Ann. Inst. Statist. Math.
Vol. 44, No. 4, 703–720 (1992)

SUBSAMPLE AND HALF-SAMPLE METHODS*

GUTTI JOGESH BABU

Department of Statistics, 319 Classroom Building, The Pennsylvania State University, University Park, PA 16802, U.S.A.

(Received October 31, 1990; revised November 18, 1991)

Abstract. Hartigan's subsample and half-sample methods are both shown to be inefficient methods of estimating the sampling distributions. In the sample mean case the bootstrap is known to correct for skewness. But irrespective of the population, the estimates based on the subsample method, have skewness factor zero. This problem persists even if we take only samples of size less than or equal to half of the original sample. For linear statistics it is possible to correct this by considering estimates based on subsamples of size λn, when the sample size is n. In the sample mean case λ can be taken as $0.5(1-1/\sqrt{5})$. In spite of these negative results, the half-sample method is useful in estimating the variance of sample quantiles. It is shown that this method gives as good an estimate as that given by the bootstrap method. A major advantage of the half-sample method is that it is shown to be robust in estimating the mean square error of estimators of parameters of a linear regression model when the errors are heterogeneous. Bootstrap is known to give inconsistent results in this case; although, it is more efficient in the case of homogeneous errors.

Key words and phrases: Half-sample method, bootstrap, variance estimation, linear models, asymptotic relative efficiency, Bahadur's representation, quantiles.

1. Introduction

Hartigan (1969, 1975) has suggested using the subsample values of an estimator of a parameter as indicators of variability of the estimator. He has shown that asymptotically valid confidence statements can be made under fairly general conditions by choosing subsamples randomly, without replacement, from all the possible non-empty subsamples of the original sample.

The subsample methods have been in use in the literature for a long time. Mahalanobis (1946) used it under the name, "interpenetrating samples". Efron (1979) discusses Hartigan's work (see his Remark I on p. 24). While comparing the subsample method with the bootstrap, observing the first order asymptotic

* Research supported in part by NSA Grant MDA904-90-H-1001 and by NASA Grant NAGW-1917.

equivalence of the methods in several cases, Efron notes that there is no evidence to prefer the one method over the other. When the subsample size is fixed, the estimator is popularized by Wu (1986, 1990) under the name "delete-k jackknife". See also Shao and Wu (1989).

Even though both the half-sample and the subsample methods are less computationally intensive than the bootstrap, these techniques have been overshadowed by the popular bootstrap method. In some situations like sample surveys, where one deals with finite populations, conceptually the subsample method has distinct advantages over the bootstrap. On the other hand, Singh (1981) and Babu and Singh (1983, 1984a) have established that, for a wide class of statistics, bootstrap automatically corrects for skewness of the sampling distribution, thereby giving a better performance than the classical normal approximation.

The subsample and the half-sample methods are described in the next section. It is shown that both the half-sample and the subsample estimators of the sampling distribution of the sample mean are asymptotically symmetric. This holds whether or not the underlying population distribution is symmetric. As a consequence, these methods cannot correct for the skewness of the sampling distribution. It is also established that this problem can be corrected for linear statistics like sample mean by restricting to subsamples of size λn, where $\lambda = (1 - 1/\sqrt{5})/2$. However, using the results of Babu and Singh (1985) on samples from finite populations, it is established that no such correction for skewness is possible for statistics like t-statistic.

In spite of these negative results, the subsample method is useful in estimating the variance of estimators, even when methods like the ordinary jackknife fails. In the i.i.d. case Shao and Shi (1989) have established the superiority of the half-sampling method to jackknife in some situations. When the observations are independent but not identically distributed, the relative asymptotic efficiency of the bootstrap and the half-sample methods in estimating the asymptotic variance of sample quantiles is shown to be one in Section 5. By asymptotic efficiency, we mean the ratio of the variances of the estimators.

Bahadur's representation of quantiles for the half-sample method in general non-stationary case is established in Section 4. Strong representation for the half-sample variance estimator of the least squares estimate of the regression parameter is established in Section 6. The variance estimator is shown to be $\sqrt{n}$ consistent in the case of heterogeneous errors, whereas the bootstrap estimate is generally known to be inconsistent. In this connection, among others, Freedman's comments in the discussion of Wu (1986) are noteworthy. Of course there have been several variations of the standard bootstrap method in the literature, to suit specific situations to force a consistent result (see Babu (1984)). The main point here is that the subsample method, as a general method, is $\sqrt{n}$ consistent in a variety of situations.

In general, the half-sample estimate is asymptotically as efficient as the jackknife estimate, in the sense that the ratio of their variances tends to one. In the case of L_1-estimators the jackknife is known to give inconsistent variance estimators even in the homogeneous case. However, it should not be difficult to establish, as in the special case of quantiles, the $\sqrt{n}$ consistency of the half-sample variance

estimators for the L_1-estimates of the regression coefficients. For recent results on L_1-estimation (see Babu (1989) and for a review see Rao (1988)).

The messy estimates needed in the proofs are established in the Appendix.

2. Description of the subsample and bootstrap methods

Let $X_1, \ldots, X_n$ be a sample from a population. Let γ be a parameter to be estimated by $\hat{\gamma} = \hat{\gamma}(X_1, \ldots, X_n)$. Let S denote the class of all 2^n subsets of $\{1, 2, \ldots, n\}$. For any non-empty $s \in S$, let $\hat{\gamma}_s$ denote the estimate based on $\{X_i : i \in s\}$. Let $\hat{\gamma}_\phi = 0$. The empty set is included only for notational convenience. Exclusion of the empty set does not affect the asymptotic results considered in this paper, as the contribution of any single subset s in the estimation of the sampling distribution is $O(2^{-n})$. An example here would help in fixing the general ideas. If $\hat{\gamma} = \bar{X} = (1/n)\sum_{i=1}^n X_i$, then for any non-empty $s \in S$, $\hat{\gamma}_s = \bar{X}_s = |s|^{-1}\sum_{i \in s} X_i$, where $|s|$ denotes the number of elements in s. The subsample estimate of the sampling distribution of $\sqrt{n}(\hat{\gamma} - \gamma)$ is given by the histogram D_S of $\sqrt{n}(\hat{\gamma}_s - \hat{\gamma})$. That is

$$D_S(x) = \#\{s \in S : \sqrt{n}(\hat{\gamma}_s - \hat{\gamma}) \leq x\}/|S|. \tag{2.1}$$

The subsample estimate of the mean square error of $\sqrt{n}(\hat{\gamma} - \gamma)$ is given by

$$V_S = n|S|^{-1}\sum_{s \in S}(\hat{\gamma}_s - \hat{\gamma})^2. \tag{2.2}$$

Note that $|S| = 2^n$.

The half-sample method consists of considering, instead, the class of all subsets $H \subset S$ of size $[n/2]$ from $\{1, 2, \ldots, n\}$ and defining D_H and V_H as in (2.1) and (2.2), but with S replaced by H. Note that $|H| = \binom{n}{[n/2]}$.

Let $X_1^*, \ldots, X_n^*$ be a sample drawn with replacement from $X_1, \ldots, X_n$ and let $\gamma^* = \hat{\gamma}(X_1^*, \ldots, X_n^*)$. The bootstrap estimate of the sampling distribution and of the variance of $\sqrt{n}(\hat{\gamma} - \gamma)$ are given by F^* and

$$V_{B,n} = E^*(\sqrt{n}(\gamma^* - \hat{\gamma}))^2,$$

where,

$$F^*(x) = P^*(\sqrt{n}(\gamma^* - \hat{\gamma}) \leq x),$$

and P^* and E^* denote the probability and expectation (given the sample $X_1, \ldots, X_n$) induced by the bootstrap sampling mechanism. Instead of sampling from the empirical distribution one can sample from a smoothend version of the empirical distribution. The same theory goes through.

3. Subsample estimate of the distribution of the sample mean

Let N be a binomial random variable with parameter values n and $1/2$. For any non-negative integer k, let $S_k = \{s \in S : |s| = k\}$. A close examination of D_S reveals that the subsample scheme is equivalent to observing $N = k$ (say) and then choosing at random an $s \in S_k$. Clearly

$$D_S(x) = \sum_{k=0}^{n} P(N = k) G_k(x), \tag{3.1}$$

where G_k denotes the empirical distribution of $\sqrt{n}(\hat{\gamma}_s - \hat{\gamma})$ restricted to $s \in S_k$, that is

$$G_k(x) = \binom{n}{k}^{-1} \sum_{s \in S_k} I(\sqrt{n}(\hat{\gamma}_s - \hat{\gamma}) \le x). \tag{3.2}$$

Since

$$P\left(\left|N - \frac{n}{2}\right| > \sqrt{n}\log n\right) = O(n^{-6}). \tag{3.3}$$

only the subsamples of size k close to $n/2$, would significantly influence the asymptotic properties of D_S. The other samples have negligible effect on the asymptotics of D_S. To study the asymptotics of G_k we require a result from Babu and Singh (1985), which is stated as Theorem 3.1 after introducing some notation below. When the subsamples are restricted to S_k, Wu (1986, 1990) called the estimators based on S_k delete-$(n-k)$ jackknife and studied their asymptotic properties.

Let $Y_1, \ldots, Y_n$ be independent random variables from a common d-variate distribution. Let $y_1, \ldots, y_k$ be a simple random sample drawn, without replacement, from $Y_1, \ldots, Y_n$. Suppose that the distribution of Y_1 is strongly non-lattice, $E|Y_1|^{3+\delta} < \infty$ and $\delta < p_k = k/n < 1 - \delta$ for some $\delta > 0$. Let $l' = (l_1, \ldots, l_d)$ denote the transpose of the vector l. Let $L = ((L_{ij}))$ be a matrix, $\mu = E(Y_1)$, $\bar{Y} = (1/n)\sum_1^n Y_i$, $\Sigma_n = (1/n)\sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})'$, $Z_k = \sqrt{n/\{k(n-k)\}}\sum_{i=1}^k (y_i - \bar{Y})$, $\sigma_n^2 = l'\Sigma_n l$, $c_k = (1 - 2p_k)(p_k(1-p_k))^{-1/2}$, and let Σ denote the dispersion matrix of Y_1. Let ϕ and Φ denote density and distribution functions of a standard normal variable. Then from the proofs of Theorems 1 and 2 of Babu and Singh (1985) we have

THEOREM 3.1. *Let* $\max\{|l_i|, |L_{ij}|, |l_d|^{-1}\}$ *be bounded and*

$$T_k = (l'Z_k + k^{-1/2} Z_k' L Z_k)\sigma_n^{-1} + R_n,$$

where

$$P(\sqrt{n}\,|R_n| > a_n) = o(n^{-1/2})$$

for some sequence $a_n \to 0$. *Then with probability* 1,

$$P(T_k \le y \mid Y_1, \ldots, Y_n) = \Phi(y) - \xi(y)\phi(y)n^{-1/2} + o(n^{-1/2}), \tag{3.4}$$

uniformly in y *and in* $p_k \in (\delta, 1-\delta)$, *where*

$$\xi(y) = \sigma^{-1}\operatorname{tr}(\Sigma L)p_k^{-1/2} + \sigma^{-3}(y^2-1)[l'\Sigma L\Sigma l p_k^{-1/2} + (1/6)c_k E(l'(Y_1-\mu))^3], \tag{3.5}$$

and $\sigma^2 = l'\Sigma l$.

By taking $L_{11} = 0$ and $l_1 = 1$, we obtain

COROLLARY 3.1. *If* $d = 1$, *then with probability* 1,

$$\text{i)}\quad P(Z_k \le y\sigma_n \mid Y_1, \ldots, Y_n) = \Phi(y) - n^{-1/2}\phi(y)(y^2-1)c_k(6\sigma^3)^{-1}E(Y_1-\mu)^3 + o(n^{-1/2}) \tag{3.6}$$

uniformly in y *and in* $p_k \in (\delta, 1-\delta)$. *In addition if* $p_k = 1/2 + o(1)$, *then* $c_k \to 0$ *and consequently,*

$$\text{ii)}\quad P(Z_k \le y\sigma_n \mid Y_1, \ldots, Y_n) = \Phi(y) + o(n^{-1/2}) \quad a.e. \tag{3.7}$$

Note that,

$$E(Z_k^2 \mid Y_1, \ldots, Y_n) = (n/(n-1))\sigma_n^2,$$

so essentially, Z_k/σ_n is a standardized statistic. On the other hand by the Edgeworth expansions in the i.i.d. case

$$P(\sqrt{n}(\bar{Y}-\mu) \le y\sigma) = \Phi(y) - (y^2-1)\phi(y)(6\sigma^3\sqrt{n})^{-1} + o(n^{-1/2}). \tag{3.8}$$

So from (3.6) and (3.8), it follows that

$$\sup_y \sqrt{n}\,|P(\sqrt{n}(\bar{Y}-\mu) \le y\sigma) - P(Z_k \le y\sigma_n \mid Y_1, \ldots, Y_n)| = \sup_y |(y^2-1)|\phi(y)|1 - c_k|(6\sigma^3)^{-1} + o(1) \to 0,$$

only when $c_k \to 1$. This holds only when

$$p_k \to (1 - (1/\sqrt{5}))/2. \tag{3.9}$$

Using (3.3), we have

$$\begin{aligned}
&\sum_{k=1}^{n} |c_k| P(N=k) \\
&\quad \le 2\sum{}' \left|\frac{2k-n}{n}\right| (1 + O(n^{-1/2}\log n))P(N=k) + O(n^{-6}) \\
&\quad \le 2n^{-1}\sum_{k=0}^{n} |2k-n|P(N=k) + O(n^{-1/2}\log n) \\
&\quad \le 2n^{-1}\{E(2N-n)^2\}^{1/2} + O(n^{-1/2}\log n) \\
&\quad = O(n^{-1/2}\log n) = o(1)
\end{aligned} \tag{3.10}$$

where $\sum'$ denotes sum over k satisfying $|2k - n| \leq 2\sqrt{n}\log n$.

Hence, from (3.9), (3.10) and Corollary 3.1 ii) we have

THEOREM 3.2. *If Y_1 has a non-lattice distribution, then for the estimator $\hat{\gamma}$ = sample mean,*

$$\sqrt{n}\,|D_S(y\sigma_n) - \Phi(y)| \to 0$$

and

$$\sqrt{n}\,|D_H(y\sigma_n) - \Phi(y)| \to 0,$$

uniformly in y a.e. But

$$\sqrt{n}\,|D_{S_k}(y\sigma_n) - P(\sqrt{n}(\bar{Y} - \mu) \leq y\sigma)| \to 0, \tag{3.11}$$

uniformly in y a.e. provided $k = [\lambda n]$, where

$$\lambda = (1 - (1/\sqrt{5}))/2.$$

From Theorem 3.2, it is clear that neither the half-sample method nor the subsample method corrects for skewness. Consequently, if $E(Y_1 - \mu)^3 \neq 0$, then the subsample approximation is much worse than that given by the bootstrap. However (3.11) shows that if only the subsamples of size $k = [\lambda n]$ are considered with $\lambda = (1 - 1/\sqrt{5})/2$, then the approximation is as good as the one given by the bootstrap.

If T_k represents the studentized statistic, then under appropriate moment conditions (see Example 1 of Babu and Singh (1985)),

$$\begin{aligned} P(T_k \leq x \mid Y_1, \ldots, Y_n) &= \Phi(x) + E(Y_1 - \mu)^3(6\sigma^3\sqrt{n})^{-1} \\ &\quad \cdot [3x^2 + ((2p_k - 1)/q_k)(x^2 - 1)]\sqrt{q_k/p_k}\,\phi(x) \\ &\quad + o(n^{-1/2}), \end{aligned} \tag{3.12}$$

where $q_k = 1 - p_k$. On the other hand the Edgeworth expansion for t-statistic is given by

$$P(\sqrt{n}(\bar{Y} - \mu) \leq x\sigma_n) = \Phi(x) + E(Y_1 - \mu)^3(6\sigma^3\sqrt{n})^{-1}(2x^2 + 1)\phi(x) + o(n^{-1/2}). \tag{3.13}$$

The only way the coefficients of $(1/\sqrt{n})$ in (3.12) and (3.13) can be matched is by taking

$$q_k/p_k = 1 + o(1) \quad \text{and} \quad (1 - 2p_k)q_k \to 1$$

which is impossible. Consequently subsample estimates based on size $k = [\lambda n]$, for any $\lambda > 0$, cannot correct for skewness for t-statistic. Using the results of Babu and Singh (1985), Wu (1990) obtained the expressions (3.11) and (3.12)

independently. A close observation of Theorem 3.1 shows that this conclusion holds whenever T_k is not linear, that is when $L \neq 0$.

The same argument implies that even if we restrict to half-samples, the skewness factor is not taken care of. On the other hand Singh (1981) and Babu and Singh (1984*a*) have shown that the bootstrap corrects for skewness for a wide class of statistics. In this respect, both the half-sample and the subsample methods are inferior to bootstrap; they do not correct for the second order term.

It may be mentioned here that Hartigan (1969) observes that, "An empirical investigation of eigen vector analysis shows that the confidence statements based on subsample values are not as accurate as the ones based on the standard normal asymptotic methods." On the other hand, in the i.i.d. case, bootstrap confidence intervals are quite accurate (see Abramovitch and Singh (1985), Babu and Bose (1988) and Hall (1988)).

4. Bahadur's representation of sample quantiles

In this section we obtain Bahadur type representations for sample quantiles under both the subsample and the half-sample methods. We do not assume that the observations come from a single population. Occurrence of such independent but not identically distributed observations is common in medical studies. Let $X_1, \ldots, X_n$ be independent random variables with X_i having a continuous distribution H_i. Let $G_n = (1/n)\sum_{i=1}^n H_i$. For any distribution F and $0 < u < 1$, let $F^{-1}(u) = \inf\{x : F(x) \geq u\}$. Let $0 < p < 1$. Suppose G_n has derivative g_n in a neighborhood of $G_n^{-1}(p)$, and $\liminf_{n\to\infty} h_n > 0$, where $h_n = g_n(G_n^{-1}(p))$. Further suppose

$$|g_n(x) - g_n(y)| \leq K|y - x|^\theta, \tag{4.1}$$

for $x, y \in (G_n^{-1}(p) - \epsilon, G_n^{-1}(p) + \epsilon)$, for some $\epsilon > 0$, $K > 1$ and $\theta > 1/2$. For nonempty $s \in S$, let F_s denote the empirical distribution function of $\{X_i : i \in s\}$. Let F_n denote the empirical distribution function of $\{X_1, \ldots, X_n\}$. Let for $s \in S$,

$$\begin{aligned} R_n^{(s)} &= R_n(s, X_1, \ldots, X_n) \\ &= |(F_s^{-1}(p) - F_n^{-1}(p))g_n(F_n^{-1}(p)) - (p - F_s F_n^{-1}(p))| \end{aligned} \tag{4.2}$$

THEOREM 4.1. *Under the above set up for some $A > 0$, we have*

$$E(P_H(R_n(\cdot) > An^{-3/4}\log n \mid X_1, \ldots, X_n)) = O(n^{-4}) \tag{4.3}$$

and

$$E(P_S(R_n(\cdot) > An^{-3/4}\log n \mid X_1, \ldots, X_n)) = O(n^{-4}), \tag{4.4}$$

where P_H and P_S denote probability measures induced by the half-sample and the subsample methods respectively. Further, with probability 1, *we have*

$$(F_n^{-1}(p) - G_n^{-1}(p))h_n - (p - F_n G_n^{-1}(p)) = O(n^{-3/4}\log n). \tag{4.5}$$

From (4.2) and (4.3) we have for almost all samples,

$$R_n = O_{P_H}(n^{-3/4}\log n) \tag{4.6}$$

and

$$R_n = O_{P_S}(n^{-3/4}\log n). \tag{4.7}$$

Here $O_{P_S}(\cdot)$, denotes convergence in P_S probability. Similarly $O_{P_H}(\cdot)$ denotes convergence in P_H probability. The theorem still holds if $g_n(F_n^{-1}(p))$ is replaced by h_n in the definition of R_n.

An expression similar to (4.6) can be shown to hold for the bootstrap. See for example Babu and Singh (1984*b*) for the i.i.d. case. A proof of Theorem 4.1 is given in the Appendix.

Liu and Singh (1989) observed that a useful interpretation exists for the i.i.d. bootstrap confidence intervals even when the data are independent but not identically distributed. In general, even though the variance estimators are inconsistent in non i.i.d. setting, they serve useful purpose in providing lower bounds for the coverage probability. Liu and Singh (1989) have shown that the coverage probability for non-identically distributed case is at least that of i.i.d. case. One can assess the difference between the two situations by using the so called heterogeneity factor. It turns out that the same behavior is exhibited by the half-sample and the subsample methods. Furthermore the same heterogeneity factor appears in all these cases. To see this let I_S denote the subsample confidence interval, with coverage probability $(1-2\alpha)$ under the i.i.d. setting (i.e. one obtained by assuming that X_i are i.i.d. from G_n). Then using the Bahadur type representation (Theorem 4.1) it can be shown that

$$P_S(G_n^{-1}(p) \in I_S) - [2\Phi(z_{1-\alpha}(1+d_n^2)^{1/2}) - 1] \to 0,$$

for almost all samples for some sequence $\{d_n^2\}$, which can be called a heterogeneity factor. Further note that

$$[2\Phi(z_{1-\alpha}(1+d_n^2)^{1/2}) - 1] \geq 1 - 2\alpha.$$

So the asymptotic coverage probability is at least that of i.i.d. case. For details on heterogeneity factor for other statistics see Liu and Singh (1989). Suppose

$$m_n = \int [\log(1+|x|)]^4 dG_n(x) = O(1), \tag{4.8}$$

then it follows that

$$d_n^2 = (V_n^2 - v_n^2)v_n^{-2}, \tag{4.9}$$

where V_n^2 is the estimator of the asymptotic variance of the p-th sample quantile using the subsample method and v_n^2 is the variance in the non i.i.d. setting. Consistency of the half sample variance estimator (and other estimators based on

S_k) for the sample quantile in the i.i.d. case was shown in Shao and Wu (1989). Clearly by Theorem 4.1,

$$
\begin{aligned}
(4.10)\qquad h_n^2 V_n^2 &= nE_S(p - F_s F_n^{-1}(p))^2 \\
&= \sum_{k=1}^{n} \frac{(n-k)n}{k(n-1)} \binom{n}{k} 2^{-n} \left(\frac{1}{n} \sum_{i=1}^{n} (p - I(X_i \le F_n^{-1}(p)))^2 \right) \\
&= p(1-p) + O(1/n) \quad \text{a.s.}
\end{aligned}
$$

By (4.5) of Theorem 4.1, (see Babu (1986) for the i.i.d. case)

$$
\begin{aligned}
(4.11)\qquad h_n^2 v_n^2 &= nE(F_n(G_n^{-1}(p)) - p)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} E(I(X_i \le G_n^{-1}(p)) - H_i(G_n^{-1}(p)))^2 \\
&= p - \frac{1}{n} \sum_{1}^{n} (H_i(G_n^{-1}(p)))^2.
\end{aligned}
$$

Suppose

$$
(4.12)\qquad \liminf_{n \to \infty} v_n^2 > 0.
$$

From (4.9), (4.10) and (4.11), it follows that the heterogeneity factor d_n^2 is given by

$$
(4.13)\qquad d_n^2 = \frac{1}{n} \sum_{i=1}^{n} (H_i(G_n^{-1}(p)) - p)^2 \left(p - \frac{1}{n} \sum_{1}^{n} (H_i(G_n^{-1}(p)))^2 \right)^{-1} (1 + o(1)).
$$

This shows that the heterogeneity factor is essentially the same for both the subsample method and the bootstrap. Consequently the subsample method is preferable here, as it is less computationally intensive than the bootstrap as explained in Section 3. The same arguments establish that the heterogeneity factor is the same even for the half-sample method.

5. The efficiency of the variance estimator

From now on we concentrate on half-sample estimates. In view of (3.3), similar results hold for subsample estimates. In this section we go one step further and show that the relative asymptotic efficiency of the half-sample variance estimator of a sample quantile compared to that given by bootstrap in one (i.e. the ratio of the variances tends to one), even in the not identically distributed case.

Let $r = [np]$ if np is not an integer and $= np + 1$ otherwise. Let $k = [n/2]$. Let $h = [pk]$ if pk is not an integer and $= 1 + kp$ otherwise. Further let

$$
f_{n,m}(u) = n \binom{n-1}{m-1} u^{m-1} (1-u)^{n-m}
$$

and

$$p_{m,i}(u) = \binom{m}{i} u^i (1-u)^{m-i}.$$

Let p, F_n, F_s, G_n, X_i and h_n be as in Section 4. Clearly for $h \le i \le n-k+h$,

$$\begin{aligned} P_H\left\{F_s^{-1}(p) = F_n^{-1}\left(\frac{i}{n}\right) \;\middle|\; X_1, \ldots, X_n\right\} \\ &= \binom{n}{k}^{-1} \#\left\{s : |s| = k, F_s^{-1}(p) = F_n^{-1}\left(\frac{i}{n}\right)\right\} \\ &= \binom{n}{k}^{-1} \binom{i-1}{h-1}\binom{n-i}{k-h} \\ &= k\binom{k-1}{h-1}\binom{n-k}{i-h}\int_0^1 u^{i-1}(1-u)^{n-i}du \\ &= \int_0^1 f_{k,h}(u) p_{n-k,i-h}(u)du, \end{aligned} \tag{5.1}$$

and

$$P_H\left\{F_s^{-1}(p) = F_n^{-1}\left(\frac{i}{n}\right) \;\middle|\; X_1, \ldots, X_n\right\} = 0$$

otherwise. Now it is straight forward to establish the next theorem, the proof of which is omitted.

THEOREM 5.1. *Assume* (4.8). *Under the setting of Section* 4, *given the sample, the bootstrap and half-sample estimates of the variance of the p-th sample quantiles are given by*

$$V_{B,n} = \int_0^1 (F_n^{-1}(u) - F_n^{-1}(p))^2 f_{n,r}(u)du \tag{5.2}$$

and

$$\begin{aligned} V_{H,n} &= E_H(F_s^{-1}(p) - F_n^{-1}(p))^2 \\ &= \int_0^1 f_{k,h}(u)\left[\sum_{i=h}^{n-k+h}\left(F_n^{-1}\left(\frac{i}{n}\right) - F_n^{-1}(p)\right)^2 p_{n-k,i-h}(u)\right]du \\ &= \binom{n}{k}^{-1}\sum_{i=h}^{n-k+h}\binom{i-1}{h-1}\binom{n-i}{k-h}\left(F_n^{-1}\left(\frac{i}{n}\right) - F_n^{-1}(p)\right)^2. \end{aligned}$$

Maritz and Jarrett (1978) derived an expression similar to (5.2) in the sample median case. Note that by (4.8),

$$\begin{aligned} P\left(\max_{1\le i\le n}|X_i| > \exp(\sqrt{n/2}\log n)\right) &\le 2^4 n^{-1}(\log n)^{-4} m_n \\ &= O(n^{-1}(\log n)^{-4}). \end{aligned}$$

Hence with probability 1,

$$\max_{1\le i\le n} |X_i| \le \exp(\sqrt{n/2}\log n) \tag{5.3}$$

for all large n. The assumption (4.8) can be omitted in the i.i.d. case. Using simple exponential inequality, one can show that in the i.i.d. case, for some $A > 0$,

$$\max\left\{\left|F_n^{-1}\left(\frac{i}{n}\right)\right| : h \le i \le n-k+h\right\} \le A,$$

with probability 1. Using Stirling's formula and (5.3) we conclude that with probability 1,

$$nV_{B,n} = nB_n' + O((\log n)^3 n^{-1/2}) \tag{5.4}$$

and

$$nV_{H,n} = nB_n' + O((\log n)^3 n^{-1/2}), \tag{5.5}$$

where $q = 1 - p$,

$$B_n' = \int_{-\log n}^{\log n} (F_n^{-1}(p + y\sqrt{pq/n}\,) - F_n^{-1}(p))^2 \phi(y)dy.$$

Further, by letting $\theta_n = (pq/nh_n)^{1/2}$ and using Theorem 4.1, we get that

$$\begin{aligned} B_n' = &-2\theta_n \int_{-\log n}^{\log n} (F_n^{-1}(p + y\sqrt{pq/n}\,) - G_n^{-1}(p + y\sqrt{pq/n}\,))y\phi(y)dy \\ &+ \theta_n^2(1 + O(n^{-\theta/2})) + O(n^{-3/2}(\log n)^2), \end{aligned} \tag{5.6}$$

where θ is as in (4.1). Consequently with probability 1,

$$n^{1/4}(nB_n' - n\theta_n^2) = -2\sqrt{pq/h_n}\, B_n + O(n^{-(2\theta-1)/4}), \tag{5.7}$$

where

$$B_n = n^{3/4} \int_{-\log n}^{\log n} (F_n^{-1}(p + y\sqrt{pq/n}\,) - G_n^{-1}(p + y\sqrt{pq/n}\,))y\phi(y)dy. \tag{5.8}$$

We immediately have the following.

THEOREM 5.2. *Under the conditions of Theorem* 5.1, *if* (4.8) *holds and*

$$\lim_{n\to\infty} E(B_n^2) = a > 0, \tag{5.9}$$

then as $n \to \infty$,

$$[E(V_{B,n} - \theta_n^2)/E(V_{H,n} - \theta_n^2)] \to 1. \tag{5.10}$$

As a consequence both the bootstrap and half-sample estimates of variance of quantiles are equally efficient. The condition (5.9) clearly holds in the i.i.d. case.

Under (4.8), $pqh_n^{-1}E(B_n^2)$ is clearly approximated by the second moment M_n of

$$T_n = n^{3/4} \int y a_n(y)\phi(y)dy,$$

where

$$a_n(y) = G_n(F_n^{-1}(p + y\sqrt{pq/n})) - G_n(F_n^{-1}(p)) - y\sqrt{pq/n}. \tag{5.11}$$

Note that

$$M_n = n^{3/2} \iint xy\phi(x)\phi(y)E(a_n(x)a_n(y))dxdy. \tag{5.12}$$

Under suitable conditions, which include

$$\lim_{n\to\infty} \sup_{|u|\le(\log n)n^{-1/2}} \sup_{1\le j\le n} |F_j(G_n^{-1}(p+u)) - F_j(G_n^{-1}(p))| = 0, \tag{5.13}$$

it can be shown that

$$\lim_{n\to\infty} M_n = \eta^2 - \lim_{n\to\infty} \iint xy\phi(x)\phi(y)a_n(x,y)dxdy, \tag{5.14}$$

and the random variable T_n converges weakly to the normal distribution with variance η^2, where

$$\eta^2 = \iint xy \min(x,y)\phi(x)\phi(y)dxdy = 1/2\sqrt{\pi},$$

$$b_n(y) = G_n^{-1}(p + y\sqrt{pq/n})$$

and

$$a_n(x,y) = n^{-1/2} \sum_{j=1}^{n} (F_j(b_n(y)) - F_j(b_n(0)))(F_j(b_n(x)) - F_j(b_n(0))).$$

Clearly for $|y| \le \log n$

$$|a_n(x,y)| \le o(\sqrt{n}\,|x|\sqrt{pq/n}) = o(|x|).$$

It follows that the second limit in (5.14) is zero, leading to the conclusion

$$\lim_{n\to\infty} M_n = \eta^2 = 1/2\sqrt{\pi} > 0.$$

Consequently in this case (5.9) holds.

6. A linear model

Bootstrap and the half-sample methods are shown to be equally efficient, for estimating the variance of a sample quantile in the last section. This is achieved via the representations (5.4), (5.5) and (5.8). It turns out that this equivalence fails to hold for the least squares estimators of linear regression parameters when the errors are heterogeneous. Liu and Singh (1992) have shown that bootstrap is more efficient than jackknife when the errors are homogeneous, but yields inconsistent estimates in the heterogeneous case. In this section we establish that the half-sample method is equivalent to jackknife in the heterogeneous case.

The procedure is best illustrated by the simple linear model

$$Y_i = \alpha + \beta x_i + \epsilon_i \tag{6.1}$$

$i = 1, 2, \ldots, n$, where ϵ_i are independent random variables with mean zero and variance σ_i^2. Let

$$L_n = \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

$$\hat{\beta} = \sum_{i=1}^{n} (x_i - \bar{x}) Y_i L_n^{-1}.$$

Let for $s \in H$,

$$\bar{x}_s = \frac{1}{k} \sum_{i \in s} x_i, \qquad L_s = \sum_{i \in s} (x_i - \bar{x}_s)^2,$$

$$\beta_s = \sum_{i \in s} (x_i - \bar{x}_s) Y_i L_s^{-1}$$

and

$$\theta_s = 2L_s(\beta_s - \hat{\beta}).$$

Recall that H consists of all subsets of $\{1, \ldots, n\}$ of size $k = [n/2]$. Clearly

$$2E_H(L_s) = 2L_n \left(\frac{k}{n} - \frac{n-k}{n(n-2)} \right) = L_n(1 + O(n^{-1}))$$

and

$$2E_H(L_s \beta_s) = \hat{\beta} L_n (1 + O(n^{-1})).$$

Practically negligible bias above, dictated the use of weights $2L_s/L_n$. Note that L_n can be taken as $\sum x_i^2$, and $L_s = \sum_{i \in s} x_i^2$, when $\alpha = 0$. In this case the chosen weights yield unbiased estimators. See Liu and Singh (1992). The equality above gives

$$E_H(\theta_s) = \binom{n}{k}^{-1} \sum_{s \in H} \theta_s = O(n^{-1}).$$

A simple algebra leads to

$$E_H(\theta_s - E_H(\theta_s))^2 = \left[\sum_{i=1}^{n}(x_i - \bar{x})^2\epsilon_i^2 + (\hat{\beta} - \beta)^2\sum_{i=1}^{n}(x_i - \bar{x})^4 - 2(\hat{\beta} - \beta)\sum_{i=1}^{n}(x_i - \bar{x})^3\epsilon_i\right](1 + O(n^{-1})). \tag{6.2}$$

Let $\sigma_i^2 \leq c$ for all i and for some $c > 0$. Further, let

$$v_n = L_n^{-2}\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_i^2, \qquad \tilde{v}_n = \sum_{i=1}^{n}\sigma_i^2/(nL_n) \tag{6.3}$$

and given the sample let

$$V_H = E_H[2L_s(\beta_s - \hat{\beta})L_n^{-1}]^2. \tag{6.4}$$

As mentioned above, the weights $2L_s/L_n$ are chosen as they lead to negligible bias. These weights were considered by Wu (1986).

THEOREM 6.1. *Suppose $n = O(L_n)$ and for some positive δ, c, c_1 and c_2,*

$$0 < \sigma_i \leq c \quad \text{for all } i \tag{6.5}$$

$$c_1 \leq w_n = n^3L_n^{-4}\sum_{i=1}^{n}(x_i - \bar{x})^4(E(\epsilon_i^4) - \sigma_i^4) \leq c_2 \tag{6.6}$$

and

$$n^{-1-\delta}\sum_{i=1}^{n}|x_i|^{4(1+\delta)} \to 0. \tag{6.7}$$

then

$$n^{3/2}\left[(V_H - v_n) - L_n^{-2}\sum_{i=1}^{n}(x_i - \bar{x})^2(\epsilon_i^2 - \sigma_i^2)\right] = o_p(1) \tag{6.8}$$

and $n^{3/2}(V_H - v_n)w_n^{-1/2}$ tends weakly to the standard normal distribution.

PROOF. By (6.2), (6.3) and (6.4), the left hand side of (6.8)

$$= O_p\left(L_n^{-2}\sum_{1}^{n}(x_i - \bar{x})^4\sqrt{n} + nL_n^{-2}\left(\sum_{i=1}^{n}(x_i - \bar{x})^3\epsilon_i\right) + \sqrt{n}\,L_n^{-2}\sum_{i=1}^{n}(x_i - \bar{x})^2\epsilon_i^2\right) \to 0$$

in probability by (6.5) and (6.7). By (6.6), (6.7), (6.8) and Lyapounov's Theorem (see B(ii) on p. 275 of Loéve (1963)) that $n^{3/2}(V_H - v_n)w_n^{-1/2}$ converges in distribution to the standard normal. This completes the proof.

Let $V_B = E_B(\hat{\beta}_B - \hat{\beta})^2$ be the bootstrap estimate of the variance of $\hat{\beta}$. If $c_1 \leq n/L_n \leq c_2$ for some $c_1, c_2 > 0$, then Liu and Singh (1992) have shown that

$$n(V_B - \tilde{v}_n) = \sum_{i=1}^{n}(\epsilon_i^2 - \sigma_i^2)L_n^{-1} + O_p(n^{-1}).$$

(They consider the case $\alpha = 0$ in which case L_n can be taken, instead, as $\sum_{i=1}^n x_i^2$.) Theorem 6.1 gives the half-sample estimate of the variance v_n. This is asymptotically equivalent to the jackknife estimate (see Theorem 4(I) of Liu and Singh (1992)). As discussed in that paper if σ_i are not all the same, the bootstrap is not a reliable method to estimate the variance, as in this case $v_n \neq \tilde{v}_n$. This leads to inconsistent estimators. However, if $\sigma_i^2 = \sigma^2$ for all i, then indeed the bootstrap is more efficient. This can be seen by comparing the variances of $n^{3/2}(V_B - \tilde{v}_n)$ and $n^{3/2}(V_H - v_n)$. See Liu and Singh (1992).

Unlike bootstrap the half-sample method is robust against departures from the homogeneous case. Robustness of related jackknife estimators was studied by Hinkley (1977), Wu (1986) and Shao and Wu (1987). Consequently, the half-sample method is more appealing than the bootstrap. However, it should be emphasized that the half-sample method is robust but not as efficient as the bootstrap in the homogeneous case.

Appendix

We assume the conditions of Theorem 4.1. We clearly have for almost all samples,

$$0 \leq F_s(F_s^{-1}(t)) - t \leq \frac{1}{k} \tag{A.1}$$

uniformly in t and in $s \in H$.

LEMMA A.1. *For some b, b_1 and b_2 we have*

$$P(|F_n(x) - F_n(y) - G_n(x) - G_n(y)| > b_1 n^{-3/4}\log n) = O(n^{-6}) \tag{A.2}$$

uniformly for x, y in

$$(G_n^{-1}(p) - b\sqrt{(\log n)/n}, G_n^{-1}(p) + b\sqrt{(\log n)/n}),$$

and

$$P(|F_n^{-1}(t) - G_n^{-1}(t)| > b\sqrt{(\log n)/n}) = O(n^{-6}) \tag{A.3}$$

uniformly for $|t - p| \leq b_2\sqrt{(\log n)/n}$.

PROOF. The inequality (A.2) follows by using, a Bernstein type inequality, Lemma 1 of Babu (1989). The estimate (A.3) follows again by Lemma 1 of Babu (1989) on noting that for some b_3 and b_4,

$$|F_n(t) - G_n(t)| \leq b_3\sqrt{(\log n)/n} \quad \text{for} \quad |t-p| \leq b_4\sqrt{(\log n)/n}$$

implies

$$|F_n^{-1}(t) - G_n^{-1}(t)| \leq b\sqrt{(\log n)/n} \quad \text{for} \quad |t-p| \leq b_2\sqrt{(\log n)/n}.$$

This completes the proof.

We first note that from the hypothesis of Theorem 4.1, we have for any $b > 0$,

$$C(b) = \liminf_{n\to\infty} \inf\{g_n(x) : |x - G_n^{-1}(p)| \leq b\sqrt{(\log n)/n}\} > 0. \tag{A.4}$$

Let for any $b > 0$,

$$b_n = C(b)b\sqrt{(\log n)/n}. \tag{A.5}$$

LEMMA A.2. *For some* $b > 0$

$$E(P_H(|F_s^{-1}(p) - F_n^{-1}(p)| > b_n \mid X_1, \ldots, X_n)) = O(n^{-5}). \tag{A.6}$$

Lemma A.2 follows by using (A.3), (A.4), (5.1) and Stirling's formula.

PROOF OF THEOREM 4.1. We only prove (4.3). Proof of (4.4) is similar. For $s \in H$, let

$$J_s = F_s(F_s^{-1}(p)) - F_s(F_n^{-1}(p)) - F_n(F_s^{-1}(p)) + F_n(F_n^{-1}(p)).$$

Note that

$$\begin{aligned} F_s(F_n^{-1}(p)) - p - G_n(F_n^{-1}(p)) + G_n(F_s^{-1}(p)) \\ = -J_s + (F_s(F_s^{-1}(p)) - p) \\ + [F_n(F_n^{-1}(p)) - F_n(F_s^{-1}(p)) - G_n(F_n^{-1}(p)) + G_n(F_s^{-1}(p))] \\ = -J_s + I_s + II_s \quad \text{(say)}. \end{aligned} \tag{A.7}$$

Let b_n be as in (A.5). By (A.1), $0 \leq I_s \leq 1/k$ and by dividing the region $\{x : |x - G_n^{-1}(p)| \leq 2b_n\}$ into sub intervals of length $(B/C(b))n^{-3/4}\log n$ for some $B > 0$, and using (A.2) and (A.3) we obtain that

$$E(P_H(|II_s| > b_5 n^{-3/4}\log n \mid X_1, \ldots, X_n)) = O(n^{-5}), \tag{A.8}$$

for some $b_5 > 0$. On the set $|F_n^{-1}(p) - G_n^{-1}(p)| < b\sqrt{(\log n)/n}$, we have for some $b_6 > 0$,

$$|J_s| \le 2\sup\{J_s(x) : |x - G_n^{-1}(p)| \le b_6\sqrt{(\log n)/n}\}, \tag{A.9}$$

where

$$J_s(x) = |F_s(x) - F_s(G_n^{-1}(p)) - F_n(x) + F_n(G_n^{-1}(p))|. \tag{A.10}$$

As in the proof of (A.8), the result now follows from (A.7) and (A.8) if we show that, for some $b_6 > 0$ and $b_7 > 0$,

$$E(P_H(|J_s(x)| > b_7 n^{-3/4}\log n \mid X_1, \ldots, X_n)) = O(n^{-6}), \tag{A.11}$$

uniformly in $|x - G_n^{-1}(p)| \le b_6\sqrt{(\log n)/n}$. To establish (A.11), let

$$G_n^{-1}(p) - b_6\sqrt{(\log n)/n} \le x < y \le G_n^{-1}(p) + b_6\sqrt{(\log n)/n},$$
$$t_i = P(x < X_i \le y) \quad \text{and} \quad u_n = b_7 n^{-3/4}\log n.$$

Markov inequality gives that for any $t > 0$

$$\begin{aligned} &E(P_H(F_s(s) - F_s(x) - F_n(y) + F_n(x) > u_n \mid X_1, \ldots, X_n)) \\ &\quad \le e^{-tku_n}\binom{n}{k}^{-1}\sum_{s\in H} E[\exp(tk(F_s(y) - F_s(x) + F_n(x) - F_n(y)))]. \end{aligned} \tag{A.12}$$

The expectation term on the right side of (A.12) is

$$\begin{aligned} &E\left[\prod_{i\in s}\exp[t(1 - kn^{-1})I(x < X_i \le y)]\prod_{i\notin s}\exp[-tkn^{-1}I(x < X_i \le y)]\right] \\ &\quad = \prod_{i\in s}[1 + t_i(e^{t(1-k/n)} - 1)]\prod_{i\notin s}[1 + t_i(e^{-tk/n} - 1)]. \end{aligned} \tag{A.13}$$

By (A.13) it follows that (A.12) is not more than

$$\begin{aligned} &e^{-tku_n}\binom{n}{k}^{-1}\prod_{i=1}^{n}\{2 - 2t_i + t_i e^{-tk/n}(e^t + 1)\} \\ &\quad \le 2^n\binom{n}{k}^{-1}e^{-tku_n}\prod_{i=1}^{n}\{1 - t_i + t_i e^{-tk/n}(1 + e^t)2^{-1}\} \\ &\quad \le b_8\sqrt{n}\,e^{-ktu_n}\exp\left(8^{-1}t^2\sum_{i=1}^{n}t_i(1 + o(t))\right). \end{aligned}$$

The last inequality is arrived at using Stirling's formula. The result (A.11) now follows by choosing $t = b_9 n^{-1/4}\sqrt{\log n}$ for some $b_9 > 0$. This completes the proof of Theorem 4.1.

REFERENCES

Abramovitch, L. and Singh, K. (1985). Edgeworth corrected pivotal statistics and the bootstrap, *Ann. Statist.*, **13**, 116–132.

Babu, G. J. (1984). Bootstrapping statistics with linear combinations of chi-squares as weak limit, *Sankhyā Ser. A*, **46**, 85–93.

Babu, G. J. (1986). A note on bootstrapping the variance of sample quantile, *Ann. Inst. Statist. Math.*, **38**, 439–443.

Babu, G. J. (1989). Strong representations for LAD estimators in linear models, *Probab. Theory Related Fields*, **83**, 547–558.

Babu, G. J. and Bose, A. (1988). Bootstrap confidence intervals, *Statist. Probab. Lett.*, **7**, 151–160.

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

Babu, G. J. and Singh, K. (1984*a*). On one term Edgeworth correction by Efron's bootstrap, *Sankhyā Ser. A*, **46**, 219–232.

Babu, G. J. and Singh, K. (1984*b*). Asymptotic representations related to jackknifing and bootstrapping L-statistics, *Sankhyā Ser. A*, **46**, 195–206.

Babu, G. J. and Singh, K. (1985). Edgeworth expansions for sampling without replacement from finite populations, *J. Multivariate Anal.*, **17**, 261–278.

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

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals, *Ann. Statist.*, **16**, 927–953.

Hartigan, J. A. (1969). Using subsample values as typical values, *J. Amer. Statist. Assoc.*, **64**, 1303–1317.

Hartigan, J. A. (1975). Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values, *Ann. Statist.*, **3**, 573–580.

Hinkley, D. V. (1977). Jackknife in unbalanced situations, *Technometrics*, **19**, 285–292.

Liu, R. Y. and Singh, K. (1989). Interpreting i.i.d. inferences under some non i.i.d. models (preprint).

Liu, R. Y. and Singh, K. (1992). Efficiency and robustness in resampling, *Ann. Statist.*, **20**, 370–384.

Loéve, M. (1963). *Probability Theory*, 3rd ed., Van Nostrand, Princeton, New Jersey.

Mahalanobis, P. C. (1946). Report on the bihar crop survey: rabi season 1943–1944, *Sankhyā Ser. A*, **7**, 269–280.

Maritz, J. S. and Jarrett, R. G. (1978). A note on estimating the variance of the sample median, *J. Amer. Statist. Assoc.*, **73**, 194–196.

Rao, C. R. (1988). Methodology based on the L_1-norm in statistical inference, *Sankhyā Ser. A*, **50**, 289–313.

Shao, J. and Shi, X. (1989). Half-sample variance estimation (preprint).

Shao, J. and Wu, C. F. J. (1987). Heteroscedasticity-robustness of jackknife variance estimators in linear models, *Ann. Statist.*, **15**, 1563–1579.

Shao, J. and Wu, C. F. J. (1989). General theory for jackknife variance estimation, *Ann. Statist.*, **17**, 1176–1197.

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

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion), *Ann. Statist.*, **14**, 1261–1350.

Wu, C. F. J. (1990). On the asymptotic properties of the jackknife histogram, *Ann. Statist.*, **18**, 1438–1452.