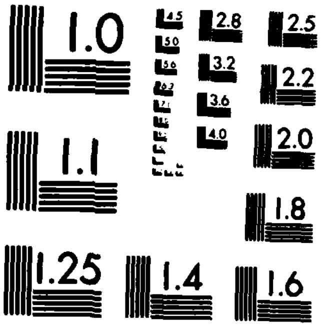
AD-A146 673 EFFICIENT SCORES VARIANCE DECOMPOSITIONS AND MONTE CARLO SWINDLES(U) STANFORD UNIV CA DEPT OF STATISTICS I JOHNSTONE ET AL. 28 AUG 84 TR-348 N00014-76-C-0475 1/1

UNCLASSIFIED F/G 12/1 NL



COPY RESOLUTION TEST CHART

EFFICIENT SCORES, VARIANCE DECOMPOSITIONS AND MONTE CARLO SWINDLES

BY

IAIN JOHNSTONE and PAUL VELLEMAN

TECHNICAL REPORT NO. 348

AUGUST 28, 1984

PREPARED UNDER CONTRACT
N00014-76-C-0475 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

EFFICIENT SCORES, VARIANCE DECOMPOSITIONS AND MONTE CARLO SWINDLES

BY

IAIN JOHNSTONE and PAUL VELLEMAN

TECHNICAL REPORT NO. 348
AUGUST 28, 1984

Prepared Under Contract
N00014-76-C-0475 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

1. INTRODUCTION

Monte Carlo studies are experiments that use simulated data. Like any experiment, they should be designed to minimize extraneous variation. However the Monte Carlo experimenter, unlike the designer of traditional field experiments, usually knows and can control the stochastic structure of the simulated data. "Swindles" or variance reduction techniques exploit this knowledge to construct more precise estimates of the unknown parameters or to reduce the number of simulation runs (and thus the cost) necessary to attain some desired level of precision. Because swindle designs use information that is not usually available in field experiments, they sometimes appear to provide an unfair reduction in the variance of estimated quantities (hence their name). Indeed, the improvement in precision and the attendant reduction in cost can be so great that a well-designed swindle can make feasible Monte Carlo studies that might otherwise have been impossible. In view of this, it is unfortunate that many Monte Carlo studies do not employ variance reduction methods. This may be due in part to the relatively restricted applicability of standard swindle methods or to a lack of awareness of the methods.

Perhaps the most common application of Monte Carlo swindles in statistics has been in estimating variances or mean squared errors. A simulation study of variances often forms the basis for comparing the small-sample efficiencies of a collection of, say, robust or resistant estimators. A typical and traditional question might be "How much more efficient is a 10% trimmed mean than the sample mean in samples of 20 from a heavy-tailed distribution?" The Princeton Robustness Study (Andrews _et al._, 1972) provides a large and well-known example. On a smaller scale, such studies are now a routine part of the scrutiny of statistical procedures.

Our principal purpose in this paper is to unify and extend the treatment of swindles for estimating variances in the hope that they may then be applied more easily and widely. To do this, we (1) propose a new swindle (based on Fisher's efficient score function) that is simpler, more effective, and more general than the current "Gaussian-over-independent" (G/I) method commonly used in variance estimation, (2) provide examples of how this swindle can be applied, and (3) describe how this and other swindles for variances fit into a familiar geometric framework (which in turn suggests further applications).

Section 3 presents the score function swindle first in the simplest case: for location equivariant estimators in the location problem with known scale. Section 4 presents a simple application to the problem of estimating the variance of the Pitman estimator of location in small samples drawn from Student's t distributions (and reports new results for this problem). In Section 5 we compare the score function swindle numerically with the standard Gaussian-over-independent swindle used in the Princeton study. We examine the relative swindle gains for the two methods for a variety of location estimators and distributions in the t-family, and find that in most cases the score function technique dominates. To facilitate the comparison, Section 2 presents some general issues in assessing swindle gains, then outlines and discusses the Gaussian-over-independent swindle.

Hammersley and Handscomb (1964) and Rubinstein (1981) discuss general principles of variance reduction methods. Simon (1976) surveys applications of swindles to simulation studies in statistical research.

The common swindles for estimating the variance of a statistic $T(\underset{\sim}{Y})$ exploit a simple variance decomposition

$$\text{Var}\,T = \text{Var}\,S + \text{var}(T - S)\ , \qquad (1.1)$$

in which S and $T - S$ are uncorrelated and $\text{Var}\,S$ is either known from the distribution of Y or can be easily approximated. Ideally S should be highly correlated with T, for then $\text{Var}(T - S)$ will be small and hence, in general, more precisely estimable. A useful way to obtain such decompositions is to identify an affine subset $\mathcal{J}$ of statistics with finite variance to which T belongs and take S as a minimum variance element of $\mathcal{J}$. This is discussed further in Section 6, where it is shown how both the score-function and G/I swindles and a number of other swindles discussed in the literature may be obtained by varying the choices of $\mathcal{J}$. Further applications of swindles based on decomposition (1.1) to statistical decision theory (frequentist and Bayesian), and of the score function swindle in particular to multivariate, discrete and bootstrap location problems are outlined in Section 7.

The 'regression estimate' of sampling theory (Cochran 1977, ch. 7) suggests a variance reduction method which in some senses is simpler and more widely applicable than the above approach. Let $\hat{\sigma}_T^2$ be unbiased for Var T. Suppose there is available another statistic S , preferably highly correlated with T , for which Var S is known and possesses an unbiased estimate $\hat{\sigma}_S^2$. Then

$$(1.2) \qquad \tilde{\sigma}_T^2 = \hat{\sigma}_T^2 + b(\hat{\sigma}_S^2 - \sigma_S^2)$$

is an unbiased estimate of σ_T^2 for all b, and has smaller variance than $\hat{\sigma}_T^2$ for some interval of b values about the optimum $b^* = -\text{Cov}(\hat{\sigma}_S^2, \hat{\sigma}_T^2)/\text{Var}(\hat{\sigma}_S^2)$. However, the optimum b* will rarely be known, and will thus need to be estimated, introducing a bias to $\tilde{\sigma}_T^2$. On the other hand, it is not necessary to have S and T - S uncorrelated for the method to be applicable as was needed for (1.1). A normal theory calculation in Remark 8A suggests that the decomposition (1.1), when available, leads to larger swindle savings than (1.2).

This paper focuses on methods for increasing the precision of variance estimates. Often comparisons of variance estimates in the form $\hat{\text{Var}}\, T_1/\hat{\text{Var}}\, T_2$ or $\hat{\text{Var}}\, T_1 - \hat{\text{Var}}\, T_2$ are sought and assessments of swindle gains will of course differ in these cases. Without attempting a systematic discussion, we give in Remark 8B an assessment of the swindle gains from the variance decomposition (1.1) in the ratio case. The crude comparison of (1.1) and the 'regression estimate' (1.2)

(Remark 8A) can be extended to the case of differences $\hat{\text{Var}}\ T_1 - \hat{\text{Var}}\ T_2$, with normal-theory calculations suggesting that (1.1) dominates when the correlation between S and T_1, and between S and T_2 is fairly strong and is of the same order or stronger than the correlation between T_1 and T_2.

ACKNOWLEDGEMENTS

The geometric perspective on construction of variance reduction methods is given in mimeographed lecture notes of Charles Stein, as is the argument leading to (8.4). We thank him, Persi Diaconis and Bruce Spencer for helpful discussions.

2. BACKGROUND

2.1 Measuring Swindle Gains

Suppose that a decomposition (1.1) holds and that Var S is known. We measure the gain in precision (or equivalently the reduction in number of experiment replications needed) by comparing $\text{Var}\,\widehat{\text{Var}}\,T$ to $\text{Var}\,\widehat{\text{Var}}(T - S)$.

More specifically, assume that $ET(\underset{\sim}{Y}) = ES(\underset{\sim}{Y}) = 0$ and that naive method-of-moments estimators are used for Var T and $\text{Var}(T - S)$. For example, Var T can be estimated by $\frac{1}{N}\sum_{J=1}^{N} T^2(\underset{\sim}{Y}^J)$, where J indexes replications of the sample $\underset{\sim}{Y} = (Y_1,\ldots,Y_n)$. Then, of course $\text{Var}\,\widehat{\text{Var}}\,T = \frac{1}{N}\,\text{Var}\,T^2(\underset{\sim}{Y})$, and from the identity

$$\text{Var}\,T^2 = (ET^2)^2[ET^4/(ET^2)^2 - 1] ,$$

we obtain

$$\frac{\text{Var}\,\widehat{\text{Var}}\,T}{\text{Var}\,\widehat{\text{Var}}(T - S)} = \left|\frac{1}{1 - \rho^2(T,S)}\right|^2 \left[\frac{\kappa(T) - 1}{\kappa(T - S) - 1}\right] . \tag{2.1}$$

Here $\kappa(T) = (E\,T^4)/(E\,T^2)^2$ denotes the kurtosis of T . The squared correlation of T and S , $\rho^2(T,S)$, equals the relative efficiency Var S/Var T whenever S is minimum variance in a linear class containing T, as it will be in our applications (Section 6 has a proof). Thus, subject to comparability of $\kappa(T)$ and $\kappa(T\text{-}S)$, the efficiency of the swindle increases quadratically as the squared correlation of T and S approaches one. Note also that the ratio (2.1) can be interpreted as the factor $N_T/N_{T\text{-}S}$ by which the number of replications to achieve a desired precision is reduced by using the swindle.

In measuring swindle gains, one should also assess the relative costs of computing T and $T-S$. Since these will typically depend on the algorithm and the machine, we will not indicate these comparisons explicitly. However if, for example, S is a Pitman estimator, the extra effort involved in finding S may be so great as to render the swindle impractical.

2.2 Gaussian Over Independent Swindle

This swindle was introduced by Dixon and Tukey (1968) and Relles (1970) and applied extensively in the Princeton study. To date it has mainly been used for location problems: Simon (1976) gives a survey discussion in this setting. We outline it here in a (more general) regression setting in which at the same time the method seems more natural (cf. also Goodfellow and Martin (1976)). Johnstone and Velleman (1984) use this (and the score function swindle of Section 3) in a small-sample comparison of several resistant simple linear regression methods.

Suppose that observations are drawn from a linear model, $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ where $\underset{\sim}{Y}$ is an $n \times 1$ column vector, X is a fixed $n \times p$ matrix of carriers, $\underset{\sim}{\beta}$ a $p \times 1$ parameter vector and $\underset{\sim}{\varepsilon}$ an $n \times 1$ vector of i.i.d. variables Z_i/W_i drawn from a Gaussian-over-independent distribution; i.e. $Z_i \sim N(0,\sigma^2)$, and the W_i are i.i.d. positive and independent of Z_i. (Table 1 lists some distributions in the Z/W family.) Suppose that $T(X,\underset{\sim}{Y})$ is a regression-invariant estimator of β: $T(X, c\underset{\sim}{Y} - X\underset{\sim}{d}) = cT(X,\underset{\sim}{Y}) - \underset{\sim}{d}$ for any $c \in \mathbb{R}_+$, $\underset{\sim}{d} \in \mathbb{R}^p$. We seek a variance decomposition for $\operatorname{Var} T$.

The denominators, W_i, constitute extra information available to the simulation (but not available in real data when only (X_i,Y_i) are observed. Here they can be used to construct an estimator with known variance. Indeed, conditional on W_i, $\underset{\sim}{\beta}$ and σ^2 can be estimated by standard weighted least squares estimates $\hat{\underset{\sim}{\beta}}_W$ and $\hat{\sigma}^2_W$, the former having covariance matrix $\sigma^2(X'\Lambda^2 X)^{-1}$ where $\Lambda = \mathrm{diag}(W_i)$. From the normal theory assumptions on Z_i and conditional on W_i, it follows that $(\hat{\beta}_W, \hat{\sigma}^2_W)$ are complete sufficient statistics for $(\underset{\sim}{\beta}, \sigma^2)$, and that the standardized residuals $\hat{\underset{\sim}{e}} = (\underset{\sim}{y} - X\hat{\underset{\sim}{\beta}}_W)/\hat{\sigma}_W$ are ancillary. Basu's sufficiency-ancillarity theorem (e.g. Simon 1976, Lehmann 1983 p.46) ensures independence of the triple $(\hat{\underset{\sim}{\beta}}_W, \hat{\sigma}^2_W, \hat{\underset{\sim}{e}})$. Using the decomposition $\underset{\sim}{y} = X\hat{\underset{\sim}{\beta}}_W + \hat{\sigma}_W\hat{\underset{\sim}{e}}$ and regression invariance of $\hat{\beta}$, we obtain

$$T(\underset{\sim}{y}) = \hat{\underset{\sim}{\beta}}_W + \hat{\sigma}_W T(\hat{\underset{\sim}{e}}) \quad .$$

Suppose also that T is unbiased. (This will happen for example if regression invariance holds for negative values of c also.) If the distribution used to generate the data satisfies $\sigma^2 = 1$, $\underset{\sim}{\beta} = 0$, then conditional on W_i and using independence,

$$\underset{\sim}{\mathrm{Var}}\ (T(\underset{\sim}{y})|\underset{\sim}{W}) = (X'\Lambda_W^2\ X)^{-1} + \underset{\sim}{\mathrm{Var}}\ (T(\underset{\sim}{e})|\underset{\sim}{W}).$$

Now take expectation with respect to $\underset{\sim}{W}$, and use the fact that $E(T(\underset{\sim}{y})|\underset{\sim}{W}) = 0$, $E(T(\underset{\sim}{e})|\underset{\sim}{W}) = 0$ to obtain finally

$$\underset{\sim}{\mathrm{Var}}\ T(\underset{\sim}{y}) = E(X'\Lambda_W^2 X)^{-1} + \underset{\sim}{\mathrm{Var}}\ T(\hat{\underset{\sim}{e}}) \quad .$$

In many cases, the first term can be evaluated analytically, numerically, or once-and-for-all by Monte Carlo, while $\underset{\sim}{\mathrm{Var}}\ T(\underset{\sim}{\hat{e}})$, being smaller than $\underset{\sim}{\mathrm{Var}}\ T(\underset{\sim}{y})$, can,in principle,be estimated more accurately (c.f. equation (2.1)).

2.3 Limitations of the Z/W Swindle

The Gaussian-over-independent swindle depends crucially on the Z/W representation for the distribution of the underlying data. While the class of distribution that arise as variance mixtures of normals is rich (see, for example, Andrews and Mallows, 1974 and Efron and Olshen 1978), it is only a subset of the symmetric distribtuions. Even within this subset the gains realized from the swindle tend to decrease for Z/W distributions with heavy tails.

This phenomenon can be accounted for in part by restrictions on the first term in the swindle gain (2.1). Heavy tailed Z/W distributions must have some W_i far from unity. Knowledge of W will then convey more information about the sample, leading the variance of $\hat{\underset{\sim}{\beta}}_W$ to fall further below the smallest attainable variance. This bounds $\rho^2(T, \hat{\underset{\sim}{\beta}}_W)(= \mathrm{Var}\ \hat{\underset{\sim}{\beta}}_W/\mathrm{Var}\ T)$ away from 1.0 .

Figure 1 illustrates this effect for Student's t distributions (for which $W_i \sim \sqrt{\chi^2_\nu/\nu}$). The Pitman variances in Figure 1 are the smallest attainable among invariant estimators of location. All other estimators would thus appear above the Pitman values. (See Appendix A for the Pitman variances and notes on how they were estimated.)

While we might ideally wish to swindle relative to the Pitman estimators (see, e.g. Andrews et. al., 1972, p. 61 and Section 5 below), the expense of computing them would cancel much of the swindle gain. However, as Figure 1 shows, often the Pitman efficiency is not far from the Cramer-Rao bound. This fact motivates the method of the next section, in which "estimators" with variance equal to the Cramer-Rao bound are used to obtain higher correlation with $T(y)$ while still possessing a variance decomposition.

3. THE SCORE FUNCTION SWINDLE

Suppose now that $Y_1,\ldots,Y_n$ are independent random variables and that Y_i has density $f_i(y,\underset{\sim}{\theta})$, $\underset{\sim}{\theta} \in \Theta \subset R^p$. (Often the f_i will be identical). We shall assume that the densities f_i are smooth enough to permit the manipulations below (the "Cramer conditions" (Cramer, 1946) would more than suffice). Suppose that $T(\underset{\sim}{Y})$ is unbiased for θ_1, at least up to a constant:

$$E_{\underset{\sim}{\theta}}\, T(\underset{\sim}{Y}) = \theta_1 + c_0 \,, \qquad \underset{\sim}{\theta} \in \Theta \ .$$

Given an arbitrary vector-valued statistic $\underset{\sim}{S}$, the linear combination $\underset{\sim}{c}^*\underset{\sim}{S}$ having maximum correlation with T is just the (population) linear regression of T on $\underset{\sim}{S}$, $\underset{\sim}{c}^* = \ddagger_S^{-1}\, \underset{\sim}{\sigma}_{TS}$, where $\ddagger_S = \text{Cov}(\underset{\sim}{S})$ and $\underset{\sim}{\sigma}_{TS} = \text{Cov}(T,\underset{\sim}{S})$. The resulting variance decomposition is

$$\text{Var}\, T = \underset{\sim}{\sigma}_{TS}'\, \ddagger_S^{-1}\, \underset{\sim}{\sigma}_{TS} + \text{Var}(T - \underset{\sim}{c}^*\, \underset{\sim}{S}) \ .$$

In general $\underset{\sim}{\sigma}_{TS}$ will be no easier to estimate than Var $\underset{\sim}{S}$. However, the score function, $\underset{\sim}{S}(\underset{\sim}{Y},\underset{\sim}{\theta}) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(y_i, \underset{\sim}{\theta})$ provides a statistic $\underset{\sim}{S}$ for which $\underset{\sim}{\sigma}_{TS}$ is simple, and yields a random vector with high (multiple) correlation with T when T has variance close to the Cramer-Rao lower bound. To see this, differentiate the relation $E_{\underset{\sim}{\theta}}\, T(\underset{\sim}{Y}) = \theta_1 + c_0$ to obtain

$$\delta_k^1 = \int T(\underset{\sim}{y}) \frac{\partial}{\partial\theta_k} \prod_{i=1}^{n} f(y_i, \underset{\sim}{\theta})\, dy_i$$

$$= E_{\underset{\sim}{\theta}}\, T(\underset{\sim}{Y})\, S_k(\underset{\sim}{Y}, \underset{\sim}{\theta}) \quad .$$

Recalling that $E_{\underset{\sim}{\theta}} S_k(\underset{\sim}{Y}, \underset{\sim}{\theta}) = 0$, we find by fixing $\underset{\sim}{\theta}$ at the value (say $\underset{\sim}{\theta}_0$) used in generating the data that $\underset{\sim}{S} = S(\underset{\sim}{Y}, \underset{\sim}{\theta}_0)$ is a "statistic" with the property $\underset{\sim}{\sigma}_{\underset{\sim}{TS}} = \underset{\sim}{e}_1 = (1,0,\ldots,0)$, and hence

$$(3.1) \qquad \mathrm{Var}\, T = \underset{\sim}{e}_1' \ddagger_S^{-1} \underset{\sim}{e}_1 + \mathrm{Var}(T - \underset{\sim}{e}_1' \ddagger_S^{-1} \underset{\sim}{S}) \quad .$$

Since $\ddagger_S$ depends only on the densities f_i it is, in principle, known or calculable, and the Monte Carlo can be restricted to estimation of $\mathrm{Var}(T - \underset{\sim}{e}_1' \ddagger_S^{-1} \underset{\sim}{S})$.

Remarks: a) This approach can be extended by taking higher derivatives of the likelihood function. We discuss in Remark 7C the amount gained by the more refined swindles that result.

b) Some proposals for the use of score functions and Cramer-Rao bounds are discussed in the setting of simultaneous simulation of two variances in Appendix B of Andrews et al. (1972).

Examples

1) Location. Let Y_i have density $f(y - \theta)$ for f positive and piecewise C^1 on $\mathbb{R}$. Let $T(\underset{\sim}{Y})$ be a location equivariant estimator: $T(Y_1 + c,\ldots,Y_n + c) = T(\underset{\sim}{Y}) + c$. Then clearly $E_\theta T(\underset{\sim}{Y}) = \theta + c_0$, where

where $c_0 = E_0(T(\underset{\sim}{Y})$ and $\theta_0 = 0$, $S = -\sum_1^n f'/f(Y_i)$ and $\ddagger_S = n\,I(f)$, the Fisher information (for location) of the density f. Thus we have the decomposition

$$(3.2) \qquad \mathrm{Var}\ T(\underset{\sim}{Y}) = 1/(nI(f)) + \mathrm{Var}\ T(\underset{\sim}{Y} - \tilde{\theta}(\underset{\sim}{Y})\underset{\sim}{1})$$

where $\tilde{\theta}(\underset{\sim}{Y}) = (nI(f))^{-1}S$. Thus if one knows both f'/f and $I(f) = E(f'/f)^2$, then the swindle simply bases the Monte Carlo estimator on the data $\underset{\sim}{Y}$ centered by $\tilde{\theta}(\underset{\sim}{Y})$. Note that S is not in general a location-equivariant estimator itself (in fact it will be if, and only if, f is Gaussian!), but this is irrelevant to the swindle calculation. Significantly, there is no need for f to be symmetric.

The score function swindle includes situations in which the data are not an i.i.d. sample. A common example is the "one-wild" sampling scheme in location problems (Andrews et.al., 1972; Hoaglin, Mosteller and Tukey, 1983, Chs. 10,11) in which $n-1$ observations are drawn from $f(y)$ and one from $(1/\sigma_0)f(y/\sigma_0)$ for σ_0 known and large. The score function $S(\underset{\sim}{Y},\theta) = -\phi(y_1/\sigma_0)/\sigma_0 - \Sigma_2^n\,\phi(y_i)$ and $\ddagger_S = (n-1+\sigma_0^{-2})\ I(f)$.

Suppose now that f includes scale as a nuisance parameter, Y_i having density $f(\frac{Y-\mu}{\sigma})$. Now $\underset{\sim}{\theta} = (\mu,\sigma)$ and if $T(\underset{\sim}{Y})$ is a location and scale equivariant estimator of μ then $E_\theta T(\underset{\sim}{Y}) = \mu + \sigma\, E_{0,1}\,T(\underset{\sim}{Y})$, where $\underset{\sim}{\theta}_0 = (0,1))$. Here the score function $\underset{\sim}{S} = (\Sigma_1^n\,\phi(Y_i),\ \Sigma_1^n\,Y_i\,\phi(Y_i))$, where $\phi(y) = -f'/f(y)$, and

$$\ddagger_S = n \begin{pmatrix} E\,\phi^2 & E\,Y\phi^2 \\ E\,Y\phi^2 & E\,Y^2\phi^2 \end{pmatrix}$$

Thus, $\operatorname{Var} T_1(\underset{\sim}{Y}) = \frac{1}{n} \dfrac{E\,Y^2\,\phi^2}{E\,\phi^2\,E\,Y^2\,\phi^2 - (E\,Y\,\phi^2)^2}$

$$+ \operatorname{Var} T(\underset{\sim}{Y} - e_1' \Sigma_S^{-1} \underset{\sim}{S}\,\underset{\sim}{1}) \quad . \tag{3.3}$$

In general, one would expect that as the number of nuisance parameters increases and the unbiasedness condition becomes more stringent, the Cramer-Rao variance bound -- the first item in (3.1) -- would increase, thus giving a better swindle. For example, if $f(y) = ce^y\, I\{y<0\} + ce^{-y^2/2}\, I\{y\geq 0\}$, the Cramer-Rao bound is easily calculated to be 1.516 times the bound obtained in (3.1) without the nuisance parameter for scale.

Note, however, that if f is <u>symmetric</u> about 0, then $E\,Y\phi^2(Y) = 0$ and the above decomposition reduces to (3.2), so that the swindle does not gain by including the scale parameter. This is an instance of a more general phenomenon: if the orginal estimation problem for θ_1 satisfies Stein's necessary condition for adaptation, then the swindle cannot be improved by adding a finite number of nuisance parameters to the model. In the present context, Stein's condition simply requires that $\operatorname{Cov}(S_1, S_k) = 0$ for $k \geq 2$ where $(S_1, S_2, \ldots, S_k)$ is the score function vector for the augmented model. For more information on adaptive (asymptotic) estimation see Stein (1956), Bickel (1982).

2) <u>Regression</u>. For simplicity we discuss here only estimation of slope in simple linear regression, though the ideas generalize to arbitrary linear models. Suppose then that we draw n observations from the model $Y_i = \alpha + \beta(x_i - \overline{x}) + \epsilon_i$ where the x_i are fixed and

ε_i are i.i.d. according to some smooth positive density f. (If f is symmetric, we would gain nothing by including a nuisance parameter for scale). Suppose that $T(\underset{\sim}{y})$ is a regression-invariant estimator of β: $T(\underset{\sim}{y} - b(\underset{\sim}{x} - \bar{x}\underset{\sim}{1})) = T(\underset{\sim}{y}) - b$. In previous notation, $\underset{\sim}{\theta} = (\beta, \alpha)$, $\underset{\sim}{\theta}_0 = (0,0)$, $\underset{\sim}{S} = -(\Sigma_1^n (x_i - \bar{x})\, \phi(y_i),\ \Sigma_1^n \phi(y_i))$, and because the x_i are centered, Cov $\underset{\sim}{S}$ = diag$(n\, \sigma_X^2\, I(f),\ nI(f))$, where $\sigma_X^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$. Thus $\underset{\sim}{e}_1' \ddagger_S^{-1} \underset{\sim}{S} = -(n\, \sigma_X^2\, I(f))^{-1}\, \Sigma_1^n (x_i - \bar{x})\, \phi(y_i)$, which in the special case $\varepsilon_i \sim N(0,1)$ is just the least squares estimate of β. Finally $\underset{\sim}{e}_1' \ddagger_S^{-1} \underset{\sim}{e}_1 = [n\, \sigma_X^2 I(f)]^{-1}$. The swindle has obvious extensions to heteroscedastic situations in which, say, Var(ε) varies with x, or to cases in which the X_i *themselves* are a random sample from a distribution. This method was used extensively in the regression study of Johnstone and Velleman (1984).

3) <u>Scale estimation</u>. If $S(Y_1, \ldots, Y_n)$ is a scale-equivariant estimate: $S(cY_1, \ldots, cY_n) = cS(\underset{\sim}{Y})$ and the Y_i are i.i.d. from a density $f(y/\sigma)$ with location known, then $E_\sigma \log S(\underset{\sim}{Y}) = \log \sigma + E_1 \log S(\underset{\sim}{Y})$ and $\log S(\underset{\sim}{Y})$ is unbiased (up to a constant) for $\log \sigma$.

Now it is often argued that Var $\log(\underset{\sim}{Y})$ is an informative measure of performance of $S(\underset{\sim}{Y})$ (see e.g. Simon, 1976, §5). To apply the location score function swindle, set $\sigma' = \log \sigma$ and let $p(y, \sigma') = f(y\, e^{-\sigma'})$. The score function evaluated for $\sigma_0' = 0$ equals $-\sum_1^n y_i f'/f(y_i)$ and the variance decomposition becomes

$$\text{Var} \log S(\underset{\sim}{Y}) = \frac{1}{n\, E\, Y^2 \phi^2(Y)}$$

$$+ \text{Var} \left\{ \log \left[S(\underset{\sim}{Y}) + \frac{\Sigma_i^n Y_i\ \phi(Y_i)}{n\, E\, Y^2 \phi^2(Y)} \right] \right\} .$$

If location is treated as a nuisance parameter, the above analysis can be modified along the lines of the second part of Example 1.

4) <u>Exponential Regression Models</u>. Suppose that $T_1,\ldots,T_n$ are (positive) survival times with associated (scalar, for convenience) covariates $z_1,\ldots,z_n$ and exponential hazard rate

$$\lambda(z,t) = \exp(z\beta) \quad .$$

An appropriate notion of equivariance for an estimator $\tilde{\beta}(z_1,\ldots,z_n; t_1,\ldots,t_n)$ of β is

$$(3.4) \qquad \tilde{\beta}(\underset{\sim}{z};\ e^{\gamma \underset{\sim}{z}}\ \underset{\sim}{t}) = \tilde{\beta}(\underset{\sim}{z},\underset{\sim}{t}) - \gamma\ , \qquad \gamma \in \mathbb{R}$$

where $e^{\gamma \underset{\sim}{z}}\ \underset{\sim}{t}$ denotes $(e^{\gamma z_1} t_1,\ldots,e^{\gamma z_n} t_n)$. This is, for example, satisfied by the MLE, which solves the likelihood equation

$$\sum_1^n z_i(1 - e^{\beta z_i} t_i) = 0 \ .$$

it follows for any estimator $\tilde{\beta}$ satisfying (3.4), that for fixed $\underset{\sim}{z}$, $E_\beta\, \tilde{\beta}(\underset{\sim}{z},\underset{\sim}{T}) = c_o + \beta$ and hence from (3.1)

$$(3.5) \qquad \mathrm{Var}_\beta\, \tilde{\beta} = \mathrm{Var}_\beta\, S + \mathrm{Var}_\beta(\tilde{\beta} - S)$$

where $S = (\Sigma\, z_i^2)^{-1}\, \Sigma\, z_i(1 - t_i\, e^{\beta z_i})$ and $\mathrm{Var}\, S = (\sum_i^n z_i^2)^{-1}$.

Extension to arbitrary (but known) baseline hazard rate is straightforward but perhaps restrictive: if $\lambda(z,t) = \theta(t)e^{z\beta}$, $t > 0$ for $\theta > 0$ known and $\Theta(t) = \int_0^t \theta(s)\, ds$, then (3.5) remains valid for estimators $\ddot{\tilde{\beta}}$ (such as the MLE) which satisfy $\tilde{\beta}(\underset{\sim}{z}, \Theta^{-1}(e^{\gamma \underset{\sim}{z}}\, \Theta(\underset{\sim}{t}))) = \tilde{\beta}(\underset{\sim}{z},\underset{\sim}{t}) - \gamma$, if t_i in the definition of S is replaced by $\Theta(t_i)$.

4. AN EXAMPLE: PITMAN VARIANCES

This section illustrates the use of the score function swindle in a simple but instructive situation: the computation of the variances of the Pitman estimators of location under sampling from distributions in the $t_{(\nu)}$-family. The Pitman estimator of location θ based on n i.i.d. observations from a density $f(x-\theta)$ on $\mathbb{R}$ (see (4.1) below) has minimum variance amongst all location equivariant estimators (Pitman, 1939). It is thus a natural baseline against which to measure the relative efficiencies of other location-equivariant estimators. In general, however, the variances of Pitman estimators cannot be evaluated analytically. Hoaglin (1975) reports on numerical evaluations of Pitman variances for selected small sample sizes from three particular distributions (including the Cauchy, or $t_{(1)}$). For the $t_{(\nu)}$ family used in our swindle comparison experiments in Section 5, no other estimates of Pitman variances seem to be available in the literature.

Our Monte Carlo trials to obtain the Pitman variances are in principle a straightforward application of the score-function swindle in the form (3.2). In fact, (1.1) and (2.1) reveal that we are in the situation where this swindle is most effective: since the Pitman estimator is the minimum variance location estimator, it has maximum possible correlation (among location estimators) with the score function statistic, whose variance is much more readily evaluated.

The Pitman estimator of θ based on n observations from the $t_{(\nu)}$ distribution is given by

$$(4.1) \qquad d_p(\underset{\sim}{y}) = \int \theta \, \Pi_{i=1}^n \, f_\nu(y_i-\theta)d\theta / \int \Pi_{i=1}^n \, f_\nu(y_i-\theta)d\theta \ .$$

where $f_\nu(y) = c_r(1+y^2/\nu)^{-(\nu+1)/2}$ and $c_\nu = \Gamma(\frac{\nu+1}{2})/\Gamma(\nu/2)\sqrt{\pi\nu}$. The score function and Fisher information are

$$S = \frac{\nu+3}{n}\Sigma_1^n \frac{x_i}{\nu+x_i^2}, \qquad I(f_\nu) = \frac{\nu+1}{\nu+3} .$$

Thus it is only necessary to estimate the variance of d_p when applied to samples after centering at S, and then to add this to $1 + 2/(\nu+1)$ to obtain an estimate of the Pitman variance.

The following section (especially Table 3) documents the dramatic increase in precision (in terms of sampling variability) of these variance estimates over those obtained by the standard G/I swindle.

TABLE 2

VARIANCES OF THE PITMAN ESTIMATES OF LOCATION FOR SMALL SAMPLES FROM STUDENT'S t POPULATION

Variance (standard error in units of last reported decimal place)

	10*	20†	40†
t_1	.2624 (49)	.1157 (21)	.0536 (4)
t_2	.1855 (14)	.0880 (5)	.0428 (1)
t_4	.1460 (4)	.0716 (0.4)	.0354 (0.3)
t_8	.1238 (1)	.0616 (0.4)	.0306 (0.07)
t_{16}	.1122 (0.2)	.0560 (0.4)	.0280 (0.03)

* 2000 replications
† 1000 replications

5. A NUMERICAL COMPARISON

The performance of variance decomposition swindles depends on the three measures in equation (2.1): (i) ρ^2, the squared correlation between the statistic T and the control S; (ii) $\kappa(T)$, the kurtosis of the sampling distribution of T on samples of size n from the underlying distribution of the data, F; and (iii) $\kappa(T-S)$, the kurtosis of the sampling distribution of $T-S$ (or equivalently, of T applied to the residuals after removing S). These values depend on T, on S, on the underlying distribution F, and on the sample size, n. We describe the results of a simulation comparison of the score function and Gaussian-over-independent swindles in selected location problems.

5.1. Correlation Term

We have previously noted that the relative size of Pitman, Cramer-Rao and optimal weighted least squares variances limits the maximum possible correlation between T and the control S. However, the behavior depicted in Figure 1 is itself dependent on sample size. Figure 2 shows the effect of sample size on the Pitman, Cramer-Rao, and optimal weighted least squares variances for Cauchy data. The score function swindle offers relatively little advantage in samples smaller than 10, but substantial advantages in samples larger than 20 where the computational effort needed for the naive variance estimates is of course much greater. (The advantages will generally also be greater for less extreme distributions.)

We emphasize that the variance decomposition swindles will generally perform better when applied to more efficient statistics. Thus for the same computing expense we will learn more about the better performing (and thus usually more interesting) statistics. This phenomenon was used to advantage in estimating the Pitman variances of the previous section.

5.2. Kurtosis Terms

The kurtosis ratio that forms the second factor in equation (2.1) makes it desirable that the kurtosis of $(T - S)$ not be substantially greater than the kurtosis of T. Often T will be asymptotically normal and even its small-sample sampling distribution will be very nearly normal. (This is true, for example, of many robust estimators of location or regression even at very heavy-tailed densities.) Unfortunately, the sampling distribution of $(T - S)$ can be very leptokurtic. In general, the kurtosis of $(T - S)$ tends to be higher when T and S are highly correlated (thus counteracting the advantage of the high correlation somewhat), when the underlying density is itself leptokurtic, and when the sample size is small. We have no good way to predict the kurtosis ratio, however we have estimated it in Monte Carlo experiments by accumulating ΣT^4 as well as $\Sigma(T - S)^4$.

The degradation of variance estimates for leptokurtic densities is well known. (See, for example, Yule and Kendall 1950, p. 443.) Briefly, the more extreme instances provide much of the information about the variance, thus reducing the effective sample size. In the swindle, when $\mathrm{var}(T - S)$ is very small, a few extraordinary samples with large $(T - S)$ can dominate the variance estimate.

5.3 Performance

Table 3 summarizes the performance of these variance decomposition swindles in a variety of situations. For location estimators the swindle gains are smallest for the most extreme population distributions (i.e., t on small degrees of freedom) and increase as the distributions approach the Gaussian. The larger swindle gains reflect high efficiencies of particular estimators at particular distributions. Figure 3 adds swindle gain information to Figure 1. The dependence of swindle gain on efficiency can be seen especially clearly at t_2.

Only rarely in these trials was the Gaussian-over-Independent swindle more effective than the score function swindle. Usually the latter was 10 to 50 times more effective.

The larger swindle gains deliver results with precision simply not obtainable by naive methods. A typical trial of 1000 replications required over 100 seconds of CPU time on an IBM 370/168. In the most extreme case (5% trimmed mean for samples of 40 from t_{16}) naive methods would have required over 15 CPU days of computing time for equivalent precision. The results for the Pitman variance in the same situation would have required 158 CPU days. Of course these last figures should not be taken too literally, as other sources of error (numerical, rounding) have not been assessed. What is clear is that <u>sampling</u> variability can be substantially reduced (or even effectively eliminated in efficient cases).

Although the G/I method does not apply to asymmetric densities, the score function swindle does. Trials on the rather extreme absolute Cauchy distribution yielded swindle gains of up to 20 in samples of 40 and up to 10 in samples of 20.

6. A SIMPLE FRAMEWORK

The simple geometrical setting given here provides a way to think about swindles for variances that can suggest new applications -- including some of these discussed in Section 7. The result (6.1) below is standard in estimation theory (Rao, 1973, Lehmann, 1983) but its role in Monte-Carlo studies is explicitly noted in unpublished lecture notes of Charles Stein.

Suppose that $T(\underset{\sim}{Y})$ belongs to an affine subset $\mathcal{J}$ (translate of a linear subspace) of the class of all estimators having finite variance under the distribution P_0 generating the data. Suppose also that $S(\underset{\sim}{Y})$ is the best (minimum variance) estimator belonging to $\mathcal{J}$. Then a version of Pythagoras' theorem gives the variance decomposition

$$\text{(6.1)} \qquad \mathrm{var}_{P_0} T = \mathrm{var}_{P_0} S + \mathrm{var}_{P_0}(T - S) \ .$$

One way to see this is to note that since $\mathcal{J}$ is affine, $S + \varepsilon(T - S)$ lies in $\mathcal{J}$ for each ε , so that $\mathrm{var}_{P_0}(S + \varepsilon(T - S))$ is minimized at $\varepsilon = 0$. Differentiation shows that S and $T - S$ are uncorrelated, which amounts to (6.1). We note in passing that expanding $\mathrm{var}_{P_0}(T\text{-}S)$ shows that $\rho^2(T,S) = \mathrm{Var}\ S/\mathrm{Var}\ T$, as remarked in Section 2.1.

The usefulness of (6.1) hinges on the choice of $\mathcal{J}$ since as $\mathcal{J}$ increases, $\mathrm{var}_{P_0} S$ decreases. Recall that we want $\mathrm{var}_{P_0} S$ to be both known (or easily calculated) and large. We illustrate this first by seeing how a number of swindles in the literature on location estimation fit into this framework.

Suppose that the data, $\underset{\sim}{Y}$, consists of n i.i.d. observations from a density $f(y-\theta)$ on R. Reasonable estimators, T, of θ are (at least) location equivariant, so to estimate the variance of T, with no loss of generality choose P_0 above to correspond to $\theta = 0$.

(i) (Stein) a) Let $\mathcal{J}$ be the class of all unbiased location equivariant estimators. Then S is the Pitman estimator

$$S(\underset{\sim}{y}) = \int \theta \, \Pi_1^n f(y_i - \theta) \, d\theta / \int \Pi_1^n f(y_i - \theta) \, d\theta \quad .$$

Typically, var S is not known analytically and must also be estimated (see Appendix A for discussion). An estimator suggested by Stein is

$$(6.2) \qquad \hat{\text{var}}\ T = \frac{1}{N} \sum_{J=1}^{N} E[(S^{(J)})^2 | \underset{\sim}{V}^{(J)}] + \frac{1}{N} \sum_1^N (T^{(J)} - S^{(J)})^2 \quad .$$

Here the superscript J refers to the J^{th} replication of the i.i.d. sample $(Y_1,\ldots,Y_n)$ from P_0, and $\underset{\sim}{V}^{(J)}$ to a maximal invariant $(Y_1^{(J)} - Y_n^{(J)},\ldots,Y_{n-1}^{(J)} - Y_n^{(J)})$. The right side of (6.2) is the conditional expectation of the naive estimate $1/N \, \Sigma_1^N (T^{(J)})^2$ given $\underset{\sim}{V}^{(1)},\ldots,\underset{\sim}{V}^{(N)}$, so $\hat{\text{var}}\ T$ is certainly a more precise (i.e., lower variance) estimate of var T than the naive one. Of course, a (univariate) numerical integration is needed to compute each $S^{(J)}$ and $E[(S^{(J)})^2 | \underset{\sim}{V}^{(J)}]$, but these can then be used repeatedly in estimating the variances of many equivariant estimators.

b) The same program is possible if $\mathcal{J}$ is restricted to the class of location <u>and</u> scale equivariant estimators, with the Pitman location-scale estimator serving as the "control function" S. Of course, bivariate numerical integrations are now necessary.

(ii) (Takeuchi, 1971) $\mathfrak{J}$ = unbiased linear combinations of order statistics $Y_{(i)}$ with weights c_i (not necessarily positive) summing to 1. Then $S(\underset{\sim}{Y}) = \Sigma\, c_i^* Y_{(i)}$, where $c^* = \Sigma_f^{-1}\underset{\sim}{1}/(\underset{\sim}{1}'\Sigma_f^{-1}\underset{\sim}{1})$ and $\Sigma_{f,ij} = \mathrm{cov}_f(Y_{(i)}, Y_{(j)})$. Thus, once Σ_f is known, both S and var S are readily computable.

(iii) <u>Gaussian over independent swindle</u>. (Andrews et. al., 1972; Hodges, 1967). A special assumption on f is needed, namely that the observations Y_i, as in Section 2, have the form $Y_i - \theta \overset{\mathcal{D}}{=} Z_i/W_i$ where $Z_i \sim N(0,1)$ and W_i is independent of Z_i. Now $\mathfrak{J}$ is the class of <u>unbiased</u> location-scale equivariant estimators. Here the variance decomposition is performed conditional on $(W_1,\ldots,W_n)$, so the problem of estimating θ becomes that of estimating the slope in the normal theory regression model $\tilde{Y}_i = W_i X_i = \theta W_i + Z_i$, where the W_i are known but the Z_i are not. The definition $\tilde{T}(\underset{\sim}{W},\underset{\sim}{\tilde{Y}}) = T(\tilde{Y}_1/W_1, \ldots, \tilde{Y}_n/W_n)$ associates a slope estimator $\tilde{T}$ in the regression model with $T \in \mathfrak{J}$; further, since T is location-scale equivariant, $\tilde{T}$ is regression equivariant (in the sense of Section 2) and unbiased for θ. To apply the decomposition (6.1), let $\tilde{S}(\underset{\sim}{W},\underset{\sim}{\tilde{Y}}) = \Sigma\, W_i\tilde{Y}_i/\Sigma W_i^2$ be the minimum variance unbiased estimator of θ. Thus conditional on $\underset{\sim}{W}$,

$$\mathrm{Var}\ T = \mathrm{Var}\ \tilde{T} = \frac{1}{\Sigma W_i^2} + \mathrm{Var}(\tilde{T} - \tilde{S}) .$$

Finally, since $E(T|\underset{\sim}{W}) = 0$ under $\theta = 0$, take expectations over $\underset{\sim}{W}$ to express the unconditional variance of T as

$$\mathrm{Var}\ T = E_0(1/\Sigma W_i^2) + E_0\ T^2(X_i - \tilde{S}) \quad .$$

(In fact, the second term on the right can be further decomposed slightly by exploiting the independence of the normal theory mean and variance estimators (cf. Simon, 1976), but the extra improvement in precision that results in small relative to that to that obtained here.)

(iv) The score function swindle. Define T to be locally unbiased for θ at θ_0 if $E_{\theta_0}\ T = \theta_0$ and $\partial/\partial\theta\ E_\theta T|_{\theta=\theta_0} = 1$. Choose $\mathfrak{J}$ as the affine space of estimators that are locally unbiased at $\theta_0 = 0$. As in Section 3, under appropriate regularity conditions, we have for statistics $T(\underset{\sim}{Y})$ of finite variance

$$\frac{\partial}{\partial\theta}\ E_\theta\ T|_{\theta=0} = E_\theta S_0 T \quad , \tag{6.3}$$

where S_0 is the score function for location. Normalizing S_0 to give $S = S_0/E_0 S_0^2$ ensures from (6.3) that S belongs to $\mathfrak{J}$, and further that it is uncorrelated with $T - S$ for $T \in \mathfrak{J}$. This yields the variance decomposition (6.1).

An analogous treatment is possible to the (successively smaller) affine spaces $\mathfrak{J}_k$ consisting of estimators locally unbiased up to order k: i.e. in addition to the properties above, we require that $\partial^j/\partial\theta^j\ E_{\theta_0}\ T = 0$ for $j = 2,\ldots,k$.

7. FURTHER APPLICATIONS

Stein effect and Bayesian robustness. Consider estimation of $\underset{\sim}{\theta} = (\theta_1,\ldots,\theta_p)$ using independent observations $X_i \sim N(\theta_i, \sigma_i^2)$ $i = 1,\ldots,p$, when loss in estimation of $\underset{\sim}{\theta}$ by $\underset{\sim}{\delta}(\underset{\sim}{x}) = (\delta_1(\underset{\sim}{x}),\ldots,\delta_p(\underset{\sim}{x}))$ is measured by $|\underset{\sim}{\delta}(\underset{\sim}{x}) - \underset{\sim}{\theta}|^2 = \Sigma_1^p(\delta_i(\underset{\sim}{x}) - \theta_i)^2$ and risk by $R(\underset{\sim}{\theta},\underset{\sim}{\delta}) = E_{\underset{\sim}{\theta}}|\underset{\sim}{\delta}(\underset{\sim}{X}) - \underset{\sim}{\theta}|^2$. It is often of interest to study the integrated risk $r(\pi,\underset{\sim}{\delta}) = \int R(\underset{\sim}{\theta},\underset{\sim}{\delta})\pi(d\underset{\sim}{\theta})$ of a rule $\underset{\sim}{\delta}$ and the Bayes risk $r(\pi) = \inf r(\pi,\underset{\sim}{\delta})$ relative to a prior measure $\pi(d\underset{\sim}{\theta})$. For example, Efron and Morris (1972) and Berger (1982) have used $r(\pi,\underset{\sim}{\delta})$ and $r(\pi,\underset{\sim}{\delta}) - r(\pi)$ in studying the "relative savings risk" of Stein-type estimator from empirical and robust Bayesian viewpoints respectively.

A fixed prior $\pi(d\underset{\sim}{\theta})$ determines, in conjuction with the sampling model, a marginal measure $\Pi(d\underset{\sim}{x})$ and, under the quadratic loss function L, an L^2 decomposition analogous to (5.1), namely

$$r(\pi,\underset{\sim}{\delta}) = r(\pi) + \int |\underset{\sim}{\delta}(\underset{\sim}{x}) - \underset{\sim}{\delta}_{\pi}(\underset{\sim}{x})|^2 \, \Pi(d\underset{\sim}{x}) \quad , \tag{7.1}$$

where $\underset{\sim}{\delta}_{\pi}(\underset{\sim}{x}) = E[\underset{\sim}{\theta}|\underset{\sim}{x}]$ is the Bayes rule minimizing $r(\pi,\underset{\sim}{\delta})$. The integral above is much easier to simulate (or evaluate numerically) than $r(\pi,\delta) = \int R(\underset{\sim}{\theta},\underset{\sim}{\delta})\, d\pi(\theta) = \int E[L(\underset{\sim}{\theta},\underset{\sim}{\delta})|\underset{\sim}{x}]\, \Pi(d\underset{\sim}{x})$. There is a further saving if one compares several $\underset{\sim}{\delta}$ for a fixed π (as done in Berger (1982)), since $r(\pi)$ need only be evaluated once. Although analytic expressions for $R(\underset{\sim}{\theta},\underset{\sim}{\delta})$ are available for many of rules $\underset{\sim}{\delta}$ of interest in the Gaussian case, this special feature disappears for other location densities, whereas the decomposition (6.1) persists.

Multivariate Location

Let $T(y_1,\ldots,y_n)$ be an be an unbiased, location equivariant estimate of the location parameters $\underset{\sim}{\theta} \in \mathbb{R}^d$ based on n i.i.d. observations x_i from a smooth density $f(\underset{\sim}{y} - \underset{\sim}{\theta})$ in $\mathbb{R}^d$. In principle the score function swindle extends directly, but we mention a couple of interesting features. The score function 'statistic' is now a vector with components $S_k^{(0)} = -\Sigma_{i=1}^{n} D_k f(y_i)/f(y_i)$ having mean $\underset{\sim}{0}$ and covariance matrix $\Sigma_{S^{(0)}}$. Defining $S = \Sigma_{S^{(0)}}^{-1} S^{(0)}$ leads to the matrix decomposition

$$\Sigma_T = \Sigma_S + \Sigma_{T-S} \quad .$$

where Σ_T is the covariance matrix of T. Note therefore that covariances of the components of T can be swindled in addition to the variances of the individual T_k. Such a swindle could be used to study efficiency properties of, for example, the computationally costly high-breakdown, affine-equivariant estimates of multivariate location proposed by Donoho (1982) and Stahel (1982).

Discrete Parameter Spaces

A discrete parameter version of the Cramer-Rao inequality (the Hammersley-Chapman-Robbins inequality) leads to a natural analog of the score function swindle. Suppose that Y has density $p(y,\theta) > 0$ for $y \in \mathcal{Y}$ and $\theta \in \Theta$. Fix $\theta_0 \in \Theta$, Δ such that $\theta_0 + \Delta \in \Theta$. Let $\mathcal{J} = \{T: E_{\theta_0+\Delta} T - E_{\theta_0} T = c\}$. Then the decomposition (5.1) holds with $S = c\psi/E\psi^2$ and $\psi = \{p(x, \theta_0 + \Delta)/p(x,\theta_0)\} - 1$. This version of the swindle can be of use, for example, in settings where the parameter θ is restricted to lie in a lattice such as the integers, as in the problem of estimation of molecular weight discussed by Hammersley (1950).

Bootstrap Estimates of Variance

To take a specific example, suppose that observations $x_1,\ldots,x_n$ are taken i.i.d. fromdistribution $F(x-\theta)$, and we wish to use a translation invariant estimate $T(x_1,\ldots,x_n)$ to estimate θ. In constrast with the simulation contexts considered earlier, it is not assumed here that F is known. Bootstrap estimates of $\mathrm{Var}\ {}_F T(x_1,\ldots,x_n)$ are obtained by replacing F by (some function of) its empirical distribution function F_n, say F_n^*, and estimating $\mathrm{Var}_{F_n^*}\ T(x_1,\ldots,x_n)$ by drawing N i.i.d. samples form F_n^* and then using the usual variance estimator.

Consider the following modification of the score function procedure to reduce the number N of "boots" required. Construct a density estimate f_n from F_n (say by using kernel methods) such that an empirical score function f_n'/f and estimated information $\int (f_n'/f_n)^2 f_n dx$ can be easily evaluated. Write F_n^* for the cdf corresponding to density f_n. Now draw i.i.d. samples from F_n^* and apply the location score function swindle, thus estimating only $\mathrm{Var}_{F^*}\ T(\underset{\sim}{X}-\tilde{\theta}(\underset{\sim}{X})\,\underset{\sim}{1})$. This proposal is speculative at present: work is in progress to evaluate the improvements obtained in particular situations.

8. DISCUSSION

8.A. Comparison of (1.1) and (1.2)

Suppose that the control statistic S satisfies (1.1) and let us compare the swindle gain from the variance decomposition (1.1) (given in (2.1)) with the swindle gain from the "regression estimate" (1.2) in the overly-optimistic situation that $b^* = -\mathrm{Cov}(\hat{\sigma}_S^2, \hat{\sigma}_T^2)/\mathrm{Var}(\hat{\sigma}_S^2)$ is known. In this case, assuming as in §2.1 that $ET = ES = 0$ and that $\hat{\sigma}_S^2$ and $\hat{\sigma}_T^2$ are given by naive method-of-moments estimators, we have

$$\frac{\mathrm{Var}(\hat{\sigma}_T^2)}{\mathrm{Var}(\breve{\sigma}_T^2)} = [1 - \rho^2(\hat{\sigma}_S^2, \hat{\sigma}_T^2)]^{-1} = [1 - \rho^2(S^2, T^2)]^{-1} \quad .$$

To render the calculations simple, suppose to a first order approximation that T and S are jointly normal with the same variance and with correlation η. Then $\rho(S^2, T^2) = \eta^2$, and it is easily checked from (2.1) that

$$\frac{\mathrm{Var}(\hat{\sigma}_T^2)}{\mathrm{Var}(\hat{\sigma}_{T-S}^2)} = \frac{1}{(1-\eta^2)^2} > \frac{1}{1-\eta^4} = \frac{\mathrm{Var}(\hat{\sigma}_T^2)}{\mathrm{Var}(\breve{\sigma}_T^2)} \quad ,$$

so that under these crude conditions the swindle gains from (1.1) are better than those from (1.2) by a factor of 5 when $\eta = .8$ and by a factor of 10 when $\eta = .9$.

Suppose now that we wish to estimate $\mathrm{Var}\ T_1 - \mathrm{Var}\ T_2$ for T_1, T_2 satisfying (1.1) for the same S. Again for simplicity, assume that

T_1, T_2 and S are jointly Gaussian with means 0, the same variances, and $\rho(T_1,S) = \rho(T_2,S) = w$, $\rho(T_1,T_2) = \rho$. It follows that the optimal $b^*_{T_1}$ and $b^*_{T_2}$ above will be equal, and hence that $\tilde{\sigma}^2_{T_1} - \tilde{\sigma}^2_{T_2} = \hat{\sigma}^2_{T_1} - \hat{\sigma}^2_{T_2}$, so that the regression swindle per se offers no improvement. However, since T_1 and T_2 are correlated

$$(8.1) \qquad N \operatorname{Var}(\hat{\sigma}^2_{T_1} - \hat{\sigma}^2_{T_2}) = \sigma^2_{T_1^2} + \sigma^2_{T_2^2} - 2\sigma_{T_1^2,T_2^2} = 4(1-\rho^2) \ ,$$

so that the precision of the difference is greater than that of each individual term. (here N is the number of Monte-Carlo trials). Does the variance decomposition swindle (1.1) help here? Easy normal theory calculations show that

$$(8.2) \qquad N \operatorname{Var}(\hat{\sigma}^2_{(T_1-S)} - \hat{\sigma}^2_{(T_2-S)}) = 4\{4(1-w)^2 - (1+\rho-2w)^2\} \ .$$

Denoting w/ρ by α, then (8.2) is smaller than (8.1) exactly when $\rho > 1/2\alpha$. Thus, for example, if all of T_1, T_2 and S are equally correlated, the variance decomposition dominates the simple difference when that correlation exceeds 1/2.

8.B. Swindles for Variance Ratios

Sometimes we are primarily interested in estimating an efficiency Var T/Var S, where S has minimum variance amongst all estimators in $\mathcal{J}$. If Var S is known, then the improvement achieved by a variance decomposition swindle can be measured simply by comparing Var $\widehat{\operatorname{Var}}$(T-S) with Var $\hat{\text{Var}}$ T as above. If Var S must also be estimated, then we can give a crude indication of the improvement attained as follows. We continue to assume ES = ET = 0 and to use moment estimators for Var S, Var T and Var(T-S). Thus, using the generic labels s and t for replications $S(X^I)$ and $T(X^I)$ for $I = 1,\ldots,N$, we seek to compare the variances of

$$(8.3) \qquad \hat{e}_1 = \frac{\Sigma t^2}{\Sigma s^2}, \quad \hat{e}_2 = \frac{\Sigma s^2 + \Sigma(t-s)^2}{\Sigma s^2},$$

(the naive and swindled estimates respectively). Making the rather crude assumption that the values of t-s (small compared to s) are independent of s and symmetrically distributed about zero, one finds that

$$(8.4) \qquad E(t^2 | (t-s)^2 + s^2) = s^2 + (t-s)^2,$$

so that certainly $\text{Var}(\Sigma t^2) \geq \text{Var}(\Sigma s^2 + \Sigma(t-s)^2)$. A more explicit expression for the difference in variances of $\hat{e}_1$ and $\hat{e}_2$ follows by expanding t^2 as $s^2 + 2s(t-s) + (t-s)^2$ and conditioning on all $s_I = S(X^I)$ values. From the independence and distributional assumptions on t-s, we have $E(\hat{e}_1 | s_1,\ldots,s_N) = E(\hat{e}_2 | s_1,\ldots,s_N)$, so

$$\begin{aligned}\text{Var}(\hat{e}_1) - \text{Var}(\hat{e}_2) &= E\,\text{Var}\left[\frac{2\Sigma s(t-s) + \Sigma(t-s)^2}{\Sigma s^2} \middle| \underset{\sim}{s}\right] - E\,\text{Var}\left[\frac{\Sigma(t-s)^2}{\Sigma s^2} \middle| \underset{\sim}{s}\right] \\ &= E\,\frac{1}{(\Sigma s^2)^2}\,\text{Var}[2\Sigma s(t-s) | \underset{\sim}{s}] = 4\,\text{Var}(t-s)\,E\,\frac{1}{(\Sigma s^2)} \geq 0.\end{aligned}$$

A further point to note is that $\hat{e}_2 \geq 1$, whereas it is certainly possible for $\hat{e}_1$ to be less than 1, in contradiction with the optimality of S.

8.C. Bhattacharya Bounds

The Cramer-Rao bound is the first of a sequence of lower bounds to the variance of an estimator that can be obtained by using successively higher derivatives of the likelihood to build control functions. In estimation of a single parameter θ, these Bhattacharya bounds take the form (e.g. Lehmann, 1983, p. 129)

$$\text{var}_\theta\, \delta \geq \underset{\sim}{\alpha}'\, \kappa^{-1}(\theta)\, \underset{\sim}{\alpha} = B_p \,, \tag{8.5}$$

where $\psi^{(j)} = [f_\theta(y)]^{-1}\, \partial^j/\partial\theta^j\, f_\theta(\underset{\sim}{y})$, $\underset{\sim}{\alpha}'$ is a row matrix with entries

$$\frac{\partial^j}{\partial\theta^j} E_\theta\, \delta(\underset{\sim}{Y}) = \text{cov}(\delta, \psi^{(j)}) \qquad j = 1, \ldots, p \,,$$

and $\kappa_{ij}(\theta) = \text{Cov}_\theta(\psi^{(i)}, \psi^{(j)})$. If $\delta(\underset{\sim}{Y})$ is unbiased for θ (at least up to a constant), then the lower bound becomes $[\kappa^{-1}(\theta)]_{11}$, which by standard matrix theory is an increasing function of p. When the Bhattacharya bounds are strictly closer to the Pitman bound, they lead in principle to more effective swindles. In practice, the cases discussed below suggest that for $p \geq 3$ or for moderate to large n (when the C-R bound becomes quite good anyway), the improvement is not very significant.

Consider the location problem, initially with $p = 2$, and $f_\theta(\underset{\sim}{y}) = \Pi_{i=1}^{n} f(y_i - \theta)$. If δ is unbiased, then we easily calculate the percentage improvement in the lower bound from

$$B_2/B_1 = \frac{\kappa_{11}\,\kappa_{22}}{\kappa_{11}\kappa_{22} - \kappa_{12}^2} = [1 - \rho^2(\psi^{(1)}, \psi^{(2)})]^{-1} .$$

In fact B_2/B_1 is independent of θ in the location problem, and if $\theta = 0$, and $\phi(y_i) = - f'/f(y_i)$, then

$$\psi^{(1)}(\underset{\sim}{y}) = \Sigma_1^n \phi(y_i) \qquad \psi^{(2)}(\underset{\sim}{y}) = - \Sigma\phi'(y_i) + [\Sigma\phi(y_i)]^2 .$$

Now if f is symmetric about 0, then f, ϕ^2 and ϕ' are even functions while ϕ is odd, so that $\mathrm{Cov}(\psi^{(1)}, \psi^{(2)}) = 0$. Thus the second-order Bhattacharya bound offers no improvement.

Even when f is asymmetric, the gain decreases inversely with n. Indeed

$$(8.6) \qquad \rho^2(\psi^{(1)}, \psi^{(2)}) = \frac{(E\,\phi\,\eta)^2}{E\phi^2[E\eta^2 + 2(n-1)(E\phi^2)^2]} ,$$

where $\eta = \eta(y_1) = (\phi^2 - \phi')(y_1)$. For a specific example, let $f(y) = c_\sigma\, n(y)$ for $y > 0$, and $c_\sigma\, n(y/\sigma)$ for $y < 0$, where $n(y) = (2\pi)^{-\frac{1}{2}} e^{-y^2/2}$ and $c_\sigma = (1+\sigma)/2$. Then

$$\rho^2(\psi^{(1)}, \psi^{(2)}) = \frac{1}{\pi[1 + n\sigma/(\sigma-1)^2]} .$$

In the symmetric location case one is forced to look at third order bounds, and it is easily shown that

$$B_3/B_1 = [1-\rho^2(\psi^{(1)}, \psi^{(3)})]^{-1} , \qquad \text{and}$$

$$\rho^2(\psi^{(1)},\psi^{(3)}) = \frac{(E\ \phi\zeta)^2}{E\phi^2[\zeta^2+9(n-1)E\phi^2E\eta^2+6(n-1)(n-2)(E\phi^2)^3]} ,$$

where $\zeta = \phi'' - 3\phi\phi' + \phi^3$. If f is Cauchy, then calculation shows that

$$\rho^2(\psi^{(1)}, \psi^{(3)}) = \frac{2}{n^2+3n+5} .$$

Thus the improvement will typically be quite small: in this case, for $n = 5$, $B_3/B_1 = 1.0465$ for example.

REFERENCES

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J.,. Rogers, W. H., and Tukey, J. W. (1972) Robust Estimators of Location: Survey and Advances, Princeton: Princeton Univ. Press.

Andrews, D. F. and Mallows, C. P. (1974) "Scale Mixtures of Normal Distributions," J. Roy. Stat. Soc. A, 99-102.

Berger, J. O. (1982) Bayesian robustness and the Stein effect. J. Amer. Statist. Assoc. 77, 358-368.

Bickel, P. J. (1982) "On Adaptive Estimation," Ann. Stat., 647-671.

Cochran, W. G. (1977) Sampling Techniques, 3rd. Edition. New York: John Wiley and Sons.

Cramer, H. (1946) Mathematical Methods of Statistics. Princeton Univ. Press, Princeton, New Jersey.

Dixon, W. J. and Tukey, J. W. (1968) "Approximate Behavior of the Distribution of Winsorized t (Trimming/Winsorization 2) Technometrics 10: 1 83-98.

Donoho, D. L. (1982) Breakdown properties of multivariate location estimators. Ph.D. Qualifying Paper, Dept. of Statistics, Harvard Univ.

Efron, B. and Morris, C. (1971/2) Limiting the risk of Bayes and empirical Bayes estimators. J. Amer. Statist. Assoc. 66, 807-815, 67 130-139.

Efron, B. and Olshen, R. A. (1978) How broad is the class of normal scale mixtures? Ann. Statist. 6, 1159-64.

Goodfellow, D. M. and Martin, R. D. (1976) "Monte Carlo Swindle Techniques in Robust Estimation" EE Technical Report, Department of Electrical Engineering, Univ. of Washington.

Hammersley, J. M. (1950) On estimating restricted parameters.. J. Roy. Statist. Soc., 12, 192-240.

Hammersley, J. M. and Handscomb, D. C. (1964) Monte Carlo Methods. London: Methuen, New York: Wiley.

Hoaglin, D. C. (1975) "The Small-Sample Variance of the Pitman Location Estimators," JASA, 70, 880-888.

Hoaglin, D. C., Mosteller, F. and Tukey, J. W. (1983) Understanding Robust and Exploratory Data Analysis. New York: Wiley.

Hodges, J. L. (1967) Efficiency in normal samples and tolerance of extreme values for some estimates of location. Proc. 5th Berkeley Symp. Math. Stat. and Probab. 1, 163-186.

Johnstone, I. M. and Velleman, P. F. (1984) The Resistant Line and Related Regression Methods. To appear in JASA.

Lehmann, E. (1983) The Theory of Point Estimation. New York: John Wiley & Sons.

Relles, D. A. (1970) "Variance Reduction Techniques for Monte Carlo Sampling from Student Distributions," Technometrics, 12: 499-

Pitman, E. J. G. (1939) "The estimation of the location and scale Parameters of a continuous population of any given form" Biometrika, 30, 391-42.

Rao, C. R. (1973) Linear Statistical Inference and Its Applications Second edition. New York, John Wiley & Sons.

Rubinstein , Reuven Y. (1981) Simulation and the Monte Carlo Method, New York, John Wiley & Sons.

Simon, G. (1976) "Computer Simulation Swindels with Applications to Estimates of Location and Dispersion," Appl. Statist. 25: 266-274.

Stahel, W. A. (1981) Robuste Schätzungen: Infinitesimale Optimalität und Schatzungen von Kovarianzmatrizen. Ph.D. Thesis, Swiss Federal Institute of Technology, Zürich.

Stein, C. (1956) "Efficient Nonparametric Testing and Estimation." Proc. 3rd. Berk. Symp. Math. Statist. and Prob., Vol. 1, 187-196 Univ. California Press.

Takeuchi, K. (1971) A uniformly asymptotically efficient estimator of a location parameter. J. Amer. Statist. Assoc., 66, 292-301.

Yule, G. U. and Kendall, M. G. (1950) An Introduction to the Theory of Statistics. New York: Hafner.

TABLE 1

GAUSSIAN/INDEPENDENT (Z/W) DISTRIBUTIONS
(from Simon (1976))

Distribution	W Drawn Form
$N(0, \sigma^2)$	$W = 1/\sigma^2$
t_ν	$W \sim \sqrt{\chi^2_{(\nu)}/\nu}$
(Cauchy = t_1)	$W = \lvert\phi(W)\rvert$
"Contaminated normal"	$W = \begin{cases} 1 & \text{with prob} = p \\ 1/\sigma^2 & \text{with prob} = 1 - p \end{cases}$
"Slash"	$W \sim U[0,1]$
Laplace (Double Exponential)	$f(w) = w^{-3} \exp(-w^{-2}/2)$

TABLE 3: SWINDLE GAIN FACTORS FOR TWO VARIANCE DECOMPOSITION SWINDLES

Population Distribution: t on Indicated Degrees of Freedom

	n	1*		2		4		8		16		ABSOLUTE CAUCHY
		SCOR	Z/W	SCOR	Z/W	SCOR	Z/W	SCOR	Z/W	SCOR	Z/W	SCOR
BIWEIGHT	10	1.8	1.9	10.6	3.4	48.9	7.9	35.3	11.6	20.8	13.0	—
	20	6.4	1.8	47.5	3.3	105.7	9.1	74.5	20.1	35.8	16.8	10.3
	40	12.6	2.9	113.0	6.1	308.9	7.6	136.9	16.9	58.4	22.1	15.9
HUBER	10	1.2	1.4	2.4	2.3	28.9	6.9	255.8	17.5	747.9	42.7	—
	20	1.6	1.3	11.4	2.0	59.3	8.7	867.3	34.9	2081.3	49.8	3.6
	40	2.6	1.6	11.2	3.0	126.9	6.2	3129.7	25.8	2032.6	52.8	6.2
MEDIAN	10	2.6	1.7	9.2	2.9	32.8	5.9	20.8	7.5	15.0	9.4	—
	20	8.0	1.8	16.2	2.6	26.6	6.0	20.2	10.1	13.5	8.1	7.8
	40	11.5	2.4	27.4	4.5	21.9	4.8	15.7	6.3	11.6	6.9	16.5
20% TRIMMED MEAN	10	1.0	5.0	1.0	1.0	2.6	2.2	8.8	4.8	10.6	7.7	—
	20	1.2	1.6	47.4	3.1	176.2	8.8	185.2	22.8	64.9	20.0	2.2
	40	4.7	2.2	79.0	5.4	907.1	8.3	445.2	18.1	147.8	27.8	19.7
10% TRIMMED MEAN	10	1.0	0.4	2.6	2.2	47.2	7.7	498.6	19.7	1390.6	38.6	—
	20	1.2	1.0	9.4	2.0	96.5	9.2	1966.1	39.3	1838.2	50.7	2.8
	40	1.9	1.5	10.5	3.0	265.5	7.3	7066.0	24.7	2183.0	47.0	5.7
5% TRIMMED MEAN	10	1.0	5.0	1.0	1.0	4.6	3.2	26.9	10.7	247.6	35.2	—
	20	1.0	1.0	3.6	1.5	29.7	7.1	845.6	36.0	7744.0	61.0	1.3
	40	1.3	1.2	3.2	1.8	5.5	5.5	1368.7	23.7	13654.0	57.9	2.0
$\bar{x}$	10	1.0	5.0	1.0	1.0	4.9	3.2	32,9	10.6	407.6	36.0	—
	20	1.0	1.6	1.0	1.4	4.3	2.9	93.6	21.4	645.9	47.9	1.0
	40	1.0	0.8	1.1	0.5	6.9	3.0	78.3	15.1	993.3	48.0	1.0
PITMAN	10	5.1	2.1	22.7	4.2	148.4	9.2	1256.7	19.7	18,884.6	45.6	
	20	9.6	2.5	64.7	3.4	300.8	10.3	4942.6	40.7	39,607.6	60.6	
	40	51.6	3.8	328.4	6.7	2298.0	9.4	35594.6	26.4	136,544.6	58.0	

* t_1 = Cauchy

SCOR = Score Function Swindle

Z/W = Gaussian-over-Independent Swindle

n = 10	2000 replications
n = 20,40	1000 replications

Figure 1

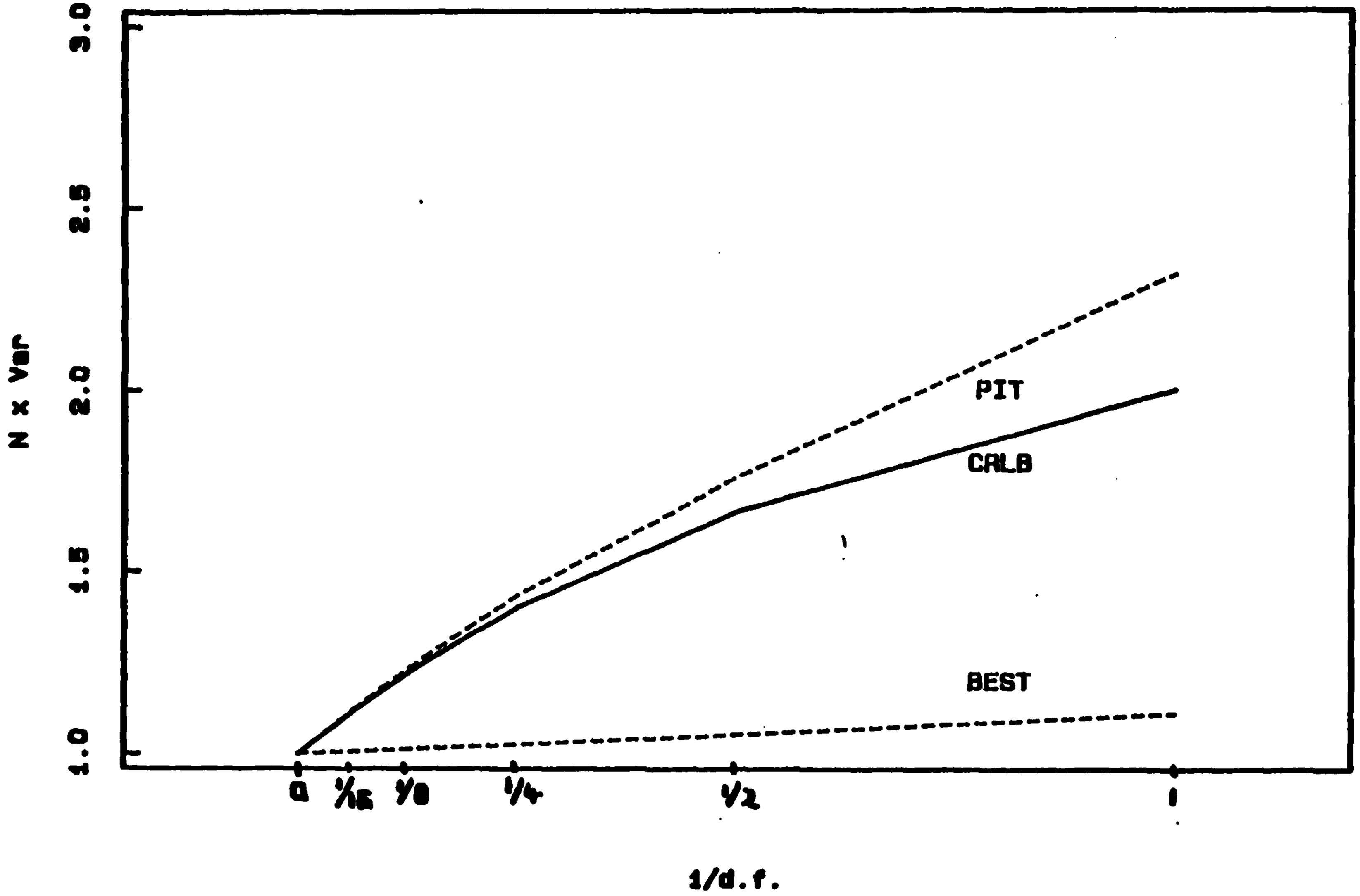


Figure 1. Nx Var for Pitman and Best Weighted Least Squares Estimates compared to Cramer Rao Bound (CRLB) for Student's *t* Distributions on $\nu = 1,2,4,8,16,\infty$ d.f. and N = 20.

Figure 2

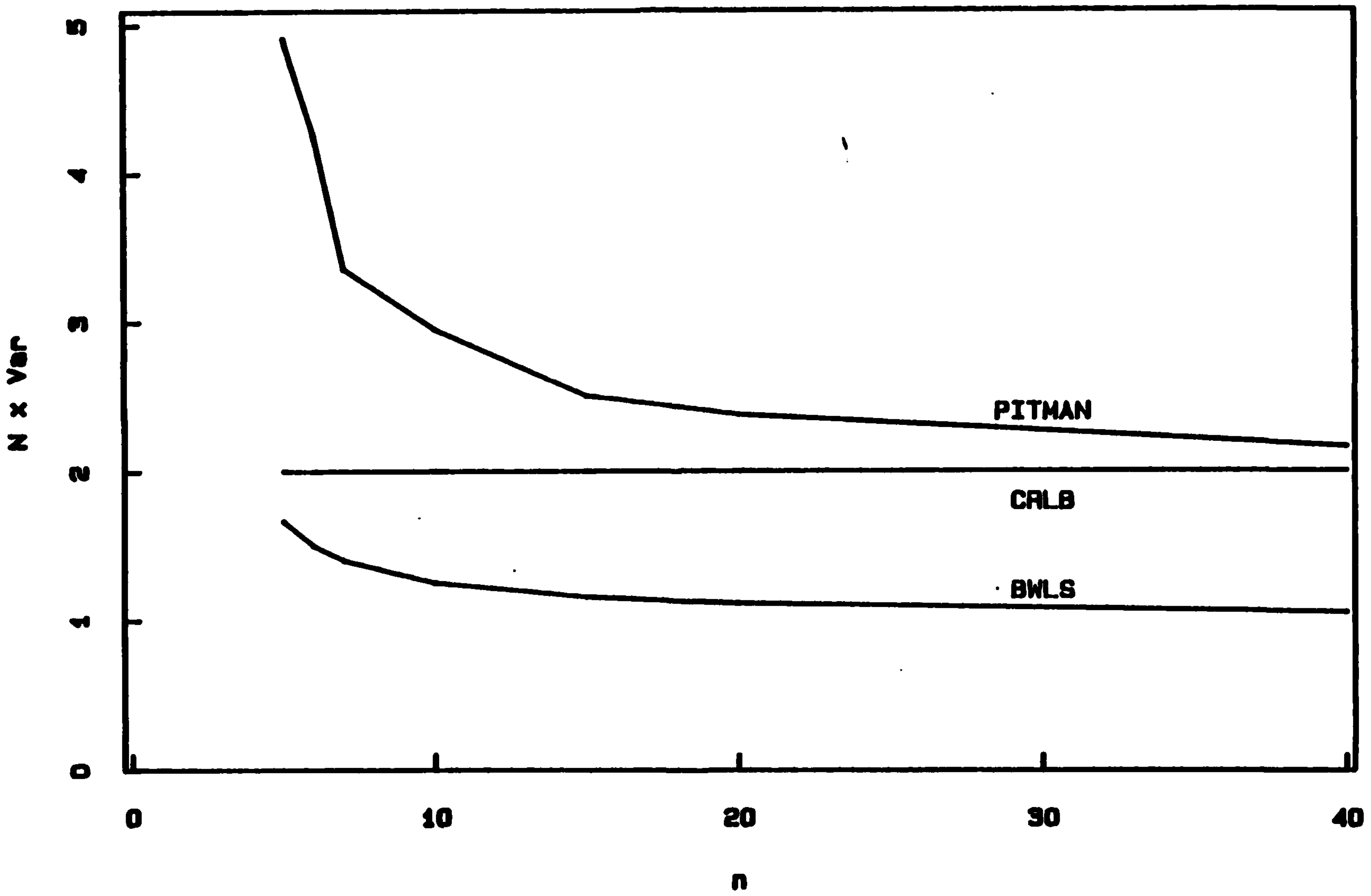


Figure 2. Nx Var for Pitman and Best Weighted L.S. Estimates Compared to Cramer Rao Lower Bound (CRLB) for Cauchy Distribution and n = 5,6,7,10,15,20,40.

Figure 3

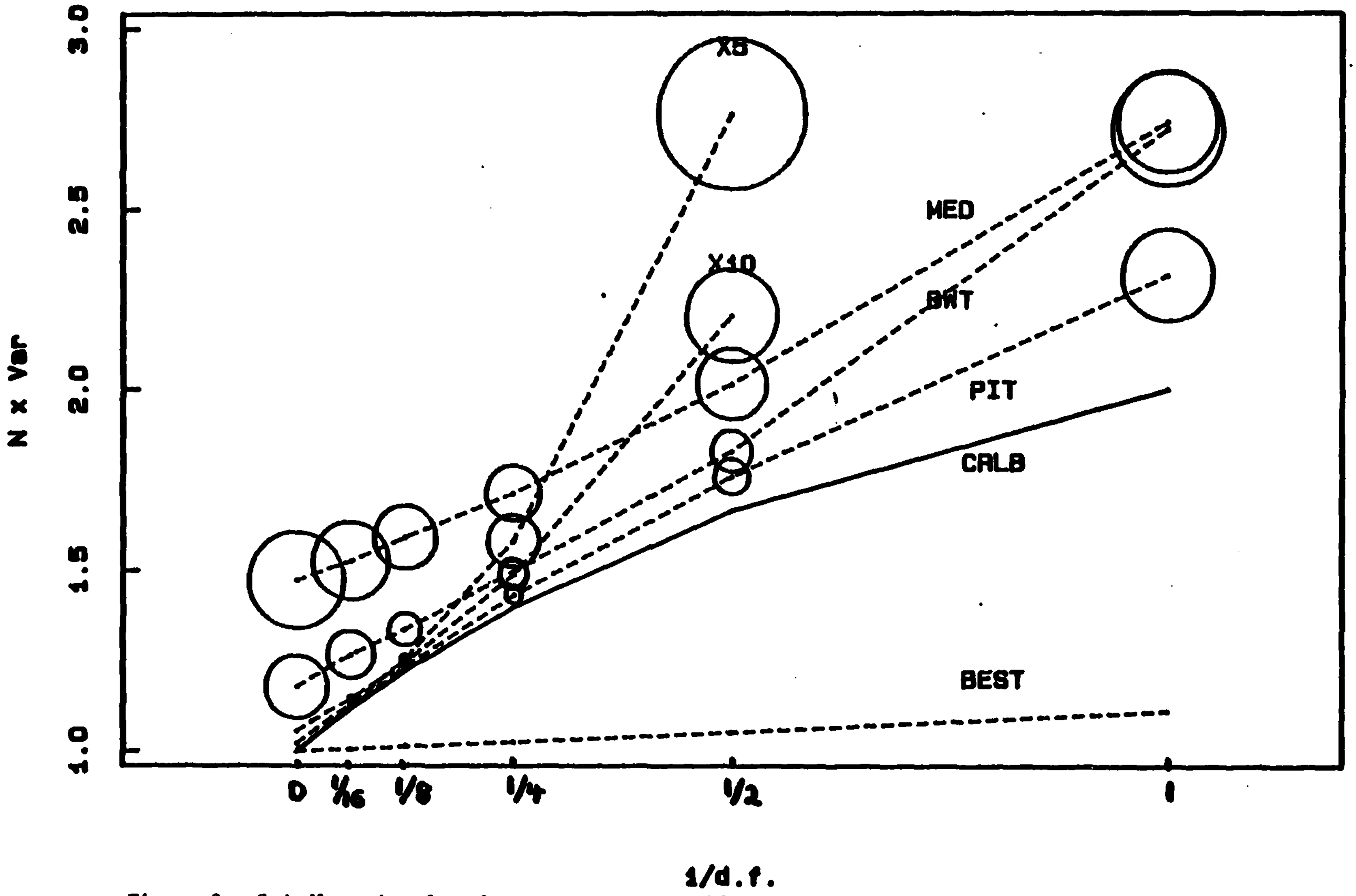


Figure 3. Swindle gains for the Estimators in Table 3 are superimposed on Table 1. Area of circles are inversely proportional to swindle gains (zero gain would give a circle with about twice the radius of the largest circle for x5). Note that swindle gains increase with efficiency of the

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 348	2. GOVT ACCESSION NO. AD-A146673	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Efficient Scores, Variance Decompositions And Monte Carlo Swindles		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Iain Johnstone and Paul Velleman		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0475
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 411SP		12. REPORT DATE August 28, 1984
		13. NUMBER OF PAGES 45
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Also issued as Technical Report No. 221 under National Science Foundation Grant MCS80-24649, Department of Statistics, Stanford University.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Efficient scores, Variance decompositions, Monte Carlo swindles.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

PLEASE SEE REVERSE SIDE

REPORT NO. 348

ABSTRACT

Monte Carlo "swindles" or variance reduction techniques exploit the experimenter's knowledge of the stochastic structure governing the simulated data to construct more precise estimates of unknown parameters. Alternatively, one can reduce the number of replications (and thus the cost) needed to gain a desired level of precision. This paper reviews the common case of swindles based on variance decompositions for estimating efficiencies and variances of location and regression estimators. ~~We~~ It then proposes a new swindle based on Fisher's efficient score function that can be applied to a much wider range of situations than can the Gaussian-over-independent swindles used in many studies of robust estimators. ~~We~~ The authors compare these methods by performing simulations for the efficiencies of location estimates and by placing them in a simple geometric framework. ~~We~~ They illustrate the use of the score function swindle in estimating the variances of Pitman estimates of location for samples from the t-distribution at selected degrees of freedom. Finally, we sketch applications to scale estimation, exponential regression, statistical decision theory, and bootstrap computations. are sketched.

END

FILMED

11-84

DTIC