the standard error. If one wants to

know how accurate those measures of accuracy are, a straightforward approach consists in bootstrapping them to obtain for instance the bootstrap estimate of their standard error (i.e. we have to bootstrap the bootstrap estimates). This procedure may not only imply an additional source of variability but it is also highly (CPU) time consuming, especially in the case of complicated estimates. A different approach to efficient bootstrapping is proposed here, which essentially consists in a way to measure the accuracy of the bootstrap estimates (of accuracy) with no additional resampling.

This approach is based on the use of *jackknife influence functions and estimates of standard error*: If we indicate with $\mathbf{X}_{(i)}$ the data set remaining after the deletion of the ith point:

$$\mathbf{X}_{(i)} = (X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n), \tag{7.1}$$

and let $S_{(i)} = S(\mathbf{X}_{(i)})$ be the corresponding deleted-point value of the statistic of interest, $S(\mathbf{X})$, then we may define the *jackknife influence function* for S as:

$$u_i\{S\} = (n-1)(S_{(\,)} - S_{(i)}), \quad (S_{(\cdot)} \equiv \sum_{i=1}^{n} S_{(i)}/n). \tag{7.2}$$

and the *relative* jackknife influence function as:

$$u_i^{\dagger}\{S\} = u_i\{S\} \Big/ [\sum_j u_j\{S\}^2/(n-1)]^{\frac{1}{2}}. \tag{7.3}$$

Tukey's *jackknife estimate for standard error* of S($\mathbf{X}$) is:

$$se_{jack}\{S\} \equiv \left[\sum_{i=1}^{n} u_i\{S\}^2/(n(n-1))\right]^{\frac{1}{2}} \tag{7.4}$$

Let us call $S(\mathbf{X})$ a *primary statistic*. Some *bootstrap statistics*, to measure its accuracy are:

- The bootstrap estimate of standard error, $\overline{se}_B$, i.e. the square root of the bootstrap estimate of the variance defined in (1.9).

- The bootstrap estimate of bias, $\overline{bias}_B\{S\}$ defined in (1.8)
- The length of a confidence interval,

(7.5) $$L \equiv S^{*(.95)} - S^{*(.05)},$$

where $S^{*(\alpha)}$ is the bootstrap estimate of the αth percentile as in (1.10)

- The shape of a confidence interval:

(7.6) $$Sh \equiv \log\{(S^{*(.95)} - S^{*(.50)})/(S^{*(.50)} - S^{*(.05)})$$

- and the 95th percentile for the *bootstrap-t* (see Efron (1982) [7] sec. 10.10).

We shall call them *secondary statistics*. If $T(\mathbf{X}, F)$ is some random variable of interest –such as the bias or standard error of a statistic S($\mathbf{X}$)– depending on the sample and on the underlying distribution F, and we write $\phi[T(\mathbf{X}^*, \hat{F})]$ to indicate some measure of accuracy for T, then the general form of a bootstrap statistic or *secondary statistic* will be:

(7.7) $$\hat{\gamma}(\mathbf{X}) = \phi[T(\mathbf{X}^*, \hat{F})]$$

The efficient measures of accuracy developed in the following are based on:

- Jackknife influence functions and standard errors.
- Delta method influence functions.
- Calculation of internal errors (which also allows us to measure the accuracy of the two former measures)
- A parametric bootsrap approach

It is worth noting that all these measurements of accuracy have been previously developed as *first-level error measures*, though they are applied here at a second (or a third) level. Efron (1982) [7], chapter 6, reviews them, as well as their inter-relations.

Tukey's jackknife for bootstrap statistics. To obtain the jackknife estimate of standard error of $\hat{\gamma}(\mathbf{X})$ we need to compute the jackknife influence function $u_{(i)}\{\hat{\gamma}\}$, and to do this we need to calculate the deleted-point values $\hat{\gamma}_{(i)} \equiv \hat{\gamma}(\mathbf{X}_{(i)})$. Let $\hat{F}$ denote the ordinary empirical c.d.f., putting probability $1/n$ on each point of the sample and $\hat{F}_{(i)}$ the *deleted-point empirical c.d.f.*,

(7.8) $\hat{F}_{(i)}$: probability $\frac{1}{n-1}$ on x_j, $\quad j = 1, 2, \ldots, i-1, i+1, \ldots, n.$

The following result (Lemma 1 in Efron (1990) [9] enables us to calculate $\hat{\gamma}_{(i)}$:

An i.i.d. sample of size n from $\hat{F}_{(i)}$,

(7.9) $(X_1, X_2, \ldots, X_n) \overset{i.i.d.}{\sim} F_{(i)}$

has the same distribution as a bootstrap sample from $\hat{F}$ in which *none of the* X_j^* *values equals* X_i.

Let $\mathbf{P}$ denote the resampling vector as defined in (2.2). Then, from (7.9) it follows that

$$\hat{\gamma}_{(i)} = \phi[T(\mathbf{X}^*, \hat{F})|P_i = 0], \tag{7.10}$$

where $[T(\mathbf{X}^*, \hat{F})|P_i = 0]$ indicates the conditional distribution of $T(\mathbf{X}, \hat{F})$ given that $P_i = 0$.

Let

$$\tilde{\gamma}(\mathbf{X}) = \phi[T(\mathbf{X}^{*b}, \hat{F}),\ b = 1, 2, \ldots, B] \tag{7.11}$$

indicate the Monte Carlo approach –i.e. with a finite number of replications B– to $\hat{\gamma}(\mathbf{X})$ (in the same way we approximated (1.4), (1.5) and (1.6) by (1.8), (1.9) and (1.10). Equivalently, the Monte Carlo approach to (7.10) is:

$$\tilde{\gamma}_{(i)} = \phi[T(\mathbf{X}^{*b}, \hat{F}_{(i)}),\ b,\ \text{s.t. } P_i^b = 0] \tag{7.12}$$

Once we have obtained $\tilde{\gamma}_{(i)}$ for any bootstrap statistic we may calculate the influence functions, which will enable us to see which is the influence of every point in the variability of the statistic, and to compute the jackknife estimate of standard error.

The Delta Method for Bootstrap Statistics. The delta method, another well-known device used to derive first-level estimates of standard error for a statistic, may under certain circumstances give more accurate results than the jackknife estimates when applied to a bootstrap statistic. It will only apply, however, to *functional statistics.*

As in the former case we first define a variant of the empirical c.d.f. (that is to say a different estimate of F to $\hat{F}$), and a *delta method influence function* from where we obtain a *delta method estimate of the standard error.*

Let

$$(7.13) \qquad \hat{F}_{\varepsilon,i} \ : \ \text{probability} \begin{cases} \dfrac{1-\varepsilon}{n} + \varepsilon & \text{on } x_i \\ \dfrac{1-\varepsilon}{n} & \text{on } x_j, \ j \neq i. \end{cases}$$

The *delta method influence function*, also called empirical influence function or infinitessimal jackknife for S is the derivative:

$$(7.14) \qquad U_i\{S\} = \left.\frac{\partial S\{\hat{F}_{\varepsilon,i}\}}{\partial \varepsilon}\right|_{\varepsilon=0} = \lim_{\varepsilon \to 0} \frac{S(\hat{F}_{\varepsilon,i}) - S(\hat{F})}{\varepsilon}$$

and may be approximated by putting a small value of ε in the expression. Finally the delta-method estimate of standard error is:

$$(7.15) \qquad se_{delta}\{S\} = [\sum_{i=1}^{n} U_i\{S\}^2/n^2]^{\frac{1}{2}}$$

where the n^2 is arbitrarily put, in the denominator instead of $n(n-1)$ as in (7.4) in order to agree with the usual nonparametric delta method estimate of standard error.

For convenience of notation the calculation of the influence function is first restricted to bootstrap statistics of the *expectation form*:

$$(7.16) \qquad \hat{\gamma}(\mathbf{X}) = \gamma(\hat{F}) = E_{\hat{F}}\{r(T(\mathbf{X}^*, \hat{F}))\}$$

where $r(T)$ is some differentiable function of T and $E_{\hat{F}}$ indicates the ordinary bootstrap expectation. The delta method influence function may be calculated through theorem 1 of Efron (1990) [9], which establishes that it has the form:

$$(7.17) \qquad U_i\{\gamma\} = n^2 cov_{\hat{F}}\{P_i, r^*\} + E_{\hat{F}}\{(r'^*)U_i\{T(\mathbf{X}^*, \hat{F})\}\}$$

where the subindex $\hat{F}$ indicates the calculations done under bootstrap sampling, and the empirical c.d.f. $\hat{F}$ is related to (7.13) by lemma 2 of Efron (1990) [9], which we omit.

Once we have the influence functions the calculation of the delta method estimate of standard error is straigthforward.

Efron (1990) [9] applies this formula to obtain the delta-method estimates of bootstrap statistics mentioned in the introduction. He also derives an alternative estimate of (7.17) based on the ANOVA decomposition of a bootstrap statistic (2.7), i.e. by means of the ***bootstrap influence function.*** This formula is more efficient than (7.17) when applied to measure the accuracy of the bootstrap bias estimate, but not much more when applied to the standard error estimate, which may be explained by the fact that being based on a linearization, its advantage comes from the amount of linearity that the statistic has. A comparison between the jackknife and delta method estimates is also given. For the bias estimate the (improved) delta method estimate works better, whereas for the jackknife estimate there are no remarkable differences. In fact in the case of linear functionals there is no difference between $U_i\{\hat{\gamma}\}$ and $u_i\{\hat{\gamma}\}$, and they are equal to the bootstrap influence functions.

Internal errors. In the search for cheap methods to measure the accuracy of bootstrap statistics, a further step comes from realizing that we can not really know $U_i\{\hat{\gamma}\}$ and $u_i\{\hat{\gamma}\}$, but only their Monte Carlo estimates, $\tilde{U}_i\{\hat{\gamma}\}$ and $\tilde{u}_i\{\hat{\gamma}\}$. Efron (1990) [9] calls the difference between the ideal and the Monte Carlo estimates the *internal errors.*

Roughly speaking, he proceeds, with assistance of a lemma that he introduces, and some matrix algebra, to apply Tukey's jackknife covariance formula (see Efron(1982) [7]), to obtain the covariance matrix of $U_i\{\hat{\gamma}\}$ and $u_i\{\hat{\gamma}\}$,

$$\text{(7.18)} \qquad \text{cov}_{\text{intern}} \begin{pmatrix} \tilde{\mathbf{u}}\{\hat{\gamma}\} \\ \tilde{\mathbf{U}}\{\hat{\gamma}\} \end{pmatrix},$$

where the square roots of the diagonal elements give the standard errors of $\tilde{\mathbf{u}}\{\hat{\gamma}\}$ and $\tilde{\mathbf{U}}\{\hat{\gamma}\}$ respectively. Using these elements, the following result is established: (7.19)

$$E_{\hat{F}}\{\tilde{se}_{jack}\{\hat{\gamma}\}^2\} = se_{jack}\{\hat{\gamma}\}^2 + \text{trace}(\text{cov}_{\text{intern}}\{\tilde{\mathbf{u}}\})/(n \cdot (n-1)) + O(1/B)$$

from which a corrected estimate for $\tilde{se}_{jack}$ will be:

$$\text{(7.20)} \qquad [\tilde{se}_{jack} - \text{trace}(\text{cov}_{\text{intern}}\{\tilde{\mathbf{u}}\})/(n \cdot (n-1))]^{\frac{1}{2}}.$$

The corresponding delta estimates are directly obtained by making the appropriate substitutions and changing $n(n-1)$ for n^2 in (7.19).

When the sample size, n, increases so do the internal errors, as a result from the fact that the internal coefficient of variation is $O(n/\sqrt{B})$ (see remark H of Efron (1990) [9]).

Parametric Bootstrap Sometimes the bootstrap is applied parametrically. In general this will mean resampling from $F(\mathbf{X}, \hat{\theta})$ instead of from $\hat{F}(\mathbf{X}, \theta)$, i.e. we first estimate the parameter from the sample, and then generate samples from F depending on the estimated $\hat{\theta}$.

Let $\hat{\gamma}(\mathbf{X})$ be a bootstrap statistic,which we estimate, as in (7.11) using $F(\mathbf{X}, \hat{\theta})$ instead of $\hat{F}$:

$$\tilde{\gamma}(\mathbf{X}) = \phi[T(\mathbf{X}^{*b}, \hat{\theta}), \quad b = 1, \ldots, B] \tag{7.21}$$

To apply the methods we have discussed up to now we need to estimate $\hat{\gamma}_{(i)}$, but now formula (7.12) no longer makes sense. It is replaced by an *importance sampling estimate*. The idea consists of first obtaining a deleted point estimate $\hat{\theta}_{(i)}$ of θ based on the deleted point data-set (7.1), and then resampling from the original sample but assigning a probability to each point x_i

$$R_i(\mathbf{X}^*) \equiv \frac{f_{\hat{\theta}_{(i)}}(\mathbf{X}^*)}{f_{\hat{\theta}}(\mathbf{X}^*)}. \tag{7.22}$$

Lemma 3 in Efron (1990) [9] establishes the equivalence between ordinary bootstrap resampling with those probabilities and deleted-point sampling (equivalence here meaning equality of the expectation of $r(T(\mathbf{X}^*,))$, as in (7.16) under both sampling schemes), and suggests the estimate:

$$\tilde{\gamma}_{(i)} = \phi[T(\mathbf{X}^{*b}, \hat{\theta}_{(i)}), \text{ with probabilities } R_i(\mathbf{X}^{*b})/B] \tag{7.23}$$

where probability $R_i(\mathbf{X}^{*b})/B$ is put on each $T(\mathbf{X}^{*b}, \hat{\theta}_{(i)})$.

An example with the parametric bootstrap-t percentiles shows that they have smaller standard errors than the non-parametric ones.

Efron (1990) [9] develops Delta method estimates and influence functions for the parametric case.

References

[1] **Beran, R.** (1987). Pre-pivoting to reduce the error level of confidence sets. *Biometrika* **74** (3), 457-468.

[2] **Bratley, P. Fox, B.L., Schrage, L.E.** (1987). *A Guide to Simulation.* 2nd edition. Springer-Verlag: New York.

[3] **Davison, A.C.** (1988) Discussion of the papers by Hinkley and DiCiccio and Romano.*J. of the R.Stat.Soc. B.***50** pp. 356-357

[4] **Davison, A.C., Hinkley, D.V., Schechtman** (1986). Efficient bootstrap simulation. *Biometrika* **23**

[5] **Di Ciccio, T.J. and Romano, J.P.** (1988). A review of bootstrap confidence intervals. *J.R. Statist. Soc.* B. **50**, 338-354.

[6] **Efron B.** (1979). Bootstrap methods. Another look at the jackknife. *Annals of Statist.* **7**, 1-26.

[7] **Efron B.** (1982). *The Jackknife, the Bootstrap and other Resampling Plans.* Reg. Conf. Ser. in Appl. Math., No 38, Philadelphia: SIAM.

[8] **Efron B.** (1988). More efficient bootstrap computations. Technical report num 294. Stanford University.

[9] **Efron B.** (1990). Jackknife-after-bootstrap standard errors and influence functions. Technical report num 339. Stanford University.

[10] **Efron, B. Tibishirani, R.** (1986) Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy.*Statistical Science* **1** 54-77

[11] **Efron B., Stein C.** (1981) The jackknife estimate of variance. *Ann. Stat.***9** 586-596

[12] **Fernholz, L. T.** (1983). *von Mises Calculus for Statistical Functionals,* Lecture Notes in Statistics, 19. Springer-Verlag. New York.

[13] **Gleason, J.R.** (1988). Algorithms for Balanced Bootstrap Simulations. *American Statistician* **42**, 4, 263-266.

[14] **Graham, R.L., Hinkley, D.V., John P.W.M., Shi, S.** (1987). Balanced design of bootstrap simulation. Technical report num 48, UNiversity of Texas, Austin.

[15] **Haldane, J.B.S.** (1940). The mean and the variance of χ^2, when used as a test of homogeneity, when expectations are small. *BIometrika*, **31**, 346-55.

[16] **Hall, P.** (1988) Theoretical comparison of Bootstrap confidence intervals. *Annals of Statistics* **16** 927-953.

[17] **Hall, P.** (1989). On efficient bootstrap simulation. *Biometrika*, **76**, 3, 613-617.

[18] **Hall, P.** (1989). Antithetic resampling for the bootstrap. *Biometrika*, **76**, 4, 713-24.

[19] **Hall, P., Heyde, C.C.** (1980). *Martingale Limit theory and its Applications.* New York: Wiley.

[20] **Hall, P., Martin, M.** (1988) On bootstrap re-sampling and iteration. *Biometrika* **75**, 4, 661-671.

[21] **Hammersley, I. M., and Handscomb, D.C.** (1964). *Monte Carlo Methods*, Chapman and Hall, London.

[22] **Hammersley, J.M. and Mauldon, J.G.** (1956). General principles of antithetic variates. *Proc. Camb. Phil. Soc.* **52**, 476-81.

[23] **Hampel, F.R.** (1968). *Contributions to the theory of Robust Estimation.* Unpublished Ph. D. Thesis, University of California, Berkeley, September 1968.

[24] **Hampel, F. R.** (1974). The Influence Curve and its Role in Robust Estimation. *J.A.S.A.* **69, 346** 383-393.

[25] **Hampel, F., Ronchetti, E., Rouseeuw, P. and Stahel, W.** (1986). *Robust Statistics, The Approach Based on Influence Functions.* Wiley, New York.

[26] **Hesterberg, T.** (1988). Advances in importance sampling. Ph.D. thesis. Dept of statistics, Stanford University.

[27] **Hinkley, D.V.** and Schechtman, E. (1987). Conditional bootstrapping methods in the mean-shift model. *Biometrika* **74**

[28] **Hinkley, D.V. and Shi, S.** (1989). Importance sampling and the nested bootstrap. *Biometrika* **76** 435-446.

[29] **Hinkley, D.V., Wei, B.C.** (1984). Improvements on jackknife confidence limit methods. *Biometrika* **71**, 331-9.

[30] **Huber, P.J.** (1977), *Robust Statistical Procedures*, CBMS Regional Conferences in Applied Mathematics 27, SIAM, Philadelphia.

[31] **Johns, M.V.** (1988). Importance Sampling for bootstrap confidence intervals. *J.A.S.A.* **83**, 709-714.

[32] **Ogbonmwan, S.M., Wynn, H.P.** (1986). Contribution to discussion of the paper by C.F.J. Wu. *Annals of Statistics* **14**, 1340-1343.

[33] **Serfling, R. J.** (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons. New York.

[34] **Therneau, T.M.** (1983). Variance reduction techniques for the bootstrap. Ph.D. thesis, Department of Statistics, Stanford University.

[35] **Tukey, J.W., Brillinger, D.R. and Jones, L.V.** (1978). *The Management of Water Ressources 2.*. Weather Modification Advisory Board, Statistical Task Force, U.S. GPO, Washington.

[36] **von Mises, Richard** (1947). On the Asymptotic distribution of Differentiable Statistical Functions. *Annals of Mathematical Statistics*, **18**, 309-48.

APPENDIX

A. STATISTICAL FUNCTIONALS, VON MISES EXPANSIONS AND INFLUENCE CURVES

Several concepts such as Influence functions and von Mises expansions appear repeatedly, and in different situations, in the context of resampling methods. Influence functions are actually mainly known from robust Statistics (Hampel (1974)[24], Huber (1977)[30] and Hampel et al (1986)[25]) but the basic ideas come from the work of von Mises (1947)[36] who made a "differential approach" for deriving the *asymptotic distribution theory* of statistical functionals.

A.1 Statistical functionals

Let $(X_1, X_2, \ldots, X_n)$ be a sample from a population with distribution function (d.f.) F and let $T_n = T_n((X_1, X_2, \ldots, X_n))$ be a statistic. If T_n can be written as a functional of the empirical d.f. F_n, $T_n = T(F_n)$ *where* T *does not depend on* n then T will be called a **statistical functional**.

Examples:

1. For any function $h(x)$ let

(A.1) $$T_n((X_1, X_2, \ldots, X_n)) = n^{-1} \sum_{i=1}^{n} h(X_i).$$

Then for a general d.f. G the functional defined by:

(A.2) $$T(G) = \int h(x) dG(x)$$

satisfies $T_n((X_1, X_2, \ldots, X_n)) = T(F_n)$.

2. Let Ψ be a real valued function of two variables and let T_n be defined implicitily by

(A.3) $$\sum_{i=1}^{n} \Psi(X_i, T_n) = 0.$$

The corresponding functional is defined as a solution $T(G) = \theta$ of

(A.4) $$\int \Psi(x, \theta) dG(x) = 0.$$

Estimators of this form are called M-estimators.

Functionals of the form

(A.5) $$T(G) = \int h(x) dG(x)$$

are called **linear statistical functionals** or simply linear functionals.

An application of the central limit theorem shows that for a linear functional T,

$$\text{(A.6)} \qquad \sqrt{n}(T(F_n) - T(F)) \to N(0, \sigma^2) \text{ weakly}$$

provided that

$$\text{(A.7)} \qquad 0 < \int h^2(x) dG(x) - (\int h(x) dG(x))^2 = \sigma^2 < \infty.$$

The central idea behind the von Mises Method is to extend this asymptotic normality result to statistical functionals which are not linear by means of an approximation by linear functionals.

A.2 Von Mises expansions

Von Mises (1947)[36] proposed that a Taylor expansion could be used to approximate statistical functions by statistical functions of simpler form, and that this result could be applied to obtain results about its asymptotic distribution. The main result is established informally in the following theorem of von Mises that appears in Serfling (1980)[33], p. 212:

Theorem 0.1 *The type of asymptotic distribution of a differentiable statistical functional $T_n = T(F_n)$ depends upon which is the first nonvanishing term in the Taylor development of the functional $T(\cdot)$ at the distribution F of the observations. If it is the linear term, the limit distribution is normal (under the usual restrictions corresponding to the central limit theorem). In other cases, "higher" types of limit distributions result.*

Serfling (1980)[33] and Fernholz (1983)[12] provide schemes for the analysis of $T(F_n)$ in this framework in order to derive its asymptotic distribution.

Roughly it starts by considerating a Taylor expansion of $T(F_n) - T(F)$:

$$\text{(A.8)} \quad T(F_n) - T(F) = d_1 T(F; F_n - F) + \frac{1}{2!} d_2 T(F; F_n - F) + \cdots$$

Analysis of $T(F_n) - T(F)$ is to be carried out by reduction to

$$\text{(A.9)} \qquad \sum_{j=1}^{m} \frac{1}{j!} d_j T(F; F_n - F)$$

for an appropriate choice of m. The reduction step is performed by dealing with the remainder term

$$R_{mn} = T(F_n) - T(F) - V_{mn} \tag{A.10}$$

and the properties of $T(F_n) - T(F)$ then are obtained from an *m-linear* structure typically possessed by V_{mn}.

A.8 will be called the von Mises expansion of T at F.

The existence of this series expansion A.8 depends on differentiability properties of statistical functionals. To deal with them, a derivative for functionals is introduced as *the von Mises derivative* also called the *Gateaux differential of T at F in the direction of G.*

Let F and G be two points in the space $\mathcal{F}$ of all distribution functions. The "line segment" in $\mathcal{F}$ joinig F and G consists of the set of distribution functions $\{(1-\lambda)F+\lambda G,\ 0 \leq \lambda \leq 1\}$, also written as $\{F+\lambda(G-F),\ 0 \leq \lambda \leq 1\}$. Consider a functional T defined on $F + \lambda(G - F)$ for all sufficiently small λ. If the limit

$$d_1\, T(F; G - F) = \lim_{\lambda \to 0+} \frac{T(F + \lambda(G - F)) - T(F)}{\lambda} \tag{A.11}$$

exists, it is called the *Gateaux differential of T at F in the direction of G.* (Note that $d_1\,(F; G - F)$ is simply the ordinary right-hand derivative, at $\lambda = 0$ of the function $Q(\lambda) = T(F + \lambda(G - F))$ of the real variable λ).

A.3 Statistical Interpretations of the Derivative of a Statistical Functional: The Influence Curve.

In the case of a statistical functional having nonvanishing first derivative (implying asymptotic normality under mild restrictions), a variety of important features of the estimator, such as the *asymptotic variance parameter* and certain *stability properties* may be characterized in terms of this derivative. In typical cases the Gateaux derivative is linear: there exists a function $T_1[F; x]$ such that:

$$d_1\, T(F; G - F) = \int T_1[F; x]\, d[G(x) - F(x)]. \tag{A.12}$$

If we put

(A.13) $$h(F;x) = T_1[F;x] - \int T_1[F;x]dF(x)$$

the reduction methodology based on the analysis of the remainder term A.10 shows that the *error of estimation* in estimating $T(F)$ by $T(F_n)$ is given approximately by:

(A.14) $$\frac{1}{n}\sum_{i=1}^{n} h(F;X_i).$$

Thus $h(F;X_i)$ represents the approximate contribution or "influence," of the observation X_i towards the estimation error $T(F_n) - T(F)$. It was introduced by Hampel (1968, 1974)[23,24] who calls $h(F;x)$, $-\infty < x < +\infty$ the *influence curve, or influence function* of the estimator $T(F_n)$ for $T(F)$. Note that the curve may be defined directly by:

(A.15) $$IC(x;F,T) = \left.\frac{dT[F + \lambda(\delta_x - F)]}{d\lambda}\right|_{\lambda=0}, \quad -\infty < x < +\infty$$

(A.16) $$= \lim_{\lambda \to 0} \frac{T[(1-\lambda)F + \lambda\delta_x] - T[F]}{\lambda}$$

where δ_x means a point mass 1 at x.

Example Let the mean of F exist and be equal to μ. The influence curve of the sample mean, $T = \int x dF(x)$ is:

(A.17) $$IC(x;F,T) = \lim_{\lambda \to 0} \frac{(1-\lambda)\mu + \lambda x - \mu}{\lambda} = x - \mu$$

The expansion A.8 may now be rewritten as:

(A.18) $$T(F_n) = T(F) + \int IC(x;F,T)d(F_n - F)(x) + V_{mn}$$

(A.19) $$= T(F) + \int IC(x;F,C)dF_n(x) + V_{mn}$$

since $\int IC(x;F,T)dF(x) = 0$.

A.19 is the expression of the the von Mises expansion of T at F by means of influence functions.

The influence curve is closely related to the asymptotic variance of T (as may be easily derived from A.19). If $(X_1, X_2, \ldots, X_n) \overset{i.i.d.}{\sim} F$ then according to the Glivenko–Cantelli theorem F_n will tend to F. If we suppose that $T_n(X_1, X_2, \ldots X_n) = T_n(F_n) \simeq T(F_n)$ (i.e. that we may approximate the functional depending on the sampling distribution, F_n by that evaluated on the empirical distribution) then we may write:

$$T_n(F_n) = T(F) + \int IC(x; F, C) dF_n(x) + V_{mn} \tag{A.20}$$

and evaluating the integral over F_n and rewriting yields

$$\sqrt{n}(T_n - T(F)) \simeq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IC(X_i; T, F) + V_{mn}. \tag{A.21}$$

By the CLT, the first term on the right-hand side is asymptotically normal. In most cases, as $n \to \infty$ the remainder, $V_{mn} \to 0$ so that T_n is assymptotically normal, i.e.

$$J_n(\sqrt{n}(T_n - T(F)) \to N(0, var(T, F)) \tag{A.22}$$

where

$$var(T, F) = \int IC^2(x; T, F) dF(x). \tag{A.23}$$

This formula emphasizes the interest of influence functions, not only as a general tool for assymptotic expansions but also as an approximation to the variance of an estimate.