Stochastic Inequalities
IMS Lecture Notes – Monograph Series
Volume 22 (1993)

ON THE BIAS OF THE JACKKNIFE ESTIMATE OF VARIANCE[1]

By RICHARD A. VITALE

University of Connecticut

Using machinery developed earlier for the covariances of symmetric statistics, we consider various aspects of the bias of the jackknife estimate of variance.

1. Introduction

The jackknife estimate of variance (Quenouille (1949, 1956), Tukey (1958)) can be described as follows. Given a symmetric function h of iid arguments $X_1, X_2, \ldots, X_m$, it is desired to estimate $\sigma^2 = \text{Var } h$. With an augmented supply $X_1, X_2, \ldots, X_n$ where $n = m+1$, or more generally $n \geq m+1$, one forms $Q = \binom{n-1}{m}^{-1} \sum_{|I|=m} [h(X_I) - \bar{h}]^2$ where $\bar{h} = \binom{n}{m}^{-1} \sum_{|I|=m} h(X_I)$ and $X_I \equiv (X_{i_1}, X_{i_2}, \ldots, X_{i_m})$ with $I = \{i_1, i_2, \ldots, i_m\}$. Several papers (Efron and Stein (1981), Karlin and Rinott (1982), Bhargava (1983), Vitale (1984), Steele (1986)) have considered the bias relation

$$\sigma^2 \leq EQ, \tag{1.1}$$

which has come to be known as the Efron–Stein inequality. Our purpose here is to investigate aspects of (1.1) including (i) an alternate proof with variant forms of the condition for equality, (ii) a sharpening, (iii) a complementary upper bound, and (iv) a consideration of Q as an estimator which is "contaminated" by estimators of other parameters.

2. Preliminaries

If $X_1, X_2, \ldots, X_n$ are iid random variables and h is a symmetric function of m of them with $Eh^2 < \infty$, then we set

$$r_k = \text{Cov}[h(X_I), h(X_J)]$$

[1] Work supported in part under grants Office of Naval Research N00014–90–J–1641 and National Science Foundation DMS–9002665.

AMS 1991 *subject classifications.* Primary 60B15; Secondary 62M10, 62M15.

Key words and phrases. Efron–Stein inequality, jackknife estimate of variance, symmetric statistic, U–statistic.

where $k = |I \cap J|$, $I = \{i_1, \ldots, i_m\}$, $J = \{j_1, \ldots, j_m\}$. Setting

$$h_k(X_1, \ldots, X_k) = E[h(X_1, \ldots, X_m) \mid X_1, \ldots, X_k] \qquad k = 1, \ldots, m$$

and

$$g_{|J|}(X_J) = \sum_{I \subseteq J} \{(-1)^{|J|-|I|}\} h_{|I|}(X_I)$$

leads to Hoeffding's (1948) ANOVA–type expansion:

$$h(X_1, \ldots, X_m) = \sum_{J \subseteq \{1, \ldots, m\}} g_{|J|}(X_J).$$

Here different terms are uncorrelated and

$$v_k \equiv \operatorname{Var} g_k(X_1, \ldots, X_k) = \sum_{\ell=1}^{k} (-1)^{k-\ell} \binom{k}{\ell} r_\ell$$

with the inverse relation

$$r_k = \sum_{\ell=1}^{k} \binom{k}{\ell} v_\ell.$$

The *index* and *dual index* of h are defined to be $\min\{k \mid v_k > 0\}$ and $\max\{k \mid v_k > 0\}$ respectively (Vitale (1992)). These parameters bracket the orders of interaction among $X_1, \ldots, X_m$ which appear in the expansion for h. This is seen clearly, for example, when h is the k^{th} $(1 \leq k \leq m)$ symmetric polynomial in $\varphi(X_1), \ldots, \varphi(X_m)$, where $E\varphi(X_1) = 0$.

3. The Efron–Stein Inequality

As noted in the introduction, several proofs have been given for the inequality. Here we give one based on an explicit representation for the bias (cf. Karlin and Rinott (1982, Eqn. 5.4) for an alternate form).

THEOREM 1 *If h has index c and dual index c', then the bias of the jackknife estimate of variance is nonnegative and given by*

$$EQ - \sigma^2 = \frac{1}{\binom{n-1}{m}} \sum_{k=1}^{m-1} \binom{m}{k} \binom{n-m}{m-k} \left(\frac{k}{m} r_m - r_k \right), \tag{3.1}$$

equivalently,

$$= \sum_{\ell=c}^{c'} v_\ell \binom{m}{\ell} \left[\frac{m}{n-m} - \frac{\binom{n-\ell}{m-\ell}}{\binom{n-1}{m}} \right]. \tag{3.2}$$

The bias is zero if and only if $c' = 1$, which is equivalent to each of the following three conditions:

$$\text{(3.3)} \qquad \frac{r_k}{k} = \frac{r_m}{m} \qquad k = 1, \ldots, m-1$$

$$\text{(3.4)} \qquad v_\ell = 0 \qquad \ell = 2, \ldots, m$$

(3.5) $h(X_1, \ldots, X_m) = h^*(X_1) + h^*(X_2) + \cdots + h^*(X_m)$ *for some* h^*.

PROOF The identity $\sum_{|I|=m}(h(X_1) - \overline{h})^2 = [2\binom{n}{m}]^{-1} \sum_{|I|=|J|=m}[h(X_I) - h(X_J)]^2$ allows Q to be written as $\left[2\binom{n-1}{m}\binom{n}{m}\right]^{-1} \sum_{|I|=|J|=m}[h(X_I) - h(X_J)]^2$ and thus

$$EQ = \left[\binom{n-1}{m}\binom{n}{m}\right]^{-1} \sum_{|I|=|J|=m} [r_m - r_{|I \cap J|}].$$

Making use of the fact that the number of pairs (I, J) with $|I \cap J| = k$ is $\binom{n}{m}\binom{m}{k}\binom{n-m}{m-k}$, we have

$$\text{(3.6)} \qquad EQ = \binom{n-1}{m}^{-1} \left[\binom{n}{m} r_m - \sum_{k=1}^{m} \binom{m}{k}\binom{n-m}{m-k} r_k\right].$$

Adding and subtracting the value

$$\frac{m\binom{n}{m}}{n} r_m = \left(\sum_{k=1}^{m} \binom{m}{k}\binom{n-m}{m-k}\frac{k}{m}\right) r_m$$

and re–arranging terms yields (3.1). To get (3.2), we insert the expression $r_k = \sum_{\ell=1}^{k} \binom{k}{\ell} v_\ell$ into (3.1) and reverse the resulting double summation to obtain

$$\binom{n-1}{m}^{-1} \sum_{\ell=1}^{m} v_\ell \sum_{k=1}^{m} \binom{m}{k}\binom{n-m}{m-k}\left[\binom{m}{\ell}\frac{k}{m} - \binom{k}{\ell}\right].$$

The inner summation, which vanishes for $\ell = 1$, is the difference of $\binom{m}{\ell}\sum_{k=1}^{m}\binom{m}{k}\binom{n-m}{m-k}\frac{k}{m} = \binom{m}{\ell}\frac{m}{n}\binom{n}{m}$ and $\sum_{k=1}^{m}\binom{m}{k}\binom{n-m}{m-k}\binom{k}{\ell} = \binom{m}{\ell}\binom{n-\ell}{m-\ell}$, and (3.2) follows. The bracketed expression in (3.2) is easily checked to be nonnegative, and thus the bias is nonnegative. Conditions for equality follow from Vitale (1992, Theorem 5.2).

4. A Sharpening, and Complementary Inequality

Here we show that by refining the proof of Theorem 1 it is possible to get more precise results.

THEOREM 2 *Suppose that h has index greater than or equal to c and dual index less than or equal to c'. Then* (1.1) *can be improved to*

$$\left[\frac{n}{n-m}-\frac{\binom{n-c}{m-c}}{\binom{n-1}{m}}\right]\sigma^2 \le EQ \tag{4.1}$$

with the complementary inequality

$$EQ \le \left[\frac{n}{n-m}-\frac{\binom{n-c'}{m-c'}}{\binom{n-1}{m}}\right]\sigma^2. \tag{4.2}$$

PROOF Without loss of generality, assume that h has index and dual index precisely c and c' respectively. As before,

$$EQ-\sigma^2=\sum_{\ell=c}^{c'} v_\ell\binom{m}{\ell}\left[\frac{m}{n-m}-\frac{\binom{n-\ell}{m-\ell}}{\binom{n-1}{m}}\right].$$

We observe now that the bracketed factor is not simply nonnegative but *nondecreasing in ℓ* (vanishing at $\ell = 1$). Accordingly, we have the bound

$$EQ-\sigma^2 \ge \left[\sum_{\ell=c}^{c'} v_\ell\binom{m}{\ell}\right]\cdot\left[\frac{m}{n-m}-\frac{\binom{n-c}{m-c}}{\binom{n-1}{m}}\right] \ge \sigma^2\left[\frac{m}{n-m}-\frac{\binom{n-c}{m-c}}{\binom{n-1}{m}}\right],$$

which becomes (4.1) upon re–arrangement. The upper bound (4.2) is found in a similar manner.

It can be easily verified that (4.1) coincides with (1.1) when $c = 1$ and gives a strict improvement when $c > 1$. As another case, consider $m = n - 1$ (the situation originally treated by Efron and Stein). Then (4.1) and (4.2) become

$$c\sigma^2 \le EQ \le c'\sigma^2.$$

5. "Explaining" the Bias

We conclude by showing that the bias of Q can be understood as the result of an unbiased estimator of σ^2 being weighted against estimators of lower order covariances. Once again write Q in the form

$$Q=\left[2\binom{n-1}{m}\binom{n}{m}\right]^{-1}\sum_{|I|=|J|=m}[h(X_I)-h(X_J)]^2. \tag{5.1}$$

Let $\sum^{(k)}$ stand for summation over pairs (I, J) such that $|I \cap J| = k$, and recall that $N_k = \sum^{(k)} 1 = \binom{n}{m}\binom{m}{k}\binom{n-m}{m-k}$. The relations $E[h(X_I) - h(X_J)]^2 = 2[r_m - r_{|I \cap J|}]$ provide $\hat{r}_k$ as an unbiased estimator of r_k, where

$$\hat{r}_m = (2N_0)^{-1} \sum\nolimits^{(0)} [h(X_I) - h(X_J)]^2$$

and

$$r_k = \hat{r}_m - (2N_k)^{-1} \sum\nolimits^{(k)} [h(X_I) - h(X_J)]^2 \qquad k = 1, \ldots, m-1.$$

Substituting these expressions into (5.1) gives

$$\begin{aligned} Q &= \hat{r}_m + \sum_{k=1}^{m-1} \binom{m}{k}\binom{n-m}{m-k}\left(\frac{k}{m}\hat{r}_m - \hat{r}_k\right) \\ &= \frac{n}{n-m}\hat{r}_m - \binom{n-1}{m}^{-1} \sum_{k=1}^{m-1} \binom{m}{k}\binom{n-m}{m-k}\hat{r}_k, \end{aligned}$$

which displays the effects of the lower-order estimators.

References

Bhargava, R. P. (1983). A property of the jackknife estimation of the variance when more than one observation is omitted. *Sankhyā Ser. A* **45** 112–119.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9** 586–596.

Hoeffding, W. (1984). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** 293–325.

Karlin, S. and Rinott, Y. (1982). Applications of ANOVA type decompositions for comparisons of conditional variance statistics including jackknife estimates. *Ann. Statist.* **10** 485–501.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *J. Royal Statist. Soc. B* **11** 68–84.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika* **43** 353–360.

Steele, J. M. (1986). An Efron–Stein inequality for nonsymmetric statistics. *Ann. Statist.* **14** 753–758.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples (abstract). *Ann. Math. Statist.* **29** 614.

Vitale, R. A. (1984). An expansion for symmetric statistics and the Efron–Stein inequality. In *Inequalities in Statistics and Probability.* Y. L. Tong, ed. Institute of Mathematical Statistics, Hayward, CA. 112–114.

Vitale, R. A. (1992). Covariances of symmetric statistics. *J. Multivariate Anal.* **41** 14–26.

Department of Statistics
University of Connecticut
Storrs, CT 06269