Statistica Sinica **5**(1995), 129-139

SOME STATISTICAL MEAN VALUE THEOREMS RELATED TO THE BOOTSTRAP

M. T. Chao and S. H. Lo

Academia Sinica and Columbia University

Abstract. Given a general statistic $T_n(\mathbf{X}, \theta) = T_n(X_1, \ldots, X_n, \theta)$, a representation is given for the difference between the bootstrapped statistic T_n^* and a replica of its own image $T_n^{'}$. Except for a high order error term, the difference, which explains the validity of the bootstrap method, consists of 3 components. The first component is the difference $\tilde{\theta} - \theta$, where $\tilde{\theta}$ is used in $F_{\tilde{\theta}}(\cdot)$ as the bootstrap resampling base. The other two components depend on the model $F_\theta(\cdot)$ and the statistic T_n only, and they appear in the form of an inner product and behave like a derivative of T_n with respect to θ. This representation is an application of the classical mean value theorem and it supports the superiority of the maximum likelihood summary as explored by Efron (1982b).

Key words and phrases: Bootstrap, bootstrap representation, estimation equation, maximum likelihood summary, mean value theorem, parametric bootstrap, root statistic.

1. Introduction

In a 1982 paper, Efron emphasizes, through persuasive argument, the concept of maximum likelihood summary (MLS). Under the framework of traditional parametric inference, if $f(\cdot, \theta)$ is the data generating mechanism, then $f(\cdot, \hat{\theta})$ can be used to compute or simulate the sampling distribution of any statistic $T(\mathbf{X}, \theta)$, where $\hat{\theta}$ is the MLE of θ. In fact, since θ is unknown, $f(\cdot, \hat{\theta})$ is about the best random mechanism one can hope for. This is the backbone of the bootstrap resampling technique, and is perceived, as usual, by R. A. Fisher.

Efron's argument is geometric and the conclusion is based on sample sizes ratio. Whereas this provides a strong basis for the claim that MLS is superior, it is not a formal proof. In particular, it is far from clear why the sampling distribution of any statistic $T(\mathbf{X}, \theta)$ can be optimally approximated via $f(\cdot, \hat{\theta})$.

In this paper, we develop a simple formula that supports the superiority of MLS in one line. This formula, suitably explained, is only an application of the classical mean value theorem. More precisely, this paper presents a class of representations for general statistics when the observations are sampled from a certain parametric family. We postulate a parametric family with density

$f(x,\theta)$ depending on a parameter θ. The discussions and results presented here are restricted to the case that both x and θ are univariate. These restrictions greatly simplify the presentation and derivations of the main results in this paper. However, with additional effort, we anticipate that similar results can be obtained along the same line in a more general set up.

Let $\tilde{\theta}_n$ be a reasonable estimator of the unknown parameter θ based on the random sample $\mathbf{X} = (X_1,\ldots,X_n)$ of size n from $f(x,\theta)$. If $\tilde{\theta}_n$ is used for summarizing purpose, $f(\cdot,\tilde{\theta}_n)$ is used as the sampling base. In doing so, we perform a (parametric) bootstrap. A general $\tilde{\theta}$ is deliberately used instead of the traditional maximum likelihood estimator $\hat{\theta}$; the purpose of this will be clear later. For simplicity, we drop the subscript n from θ and other quantities when there is no confusion. The (parametric) bootstrap sample $\mathbf{X}^* = (X_1^*,\ldots,X_n^*)$ is generated from the sampling base $F_{\tilde{\theta}}$, where F_θ denotes the cumulative distribution function (cdf) of $f(x,\theta)$. We are interested in the behavior of statistic $T_n = T_n(X_1,\ldots,X_n,\theta) = T_n(\mathbf{X},\theta)$ under $f(x,\theta)$. Let $T_n^* = T_n(\mathbf{X}^*,\tilde{\theta})$. The basic idea of the bootstrap method, parametric or nonparametric, is to claim (Efron (1979, 1982a)) or prove (Singh (1981), Bickel and Freedman (1981)) that T_n and T_n^* have similar sampling distributions.

The fundamental goal of this work is to study the relationship between $T_n^{'}$ and T_n^*, where $T_n^{'}$ is an iid copy of T_n and is defined as follows. Let $\mathbf{X}^{'} = (X_1^{'},\ldots,X_n^{'})$ be an (unobservable) iid replica of $(X_1,\ldots,X_n)$, and let $T_n^{'} = T_n(\mathbf{X}^{'},\theta)$. Furthermore, $X_i^{'}$ and X_i^* are associated by the relation $F_\theta(X_i^{'}) = F_{\tilde{\theta}}(X_i^*)$. The parametric bootstrap (PB) applied to this problem will lead to a representation which can be expressed as

$$T_n(\mathbf{X}^*,\tilde{\theta}) = T_n(\mathbf{X}^{'},\theta) + \triangle_n(\mathbf{X},\mathbf{X}^{'}). \tag{1}$$

It is clear that $T_n(\mathbf{X}^{'},\theta)$ has the same distribution as $T_n(\mathbf{X},\theta)$ but is independent of $T_n(\mathbf{X},\theta)$. Under some smoothness conditions, it is found in this paper that $\triangle_n(\mathbf{X},\mathbf{X}^{'})$ can be expressed as $(\tilde{\theta}-\theta)\bigtriangledown T(\mathbf{X}^{'},\theta)$ asymptotically, where $\bigtriangledown$ denotes the gradient operator and $\bigtriangledown T(\mathbf{X}^{'},\theta)$ plays the role of the derivative of T_n at θ and does not depend on $\mathbf{X}$. It is this form which justifies the title of this article.

The parametric bootstrap (PB) was mentioned and suggested by Efron in his pioneer work (1979, 1982a). However, compared with the ordinary bootstrap (nonparametric bootstrap), the PB has attracted much less attention in the literature. Like the nonparametric bootstrap (NB), the behavior of the bootstrapped statistics under the PB may be studied in terms of the central limit theorem (CLT) and Edgeworth expansions. This approach is indirect and the same difficulties arise, however, when the limiting distributions of T_n^* and T_n are difficult to derive or the distributions of T_n^* and T_n do not admit the CLT or

Edgeworth expansions. As a result, only sufficiently regular estimators can be justified through this approach, which, from a practical point of view, seems to be unnecessarily limited. Our approach is in the same spirit of Lo (1989), where nonparametric bootstrap is considered.

There are three components in the formulation (1). The estimator $\tilde{\theta}$ is used as a general resampling base to generate $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)$. The model $F_\theta(\cdot)$ represents the frequentist's view of the original data. The statistic T_n represents the inference introduced by the statistician.

As an important application, the representations obtained here are relevant to Efron's discussion (Efron (1982b)) of the superiority of the maximum likelihood summary (MLS) $\hat{f} = f(\cdot, \hat{\theta})$ as a distribution for bootstrapping, where $\hat{\theta}$ denotes the MLE. Since the representations for the MLS $f(\cdot, \hat{\theta})$ have analogous representations for an alternative summary $f(\cdot, \tilde{\theta})$ if another estimator $\tilde{\theta}$ is available, it is possible to compare the effectiveness of using $f(\cdot, \hat{\theta})$ and $f(\cdot, \tilde{\theta})$ for parametric bootstrap distribution to evaluate a third common statistic T. From (1), we see that if we want to keep $T_n(\mathbf{X}, \tilde{\theta}_n)$ and $T_n(\mathbf{X}^*, \theta)$ close, the only hope is to use a $\tilde{\theta}$ that is close to the true θ. Hence (1) strongly suggests that the MLS is about the optimal choice for bootstrapping whenever $\hat{\theta}$ is asymptotically efficient.

In Section 2, we introduce the familiar probability integral transform that is relevant to our representation. Let x be the datum from $f(\cdot, \theta)$. If θ changes to a nearby $\theta^{'}$, a specific feature, or technical device, of our development is that we do not hold x fixed, but to allow x changing to a nearby x', related by $F_\theta(x) = F_{\theta'}(x^{'})$, where F, with suitable subscript, denotes the cdf of f. In this way x is also a function of θ, and the basic relation (1) is only the first term of a Taylor expansion.

In Section 3, we derive a simple mean value theorem for general root statistics $R(\mathbf{X}, \theta)$ under the PB. By root we mean a real-valued random variable $R(\mathbf{X}, \theta)$ which is a function of θ and the data $\mathbf{X} = (X_1, \ldots, X_n)$ such that $E_\theta R(\mathbf{X}, \theta)$ is a constant for all $\theta \in \Theta$. Without loss of generality, one may always assume the constant mentioned above is zero. In this case a root is obtained from an estimating equation, introduced by Godambe (1960) and discussed extensively in a series of papers. With a specific, but natural, loss structure, Godambe demonstrated that the estimating equation is a natural concept in statistical inference. In contrast to his approach, however, our representation does not need the concept of a loss function. Relation (10) provides a general representation for bootstrapped root statistics. In this representation, the difference between a copy of the theoretical root and the bootstrapped root is decomposed into a product of three independent components plus a higher-order error term. The first component is $\tilde{\theta} - \theta$, which indicates the necessity of a good initial estimator for θ in order to obtain a better bootstrap approximation. Since the second and the third components depend on the model and the nature of the underlying root statistic only, no better approximation can be expected by adjusting these two components.

Estimators are determined by certain roots, either directly or implicitly. If an estimator is determined by a root directly, the results obtained in Section 3 can apply. If an estimator can only be found by solving a certain equation $Q(\mathbf{X}, \theta) = 0$, it is shown in Section 4 that such an estimator still admits an appropriate mean value theorem. Furthermore, the specific structure of the error term can also be found, which strongly suggests that the best approximation can be obtained by choosing an efficient estimator $\tilde{\theta}$ for bootstrapping in the first place. Since maximum likelihood estimators are asymptotically efficient under standard regularity conditions, this provides some direct evidence that the MLS is the optimal choice for bootstrapping, and which is also a point argued heuristically in Efron (1982b).

If Q can be expressed as an independent summand, in addition to giving an appropriate mean value theorem for $\tilde{\theta}_Q$, which is the solution of $Q = 0$, we show in Section 4 that the limiting distribution of the "derivative" is normal with mean zero and a specific variance. Despite the provision of an explicit expression for variance, a geometric interpretation for the variance term is still lacking.

2. The Probability Integral Transform

If X has a continuous cdf F, then $F(X)$ is $R(0,1)$, the uniform distribution over the interval $(0,1)$. This is the well known probability integral transform. Conversely, if U is $R(0,1)$, then $F^{-1}(U)$ has distribution F. This is frequently used as a general purpose random number generator. One way of looking at our approach is to pretend that Nature generates its random numbers the same way.

More specifically, we pretend that Nature first generates a U from $R(0,1)$ and then waiting for specific orders. If the request is F_θ, it gives $X = F_\theta^{-1}(U)$; if the request is $F_{\theta'}$, it gives $X' = F_{\theta'}^{-1}(U)$. Since U is common, we have

$$F_\theta(X) = F_{\theta'}(X^{'}). \tag{2}$$

Note that, under this formulation, X is a function of θ.

The traditional frequentist's view pretends that $\mathbf{X}$ is observed from F_θ. If θ changes to a nearby $\theta^{'}$, do we expect to observe an $\mathbf{X}^{'}$ slightly different from $\mathbf{X}$? In standard treatment, this $\mathbf{X}^{'}$ never comes into the picture. For example, the score function is obtained by differentiating the log likelihood function with respect to θ while keeping the data $\mathbf{X}$ fixed. Since in practice only the data $\mathbf{X}$ is given and no other data is available in statistical analysis, such a view seems to be natural and prevailing. All statistical theories, frequentist or Bayesian alike, are derived under such a premise.

But under a bootstrap setup, we expect the resampling data $\mathbf{X}^*$ to be generated according to $f(\cdot, \tilde{\theta})$. Relation (2) may be unexpectedly useful. In particular, let $T(\mathbf{X}, \theta)$ be a general statistic under investigation. Using (2) we may treat T as a function of θ only and write

$$T(\theta) = T(\theta_0) + (\theta - \theta_0) \cdot T^{'}(\theta_0) + \cdots \tag{3}$$

for θ near θ_0. Our basic relation (1) or (10), although expressed somewhat differently, is nothing but Equation (3).

As a sideline remark to (2), let Y and Y' be any pair of random variables with marginal distributions F_θ and $F_{\theta'}$ respectively, then (Bártfai (1970))

$$E(X - X^{'})^2 \leq E(Y - Y^{'})^2.$$

Hence (X, X') is the closest pair in mean square error sense.

3. The Representation of a Root under the PB

Suppose that $X_1, \ldots, X_n$ are iid from $f_\theta(x)$. Let $R(\mathbf{X}, \theta)$ be a root such that $E_\theta R(\mathbf{X}, \theta) = 0$ for all θ, and let $\tilde{\theta} = \tilde{\theta}_n = \tilde{\theta}_n(\mathbf{X})$ be a consistent estimator of θ. In this section we give a representation of $R(\mathbf{X}^*, \tilde{\theta}_n)$ in terms of $R(\mathbf{X}^{'}, \theta)$, where $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)$ is the bootstrap sample from $f(\cdot, \tilde{\theta})$. The sample $\mathbf{X}^{'} = (X_1^{'}, \ldots, X_n^{'})$ is unobservable and associated with the bootstrap sample by the relation $F_{\tilde{\theta}}(X_i^*) = F_\theta(X_i^{'})$ for all i, $1 \leq i \leq n$.

To simplify the presentation, we shall assume throughout this section that both $F_\theta(x)$ and $R(\mathbf{X}, \theta)$ are smooth enough to allow differentiation in both θ and the components of $\mathbf{X}$; also, all expansions in higher order terms are assumed valid.

For the location parameter case that $f(x, \theta) = f(x - \theta)$, let $R(\mathbf{X}, \theta) = \hat{\theta} - E_\theta(\hat{\theta})$, where $\hat{\theta}$ is a location invariant estimator of θ. It is easy to check that

$$R(\mathbf{X}^*, \hat{\theta}^*) = R(\mathbf{X}^{'}, \theta),$$

where $\hat{\theta}^* = \hat{\theta}(\mathbf{X}^*)$. This shows that the location root R is pivotal in the sense that the distribution of $R(\mathbf{X}, \theta)$ does not depend on θ. The scale parameter case $f(x, \theta) = e^{-\theta} f(e^{-\theta} x)$ can be dealt similarly. In both cases, $R(\mathbf{X}, \theta)$ is pivotal and (1) is exact with $\triangle \equiv 0$.

The following simple example, which is less trivial than the pivotal cases, shows that our general approach for the bootstrap representation is different from, and perhaps better than, the traditional CLT or Edgeworth expansion approaches.

Let $X_1, \ldots, X_n$ be iid F_θ, the uniform distribution over $(0, \theta)$. Let $X_1^{'}, \ldots, X_n^{'}$ be an iid replica. To fix ideas, let P denote the probability measure generated by the $\mathbf{X}$; $P^{'}$ denote the probability measure generated by the $\mathbf{X}^{'}$. We shall first work with the $P \times P^{'}$ measure. The MLE $\hat{\theta}_n$ of θ is $X_{(n)} = \max\{X_1, \ldots, X_n\}$. We will use $R(\mathbf{X}, \theta) = \hat{\theta}_n - \theta$ in (1).

Using the simple fact that F_θ is linear over $(0, \theta)$, a direct calculation shows that the error term $\triangle$ of (1) in this case is

$$\begin{aligned}\triangle_n(\mathbf{X}, \mathbf{X}^{'}) &= (\hat{\theta}_n^* - \hat{\theta}_n) - (\hat{\theta}_n^{'} - \theta) \\ &= F_{\hat{\theta}_n}^{-1}\left(F_\theta(X_{(n)}^{'})\right) - X_{(n)}^{'} - \left(X_{(n)} - \theta\right) \\ &= \left(\frac{\hat{\theta}_n}{\theta} X_{(n)}^{'} - X_{(n)}^{'}\right) - (X_{(n)} - \theta) \\ &= \left(\frac{X_{(n)}}{\theta} - 1\right) X_{(n)}^{'} - (X_{(n)} - \theta) \\ &= \frac{1}{\theta}(X_{(n)} - \theta)(X_{(n)}^{'} - \theta).\end{aligned}$$

Hence it is of order $O(n^{-2})$ with respect to the probability $P \times P^{'}$. Note that accuracy to this order cannot be obtained through the CLT or Edgeworth expansion approaches.

Although (1) is a relation over $P \times P^{'}$, there is no difficulty to arrange it in terms of the traditional conditional version. Using the uniform example again, the difference between the bootstrap and the true distribution can be expressed as, for any real t,

$$\begin{aligned}& P^*\left(n(\hat{\theta}_n^* - \hat{\theta}_n) \leq t|\mathbf{x}\right) - P\left(n(\hat{\theta}_n - \theta) \leq t\right) \\ &= P^{'}\left(n(\hat{\theta}_n - \theta) + n\Big(\frac{x_{(n)}}{\theta} - 1\Big)(\hat{\theta}_n^{'} - \theta) \leq t\right) - P^{'}\left(n(\hat{\theta}_n^{'} - \theta) \leq t\right) \\ &= P^{'}\left(n(\hat{\theta}_n^{'} - \theta)\frac{x_{(n)}}{\theta} \leq t\right) - P^{'}\left(n(\hat{\theta}_n^{'} - \theta) \leq t\right) \\ &= P^{'}\left(n(\hat{\theta}_n^{'} - \theta) \leq \frac{t\theta}{x_{(n)}}\right) - P^{'}\left(n(\hat{\theta}_n^{'} - \theta) \leq t\right) \\ &= O\left(\frac{\theta}{x_{(n)}} - 1\right)\end{aligned}$$

uniformly on $t \in R$ which is $O_P(1/n)$, according to the usual conditional approach.

Hereafter in this article, we shall focus on the discussion of unconditional approach only (on $P \times P^{'}$ measure).

With this convention in mind, we now return to the discussion of $F_{\tilde{\theta}}(X^*)$ and the root statistic $R(\mathbf{X}^*, \tilde{\theta})$. Applying a bivariate Taylor expansion to $F_{\tilde{\theta}}(X_i^*)$ yields

$$\begin{aligned} F_{\tilde{\theta}}(X_i^*) &= F_\theta(X_i^{'}) + (X_i^* - X_i^{'}, \tilde{\theta} - \theta) \left(\frac{\partial F_\theta(X)}{\partial X}, \frac{\partial F_\theta(X)}{\partial \theta} \right)^T \\ &\quad + \frac{1}{2} (X_i^* - X_i^{'}, \tilde{\theta} - \theta) \begin{bmatrix} \frac{\partial^2 F_\theta(X)}{\partial X^2} & \frac{\partial^2 F_\theta(X)}{\partial X \partial \theta} \\ \frac{\partial^2 F_\theta(X)}{\partial \theta \partial X} & \frac{\partial^2 F_\theta(X)}{\partial \theta^2} \end{bmatrix} \\ &\quad \cdot \ (X_i^* - X_i^{'}, \tilde{\theta} - \theta)^T + \text{higher order terms}, \end{aligned} \tag{4}$$

where and hereafter, the derivatives are all evaluated at $(X_i^{'}, \theta)$ or $(\mathbf{X}^{'}, \theta)$. Using the identity $F_{\tilde{\theta}}(X_i^*) = F_\theta(X_i^{'})$, $X_i^* - X_i^{'}$ can be expressed in terms of a power series in $\tilde{\theta} - \theta$:

$$X_i^* - X_i^{'} = \alpha_i (\tilde{\theta} - \theta) + \beta_i (\tilde{\theta} - \theta)^2 + (\tilde{\theta} - \theta)^3, \tag{5}$$

where

$$\begin{aligned} \alpha_i &= \ - \left(\frac{\partial F_\theta(X)}{\partial \theta} \right) \Big/ \left(\frac{\partial F_\theta(X)}{\partial X} \right), \\ \beta_i &= \ - \frac{1}{2} (\alpha_i, 1) \begin{bmatrix} \frac{\partial^2 F_\theta(X)}{\partial X^2} & \frac{\partial^2 F_\theta(X)}{\partial X \partial \theta} \\ \frac{\partial^2 F_\theta(X)}{\partial \theta \partial X} & \frac{\partial^2 F_\theta(X)}{\partial \theta^2} \end{bmatrix} \begin{pmatrix} \alpha_i \\ 1 \end{pmatrix} \Big/ \frac{\partial F_\theta(X)}{\partial X}. \end{aligned}$$

Relation (5), which provides a local linear rate between the differences $X_i^* - X_i^{'}$ and $\tilde{\theta} - \theta$, is deterministic in nature and is crucial in establishing our results. If $R(\mathbf{X}, \theta)$ is smooth enough to allow differentiations, we may express $R(\mathbf{X}^*, \tilde{\theta})$ in terms of $R(\mathbf{X}^{'}, \theta)$ plus a small remainder term. A multivariate Taylor expansion yields

$$R(\mathbf{X}^*, \tilde{\theta}) = R(\mathbf{X}^{'}, \theta) + U \cdot \bigtriangledown R + \frac{1}{2} U \cdot \bigtriangledown^2 R \cdot U^T + \text{high order terms}, \tag{6}$$

where

$$\begin{aligned} U &= (X_1^* - X_1^{'}, \ldots, X_n^* - X_n^{'}, \tilde{\theta} - \theta), \\ \bigtriangledown R &= \left(\frac{\partial R}{\partial X_1}, \ldots, \frac{\partial R}{\partial X_n}, \frac{\partial R}{\partial \theta} \right)^T, \end{aligned}$$

and $\bigtriangledown^2 R$ denotes the matrix of second order derivatives of R evaluated at $\mathbf{X}^{'}$ and θ.

From (5), we may express $U \cdot \bigtriangledown R$ in (6) as

$$\begin{aligned} U \cdot \bigtriangledown R &= (\alpha_1, \ldots, \alpha_n, 1) \bigtriangledown R \cdot (\tilde{\theta} - \theta) \\ &\quad + (\beta_1, \ldots, \beta_n) \left(\frac{\partial R}{\partial X_1}, \ldots, \frac{\partial R}{\partial X_n} \right)^T (\tilde{\theta} - \theta)^2 + O_{P \times P'}((\tilde{\theta} - \theta)^3). \end{aligned}$$

Likewise, one can express $(1/2)U\cdot\bigtriangledown^2R\cdot U^T$ in (6) as

$$\begin{aligned}\frac{1}{2}U\cdot\bigtriangledown^2R\cdot U^T=&\ \frac{1}{2}(\alpha_1,\ldots,\alpha_n,1)\cdot\bigtriangledown^2R\cdot(\alpha_1,\ldots,\alpha_n,1)^T\cdot(\tilde\theta-\theta)^2\\&+O_{P\times P'}((\tilde\theta-\theta)^3).\end{aligned}$$

Finally, we can rewrite $R(\mathbf{X}^*,\tilde\theta)$ in (6) as

$$\begin{aligned}&R(\mathbf{X}^*,\tilde\theta)\\&=R(\mathbf{X}',\theta)+(\tilde\theta-\theta)\left(\frac{\partial R}{\partial\theta}+\sum_{i=1}^n\alpha_i\frac{\partial R}{\partial X_i}\right)\\&\quad+(\tilde\theta-\theta)^2\Big[\sum_{i=1}^n\beta_i\frac{\partial R}{\partial X_i}+\frac{1}{2}(\alpha_1,\ldots,\alpha_n,1)\cdot\bigtriangledown^2R\cdot(\alpha_1,\ldots,\alpha_n,1)^T\Big]\\&\quad+O_{P\times P'}((\tilde\theta-\theta)^3).\end{aligned}\tag{7}$$

Furthermore, it can be shown that

$$E_\theta(\boldsymbol{\alpha}\cdot\bigtriangledown R)=E_\theta\Big(\frac{\partial R}{\partial\theta}+\sum_{i=1}^n\alpha_i\frac{\partial R}{\partial X_i'}\Big)=0\tag{8}$$

and

$$E_\theta\left(\sum_{i=1}^n\beta_i\frac{\partial R}{\partial X_i}+\frac{1}{2}\boldsymbol{\alpha}\cdot\bigtriangledown^2R\cdot\boldsymbol{\alpha}^T\right)=0,\tag{9}$$

where $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_n,1)$. In addition, if $R(\mathbf{X},\theta)=\sum_{i=1}^n g(X_i,\theta)$ for some smooth function g, then

$$\begin{aligned}R(\mathbf{X}^*,\tilde\theta)=&\ R(\mathbf{X}',\theta)+(\tilde\theta-\theta)(\boldsymbol{\alpha}\cdot\bigtriangledown R)\\&+O_{P\times P'}((\tilde\theta-\theta)^2n^{-\frac{1}{2}})+O_{P\times P'}((\tilde\theta-\theta)^3).\end{aligned}\tag{10}$$

The proofs of (8) and (9) are elementary but tedious. They are based on repeated use of integration by parts and the Fubini Theorem. We shall omit them. The expression (10) follows from (9) and the fact that the coefficient of $(\tilde\theta-\theta)^2$ in (7) is of order $O_{P'}(n^{-1/2})$. If $\tilde\theta-\theta=O_P(n^{-1/2})$, it is easy to see the remainder on the right hand side of (10) is of order $O_{P\times P'}(n^{-3/2})$.

Relation (10) is our basic result for the structure of a "regular" root R under PB. It shows that the difference (the error) between a copy of the true root R and the bootstrapped root R^* can be expressed as a product of three components. The first component is $\tilde\theta-\theta$, indicating the necessity of a good estimator $\tilde\theta$ for bootstrapping. The second component is the vector $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_n,1)$, which depends on the model F_θ alone. The third component is the gradient

$\bigtriangledown R = (\frac{\partial R}{\partial X_1}, \ldots, \frac{\partial R}{\partial X_n}, \frac{\partial R}{\partial \theta})$ which expresses the role played by the root R. Note that these three components act independently in the sense that a change in any component does not affect the others. Thus, for a given model $F_\theta(\cdot)$, we are in a position to quantitatively judge the effectiveness of different versions of the bootstrap (i.e. choices of $\tilde{\theta}$) against all possible inferences (choices of the statistic T_n).

It is worth noting that once the model and the root are chosen, the only factor which can improve the accuracy of the approximations is selection of the best initial estimator $\tilde{\theta}$ for bootstrapping. It is not surprising that, under standard regularity conditions, the maximum likelihood estimator provides an optimal choice, although it is also clear that the optimality will still be achieved by choosing any other efficient estimator for bootstrapping.

4. The Representation of an Estimator Derived via a Root

Estimators sometimes are determined by roots, either directly or implicitly. If an estimator has a specific form, like sample mean $\overline{X}_n$ or sample correlation coefficient, it is easy to form a root by subtracting its mean from the estimator. In this case, the estimator is determined by the relevant root directly. The representations of the bootstrap counterpart ($R^* = \hat{\theta}^* - E_{\tilde{\theta}}(\hat{\theta}^*)$) can thus be derived using the results obtained in the previous section.

When the estimator does not have a specific form or the estimator can only be obtained by solving certain equations numerically, then the estimator is determined by a root implicitly. Typical examples are the maximum likelihood estimators or the minimum distance estimators. In these cases, the estimator $\hat{\theta}_Q$ is the solution of an equation

$$Q(\mathbf{X}, \theta) = 0, \tag{11}$$

and $Q(\mathbf{X}, \theta)$ is a root statistic. A usual situation is that the formula Q is in closed form but $\hat{\theta}_Q$ has to be solved by numerical methods. Our objective in this section is to derive a representation of $\hat{\theta}_Q(\mathbf{X}^*) - \tilde{\theta}(\mathbf{X})$ in terms of $\hat{\theta}_Q(\mathbf{X}') - \theta$ and Q.

First we assume that Q has a specific form. More precisely, we assume $Q(\mathbf{X}, \theta) = \sum_{i=1}^n g(X_i, \theta)$ for some smooth function g. For example, with $g(x, \theta) = [\frac{\partial f(x,\theta)}{\partial \theta}] f(x, \theta)^{-1}$, one immediately obtains $Q(\mathbf{X}, \theta) = \frac{\partial \ell(\mathbf{X}, \theta)}{\partial \theta} = \sum_{i=1}^n \frac{\partial \ell(X_i, \theta)}{\partial \theta}$, the usual derivative of the log likelihood function. Here $\ell(\mathbf{X}, \theta)$ and $\ell(X_i, \theta)$ stand for $\log \prod_{i=1}^n f(X_i, \theta)$ and $\log f(X_i, \theta)$, respectively. This leads to the usual MLE. The assumption (about Q) can be removed and more general results expected but with additional effort. However, we have chosen a less general case for clarity and simplicity. In fact, many important cases (including the MLE) do satisfy this assumption. The proof is lengthy and is omitted; we merely summarize our findings:

Assuming that $\tilde{\theta}$ is a consistent estimator of θ, one can write

$$\hat{\theta}^*_Q - \tilde{\theta} = (\hat{\theta}'_Q - \theta) - (\tilde{\theta} - \theta)\frac{1}{\sqrt{n}}[Z'_{n1} + Z'_{n2}] + o_{P\times P'}((\tilde{\theta} - \theta)^2), \qquad (12)$$

where Z'_{n1}, Z'_{n2} are functions of $\mathbf{X}'$; and as $n \to \infty$,

$$Z'_{n1} \xrightarrow{\mathcal{L}} N(0, \sigma_1^2),$$
$$Z'_{n2} \xrightarrow{\mathcal{L}} N(0, \sigma_2^2),$$

with

$$\sigma_1^2 = \left[\frac{\partial E_\theta\left(\frac{\partial g(X_1,\theta)}{\partial\theta}\right)}{\partial\theta}\right]^2 \left[-E_\theta\Big(\frac{\partial g(X_1,\theta)}{\partial\theta}\Big)\right]^{-3}$$

$$\sigma_2^2 = E_\theta\left[\frac{\partial g(X_1,\theta)}{\partial\theta} + \alpha_1\frac{\partial g(X_1,\theta)}{\partial X_1}\right]^2 \left[E_\theta\left(\frac{\partial g(X_1,\theta)}{\partial\theta}\right)\right]^{-2}.$$

Acknowledgements

The first author's research was partially supported by the National Science Council of ROC; the second author's research was partially supported by Air Force Grant AFOSR 49620. Part of the current work was done while the second author was visiting the Institute of Statistical Science, Academia Sinica at Taipei in 1989. He wishes to express his deep gratitude to the Institute for all the support he received during this visit. He also wishes to thank Professor Herman Chernoff for his helpful discussions and constructive comments.

References

Bártfai, P. (1970). Über die Entfernung der Irrfahrtswege. Studis Sci. Math. Hungar **5**, 41-59.

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

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

Efron, B. (1982a). The Jackknife, the Bootstrap and Other Resampling Plans. SIAM CBMS-NSF Regional Conference Series in Applied Mathematics, no.38, Philadelphia, PA.

Efron, B. (1982b). Maximum likelihood and decision theory. Ann. Statist. **10**, 340-356.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. Ann. Math. Statist. **31**, 1208-1211.

Lo, S. H. (1989). On some representations of the bootstrap. Probab. Theory Related Fields **82**, 411-418.

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

Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan.

Department of Statistics, Columbia University, New York, NY 10027, U.S.A.

(Received October 1991; accepted August 1994)