A PSEUDO-REPLICATE ESTIMATE OF VARIANCE FOR A FUNCTION OF SAMPLE MEANS

Alan Fenech, University of California, Davis

1. INTRODUCTION

Consider a finite population of vectors composed of L strata, together with a parameter $\theta = f(Y_{1.}, Y_{2.}, \ldots Y_{L.})$; $Y_{i.}$ is the average of the vectors in stratum i and f is a known real-valued function. Many parameters are of this form: the combined ratio, the mean of a domain, and the correlation coefficient are examples. Suppose that a random sample of size n is taken from each stratum, with the objective of constructing a confidence interval for θ. Let $y_{i.}$ be the average of the vectors sampled from stratum i. Then a reasonable estimate of θ is $f(y_{1.}, y_{2.}, \ldots, y_{L.}) = \hat{\theta}(n)$. Denote the variance of $\hat{\theta}(n)$ by $\sigma^2(n)$. If $(\hat{\theta}(n)-\theta)/\sqrt{\sigma^2(n)}$ is approximately standard normal, then a confidence interval for θ can be constructed by finding an accurate estimate of $\sigma^2(n)$. Currently there are three approaches used to estimate $\sigma^2(n)$. First, by consideration of the specific function f it is possible that an estimable expression (often approximate) can be derived for $Var[f(y_{1.}, y_{2.}, \ldots, y_{L.})]$. Second, by explicit or numerical calculation of the appropriate partial derivatives of f, the delta method can be used to estimate $\sigma^2(n)$; see Tepping [7] and Woodruff [8]. Third, there are pseudo-replication methods for estimating $\sigma^2(n)$; see Kish and Frankel [5] and McCarthy [6]. Jackknife procedures applied to $\hat{\theta}(n)$ provide a fourth approach to constructing a confidence interval for θ.

In this paper, an estimator of $\sigma^2(n)$ based on pseudo-replicates is introduced. Some theoretical properties of this estimator are derived. Also the results of simulation experiments are presented, in which the interval based on the pseudo-replicate estimate of $\sigma^2(n)$ compared favorably with three other interval procedures. Specifically, in Section 2, the problem and notation of this paper are introduced. In Section 3, a family of pseudo-replicate estimates of θ is defined. An estimator $\hat{\sigma}^2_p(n)$ is proposed, which is based on the pseudo-replicate estimates of θ. In Section 4, it is shown that if $f(\cdot)$ is linear, $\hat{\sigma}^2_p(n)$ is identical to the standard unbiased estimate of $\sigma^2(n-1)$. Also, a theorem is presented which suggests that $(\hat{\theta}(n)-\theta)/\sqrt{\hat{\sigma}^2_p(n)}$ has approximately a standard normal distribution. The results of simulation experiments are presented, which show the interval based on $\hat{\sigma}^2_p(n)$ performing well in comparison to other standard interval procedures. In Section 5, remarks are made on the practical use of $\hat{\sigma}^2_p(n)$. Also, the jackknife interval used in the simulation studies behaved erratically relative to the other intervals. This behavior is displayed.

2. NOTATION

In this section, the notation for this paper is developed. Consider a finite population P of vectors from $\mathbb{R}^m$ divided into L strata. Denote the members of P by $Y_{ij} = (Y_{ij1}, Y_{ij2}, \ldots, Y_{ijm})'$. Here $1 \leq i \leq L$ and $1 \leq j \leq N_i$; i identifies the stratum and j a particular vector among the N_i which belong to stratum i. Denote $\Sigma_{j=1}^{N_i} Y_{ijk}/N_i$ by $Y_{i \cdot k}$ and denote the mean of the vectors in stratum i, $(Y_{i \cdot 1}, Y_{i \cdot 2}, \ldots, Y_{i \cdot m})'$ by $Y_{i.}$. Also define the matrix D_i by

$$[D_i]_{k\ell} = \frac{\sum_{j=1}^{N_i} (Y_{ijk}-Y_{i \cdot k})(Y_{ij\ell}-Y_{i \cdot \ell})}{(N_i-1)}$$

Suppose Z is a vector selected at random from stratum i. Then $E[Z]=Y_{i.}$ and $((N_i-1)/N_i)D_i = Cov[Z]$.

Suppose θ is an unknown real parameter, where $\theta = f(Y_{1.}, Y_{2.}, \ldots, Y_{L.})$ and $f(\cdot)$ is a known continuous function. Consider drawing a random sample of size n without replacement from each stratum; denote the vector outcome of draw j from stratum i by $y_{ij} = (y_{ij1}, y_{ij2}, \ldots, y_{ijm})'$. An estimate of θ is $\hat{\theta}(n) = f(y_{1.}, y_{2.}, \ldots, y_{L.})$, where $y_{i.} = \Sigma_{j=1}^{n} y_{ij}/n$. Define $b(n)$ and $\sigma^2(n)$ by $E[\hat{\theta}(n)] = \theta + b(n)$ and $Var[\hat{\theta}(n)] = \sigma^2(n)$, and the matrix $\hat{D}_i$ by

$$[\hat{D}_i]_{k\ell} = \frac{\sum_{j=1}^{n} (y_{ijk}-y_{i \cdot k})(y_{ij\ell}-y_{i \cdot \ell})}{(n-1)} .$$

Of course, the bias and variance of $\hat{\theta}(n)$ depend on P and $f(\cdot)$, in addition to n; since P and $f(\cdot)$ are kept fixed in the results of this paper, this dependence is not made explicit in the notation.

3. A PSEUDO-REPLICATE ESTIMATE OF $\sigma^2(n)$

In this section, a collection of pseudo-replicate estimates of θ is defined for subsamples of size (n-1) from the strata samples of size n. Then two results are developed. First, the average of the squared differences between these pseudo-replicates and $\hat{\theta}(n)$ equals the product of a factor greater than one and an unbiased estimate of $\sigma^2(n-1)(1-\rho(n))$. Here $\sigma^2(n-1)$ is the variance of $f(y_{1.}, y_{2.}, \ldots y_{L.})$ when the stratum sample size is (n-1); $\rho(n)$ is a correlation coefficient which will be defined. Second, under assumptions on the N_i and $f(\cdot)$, an explicit expression for $\rho(n)$ is derived. As a consequence of these two results, an estimator of $\sigma^2(n)$, $\hat{\sigma}^2_p(n)$, is developed.

Imitating McCarthy [6], we proceed to define a collection of pseudo-replicate estimates of θ. The data generated by drawing a sample of size n from each stratum can be displayed as follows.

Stratum	1	y_{11}	y_{12}	y_{13} ..	y_{1n}
	2	y_{21}	y_{22}	y_{23} ..	y_{2n}
	⋮	⋮	⋮	⋮	⋮
	L	y_{L1}	y_{L2}	y_{L3} ..	y_{Ln}

Denote this data matrix by y. There are $t=(n)^L$ ways to select a subsample of size (n-1) from each of the rows of y. These subsamples can be identified by indicating the element in each row of y that is not selected; we use the vector $(4,6,\ldots,7)' \in \mathbb{R}^L$ to denote the subsample which deletes $y_{14}, y_{26}, \ldots, y_{L7}$ from y. The vectors which in this fashion represent all possible subsamples are ordered lexicographically and the corresponding subsamples are numbered from 1 to t. Let $y_{i(j)}$ be the element of row i of y that is not in subsample j, $1 \leq j \leq t$. Now let $(n y_{i.} - y_{i(j)})/(n-1) = w_{ij}$. Finally define $\hat{\theta}_j(n) = f(w_{1j}, w_{2j}, \ldots, w_{Lj})$ for $1 \leq j \leq t$; the $\hat{\theta}_j(n)$ are the pseudo-replicate estimates of θ.

Define $\mathbb{R}^2(y)$ by

$$\mathbb{R}^2(y) = \sum_{j=1}^{t} (\hat{\theta}_j(n) - \hat{\theta}(n))^2/t$$

The magnitude of $\mathbb{R}^2(y)$ is an indicator of the magnitude of $\sigma^2(n)$. The following theorem is specific on this point.

Theorem 3.1 $(n/(n-1))\, \mathbb{R}^2(y) = S^2(y)\ (1 + T(y))$ where $E[S^2(y)] = \sigma^2(n-1)(1-\rho(n))$ and $T(y) \geq 0$.

Comments: the random variables $S^2(y)$ and $T(y)$, and the parameter $\rho(n)$, are conveniently defined in the proof of Theorem 3.1. Note in the proof that $S^2(y)$ is a sufficiency improvement of an unbiased estimate of $\sigma^2(n-1)\ (1-\rho(n))$, and that $T(y)$ looks as if it should be near zero.

Proof. Let $\hat{\theta}_k$, for $1 \leq k \leq n$, be the pseudo-replicate estimate of θ corresponding to the subsample of y obtained by deleting the k-th column of y. The $\hat{\theta}_k$ are identically distributed, having the same distribution as $\hat{\theta}(n-1)$. Therefore $\theta + b(n-1) = E[\hat{\theta}_k]$ and $\mathrm{Var}[\hat{\theta}_k] = \sigma^2(n-1)$. By symmetry, the $\hat{\theta}_k$ have common correlation; denote it by $\rho(n)$. Denote $\sum_{k=1}^{n} \hat{\theta}_k/n$ by $\hat{\theta}_{\cdot}$. Standard linear model theory yeilds that $\sum_{k=1}^{n}(\hat{\theta}_k - \hat{\theta}_{\cdot})^2/(n-1)$ is an unbiased estimate of $\sigma^2(n-1)\ (1-\rho(n))$. Write

$$\sum_{k=1}^{n} \frac{(\hat{\theta}_k - \hat{\theta}(n))^2}{(n-1)} = \sum_{k=1}^{n} \frac{(\hat{\theta}_k - \hat{\theta}_{\cdot})^2}{(n-1)} + \sum_{k=1}^{n} \frac{(\hat{\theta}_{\cdot} - \hat{\theta}(n))^2}{(n-1)} \qquad (3.1)$$

There are $u = (n!)^L$ data matrices y which can be constructed from the original y by permuting the elements in each row of the original y. Averaging the three terms in (3.1) over all these u data matrices $(1 \leq j \leq u)$ yields

$$\frac{n}{(n-1)} \sum_{j=1}^{u} \sum_{k=1}^{n} \frac{(\hat{\theta}_k - \hat{\theta}(n))^2}{un} \qquad (3.2)$$

$$= \sum_{j=1}^{u} \sum_{k=1}^{n} \frac{(\hat{\theta}_k - \hat{\theta}_{\cdot})^2}{u(n-1)} \left[1 + \frac{\sum_{j=1}^{u} \sum_{k=1}^{n} (\hat{\theta}_{\cdot} - \hat{\theta}(n))^2}{\sum_{j=1}^{u} \sum_{k=1}^{n} (\hat{\theta}_{\cdot} - \hat{\theta}_k)^2} \right]$$

Define $S^2(y)$ and $T(y)$ in the obvious way so that the right side of (3.2) equals $S^2(y)\ (1 + T(y))$. It is not difficult to show that the left side of equation (3.2) equals $(n/(n-1))\, \mathbb{R}^2(y)$. ■

Assuming $T(y) \doteq 0$ and $\sigma^2(n-1) \doteq \sigma^2(n)$, Theorem 3.1 suggests

$$[n/(n-1)]\,[\mathbb{R}^2(y)/(1-\rho(n)] \qquad (3.3)$$

as an estimator of $\sigma^2(n)$. But the value of $\rho(n)$ is not known, and in general depends upon $f(\cdot)$, n, and P. In this paragraph, an expression for $\rho(n)$ is derived under assumptions on f and the N_i. Then an estimator of $\sigma^2(n)$ is produced by substituting this expression into (3.3).

Theorem 3.2 Suppose that $f(Y_{1.}, Y_{2.}, \ldots, Y_{L.}) = \sum_{i=1}^{L} a_i' Y_{i.} + a_0$ where $a_0 \in \mathbb{R}$ and $a_i \in \mathbb{R}^m$ for $1 \leq i \leq L$. Then $\rho(n)$ is

$$\frac{\sum_{i=1}^{L} \left[-\frac{1}{N_i(n-1)^2} + \frac{(n-2)}{(n-1)^2} \frac{(N_i - n)}{N_i} \right] a_i' D_i a_i}{\sum_{i=1}^{L} \frac{1}{(n-1)} \frac{N_i - (n-1)}{N_i} a_i' D_i a_i}$$

Proof. $\rho(n) = \mathrm{Cor}(\hat{\theta}_1, \hat{\theta}_2)$, where $\hat{\theta}_i$ is defined in the proof of Theorem 3.1. $\mathrm{Cov}(\hat{\theta}_1, \hat{\theta}_2)$ equals

$$\mathrm{Cov}\left[\sum_{i=1}^{L} a_i' \frac{(n y_{i.} - y_{i1})}{(n-1)}, \sum_{i=1}^{L} a_i' \frac{(n y_{i.} - y_{i2})}{(n-1)} \right].$$

Since the draws from different strata are independent, this simplifies to

$$\frac{\sum_{i=1}^{L} \mathrm{Cov}\left[a_i' y_{i2} + \sum_{j=3}^{n} a_i' y_{ij}, a_i' y_{i1} + \sum_{j=3}^{n} a_i' y_{ij} \right]}{(n-1)^2}$$

Using $\mathrm{Cov}(a_i' y_{ij}, a_i' y_{i\ell}) = -(1/N_i) a_i' D_i a_i$ for $j \neq \ell$, this expression further simplifies to the numerator of $\rho(n)$. Now $\mathrm{Var}(\hat{\theta}_1) = \mathrm{Var}(\hat{\theta}_2) = \sum_{i=1}^{L} a_i' \,\mathrm{Cov}[(n y_{i.} - y_{i2})/(n-1)]\, a_i$ which simplifies to the denominator of $\rho(n)$. ■

The following theorem gives a useful expression for $\rho(n)$, in that it does not involve the D_i.

Theorem 3.3 Suppose $f(Y_{1.}, Y_{2.}, \ldots, Y_{L.}) = a_0 + \sum_{i=1}^{L} a_i' Y_{i.}$ as in Theorem 3.2, and $N_i = N$ for $1 \leq i \leq L$. Then $\rho(n) = [(N-n)(n-2) - 1]/[(n-1)(N-n+1)]$.

Proof. Simply examine $\rho(n)$ as given by Theorem 3.2. ■

Suppose that $f(\cdot)$ is approximately linear over the range of possible values of the random vectors $y_{i.}$ and that stratum sizes are equal. Then, substituting the result of Theorem 3.3 into (3.3), we have the following estimator of $\sigma^2(n)$.

$$\hat{\sigma}_p^2(n) = n\, \mathbb{R}^2(y)\, \frac{N-n+1}{N}$$

This estimator will be called the pseudo-replicate estimator of $\sigma^2(n)$.

4. RESULTS ON $\hat{\sigma}_p^2(n)$

In this section, three results are presented on $\hat{\sigma}_p^2(n)$. First, when $f(\cdot)$ is linear, $\hat{\sigma}_p^2(n)$ is identical to the standard unbiased estimate of $\sigma^2(n-1)$. Second, under the assumption of with replacement sampling, $(\hat{\theta}(n) - \theta)/\sqrt{n\, \mathbb{R}^2(y)}$ converges in distribution to a standard normal distribution. Third, the results of a simulation experiment are presented in which a confidence interval procedure for θ based on $\hat{\sigma}_p^2(n)$ performed well in comparison to three other interval procedures.

Suppose $f(Y_{1.}, Y_{2.}, \ldots, Y_{L.}) = a_0 + \sum_{i=1}^{L} a_i' Y_{i.}$. Then $\sigma^2(n) = \sum_{i=1}^{L} a_i' \,\mathrm{Cov}(y_{i.}) a_i$, which equals $\sum_{i=1}^{L} [(N-n)/(Nn)]\, a_i' D_i a_i$. The standard unbiased estimate of $\sigma^2(n)$ is produced by replacing D_i by $\hat{D}_i$, yielding

$$\sum_{i=1}^{L} \frac{N-n}{N} \frac{1}{n} a_i' \hat{D}_i a_i \qquad (4.1)$$

The following theorem shows that (4.1) is closely related to $\hat{\sigma}_p^2(n)$.

Theorem 4.1 Suppose $f(Y_{1\cdot},Y_{2\cdot},\ldots,Y_{L\cdot}) = a_0+\Sigma_{i=1}^{L} a_i' Y_{i\cdot}$. Then $\hat{\sigma}_p^2(n)$ equals

$$\frac{n}{n-1} \frac{N-(n-1)}{N-n} \sum_{i=1}^{L} \frac{N-n}{N} \frac{1}{n} a_i' D_i a_i \qquad (4.2)$$

Proof. The theorem follows from the identity $\mathbb{R}^2(y) = \Sigma_{i=1}^{L} a_i' \hat{D}_i a_i/(n(n-1))$, which we proceed to verify. By definition, $\mathbb{R}^2(y) = \Sigma_{j=1}^{t}(\theta_j(n)-\hat{\theta}(n))^2/t$ where $\hat{\theta}_j(n) = a_0+\Sigma_{i=1}^{L} a_i'(n\, y_{i\cdot} - y_{i(j)})/(n-1)$ and $\hat{\theta}(n) = a_0 + \Sigma_{i=1}^{L} a_i' y_{i\cdot}$. With the data matrix y fixed, consider the experiment of selecting one of the t pseudo-replicates at random. Denote the result by $\hat{\theta}$. Clearly $E[\hat{\theta}]=\hat{\theta}(n)$, so $\mathrm{Var}[\hat{\theta}]=\mathbb{R}^2(y)$. But it is legitimate to regard the random variable $\hat{\theta}$ as the pseudo-replicate estimate of θ produced by selecting n-1 vectors at random from each row of y. Denote by $\bar{y}_{i\cdot}$ the average of the (n-1) selected from the i-th row of y. From this point $\mathrm{Var}(\hat{\theta})$ equals

$$\sum_{i=1}^{L} a_i' \,\mathrm{Cov}(\bar{y}_{i\cdot}) a_i = \sum_{i=1}^{L} a_i' \left[\frac{n-(n-1)}{n} \frac{1}{n-1} \hat{D}_i\right] a_i$$

$$= \sum_{i=1}^{L} \frac{a_i' \hat{D}_i a_i}{n(n-1)} \qquad ∎$$

It should be noted that (4.2) simplifies to the standard unbiased estimate of $\sigma^2(n-1)$. This is consistent with Theorem 3.1, in light of the fact that T(y)=0 when f(·) is linear. Theorem 4.1 is comforting, in that it shows $\hat{\sigma}_p^2(n)$ to be a simple multiple of the standard unbiased estimator of $\sigma^2(n)$, when f(·) is linear.

In this paragraph, a theorem is presented which suggests that for nonlinear f(·), if $n \ll N$ and n is large, $(\hat{\theta}(n)-\theta)/\sqrt{\hat{\sigma}_p^2(n)}$ is approximately normal in distribution.

Theorem 4.2 Suppose that the stratum samples of size n are drawn with replacement. Then $(\hat{\theta}(n)-\theta)/\sqrt{n\,\mathbb{R}^2(y)}$ converges in distribution to a standard normal distribution, under the following condition. Regarding $f(x_1,x_2,\ldots x_L)$, $x_i \in \mathbb{R}^m$, as a function from $\mathbb{R}^{mL}$ to $\mathbb{R}$, f(·) has continuous partials and second partials in a neighborhood of $x_i=Y_{i\cdot}$, $1 \le i \le L$.

Proof. It is well-known that $(\hat{\theta}(n)-\theta)/\sqrt{\sigma^2(n)}$ converges in distribution to a standard normal distribution. The theorem follows from a lengthy argument presented by Fenech (3) that $\sigma^2(n)/n\mathbb{R}^2(y)$ converges stochastically to one. ∎

Theorem 4.2 suggests that an approximate 100β% confidence interval for θ is

$$\hat{\theta}(n) \pm Z(\beta)\sqrt{\hat{\sigma}_p^2(n)} \qquad (4.3)$$

where $P\{|N(0,1)| \le Z(\beta)\}=\beta$. Immediate questions are: how close is the actual level of confidence of (4.3) to the nominal 100β%? How does this difference compare with those which result with other standard methods of constructing an interval for θ? Using simulation experiments, these questions were studied. In the remaining paragraphs of this section, three methods other than (4.3) which can be used to construct an interval for θ are described, and a simulation experiment testing the four interval procedures is described and its results presented.

The three confidence interval methods other than (4.3) are: the standard method, based on direct derivation of an estimable expression for $\sigma^2(n)$; the delta method; and the jackknife method. Specifically $\sigma^2(n)=\mathrm{Var}[f(y_{1\cdot},y_{2\cdot},\ldots,y_{L\cdot})]$. Since f(·) is known, a (usually approximate) expression for $\sigma^2(n)$ is derived which can be estimated from the data. Denoting this estimate by $\hat{\sigma}_s^2(n)$, the following is the standard interval for θ:

$$\hat{\theta}(n) \pm Z(\beta)\sqrt{\hat{\sigma}_s^2(n)} \ . \qquad (4.4)$$

In the delta method, an estimate of the variance of the linear approximation to f(·) is used to estimate $\sigma^2(n)$, yielding

$$\hat{\theta}(n) \pm Z(\beta)\sqrt{\hat{\sigma}_d^2(n)} \ . \qquad (4.5)$$

In the jackknife method, the pseudo-values $\hat{\theta}_{-k}=n\hat{\theta}(n)-(n-1)\hat{\theta}_k$ for $1\le k\le n$ are regarded as i.i.d estimates of θ. (The $\hat{\theta}_k$ are defined in the proof of Theorem 3.1.) Then, denoting $\Sigma_{k=1}^{n}\hat{\theta}_{-k}/n$ by $\hat{\theta}_{-\cdot}$ and $\Sigma_{k=1}^{n}(\hat{\theta}_{-k}-\hat{\theta}_{-\cdot})^2/n(n-1)$ by $\hat{\sigma}_j^2(n)$, the jackknife interval for θ is:

$$\hat{\theta}_{-\cdot} \pm Z(\beta)\sqrt{\hat{\sigma}_j^2(n)} \ . \qquad (4.6)$$

This jackknife is suggested by Brillinger [1].

In the simulation experiment, the performances of the four interval procedures were determined for the problem of constructing a confidence interval for the mean of a domain. For a description of and the standard solution to this problem see Cochran [2], pages 146-149. An artificial population

$$P = \left\{(X(i,j),Y(i,j)) : 1 \le i \le 4,\ 1 \le j \le 500\right\}$$

composed of 4 strata of size 500 was constructed, where X(i,j) equals 1 or 0 according to whether Y(i,j) belongs to the domain of interest or not, and where $Y(i,j)=i+\sqrt{i}\,\epsilon(i,j)$. The $\epsilon(i,j)$ are pseudo-random N(0,1) numbers. Five sample sizes (n=2,6(1)) were considered. For each sample size n, six hundred random samples were selected from P; each sample is composed of a random sample of size n from each of the four strata. For each of the six hundred samples, twenty-one intervals for the domain mean were constructed, one for each of three methods at seven nominal levels of confidence. The entries of Table I give the proportion of each group of six hundred intervals that in fact covered the domain mean. Three comments: first, the standard interval for a domain mean, labeled RATIO in Table I, is given in Cochran [2], pages 148-149; second, the delta interval and the standard interval turn out to be algebraically identical for this problem; third, the standard error of a sample proportion based on six hundred trials is at most .0204.

In this simulation experiment, the pseudo-replication interval performed very well in comparison with the other intervals. For example, the difference between nominal and estimated confidence level is generally smallest for the pseudo-replication interval, substantially so for the smaller sample sizes. Five additional simulation experiments, similar to the one described, were carried out and yielded similar results. These limited simulation studies suggest that the pseudo-replication interval procedure (4.3) is valid for small/moderate n.

5. CONCLUDING REMARKS

The practical use of $\hat{\sigma}_p^2(n)$ is constrained by the assumptions of common stratum size and common sample size. Suppose these assumptions are not met; denote by N_i and n_i the size and sample size for the i-th stratum, respectively. If $n_i << N_i$ and the n_i are approximately equal, in the sense that n_i/n_j is near 1 for all i and j pairs, then the developments in Section 3 can still be carried out, replacing equalities by approximate equalities. This leads to an estimator $\bar{n}\,\mathbb{R}^2(y)$, where $\bar{n}$ is a measure of location of the n_i (say the mean) and $\mathbb{R}^2(y)$ is based on the $\Pi_{i=1}^{L} n_i$ possible pseudo-replicates. Under these circumstances, the interval procedure

$$\hat{\theta}(n) \pm Z(\beta)\sqrt{\bar{n}\,\mathbb{R}^2(y)} \qquad (5.1)$$

will be valid. If the n_i are quite unequal, the development of Section 3 breaks down because the expression for ρ analogous to (3.4) does not allow the $a_i' D_i a_i$ to even approximately cancel out. In this case, it is not clear how to use $\mathbb{R}^2(y)$ to estimate $\mathrm{Var}[f(y_{1\cdot},\ldots,y_{L\cdot})]$.

Another apparent constraint to the use of $\hat{\sigma}_p^2(n)$ is the calculation of $\mathbb{R}^2(y)$, as it involves calculating n^L pseudo-estimates. n^L may be prohibitively large. If so, select a large random sample of the n^L subsamples, and use the associated pseudo-replicates estimates to estimate $\mathbb{R}^2(y)$. McCarthy's [6] clever balanced selection scheme for n=2 suggests that there may be a systematic scheme for selecting a modest number of the n^L subsamples whose pseudo-replicate estimates can be used to estimate $\mathbb{R}^2(y)$ well also. Parenthetically, an estimator of $\sigma^2(n)$ could be based on $S^2(y)$, rather than on $\mathbb{R}^2(y)$. But note that calculating $S^2(y)$ involves many more operations than calculating $\mathbb{R}^2(y)$.

One final remark. In the simulation studies the jackknife interval is occasionally very misleading, in the sense that $(\hat{\theta}_{-\cdot} - \theta)/\sqrt{\hat{\sigma}_J^2(n)}$ is very large in magnitude. This kind of behavior was not exhibited by the pseudo-replication interval or the standard interval. These points are reflected by the entries in Table II. The entries come from the experiment associated with Table I. For each sample size (n=2,6(1)), the following triple of numbers were listed for the first one hundred samples: $(\hat{\theta}(n)-\theta)/\sqrt{\hat{\sigma}_S^2(n)}$, $(\hat{\theta}(n)-\theta)/\sqrt{\hat{\sigma}_p^2(n)}$, and $(\hat{\theta}_{-\cdot}-\theta)/\sqrt{\hat{\sigma}_J^2(n)}$. Table II lists six of these triples for each sample size, from among those triples at least one of whose members is 2.0 or larger in magnitude. The poor performance of the jackknife interval is apparent particularly at small sample sizes. In no case in these simulation experiments are either of the other intervals very misleading relative to the jackknife interval. Hinkley [4] shows that a jackknifed estimator of the correlation coefficient displays the same erratic behavior.

References

[1] Brillinger, David R., "The Application of the Jackknife to the Analysis of Sample Surveys," Commentary, 8 (1966), 74-80.

[2] Cochran, William G., "Sampling Techniques," 2nd Ed., New York: John Wiley and Sons, Inc., 1963.

[3] Fenech, Alan P., "A Pseudo-Replicate Estimator of Variance," submitted.

[4] Hinkley, David V., "Improving the jackknife with special reference to correlation estimation," Biometrika, 65 (1978), 13-21.

[5] Kish, L. and Frankel, M., "Balanced Repeated Replications for Standard Errors," J.A.S.A., 65 (1970), 1071-1094.

[6] McCarthy, Philip, "Pseudo-replications: Half Samples," Rev. of the I.S.I., 37 (1969), 239-264.

[7] Tepping, Benjamin J., "Variance Estimation in Complex Surveys," Proc. of the Soc. Stat. Sect. of ASA, (1968), 11-18.

[8] Woodruff, Ralph S. and Causey, Beverly D., "Computerized Method for Approximating the Variance of a Complicated Estimate," J.A.S.A. 71 (1976), 315-321.

Table I Nominal Confidence Level

		80%	85%	90%	92%	94%	95%	99%
	RATIO	.7233	.7550	.8083	.8300	.8433	.8516	.9150
n=2	PSEUDO	.8616	.8850	.9100	.9250	.9400	.9416	.9750
	JACK.	.6000	.6283	.6733	.6983	.7066	.7233	.7883
	RATIO	.7500	.7866	.8450	.8550	.8766	.8883	.9450
n=3	PSEUDO	.8483	.8683	.9100	.9283	.9366	.9433	.9750
	JACK.	.6883	.7216	.7600	.7783	.7916	.8083	.8733
	RATIO	.7583	.8200	.8600	.8833	.9100	.9183	.9666
n=4	PSEUDO	.8350	.8800	.9183	.9350	.9450	.9500	.9866
	JACK.	.7383	.7716	.8116	.8333	.8583	.8733	.9250
	RATIO	.7500	.7983	.8583	.8883	.9133	.9283	.9750
n=5	PSEUDO	.8050	.8550	.9133	.9383	.9500	.9566	.9883
	JACK.	.7316	.7650	.8166	.8366	.8483	.8533	.9350
	RATIO	.7950	.8400	.8950	.9083	.9200	.9283	.9716
n=6	PSEUDO	.8350	.8850	.9116	.9283	.9400	.9450	.9833
	JACK.	.7716	.8316	.8733	.8916	.9050	.9133	.9533

Table II $(\hat{\theta} - \theta)/\sqrt{\hat{\sigma}^2(n)}$

	RATIO	- 2.7209	- 1.4934	- 3.7136	0.7503	1.3030	- 1.0643
n=2	PSEUDO	- 1.8639	- 0.9931	- 2.0315	0.4902	0.7503	- 0.5558
	JACK.	-35.4063	-48.3115	-17.3102	27.0443	17.4852	-19.5939
	RATIO	- 3.9695	- 5.3978	- 2.7540	- 1.9705	- 4.7656	- 2.6943
n=3	PSEUDO	- 3.2193	- 4.3597	- 2.1649	- 1.5389	- 3.8533	- 1.9682
	JACK.	- 4.7950	- 4.1977	-11.1347	- 5.8058	- 6.1411	- 4.3014
	RATIO	2.0650	- 2.5770	- 3.1543	- 2.3468	1.3672	- 1.9596
n=4	PSEUDO	1.7180	- 2.1787	- 2.6850	- 1.9992	1.1539	- 1.6688
	JACK.	2.3439	- 3.7330	- 2.9778	- 1.9311	2.9534	- 3.0353
	RATIO	- 2.8511	- 2.0258	1.8956	2.9404	- 1.3670	- 4.3412
n=5	PSEUDO	- 2.5134	- 1.7273	1.6688	2.5353	- 1.2081	- 3.7903
	JACK.	- 4.1252	- 3.2102	4.4961	5.3570	- 5.1943	- 3.9459
	RATIO	- 1.8884	- 2.3032	- 2.7854	- 1.1550	- 2.9543	- 2.4633
n=6	PSEUDO	- 1.6983	- 2.0734	- 2.5333	- 1.0293	- 2.6668	- 2.2107
	JACK.	- 2.0499	- 8.7949	- 3.0168	- 2.7209	- 2.9590	- 2.6132