ASYMPTOTIC NORMALITY AND THE BOOTSTRAP IN STRATIFIED SAMPLING

BY

P. J. BICKEL[1] AND D. A. FREEDMAN[2]

TECHNICAL REPORT NO. 29
JANUARY 1984

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY

RESEARCH SUPPORTED IN PART BY
[1]OFFICE OF NAVAL RESEARCH CONTRACT N00014-80-C-0163
[2]NATIONAL SCIENCE FOUNDATION GRANT MCS80-02535

Prepared for Ann. Statist.

Asymptotic normality and the bootstrap in stratified sampling

P. J. Bickel[1]
D. A. Freedman[2]
University of California, Berkeley

Abstract

This paper is about the asymptotic distribution of linear combinations of stratum means in stratified sampling, with and without replacement. Both the number of strata and their size is arbitrary. Lindeberg conditions are shown to guarantee asymptotic normality and consistency of variance estimators. The same conditions also guarantee the validity of the bootstrap approximation for the distribution of the t-statistic. Via a bound on the Mallows distance, situations will be identified in which the bootstrap approximation works even though the normal approximation fails. Without proper scaling, the naive bootstrap fails.

[1]This work was performed with the partial support of Office of Naval Research Contract N00014-80-C-0163. The hospitality of the Hebrew University, Jerusalem is also gratefully acknowledged.

[2]Research partially supported by National Science Foundation Grant MCS80-02535.

1. Introduction

Consider the problem of estimating a linear combination $\gamma = \sum_{i=1}^{p} c_i\mu_i$ of the means $\mu_1,\ldots,\mu_p$ of p numerical populations $X_1,\ldots,X_p$ with corresponding distributions $F_1,\ldots,F_p$. For each $i = 1,\ldots,p$ there is a sample X_{ij} from population X_i; the sample elements are indexed by $j = 1,\ldots,n_i$. Thus, n_i is the size of the sample from the i^{th} population. Two situations will be discussed:

(a) The populations X_i are assumed arbitrary and the sampling is with replacement: X_{ij} for $j = 1,\ldots,n_i$ are identically distributed with common distribution F_i; all the X_{ij} are independent.

(b) The populations are assumed finite; X_i has known size N_i; sampling is without replacement and independent in i; in this case, F_i is uniform. Enumerate X_i as $\{x_{i1},\ldots,x_{iN_i}\}$.

For simplicity, the populations are supposed univariate.

The natural unbiased estimate of γ is

$$\hat{\gamma} = \sum_{i=1}^{p} c_i X_{i\cdot}. \tag{1}$$

Here, the dot is the averaging operator.

Let τ_a^2 or τ_b^2 denote the variance of $\hat{\gamma}$ under sampling schemes (a) and (b) respectively. Let $\hat{\tau}_a^2$ or $\hat{\tau}_b^2$ be the customary unbiased variance estimates. Inference about γ can be based either on the normal approximation to the distribution of $(\hat{\gamma}-\gamma)/\hat{\tau}$ or on bootstrap approximations. This paper will discuss the validity of these approximations when the total sample size tends to ∞ in any way whatsoever, e.g., many small samples or a few large samples or some combination thereof. More precisely: suppose p, the c_i, the populations, the

N_i, and n_i all depend on an index ν such that $n(\nu) = n_1(\nu) + \cdots + n_p(\nu)$ $\to \infty$ as $\nu \to \infty$. This index will be suppressed in the sequel.

Here are two examples.

(a) The X_{ij} are unbiased measurements of the same quantity μ, taken with p different instruments. So the precision of X_{ij}, viz.,

$$\sigma_i^2 = \int (x-\mu)^2 dF_i(x)$$

depends on i. If σ_i^2 is known to be proportional to r_i, then

$$\hat{\gamma} = \sum \frac{n_i}{r_i} X_{i\cdot} / \sum \frac{n_i}{r_i}$$

is the natural estimate of μ.

(b) In the classical stratified sampling model a population X of size N is broken up into disjoint strata $X_1, \ldots, X_p$ of sizes $N_1, \ldots, N_p$ respectively; $\sum_{i=1}^p N_i = N$. From stratum i the sample X_{ij} for $j = 1, \ldots, n_i$ is taken without replacement. Enumerate the ith stratum as $\{x_{i1}, \ldots, x_{iN_i}\}$. The population mean is

$$\gamma = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^{N_i} x_{ij} = \sum_{i=1}^p N_i x_{i\cdot}/N$$

and $\hat{\gamma} = \sum_{i=1}^p N_i X_{i\cdot}/N$ is the usual estimate of γ.

We first take up the normal approximation in case (a). Suppose

$$\int x^2 dF_i < \infty \quad \text{and} \quad n_i \geq 2 \quad \text{for} \quad i = 1, \ldots, p \tag{2}$$

Then

$$\tau_a^2 = \sum_{i=1}^p c_i^2 \sigma_i^2 / n_i \quad \text{where} \quad \sigma_i^2 = \text{var}\, X_{ij}$$

and

$$\hat{\tau}_a^2 = \sum_{i=1}^p c_i^2 s_i^2 / n_i$$

where

$$s_i^2 = (n_i-1)^{-1} \sum_{j=1}^{n_i} (X_{ij}-X_{i\cdot})^2$$

Let

$$\phi(x,\varepsilon) = x \quad \text{for } |x| \geqq \varepsilon$$
$$= 0 \quad \text{otherwise}$$
$$\bar{\phi}(x,\varepsilon) = x - \phi(x,\varepsilon)$$

Suppose that for all $\varepsilon > 0$,

(3) $$\tau_a^{-2} \sum_{i=1}^p n_i^{-1} c_i^2 \, E\{\phi^2(X_{ij}-\mu_i, \varepsilon n_i \tau_a |c_i|^{-1}\} \to 0$$

By the Lindeberg-Feller theorem, $(\hat{\gamma}-\gamma)/\tau_a$ converges in law to $N(0,1)$, the standard normal distribution.

According to the first main theorem of this paper, conditions (2) and (3) are also sufficient to guarantee that $\hat{\tau}_a^2$ has the right limiting behavior. However, before giving a precise statement, it may be helpful to reformulate condition (3). Let $Y_{ij} = (X_{ij}-\mu_i)/\sigma_i$. Define the "variance weight" of the i^{th} stratum by

$$w_i^2 = c_i^2 \sigma_i^2 / n_i \tau_a^2 = \text{var}\{c_i X_{i\cdot}/\tau_a\}$$

Clearly, $$\sum_{i=1}^p w_i^2 = 1 .$$

Condition (3) can then be written

(4) $$\sum_{i=1}^{p} E\{\phi^2(w_i Y_{ij}, \varepsilon\sqrt{n_i})\} \longrightarrow 0 \quad \text{for all} \quad \varepsilon > 0$$

THEOREM 1. *If* (2) *and* (4) *hold in case (a), then* $\hat{\tau}_a^2/\tau_a^2 \to 1$ *in probability.*

The proof is deferred.

COROLLARY. $(\hat{\gamma}-\gamma)/\hat{\tau}_a$ *tends to* $N(0,1)$ *in law.*

We consider next the bootstrap approximation in case (a); also see Babu and Singh (1983). For $i = 1,\ldots,p$, let $\hat{F}_i$ be the empirical distribution of X_{ij} for $j = 1,\ldots,n_i$. Take samples of size n_i with replacement from $\hat{F}_i$. That is, let $\{X_{ij}^*\}$ be conditionally independent given F, the σ-field spanned by $\{X_{ij}\}$; let X_{ij}^* have common distribution $\hat{F}_i$ for $j = 1,\ldots,n_i$. Let

$$\hat{\gamma}^* = \sum_{i=1}^{p} c_i X_{i\cdot}^*$$

$$s_i^{*2} = (n_i-1)^{-1} \sum_{j=1}^{n_i} (X_{ij}^* - X_{i\cdot}^*)^2$$

$$\hat{\tau}_a^{*2} = \sum_{i=1}^{p} c_i^2 s_i^{*2}/n_i$$

$$\tilde{\tau}_a^2 = \sum_{i=1}^{p} c_i^2 \, (n_i-1)\, s_i^2/n_i^2$$

THEOREM 2. *If* (2) *and* (4) *hold in case (a), then the conditional distribution of* $(\hat{\gamma}^*-\hat{\gamma})/\tilde{\tau}_a$ *converges weakly to* $N(0,1)$ *in probability, and* $\hat{\tau}_a^*/\tilde{\tau}_a$ *converges to* 1 *in probability.*

The proof is deferred. The theorem points to a problem in using the bootstrap to make inferences: the scaling may go wrong. This is because $X^*_{i\cdot}$ has variance $(n_i-1)s_i^2/n_i^2$, not s_i^2/n_i. To fix ideas, suppose there are many small strata: more particularly, that $n_i \leqq k$ for all i. Now

$$\tilde{\tau}_a^2 \leqq (k-1)/k \cdot \hat{\tau}_a^2 \approx (k-1)/k \cdot \tau_a^2$$

The bootstrap distribution of $\hat{\gamma}^* - \hat{\gamma}$ has asymptotic scale $\tilde{\tau}_a$, while $\hat{\gamma} - \gamma$ has the scale τ_a.

We take up next the normal approximation in case (b). Suppose

$$2 \leqq n_i \leqq N_i - 1 \tag{5}$$

Then

$$\tau_b^2 = \sum_{i=1}^p c_i^2 \frac{\sigma_i^2}{n_i} \frac{(N_i-n_i)}{N_i-1}$$

and

$$\hat{\tau}_b^2 = \sum_{i=1}^p c_i^2 \frac{s_i^2}{n_i} \frac{(N_i-n_i)}{N_i}$$

To state the regularity condition, let v_i^2 be the "variance weight" in case (b): $v_i^2 = c_i^2\sigma_i^2(N_i-n_i)/n_i\tau_b^2(N_i-1) = \text{var}\,\{c_iX_{i\cdot}/\tau_b\}$. Let ρ_i be "the effective sample size:" $\rho_i = n_i(N_i-1)/(N_i-n_i)$.

Let $y_i = \{y_{i1},\ldots,y_{iN_i}\}$ where $y_{ij} = (x_{ij}-\mu_{i\cdot})/\sigma_i$ and $\sigma_i^2 = N_i^{-1}\sum_{j=1}^{N_i}(x_{ij}-\mu_{i\cdot})^2$. So $Y_{ij} = (X_{ij}-\mu_i)/\sigma_i$ are sampled from y_i.

The condition is

$$(6) \qquad \sum_{i=1}^{p} N_i^{-1} \sum_{j=1}^{N_i} \phi^2(v_i y_{ij}, \varepsilon\sqrt{\rho_i}) \to 0$$

This may be compared with condition (4).

If $\sup_{1\le i\le p} E|Y_i|^3$ is uniformly bounded independent of the hidden index ν, the Lindeberg conditions (4) and (6) are implied respectively by the natural conditions $\max_i w_i/\sqrt{n_i} \to 0$ or $\max_i v_i/\sqrt{\rho_i} \to 0$. Thus if the standardized populations have reasonably light tails, asymptotic normality holds if for each stratum the variance weight contribution is small or the stratum is heavily sampled.

THEOREM 3. *If* (5) *and* (6) *hold in case (b), then*

i) $$(\hat{\gamma}-\gamma)/\tau_b \to N(0,1) \quad \textit{in law}$$

and

ii) $$\hat{\tau}_b/\tau_b \to 1 \quad \textit{in probability}$$

The proof is deferred.

COROLLARY. $(\hat{\gamma}-\gamma)/\hat{\tau}_b \to N(0,1)$ *in law.*

Finally, we consider the bootstrap in case (b). If $N_i/n_i = k_i$ an integer for each i, the natural bootstrap procedure was suggested by Gross (1980): given $\{X_{ij}\}$, to create populations $\hat{X}_i$ consisting of k_i copies of each X_{ij} for $j=1,\ldots,n_i$; then X^*_{ij} for $j=1,\ldots,n_i$ are generated as a sample without replacement from $\hat{X}_i$; the samples being independent for different $i=1,\ldots,p$. In general, if $N_i = k_i n_i + r_i$ with $0 \leqq r_i < n_i$, form populations $\hat{X}_{i0}$ and $\hat{X}_{i1}$, where $\hat{X}_{i0}$ consists of k_i copies of each X_{ij}, for $j = 1,\ldots,n_i$; while X_{i1} consists of k_i+1 copies. Let

$$\alpha_i = (1 - \frac{r_i}{n_i})(1 - \frac{r_i}{N_i-1})$$

With probability α_i, let $(X^*_{i1},\ldots,X^*_{in_i})$ be a sample without replacement of size n_i from $\hat{X}_{i0}$; with probability $1-\alpha_i$, let $(X^*_{i1},\ldots,X^*_{in_i})$ be a sample without replacement of size n_i from $\hat{X}_{i1}$. The virtue of this scheme is that both $\hat{X}_{i0}$ and $\hat{X}_{i1}$ have the same distribution $\hat{F}_i$ and

$$\mathrm{Var}(X^*_{i\cdot}|\{X_{ij}\}) = \frac{n_i-1}{n_i^2}\, s_i^2(\frac{N_i-n_i}{N_i-1})$$

The proof of the following theorem is similar to that of Theorem 2 and is omitted. Define $\hat{\gamma}^*$ as before, and $\hat{\tau}_b^{*2}$ by substituting X^*_{ij} for X_{ij} in $\hat{\tau}_b^2$.

THEOREM 4. *Let* $\tilde{\tau}_b^2$ *be the variance of* $\hat{\gamma}^*$ *given the data. Then, if* (5) *and* (6) *hold in case* (b), *the conditional distribution of* $(\hat{\gamma}^*-\hat{\gamma})/\tilde{\tau}_b$ *converges weakly to* $N(0,1)$ *and* $\hat{\tau}^*_b/\tilde{\tau}_b \to 1$ *in probability.*

The same inference problem arises as in the case of Theorem 2. The variance of $\hat{\gamma}^*$ given the data is an inconsistent estimate of the variance of $\hat{\gamma}$. We have side-stepped the issue by computing the scale externally to the bootstrap process. Other patches could be made: one is to rescale the elements of X_i ; another is to adjust the constants c_i . These fixes are all a bit ad hoc.

If γ stays bounded as $\nu \to \infty$ our results extend easily to pivots

$$\frac{g(\hat{\gamma}) - g(\gamma)}{g'(\gamma)\hat{\tau}_b}$$

where g is nonlinear continuously differentiable. The same issue as before arises a fortiori for nonlinear functions. Neither the variance of $g(\hat{\gamma}^*)$ given the data nor its natural approximation $[g'(\hat{\gamma})]^2 \tilde{\tau}_b^2$ are consistent estimates of the asymptotic variance of $g(\hat{\gamma})$. A fix which works if $\sum_{i=1}^{p} |c_i \mu_i|$ stays bounded is as before to rescale the elements of X_i or the c_i before applying the bootstrap. Alternatives (the jackknife, linearization, BRR) are discussed in Krewski and Rao (1981). For the case of one stratum, Theorem 4 was derived independently by Chao and Lo (1983).

The bootstrap can work even when Theorem 4 fails but the circumstances are artificial. Suppose we have only one stratum and $N_1 - n_1 = k$ for all ν i.e., all but k members are sampled. Since $\Sigma_{j=1}^{N_1}(x_{1j}-\mu_1) = 0$, the pivot $(\hat{\gamma}-\gamma)/\tau_b$ is distributed as the standardized mean of a sample of size k taken without replacement from the population $\mathcal{Y}_1$. No matter how large N_1 is, if k is small and $\mathcal{Y}_1$ nonnormal we would not expect the normal approximation to apply to $\hat{\gamma}$. To be specific let F_ν be the uniform distribution on $\mathcal{Y}_1$ and suppose

(7) F_ν converges to F in the Mallows d_2-metric,

i.e., $F_\nu \to F$ weakly and $\int x^2 dF_\nu \to \int x^2 dF$. Then $(\hat{\gamma}-\gamma)/\tau_b$ is distributed in the limit as the standardized mean of k independent variables identically distributed according to F. On the other hand since we have sampled nearly the whole population we expect the bootstrap to work.

THEOREM 5. *If* (7) *holds the conditional distribution of* $(\hat{\gamma}^*-\hat{\gamma})/\tilde{\tau}_b$ *converges weakly in probability to the same limit as that of the unconditional distribution of* $(\hat{\gamma}-\gamma)/\tau_b$. *Moreover,* $\hat{\tau}_b/\tau_b$ *and* $\hat{\tau}^*_b/\tau_b$ *both tend to* 1 *in probability.*

We can extend this result somewhat by replacing (7) with a compactness-in-d_2 condition on $\{F_\nu\}$

$$\overline{\lim_{m\to\infty}}\ \overline{\lim_{\nu}}\ N_1^{-1} \Sigma_{j=1}^{N_1} \phi^2(v_1 y_{1j}, m) = 0$$

This condition is evidentally weaker than (6) for $p = 1$. The conclusion now is that the d_2-distance between the conditional distribution of $(\hat{\gamma}^* - \hat{\gamma})/\hat{\tau}_b^*$ and the unconditional distribution of $(\hat{\gamma} - \gamma)/\hat{\tau}_b$ tends in probability to 0. A further extension to an arbitrary number of strata which includes both Theorems 4 and 5 is also possible but not worthwhile.

2. Some lemmas

Recall the truncation operator ϕ from section 1.

LEMMA 1. a) $|\phi(\frac{1}{k}\sum_{i=1}^k y_i, \varepsilon)| \leqq \sum_{i=1}^k |\phi(y_i, \varepsilon/k)|$; equivalently,
$|\phi(\sum_1^k y_i, \varepsilon)| \leqq k\sum_{i=1}^k |\phi(y_i, \varepsilon/k^2)|$

b) *Let* $Y_1, Y_2, \ldots$ *be independent and identically distributed. Then* $E\{\phi^2(\frac{1}{k}\sum_{i=1}^k Y_i, \varepsilon)\} \leqq k^2 E\{\phi^2(Y_i, \varepsilon/k)\}$

Proof. *Claim* a). As is easily verified,

$$|\phi(\frac{1}{k}\sum_{i=1}^k y_i, \varepsilon)| \leqq \phi(\frac{1}{k}\sum_{i=1}^k |y_i|, \varepsilon)$$

Without loss of generality, suppose all $y_i \geqq 0$. Let $y_{(k)}$ be the largest y_i. If $y_{(k)} < \varepsilon/k$, both sides of the inequality vanish. If $y_{(k)} \geqq \varepsilon/k$, the left side is the average of the y_i, or zero; the right side is at least the maximum $y_{(k)}$.

Claim b) follows by the Cauchy-Schwarz inequality. □

LEMMA 2. *Let* $(X_1', \ldots, X_n')$ *and* $(X_1, \ldots, X_n)$ *be distributed respectively as samples with and without replacement from a finite population. Then*

$$E\{\phi^2(\sum_{i=1}^{n} X_i,\varepsilon)\} \leqq E\{\phi^2(\sum_{i=1}^{n} X_i',\tfrac{1}{2}\varepsilon)\}$$

Proof. By a theorem of Hoeffding (1963), if g is convex, then

$$E\{g(\sum X_i)\} \leqq E\{g(\sum X_i')\}$$

Let

$$\begin{aligned} g(x,\varepsilon) &= x^2 && \text{for } |x| \geqq \varepsilon \\ &= 2\varepsilon|x| - \varepsilon^2 && \text{for } \tfrac{1}{2}\varepsilon \leqq |x| \leqq \varepsilon \\ &= 0 && \text{otherwise.} \end{aligned}$$

Then g is convex and

$$\phi^2(x,\varepsilon) \leqq g(x,\varepsilon) \leqq \phi^2(x,\tfrac{1}{2}\varepsilon)$$

So

$$E\{\phi^2(\sum X_i,\varepsilon)\} \leqq E\{g(\sum X_i,\varepsilon)\} \leqq E\{g(\sum X_i',\varepsilon)\} \leqq E\{\phi^2(\sum X_i',\tfrac{1}{2}\varepsilon)\}$$ □

The next result involves the Mallows metric d_2; see Mallows (1972) or Bickel and Freedman (1981).

LEMMA 3. *Let* X *and* Y *be two finite populations of real numbers, of the same size* N. *Let* F *and* G *be the uniform distributions on* X *and* Y. *Suppose* F *and* G *have the same means. Let* $X_1,\ldots,X_n$ *be a sample of size* n, *drawn at random without replacement from* X; *let* $F_{(n)}$ *be the law of* $X_1+\cdots+X_n$. *Likewise for* $Y_1,\ldots,Y_n$ *and* $G_{(n)}$. *Then*

$$d_2[F_{(n)},G_{(n)}]^2 \leqq \frac{n(N-n)}{N-1}\, d_2(F,G)^2$$

Proof. Enumerate $\mathcal{X}$ as $x_1 \leqq x_2 \leqq \cdots \leqq x_N$ and $\mathcal{Y}$ as $y_1 \leqq y_2 \leqq \cdots \leqq y_N$. Then

$$\frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2 = d_2(F,G)^2$$

This follows from Bickel and Freedman (1981, Lemmas 8.2 and 8.3). Let $Z = \{1,\ldots,N\}$. Let $Z_1,\ldots,Z_n$ be a sample of size n, drawn at random without replacement from Z. Set $X_i = x_{Z_i}$ and $Y_i = y_{Z_i}$. Now

$$\begin{aligned} d_2[F_{(n)},G_{(n)}]^2 &\leqq E\{[\textstyle\sum_{i=1}^{n}(X_i-Y_i)]^2\} \\ &= \frac{n(N-n)}{N-1} E[(X_i-Y_i)^2] \\ &= \frac{n(N-n)}{N-1} d_2(F,G)^2 \end{aligned}$$

□

Here is an easy generalization of Lemma 3.

LEMMA 4. *For* $i = 0,1$ *let* $\mathcal{X}_i = \{x_{i1},\ldots,x_{iN_i}\}$ *be finite populations and* F_i *the associated uniform distributions on* $\mathcal{X}_i$. *Let* F_{ni} *be the distribution of* $\sum_{j=1}^{n} X_j$ *when* $X_1,\ldots,X_n$ *is a sample without replacement from* $\mathcal{X}_i$. *Let* $n \leqq N_0 \leqq N_1$. *If* J *is a subset of* $\{1,\ldots,N_1\}$, *let* F_{1J} *be the uniform distribution on* $\{x_{1j}: j \in J\}$. *Then,*

$$d_2(F_{n0},F_{n1})^2 \leq \frac{n(N_0-n)}{N_0-1} \frac{1}{\binom{N_1}{N_0}} \sum_J \{d_2(F_0,F_{1J})^2 : |J| = N_0\}$$

LEMMA 5. *For* $\nu \geq 1$ *let* X_ν *be a finite population of size* N_ν, F_ν *the uniform distribution on* $X_\nu, X_1, \ldots, X_{n_\nu}$ *, a sample without replacement from* X_ν, $\hat{F}_\nu$ *the empirical df. of the sample. If for some* F, $d_2(F_\nu, F) \to 0$ *as* $\nu \to \infty$ *and* $n_\nu \to \infty$ *then* $d_2^2(\hat{F}_\nu, F) \to 0$ *in probability.*

Proof. If g is continuous and bounded

$$E \int g(x) d\hat{F}_\nu(x) = \int g(x) dF_\nu(x) \to \int g(x) dF(x) , \quad \operatorname{Var}\left(\int g(x)\, d\hat{F}_\nu(x)\right) \to 0$$

So,

$$\int g(x) d\hat{F}_\nu(x) \to \int g(x) dF(x) \tag{8}$$

in probability. Moreover,

$$\overline{\lim}_\nu E \int \phi(x,M)^2 \, d\hat{F}_\nu(x) = \int \phi(x,M)^2 dF(x)$$

by lemma 8.3c) of Bickel and Freedman (1981). Since we can make $\int \phi(x,M)^2 dF(x)$ small for M large we conclude that (8) holds for $g(x) = x^2$ also and the lemma follows. □

3. Proving the theorems in case (a)

Proof of Theorem 1. Recall the variance weights w_i from section 1. As is easily verified, $\hat{\tau}_a^2/\tau_a^2 = 1+\xi-\zeta$, where

(9a) $$\xi = \sum_{i=1}^{p} w_i^2(n_i-1)^{-1} \sum_{j=1}^{n_i}(Y_{ij}^2-1)$$

(9b) $$\zeta = \sum_{i=1}^{p} w_i^2(n_i-1)^{-1}(n_i Y_{i\cdot}^2-1)$$

To prove the theorem, it is enough to show that ξ and ζ are both small. But $\xi = \xi_1+\xi_2$, where

(10a) $$\xi_1 = \sum_{i=1}^{p}(n_i-1)^{-1} \sum_{j=1}^{n_i}\{[\bar{\phi}^2(w_i Y_{ij},\varepsilon\sqrt{n_i}) - E\{\bar{\phi}^2(w_i Y_{ij},\varepsilon\sqrt{n_i})\}]$$

(10b) $$\xi_2 = \sum_{i=1}^{p}(n_i-1)^{-1} \sum_{j=1}^{n_i}[\phi^2(w_i Y_{ij},\varepsilon\sqrt{n_i}) - E\{\phi^2(w_i Y_{ij},\varepsilon\sqrt{n_i})\}]$$

Now

$$\begin{aligned}
E(\xi_1^2) &= \operatorname{var} \xi_1 \\
&= \sum_{i=1}^{p}(n_i-1)^{-2} \sum_{j=1}^{n_i} \operatorname{var}\{\bar{\phi}^{2}(w_i Y_{ij},\varepsilon\sqrt{n_i})\} \\
&\leqq \sum_{i=1}^{p}(n_i-1)^{-2} n_i E\{\bar{\phi}^4(w_i Y_{ij},\varepsilon\sqrt{n_i})\} \\
&\leqq \varepsilon^2 \sum_{i=1}^{p}(n_i-1)^{-2} n_i^2 E\{\bar{\phi}^2(w_i Y_{ij},\varepsilon\sqrt{n_i})\} \\
&\leqq \varepsilon^2 \sum_{i=1}^{p}(n_i-1)^{-2} n_i^2 w_i^2 E\{Y_{ij}^2\} \\
&\leqq 4\varepsilon^2 \sum_{i=1}^{p} w_i^2 \\
&= 4\varepsilon^2
\end{aligned}$$

On the other hand, $E\{|\xi_2|\} \to 0$ for each $\varepsilon > 0$, by (4). This disposes of ξ.

The term ζ in (9b) can be decomposed according as to whether $n_i > M$ or $n_i \leqq M$. Since

$$\sum_i \{(n_i-1)^{-1} w_i^2 \colon n_i \geqq M+1\} \leqq M^{-1}$$

and $E\{n_i Y_{i\cdot}^2\} = 1$, the strata i with $n_i \geqq M+1$ are negligible. For the i with $n_i \leqq M$, $\zeta = \zeta_1 + \zeta_2$ where

(11a) $$\zeta_1 = \sum_i \frac{n_i}{n_i-1}[\bar{\phi}^2(w_i Y_{i\cdot}, \varepsilon\sqrt{n_i}) - E\{\bar{\phi}^2(w_i Y_{i\cdot}, \varepsilon\sqrt{n_i})\}]$$

(11b) $$\zeta_2 = \sum_i \frac{n_i}{n_i-1}[\phi^2(w_i Y_{i\cdot}, \varepsilon\sqrt{n_i}) - E\{\phi^2(w_i Y_{i\cdot}, \varepsilon\sqrt{n_i})\}]$$

The sums need be extended only over i with $2 \leqq n_i \leqq M$. Now whatever n_i may be, as for ξ_1,

(12) $$E\{\zeta_1^2\} \leqq 4\varepsilon^2$$

is small. Next,

(13) $$E\{|\zeta_2|\} \leqq 2 \sum_i \frac{n_i}{n_i-1} E\{\phi^2(w_i Y_{i\cdot}, \varepsilon\sqrt{n_i})\}$$

$$\leqq 4M^2 \sum_i E\{\phi^2(w_i Y_{ij}, \varepsilon\sqrt{n_i}/M)\}$$

because $2 \leqq n_i \leqq M$; see Lemma 1. So ζ_2 is small too, by condition (4). □

Proof of Theorem 2. The Lindeberg condition is applied, given F. It is enough to check that for every $\varepsilon > 0$,

(14) $$\tilde{\tau}_a^{-2} \sum_{i=1}^p n_i^{-1} c_i^2 E\{\phi^2(X_{ij}^* - X_{i\cdot}, \varepsilon n_i \tilde{\tau}_a |c_i|^{-1}) | F\} \to 0$$

in probability, where $\tilde{\tau}_a^2 = \sum_{i=1}^p c_i^2(n_i-1)s_i^2/n_i^2$ is the conditional variance of $\hat{\gamma}^*$ given F. For then, Theorem 1 can be applied to X_{ij}^*.

Since $n_i \geqq 2$,

$$\frac{1}{2}\hat{\tau}_a \leqq \tilde{\tau}_a \leqq \hat{\tau}_a \tag{15}$$

Thus $\hat{\tau}_a$ and hence τ_a may be substituted in (14) for $\tilde{\tau}_a$. So (14) reduces to

$$\tau_a^{-2} \sum_{i=1}^p c_i^2 n_i^{-2} \sum_{j=1}^{n_i} \phi^2(X_{ij}-X_{i.}, \varepsilon n_i \tau_a |c_i|^{-1}) \to 0$$

in probability. This in turn reduces to

$$\sum_{i=1}^p n_i^{-1} \sum_{j=1}^{n_i} \phi^2[w_i(X_{ij}-X_{i.})/\sigma_i, \varepsilon\sqrt{n_i}] \to 0 \tag{16}$$

in probability.

Now $(X_{ij}-X_{i.})/\sigma_i = Y_{ij}-Y_{i.}$. Use Lemma 1a) with $k = 2$ to see that (16) follows from (17) and (18):

$$\sum_{i=1}^p n_i^{-1} \sum_{j=1}^{n_i} \phi^2(w_i Y_{ij}, \tfrac{1}{4}\varepsilon\sqrt{n_i}) \to 0 \quad \text{in probability} \tag{17}$$

and

$$\sum_{i=1}^p \phi^2(w_i Y_{i.}, \tfrac{1}{4}\varepsilon\sqrt{n_i}) \to 0 \quad \text{in probability} \tag{18}$$

Clearly, (17) follows from (4). We bound the expected value of the left side of (18). Take first those i with $n_i \leq M$. In view of Lemma 1b), the sum over such i is bounded above by

$$M^2 \sum_i E\{\phi^2(w_i Y_{ij}, \tfrac{1}{4}\varepsilon\sqrt{n_i}/M)\}$$

which tends to zero by condition (4). Take next those i with $n_i > M$. The sum over such i is bounded above by

$$\begin{aligned}\sum_i E\{(w_i Y_{i.})^2\} &= \sum_i w_i^2 n_i^{-1} \\ &< M^{-1} \sum_i w_i^2 \\ &\leqq M^{-1}\end{aligned}$$

which is small for M large.

That $\tilde{\tau}_a^*/\tilde{\tau}_a \to 1$ follows from Theorem 1. □

REMARKS. (i) The Lindeberg-Feller theorem can be supplemented by direct bounds generalizing those of Berry-Esseen; see Petrov (1975, Theorem 3, p.111 or Theorem 8, p.118). These bounds may give estimates on the discrepancy between the bootstrap distribution and the true distribution.

(ii) The difference between the distribution of $(\hat{\gamma}-\gamma)/\tau_a$ and the bootstrap distribution of $(\hat{\gamma}^*-\hat{\gamma})/\tilde{\tau}_a$ can be estimated using the Mallows metric as in equation (2.2) of Bickel and Freedman (1981). The condition needed to push this through is stronger than (4).

(iii) The results can be extended in an obvious way to vector X_{ij}, and under further conditions to nonlinear statistics such as $\sum_{i=1}^p [g_i(X_{i.})-g_i(\mu_i)]$; this covers ratio estimates.

4. Proving the theorems in case (b)

Proof of Theorem 3. The Lindeberg-Feller theorem does not apply to give us i) directly here, since the X_{ij} are dependent for fixed i; however, essestially the same ideas can be used. The proof we give is a bit complicated; an alternative but we believe no simpler approach is, given by Dvoretzky (1971). Our argument is by cases, and the focus is on asymptotic normality. Without loss of generality, assume $\mu_i \equiv 0, c_i \equiv 1$. In outline, the argument is as follows.

Case 1: there is only one stratum, and $n \leqq \frac{1}{2}N$; we drop the unnecessary stratum subscript i. Then ρ^2 is of order n, and asymptotic normality follows from Erdös-Renyi (1959). Also see Rosén (1967), Dvoretzky (1971).

Case 2: there is only one stratum, and $n > \frac{1}{2}N$. Apply Case 1 to the "co-sample" consisting of the objects not in the sample.

Case 3: the number of strata is bounded; no variance weight tends to zero. Case 1 or Case 2 applies to each statum individually.

Case 4: there are many strata, each of small variance weight; in each stratum, $n_i \leqq \frac{1}{2}N_i$. Then $\hat{\gamma}/\tau_b$ is the sum of p independent u.a.n. summands: $\text{var}\,\{X_{i\cdot}/\tau_b\} = v_i^2$ being uniformly small by assumption. We must verify the Lindeberg condition on $X_{i\cdot}/\tau_b$, and do so by an indirect argument. Let X'_{ij} be sampled with replacement from X_i. And let

$$\hat{\gamma}' = \sum_{i=1}^{p} \frac{1}{n_i} \sum_{j=1}^{n_i} X'_{ij}$$

Since $n_i \leq \frac{1}{2}N_i$, the variance weights v_i^2 and w_i^2 are of the same order, as are the total variances τ_a^2 and τ_b^2. In particular, condition (6) implies (4). Thus, the Lindeberg condition holds for the individual summands in $\hat{\gamma}'/\tau_a$, viz., $X'_{ij}/n_i\tau_a$, and asymptotic normality of $\hat{\gamma}'$ follows. By the converse to Lindeberg's theorem, his condition holds for the stratum averages $\frac{1}{n_i}\sum_{j=1}^{n_i} X'_{ij}/\tau_a$. Hence, by Lemma 2, the condition holds for the stratum averages taken without replacement, viz., $\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}/\tau_b$. Now a second application of the direct Lindeberg theorem gives asymptotic normality of $\hat{\gamma}$.

Case 5: there are many strata, each of small variance weight; on each stratum, $n_i > \frac{1}{2}N_i$. Apply Case 4 to the co-samples.

Case 6: there are many strata, each of small variance weight. Consider two groups of strata: in the first, $n_i \leq \frac{1}{2}N_i$; in the second, $n_i > \frac{1}{2}N_i$. Case 4 applies to the first group, Case 5 to the second. (One of the two groups may be negligible.)

The general case: We combine cases 3 and 6. Let

$$J_k(\nu) = \{i : v_i \geq \tfrac{1}{k}\}$$

$$V_k(\nu) = \Sigma\{v_i^2 : i \in J_k(\nu)\}$$

where dependence on the hidden index is made explicit. Given any subsequence of $\{\nu\}$ we can extract a subsubsequence $\{\nu_r\}$ such that for all k, as $r \to \infty$, $V_k(\nu_r)$ tends to a finite limit ν_k. If $V_k = 0$ for all k, there must be $k_r \to \infty$ such that $V_{k_r}(\nu_r) \to 0$. Hence, as $r \to \infty$,

(19) $$\Sigma\{X_{i\cdot}/\tau_b : i \in J_{k_r}(\nu_r)\} \to 0 \text{ in probability.}$$

But, $\max\{v_i : i \notin J_{k_r}(\nu_r)\} \leqq 1/k_r \to 0$. So we can apply case 6 to get that

(20) $$\Sigma\{X_{i\cdot}/\tau_b : i \notin J_{k_r}(\nu_r)\} \text{ is asymptotically } N(0,1).$$

Combining (19) and (20), we get

(21) $$\Sigma\, X_{i\cdot}/\tau_b \text{ is asymptotically } N(0,1), \text{ as } r \to \infty.$$

On the other hand, suppose $V_k > 0$ for some k. Since $J_k(\nu_r)$ has at most k^2 members, we can apply case 3 to see that for all k, as $r \to \infty$,

$$\Sigma\{X_{i\cdot}/\tau_b : i \in J_k(\nu_r)\} \text{ is asymptotically } N(0,V_k)$$

By a standard argument, there are $k_r \to \infty$ such that

(22) $$\Sigma\{X_{i\cdot}/\tau_b : i \in J_{k_r}(\nu_r)\} \text{ is asymptotically } N(0, \sup_k V_k).$$

Applying case 6 as above,

(23) $$\Sigma\{X_{i\cdot}/\tau_b : i \notin J_{k_r}(\nu_r)\} \text{ is asymptotically } N(0,1 - \sup_k V_k).$$

Combining (22) and (23) we obtain (21) in this case also. Part (i) of the theorem follows by a standard compactness argument. The proof of (ii) follows the pattern of that of Theorem 1 and is omitted. □

Proof of Theorem 5: We simplify the argument by supposing n_1 divides N_1 so we can use the naive bootstrap. (The general argument uses lemma 4.) Moreover, without loss of generality let $\mu_1 = 0$, $\sigma_1 = 1$. Since $p = 1$ we want to compare the distributions of standardized means of a sample size n_1 from the populations $\mathcal{Y}_1$ and that composed of N_1/n_1 copies of the standardized sample: $(X_{ij}-\hat{\mu}_1)/\hat{\sigma}_1$, $1 \leqq j \leqq n_1$, where $\hat{\mu}_1$ are the sample mean and sample standard deviation respectively. So by lemma 3,

$$d_2^2\{\mathscr{L}(\frac{\hat{\gamma}-\gamma}{\tau_b}), \mathscr{L}(\frac{\hat{\gamma}^*-\hat{\gamma}}{\tilde{\tau}_b})|X_{1j}\ , 1 \leqq j \leqq n_1\} \leqq d_2^2\{F_\nu, \hat{F}_\nu(\hat{\sigma}_1 x+\hat{\mu}_1)\} \ .$$

By lemma 5, $d_2^2(F_\nu,\hat{F}_\nu)$, $\hat{\mu}_1$, and $\hat{\sigma}_1 - 1$ all tend in probability to 0 as $\nu \to \infty$. A truncation argument of the type we have all ready used shows that $\hat{\tau}_b/\tau_b$ and $\hat{\tau}_b^*/\hat{\tau}_b$ both tend in probability to 1. The theorem follows. □

REFERENCES

Babu, G. J. and Singh, K. (1983). Inference on means using the bootstrap. *Ann. Statist.* **11**, 999-1003.

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

Bickel, P. J. and Krieger, A. (1983). Using the bootstrap to set confidence bands for a distribution function: Monte Carlo and some theory. In preparation.

Chao, M. T. and Lo, S. H. (1983). A bootstrap method for finite population. Preprint.

Dvoretzky, A. (1971). Asymptotic normality for sums of dependent random variables. *Proc. Sixth Berkeley Symp.* II (Ed: Le Cam, Neyman, Scott) Univ. of Calif. Press, Berkeley, p 513.

Erdös, P. and Renyi, A. (1959). On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hung. Acad. Sci.* **4**, 49-61.

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

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13-30.

Krewski, D. and Rao, J. N. K. (1981). Inference from stratified samples: properties of linearization, jackknife, and balanced repeated replication. *Ann. Statist.* **9**, 1010-1019.

Mallows, C. (1972). A note on asymptotic joint normality. *Ann. Math. Statist.* **43**, 508-515.

Petrov, V. (1975). Sums of Independent Random Variables. Springer-Verlag. Berlin-Heidelberg-New York.

Rosén, B. (1967). On the central limit theorem for sums of dependent random variables. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **7**, 48-82.