Data Analysis Using Stein's Estimator and its Generalizations

Bradley Efron; Carl Morris

Journal of the American Statistical Association, Vol. 70, No. 350. (Jun., 1975), pp. 311-319.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197506%2970%3A350%3C311%3ADAUSEA%3E2.0.CO%3B2-1

Journal of the American Statistical Association is currently published by American Statistical Association.

Data Analysis Using Stein's Estimator and Its Generalizations

BRADLEY EFRON and CARL MORRIS*

In 1961, James and Stein exhibited an estimator of the mean of a multivariate normal distribution having uniformly lower mean squared error than the sample mean. This estimator is reviewed briefly in an empirical Bayes context. Stein's rule and its generalizations are then applied to predict baseball averages, to estimate toxomosis prevalence rates, and to estimate the exact size of Pearson's chi-square test with results from a computer simulation. In each of these examples, the mean square error of these rules is less than half that of the sample mean.

1. INTRODUCTION

Charles Stein [15] showed that it is possible to make a uniform improvement on the maximum likelihood estimator (MLE) in terms of total squared error risk when estimating several parameters from independent normal observations. Later James and Stein [13] presented a particularly simple estimator for which the improvement was quite substantial near the origin, if there are more than two parameters. This achievement leads immediately to a uniform, nontrivial improvement over the least squares (Gauss-Markov) estimators for the parameters in the usual formulation of the linear model. One might expect a rush of applications of this powerful new statistical weapon, but such has not been the case. Resistance has formed along several lines:

1. Mistrust of the statistical interpretation of the mathematical formulation leading to Stein's result, in particular the sum of squared errors loss function;
2. Difficulties in adapting the James-Stein estimator to the many special cases that invariably arise in practice;
3. Long familiarity with the generally good performance of the MLE in applied problems;
4. A feeling that any gains possible from a "complicated" procedure like Stein's could not be worth the extra trouble. (J.W. Tukey at the 1972 American Statistical Association meetings in Montreal stated that savings would not be more than ten percent in practical situations.)

We have written a series of articles [5, 6, 7, 8, 9, 10, 11] that cover Points 1 and 2. Our purpose here, and in a lengthier version of this report [12], is to illustrate the methods suggested in these articles on three applied problems and in that way deals with Points 3 and 4. Only one of the three problems, the toxoplasmosis data, is "real" in the sense of being generated outside the statistical world. The other two problems are contrived to illustrate in a realistic way the genuine difficulties and rewards of procedures like Stein's. They have the added advantage of having the true parameter values available for comparison of methods. The examples chosen are the first and only ones considered for this report, and the favorable results typify our previous experience.

To review the James-Stein estimator in the simplest setting, suppose that for given θ_i

$$X_i|\theta_i \overset{\text{ind}}{\sim} N(\theta_i, 1), \quad i = 1, \cdots, k \geq 3 , \tag{1.1}$$

meaning the $\{X_i\}$ are independent and normally distributed with mean $E_{\theta_i}X_i \equiv \theta_i$ and variance $\text{Var}_{\theta_i}(X_i) = 1$. The example (1.1) typically occurs as a reduction to this canonical form from more complicated situations, as when X_i is a sample mean with known variance that is taken to be unity through an appropriate scale transformation. The unknown vector of means $\boldsymbol{\theta} \equiv (\theta_1, \cdots, \theta_k)$ is to be estimated with loss being the sum of squared component errors

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \equiv \sum_{i=1}^{k} (\hat{\theta}_i - \theta_i)^2 , \tag{1.2}$$

where $\hat{\boldsymbol{\theta}} \equiv (\hat{\theta}_1, \cdots, \hat{\theta}_k)$ is the estimate of $\boldsymbol{\theta}$. The MLE, which is also the sample mean, $\boldsymbol{\delta}^0(\mathbf{X}) \equiv \mathbf{X} \equiv (X_1, \cdots, X_k)$ has constant risk k,

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}^0) \equiv E_{\boldsymbol{\theta}} \sum_{i=1}^{k} (X_i - \theta_i)^2 = k , \tag{1.3}$$

$E_{\boldsymbol{\theta}}$ indicating expectation over the distribution (1.1). James and Stein [13] introduced the estimator $\boldsymbol{\delta}^1(\mathbf{X}) = (\delta_1{}^1(\mathbf{X}), \cdots, \delta_k{}^1(\mathbf{X}))$ for $k \geq 3$,

$$\delta_i{}^1(\mathbf{X}) \equiv \mu_i + (1 - (k-2)/S)(X_i - \mu_i) , \quad i = 1, \cdots, k \tag{1.4}$$

with $\boldsymbol{\mu} \equiv (\mu_1, \cdots, \mu_k)'$ any initial guess at $\boldsymbol{\theta}$ and $S \equiv \sum (X_j - \mu_j)^2$. This estimator has risk

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}^1) \equiv E_{\boldsymbol{\theta}} \sum_{i=0}^{k} (\delta_i{}^1(\mathbf{X}) - \theta_i)^2 \tag{1.5}$$

$$\leq k - \frac{(k-2)^2}{k - 2 + \sum (\theta_i - \mu_i)^2} < k , \tag{1.6}$$

being less than k for all $\boldsymbol{\theta}$, and if $\theta_i = \mu_i$ for all i the risk is two, comparing very favorably to k for the MLE.

* Bradley Efron is professor, Department of Statistics, Stanford University, Stanford, Calif. 94305. Carl Morris is statistician, Department of Economics, The RAND Corporation, Santa Monica, Calif. 90406.

© Journal of the American Statistical Association
June 1975, Volume 70, Number 350
Applications Section

The estimator (1.4) arises quite naturally in an empirical Bayes context. If the $\{\theta_i\}$ themselves are a sample from a prior distribution,

$$\theta_i \overset{\text{ind}}{\sim} N(\mu_i, \tau^2), \quad i = 1, \cdots, k \ , \tag{1.7}$$

then the Bayes estimate of θ_i is the *a posteriori* mean of θ_i given the data

$$\delta_i^*(X_i) = E\theta_i | X_i = \mu_i + (1 - (1 + \tau^2)^{-1})(X_i - \mu_i) \ . \tag{1.8}$$

In the empirical Bayes situation, τ^2 is unknown, but it can be estimated because marginally the $\{X_i\}$ are independently normal with means $\{\mu_i\}$ and

$$S = \sum (X_j - \mu_j)^2 \sim (1 + \tau^2)\chi_k^2 \ , \tag{1.9}$$

where χ_k^2 is the chi-square distribution with k degrees of freedom. Since $k \geq 3$, the unbiased estimate

$$E(k - 2)/S = 1/(1 + \tau^2) \tag{1.10}$$

is available, and substitution of $(k - 2)/S$ for the unknown $1/(1 + \tau^2)$ in the Bayes estimate δ_i^* of (1.8) results in the James-Stein rule (1.4). The risk of δ_i^1 averaged over both $\mathbf{X}$ and $\boldsymbol{\theta}$ is, from [6] or [8],

$$E_\tau E_\theta(\delta_i^1(\mathbf{X}) - \theta_i)^2 = 1 - (k - 2)/k(1 + \tau^2) \ , \tag{1.11}$$

E_τ denoting expectation over the distribution (1.7). The risk (1.11) is to be compared to the corresponding risks of 1 for the MLE and $1 - 1/(1 + \tau^2)$ for the Bayes estimator. Thus, if k is moderate or large δ_i^1 is nearly as good as the Bayes estimator, but it avoids the possible gross errors of the Bayes estimator if τ^2 is misspecified.

It is clearly preferable to use min $\{1, (k - 2)/S\}$ as an estimate of $1/(1 + \tau^2)$ instead of (1.10). This results in the simple improvement

$$\delta_i^{1+}(\mathbf{X}) = \mu_i + (1 - (k - 2)/S)^+(X_i - \mu_i) \tag{1.12}$$

with $a^+ \equiv \max(0, a)$. That $R(\boldsymbol{\theta}, \boldsymbol{\delta}^{1+}) < R(\boldsymbol{\theta}, \boldsymbol{\delta}^1)$ for all $\boldsymbol{\theta}$ is proved in [2, 8, 10, 17]. The risks $R(\boldsymbol{\theta}, \boldsymbol{\delta}^1)$ and $R(\theta, \delta^{1+})$ are tabled in [11].

2. USING STEIN'S ESTIMATOR TO PREDICT BATTING AVERAGES

The batting averages of 18 major league players through their first 45 official at bats of the 1970 season appear in Table 1. The problem is to predict each player's batting average over the remainder of the season using only the data of Column (1) of Table 1. This sample was chosen because we wanted between 30 and 50 at bats to assure a satisfactory approximation of the binomial by the normal distribution while leaving the bulk of at bats to be estimated. We also wanted to include an unusually good hitter (Clemente) to test the method with at least one extreme parameter, a situation expected to be less favorable to Stein's estimator. Batting averages are published weekly in the *New York Times*, and by April 26, 1970 Clemente had batted 45 times. Stein's estimator requires equal variances,[1] or in this situation, equal at bats, so the remaining 17 players are all whom either the April 26 or May 3 *New York Times* reported with 45 at bats.

Let Y_i be the batting average of Player i, $i = 1, \cdots, 18$ ($k = 18$) after $n = 45$ at bats. Assuming base hits occur according to a binomial distribution with independence between players, $nY_i \overset{\text{ind}}{\sim}$ Bin (n, p_i) $i = 1, 2, \cdots, 18$ with p_i the true season batting average, so $EY_i = p_i$. Because the variance of Y_i depends on the mean, the arc-sin transformation for stabilizing the variance of a binomial distribution is used: $X_i \equiv f_{45}(Y_i)$, $i = 1, \cdots, 18$ with

$$f_n(y) \equiv (n)^{\frac{1}{2}} \arcsin(2y - 1) \ . \tag{2.1}$$

Then X_i has nearly unit variance[2] independent of p_i. The mean[3] θ_i of X_i is given approximately by $\theta_i = f_n(p_i)$. Values of X_i, θ_i appear in Table 1. From the central limit theorem for the binomial distribution and continuity of f_n we have approximately

$$X_i | \theta_i \overset{\text{ind}}{\sim} N(\theta_i, 1), \quad i = 1, 2, \cdots, k \ , \tag{2.2}$$

the situation described in Section 1.

We use Stein's estimator (1.4), but we estimate the common unknown value $\mu = \sum \mu_i/k$ by $\bar{X} = \sum X_i/k$, shrinking all X_i toward $\bar{X}$, an idea suggested by Lindley [6, p. 285–7]. The resulting estimate of the ith component θ_i of $\boldsymbol{\theta}$ is therefore

$$\tilde{\delta}_i^1(\mathbf{X}) = \bar{X} + (1 - (k - 3)/V)(X_i - \bar{X}) \tag{2.3}$$

with $V \equiv \sum (X_i - \bar{X})^2$ and with $k - 3 = (k - 1) - 2$ as the appropriate constant since one parameter is estimated. In the empirical Bayes case, the appropriateness of (2.3) follows from estimating the Bayes rule (1.8) by using the unbiased estimates $\bar{X}$ for μ and $(k - 3)/V$ for $1/(1 + \tau)^2$ from the marginal distribution of $\mathbf{X}$, analogous to Section 1 (see also [6, Sec. 7]). We may use the Bayesian model for these data because (1.7) seems at least roughly appropriate, although (2.3) also can be justified by the non-Bayesian from the suspicion that $\sum (\theta_i - \bar{\theta})^2$ is small, since the risk of (2.3), analogous to (1.6), is bounded by

$$R(\boldsymbol{\theta}, \tilde{\boldsymbol{\delta}}^1) \leq k - \frac{(k - 3)^2}{k - 3 + \sum (\theta_i - \bar{\theta})^2}, \quad \bar{\theta} \equiv \sum \theta_i/k \ . \tag{2.4}$$

For our data, the estimate of $1/(1 + \tau^2)$ is $(k - 3)/V = .791$ or $\hat{\tau} = 0.514$, representing considerable *a priori* information. The value of $\bar{X}$ is -3.275 so

$$\tilde{\delta}_i^1(\mathbf{X}) = \hat{\theta}_i = .791\bar{X} + .209X_i = .209X_i - 2.59 \ . \tag{2.5}$$

---

[1] The unequal variances case is discussed in Section 3.

[2] An exact computer computation showed that the standard deviation of X_i is within .036 of unity for $n = 45$ for all p_i between 0.15 and 0.85.

[3] For most of this discussion we will regard the values of p_i of Column 2, Table 1 and θ_i as the quantities to be estimated, although we actually have a prediction problem because these quantities are estimates of the mean of Y_i. Accounting for this fact would cause Stein's method to compare even more favorably to the sample mean because the random error in p_i increases the losses for all estimators equally. This increases the errors of good estimators by a higher percentage than poorer ones.

1. 1970 Batting Averages for 18 Major League Players and Transformed Values X_i, θ_i

i	Player	Y_i = batting average for first 45 at bats (1)	p_i = batting average for remainder of season (2)	At bats for remainder of season (3)	X_i (4)	θ_i (5)
1	Clemente (Pitts, NL)	.400	.346	367	−1.35	−2.10
2	F. Robinson (Balt, AL)	.378	.298	426	−1.66	−2.79
3	F. Howard (Wash, AL)	.356	.276	521	−1.97	−3.11
4	Johnstone (Cal, AL)	.333	.222	275	−2.28	−3.96
5	Berry (Chi, AL)	.311	.273	418	−2.60	−3.17
6	Spencer (Cal, AL)	.311	.270	466	−2.60	−3.20
7	Kessinger (Chi, NL)	.289	.263	586	−2.92	−3.32
8	L. Alvarado (Bos, AL)	.267	.210	138	−3.26	−4.15
9	Santo (Chi, NL)	.244	.269	510	−3.60	−3.23
10	Swoboda (NY, NL)	.244	.230	200	−3.60	−3.83
11	Unser (Wash, AL)	.222	.264	277	−3.95	−3.30
12	Williams (Chi, AL)	.222	.256	270	−3.95	−3.43
13	Scott (Bos, AL)	.222	.303	435	−3.95	−2.71
14	Petrocelli (Bos, AL)	.222	.264	538	−3.95	−3.30
15	E. Rodriguez (KC, AL)	.222	.226	186	−3.95	−3.89
16	Campaneris (Oak, AL)	.200	.285	558	−4.32	−2.98
17	Munson (NY, AL)	.178	.316	408	−4.70	−2.53
18	Alvis (Mil, NL)	.156	.200	70	−5.10	−4.32

The results are striking. The sample mean $\mathbf{X}$ has total squared prediction error $\sum (X_i - \theta_i)^2$ of 17.56, but $\tilde{\boldsymbol{\delta}}^1(\mathbf{X}) \equiv (\tilde{\delta}_1{}^1(\mathbf{X}), \cdots, \tilde{\delta}_k{}^1(\mathbf{X}))$ has total squared prediction error of only 5.01. The efficiency of Stein's rule relative to the MLE for these data is defined as $\sum (X_i - \theta_i)^2 / \sum (\tilde{\delta}_i{}^1(\mathbf{X}) - \theta_i)^2$, the ratio of squared error losses. The efficiency of Stein's rule is 3.50 (=17.56/5.01) in this example. Moreover, $\tilde{\delta}_i{}^1$ is closer than X_i to θ_i for 15 batters, being worse only for Batters 1, 10, 15. The estimates (2.5) are retransformed in Table 2 to provide estimates $\hat{p}_i{}^1 = f_n{}^{-1}(\hat{\theta}_i)$ of p_i.

Stein's estimators achieve uniformly lower aggregate risk than the MLE but permit considerably increased risk to individual components of the vector $\boldsymbol{\theta}$. As a function of $\boldsymbol{\theta}$, the risk for estimating θ_1 by $\tilde{\delta}_1{}^1$, for example, can be as large as $k/4$ times as great as the risk of the MLE X_1. This phenomenon is discussed at length in [5, 6], where "limited translation estimators" $\tilde{\boldsymbol{\delta}}^s(\mathbf{X})$ $0 \leq s \leq 1$ are introduced to reduce this effect. The MLE corresponds to $s = 0$, Stein's estimator to $s = 1$. The estimate $\tilde{\delta}_i{}^s(\mathbf{X})$ of θ_i is defined to be as close as possible to $\tilde{\delta}_i{}^1(\mathbf{X})$ subject to the condition that it not differ from X_i by more than $[(k-1)(k-3)/kV]^{\frac{1}{2}} D_{k-1}(s)$ standard deviations of X_i, $D_{k-1}(s)$ being a constant taken from [6, Table 1]. If $s = 0.8$, then $D_{17}(s) = 0.786$, so $\tilde{\delta}_i{}^{0.8}(\mathbf{X})$ may differ from X_i by no more than

$$0.786\ (17 \times 0.791/18)^{\frac{1}{2}} = .68$$

This modification reduces the maximum component risk of 4.60 for $\tilde{\delta}_i{}^1$ to 1.52 for $\tilde{\delta}_i{}^{0.8}$ while retaining 80 percent of the savings of Stein's rule over the MLE. The retransformed values $\hat{p}_i{}^{0.8}$ of the limited translation estimates $f_n{}^{-1}(\tilde{\delta}_i{}^{0.8}(\mathbf{X}))$ are given in the last column of Table 2, the estimates for the top three and bottom two batters being affected. Values for $s = 0.9$ are also given in Table 2.

2. Batting Averages and Their Estimates

i	Batting average for season remainder p_i	Maximum likelihood estimate Y_i	Retransform of Stein's estimator $\hat{p}_i{}^1$	Retransform of $\tilde{\delta}^{0.9}$ $\hat{p}_i{}^{0.9}$	Retransform of $\tilde{\delta}^{0.8}$ $\hat{p}_i{}^{0.8}$
1	.346	.400	.290	.334	.351
2	.298	.378	.286	.313	.329
3	.276	.356	.281	.292	.308
4	.222	.333	.277	.277	.287
5	.273	.311	.273	.273	.273
6	.270	.311	.273	.273	.273
7	.263	.289	.268	.268	.268
8	.210	.267	.264	.264	.264
9	.269	.244	.259	.259	.259
10	.230	.244	.259	.259	.259
11	.264	.222	.254	.254	.254
12	.256	.222	.254	.254	.254
13	.303	.222	.254	.254	.254
14	.264	.222	.254	.254	.254
15	.226	.222	.254	.254	.254
16	.285	.200	.249	.249	.242
17	.316	.178	.244	.233	.218
18	.200	.156	.239	.208	.194

Clemente ($i = 1$) was known to be an exceptionally good hitter from his performance in other years. Limiting translation results in a much better estimate for him, as we anticipated, since $\tilde{\delta}_1{}^1(\mathbf{X})$ differs from X_1 by an excessive 1.56 standard deviations of X_1. The limited translation estimators are closer than the MLE for 16 of the 18 batters, and the case $s = 0.9$ has better efficiency (3.91) for these data relative to the MLE than Stein's rule (3.50), but the rule with $s = 0.8$ has lower efficiency (3.01). The maximum component error occurs for Munson ($i = 17$) with all four estimators. The Bayesian effect is so strong that this maximum error $|\hat{\theta}_{17} - \theta_{17}|$ decreased from 2.17 for $s = 0$, to 1.49 for $s = 0.8$, to 1.25 for $s = 0.9$ to 1.08 for $s = 1$. Limiting translation

therefore increases the worst error in this example, just opposite to the maximum risks.

3. A GENERALIZATION OF STEIN'S ESTIMATOR TO UNEQUAL VARIANCES FOR ESTIMATING THE PREVALENCE OF TOXOPLASMOSIS

One of the authors participated in a study of toxoplasmosis in El Salvador [14]. Sera obtained from a total sample of 5,171 individuals of varying ages from 36 El Salvador cities were analyzed by a Sabin-Feldman dye test. From the data given in [14, Table 1], toxoplasmosis prevalence rates X_i for City i, $i = 1, \cdots, 36$ were calculated. The prevalence rate X_i has the form (observed minus expected)/expected, with "observed" being the number of positives for City i and "expected" the number of positives for the same city based on an indirect standardization of prevalence rates to the age distribution of City i. The variances $D_i = \text{Var}\,(X_i)$ are known from binomial considerations and differ because of unequal sample sizes.

These data X_i together with the standard deviations $D_i^{\frac{1}{2}}$ are given in Columns 2 and 3 of Table 3. The prevalence rates satisfy a linear constraint $\sum d_i X_i = 0$ with known coefficients $d_i > 0$. The means $\theta_i = EX_i$, which also satisfy $\sum d_i\theta_i = 0$, are to be estimated from the $\{X_i\}$. Since the $\{X_i\}$ were constructed as sums of independent random variables, they are approximately normal; and except for the one linear constraint on the $k = 36$ values of X_i, they are independent. For simplicity, we will ignore the slight improvement in the independence approximation that would result from applying our methods to an appropriate 35-dimensional subspace and assume that the $\{X_i\}$ have the distribution of the following paragraph.

3. Estimates and Empirical Bayes Estimates of Toxoplasmosis Prevalence Rates

i	X_i	$\sqrt{D_i}$	$\delta_i(\mathbf{X})$	$\hat{A}_i$	$\hat{k}_i$	$\hat{B}_i$
1	.293	.304	.035	.0120	1334.1	.882
2	.214	.039	.192	.0108	21.9	.102
3	.185	.047	.159	.0109	24.4	.143
4	.152	.115	.075	.0115	80.2	.509
5	.139	.081	.092	.0112	43.0	.336
6	.128	.061	.100	.0110	30.4	.221
7	.113	.061	.088	.0110	30.4	.221
8	.098	.087	.062	.0113	48.0	.370
9	.093	.049	.079	.0109	25.1	.154
10	.079	.041	.070	.0109	22.5	.112
11	.063	.071	.045	.0111	36.0	.279
12	.052	.048	.044	.0109	24.8	.148
13	.035	.056	.028	.0110	28.0	.192
14	.027	.040	.024	.0108	22.2	.107
15	.024	.049	.020	.0109	25.1	.154
16	.024	.039	.022	.0108	21.9	.102
17	.014	.043	.012	.0109	23.1	.122
18	.004	.085	.003	.0112	46.2	.359
19	−.016	.128	−.007	.0116	101.5	.564
20	−.028	.091	−.017	.0113	51.6	.392
21	−.034	.073	−.024	.0111	37.3	.291
22	−.040	.049	−.034	.0109	25.1	.154
23	−.055	.058	−.044	.0110	28.9	.204
24	−.083	.070	−.060	.0111	35.4	.273
25	−.098	.068	−.072	.0111	34.2	.262
26	−.100	.049	−.085	.0109	25.1	.154
27	−.112	.059	−.089	.0110	29.4	.210
28	−.138	.063	−.106	.0110	31.4	.233
29	−.156	.077	−.107	.0112	40.0	.314
30	−.169	.073	−.120	.0111	37.3	.291
31	−.241	.106	−.128	.0114	68.0	.468
32	−.294	.179	−.083	.0118	242.4	.719
33	−.296	.064	−.225	.0111	31.9	.238
34	−.324	.152	−.114	.0117	154.8	.647
35	−.397	.158	−.133	.0117	171.5	.665
36	−.665	.216	−.140	.0119	426.8	.789

To obtain an appropriate empirical Bayes estimation rule for these data we assume that

$$X_i|\theta_i \overset{\text{ind}}{\sim} N(\theta_i, D_i), \quad i = 1, \cdots, k \tag{3.1}$$

and

$$\theta_i \overset{\text{ind}}{\sim} N(0, A), \quad i = 1, \cdots, k\ , \tag{3.2}$$

A being an unknown constant. These assumptions are the same as (1.1), (1.7), which lead to the James-Stein estimator if $D_i = D_j$ for all i, j. Notice that the choice of *a priori* mean zero for the θ_i is particularly appropriate here because the constant $\sum d_i\theta_i = 0$ forces the parameters to be centered near the origin.

We require $k \geq 3$ in the following derivations. Define

$$B_i \equiv D_i/(A + D_i)\ . \tag{3.3}$$

Then (3.1) and (3.2) are equivalent to

$$\theta_i|X_i \overset{\text{ind}}{\sim} N((1 - B_i)X_i, D_i(1 - B_i)), \quad i = 1, \cdots, k\ . \tag{3.4}$$

For squared error loss[4] the Bayes estimator is the *a posteriori* mean

$$\delta_i^*(X_i) = E\theta_i|X_i = (1 - B_i)X_i\ , \tag{3.5}$$

with Bayes risk $\text{Var}\,(\theta_i|X_i) = (1 - B_i)D_i$ being less than the risk D_i of $\hat{\theta}_i = X_i$.

Here, A is unknown, but the MLE $\hat{A}$ of A on the basis of the data $S_j \equiv X_j^2 \sim (A + D_j)\chi_1^2$, $j = 1, 2, \cdots, k$ is the solution to

$$\hat{A} = \sum_{j=1}^{k} (S_j - D_j)I_j(\hat{A}) \Big/ \sum_{j=1}^{k} I_j(\hat{A}) \tag{3.6}$$

with

$$I_j(A) \equiv 1/\text{Var}\,(S_j) = 1/[2(A + D_j)^2] \tag{3.7}$$

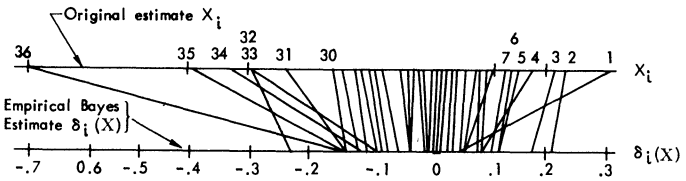
being the Fisher information for A in S_j. We could use $\hat{A}$ from (3.6) to define the empirical Bayes estimator of θ_i as $(1 - D_i/(\hat{A} + D_i))X_i$. However, this rule does not reduce to Stein's when all D_j are equal, and we instead use a minor variant of this estimator derived in [8] which does reduce to Stein's. The variant rule estimates a different value $\hat{A}_i$ for each city (see Table 3). The difference between the rules is minor in this case, but it might be important if k were smaller.

Our estimates $\delta_i(\mathbf{X})$ of the θ_i are given in the fourth column of Table 3 and are compared with the unbiased

[4] Or for any other increasing function of $|\theta_i - \hat{\theta}_i|$.

estimate X_i in Figure A. Figure A illustrates the "pull in" effect of $\delta_i(\mathbf{X})$, which is most pronounced for Cities 1, 32, 34, 35, and 36. Under the empirical Bayes model, the major explanation for the large $|X_i|$ for these cities is large D_i rather than large $|\theta_i|$. This figure also shows that the rankings of the cities on the basis of $\delta_i(\mathbf{X})$ differs from that based on the X_i, an interesting feature that does not arise when the X_i have equal variances.

A. Estimates of Toxoplasmosis Prevalence Rates



The values $\hat{A}_i$, $\hat{k}_i$, and $\hat{B}_i(S)$ defined in [8] are given in the last three columns of Table 3. The value $\hat{A}$ of (3.6) is $\hat{A} = 0.0122$ with standard deviation $\sigma(\hat{A})$ estimated as 0.0041 (if $A = 0.0122$) by the Cramér-Rao lower bound on $\sigma(\hat{A})$. The preferred estimates $\hat{A}_i$ are all close to but slightly smaller than $\hat{A}$, and their estimated standard deviations vary from 0.00358 for the cities with the smallest D_i to 0.00404 for the city with the largest D_i.

The likelihood function of the data plotted as a function of A (on a log scale) is given in Figures B and C as LIKELIHOOD. The curves are normalized to have unit area as a function of $\alpha = \log A$. The maximum value of this function of α is at $\hat{\alpha} = \log(\hat{A}) = \log(.0122) = -4.40 \equiv \mu_\alpha$. The curves are almost perfectly normal with mean $\hat{\alpha} = -4.40$ and standard deviation $\sigma_\alpha \equiv .371$. The likely values of A therefore correspond to a α differing from μ_α by no more than three standard deviations, $|\alpha - \mu_\alpha| \leq 3\sigma_\alpha$, or equivalently, $.0040 \leq A \leq .0372$.

In the region of likely values of A, Figure B also graphs two risks: BAYES RISK and EB RISK (for empirical Bayes risk), each conditional on the data $\mathbf{X}$. EB RISK[5] is the conditional risk of the empirical Bayes rule defined (with $D_0 \equiv (1/k)\sum_{i=1}^{k} D_i$) as

$$E_A \frac{1}{kD_0} \sum_{i=0}^{k} (\delta_i(\mathbf{X}) - \theta_i)^2 | \mathbf{X} , \tag{3.8}$$

and BAYES RISK is

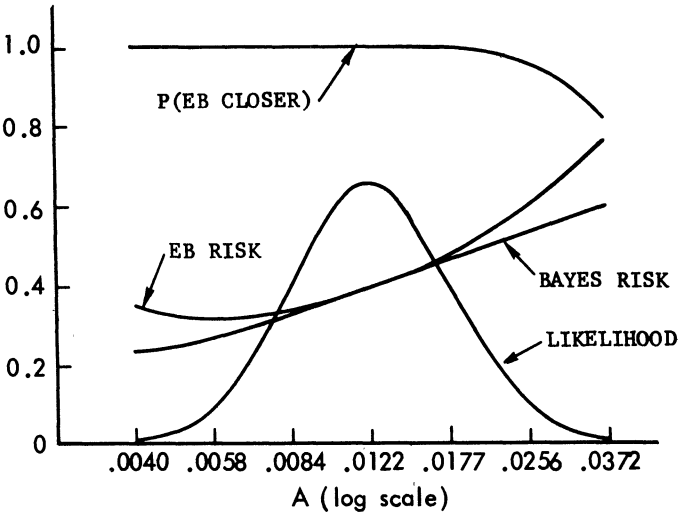
$$E_A \frac{1}{kD_0} \sum_{i=1}^{k} \left(\frac{A}{A + D_i} X_i - \theta_i \right)^2 \Bigg| \mathbf{X} . \tag{3.9}$$

Since A is not known, BAYES RISK yields only a lower envelope for empirical Bayes estimators, agreeing with EB RISK at $A = .0122$. Table 4 gives values to supplement Figure B. Not graphed because it is too large to fit in Figure B is MLE RISK, the conditional risk of the MLE, defined as

$$E_A \frac{1}{kD_0} \sum_{i=1}^{k} (X_i - \theta_i)^2 | \mathbf{X} . \tag{3.10}$$

MLE RISK exceeds EB RISK by factors varying from 7 to 2 in the region of likely values of A, as shown in Table 4. EB RISK tends to increase and MLE RISK to decrease as A increases, these values crossing at $A = .0650$, about $4\frac{1}{2}$ standard deviations above the mean of the distribution of $\hat{A}$.

B. Likelihood Function of A and Aggregate Operating Characteristics of Estimates as a Function of A, Conditional on Observed Toxoplasmosis Data



4. Conditional Risks for Different Values of A

Risk	A				
	.0040	.0122	.0372	.0650	∞
EB RISK	.35	.39	.76	1.08	2.50
MLE RISK	2.51	1.87	1.27	1.08	1.00
P(EB CLOSER)	1.00	1.00	.82	.50	.04

The remaining curve in Figure B graphs the probability that the empirical Bayes estimator is closer to $\boldsymbol{\theta}$ than the MLE $\mathbf{X}$, conditional on the data $\mathbf{X}$. It is defined as

$$P_A[\sum (\delta_i(\mathbf{X}) - \theta_i)^2 < \sum (X_i - \theta_i)^2 | \mathbf{X}] . \tag{3.11}$$

This curve, denoted P(EB CLOSER), decreases as A increases but is always very close to unity in the region of likely values of A. It reaches one-half at about $4\frac{1}{2}$ standard deviations from the mean of the likelihood function and then decreases as $A \to \infty$ to its asymptotic value .04 (see Table 4).

The data suggest that almost certainly A is in the interval $.004 \leq A \leq .037$, and for all such values of A, Figure B and Table 4 indicate that the numbers $\delta_i(\mathbf{X})$ are much better estimators of the θ_i than are the X_i. Non-Bayesian versions of these statements may be based on a confidence interval for $\sum \theta_i^2/k$.

Figure A illustrates that the MLE and the empirical Bayes estimators order the $\{\theta_i\}$ differently. Define the

[5] In (3.8) the $\delta_i(\mathbf{X})$ are fixed numbers—those given in Table 3. The expectation is over the *a posteriori* distribution (3.4) of the θ_i.

correlation of an estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ by

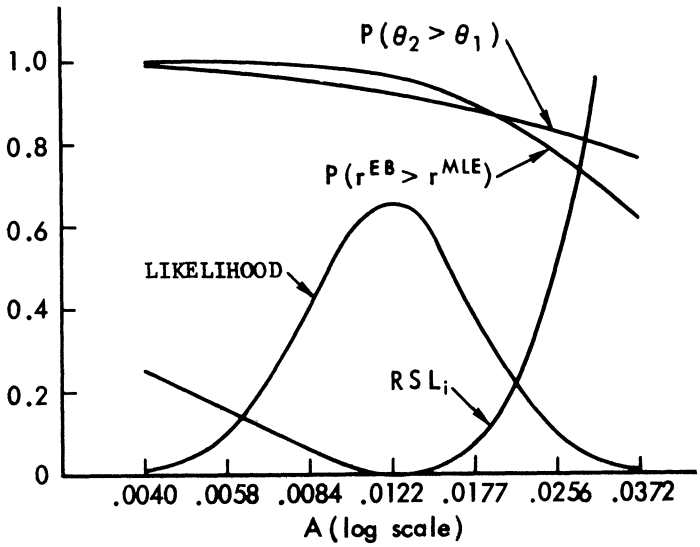
$$r(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sum \hat{\theta}_i \theta_i / (\sum \hat{\theta}_i^2 \sum \theta_i^2)^{\frac{1}{2}} \quad (3.12)$$

as a measure of how well $\hat{\boldsymbol{\theta}}$ orders $\boldsymbol{\theta}$. We denote $P(r^{EB} > r^{MLE})$ as the probability that the empirical Bayes estimate $\boldsymbol{\delta}$ orders $\boldsymbol{\theta}$ better than $\mathbf{X}$, i.e., as

$$P_A\{r(\boldsymbol{\delta}, \boldsymbol{\theta}) > r(\mathbf{X}, \boldsymbol{\theta}) \,|\, \mathbf{X}\} \ . \quad (3.13)$$

The graph of (3.13) given in Figure C shows that $P(r^{EB} > r^{MLE}) > .5$ for $A \leq .0372$. The value at $A = \infty$ drops to .046.

C. Likelihood Function of A and Individual and Ordering Characteristics of Estimates as a Function of A, Conditional on Observed Toxoplasmosis Data



Although $X_1 > X_2$, the empirical Bayes estimator for City 2 is larger, $\delta_2(\mathbf{X}) > \delta_1(\mathbf{X})$. This is because $D_1 \gg D_2$, indicating that X_1 is large under the empirical Bayes model because of randomness while X_2 is large because θ_2 is large. The other curve in Figure C is

$$P_A(\theta_2 > \theta_1 | \mathbf{X}) \quad (3.14)$$

and shows that $\theta_2 > \theta_1$ is quite probable for likely values of A. This probability declines as $A \to \infty$, being .50 at $A = .24$ (eight standard deviations above the mean) and .40 at $A = \infty$.

4. USING STEIN'S ESTIMATOR TO IMPROVE THE RESULTS OF A COMPUTER SIMULATION

A Monte Carlo experiment is given here in which several forms of Stein's method all double the experimental precision of the classical estimator. The example is realistic in that the normality and variance assumptions are approximations to the true situation.

We chose to investigate Pearson's chi-square statistic for its independent interest and selected the particular parameters ($m \leq 24$) from our prior belief that empirical Bayes methods would be effective for these situations. Although our beliefs were substantiated, the outcomes in this instance did not always favor our pet methods.

The simulation was conducted to estimate the exact size of Pearson's chi-square test. Let Y_1 and Y_2 be independent binomial random variables, $Y_1 \sim$ bin (m, p'), $Y_2 \sim$ bin (m, p'') so $EY_1 = mp'$, $EY_2 = mp''$. Pearson advocated the statistic and critical region

$$T = \frac{2m(Y_1 - Y_2)^2}{(Y_1 + Y_2)(2m - Y_1 - Y_2)} > 3.84 \quad (4.1)$$

to test the composite null hypothesis $H_0: p' = p''$ against all alternatives for the nominal size $\alpha = 0.05$. The value 3.84 is the 95th percentile of the chi-square distribution with one degree of freedom, which approximates that of T when m is large.

The true size of the test under H_0 is defined as

$$\alpha(p, m) \equiv P(T > 3.84 | p, m) \ , \quad (4.2)$$

which depends on both m and the unknown value $p \equiv p' = p''$. The simulation was conducted for $p = 0.5$ and the $k = 17$ values of m with $m_j = 7 + j$, $j = 1, \cdots, k$. The k values of $\alpha_j \equiv \alpha(0.5, m_j)$ were to be estimated. For each j we simulated (4.1) $n = 500$ times on a computer and recorded Z_j as the proportion of times H_0 was rejected. The data appear in Table 5. Since $nZ_j \sim$ bin (n, α_j) independently, Z_j is the unbiased and maximum likelihood estimator usually chosen[6] to estimate α_j.

5. Maximum Likelihood Estimates and True Values for p = 0.5

	MLE		True values
j	m_j	Z_j	α_j
1	8	.082	.07681
2	9	.042	.05011
3	10	.046	.04219
4	11	.040	.05279
5	12	.054	.06403
6	13	.084	.07556
7	14	.036	.04102
8	15	.036	.04559
9	16	.040	.05151
10	17	.050	.05766
11	18	.078	.06527
12	19	.030	.05306
13	20	.036	.04253
14	21	.060	.04588
15	22	.052	.04896
16	23	.046	.05417
17	24	.054	.05950

Under H_0 the standard deviation of Z_j is approximately $\sigma = \{(.05)(.95)/500\}^{\frac{1}{2}} = .009747$. The variables $X_j \equiv (Z_j - .05)/\sigma$ have expectations

$$\theta_j \equiv EX_j = (\alpha_j - .05)/\sigma$$

[6] We ignore an extensive bibliography of other methods for improving computer simulations. Empirical Bayes methods can be applied simultaneously with other methods, and if better estimates of α_j than Z_j were available then the empirical Bayes methods could instead be applied to them. But for simplicity we take Z_j itself as the quantity to be improved.

and approximately the distribution

$$X_j|\theta_j \overset{\text{ind}}{\sim} N(\theta_j, 1), \quad j = 1, 2, \cdots, 17 = k \ , \quad (4.3)$$

described in earlier sections.

The average value $\bar{Z} = .051$ of the 17 points supports the choice of the "natural origin" $\bar{\alpha} = .05$. Stein's rule (1.4) applied to the transformed data (4.3) and then retransformed according to $\hat{\alpha}_j = .05 + \sigma\hat{\theta}_j$ yields

$$\hat{\alpha}_j = (1 - \hat{B})Z_j + .05\hat{B} \ , \quad \hat{B} = .325 \ , \quad (4.4)$$

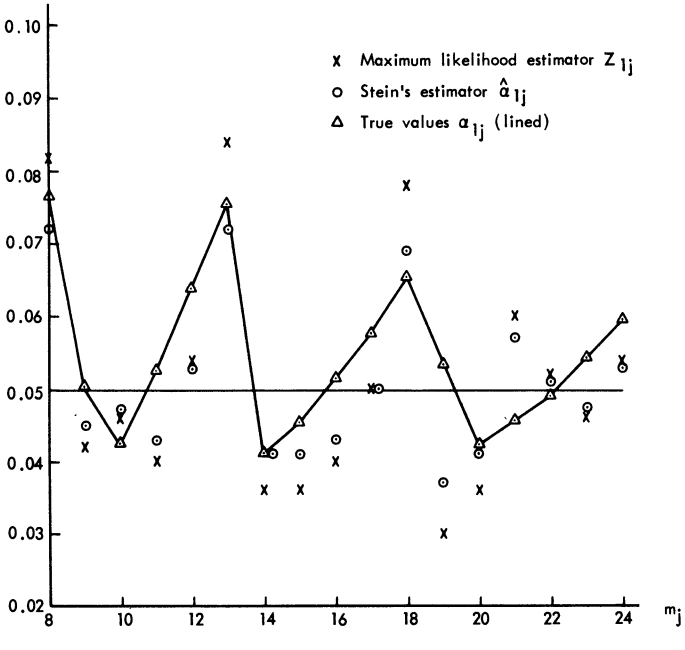
where $\hat{B} \equiv (k - 2)/S$ and

$$S \equiv \sum_{j=1}^{17} (Z_j - .05)^2/\sigma^2 = 46.15 \ .$$

All 17 true values α_j were obtained exactly through a separate computer program and appear in Figure D and Table 5, so the loss function, taken to be the normalized sum of squared errors $\sum (\hat{\alpha}_j - \alpha_j)^2/\sigma^2$, can be evaluated.[7] The MLE has loss 18.9, Stein's estimate (4.4) has loss 10.2, and the constant estimator, which always estimates α_j as .05, has loss 23.4. Stein's rule therefore dominates both extremes between which it compromises.

Figure D displays the maximum likelihood estimates, Stein estimates, and true values. The true values show a surprising periodicity, which would frustrate attempts at improving the MLE by smoothing.

D. MLE, Stein Estimates, and True Values for $p = 0.5$



On theoretical grounds we know that the approximation $\alpha(p, m) = .05$ improves as m increases, which suggests dividing the data into two groups, say $8 \leq m \leq 16$ and $17 \leq m \leq 24$. In the Bayesian framework [9] this disaggregation reflects the concern that A_1, the expectation of $A_1^* \equiv \sum_{j=1}^{9} (\alpha_j - .05)^2/9\sigma^2$ may be much larger than A_2, the expectation of $A_2^* = \sum_{j=10}^{17} (\alpha_j - .05)^2/8\sigma^2$, or equivalently that the pull-in factor $B_1 = 1/(1 + A_1)$ for Group 1 really should be smaller than $B_2 = 1/(1 + A_2)$ for Group 2.

The combined estimator (4.4), having $\hat{B}_1 = \hat{B}_2$, is repeated in the second row of Table 6 with loss components for each group. The simplest way to utilize separate estimates of B_1 and B_2 is to apply two separate Stein rules, as shown in the third row of the table.

6. Values of $\hat{B}$ and Losses for Data Separated into Two Groups, Various Estimation Rules

Rule	$8 \leq m \leq 16$ $\hat{B}_1$	Group 1 loss	$17 \leq m \leq 24$ $\hat{B}_2$	Group 2 loss	Total loss
Maximum Likelihood Estimator	.000	7.3	.000	11.6	18.9
Stein's rule, combined data	.325	4.2	.325	6.0	10.2
Separate Stein rules	.232	4.5	.376	5.4	9.9
Separate Stein rules, bigger constant	.276	4.3	.460	4.6	8.9
All estimates at .05	1.000	18.3	1.000	5.1	23.4

In [8, Sec. 5] we suggest using the bolder estimate

$$\hat{B}_i = (k_i - .66)/S_i \ , \quad S_1 \equiv \sum_{j=1}^{9} (Z_j - .05)^2/\sigma^2 \ ,$$

$$S_2 \equiv S - S_1 \ , \quad k_1 = 9 \ , \quad k_2 = 8 \ .$$

The constant $k_i - .66$ is preferred because it accounts for the fact that the positive part (1.12) will be used, whereas the usual choice $k_i - 2$ does not. The fourth row of Table 6 shows the effectiveness of this choice.

The estimate of .05, which is nearly the mean of the 17 values, is included in the last row of the table to show that the Stein rules substantially improve the two extremes between which they compromise.

The actual values are

$$A_1^* = \sum_{j=1}^{9} (\alpha_j - .05)^2/9\sigma^2 = 2.036$$

for Group 1 and

$$A_2^* = \sum_{j=10}^{17} (\alpha_j - .05)^2/8\sigma^2 = .635 \ ,$$

so $B_1^* = 1/(1 + A_1^*) = .329$ and $B_2^* = 1/(1 + A_2^*) = .612$. The true values of B_1^* and B_2^* are somewhat different, as estimates for separate Stein rules suggest. Rules with $\hat{B}_1$ and $\hat{B}_2$ near these true values will ordinarily perform better for data simulated from these parameters $p = 0.5$, $m = 8, \cdots, 24$.

5. CONCLUSIONS

In the baseball, toxoplasmosis, and computer simulation examples, Stein's estimator and its generalizations increased efficiencies relative to the MLE by about 350 percent, 200 percent, and 100 percent. These examples

[7] Exact rejection probabilities for other values of p are given in [12].

were chosen because we expected empirical Bayes methods to work well for them and because their efficiencies could be determined. But we are aware of other successful applications to real data[8] and have suppressed no negative results. Although blind application of these methods would gain little in most instances, the statistician who uses them sensibly and selectively can expect major improvements.

Even when they do not significantly increase efficiency, there is little penalty for using the rules discussed here because they cannot give larger total mean squared error than the MLE and because the limited translation modification protects individual components. As several authors have noted, these rules are also robust to the assumption of the normal distribution, because their operating characteristics depend primarily on the means and variances of the sampling distributions and of the unknown parameters. Nor is the sum of squared error criterion especially important. This robustness is borne out by the experience in this article since the sampling distributions were actually binomial rather than normal. The rules not only worked well in the aggregate here, but for most components the empirical Bayes estimators ranged from slightly to substantially better than the MLE, with no substantial errors in the other direction.

Tukey's comment, that empirical Bayes benefits are unappreciable (Section 1), actually was directed at a method of D.V. Lindley. Lindley's rules, though more formally Bayesian, are similar to ours in that they are designed to pick up the same intercomponent information in possibly related estimation problems. We have not done justice here to the many other contributors to multiparameter estimation, but refer the reader to the lengthy bibliography in [12]. We have instead concentrated on Stein's rule and its generalizations to illustrate the power of the empirical Bayes theory, because the main gains are derived by recognizing the applicability of the theory, with lesser benefit attributable to the particular method used. Nevertheless, we hope other authors will compare their methods with ours on these or other data.

The rules of this article are neither Bayes nor admissible, so they can be uniformly beaten (but not by much; see [8, Sec. 6]). There are several published, admissible, minimax rules which also would do well on the baseball data, although probably not much better than the rule used there, for none yet given is known to dominate Stein's rule with the positive part modification. For applications, we recommend the combination of simplicity, generalizability, efficiency, and robustness found in the estimators presented here.

The most favorable situation for these estimators occurs when the statistician wants to estimate the parameters of a linear model that are known to lie in a high dimensional parameter space H_1, but he suspects that they may lie close to a specified lower dimensional parameter space $H_0 \subset H_1$.[9] Then estimates unbiased for every parameter vector in H_1 may have large variance, while estimates restricted to H_0 have smaller variance but possibly large bias. The statistician need not choose between these extremes but can instead view them as endpoints on a continuum and use the data to determine the compromise (usually a smooth function of the likelihood ratio statistic for testing H_0 versus H_1) between bias and variance through an appropriate empirical Bayes rule, perhaps Stein's or one of the generalizations presented here.

We believe many applications embody these features and that most data analysts will have good experiences with the sensible use of the rules discussed here. In view of their potential, we believe empirical Bayes methods are among the most under utilized in applied data analysis.

[*Received October 1973. Revised February 1975.*]

REFERENCES

[1] Anscombe, F., "The Transformation of Poisson, Binomial and Negative-Binomial Data," *Biometrika*, 35 (December 1948), 246–54.

[2] Baranchik, A.J., "Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution," Technical Report No. 51, Stanford University, Department of Statistics, 1964.

[3] Carter, G.M. and Rolph, J.E., "Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities," *Journal of the American Statistical Association*, 69, No. 348 (December 1974), 880–5.

[4] Efron, B., "Biased Versus Unbiased Estimation," *Advances in Mathematics*, New York: Academic Press (to appear 1975).

[5] ——— and Morris, C., "Limiting the Risk of Bayes and Empirical Bayes Estimators—Part I: The Bayes Case," *Journal of the American Statistical Association*, 66, No. 336 (December 1971), 807–15.

[6] ——— and Morris, C., "Limiting the Risk of Bayes and Empirical Bayes Estimators—Part II: The Empirical Bayes Case," *Journal of the American Statistical Association*, 67, No. 337 (March 1972), 130–9.

[7] ——— and Morris, C., "Empirical Bayes on Vector Observations—An Extension of Stein's Method," *Biometrika*, 59, No. 2 (August 1972), 335–47.

[8] ——— and Morris, C., "Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach," *Journal of the American Statistical Association*, 68, No. 341 (March 1973), 117–30.

[9] ——— and Morris, C., "Combining Possibly Related Estimation Problems," *Journal of the Royal Statistical Society*, Ser. B, 35, No. 3 (November 1973; with discussion), 379–421.

[10] ——— and Morris, C., "Families of Minimax Estimators of the Mean of a Multivariate Normal Distribution," P-5170, The RAND Corporation, March 1974, submitted to *Annals of Mathematical Statistics* (1974).

[11] ——— and Morris, C., "Estimating Several Parameters Simultaneously," to be published in *Statistica Neerlandica*.

[12] ——— and Morris, C., "Data Analysis Using Stein's Estimator and Its Generalizations," R-1394-OEO, The RAND Corporation, March 1974.

[13] James, W. and Stein, C., "Estimation with Quadratic Loss,"

[8] See, e.g., [3] for estimating fire alarm probabilities and [4] for estimating reaction times and sunspot data.

[9] One excellent example [17] takes H_0 as the main effects in a two-way analysis of variance and $H_1 - H_0$ as the interactions.

Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, Berkeley: University of California Press, 1961, 361–79.

[14] Remington, J.S., *et al.*, "Studies on Toxoplamosis in El Salvador: Prevalence and Incidence of Toxoplasmosis as Measured by the Sabin-Feldman Dye Test," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 64, No. 2 (1970), 252–67.

[15] Stein, C., "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley: University of California Press, 1955, 197–206.

[16] ———, "Confidence Sets for the Mean of a Multivariate Normal Distribution," *Journal of the Royal Statistical Society*, Ser. B, 24, No. 2 (1962), 265–96.

[17] ———, "An Approach to the Recovery of Inter-Block Information in Balanced Incomplete Block Designs," in F.N. David, ed., *Festschrift for J. Neyman*, New York: John Wiley & Sons, Inc., 1966, 351–66.