The Annals of Statistics
1986, Vol. 14, No. 2, 753–758

AN EFRON–STEIN INEQUALITY FOR NONSYMMETRIC STATISTICS[1]

BY J. MICHAEL STEELE

Princeton University

If $S(x_1, x_2, \ldots, x_n)$ is any function of n variables and if X_i, $\hat{X}_i$, $1 \le i \le n$ are $2n$ i.i.d. random variables then

$$\operatorname{var} S \le \tfrac{1}{2} E \sum_{i=1}^{n} (S - S_i)^2,$$

where $S = S(X_1, X_2, \ldots, X_n)$ and S_i is given by replacing the ith observation with $\hat{X}_i$, so $S_i = S(X_1, X_2, \ldots, \hat{X}_i, \ldots, X_n)$. This is applied to sharpen known variance bounds in the long common subsequence problem.

1. Introduction. In Efron and Stein (1981) the following result was established: If $S(x_1, x_2, \ldots, x_{n-1})$ is a symmetric function of $n-1$ variables and $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables then for $S_i = S(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ and $\bar{S} = n^{-1}\sum_{i=1}^{n} S_i$, one has

$$\operatorname{var} S(X_1, X_2, \ldots, X_{n-1}) \le E \sum_{i=1}^{n} (S_i - \bar{S})^2. \tag{1.1}$$

This inequality was motivated by a desire to understand the nature of the bias in the jackknife estimate of variance, but it has also proved useful in the probabilistic analysis of algorithms, Steele (1981, 1982b). There have been extensions of the Efron–Stein inequality to the case where one drops out more than one observation from S (Bhargava (1980)), and there have been new proofs of the result by Karlin and Rinott (1982) and Vitale (1984).

The purpose of this note is to establish an analogue to (1.1) which is not burdened by a symmetry hypothesis. It will be proved using the Hilbert space technique introduced in this context by Vitale (1984).

Finally the inequality is applied to a problem of string comparisons by means of long common subsequences, a problem considered at length in Sankoff and Kruskal (1983). The best known bound on the variance of the longest common subsequence is improved, and a new k string comparison problem is introduced.

2. Main results. Let $S(x_1, x_2, \ldots, x_n)$ be any function of n arguments and consider the statistics formed by

$$S = S(X_1, X_2, \ldots, X_n)$$

and $S_i = S(X_1, X_2, \ldots, X_{i-1}, \hat{X}_i, X_{i+1}, \ldots, X_n)$ where the X_i and $\hat{X}_i$ are $2n$

Received June 1985; revised July 1985.

[1]Research partially supported by NSF Grant DMS-84-14069.

AMS 1980 *subject classifications.* Primary 60E15; secondary 62H20.

Key words and phrases. Efron–Stein inequality, variance bounds, tensor product basis, long common subsequences.

independent random variables with the distribution F. In other words, the S_i are formed by redrawing the ith datum independently and then recalculating S. We will prove the following inequality:

$$\operatorname{var} S \le \tfrac{1}{2} E \sum_{i=1}^{n} (S - S_i)^2. \tag{2.1}$$

First we can check that there is no loss of generality in assuming that $ES^2 < \infty$. To do this, consider new variables which resample the first i observations, so $\hat{S}_i = S(\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_{i-1}, \hat{X}_i, X_{i+1}, \ldots, X_n)$, where X_i and $\hat{X}_i$, $1 \le i \le n$ are $2n$ i.i.d. random variables. Setting $\hat{S}_0 = S$, one has by Schwarz's inequality that

$$\begin{aligned} \operatorname{var}(S) &= \tfrac{1}{2} E(S - \hat{S}_n)^2 \\ &= \tfrac{1}{2} E\left(\sum_{i=0}^{n-1} (\hat{S}_i - \hat{S}_{i+1}) \right)^2 \\ &\le \left(\frac{n}{2} \right) \sum_{i=0}^{n-1} E(\hat{S}_i - \hat{S}_{i+1})^2. \end{aligned} \tag{2.2}$$

Since $\hat{S}_i - \hat{S}_{i+1}$ has the same distribution as $S - S_{i+1}$, we see from (2.2) that the right-hand side of (2.1) is infinite unless $\operatorname{var} S < \infty$. This shows (2.2) holds when $ES^2 = \infty$ and thus lets us focus on the case $ES^2 < \infty$.

By elementary Hilbert space theory we know that there are functions ϕ_k such that $\phi_0(x) \equiv 1$ and $E\phi_k(X)\phi_l(X) = \delta_{kl}$, i.e., we choose ϕ_k, $0 \le k < \infty$ to be an orthonormal basis for $L^2(dF)$. Further, if we let $\mathbf{k} = (k_1, k_2, \ldots, k_n)$ denote a multiindex then the variables defined by

$$\phi_{\mathbf{k}}(X) = \phi_{\mathbf{k}}(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \phi_{k_i}(X_i)$$

are orthonormal and there are constants $c(\mathbf{k})$ such that

$$S(X_1, X_2, \ldots, X_n) = \sum_{\mathbf{k}} c(\mathbf{k}) \phi_{\mathbf{k}}(X) \tag{2.3}$$

holds almost everywhere. Here we have just expressed S in what is sometimes called the tensor product basis for $L^2(dF\,dF \cdots dF)$. By orthonormality we have $ES^2 = \sum_{\mathbf{k}} c^2(\mathbf{k})$ and $\operatorname{var} S = \sum_{\mathbf{k} \ne \mathbf{0}} c^2(\mathbf{k})$. All we need now is to relate these identities to the right-hand side of (2.2).

Without loss of generality we can assume that $ES = 0$ so $c(\mathbf{k}) = 0$ if $\mathbf{k} = (0, 0, \ldots, 0)$. We first note that $E(S - S_i)^2 = 2ES^2 - 2ESS_i$. When $\hat{X}_i$ is substituted into (2.3) to give an expansion for S_i we see that

$$E(SS_i) = \sum_{\mathbf{k}:\, k_i = 0} c^2(\mathbf{k}), \tag{2.4}$$

since the orthonormality of the $\phi_{\mathbf{k}}(X_i)$ and the independence of $X_1, X_2, \ldots, X_i, \ldots, X_n$ and $\hat{X}_i$ cause all other summands to have expectation

zero. Summing over i we have

$$E\sum_{i=1}^{n}(S-S_i)^2 = 2nES^2 - 2\sum_{\mathbf{k}}c^2(\mathbf{k})z(\mathbf{k}), \tag{2.5}$$

where $z(\mathbf{k}) = \sum_{i=1}^{n} I(k_i = 0)$, i.e., $z(\mathbf{k})$ is equal to the number of indices k_i of the multiindex $\mathbf{k}$ which equal zero. Since we were able to assume that $c(\mathbf{0}) = 0$, we have $z(\mathbf{k}) \le n-1$ for all $c(\mathbf{k}) \neq 0$. This is a crucial observation, which applied to (2.5) gives us

$$E\sum_{i=1}^{n}(S-S_i)^2 \ge 2nES^2 - 2(n-1)\sum_{\mathbf{k}}c^2(k) = 2ES^2 \tag{2.6}$$

which completes the proof of the main result.

One can easily extend this result to the situation where one redraws m observations, i.e., one considers $S(X_1, \hat{X}_{i_1}, \ldots, \hat{X}_{i_m}, \ldots, X_n)$. We let $[m]$ denote a subset of $\{1, 2, \ldots, n\}$ of cardinality m and denote by $S_{[m]}$ the statistic obtained by replacing X_{i_j} by $\hat{X}_{i_j}$ for $i_j \in [m]$. The extension we seek for (2.1) is the following:

$$\binom{n-1}{m-1}\operatorname{var} S \le \tfrac{1}{2}E\sum_{[m]}(S-S_{[m]})^2. \tag{2.7}$$

The observations that we can assume $ES^2 < \infty$, $ES = 0$, and $S = \sum_{\mathbf{k}}c(\mathbf{k})\phi_{\mathbf{k}}(X)$ go on just as before. Now in calculating $ESS_{[m]}$ we get $\sum_{\mathbf{k}\in G[m]}c^2(\mathbf{k})$ where $G[m]$ is the set of all $\mathbf{k}$ such that $k_i = 0$ if $i \in [m]$. Hence using $E(S-S_{[m]})^2 = 2ES^2 - 2ESS_{[m]}$ and summing overall m subsets $[m]$ contained in $\{1, 2, \ldots, n\}$, we get

$$\begin{aligned}\tfrac{1}{2}\sum_{[m]}(S-S_{[m]})^2 &= \binom{n}{m}ES^2 - \sum_{\mathbf{k}}c^2(\mathbf{k})\binom{z(\mathbf{k})}{m} \\ &\ge \binom{n}{m}ES^2 - \binom{n-1}{m}\sum_{\mathbf{k}}c^2(\mathbf{k}) = \binom{n-1}{m-1}ES^2.\end{aligned}$$

This completes the proof of inequality (2.7).

Before attending to applications it is worth recording two key remarks.

(1) The arguments x_i of S need not be real numbers, or for that matter even vectors, although the case of vectors is by far the most important. The proofs of (2.1) and (2.7) depend on the structure of the arguments x_i only in the very shallow sense that we need $L^2(dF)$ to have a countable basis.

(2) The inequalities (2.1) and (2.7) are both sharp as one can see from choosing $S(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n}x_i$.

3. Applications to long common subsequences. A benefit of the two variance bounds of the last section is that they provide reasonably tight bounds on the variances of statistics which may be computationally difficult or even intractable. One illustration comes from the theory of string comparison.

The length of the longest common subsequence of the two random strings is a statistic which arises in an amazing variety of fields from biology to computer

science (see, e.g., Sankoff and Kruskal (1983)). Formally, we consider $2n$ independent, identically distributed random variables X_i and X_i', $1 \le i \le n$, which take values in a finite alphabet $\mathscr{A}$, and let

$$L_n := \max\{k\colon X_{i_1} = X'_{j_1}, X_{i_2} = X'_{j_2}, \ldots, X_{i_k} = X'_{j_k} \text{ where } 1 \le i_1 < i_2 < \cdots < i_k \le n \text{ and } 1 \le j_1 < j_2 < \cdots < j_k \le n\}.$$

Chvátal and Sankoff (1974) initiated the asymptotic study of EL_n, proved that $EL_n \sim cn$, and established some bounds on c. Much subsequent work has been done on c by Deken (1979) and Gordon and Arratia (unpublished).

The most intriguing special case is that of coin-flip sequences; that is where X_i and X_i' are independent random variables with success probability $\frac{1}{2}$. In that case the value $c = 2/(1 + \sqrt{2})$ is consistent with all of the known bounds and computational experience. This tidy expression was put forth by Richard Arratia and subsequently we found the following suggestive heuristic.

By a "good k pair" we will denote any pair of subsequences of length k of the X_i and the X_i' which coincide. We let Z denote the total number of good k pairs which can be found in the two n strings of the X_i and X_i', $1 \le i \le n$. The expectation of Z is easily determined,

$$E(Z) = 2^{-k}\binom{n}{k}^2$$

and, by considering ratios of successive choices of k, it is easy to see that $E(Z)$ is unimodal and the mode occurs for that integer value of k nearest $n/(1 + \sqrt{2})$.

We can also get a handle on the number of good k pairs by noting that every k subsequence of the longest common subsequence gives a rise to good k pair, so there are at least $\binom{L_n}{k}$ of them. Now by the usual unimodality of the binomial coefficients, as k varies this sequence is unimodal with mode equal to the nearest integer to $L_n/2$. The heuristic leap of faith is that in expectation these two modes are within a distance of $o(n)$ of each other. A proof of that leap would prove that Arratia's suggested value for c is the correct one.

A second interesting problem concerning L_n is the conjecture of Chvátal and Sankoff (1974) that $\operatorname{var}(L_n) = O(n^2/3)$. It was put forth in Steele (1982a) that $\operatorname{var}(L_n) \le (n^{1/2} + 1)^2$. As an illustration of the power of (2.1), we can now give an easy proof of the stronger result,

$$\operatorname{var} L_n \le n; \tag{3.1}$$

in fact we can show

$$\operatorname{var} L_n \le n\Big(1 - \sum_{\alpha \in \mathscr{A}} p_\alpha^2\Big), \tag{3.2}$$

where $p_\alpha = P(X_i = \alpha) = P(\hat{X}_i = \alpha)$, for $\alpha \in \mathscr{A}$. Since $\sum_{\alpha \in \mathscr{A}} p_{\alpha=1}$ we have $\sum_{\alpha \in \mathscr{A}} p_\alpha^2 \ge |\mathscr{A}|^{-1}$. To get (3.2) from (2.1) we consider $S = L(X_1, X_2, \ldots, X_n, X_1', X_2', \ldots, X_n')$ and consider S as a (nonsymmetric) function of $2n$ variables. Changing any one of those variables will change S by at most one. Moreover if, say, X_i is replaced by $\hat{X}_i$ then $P(X_i = \hat{X}_i) = \sum p_\alpha^2$ so $P(S = S_i) \ge \sum p_\alpha^2$. These

two facts give us $E(S - S_i)^2 \leq 1 - \Sigma p_\alpha^2$ and there are $2n$ such summands, so we have established (3.2). This is a long way from the conjectured $\operatorname{var} L_n = o(n^{2/3})$, but it is the sharpest known result. The ease with which it comes from (2.1) is surprising if one initially studies L_n from a combinatorial perspective like that of Chvátal and Sankoff (1974).

It is tempting to try to improve (3.2) by use of the change-m inequality (2.7). To do so would require improving the naive bound

$$E(S - S_{[m]})^2 \leq m^2,$$

since this bound leads only to a variance bound given by the ratio $m(n-1)(n-2)\cdots(n-m+1)/(m-1)!$. This bound is not even linear in n for fixed $m \geq 2$, and trying to optimally choose m for fixed n does not help since the bound is minimized for $m = 1$. One seems to need new combinatorial insights to improve the naive bound on $(S - S_{[m]})^2$, and thus to use (2.7) with effect.

It is likely that (2.7) is in fact sharper than (2.1). Karlin and Rinott (1982) found that to be the case with the Bhargava (1980) version of the original Efron–Stein inequality and there is no reason to expect our version to break with tradition.

The hard part of using (2.7) is not a lack of sharpness but rather an excess of complexity. One has to find a way to get strong information on how L_n changes as one changes a substantial part of the sample. This is harder than getting a decent bound on the possible variation due to changing a single observation.

The longest common subsequence problem has a natural analogue for k strings for which much of the preceding theory goes through with little change. One benefit of the k string analogue is that it gives a second handle on the constant c.

To define the simplest incidence of the k-sequence problem we consider $k = 3$; let $Y_i = (X_i, X_i', X_i'')$ and set

$$L_n := \max\{t\colon X_{i_1} = X'_{j_1} = X''_{k_1}, \ldots, X_{i_t} = X'_{j_t} = X''_{k_t}\},$$

where $1 \leq i_1 < i_2 < \cdots < i_t \leq n$, $1 \leq j_1 < \cdots < j_t \leq n$, $1 \leq k_1 < \cdots < k_t \leq n$. The same proof that Chvátal and Sankoff (1974) give for $k = 2$ will show that

$$\lim_{n\to\infty} EL_n/n = c_3$$

and the same proof given by Deken (1979) shows that $L_n/n \to c_3$, almost surely. It would be of interest to relate c_3 to c, and one is tempted to speculate that $c_3 = c^2$ (and more generally that $c_k = c^{k-1}$). Computational evidence does not yet rule this out. The application of (2.7) to this new functional gives $\operatorname{var} L_n \leq \frac{3}{2}n(1 - \Sigma p_\alpha^2)$ in the case $k = 3$. Again, this seems difficult to improve.

Acknowledgments. I am indebted to Michael Waterman and Louis Gordon for stimulating this work, and to Richard Arratia for his comments on $c = 2/(1 + \sqrt{2})$.

REFERENCES

BHARGAVA, R. P. (1980). A property of the jackknife estimation of the variance when more than one observation is omitted. Technical Report No. 140, Dept. of Statistics, Stanford University.

CHVÁTAL, V. and SANKOFF, D. (1975). Longest common subsequences of two random sequences. *J. Appl. Probab.* **12** 306–315.

DEKEN, J. P. (1979). Some limit results for longest common subsequences. *Discrete Math.* **26** 17–31.

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

KARLIN, S. and RINOTT, Y. (1982). Application of ANOVA type decomposition for comparisons of conditional variance statistics including jackknife estimates. *Ann. Statist.* **10** 485–501.

SANKOFF, D. and KRUSKAL, J. B., eds. (1983). *Time Warps, String Edits, and Macromolecules: The theory and practice of sequence comparison.* Addison-Wesley, Reading, Mass.

STEELE, J. M. (1981). Complete convergence of short paths and Karp's algorithm for the TSP. *Math. Oper. Res.* **6** 374–378.

STEELE, J. M. (1982a). Long common subsequences and the proximity of two random strings. *SIAM J. Appl. Math.* **42** 731–737.

STEELE, J. M. (1982b). Optimal triangulations of random samples in the plane. *Ann. Probab.* **10** 548–553.

VITALE, R. A. (1984). An expansion for symmetric statistics and the Efron–Stein inequality. In *Inequalities in Statistics and Probability*, IMS Lecture Notes—Monograph Series **5** (Y. L. Tong, ed.) 112–114. IMS, Hayward, Calif.

DEPARTMENT OF STATISTICS
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544