Statistica Sinica 4(1994), 389-406

MAXIMUM LIKELIHOOD SUMMARY AND THE BOOTSTRAP METHOD IN STRUCTURED FINITE POPULATIONS

Min-Te Chao and Shaw-Hwa Lo

Academia Sinica and Columbia University

Abstract: Based on a few basic requirements on bootstraps, we derive the proper bootstrap resampling scheme for SRS (simple random sampling without replacement in a finite population). This concept is then extended to unequal probability samplings. We show that these definitions lead to no ambiguous bootstraps whether we treat the stratified populations as a whole population or through structured conditioning. Stratified and two stage samplings are included.

Key words and phrases: Bootstrap, maximum likelihood summary, simple random sampling, stratified random sampling, two stage sampling.

1. Introduction

Efron (1979) introduced the bootstrap method as a general purpose nonparametric tool to approximate the sampling distributions of estimators. Since the estimators are unspecified, the same bootstrap resampling method works for a class of estimators and plays the role of a summarizer. Thus, if $(F(\cdot,\theta))^n$ is the underlying data generating mechanism, then $((F(\cdot,\hat{\theta}))^n$ is used to generate the bootstrap sample, where $\hat{\theta}$ is the parametric MLE or $\hat{\theta} = \hat{F}$, the nonparametric MLE. A key feature is that $(F(\cdot,\theta))^n$ and $(F(\cdot,\hat{\theta}))^n$ have the same (product) structure. Being a resampling technique, the bootstrap method shares with other resampling schemes, e.g., the jackknife, the common implicit assumption that they can only be applied to essentially independent observations. For the iid case, there is no dispute regarding the resampling mechanism. Attempts to extend the bootstrap methods to dependent data roughly fall into two categories. One is objective-specific: new bootstrap methods are suggested on the basis that they "work" for a certain class of problems. The trouble of this approach is that the resulting methods may not work for some other cases, so that the modified version is no longer a general purpose summarizing tool, as it should be. The other approach is more fundamental. One tries to ask what is the "correct" bootstrap, in the sense that it is a natural extension of Efron's original bootstrap.

Studies of bootstrap methods to survey data came to the picture slowly.

Survey data, being dependent, is a good candidate to consider when one wants to suggest new bootstrap methods. Gross (1980) suggested the bootstrap methods with respect to simple random sampling without replacement. For an extensive study of its properties, see Chao and Lo (1985). For the large sample result, see Bickel and Freedman (1984). McCarthy and Snowden (1985) discuss the bootstrap methods in finite population sampling in more detail. Rao and Wu (1988) extend the bootstrap method to the case of complex survey data. For a Bayesian bootstrap applied to finite populations, see Lo (1988).

Except for the paper by Lo (1988), there are basically two approaches to treat the bootstrap methods in finite populations. The first type is operational. Under this approach, one tries to suggest different bootstrap resampling methods to approximate the sampling distributions of specific class of statistics. If the class of estimators is specified, then it may happen that the proposed resampling method, tailor made for such cases, do provide useful approximations to the desired sampling distribution. A typical example is Rao and Wu (1988), where attention is restricted to parameters defined by functions of population means only. Another approach is to treat the more general bootstrap as the natural extensions of the iid bootstrap. For this approach, one is more inclined to check how many of the original features of the iid bootstrap are preserved, and the utility for specific class of problems is not a major concern. An example is Chao and Lo (1985).

In this paper, we use the second approach and go one step further to explore the bootstrap method with respect to more complex probability sampling plans. By dealing with unbalanced situations with no empirical distribution to facilitate the large sample theory, we hope to explore the essential features of bootstrap methods which are easily masked by the ideal case of iid observations.

The fundamental message of this paper is that the bootstrap method can be formally formulated as a non-parametric ML method under the most general fixed sample scheme. There is no operational reason why we have to adopt to this principle except it is the natural thing to suggest in view of its optimal summarizing power as discussed and suggested in Efron (1982). A key message of this work is that an integral part of the bootstrap methods is to construct a population to estimate the population where the observed sample is coming from. In any case, the resampling scheme is *identical* to the original sampling plan except it is applied to the bootstrap population. Hence, a bootstrap scheme consists of two essential parts: a population that serves as a summarizer, and a random mechanism to generate the bootstrap sample. In this way, we maintain the tradition that the bootstrap method *mimics* the original sampling scheme.

Being an ML-based method, our approach suffers naturally from the general weakness of ML. For non-standard situations, we cannot always expect good resampling approximations under our general setup. For example, in small area survey, a discrete version of the Neyman-Scott problem, our approach needs adjustment. These are the exceptions to be expected, however.

2. The Bootstrap Method

Although not very popular in statistical literature, the bootstrap methods in finite populations have been discussed by several authors, see Chao and Lo (1985), Bickel and Freedman (1984). For the Bayesian version, see Lo (1988). Except for the work of Rao and Wu (1988), which deals with bootstrap methods in complex surveys for the class of estimators which are functions of the sample means only, essentially all such works are dealing with balanced situations; e.g., simple random sampling without replacement from a finite population (SRS). The bootstrap method of Rao and Wu (1988) is estimator-dependent and is not a general purpose tool.

The justification of these SRS based methods fall into two categories. Chao and Lo (1985) show that essentially all of Efron's earlier examples (1979) with respect to iid observations have parallel versions in an SRS setting with a natural finite correction factor attached to the estimated variances. Bickel and Freedman (1984) provide large sample results. These works tell us how bootstrap methods should be applied with respect to SRS and indicate the properties to be expected if such methods are applied. But little indication is given as to why the proposed method is the proper thing to do, and consequently extensions to general sampling schemes is hardly straightforward.

In this section we show that the finite population method of Gross (1980) is a logical thing to suggest if certain minimum properties on the proposed bootstrap method are required. We explore and suggest the concept that a design-based finite population bootstrap consists of two parts: an estimator Ω^* of the population Ω, and a sampling scheme $\{p(s), s \subset \Omega\}$ which is used as a bootstrap sample generator on Ω^*. Thus, the bootstrap mimics the original structure of the statistical problem.

2.1. Essential features of iid bootstrap

To extend the bootstrap method to non-iid cases, we first investigate the essential features of Efron's original bootstrap. If an extension is suggested, we hope that the proposed method still retain these features.

If $X_1, X_2, \ldots, X_n$ is an iid sample form F, for any statistic $R(X_1, X_2, \ldots, X_n, F)$ the bootstrap method suggests that the sampling distribution of $R^* = R(X_1^*, X_2^*, \ldots, X_n^*, \hat{F})$ can be used to approximate the sampling distribution of R, where $\mathbf{X}^* = \{X_1^*, X_2^*, \ldots, X_n^*\}$ is an iid sample from $\hat{F}$.

Under this setup, R is arbitrary. This means that the bootstrap is a general purpose tool in that it can be applied to all statistic R, not just a limited class of statistics. In this sense, $\hat{F}$ *summarizes* what we know about the sampling distribution of R for all R via the random mechanism $\hat{F}$.

For the iid case, $\hat{F}$ represents both a population (of infinite size) and a random mechanism from which samples can be drawn. The standard bootstrap instruction "take a sample of size n from $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ without replacement" combines these two points neatly.

As mentioned earlier, bootstrap mimics the original sampling scheme. For any statistic $R(\mathbf{X}, F)$, we use $R^* = R(\mathbf{X}^*, \hat{F})$. In this process, a simple substitution is employed. In order to make such a simple substitution *valid*, a restriction usually overlooked is that the support of $\hat{F}$ cannot include points outside the support of F. This will cause trouble if, for example, F is discrete.

Example 1. If F has jumps at 1, 2, and 3 only, then $X \sim F$ is trinomial with parameters $p_1, p_2, p_3 = 1 - p_1 - p_2$. The bootstrap method in this case is parametric, and we only need the MLE $(\hat{p}_1, \hat{p}_2, \hat{p}_3)$ to generate the bootstrap sample and this is $\hat{F}$. Intuitively, other sampling bases will be less efficient.

In fact, we have more serious problems. Let $T = X - 2$. Then $T = T^3$. Let

$$R(t) = a_0 + a_1 t + a_2 t^2 + \cdots \tag{2.1}$$

then

$$R(T) = a_0 + (a_1 + a_3 + \cdots)T + (a_2 + a_4 + \cdots)T^2 \tag{2.2}$$

so any function of $X (= T + 2)$ is a polynomial of degree 2, since X takes the values 1, 2 and 3 only. However (2.1) and (2.2) are not equivalent if T takes values outside $\{-1, 0, 1\}$. If (2.1) and (2.2) are not the same, we have a fundamental problem that R^* may not be well-defined, because using (2.1) and (2.2), which are equivalent for the original sample, direct substitution may give us different values.

The finite population problems are discrete, and we shall anticipate similar problems if we are not careful. The next result is general enough to cover all finite population case.

Lemma 2.1. *Let $S = \{s_1, s_2, \ldots, s_k\}$ be a finite set of real numbers. If $S^* \neq S$ then for any function R on S there exists a R^* on $S^* \cup S$ such that $R = R^*$ on S but $R \neq R^*$ on $S^* - S$.*

Proof. For simplicity, let $S^* \cup S = \{s_1, s_2, \ldots, s_k, t_1, \ldots, t_h\}$. The function R passes through k points $(s_i, R(s_i)), i = 1, 2, \ldots, k$ on the Euclidean plane. One way to express R is to use a $(k-1)$ degree polynomial, which can be determined by the Lagrange interpolation formula. Now consider the $k+h$ points $\{(s_i, R(s_i)), i = 1, \ldots, k; (t_i, R(t_i)+1), j = 1, 2, \ldots, h\}$. Construct a $(k+h-1)$ degree polynomial that passes through these $k+h$ points. Now $R^*(s_i) = R(s_i)$ by construction, but $R^*(t_j) = R(t_j) + 1 \neq R(t_j)$. This completes the proof.

If R^* is defined outside the domain of R, we have to extend the domain of definition of R to allow for proper substitution. Lemma 2.1 says that we can always extend R to R^* or R^{**} so that $R^* \neq R^{**}$ outside the domain of definition of R. In terms of bootstrap substitution, this means that for all R, the corresponding R^* is not mathematically well defined.

Another important bootstrap feature is that the probability structure of $\{X_1^*, X_2^*, \ldots, X_n^*\}$ should be identical to that of $\{X_1, X_2, \ldots, X_n\}$, except for different parametric values; namely,

$$F(x_1^*, \ldots, x_n^*, \hat{\theta}) \text{ and } F(x_1, \ldots, x_n, \theta)$$

both use the *same* joint distribution form F. Note that F is only a symbol to denote joint distribution. For a finite population problem, F includes information such as population structure together with the proposed sampling scheme. For example, F may denote the probability structure of a stratified sampling plan with Neyman allocation. In this sense the interpretation of (2.3) below is that the sampling distribution of $R(X_1, \ldots, X_n)$, where the X's are drawn from a stratified plan with the Neyman allocation, is approximated by the sampling distribution of $R(X_1^*, \ldots, X_n^*)$, where the X^*'s are drawn the same stratified plan but perhaps with a different parameter $\hat{\theta}$. Later, we shall see that we treat $\theta = \Omega$, the whole unknown population and $\hat{\theta} = \Omega^*$ is a set-valued MLE of Ω.

We do not have a formal proof for this requirement. But the heuristic is clear. Since, in the bootstrap, we anticipate

$$d[R(X_1^*, \ldots, X_n^*, \hat{\theta})] \sim d[R(X_1, \ldots, X_n, \theta)] \tag{2.3}$$

for any R, we see that in the expression

$$\int \exp\{\text{it} \cdot R(x_1^*, \ldots, x_n^*, \hat{\theta})\} dF(x_1^*, \ldots, x_n^*, \hat{\theta})$$

the x^*'s play the role of dummy variables. Hence we may drop the asterisks in the above expression, and it reduces to the characteristic function of R, except θ is at $\hat{\theta}$. The closeness of these two distributions is influenced *only* by the fact that $\hat{\theta}$ may not be equal to θ. In particular, if $\hat{\theta} = \theta$, then (2.3) is exact. This is the basic consistency requirement. If the joint distribution of $\{X_1^*, X_2^*, \ldots, X_n^*\}$ has any different structure than the joint distribution of $\{X_1, X_2, \ldots, X_n\}$, it is easy to construct R to amplify this difference.

Moreover, $F(x_1, x_2, \ldots, x_n, \theta)$ is equivalent to the likelihood function. All important statistical concepts, e.g., sufficiency, completeness, ancillarity, information, score function, etc. are defined via the likelihood function. If the resampling scheme uses an essentially different F^*, it may happen that some important features will be lost in the new structure.

2.2. Simple random sampling

In the previous subsection, we have observed that in order for the bootstrap technique to be a truly general purpose tool, namely, applicable to any statistic R, it is necessary that (a) the domain of definition of R^* cannot go beyond that of R, and (b) except for different parametric values, $\{X_1^*, X_2^*, \ldots, X_n^*\}$ and $\{X_1, X_2, \ldots, X_n\}$ have the same joint probability structure. Based on these two restrictions, we proceed to suggest that the bootstrap method of Chao and Lo (1985) or Gross (1980), is the only bootstrap method that can be suggested for the SRS without replacement.

Suppose an SRS $s = \{x_1, x_2, \ldots, x_n\}$ is selected from $\Omega = \{y_1, y_2, \ldots, y_N\}$. If $k = N/n$ is an integer, we define the bootstrap method as follows.

(i) Consider a population Ω^* where each element of the sample s is duplicated k times to form a sampled image of Ω. This is our bootstrap population.

(ii) Draw an SRS $s^* = \{x_1^*, x_2^*, \ldots, x_n^*\}$ of size n from Ω^*.

(iii) Let $R = R(\delta, \Omega)$ be any statistics (which may depend upon the unknown population Ω) and let $\delta^* = \delta(s^*)$. Then we shall use the sampling distribution of $R^* = R(\delta^*, \Omega^*)$ with respect to the random mechanism in (ii), obtained by analytic methods or simulation, to approximate the sampling distribution of R.

We proceed to argue that this bootstrap scheme is the only logical thing to do for SRS. In this process, only the method of moments is employed.

From the basic consistency requirement, $\{X_1^*, X_2^*, \ldots, X_n^*\}$ should have the same probability structure as $\{X_1, X_2, \ldots, X_n\}$; i.e., an SRS without replacement from a certain finite population, Ω^*, say. From Lemma 2.1, in order to avoid the problem that R^* may not be well defined, Ω^* cannot contain points not in Ω. In a typical survey problem in a finite population, all we see is the sample $\{x_1, x_2, \ldots, x_n\}$. Hence

$$\{x_1, x_2, \ldots, x_n\} \subset \Omega^*.$$

Now consider $\Omega^* = \{x_1, x_2, \ldots, x_n, y_{n+1}, \ldots, y_N\}$, where, for notational convenience, the x's denote the sampled points and the y's denote the unknown population values (parameters). We shall consider the problem of predicting y_j, using $s = \{x_1, x_2, \ldots, x_n\}$.

Let

$$Y^{(r)} = (x_1^r + x_2^r + \cdots + x_n^r + y_{n+1}^r + \cdots + y_N^r)/N.$$

Then $Y^{(r)}$ is the rth central moment of Ω^*. Using the method of moments, by equating population moments with the corresponding sample moments, we have

$$Y^{(r)} = (x_1^r + x_2^r + \cdots + x_n^r)/n$$

or

$$\frac{1}{n}(x_1^r + x_2^r + \cdots + x_n^r) = \frac{1}{N-n}(y_{n+1}^r + y_{n+2}^r + \cdots + y_N^r)$$

for $r = 1, 2, \ldots$. If $N = kn$ and the x's are all distinct, the above equalities hold if and only if the y's assume the values $x_1, x_2, \ldots$ only, with each x duplicated $(k-1)$-times.

The above construction is EM (expectation maximizing (Dempster, Laird and Rubin (1977)) in nature. For a finite population, the first N moments determine the values of all N units. This EM view may provide a constructive way to propose bootstrap methods in more complicated sampling schemes. For truly unbalanced situations, however, it is not clear how these equations can be established. One reason is that for unbalanced situations the population and sample moments no longer play a natural role to identify the population Ω. Hence, the EM algorithm with respect to the moments does not work in general.

We may also adopt the traditional parametric point of view. Consider an SRS of size n from a population of N as before. But let us assume that units in Ω take values in a finite set $T = \{t_1, t_2, \ldots, t_r\}$. Let M_j denote the number of units in Ω that takes the value t_j, and let N_j denote the number of units in s that takes the value t_j. Then $\boldsymbol{N} = (N_1, N_2, \ldots, N_r)$ has an r-variate hypergeometric distribution with the population fixed at $\boldsymbol{M} = (M_1, \ldots, M_r)$

$$P(\boldsymbol{N} = (N_1, N_2, \ldots, N_r) \mid \boldsymbol{M} = (M_1, \ldots, M_r)) = \frac{\binom{M_1}{n_1}\binom{M_2}{n_2}\cdots\binom{M_r}{n_r}}{\binom{N}{n}}.$$

The MLE of $\boldsymbol{M}$ is easily found to be

$$\hat{M}_j = n_j N/n,$$

$j = 1, 2, \ldots, r$. The above holds for all r and any arbitrary set T. If $r \to \infty$ and $s \subset T$, we see that n_j becomes either 0 or 1 (if elements of T are distinct) and $\hat{M}_j = N/n = k$ for the interval j that contains a value of x_j in the sample. Thus, the MLE $\hat{M}_j$ implies that the population should consist of k duplicates of the sample.

We have derived for the SRS case, the bootstrap population Ω^* to be the one that suitably duplicates the sampled values. The bootstrap population so constructed should have the same structure as Ω except for the values of its units. It is clear that the well-defining problem explained in Lemma 2.1 does not exist in Ω^*. In fact, Ω^* is the MLE of Ω, and it satisfies the substitution principle:

$$\text{MLE of } g(\theta) = g(\text{MLE of } \theta)$$

for any g.

If N/n is not an integer, we randomize Ω^*. For example, if $N = 12$ and $n = 5$, we will use

$$\Omega^* = \{x_1, \ldots, x_5, x_1, \ldots, x_5, r_1, r_2\},$$

where

$$r_1 = \sum_1^5 I_i x_i = \boldsymbol{I} \cdot \boldsymbol{x}$$

and

$$r_2 = \sum_1^5 J_i x_i = \boldsymbol{J} \cdot \boldsymbol{x},$$

and the vectors $\boldsymbol{I}, \boldsymbol{J}$ are iid 5-dimensional multinomials with cell probabilities 1/5 each. The general randomization scheme is clear. Randomization was first discussed by Chao and Lo (1985). However, their method tries to match the variances and is not consistent with the method of moments.

2.3. Unequal probability fixed size sampling

Armed with the ML principle, it is not difficult to define the bootstrap methods for general setup. To fix ideas, we only consider the case that a sample of fixed size n is selected from Ω with N items. The sampling plan is otherwise arbitrary. The following example should make the general definition easy to understand.

Example 2. Let $\Omega = \{y_1, y_2, \cdots, y_5\}$ and choose a sample of size 2 from Ω by Murthy's plan described in Cochran (1977) with $z_1 = z_2 = .1, z_3 = .2, z_4 = z_5 = .3$. The second order inclusion probabilities are

$$\pi_{ij} = \frac{z_i z_j (2 - z_i - z_j)}{(1 - z_i)(1 - z_j)}$$

for $i \neq j$. If $\{1, 2\}$ is selected, we find y_3, y_4, y_5 so that

$$P[\text{observe } \{x_1, x_2\} \text{ by Murthy's plan} \mid \Omega = \{x_1, x_2, y_3, y_4, y_5\}]$$

is maximum. There is no algorithm which we know of to effectively find the maximizing y_3, y_4, y_5. An exhaustive search is not out of reach in our case, however.

It is clear that if y_j $(j \geq 3)$ takes any value other than those already observed, the likelihood will decrease. Thus, we may restrict y_j to $\{x_1, x_2\}$ only. For the 8 possible cases, Table 1 shows that there are 4 possible choices of Ω^*, marked by $(*)$ in column 3. These are MLE's.

Table 1. MLE of Ω with respect to Murthy's plan when the first two elements of Ω are sampled.

Ω	value of likelihood	MLE
1,2,1,1,1	.1996	
1,2,1,1,2	.6175	*
1,2,1,1,2	.6175	*
1,2,1,2,2	.5210	
1,2,2,1,1	.5210	
1,2,2,1,2	.6175	*
1,2,2,2,1	.6175	*
1,2,2,2,2	.1996	

In unequal probability sampling, the labels of y are usually associated with covariates z. Hence the positions taken by the parameter values cannot be ignored. For example, 1,2,1,1,2 in Table 1 means that $\Omega^* = \{x_1, x_2, x_1, x_1, x_2\}$, in that order. More precisely, we have $\hat{y}_3 = x_1, \hat{y}_4 = x_1, \hat{y}_5 = x_2$ when $\{x_1, x_2\}$ at $\{1, 2\}$ are observed.

Having constructed Ω^*, which has the same structure as Ω except for some of its parameter values, we shall resample from Ω^*, using the same Murthy's plan and the same π_{ij}. For any statistic $R = R(\delta, \Omega)$ we approximate its sampling distribution by the sampling distribution of $R^* = R(\delta^*, \Omega^*)$. This completes our bootstrap construction.

The general definition is only slightly more complicated. Let $\Omega = \{y_1, y_2, \ldots, y_N\}$ be a finite population to be sampled. For each $s \subset \Omega$, let $p(s) = P(s|\Omega)$ denote the probability that the subset s is sampled. A sampling plan corresponds to a probability measure over the subset of Ω such that $\sum p(s) = 1$. For example, a fixed size $(= n)$ plan is one for which $p(s) = 0$ if s does not contain exactly n units.

Suppose $s = \{x_1, x_2, \ldots, x_n\}$ is sampled. Treating Ω as the population, the likelihood is obtained as follows:

$$L(\Omega|s) = P(s|\Omega).$$

The MLE Ω^* is defined by

$$P(s|\Omega^*) \geq P(s|\Omega)$$

for all Ω satisfying the assumed population structure. We remark here that in this framework, the structure of the population is taken into account by restricting the maximizing only to the populations satisfying the assumed structure. On the other hand, the structure of the sample is automatically taken into account by the sampling plan $\{p(s), s \subset \Omega\}$. Now Ω^* is our bootstrap population. Resamplings are done in Ω^* according to the original plan $p(\cdot)$. Let s^* be its corresponding sample in Ω^*. If $R(\delta(s), \Omega)$ is any statistic with respect to the sampling plan $p(\cdot)$, the estimator δ and the population Ω, we shall approximate the sampling distribution of R by the sampling distribution of $R^* = R(\delta(s^*), \Omega^*)$. This completes our definition of the bootstrap method.

It is easy to see that the MLE Ω^* must contain elements from s only. This is because $p(s)$ will increase when Ω^* contains duplications. For example, if $\Omega = \{x_1, x_1\}$, then the sample $s = \{x_1\}$ will be observed with probability 1 instead of $\frac{1}{2}$ for any sampling plan with sample size 1.

If N, n are small, it is not difficult to carry out the bootstrap scheme. If, however, N and n are even moderate, the general bootstrap method becomes impossible to program. Fortunately, large scale pure unequal probability plans are seldom used in practice. The definition we proposed is more useful to serve as the starting and check point of other, perhaps more structured, sampling plans. It also helps demonstrating that many standard estimators in sampling theory are actually MLE's. We shall discuss these points in later sections. The following example, however, shows the limitation of the bootstrap method.

Example 3. Systematic Sampling.

Suppose Ω contains 50 elements and a systematic sample of size 5, $s = \{x_1, \ldots, x_5\}$ is taken. Then

$$\Omega^* = \{x_1, \ldots, x_1, x_2, \ldots, x_2, \ldots, x_5, \ldots, x_5\}$$

(each x is duplicated 10 times, in that order) is the bootstrap population of Ω because under Ω^* and the systematic sampling, we shall observe s with probability 1. For the same reason, R^* is a constant for all δ^* and one cannot expect to approximate the sampling distribution of R by that of R^* except the trivial location match $E^*R^* = ER$, where E^* denotes the expectation with respect to the sampling plan applied to Ω^*. Hence unbiased estimators can be constructed but nothing more, even with the help of the bootstrap method. This is, of course, to be expected.

3. The Bootstrap Method for Stratified Sampling

If a sampling scheme is basically defined via blocks of SRS's, it is relatively easy to propose the corresponding bootstrap method. However, there are two views for this problem. We can either treat a complex sampling plan as a general unequal probability plan and define its bootstrap method accordingly, or consider the bootstrap method through conditioning, for example, treat the second stage plans as SRS's within each stratum and bootstrap accordingly. The latter approach is used in Bickel and Freedman (1984), but it is not at all clear whether these two approaches agree. In this section we start with the general bootstrap definition and derive the corresponding bootstrap method for the stratified plan. We show that the natural conditional approach agrees with the general bootstrap definition of Section 2. This provides positive evidence of our general definition.

Let us consider a stratified plan as follows. There are H strata. For the hth stratum Ω_h an SRS of size n_h is taken from the N_h units of Ω_h, $h = 1, 2, \ldots, H$. Let $\Omega = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_H$ and $n = n_1 + n_2 + \cdots + n_H$. Then this stratified plan can be viewed as an unequal probability plan of size n from Ω with population size $N = N_1 + N_2 + \cdots + N_H$. Let s be a sample of the form $s = s_1 \cup s_2 \cup \cdots \cup s_H = \{x_{hj} : j = 1, 2, \ldots, n_h; h = 1, 2, \ldots, H\}$, where x_{hi} is taken from Ω_h. For generality, we do not assume that elements of s are distinct. According to our general bootstrap definition, we need to construct Ω^* such that

$$P(s|\Omega^*) \geq P(s|\Omega) \tag{3.1}$$

for all Ω.

To maximize (3.1), we shall see that we should have $\Omega^* = \Omega_1^* \cup \Omega_2^* \cup \cdots \cup \Omega_H^*$ and Ω_h^* must contain elements from s_h only. Since $\Omega = \Omega_1 \cup \cdots \cup \Omega_H$, we can represent the parameters (elements of $\Omega - s$) with vectors $\boldsymbol{R}_h = (R_{h1}, \ldots, R_{hn})$, $h = 1, 2, \ldots, H$ where R_{hk} represents the number of repetitions that the kth elements in s appear in Ω_h.

In the following, we take the sample s as an unordered set of values. This means that when we see $x \in s$, we only know that $x \in \Omega_h$ for some h, but we do not know the value of h. If we also know the value of h for all $x \in s \cap \Omega_h$, $h = 1, 2, \ldots, H$, the maximizing procedure would be much simpler.

Since we can always increase the likelihood of observing s by allowing more repetitions of elements of s in Ω_h, it is clear that we can restrict our attention to those $\boldsymbol{R}_h$ such that

$$R_{h1} + R_{h2} + \cdots + R_{hn} = N_h. \tag{3.2}$$

Consider a partition π of the index set $\{1, 2, \ldots, n\}$ as follows: $\pi = (\pi_1, \pi_2, \ldots, \pi_H)$, where π_h contains exactly n_h elements. Then for Ω satisfying (3.2), we have

$$P(s|\Omega) = \frac{\sum_\pi \prod_{h=1}^H \prod_{j\in\pi_h} R_{hj}}{\prod_{h=1}^H \binom{N_h}{n_h}}. \tag{3.3}$$

Hence, to find Ω^*, it suffices to maximize

$$\sum_\pi \prod_{h=1}^H \prod_{j\in\pi_h} r_{hj} \tag{3.4}$$

subject to

$$\sum_{j=1}^H r_{hj} = 1, \quad r_{hi} \geq 0, \tag{3.5}$$

where

$$r_{hj} = R_{hj}/N_h, \quad h = 1, 2, \ldots, H. \tag{3.6}$$

Let

$$\begin{aligned} r^*_{hj} &= n_h^{-1}, \text{ for } n_1 + \cdots + n_{h-1} < j \leq n_1 + \cdots + n_h, \\ &= 0, \quad \text{otherwise.} \end{aligned} \tag{3.7}$$

We see that the configuration (3.7) represents an $\Omega^* = \Omega^*_1 \cup \Omega^*_2 \cup \cdots \cup \Omega^*_H$ in which Ω^*_h contains exactly N_h/n_h duplicates of each $x_{hj}, j = 1, 2, \ldots, n_h$; namely, Ω^* is constructed by bootstrapping each Ω^*_h by SRS.

The following theorem can be proved by pure analytic argument. It says that the Ω^* defined by (3.7) is the Ω^* that maximizes (3.1).

Theorem 3.1. *The configuration $\{r^*_{hj}\}$ defined in* (3.7) *maximizes the expression* (3.4). *Therefore the Ω^* so defined is the bootstrap population of Ω with respect to stratified sampling.*

For proof, see the Appendix. The message of this theorem is as follows. A stratified plan can be viewed as a special case of an unequal probability plan. When we take such a general view, it is not clear whether the general ML configuration of Section 2 will agree with the simple minded one stratum at a time bootstrap scheme. Theorem 3.1 confirms this important check point. In this process, we deliberately dropped the "labels" of elements in s so we directly established this result from the basic definition without using the bootstrap result in SRS. More importantly, it provides a clue to define bootstrap methods via conditioning in more complex sampling schemes.

As a by product, we have just shown that Ω^* is the MLE of Ω in stratified sampling.

Remark. If, in addition to the values of $x \in s$, we also know which stratum it belongs to, then the summation sign $\sum_\pi$ in (3.3) will disappear and

$\{1, 2, \ldots, n_1\} + \{n_1 + 1, \ldots, n_1 + n_2\} + \cdots + \{n_1 + \cdots + n_{H-1} + 1, \ldots, n\}$ is the only partition of $\{1, 2, \ldots, n\}$. Without the summation $\sum_\pi$, the proof simplifies considerably.

4. Two-Stage Sampling

In this section we start with a balanced two-stage sampling plan and discuss its corresponding bootstrap methods.

Consider a two-stage setup as follows. Let Ω consist of M primary sampling units (psu's). An SRS of size m is taken to select the first stage psu's. For each selected psu, an SRS of size n is taken to select the second stage samples. For simplicity, we assume all psu's are of the same size N.

Example 4. Consider the special case $M = 3, m = 2, N = 4$ and $n = 2$. Assume the first two psu's are selected and the second stage samples are $s_1 = \{x_1, x_2\}, s_2 = \{y_1, y_2\}$. We proceed to find Ω^*, the bootstrap population of Ω under this two-stage plan. An analytic search of Ω^* (such as the method for the stratified plan) turns out to be surprisingly difficult. But we can suggest a few possible candidates.

Consider

$$\begin{aligned}\Omega_1^* &= \{x_1, x_2, x_1, x_2\} \cup \{y_1, y_2, y_1, y_2\} \cup \{x_1, x_2, y_1, y_2\},\\ \Omega_2^* &= \{x_1, x_2, x_1, x_2\} \cup \{y_1, y_2, y_1, y_2\} \cup \{x_1, x_2, x_1, x_2\}.\end{aligned}$$

Since Ω_1^* is more "balanced" than Ω_2^* one would conjecture that the likelihood of observing $s = s_1 \cup s_2$ would be larger under Ω_1^* than under Ω_2^*. The fact, however, is

$$P(s|\Omega_1^*)/P(s|\Omega_2^*) = 3/5.$$

Furthermore, the above relation holds not just for $N = 4$, but for all $N = 2k$.

An obvious message from this example is that in constructing Ω^*, we probably should first bootstrap each sampled psu by evenly filling their sampled duplicates as in SRS. After this is done, we should fill the population of unsampled psu's with these bootstrapped psu's with respect to the first stage plan. In doing so, we treat each bootstrapped psu as an indivisible unit. It is these whole units that we use to fill the unobserved psu's. In other word, we do not try to break $\{x_1, x_2, x_1, x_2\}$ and $\{y_1, y_2, y_1, y_2\}$ in halves to construct $\{x_1, x_2, y_1, y_2\}$ as in Ω_1^* of Example 4.

We can use the multivariate hypergeometric distribution to construct the bootstrap population for the balanced two-stage plan. Assume there are K different psu's, $p_1, p_2, \ldots, p_K$, say. There are L possible values for the second stage units to take, say $t_1, t_2, \ldots, t_L$. In the hth psu, assume $N_{h\ell}$ units take the value $t_\ell, \ell = 1, 2, \ldots$, L. We have

$$\sum_{\ell=1}^{L} N_{h\ell} = N, \quad h = 1, 2, \ldots, M.$$

If Ω_h is sampled, let $n_{h\ell}$ be the number of units in that sample that take the value t_ℓ. Then

$$\sum_{\ell=1}^{L} n_{h\ell} = n, \quad h = 1, 2, \ldots, m.$$

Let M_k denote the number of psu's that are of the kth type, and let m_k be the number of psu's in the first stage sample that are of the kth type, $k = 1, 2, \ldots, K$. The random variables become $\boldsymbol{m} = (m_1, m_2, \ldots, m_K)$ and $\boldsymbol{n}_h = (n_{h1}, n_{h2}, \ldots, n_{hL})$. The parameters to be estimated become $\boldsymbol{M} = (M_1, \ldots, M_K)$ and $\boldsymbol{N}_h = (N_{h1}, \ldots, N_{hL}), h = 1, 2, \ldots, M$.

The notation $\boldsymbol{m}$ can either be written as a vector of 0's and 1's, or in terms of $i_1 < i_2 < \cdots < i_m$. For example, $\boldsymbol{m} = (1, 0, 1, 0, 1)$ is equivalent to the ordered indices $1 < 3 < 5$. With this in mind, the full likelihood of the sample $(\boldsymbol{m}, \boldsymbol{n}_h, h = 1, \ldots, m)$ is

$$\prod_{j=1}^{m} \left[\frac{\binom{N_{i_j1}}{n_{i_j1}} \binom{N_{i_j2}}{n_{i_j2}} \cdots \binom{N_{i_jL}}{n_{i_jL}}}{\binom{N}{n}} \right] \frac{\binom{M_1}{m_1} \binom{M_2}{m_2} \cdots \binom{M_K}{m_K}}{\binom{M}{m}}.$$

The maximizing point is easy to see: $\hat{M}_i = M/m$ for the sampled psu's and $\hat{M}_i = 0$ otherwise. For fixed $\boldsymbol{m}$ of the form $i_1 < i_2 < \cdots < i_m$, say,

$$\begin{aligned} \hat{N}_{h\ell} &= N/n, \ \text{ for } h = i_1, i_2, \ldots, i_n;\ \ell = 1, 2, \ldots, L, \\ &= 0, \qquad \text{otherwise.} \end{aligned}$$

This covers the two-stage bootstrap suggested in Example 4.

In general, as long as the sample sizes and psu sizes are balanced, there is no difficulty to use the same decomposition to establish the general two-stage plan. If the m psu's are selected from the M psu's with a sampling plan $P_1(\cdot)$, from the hth selected psu Ω_h a plan $P_{2h}(\cdot)$ is applied to select the units of $\Omega_h, h = 1, 2, \ldots, m$. Then we should first bootstrap each selected psu to form $\Omega_h^*, h = 1, 2, \ldots, m$. We then treat each Ω_h^* as an indivisible, sampled unit of Ω to form a bootstrap population of psu's Ω^*. In each stage, the general bootstrap definition of Section 2 is applied.

This is equivalent to considering a joint version of these two stages to form a general one stage plan of size nm from a population of size NM. The conditional approach, of course, offers many more practical implications.

5. Conclusion

The key ingredient for the finite population bootstrap method is the construction of a population, through the ML principle, from which the samples are drawn from. Hence the bootstrap population should be identical in structure to the population under study. A bootstrap procedure consists of two components: the estimated population and the method of sampling. If the sample size is infinity, we should be able to estimate the population with certainty. Hence the resampling method should produce the desired sampling distribution for any statistics if the resampling method is identical to the original one.

We have demonstrated that a key step is to estimate the true population by the ML principle, whether this population is finite or infinite, with or without the independence structure. In fact, there is no real reason that we should use the ML principle except for its well-known credibility. This provides a solid base for the investigation of bootstrap methods in other non-standard cases.

The bootstrap methods suggested in this paper are meant to be of general purpose. We are concerned mostly with the problem of a *correct* bootstrap, which works for a wide class of estimators, rather than convenient sampling schemes for specified problems. For complex surveys, if θ is a function of the means, Rao and Wu (1988) suggested special purpose sampling procedures. Their methods do not agree with the general requirement for bootstrap methods stated in Subsection 2.1. It is possible to construct samples in which it will fail.

But generality also pays a price, at least in some practical problems when no generality is required. Nevertheless statisticians struggle between practicability and generality; the former is concerned with how statistical methods are used, and the latter contributes to basic understanding.

Appendix

We prove Theorem 3.1 in this appendix. A few lemmas are needed. The first one is standard and can be proved by a few applications of the Holder's inequality.

Lemma A.1. *Assume* A_{hi} *are real,* $h = 1, 2, \ldots, H, i = 1, 2, \ldots, K$. *Then*

$$\sum_i \prod_h A_{hi} \le \prod_{h=1}^{H} \|\boldsymbol{A}_h\|_H,$$

where

$$\|\boldsymbol{A}_h\|_H = \Big(\sum_i |A_{hi}|^H\Big)^{\frac{1}{H}}.$$

Lemma A.2. *Let $a_i \geq 0$ and $\sum_{i=1}^{n} a_i = 1$. For $p = 2, 3, \ldots,$ define*

$$L = \sum_{1 \leq i_1 < \cdots < i_k \leq n} (a_{i_1} a_{i_2} \cdots a_{i_k})^p.$$

Then $L \leq k^{-kp}$. Further, this inequality is sharp.

Proof. We may assume without loss of generality that $a_1 \geq a_2 \geq \cdots \geq a_n$. Fix $a_1, \ldots, a_{n-2}$ and first differentiate L with respect to a_n. By noting that $a_{n-1} + a_n =$ constant, we have, after simplification,

$$\frac{\partial L}{\partial a_n} = B(a_n) \cdot (a_n - a_{n-1}),$$

where

$$B(a_n) = p\Big\{\sum_{S_1} (a_{i_1} a_{i_2} \cdots a_{i_{k-1}})^p (a_n^{p-2} + a_n^{p-3} a_{n-1} + \cdots + a_n a_{n-1}^{p-3} + a_{n-1}^{p-2}) - \sum_{S_2} (a_{i_1} a_{i_2} \cdots a_{i_{k-2}})^p a_{n-1}^{p-1} a_n^{p-1}\Big\},$$

where $S_1 = \{1 \leq i_1 < \cdots < i_{k-1} < n-1\}$ and $S_2 = \{1 \leq i_1 < i_2 < \cdots < i_{k-2} < n-1\}$.

We claim that L is maximized at $a_n = 0$.

Case 1. If $B(a_n) \geq 0$ for all a_n. Then $\frac{\partial L}{\partial a_n} \leq 0$, and the maximum is at the extreme point $a_n = 0$.

Case 2. $B(a_n) < 0$ for some $a_n < \frac{1}{2}(1 - a_1 - \cdots - a_{n-2}) = \frac{c}{2}$, say. Note that $B(a_n)$ is of the form

$$B(a_n) = C_1 \cdot \frac{a_n^{p-1} - (c - a_n)^{p-1}}{2a_n - c} - C_2 a_n^{p-1} (c - a_n)^{p-1},$$

where $C_1, C_2 > 0$. Let $x = a_n / c$, then

$$B(a_n) = C_3 \cdot (x^{p-1} - (1-x)^{p-1})/(2x - 1) - C_4 (x(1-x))^{p-1},$$

$0 \leq x \leq 1/2, C_3, C_4 > 0$. We shall measure B with the new x-scale. Now

$$B'(x) = \frac{C_3}{(2x-1)^2}\{(2x-1)[(p-1)x^{p-2} + (p-1)(1-x)^{p-2}] - 2[x^{p-1} - (1-x)^{p-1}]\} - C_4(p-1)[x(1-x)]^{p-2}(1 - \frac{x}{2}).$$

If $B'(x) \leq 0$, then since $B(0) > 0$, the function B changes sign only once in $(0, 1/2)$. It follows that $\frac{\partial L}{\partial a_n} = B(a_n) \cdot (2x - 1)$ is negative at $a_n = 0$, crosses the

x-axis at some point x_0 and remains to be positive until at $x = 1/2$. Hence L, as a function of x, assumes a minimum at $x = x_0$. To show $L(0)$ is maximum is equivalent to showing $L(0) > L(1/2)$. This can be verified by direct substitution.

For $p \geq 2$, Lemma A.3 shows that $B'(x) \leq 0$ so that $L(a_n)$ has maximum at $a_n = 0$. Therefore we have shown that $L(a_n)$ is maximized at $a_n = 0$.

Now we fix $a_1, \ldots, a_{n-3}, a_n = 0$ and consider L as a function of a_{n-1}. This same argument shows that L is maximized at $a_{n-1} = 0$. In this way we show that L is maximized at $a_{k+1} = \cdots = a_n = 0$; or

$$L \leq (a_1 a_2 \cdots a_k)^p, \quad a_1 + \cdots + a_k = 1.$$

The right hand side is maximized at $a_1 = \cdots = a_k = 1/k$, which yields the upper bound k^{-kp}. We need to show

Lemma A.3. *If* $p = 2, 3, 4, \ldots$, *then* $B'(x) \leq 0$.

Proof. Note that the second term of $B'(x)$ is negative for $0 \leq x \leq 1/2$. It suffices to show

$$F(x) = (p-1)x(1-x)(x^{p-3} - (1-x)^{p-3}) - (p-3)(x^{p-1} - (1-x)^{p-1}) \geq 0.$$

If $p = 2$, then $F(x) = 0$. We thus assume $p \geq 3$. Let $y = x/(1-x)$. Then $F(x) \geq 0$ is equivalent to

$$H(y) = (p-1)(y^{p-2} - y) - (p-3)(y^{p-1} - 1) \geq 0$$

for $0 \leq y \leq 1$. The last inequality holds because $H''(y) > 0, H'(0) < 0, H(0) = p - 3 \geq 0, H(1) = 0$.

Proof of Theorem 3.1. Index the partitions π from 1 to K. If $\pi = (\pi_1, \pi_2, \ldots, \pi_H)$ corresponds to the ith partition, write

$$A_{hi} = \prod_{j \in \pi_h} r_{hj} \quad i = 1, 2, \ldots, K.$$

Then (3.4) becomes

$$\sum_{i=1}^{K} \prod_{h=1}^{H} A_{hi}$$

which, by Lemma A.1, is bounded by the product of the norms. The fact that the upper bound of Lemma A.2 is achieved by $\{r^*_{hj}\}$ of (3.7) is easy to check.

Acknowledgement

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 5-22107.

References

Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Ann. Statist.* **12**, 470-482.

Chao, M. T. and Lo, S. H. (1985). A bootstrap method for finite population. *Sankhyā Ser. A* **47**, 399-405.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition. John Wiley, New York.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

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

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

Gross, S. (1980). Median estimation in sample surveys. Paper presented at 1980 ASA meeting.

Lo, A. Y. (1988). A Bayesian bootstrap for a finite population. *Ann. Statist.* **16**, 1684-1695.

McCarthy, P. J. and Snowden, M. A. (1985). The bootstrap and finite population sampling. U.S. Department of Health and Human Services Publication no. 85-1369.

Rao, J. N. K. and Wu, C. F. J. (1988). Resampling inference with complex survey data. *J. Amer. Statist. Assoc.* **83**, 231-241.

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 December 1993)