

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

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

meaning the $\{X_i\}$ are independent and normally distributed with mean $E_{\theta_i} X_i = \theta_i$ 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 $\theta = (\theta_1, \dots, \theta_k)$ is to be estimated with loss being the sum of squared component errors

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

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

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

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

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

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

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

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

being less than k for all θ , 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.

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

$\theta_i \sim N(\mu_i, \tau^2), i = 1, \dots, k,$ (1.7)

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

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

$S = \sum (X_i - \mu_i)^2 \sim (1 + \tau^2)\chi_k^2,$ (1.9)

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

$E(k - 2)/S = 1/(1 + \tau^2)$ (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^* averaged over both X and θ is, from [6] or [8],

$E_r E_\theta (\delta_i^*(X) - \theta_i)^2 = 1 - (k - 2)/k(1 + \tau^2),$ (1.11)

E_r 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^* 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

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

with $(a^+) \equiv \max(0, a)$. That $R(\theta, \delta_i^{1+}) < R(\theta, \delta_i^*)$ for all θ is proved in [2, 8, 10, 17]. The risks $R(\theta, \delta_i^*)$ and $R(\theta, \delta_i^{1+})$ 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, 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 $i, i = 1, \dots, 18$ ($k = 18$) after $n = 45$ at bats. Assuming base hits occur according to a binomial distribution with independence between players, $nY_i \sim \text{Bin}(n, p_i) i = 1, 2, \dots, 18$ with p_i the true season batting average, $EY_i = p_i$. Because the variance of Y_i depends on p_i , the mean, the arc-sin transformation for stabilizing the variance of a binomial distribution is used: $X_i = f_n(Y_i) i = 1, \dots, 18$ with

$f_n(y) = (n)^{1/2} \arcsin(2y - 1).$ (2.1)

Then X_i has nearly unit variance independent of p_i . The mean 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

$X_i | \theta_i \sim N(\theta_i, 1), i = 1, 2, \dots, k,$ (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 i th component θ_i of θ is therefore

$\delta_i^1(X) = \bar{X} + (1 - (k - 3)/V)(X_i - \bar{X})$ (2.3)

with $V = \sum (X_i - \bar{X})^2$ 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 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

$R(\theta, \delta_i^1) \leq k - \frac{(k - 3)^2}{k - 3 + \sum (\theta_i - \bar{\theta})^2}, \bar{\theta} = \sum \theta_i/k.$ (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 $-.3275$ so

$\delta_i^1(X) = \hat{\theta}_i = .791\bar{X} + .209X_i = .209X_i - 2.59$ (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.

selection of data is biased

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

Player	$Y_i =$ batting average for first 45 at bats	$p_i =$ batting average for remainder of season	At bats for remainder of season	$X_i = 45 \frac{1}{2} \sin^{-1}(2y_i - 1)$	$\theta_i = 45 \frac{1}{2} \sin^{-1}(2p_i - 1)$	
	(1)	(2)	(3)	(4)	(5)	
1	Clemente (Pitts, NL)	.400	346	367	-1.35	-2.10
2	F. Robinson (Balt, AL)	.378	4298	426	-1.66	-2.79
3	F. Howard (Wash, AL)	.356	6276	521	-1.97	-3.11
4	Johnstone (Cal, AL)	.333	16222	275	-2.28	-3.96
5	Berry (Chi, AL)	.311	7273	418	-2.60	-3.17
6	Spencer (Cal, AL)	.311	8270	466	-2.60	-3.20
7	Kessinger (Chi, NL)	.289	12263	586	-2.92	-3.32
8	L. Alvarado (Bos, AL)	.267	17210	138	-3.26	-4.15
9	Santo (Chi, NL)	.244	9269	510	-3.60	-3.23
10	Swoboda (NY, NL)	.244	14230	200	-3.60	-3.83
11	Unser (Wash, AL)	.222	10264	277	-3.95	-3.30
12	Williams (Chi, AL)	.222	13256	270	-3.95	-3.43
13	Scott (Bos, AL)	.222	3303	435	-3.95	-2.71
14	Petrocelli (Bos, AL)	.222	11264	538	-3.95	-3.30
15	E. Rodriguez (KC, AL)	.222	15226	186	-3.95	-3.89
16	Campaneris (Oak, AL)	.200	5285	558	-4.32	-2.98
17	Munson (NY, AL)	.178	2316	408	-4.70	-2.53
18	Alvis (Mil, NL)	.156	19200	70	-5.10	-4.32

$\theta_i = 45 \frac{1}{2} \sin^{-1}(2p_i - 1)$
 Note: θ_i is calculated from p_i using the formula above.

The results are striking. The sample mean \bar{X} has total squared prediction error $\sum (X_i - \theta_i)^2$ of 17.56, but $\hat{\theta}_i(\mathbf{X}) \equiv (\hat{\theta}_1^1(\mathbf{X}), \dots, \hat{\theta}_n^1(\mathbf{X}))$ 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 $\sum (X_i - \theta_i)^2 / \sum (\hat{\theta}_i^1(\mathbf{X}) - \theta_i)^2$, the ratio of squared error losses. The efficiency of Stein's rule is 3.50 (=17.56/5.01) in this example. Moreover, $\hat{\theta}_i^1$ 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 $\hat{p}_i^1 = f_n^{-1}(\hat{\theta}_i^1)$ 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 θ . As a func-

tion of θ , the risk for estimating θ_i by $\hat{\theta}_i^1$, for example, can be as large as $k/4$ times as great as the risk of the MLE X_i . This phenomenon is discussed at length in [5, 6], where "limited translation estimators" $\hat{\theta}_i^s(\mathbf{X})$ $0 \leq s \leq 1$ are introduced to reduce this effect. The MLE corresponds to $s = 0$, Stein's estimator to $s = 1$. The estimate $\hat{\theta}_i^s(\mathbf{X})$ of θ_i is defined to be as close as possible to $\hat{\theta}_i^1(\mathbf{X})$ subject to the condition that it not differ from X_i by more than $[(k-1)(k-3)/kV]^{1/2} D_{k-1}(s)$ 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 $\hat{\theta}_i^{0.8}(\mathbf{X})$ may differ from X_i by no more than

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

This modification reduces the maximum component risk of 4.60 for $\hat{\theta}_i^1$ to 1.52 for $\hat{\theta}_i^{0.8}$ 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 $f_n^{-1}(\hat{\theta}_i^{0.8}(\mathbf{X}))$ 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.

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 $\hat{\theta}_1^1(\mathbf{X})$ 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

2. Batting Averages and Their Estimates

Batting average for season remainder	Maximum likelihood estimate	Retransform of Stein's estimator	Retransform of $\hat{\theta}_i^{0.8}$	Retransform of $\hat{\theta}_i^{0.9}$
p_i	Y_i	\hat{p}_i^1	$\hat{p}_i^{0.8}$	$\hat{p}_i^{0.9}$
.346	.400	.290	.334	.351
.298	.378	.266	.313	.329
.276	.356	.281	.292	.308
.222	.333	.277	.277	.287
.273	.311	.273	.273	.273
.270	.311	.273	.273	.273
.263	.289	.268	.268	.268
.210	.267	.264	.264	.264
.269	.244	.259	.259	.259
.230	.244	.259	.259	.259
.284	.222	.254	.254	.254
.256	.222	.254	.254	.254
.303	.222	.254	.254	.254
.264	.222	.254	.254	.254
.226	.222	.254	.254	.254
.285	.200	.249	.249	.242
.316	.178	.244	.233	.218
.200	.156	.239	.208	.194

Approx into

Lower efficiency

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 , $i = 1, \dots, 36$ 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 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.

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

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

and

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

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

$$B_i = D_i / (A + D_i) = \frac{1}{1 + \frac{A}{D_i}} \quad (3.3)$$

Then (3.1) and (3.2) are equivalent to

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

For squared error loss⁴ the Bayes estimator is the a posteriori mean

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

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 $S_j \equiv X_j^2 \sim (A + D_j)X_j^2$, $j = 1, 2, \dots, k$ is the solution to

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

with

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

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

3. Estimates and Empirical Bayes Estimates of Toxoplasmosis Prevalence Rates

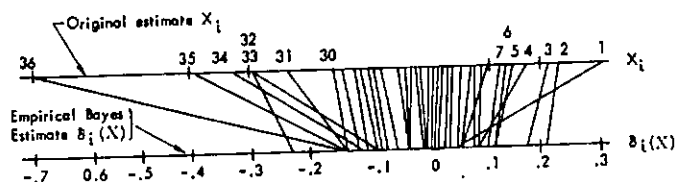
i	X_i	$\sqrt{D_i}$	$\delta_i(X)$	\hat{A}_i	k_i	$\hat{\theta}_i$
1	.293	.302	.035	.0120	1334.1	.882
2	.214	.039	.192	.0108	21.9	.102
3	.185	.047	.159	.0109	24.4	.143
4	.152	.115	.075	.0115	80.2	.509
5	.139	.081	.092	.0112	43.0	.336
6	.128	.061	.100	.0110	30.4	.221
7	.113	.061	.088	.0110	30.4	.221
8	.098	.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

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

Data Analysis Using Stein's Estimator

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 RISK⁵ is the conditional risk of the empirical Bayes rule defined (with $D_0 \equiv (1/k) \sum_{i=1}^k D_i$) as

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

and BAYES RISK is

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

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

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

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} .

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 θ than the MLE \mathbf{X} , conditional on the data \mathbf{X} . It is defined as

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

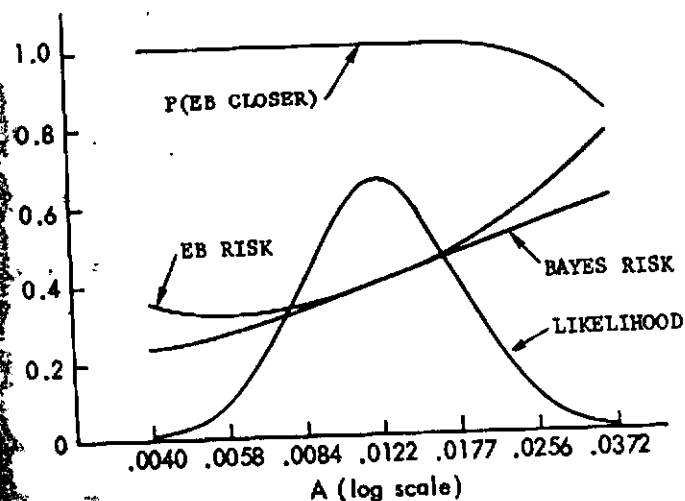
This curve, denoted $P(\text{EB CLOSER})$, decreases as A increases but is always very close to unity in the region of likely values of A . It reaches one-half at about $4\frac{1}{2}$ standard deviations from the mean of the likelihood function and then decreases as $A \rightarrow \infty$ to its asymptotic 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 α posteriori distribution (3.4) of the θ_i .

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



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

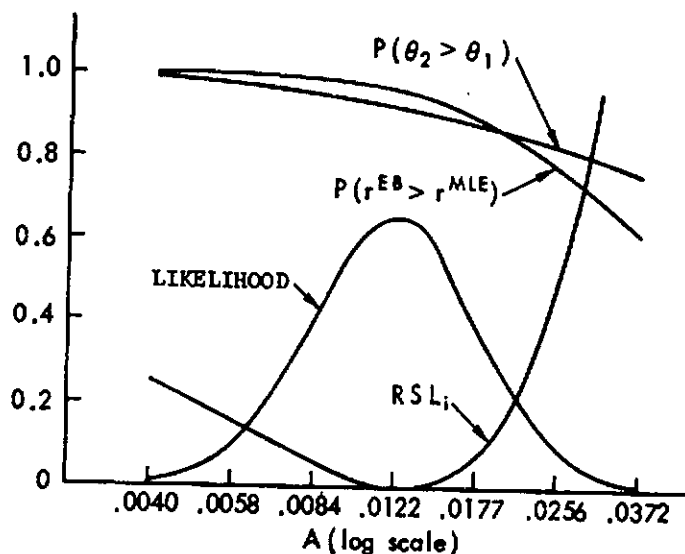
$$r(\hat{\theta}, \theta) = \frac{\sum \hat{\theta}_i \theta_i}{(\sum \hat{\theta}_i^2 \sum \theta_i^2)^{1/2}} \quad (3.12)$$

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

$$P_A\{r(\hat{\theta}, \theta) > r(X, \theta) | X\} \quad (3.13)$$

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, $\delta_2(X) > \delta_1(X)$. 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

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

and shows that $\theta_2 > \theta_1$ is quite probable for likely values of A . This probability declines as $A \rightarrow \infty$, 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, $Y_1 \sim \text{bin}(m, p')$, $Y_2 \sim \text{bin}(m, p'')$ 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 $m_j = 7 + j$, $j = 1, \dots, k$. 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 $nZ_j \sim \text{bin}(n, \alpha_j)$ independently, Z_j is the unbiased and maximum likelihood estimator usually chosen* to estimate α_j .

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

j	MLE		True values α_j
	m_j	Z_j	
1	8	.082	.07681
2	9	.042	.05011
3	10	.046	.04218
4	11	.040	.05279
5	12	.054	.06403
6	13	.084	.07556
7	14	.038	.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 $\sigma = \{(.05)(.95)/500\}^{1/2} = .009747$. The variables $X_j \equiv (Z_j - .05)/\sigma$ have expectations

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

* 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

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

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

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

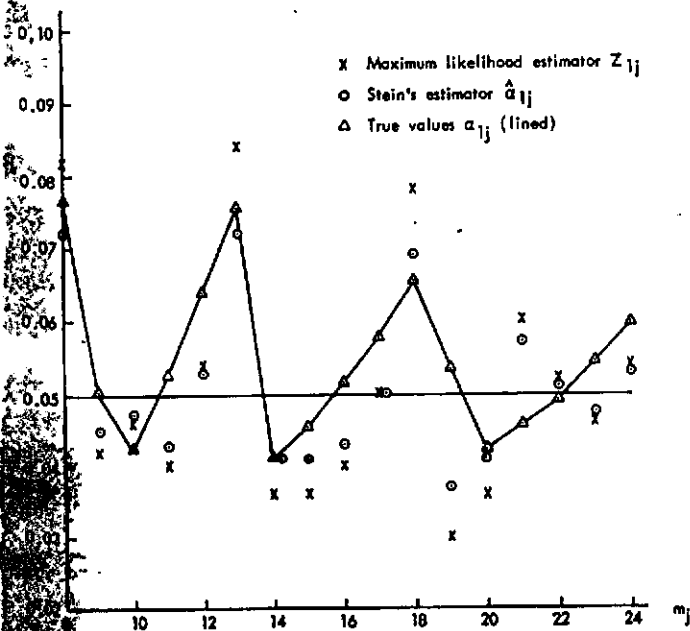
where $\hat{B} \equiv (k - 2)/S$ and

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

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. 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 aggregation reflects the concern that A_1 , the expecta-

tion 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 $A_2^* \equiv \sum_{j=10}^{17} (\alpha_j - .05)^2 / 8\sigma^2$, 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 $p = 0.5, m = 8, \dots, 24$.

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

estimation probabilities for other values of p are given in [13].

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 data⁸ 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 Toxoplasmosis 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.

Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities

GRACE M. CARTER and JOHN E. ROLPH*

An empirical Bayes approach is used to derive a Stein-type estimator of a multivariate normal mean when the components have unequal variances. This estimator is applied to estimating the probability that a fire alarm reported from a particular street box signals a structural fire rather than a false alarm or other emergency. The approach is to group alarm boxes into relatively homogeneous neighborhoods and to make empirical Bayes estimates of the "probability structural" for each box in the neighborhood from yearly (1967-1969) Bronx data. A dispatching rule based on the estimates is evaluated on 1970 data.

1. INTRODUCTION AND SUMMARY

In this article we use an empirical Bayes approach to generalize the usual James-Stein estimator [6] so that it can be used in a class of spatial analysis problems. We then present an application of our procedure to the problem of estimating spatially varying fire alarm probabilities.

The problem addressed here is estimating a parameter $\theta(\ell_1, \ell_2)$ which varies as a function of its location coordinates (ℓ_1, ℓ_2) . A sample $X(\ell_1, \ell_2)$ from each location is available for estimating $\theta(\ell_1, \ell_2)$. One possibility is using only the data from location (ℓ_1, ℓ_2) to estimate the corresponding θ so that

$$\hat{\theta}(\ell_1, \ell_2) = f[X(\ell_1, \ell_2)].$$

One can frequently improve on this approach if the values of $\theta(\ell_1, \ell_2)$ tend to be close to one another for nearby locations. For the approach taken here, we assume that the θ 's tend to vary reasonably smoothly over geography. Further, there is a natural way of creating small groups or neighborhoods of 4 to 40 locations each in which the θ 's are more homogeneous than in the whole space. In each neighborhood we develop empirical Bayes estimates of $\theta(\ell_1, \ell_2)$ which are weighted averages of an estimate based on $X(\ell_1, \ell_2)$ and an estimate based on the average value of the X 's in the neighborhood.

Examples of problems where this approach is applicable occur when the area considered is a large city. Frequently, enough is known about how θ varies with location so that an analyst knowledgeable about the city can construct reasonably homogeneous neighborhoods from looking at a map. A representative listing of

problems for which our approach is designed include

1. Estimating an average demographic characteristic such as family income by city block where the data come from a geographically stratified sample by block of family income.
2. Estimating the average amount of garbage per city block from a one week complete tally of garbage accumulation by block, to determine how frequently collections should be made for each block;
3. Estimating the needs for emergency medical care by city tract so that the ambulances may be located to improve response time;
4. Estimating the probability that a fire alarm reported from a street box signals a structural fire as a function of the location of the alarm box from several years of data for each alarm box. These estimates can be used to decide the number of fire companies to dispatch to a particular alarm. We describe this application in Section 3.

In Section 2, we use an empirical Bayes approach to derive estimators appropriate to estimating $\theta(\ell_1, \ell_2)$. In Section 3 we apply these estimators to estimating fire alarm probabilities in New York City.

2. EMPIRICAL BAYES ESTIMATORS

We begin by assuming that the area in question has been broken into neighborhoods having at least four distinct locations each. Focusing on one neighborhood with k locations, suppose the observations X_i from location i are of the form $X_i = \theta_i + \epsilon_i$ where ϵ_i

$$X_i \sim N(\theta_i, D_i) \quad i = 1, \dots, k$$

(meaning that the $\{X_i\}$ are independent, normally distributed with mean $EX_i = \theta_i$ and known variance D_i). The object is to estimate θ_i for $i = 1, \dots, k$. We give estimators of θ_i here and defer making the connection between the preceding model and the spatial analysis problem of estimating fire alarm probabilities until Section 3.

In this section we use an empirical Bayes approach to generalize the usual James-Stein estimator [7], and apply it to estimating θ_i in the preceding model. The usual maximum likelihood estimator for $\theta = (\theta_1, \dots, \theta_k)$ is $\hat{\theta} = \mathbf{X}$ where $\mathbf{X} = (X_1, \dots, X_k)$. In the situation where the variances are equal and known ($D_i = D$), the positive part version of the James-Stein estimator which

* Grace M. Carter is policy analyst, Information Sciences Department, and John E. Rolph is research statistician, Economics Department, both with the Rand Corporation, 1700 Main St., Santa Monica, Calif. 90406. This work was supported by a contract between the Fire Department of New York City and the New York City Rand Institute.

shrinks the observed values toward their mean is appropriate and is given by

$$[1 - \hat{B}(X)]X + \hat{B}(X)\bar{X}e, \quad (2.1)$$

where $\bar{X} = (1/k) \sum X_i$, $e = (1, \dots, 1)$ and $\hat{B}(X) = \min [1, (k - 3)D/S]$ with $S = \sum (X_i - \bar{X})^2$. Efron and Morris [4, 6] have shown that this type of estimator has desirable properties either with or without the assumption of a prior distribution for the θ_i 's.

When the loss function is squared error and the significance level is not too small, Selove, Morris and Radhakrishnan [8] show that (2.1) uniformly dominates the commonly used "testimator" procedure of first testing the hypothesis that $\theta_1 = \theta_2 = \dots = \theta_k$, and then using X_i or \bar{X} to estimate θ_i , depending on the outcome of the test.

We now give a generalization of the James-Stein estimator in the empirical Bayes context for use when the D_i 's are not equal. This is necessary because in our application, as well as many others, the X 's do not have equal variances. We wish to estimate the mean of each of k distributions given a sufficient statistic (think of it as one observation), from each. The observations $\{X_i; i = 1, \dots, k\}$ are independent, normally distributed with unknown means θ_i and known variances D_i . For some applications, it will be useful to derive the estimates for prior distributions of the θ_i of the form:

$$\theta_i \stackrel{\text{ind}}{\sim} N(\nu, \rho_i A) \quad (2.2)$$

where ρ_i is a known proportionality constant normalized so that $\sum \rho_i = k$. In our application either $\rho_i = 1$ or $\rho_i \propto D_i$. In general ρ_i can be chosen by the analyst to reflect the degree he wishes θ_i to be estimated by X_i as opposed to ν . Thus,

$$X_i \sim N(\nu, D_i + \rho_i A) \quad (2.3)$$

and

$$\theta_i | X_i \sim N[(1 - B_i)X_i + B_i\nu, D_i(1 - B_i)] \quad (2.4)$$

where $B_i = D_i/(\rho_i A + D_i)$. Thus, the Bayes estimate for θ_i is $(1 - B_i)X_i + B_i\nu$. To get an empirical Bayes estimate, suppose $\alpha = (\alpha_1, \dots, \alpha_k)$, $\gamma = (\gamma_1, \dots, \gamma_k)$ and $\sum \gamma_i = 1$. Define

$$S(\alpha, \gamma) = \sum_{i=1}^k \alpha_i [X_i - \bar{X}(\gamma)]^2, \quad \bar{X}(\gamma) = \sum_{i=1}^k \gamma_i X_i.$$

Then

$$[S(\alpha, \gamma)] = \sum_{i=1}^k [(\rho_i A + D_i)\alpha_i - 2(\rho_i A + D_i)\alpha_i \gamma_i + \alpha_i \sum_j (\rho_j A + D_j)\gamma_j^2]. \quad (2.5)$$

From (2.5) it is clear that for $\alpha_i \equiv 1$ and $\gamma_i \equiv 1/k$, $S(\alpha, \gamma) = \sum_{i=1}^k (X_i - \bar{X})^2$ is an unbiased estimator of $(k - 1)(A + \bar{D})$ where $\bar{D} = (1/k) \sum D_i$. The minimum variance unbiased estimator of $(k - 1)(A + \bar{D})$ is $S(\alpha_A, \gamma_A)$ where $\alpha_{A,i} = (A + \bar{D})/(\rho_i A + D_i)$ and $\gamma_{A,i} = \alpha_{A,i}/(\sum_{i=1}^k \alpha_{A,i})$. Using (2.3) we see that

$$S(\alpha_A, \gamma_A) \sim (A + \bar{D})\chi_{k-1}^2,$$

where χ_{k-1}^2 is a chi-square variable with $k - 1$ degrees of freedom. Since A is unknown, we estimate A , and thus α and γ , from the data by defining \hat{A} to be the solution to the equation

$$S(\alpha_{\hat{A}}, \gamma_{\hat{A}}) = (k - 1)(\hat{A} + \bar{D}) \quad (2.6)$$

if the solution is positive and $\hat{A} = 0$ otherwise. A method for computing \hat{A} is given in the appendix.

Letting $S = S(\alpha_{\hat{A}}, \gamma_{\hat{A}})$, an estimate of $\bar{B} = \bar{D}/(A + \bar{D})$ is $(k - 3)\bar{D}/S$. Thus, from the definition of B_i , the appropriate positive part estimator of B_i is

$$\hat{B}_i = \min \left(1, \frac{(k - 3)D_i}{\rho_i S + (k - 3)(D_i - \rho_i \bar{D})} \right)$$

and

$$\hat{\theta}_i = (1 - \hat{B}_i)X_i + \hat{B}_i \bar{X}(\gamma_{\hat{A}}). \quad (2.7)$$

The rationale for using $\bar{X}(\gamma_{\hat{A}})$ is that $\bar{X}(\gamma_A)$ is the minimum variance unbiased estimator of ν . As before, if $D_i \equiv \bar{D}$, our estimator is the usual empirical Bayes estimator given by (2.1) no matter what value \hat{A} takes.

The constant prior estimator occurs when $\rho_i \equiv 1$, so that the appropriate positive part estimator of B_i is

$$\hat{B}_i = \min \left(1, \frac{(k - 3)D_i}{S + (k - 3)(D_i - \bar{D})} \right). \quad (2.8)$$

The proportional prior estimator occurs when $\rho_i \propto D_i$. Here the prior uncertainty on θ_i is proportional to the amount of information on θ_i which the observations will yield. The Bayes estimators for this choice of ρ reduce to using $B_i = D_i/(D_i A + D_i) = (1 + A)^{-1}$. Since for any value of A , $\alpha_{A,i} = (A\bar{D} + \bar{D})/(AD_i + D_i) = \bar{D}/D_i$, the estimator becomes

$$\hat{\theta}_i = (1 - \hat{B})X_i + \hat{B}\bar{X}(\gamma_0)$$

where

$$\hat{B} = \min [1, (k - 3)\bar{D}/S(\alpha_0, \gamma_0)]. \quad (2.9)$$

The proportional prior estimator can also be derived by applying the appropriate equal variance empirical Bayes estimator to the transformed X 's, $X_i/\sqrt{D_i}$ and noting the $E(X_i/\sqrt{D_i}) = \nu/\sqrt{D_i}$.

3. APPLICATION: ESTIMATING THE PROBABILITY THAT AN ALARM SIGNALS A STRUCTURAL FIRE

An estimate of the probability of a particular alarm being a serious fire is useful for making an initial dispatch decision for a fire alarm reported by a street box. If this probability is high, more equipment should be dispatched than when it is low. For our purposes, serious fires are defined as fires in structures. We have used data from the borough of the Bronx in New York and used empirical Bayes methods to estimate the probability that a box-reported alarm signals a structural fire given the alarm box location. To evaluate our techniques, we used data from 1967-9 to develop estimates for 1970 box-reported alarms and then compared our predictions with actual

1970 data. We also used the performance of a dispatching rule based on our estimates as an operationally relevant loss function.

The obvious estimate of the conditional probability that a particular box-reported alarm signals a structural fire is the proportion of box-reported alarms in the past several years at that location that were structural fires. If "location" means alarm box, insufficient data are a problem, since at a given box there are some five to 40 alarms per year, of which about 15 percent are for structural fires. Using "location" to mean a neighborhood containing a sufficient number of boxes may solve the sample size problem, but can be inaccurate if boxes in the same neighborhood have different conditional probabilities that the box-reported alarms indicate structural fires. We use the empirical Bayes estimates developed in Section 2 to make the tradeoff between using estimates for each alarm box based on data from that box alone, and on data averaged over the neighborhood containing that box. A complete account of this work is given in [2].

3.1 Defining the Estimates

The empirical Bayes methods will perform best if applied separately to groups of alarm boxes in the Bronx which are small neighborhoods such that all boxes have similar probabilities of a box-reported alarm signaling a structural fire. Starting with a map of the Bronx and a printout showing the number of box-reported alarms and box-reported structural fires at each box for 1967-9,¹ we formed neighborhoods whose boxes had similar alarm characteristics with the following requirements:

- 1. Neighborhoods should be geographically connected;
- 2. Boxes with obvious geographical properties (those around parks, or on highways), should be grouped together;
- 3. Each neighborhood should have at least 100 alarms in the period 1967-9;
- 4. Each neighborhood should have at least four boxes unless it contains only one box with such a large number of alarms that that box alone could be used to estimate the probability that an alarm signals a structural fire.

Keeping these requirements in mind, we grouped the approximately 2,500 boxes into a set of 216 neighborhoods.

We now show how the empirical Bayes methods presented in Section 2 can be applied to each of the 216 neighborhoods. For a fixed neighborhood, let k be the number of boxes in the neighborhood with at least one alarm, let Y_i be the number of box-reported structural fires and let n_i be the number of box-reported alarms at the i th box in the neighborhood with $N = \sum_{i=1}^k n_i$.

All data referred to here are for 1967-9. Then conditional on n_i , Y_i has a binomial distribution with parameters n_i and p_i . Letting $X_i = Y_i/n_i$, then approximately, $X_i \sim N(p_i, p_i q_i/n_i)$ where $q_i = 1 - p_i$. Since $\text{Var}(X_i) = p_i q_i/n_i$ depends on n_i and on the unknown p_i , the assumptions for the equal variance empirical

Bayes model are violated. One way to handle the dependency on p_i is to use an estimate of $p_i q_i$. We used \hat{p}_i as the estimate where

$$\hat{p}_i = (1/N) \sum_{i=1}^k n_i X_i \quad \text{and} \quad \hat{q}_i = (1 - \hat{p}_i)$$

An alternative method is to make a variance stabilizing transformation. The arcsin transformation of the square root can be used to stabilize the variance. Let

$$X_i = \arcsin(\sqrt{Y_i/n_i}), \quad i = 1, 2, \dots, k,$$

$$\bar{X} = (1/N) \sum_{i=1}^k n_i X_i,$$

then $E(X_i) \doteq \arcsin(\sqrt{p_i})$ and $\text{Var}(X_i) \doteq (1/4n_i)$ where p_i is the conditional probability of a structural fire at box i .

Our unequal variance normal model given in Section 2 is approximately valid in either case. For the untransformed data, $\theta_i = p_i$ and $D_i = (\hat{p}_i \hat{q}_i/n_i)$ while when the variance stabilizing transformation is made, $\theta_i = \arcsin(\sqrt{p_i})$ and $D_i = (1/4n_i)$. Empirical Bayes estimates of θ_i can be made in either case, although using the transformed data, the estimate $\hat{\theta}_i$ of $\arcsin(\sqrt{p_i})$ must be converted to an estimate of p_i by

$$\hat{p}_i = \sin^2(\hat{\theta}_i) + \{(1 - 2 \sin^2 \hat{\theta}_i)/[4(N/k) + 2]\} \hat{B}_i$$

where \hat{B}_i depends on the empirical Bayes model used (see [1]).

We computed the constant prior (2.8), and proportional prior (2.9), empirical Bayes estimates for both the untransformed and transformed data. The resulting four estimates of p_i , as well as the box history estimate (Y_i/n_i) of p_i were compared with the proportion of box-reported alarms in 1970 which signaled structural fires at each box using likelihood ratio tests.

Any of these estimators can be modified using the Efron-Morris limited translation version of the estimator [5]. This modification ensures that the estimator of θ_i is not shifted so far from X_i that $\hat{\theta}_i$ is inconsistent with X_i . Too large a shift can occur if the value of $\hat{B}_i(X_i - \bar{X})$ is large compared to $\sqrt{D_i}$, the standard deviation of $X_i|\theta_i$. In our application, limiting the shift of the estimate means that no one fire alarm box can have its θ_i shifted too much by the surrounding neighborhood.

To use the Efron-Morris modification, we limit the amount θ_i can deviate from X_i to one standard deviation of X_i . Efron and Morris give an extensive evaluation of these methods in [5]. Rather than try the Efron-Morris modification on all four empirical Bayes methods, we arbitrarily chose two methods: the proportional prior model with no transformation and the proportional prior model with transformed data. Less than half of one percent of the box estimates were affected by the modification in either case, and almost no box estimates were changed substantially by the modification. Since the effect was small and computing costs were higher with the modification, we elected not to use it.

¹ Many neighborhoods in New York are changing so rapidly that we decided to use only three years of data.

1. Distribution of \hat{B}_i by Number of Alarms Per Box

Values* of \hat{B}_i	Number of boxes with specified alarm frequencies (1967-69)										Total
	Alarms										
	0	1-5	6-10	11-15	16-20	21-35	36-60	61-120	121-240	Over 240	
0 to .1	0 (.0)	0 (.0)	5 (.01)	6 (.02)	20 (.10)	83 (.27)	97 (.49)	125 (.64)	106 (.75)	67 (.92)	509 (.20)
.1 to .2	0 (.0)	6 (.01)	28 (.07)	84 (.27)	86 (.43)	126 (.41)	73 (.37)	43 (.22)	30 (.21)	0 (.0)	476 (.19)
.2 to .3	0 (.0)	14 (.03)	104 (.25)	112 (.37)	59 (.29)	69 (.22)	16 (.08)	22 (.11)	1 (.01)	0 (.0)	397 (.16)
.3 to .4	0 (.0)	50 (.10)	125 (.30)	57 (.19)	22 (.11)	12 (.04)	6 (.03)	4 (.02)	0 (.0)	0 (.0)	276 (.11)
.4 to .5	0 (.0)	100 (.21)	89 (.21)	25 (.08)	5 (.02)	6 (.02)	4 (.02)	0 (.0)	0 (.0)	0 (.0)	229 (.09)
.5 to .6	0 (.0)	96 (.19)	38 (.09)	7 (.02)	1 (.00)	5 (.02)	1 (.01)	0 (.0)	0 (.0)	0 (.0)	146 (.06)
.6 to .7	0 (.0)	68 (.14)	7 (.02)	6 (.02)	4 (.02)	1 (.00)	0 (.0)	0 (.0)	0 (.0)	0 (.0)	86 (.03)
.7 to .8	0 (.0)	78 (.16)	8 (.02)	4 (.01)	2 (.01)	0 (.0)	0 (.0)	0 (.0)	0 (.0)	0 (.0)	92 (.04)
.8 to .9	0 (.0)	44 (.09)	6 (.01)	0 (.0)	0 (.0)	0 (.0)	0 (.0)	0 (.0)	0 (.0)	0 (.0)	50 (.02)
.9 to 1.0	168 (1.0)	37 (.08)	11 (.03)	5 (.02)	3 (.01)	6 (.02)	2 (.01)	0 (.0)	4 (.03)	6 (.08)	232 (.09)
Mean \hat{B}_i	1.0	.60	.38	.29	.23	.18	.13	.11	.10	.08	.35
No. of boxes	158	493	419	308	202	308	199	194	141	73	2493

* Intervals for values of \hat{B}_i include upper end point. Thus .2 to .3 means $.2 < \hat{B}_i \leq .3$. Standard deviation of $\hat{B}_i = .288$.

\hat{B}_i is calculated using (2.8) assuming a constant prior distribution and no transformation.

NOTE: Numbers in parentheses give the proportions of each column.

All four of the empirical Bayes procedures performed similarly and dominated the box-history estimates when they were compared to 1970 data using likelihood ratio tests. We chose the constant prior estimator with no transformation, because it is simple, and because the constant prior distribution allows the high alarm rate boxes, and low alarm rate boxes, to have different weights attached to neighborhood information, as compared to box-specific information.

Table 1 gives the distribution of \hat{B}_i as a function of the number of box-reported alarms at the box during 1967-9 [see (2.8)]. It is included because the size of \hat{B}_i measures how much the estimate for box i uses the neighborhood estimate \bar{X} , as opposed to the box-history estimate X_i . We see from Table 1 that \hat{B}_i is larger for low alarm-rate boxes and smaller for high alarm-rate boxes. This confirms our intuition that the high alarm-rate boxes should use the box-history estimate more, while the low alarm-rate boxes should use the neighborhood estimate \bar{X} more, since their own box history is too small to give an accurate estimate of θ_i . Thus, the empirical Bayes estimates act as an insurance policy against inaccurate estimation in that they use mostly box history when there is enough of it, and use neighborhood information when there is not enough box history information.

Looking across the row labeled "Mean \hat{B}_i ," we see that the mean of \hat{B}_i is a decreasing function of the number of alarms per box. With the exception of the boxes with very large numbers of alarms, note that the relative distribution of \hat{B}_i shifts from high to low. The columns for boxes with the large number of alarms per box (over 120) show ten boxes listed with a \hat{B}_i of 1.0. These ten boxes are one-box neighborhoods, and thus an empirical Bayes procedure was not applied to those boxes.

box-history estimates of the probability that an alarm reported from Box i signals a structural fire. We compared the performance of these two procedures using a loss function which reflected the way the estimates would be used. Since the estimates will be used to define a dispatch policy, our loss function is based on the properties of the dispatch policies determined by the two estimates.

We consider the class of initial dispatch policies which consist of specifying a number, P^* , such that if for box i , the estimated probability that an alarm is a structural fire is less than P^* , then we send a reduced response to the box and send a larger initial dispatch otherwise. For example, if for Box i the estimate is less than P^* , we send one ladder truck to a box-reported alarm, and otherwise, we dispatch two ladder trucks. To compare the losses when using the empirical Bayes estimators with those of the box-history estimators, we evaluate what would have happened in 1970 if we had used the preceding dispatch policy with each set of estimates based on 1967-9 data.

The performance of each dispatch policy has two dimensions:² a savings in the number of runs to those alarms which received a reduced response and a penalty for the delayed response to those structural fires which required more equipment than was dispatched initially. Because these two effects are not comparable, we fix the number of runs getting a reduced response using each set of estimators, and then compare the number of structural fires that received a reduced response with each dispatch policy.

We ordered all the boxes in the Bronx by the box's estimated probability that an alarm signals a structural fire, using each of the two estimates. For a given number of 1970 alarms receiving a reduced response, the ordering

² It is true that the average time interval from the alarm until the initially dispatched equipment arrives at the scene will be reduced as the number of runs decreases. We may ignore this effect, however, since it is reasonable to assume that it depends primarily on the number of runs saved rather than on the identity of the boxes which receive a reduced response.

The Effect of the Estimates on Initial Dispatch Policies

Our primary goal in developing empirical Bayes estimates was to use them as an alternative to the traditional

of boxes for each estimating method yields a P^* , and thus a set of boxes which should receive a reduced response.

Table 2 compares the number of structural fires which would have received a reduced response using each estimator for the goal of saving about 10 percent (4,500) of the runs and the goal of saving about 50 percent (21,000) of the runs. For each of the two goals, Table 2 shows the performance for all boxes in the Bronx as well as for each of three groups of boxes where each group has an equal number of box-reported alarms. Groups with equal numbers of alarms were used because the effect of estimating errors for the boxes in each group depends on the number of alarms reported from the group, rather than the number of boxes in the group. This is because each alarm is an opportunity for the Fire Department to avoid a wasted overresponse. Examining these three groups separately shows more clearly where the empirical Bayes estimators are superior to the box-history estimators. We separate the boxes which receive a reduced response into the three alarm-rate groups, however, and compare the boxes within each group separately.

The actual number of runs saved is higher than the goals noted in Table 2 because the reduced responses were added in one box increments yielding the slight overshoot shown in the table.

In the three right-hand columns of Table 2, the dispatch policies are based on all the boxes, but the columns for each group refer to only the boxes in that group being compared. For example, if we choose to save 4,500 runs by dispatching only one ladder rather than two ladders to 4,500 alarms, the empirical Bayes dispatch policy will

underrespond to only 138 structural fires as compared to 183 structural fires for the box-history dispatch policy. This savings of 45 underresponses (which is statistically significant), is apportioned among the three groups as follows:

Group	Box history	E.B.
1	162	99
2	21	27
3	0	12
Total	183	138

As these numbers show, the number of underresponses saved (63) by using the empirical Bayes estimates is more than the box-history estimates in Group 1, the high alarm-rate group, although comprising only one-third of the responses, is greater than the savings (45) for the low alarm rate boxes whose low but inaccurate box-history estimates places them in the reduced response category—770 boxes as compared to the empirical Bayes policy having 392 boxes in Group 1.

When our goal is saving half the runs (21,000) the situation changes. The saving of 1,564 - 1,515 = 49 underresponses to structural fires, though still statistically significant, is spread over all three groups with the largest contribution coming from the high alarm-rate group. The goals of saving 4,500 and 21,000 runs were chosen for illustrative purposes. In practical situations the goal will depend on the alarm rates, number of fire companies available, workload and other factors. For a wide range of goals, we found that the losses as measured in number of structural fires receiving underresponses were higher for dispatch rules based on box history estimates, as compared to empirical Bayes estimates.

2. Dispatch Policy Performance^a

Item	All boxes	Group 1 ^b boxes	Group 2 ^b boxes	Group 3 ^b boxes
Goal of 4,500 runs saved				
Number of boxes with reduced response	793 (427)	770 (392)	22 (31)	1 (4)
Total runs saved	4524 (4505)	2708 (2177)	926 (1238)	103 (559)
Structural fires with reduced response	183 (138)	162 (99)	21 (27)	0 (12)
Significance probability ^c	.000	.002	.500	.500
Goal of 21,000 runs saved				
Number of boxes with reduced response	1437 (1600)	1267 (1436)	129 (122)	41 (42)
Total runs saved	21167 (21185)	9221 (9347)	6233 (5974)	5713 (5864)
Structural fires with reduced response	1584 (1515)	569 (564)	466 (435)	527 (518)
Significance probability ^c	.001	.068	.100	.000

^a Numbers in parenthesis refer to empirical Bayes estimators while other numbers refer to box-history estimates. The significance probability is based on a one-tailed test.

^b Groups 1, 2 and 3 comprise those boxes having fewer than 70 alarms, between 71 and 200 alarms and over 200 alarms, respectively, in 1967-9.

^c See [2] for a description of this calculation.

APPENDIX

The computation of \hat{A} and $S(\alpha\lambda, \gamma\lambda)$ is described for the general ρ_i case. To get an iterative method for computing the solution of the equation $S(\alpha\lambda, \gamma\lambda) = (k-1)(\hat{A} + D)$, we use the fact that for any value A_0 ,

$$E[S(\alpha_{A_0}, \gamma_{A_0})] = E\left(\sum_{i=1}^k \alpha_{A_0 i} [X_i - \bar{X}(\gamma_{A_0})]^2\right) \\ = E\left(\sum_{i=1}^k \alpha_{A_0 i} \{ (X_i - \nu)^2 - [\bar{X}(\gamma_{A_0}) - \nu]^2 \}\right) \\ = \sum_{i=1}^k [(A_0 + \bar{D}) / (\rho_i A_0 + \bar{D})] \\ \cdot [(\rho_i A + D_i) - \gamma_{A_0 i} (\rho_i A + D_i)] \\ = \sum_{i=1}^k C_{A_0 i} (\rho_i A + D_i) \quad (A.1)$$

where

$$C_{A_0 i} = \frac{A_0 + \bar{D}}{\rho_i A_0 + D_i} \left[\frac{\sum_{j=1}^k \left(\frac{1}{\rho_j A_0 + D_j} \right) - \frac{1}{\rho_i A_0 + D_i}}{\sum_{j=1}^k \frac{1}{\rho_j A_0 + D_j}} \right] \\ = \alpha_{A_0 i} (1 - \gamma_{A_0 i})$$

Note that if $A_0 = A$,

$$E[S(\alpha_{A_0}, \gamma_{A_0})] = \sum_{i=1}^k (\rho_i A + D_i) C_{A_0 i} = (k-1)(A + \bar{D})$$

Using (A.1), we define a series of estimators for A as follows:

$$\hat{A}_0 = 0 \\ \hat{A}_1 = [S(\alpha_0, \gamma_0) - \sum_{i=1}^k D_i C_{0i}] / \sum_{i=1}^k C_{0i}$$

Continuing in this manner,

$$\hat{A}_{j+1} = [S(\alpha_{\hat{A}_j}, \gamma_{\hat{A}_j}) - \sum_{i=1}^k D_i C_{\hat{A}_j}] / \sum_{i=1}^k G_{\hat{A}_j} \rho_i.$$

The rule is: if $\hat{A}_1 < 0$, set $\hat{A} = 0$, otherwise, define \hat{A} to be the limit of the sequence \hat{A}_j . Note that \hat{A}_j is an unbiased estimator of A for all j so that $S(\alpha_{\hat{A}_j}, \gamma_{\hat{A}_j})$ may be used in (2.7) to estimate B_i .

Substituting $\rho_i = 1$, yields the estimators for the constant prior case. In [4], Efron and Morris give an alternative generalization of the James-Stein estimator for the $\rho_i = 1$ case. In studying regression estimates, Dempster [3] uses estimators which he calls "Stein" and "Ridge" estimators, which are almost identical to our proportional prior and constant prior estimators, respectively. In the proportional prior case $\alpha_{A_i} = \bar{D}/D_i$ for all A , so that \hat{B} can be computed directly, without using the above iterative procedure. Although we have applied only two sets of values of ρ_i , any set given by the analyst can be used with the above procedure to get empirical Bayes estimators.

[Received September 1973. Revised August 1974.]

REFERENCES

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

[2] Carter, Grace M. and Rolph, John E., *New York City Fire Alarm Prediction Models I: Box-Reported Serious Fires*, The Rand Corporation, R-1214-NYC, May 1973.

[3] Dempster, Arthur P., "Alternatives to Least Squares in Multiple Regression," in D.G. Kabe and R.P. Gupta, eds., *Multivariate Statistical Inference*, Amsterdam: North-Holland Publishing Co., 1973.

[4] Efron, Bradley and Morris, Carl, *Data Analysis Using Stein's Estimator and Its Generalizations*, The Rand Corporation, R-1394-OEO (March 1974).

[5] ——— and Morris, Carl, "Limiting the Risk of Bayes and Empirical Bayes Estimators—Part II: The Empirical Bayes Case," *Journal of the American Statistical Association*, 67 (March 1972), 130-39.

[6] ——— and Morris, Carl, "Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach," *Journal of the American Statistical Association*, 68 (March 1973), 117-30.

[7] James, W. and Stein, Charles, "Estimation with Quadratic Loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probabilities*, Berkeley: University of California Press, 1960, 361-79.

[8] Sclove, Stanley L., Morris, Carl and Radhakrishnan, R., "Non-optimality of Preliminary-Test Estimators for the Mean of a Multivariate Normal Distribution," *Annals of Mathematical Statistics*, 43 (October 1972), 1481-90.

Proc N A S
201 cademy sciences

¹ It is amusing to note that the English translation, made under Bateson's influence, changed Mendel's careful wording "nicht zu unterschätzen" to the overstatement "cannot be overestimated."

"not to underestimate" →
could be overestimated }

² The name of August Weismann has been repeatedly misspelled in recent years. It should suggest a wise man, not a white man.

DYNAMIC PROGRAMMING AND STATISTICAL COMMUNICATION THEORY

BY RICHARD BELLMAN AND ROBERT KALABA

RAND CORPORATION, SANTA MONICA, CALIFORNIA

Communicated by P. A. Smith, May 23, 1957

1. *Introduction.*—The purpose of this paper is to present some applications of the functional equation technique of dynamic programming^{1, 2} to the study of some multistage stochastic decision processes arising in statistical communication theory.

Our starting point is a paper by Kelly³ in which it is shown that the rate of transmission, as obtained by Shannon from considerations of coding of information,⁴ can be obtained from a certain multistage process with a suitable criterion function. In this paper we shall complete a result of Kelly's and considerably extend the scope of the investigation. Further results and proofs of the theorems stated below will be presented in a subsequent publication.

2. *An M-Signal Noisy Channel.*—Consider a noisy channel which is called upon to transmit any of M different signals, which we name 1, 2, . . . , M , in succession. Let

- p_{ij} = the conditional probability that the j signal has actually been sent whenever the i signal is received;
- q_i = the probability that the i signal is received at any particular time.

A gambler, upon receiving a particular signal, is required to place bets on what he believes the transmitted signal to have been. He is allowed to bet a quantity z_i that the i signal was transmitted, subject to the restrictions that $\sum_i z_i \leq x$, his initial capital, and $z_i \geq 0$. If he bets correctly, he then receives $r_i z_i$, otherwise nothing. This process continues for N stages, with a payoff at the end of each stage. Assuming that the transmitted signals are independent of each other and that the gambler wishes to maximize the expected value of a function $\phi(w)$ of the final total at the end of the process, the problem is to determine an optimal wagering policy.

presumably better known [r_i]

3. *Dynamic Programming Formulation.*—Let us define the sequence of functions

$$f_N(x) = \text{the expected value of } \phi(w) \text{ obtained using an optimal } N\text{-stage wagering policy, starting with a capital of } x. \quad (3.1)$$

for $N = 1, 2, \dots$, and $x \geq 0$.

Then the principle of optimality yields the recurrence relations

$$f_1(x) = \sum_{i=1}^M q_i \left\{ \text{Max}_{\sum z_i \leq x} \sum_{j=1}^M p_{ij} \phi \left(r_j z_j + x - \sum_{s=1}^M z_s \right) \right\},$$

$$f_N(x) = \sum_{i=1}^M q_i \left\{ \text{Max}_{\sum z_i \leq x} \sum_{j=1}^M p_{ij} f_{N-1} \left(r_j z_j + x - \sum_{s=1}^M z_s \right) \right\}, \quad N \geq 2. \quad (3.2)$$

If we specialize the function $\phi(w)$, we obtain a noteworthy result:

THEOREM 1. *In the case where $\phi(w) = \log w$, we have*

$$f_N(x) = \log x + Nk, \quad (3.3)$$

where

$$k = \sum_{i=1}^M q_i \left\{ \text{Max}_{\sum z_i \leq 1} \sum_{j=1}^M p_{ij} \log \left(r_j z_j + 1 - \sum_{s=1}^M z_s \right) \right\}. \quad (3.4)$$

The optimal policy is independent of the number of stages remaining, independent of the quantity of money available, and independent of the sequence $\{q_i\}$. It is determined by the maximization in equation (3.4).

A particular case of the above in which it is required that $\sum_i z_i = x$ yields the expression $-\sum_i p_i \log p_i$ of Shannon. This furnishes an interesting link with information theory. The foregoing result resolves a problem left open by Kelly.

There is an analogous result if $\phi(w) = w^a$, $a > 0$.

4. *Generalizations.*—The results of the preceding section may be generalized in many directions, in particular, to time-dependent processes and to the case where there is a continuum of types of signals. In both cases the functional equation technique is applicable, and the analogue of Theorem 1 holds.

Another interesting type of process to consider is that in which the p_{ij} are fixed, but unknown, constants. For an expository account of the problems encountered in this area we refer to Robbins,⁵ cf. also Bellman⁶ and Robbins.⁷

As we shall show in a subsequent paper, a number of these problems may be treated by means of the foregoing techniques.

5. *Correlated Signals.*—Let us now consider the case where the signals are not independent. Although a large variety of questions of this type may be formulated, the following discussion of a simple process will illustrate the general method that may be employed.

Assume that there are only two types of signals, say 0 and 1, that the probability of correct transmission at any stage depends upon whether or not the preceding signal was transmitted correctly, and that the gambler bets at each stage a certain quantity of his resources that the signal he receives was actually sent. Let

$$\begin{aligned} p_k &= \text{the probability of correct transmission of the } k\text{th signal if the} \\ &\quad (k-1)\text{st signal was transmitted correctly.} \\ r_k &= \text{the probability of correct transmission of the } k\text{th signal if the} \\ &\quad (k-1)\text{st signal was transmitted incorrectly.} \end{aligned} \quad (5.1)$$

Define the sequence of functions

$$f_k(x) = \text{the expected value of the logarithm of the final capital obtained from the remaining } k \text{ stages of the original } N\text{-stage}$$

process, when one has a capital of x and the information that the $N - k$ th signal was transmitted correctly, and uses an optimal policy.

$g_k(x)$ = the corresponding expected value in the case where the $N - k$ th signal was transmitted incorrectly. (5.2)

Then

$$f_k(x) = \text{Max}_{0 \leq y \leq x} [p_{N-k+1} f_{k-1}(x+y) + (1 - p_{N-k+1}) g_{k-1}(x-y)],$$

$$g_k(x) = \text{Max}_{0 \leq y \leq x} [r_{N-k+1} f_{k-1}(x+y) + (1 - r_{N-k+1}) g_{k-1}(x-y)]. \quad (5.3)$$

It follows inductively that

$$f_k(x) = \log x + a_k, \quad g_k(x) = \log x + b_k, \quad (5.4)$$

where a_k and b_k are independent of x . The recurrence relations for the sequence $\{a_k, b_k\}$ are readily obtained from equation (5.3).

¹ R. Bellman, *Dynamic Programming* (Princeton, N.J.: Princeton University Press, 1957).

² R. Bellman, "The Theory of Dynamic Programming," *Bull. Am. Math. Soc.*, **60**, 503-515, 1954.

³ J. Kelly, "A New Interpretation of Information Rate," *Bell System Tech. J.*, **35**, 917-926, 1956.

⁴ C. Shannon, "A Mathematical Theory of Communication," *Bell System Tech. J.*, **27**, 379-423, 623-656, 1948.

⁵ H. Robbins, "Some Aspects of the Sequential Design of Experiments," *Bull. Am. Math. Soc.*, **58**, 527-536, 1952.

⁶ R. Bellman, "A Problem in the Sequential Design of Experiments," *Sankhya*, **16**, 221-229, 1956.

⁷ H. Robbins, "A Sequential Decision Problem with a Finite Memory," these PROCEEDINGS, **42**, 920-923, 1956.

SOLUTION OF THE BURNSIDE PROBLEM FOR EXPONENT 6*

BY MARSHALL HALL, JR.

DEPARTMENT OF MATHEMATICS, OHIO STATE UNIVERSITY

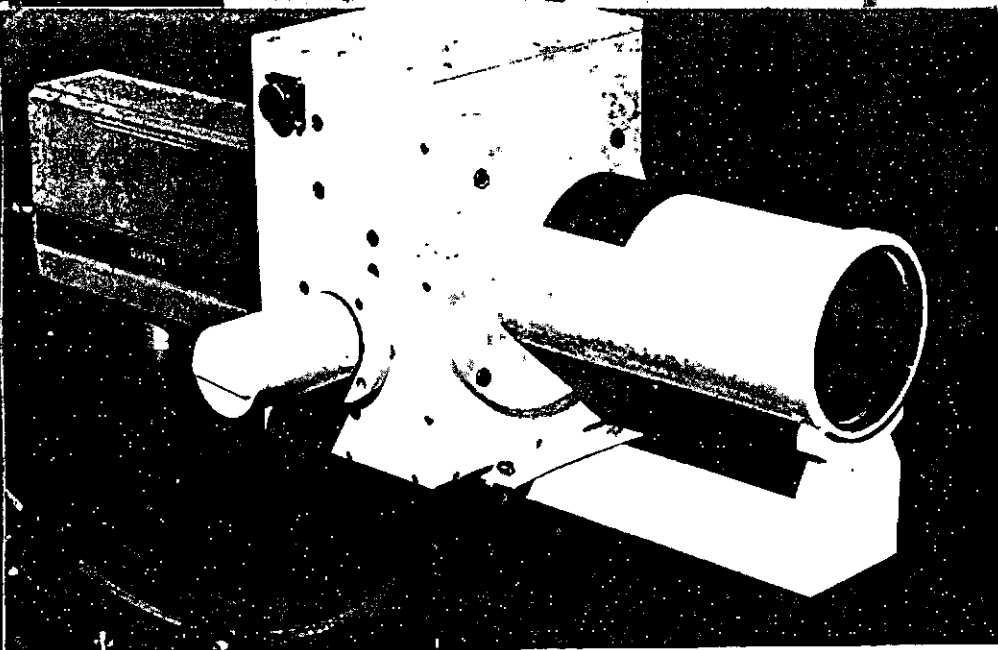
Communicated by Saunders Mac Lane, June 5, 1957

The restricted Burnside problem for exponent 6 was solved by Philip Hall and Graham Higman.¹ They even found the order of the largest finite group of exponent 6 generated by k elements, this order being

$$2^a 3^b + \binom{a}{2} + \binom{a}{3},$$

where $a = 1 + (k-1)3^k + \binom{k}{2} + \binom{k}{3}$ and $b = 1 + (k-1)2^k$. A proof is sketched here that a finitely generated group of exponent 6 is necessarily finite solving the Burnside problem for exponent 6. The proof will be published in detail elsewhere.

It has been shown by Levi and van der Waerden² that there is a group $B(3, k)$ generated by k elements and of exponent 3 whose order is 3^k , $K = k + \binom{k}{2} + \binom{k}{3}$,



QUESTAR 20-40 . . .

the tracking Questar

The Questar high-resolution optical system now provides a wide variety of instruments for accurate image reproduction in critical film and TV auto-tracking applications. For example, the Questar 20-40 operates at two par-focal, remotely controlled focal lengths: 20 inches at F5.7 and 40 inches at F11.4.

This ruggedized unit is equipped with solenoid-operated over-exposure shutter and four-port filter wheel assembly, likewise remotely controlled. Provision is made, also, for the manual insertion of additional filters. The instrument is available in aluminum, stainless steel, or Invar steel, depending on the degree of temperature compensation that is necessary; and the optical components, too, can be furnished in various temperature-stable materials: Pyrex®, quartz or CerVit®.

When you have been building Maksutov telescopes for 25 years, as we have here at Questar, you have learned a great deal about the ways in which these superlative optics can be used in special situations. The capabilities of this versatile system, on which we have built our world-wide reputation, have unfolded year by year as we have pursued the engineering challenges presented to us.

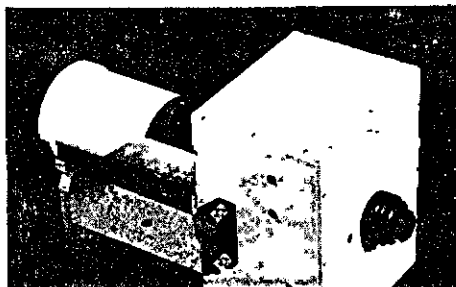
Ruggedization and temperature stability were early demands made of this miniaturized instrument in applications where space was the third limiting factor. And so, in the beginning, our off-the-shelf Questars were modified by substituting materials to meet the requirements. These, in turn, became our stock items, so that now when a sudden need arises for a telescope to put aboard an orbiting satellite, or to use with stabilizing equipment mounted in a helicopter, or to observe hot materials in a laboratory, there is usually a modified Questar instantly available.

Demands for more highly specialized equipment, such as the tracking instrument above, have kept us in the forefront of research and development in this field, where optical modification, also, is often a part of the job. If you have a special problem we may have solved it. Why not call us?

® Registered Trademark of Corning Glass Works
 ® Registered Trademark of Owens-Illinois, Inc.

SEND FOR OUR LITERATURE. OUR 1977 BOOKLET ABOUT QUESTAR, THE WORLD'S FINEST, MOST VERSATILE TELESCOPE CONTAINS BEAUTIFUL PHOTOGRAPHS BY QUESTAR OWNERS. \$1 COVERS MAILING COSTS ON THIS CONTINENT; BY AIR, TO S. AMERICA \$3; EUROPE AND N. AFRICA, \$3.50; ELSEWHERE \$4.

QUESTAR



SITY HERBARIA. Siri von Reis. Allschul. Harvard University Press, 1973.
 EMPIRICAL AZTEC MEDICINE. Bernard Ortiz de Montellano in *Science*, Vol. 188, No. 4185, pages 215-220; April 18, 1975.

RAT SOCIETIES

RATS. S. A. Barnett in *Scientific American*, Vol. 216, No. 1, pages 78-85; January, 1967.
 AGGRESSION AND SOCIAL EXPERIENCE IN DOMESTICATED RATS. David Luciano, and Richard Lore in *Journal of Comparative and Physiological Psychology*, Vol. 88, No. 2, pages 917-923; February, 1975.
 ROLE OF RESIDUAL OLFACTORY CUES IN THE DETERMINATION OF FEEDING SITE SELECTION AND EXPLORATION PATTERNS OF DOMESTIC RATS. Bennett G. Galef, Jr., and Linda Heiber in *Journal of Comparative and Physiological Psychology*, Vol. 90, No. 8, pages 727-739; August, 1976.

STEIN'S PARADOX IN STATISTICS

AN EMPIRICAL BAYES APPROACH TO STATISTICS. Herbert Robbins in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability—Vol. 1: Contributions to the Theory of Statistics*, edited by Jerzy Neymann. University of California Press, 1956.
 STEIN'S ESTIMATION RULE AND ITS COMPETITORS—AN EMPIRICAL BAYES APPROACH. Bradley Efron and Carl Morris in *Journal of the American Statistical Association*, Vol. 68, No. 341, pages 117-130; March, 1973.
 EMPIRICAL BAYES METHODS APPLIED TO ESTIMATING FIRE ALARM PROBABILITIES. Grace M. Carter and John E. Rolph in *Journal of the American Statistical Association*, Vol. 69, No. 348, pages 880-885; December, 1974.
 BIASED VERSUS UNBIASED ESTIMATION. Bradley Efron in *Advances in Mathematics*, Vol. 16, No. 3, pages 259-277; June, 1975.
 DATA ANALYSIS USING STEIN'S ESTIMATOR AND ITS GENERALIZATIONS. Bradley Efron and Carl Morris in *Journal of the American Statistical Association*, Vol. 70, No. 350, pages 311-319; June, 1975.

MATHEMATICAL GAMES

THE METHOD OF MATHEMATICAL INDUCTION. I. S. Sominskii, translated from the Russian by Hallina Moss. Blaisdell Publishing Co., 1961.
 MATHEMATICAL INDUCTION. Albert A. Blank in *Enrichment Mathematics for High School*. National Council of Teachers of Mathematics, 1963.

Sci AMER
 MAY 1977

Stein's Paradox in Statistics

The best guess about the future is usually obtained by computing the average of past events. Stein's paradox defines circumstances in which there are estimators better than the arithmetic average

by Bradley Efron and Carl Morris

*Computer in Language Methods in Statistics
Discussion "Bootstrap": May 83 p116*

Sometimes a mathematical result is strikingly contrary to generally held belief even though an obviously valid proof is given. Charles Stein of Stanford University discovered such paradox in statistics in 1955. His result undermined a century and a half of work on estimation theory, going back to Karl Friedrich Gauss and Adrien Marie Legendre. After a long period of resistance to Stein's ideas, punctuated by frequent and sometimes angry debate, the sense of paradox has diminished and Stein's ideas are being incorporated into applied and theoretical statistics.

Stein's paradox concerns the use of observed averages to estimate unobservable quantities. Averaging is the second most basic process in statistics, the first being the simple act of counting. A baseball player who gets seven hits in 20 official times at bat is said to have a batting average of .350. In computing this statistic we are forming an estimate of the player's true batting ability in terms of his observed average rate of success. Asked how well the player will do in his next 100 times at bat, we would probably predict 35 more hits. In traditional statistical theory it can be proved that no other estimation rule is uniformly better than the observed average.

The paradoxical element in Stein's result is that it sometimes contradicts this elementary law of statistical theory. If we have three or more baseball players, and if we are interested in predicting future batting averages for each of them, then there is a procedure that is better than simply extrapolating from the three separate averages. Here "better" has a strong meaning. The statistician who employs Stein's method can expect to predict the future averages more accurately no matter what the true batting abilities of the players may be.

Baseball is a sport with a large and carefully compiled body of statistics, which supplies convenient material for illustrating the workings of Stein's method. As our primary data we shall consider the batting averages of 18 ma-

ior-league players as they were recorded after their first 45 times at bat in the 1970 season. These were all the players who happened to have batted exactly 45 times the day the data were tabulated. A batting average is defined, of course, simply as the number of hits divided by the number of times at bat; it is always a number between 0 and 1. We shall denote each such average by the letter y .

The first step in applying Stein's method is to determine the average of the averages. Obviously this grand average, which we give the symbol \bar{y} , must also lie between 0 and 1. The essential process in Stein's method is the "shrinking" of all the individual averages toward this grand average. If a player's hitting record is better than the grand average, then it must be reduced; if he is not hitting as well as the grand average, then his hitting record must be increased. The resulting shrunken value for each player we designate z . This value is the James-Stein estimator of that player's batting ability, named for Stein and W. James, who together proposed a particularly simple version of the method in 1961. Stein's paradox is simply that the z values, the James-Stein estimators, give better estimates of true batting ability than the individual batting averages.

The James-Stein estimator for each player is found through the following equation: $z = \bar{y} + c(y - \bar{y})$. The quantity $(y - \bar{y})$ is the amount by which the player's batting average differs from the grand average. The equation thus states that the James-Stein estimator z differs from the grand average by this same quantity $(y - \bar{y})$ multiplied by a constant, c . The constant c is the "shrinking factor." If it were equal to 1, then the equation would state that the James-Stein estimator for a given player is identical with that player's batting average; in other words, y equals z . Stein's theorem states that the shrinking factor is always less than 1. Its actual value is determined by the collection of all the observed averages.

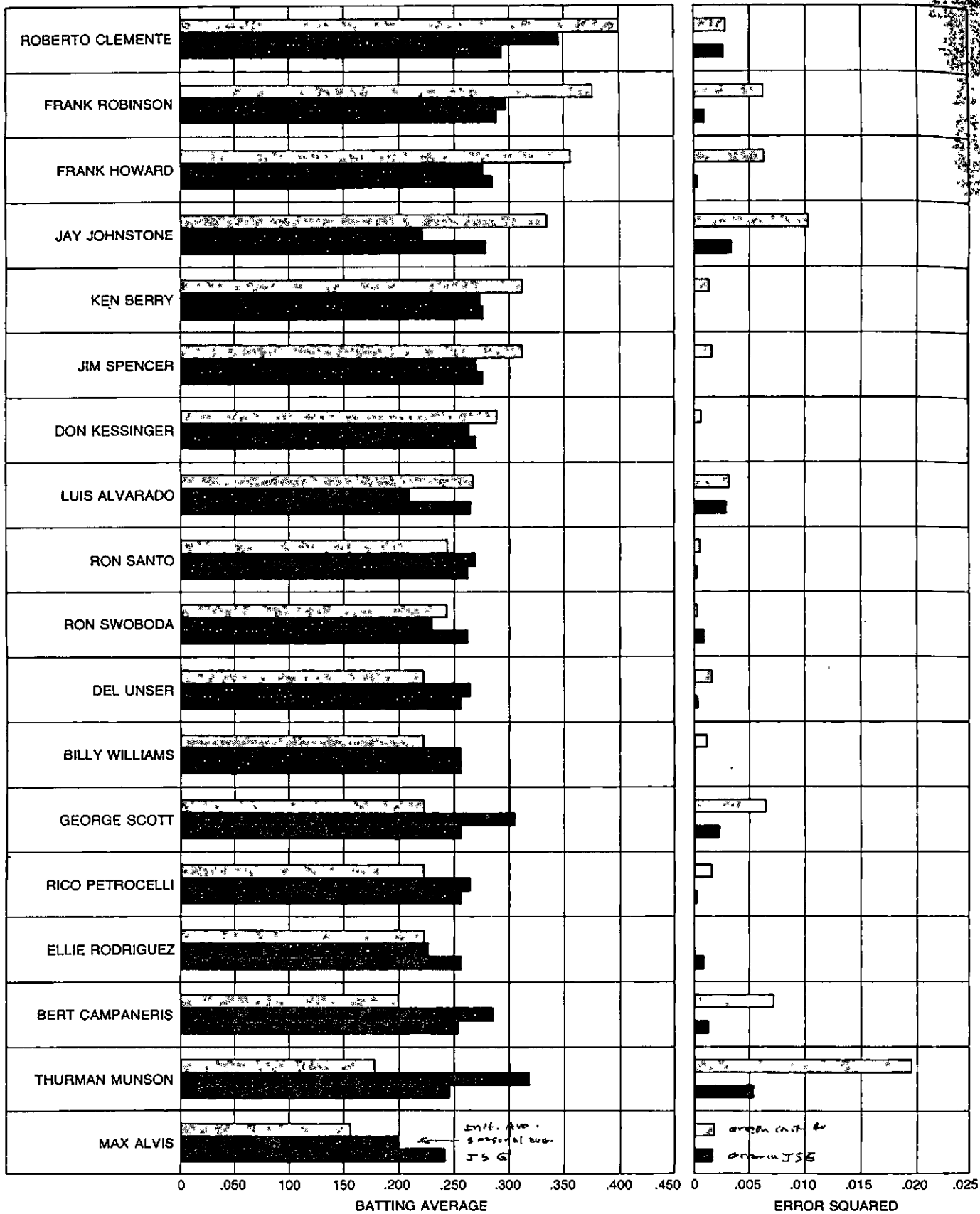
In the case of the baseball data, the grand-average \bar{y} is .265 and the shrinking




factor c is .212. Substituting these values in the equation, we find that for each player z equals $.265 + .212(y - .265)$. Because c is about .2, each average will shrink about 80 percent of the distance to the grand average, and the total spread of the averages will be reduced about 80 percent.

As an example consider the late Roberto Clemente, who was the leading batter in the major leagues when our statistics were compiled. For Clemente y is equal to .400, and z can be determined by evaluating the expression $z = .265 + .212(.400 - .265)$. The result is .294. In other words, Stein's theorem states that Clemente's true batting ability is best estimated not by .400 but lies closer to .294. Thurman Munson, in a batting slump early in the 1970 season, had an average of only .178. Substituting this value in the equation, we find that his estimated batting ability is substantially increased: the James-Stein estimator for Munson is .247.

Which set of values, y or z , is the better indicator of batting ability for the 18 players in our example? In order to answer that question in a precise way one would have to know the "true batting ability" of each player. This true average we shall designate with θ (the Greek letter theta). Actually it is an unknowable quantity, an abstraction representing the probability that a player will get a hit on any given time at bat. Although θ is unobservable, we have a good approximation to it: the subsequent performance of the batters. It is sufficient to consider just the remainder of the 1970 season, which includes about nine times as much data as the preliminary averages were based on. The expected statistical error in such a sample is small enough for us to neglect it and proceed as if the seasonal average were the "true batting ability" θ of a player. That is one reason for choosing batting averages for this example. In most problems the true value of θ cannot be determined.

One method of evaluating the two es-



 INITIAL AVERAGE
 SEASON AVERAGE
 JAMES-STEIN ESTIMATOR

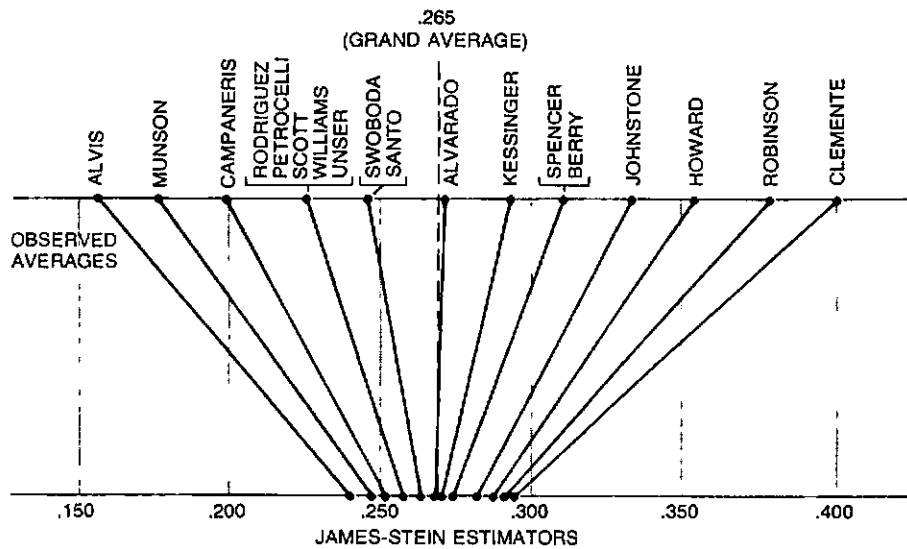
BATTING ABILITIES of 18 major-league baseball players are estimated more accurately by the method of Charles Stein and W. James than they are by the individual batting averages. The averages employed as estimators are those calculated after each player had had 45 times at bat in the 1970 season. The true batting ability of a player is an unobservable quantity, but it is closely approximated by his long-term average performance. Here the true ability is represented by the batting average maintained during the remainder of the 1970 season. For 16 of the players the initial average is inferior to another number, the James-Stein estimator, as a predictor of batting ability. The James-Stein estimators, considered as a group, also have the smaller total squared error.

imates is by simply counting their successes and failures. For 16 of the 18 players the James-Stein estimator z is closer than the observed average y to the "true," or seasonal, average θ . A more quantitative way of comparing the two techniques is through the total squared error of estimation. This is measured by first determining the actual error of each prediction, given by $(\theta - y)$ and $(\theta - z)$, for each player. Each of these quantities is then squared and the squared values are added up. The observed averages y have a total squared error of .077, whereas the squared error of the James-Stein estimators is only .022. By this comparison, then, Stein's method is 3.5 times as accurate. It can be shown that for the data given 3.5 is close to the expected ratio of the total squared errors of the two methods. We have not just been lucky.

Suppose a statistician makes a random sampling of automobiles in Chicago and finds that of the first 45 recorded nine are foreign-made and the remaining 36 are domestic. We want to estimate the true proportion of imported cars in Chicago, a quantity represented by another unobservable θ . The observed average of $9/45 = .200$ is one estimate. Another can be obtained by simply lumping this problem together with that of the 18 baseball players. Substituting the value .200 in the equation used in that problem gives a James-Stein estimator of .251 for the imported-car ratio. (Actually the addition of a 19th value changes the grand average \bar{y} and also slightly alters the shrinking factor c . The changes are small, however; the amended value of z is .249.)

In this case intuition argues strongly that the observed average and not the James-Stein estimator must be the better predictor. Indeed, the entire procedure seems silly: what could batting averages have to do with imported cars? It is here that the paradoxical nature of Stein's theorem is most uncomfortably apparent. The theorem applies as well to the 19 problems as it did to the original 18. There is nothing in the statement of the theorem that requires the component problems to have some sensible relation to one another.

The same disconcerting indifference to common sense can be demonstrated in another way. What does Clemente's 400 observed average have to do with Max Alvis, who was poorest in batting among the 18 players? If Alvis had had an early-season hitting streak, batting .444 instead of his actual .156, the James-Stein estimator for Clemente's average would have been increased from .294 to .325. Why should Alvis' success or lack of it have any influence on our estimate of Clemente's ability? (They were not even in the same league.)



JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.

It is questions of this kind that have been raised by critics of Stein's method. In order to reply to them it will be necessary to describe the method rather more carefully.

Taking an average is an easy and familiar process that seems to need no justification. Actually it is not obvious why the average is so often useful in estimating the true center of gravity of a random process. The explanation lies in the distribution that the values of the random variable tend to assume.

The distribution most common in scientific work is the "normal" distribution, described by a bell-shaped curve; it was first investigated in depth by Gauss and is sometimes called the Gaussian distribution. It is constructed by assuming that the random variable can take on any value along some axis; the probability that it falls within any given interval is then made equal to the area under the same interval of the bell-shaped curve. The curve is completely specified by two parameters: the mean, θ , which lies at the peak of the curve, and the standard deviation, which measures how closely the values are distributed around the mean. It is customary to assign the standard deviation the symbol σ (sigma). The larger the standard deviation is, the more widely dispersed the data are.

In probability theory a known mean and standard deviation are employed to predict future behavior. A problem in statistics proceeds in the opposite direction: from observed data the statistician must infer the mean θ and the standard deviation σ .

Suppose, for example, the measurement of some random variable x yields

the five successive values 10.0, 9.4, 10.3, 8.6 and 9.7. Suppose further the values are known to be part of a normal distribution with a standard deviation of 1. What is the value of the true mean θ ? In principle the mean could have any value, but some values are more likely than others. A mean of 6.5, for example, would require that all five values be under the extreme tail of the curve and that none be found near the center. Gauss showed that among all possible choices for the mean, the average \bar{x} of the observed data (which in this case has a value of 9.6) maximizes the probability of obtaining the data actually seen. In this sense the average is the most likely estimate of the mean; in fact, Gauss constructed the normal distribution just so that it would have this property.

There is a further justification, also pointed out by Gauss, for choosing the average as the best estimator of the unobservable mean θ . Gauss noted that the average of the data is an "unbiased" estimator of the mean, in the sense that it favors no selected value of θ . To be more precise, the average is unbiased because the expected value of \bar{x} equals the true θ no matter what θ may be. There are infinitely many unbiased estimators of θ , none of which estimates θ perfectly. Gauss showed that the expected squared error of estimation for the average \bar{x} is lower than that for any other linear, unbiased function of the observations. In the 1940's it was demonstrated that no other unbiased function of the data, whether it is linear or nonlinear, can estimate θ more accurately than the average, in terms of expected squared error. An essential contribution to that proof had been made in the 1920's by

R. A. Fisher, who showed that all the information about θ that can possibly be found in the data is contained in the average \bar{x} .

In the 1930's a mathematically more rigorous approach to statistical inference was undertaken by Jerzy Neyman, Egon S. Pearson and Abraham Wald; the ideas they developed are part of what is now known as statistical decision theory. They discarded the requirement of unbiased estimation and examined all functions of the data that could serve as estimators of the unknown mean θ . These estimators were compared through a risk function, defined as the expected value of the squared error for every possible value of θ .

Consider three competing estimators: the average of the data, \bar{x} ; half that average, $\bar{x}/2$, and the median of the data, or middle value. For both the average and the median the risk function is constant; that is merely another way of saying that their expected squared error in predicting the mean θ is the same no matter what the value of θ really is. Of the two constant risk functions, the one for the average \bar{x} is uniformly smaller by a factor of about two-thirds; clearly the average is the preferred estimator. In the language of decision theory the median is said to be "inadmissible" as an estimator of θ , since there is another estimator that has a smaller risk (expected squared

error) no matter what θ is. (It should be mentioned, however, that when the data have a distribution other than the normal one, it is possible for the order of preference to be reversed.)

For the estimator $\bar{x}/2$, which is biased toward the value $\theta = 0$, the risk function is not constant; this estimator is accurate if θ happens to be close to zero, but the expected squared error increases rapidly as the true mean departs from zero. The risk function describes a parabola, with the minimum point at $\theta = 0$; if the mean does happen to be zero, then the risk function for $\bar{x}/2$ is four times smaller than that for the average itself. At large values of the mean, however, the average \bar{x} regains its superiority. With other estimators we can poke down the risk function below that of the average at any point we wish to, but it always pops up again somewhere else.

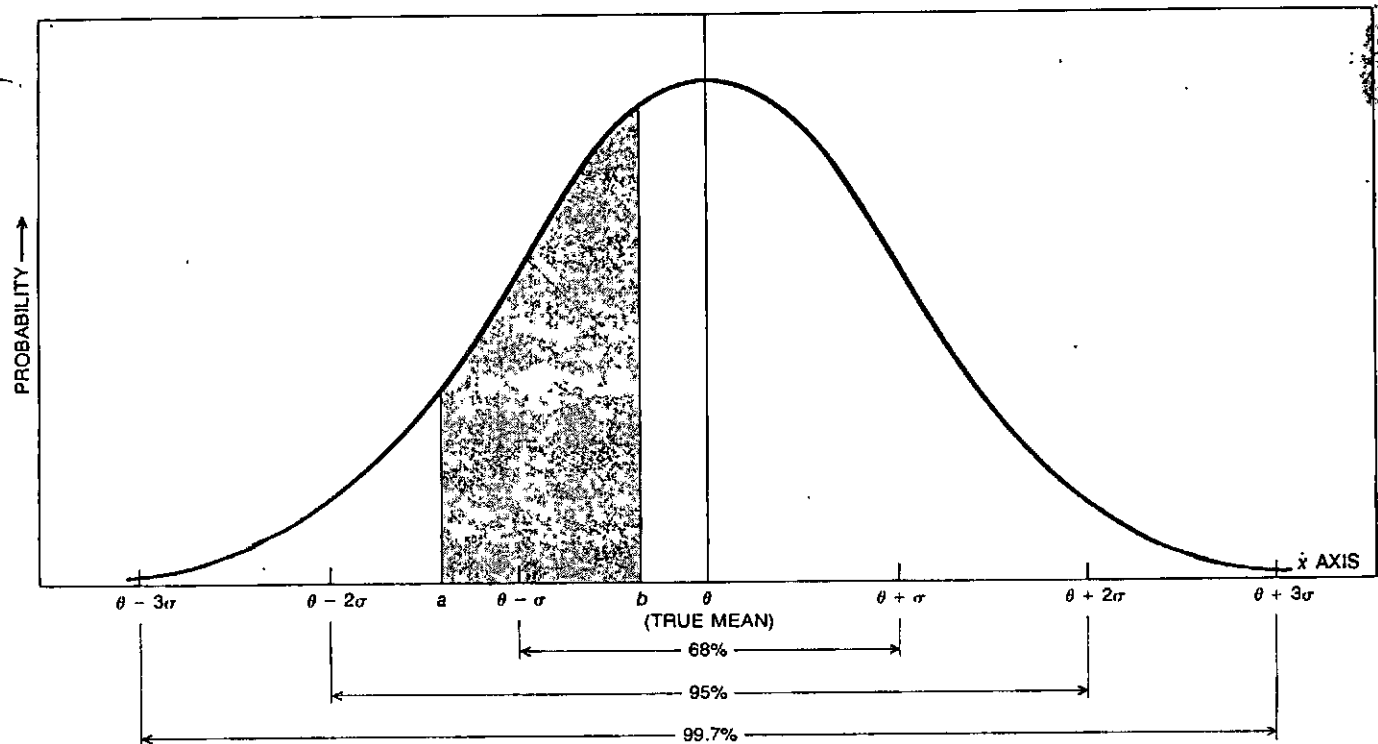
There remains the possibility that some other estimator has a risk that is uniformly lower than that of the average. In 1950 Colin R. Blyth, Erich L. Lehmann and Joseph L. Hodges, Jr., proved that no such estimator exists. In other words, the average \bar{x} is admissible, at least when it is applied to one set of observations for the purpose of estimating one unknown mean.

Stein's theorem is concerned with the estimation of several unknown means. No relation between the means need be assumed; they can be batting abilities or

proportions of imported cars. On the other hand, the means are assumed to be independent of one another. In evaluating estimators for these means it is once again convenient to employ a risk function defined as the sum of the expected values of the squared errors of estimation for all the individual means.

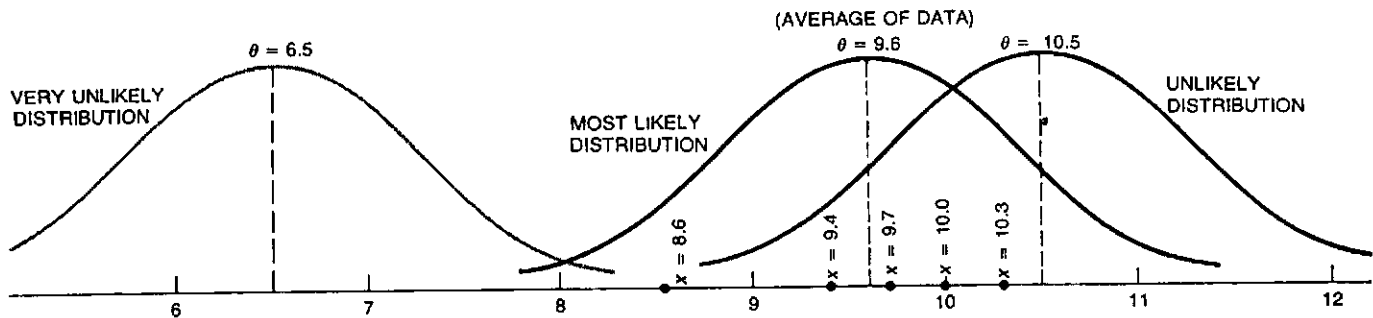
The obvious first choice of an estimator for each of several means is the average of the data related to that mean. The entire historical development of statistical theory from Gauss through decision theory argues that the average is an admissible estimator as long as there is just one mean, θ , to be estimated. Stein showed in 1955 that the average is also admissible for estimating two means. Stein's paradox is simply his proof that when the number of means exceeds two, estimating each of them by its own average is an inadmissible procedure. No matter what the values of the true means, there are estimation rules with smaller total risk.

In 1955 Stein was able to prove this proposition only in those cases where the number of means, a quantity we shall designate k , was very large. Stein's 1961 paper written in collaboration with James extended the result to all values of k greater than 2; moreover, it did so in a constructive manner. Stein and James not only showed that estimators must exist that are everywhere superior to the



NORMAL DISTRIBUTION of a random variable around the mean value of that variable provides the fundamental justification for estimation by averaging. The distribution is defined by two parameters, the mean, θ , which locates the central peak of the distribution, and the standard deviation, σ , which measures how widely scattered the

data points are. It is assumed in defining the distribution that the variable x can take on any value on the x axis. The most likely value of x , by definition, the mean θ . The probability that x lies within any given interval on the axis, such as that between the points a and b , is equal to the area under the bell-shaped curve between those points.



PROBLEM IN STATISTICS is to deduce from a set of data the true mean and standard deviation of the distribution. Even when it is known that the distribution is a normal one and that the standard deviation is 1, the mean could in principle have any value. Some values, however, are more likely than others. For example, the five data

points (x) given here could be described by a normal distribution with a mean of 6.5 only if all five points were more than two standard deviations above the mean. It can be shown that the data are most likely to be generated by a distribution with a mean equal to the observed average of the data, denoted \bar{x} . In this case the average is equal to 9.6.

averages; they were also able to provide an example of such an estimator.

The James-Stein estimator has already been defined in our investigation of batting averages. It is given by the equation $z = \bar{y} + c(y - \bar{y})$, where y is the average of a single set of data, \bar{y} is the grand average of averages and c is a "shrinking factor." There are several other expressions for the James-Stein estimator, but they differ mainly in detail. All of them have in common the shrinking factor c ; it is the definitive characteristic of the James-Stein estimator.

In the baseball problem c was treated as if it were a constant. Actually it is determined by the observed averages and therefore is not a constant. The shrinking factor is given by the equation

$$c = 1 - \frac{(k-3)\sigma^2}{\sum(y - \bar{y})^2}$$

Here k is again the number of unknown means, σ^2 is the square of the standard deviation and $\sum(y - \bar{y})^2$ is the sum of the squared deviations of the individual averages y from the grand average \bar{y} .

Let us briefly explore the meaning of this rather forbidding equation. With k and σ^2 fixed, we find that the shrinking factor c becomes smaller (and the predicted means are more severely affected by it) as the expression $\sum(y - \bar{y})^2$ gets smaller. On the other hand, c increases, approaching unity, and the shrinking is less drastic as the expression $\sum(y - \bar{y})^2$ increases.

What do these equations mean in terms of the behavior of the estimator? In effect the James-Stein procedure makes a preliminary guess that all the unobservable means are near the grand average \bar{y} . If the data support that guess in the sense that the observed averages are themselves not too far from \bar{y} , then the estimates are all shrunk further toward the grand average. If the guess is contradicted, then not much shrinking is done. These adjustments to the shrinking factor are accomplished through the

effect the distribution of averages around the grand average \bar{y} has on the equation that determines c . The number of means being estimated also influences the shrinking factor, through the term $(k-3)$ appearing in this same equation. If there are many means, the equation allows the shrinking to be more drastic, since it is then less likely that variations observed represent mere random fluctuations.

With c calculated in this manner, the risk function for the James-Stein estimator is less than that for the sample averages no matter what the true values of the means θ happen to be. The reduction of risk can be substantial, particularly when the number of means is larger than five or six. The risk function is not constant for all values of the true mean θ , as it is for the observed averages. The risk of the James-Stein estimator is smallest when all the true means are the same. As the true means depart from one another the risk of the estimator increases, approaching that of the observed averages but never quite equaling it. The James-Stein estimator does substantially better than the averages only if the true means lie near each other, so that the initial guess involved in the technique is confirmed. What is surprising is that the estimator does at least marginally better no matter what the true means are.

The expression for the James-Stein estimator that we have employed refers all observed averages to the grand average \bar{y} . This procedure is not the only one possible; other expressions for the estimator dispense with \bar{y} entirely. What cannot be avoided is the introduction of some more or less arbitrary initial guess or point of origin for the estimator. The observed averages, it will be noted, do not depend on a choice of origin. Before Stein discovered his method it was felt that such "invariant" estimators must be preferable to those whose predictions change with each choice of an origin. The theory of invariance, to which Stein had been a principal contributor, was

badly shaken by the James-Stein counterexample. From the standpoint of mathematics this is the most unsettling aspect of Stein's theorem. Indeed, the paradox was not discovered earlier largely because of a strong prejudice that the estimation problem, being stated without reference to any particular origin, should be solved in a similar way.

Applications of Stein's method tend to involve large sets of data with many unknown parameters. Some of the difficulties of such problems, as well as the practical potential of the method itself, can be illustrated by an example: an analysis of the distribution of the disease toxoplasmosis in the Central American country of El Salvador.

Toxoplasmosis is a disease of the blood that is endemic in much of Central America and in other regions of the Tropics. In El Salvador roughly 5,000 people drawn in varying numbers from 36 cities were tested for toxoplasmosis. The observed rate of incidence for each city can conveniently be expressed by comparison with the national rate (that is, with the grand average \bar{y}). A measured rate of .050, for example, denotes a city with an incidence of the disease 5 percent higher than the national average. The measured rates have an approximately normal distribution. The standard deviations of these distributions are known, but they differ from city to city, depending inversely on how large a sample population was tested in that city. It is the task of the statistician to estimate the true mean θ of the distribution for each city from the measured incidence y .

In this case the appropriate form of the James-Stein estimator is $z = cy$. The simplification, which was introduced by us, is made possible by the chosen manner of expressing the observations y . They are defined in such a way that the grand average \bar{y} is zero, and terms containing \bar{y} therefore drop out of the equation. On the other hand, the estimation

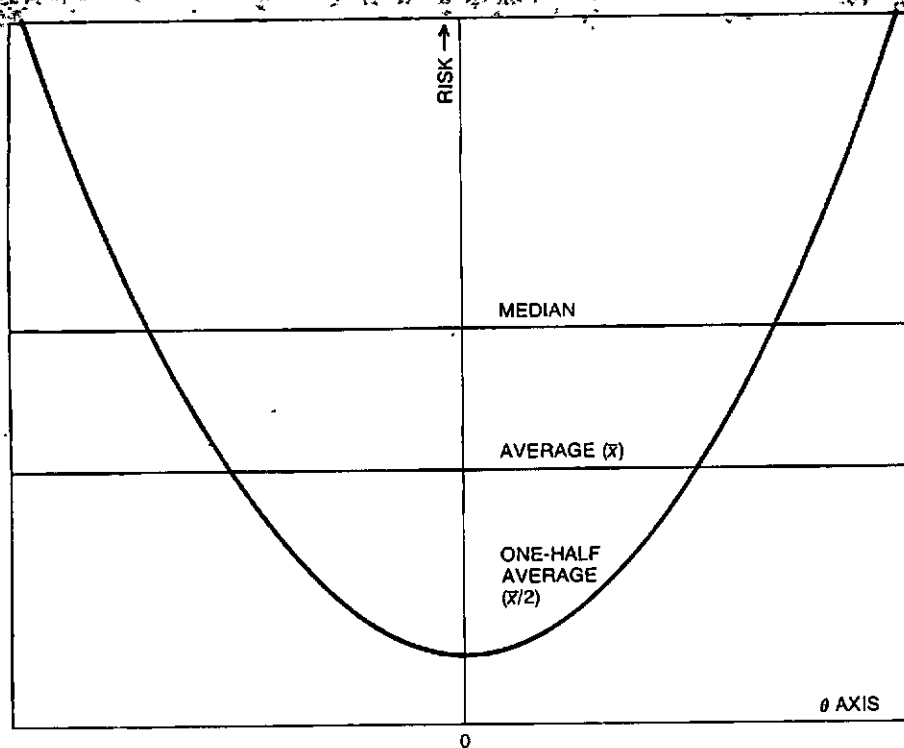
A BIG Q: What's data of σ^2 ?

Best Guess for $\sigma^2 = \frac{1}{k} \cdot \frac{1}{n} \cdot \sum \sum (y_i - \bar{y}_i)^2$

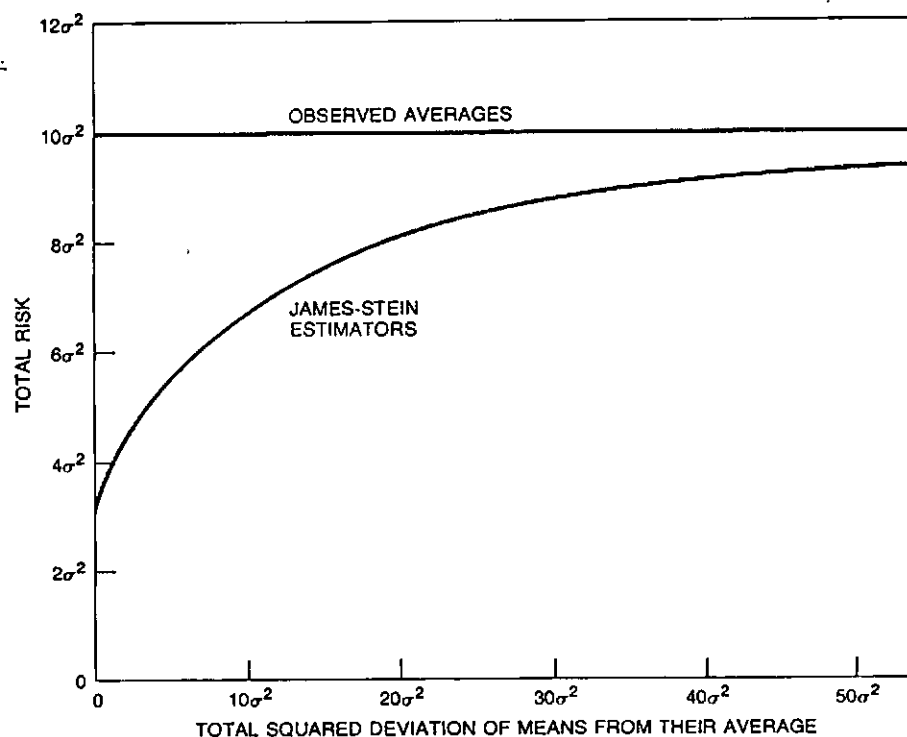
$y_i = 0.01$; The mean

variance of y_i is σ^2

What's our best guess for σ^2 ?



VARIOUS ESTIMATORS of a single true mean, θ , can be evaluated by way of a risk function. The risk is defined as the expected value of the squared error of estimation, considered as a function of the mean θ . The average of the data, \bar{x} , is an estimator with a constant risk function: no matter what the true mean is, the expected value of the squared error is the same. The median, or middle value, of the data also has constant risk, but it is everywhere greater (by a factor of 1.57) than the risk of the average. Half the average ($\bar{x}/2$) is an estimator whose risk depends on the actual value of the mean; the risk is smallest when the mean is near zero and increases rapidly when the mean departs from zero. For the estimation of a single mean there is no estimator with a risk function that is everywhere less than the risk function of the average \bar{x} .



TOTAL RISK FUNCTION for the James-Stein estimators is everywhere less than that for the individual observed averages, as long as the number of means being estimated is greater than two. In this case there are 10 unknown means. The risk is smallest when all the means are clustered at a single point. As the means depart from one another the risk of the James-Stein estimators increases, approaching that of the observed averages but never quite reaching it.

procedure is now complicated by the fact that the shrinking factor c is different for each city, varying inversely with the standard deviation of y for that city. This dependence of the shrinking factor on the standard deviation has a simple intuitive rationale. A large standard deviation implies a high degree of randomness or uncertainty in a measurement. If the measured incidence is unusually large, it can therefore be attributed more reasonably to random fluctuation within the normal distribution than to a genuinely large value of the true mean θ . It is thus proper to reduce this value drastically, that is, to apply a small shrinking factor.

The same argument can be made even more forcefully by returning for a moment to baseball. Frank O'Connor pitched for Philadelphia in the 1893 season. He batted twice in his major-league career, hitting successfully both times. His observed batting average is hence 1.000. The James-Stein rule for the 18 players considered above estimates O'Connor's true batting ability to be $.265 + .212(1.000 - .265) = .421$ (ignoring the effect of the new data on the grand average and on the shrinking factor). This is a silly estimate, although not as silly as 1.000. A perfect average after two times at bat is not at all inconsistent with a true value in the range from .242 to .294 that is estimated for the other players. The shrinking constant c applied to O'Connor's average should be severer in order to compensate for the smaller amount of data available for him.

For the El Salvador observations most of the shrinking factors are quite gentle, between .6 and .9, but a few are in the range from .1 to .3. Which set of numbers should we prefer, the James-Stein estimators or the measured rates of incidence? That depends largely on what we want to use the numbers for.

If the Minister of Health for El Salvador intends to build local hospitals for people suffering from toxoplasmosis, the James-Stein estimators probably offer the more reliable guidance. The reason is that the expected value of the total squared error is smaller for the James-Stein estimators; in fact, it is smaller by a factor of about three. The important point in this calculation is that the expected error is added up for all the cities. Any particular hospital might be the wrong size or in the wrong place, but the sum of all such mismatches would be smaller for the James-Stein estimators than for the observed rates.

The James-Stein estimators are also likely to be preferable for determining the ordering of the true means. In this regard it is notable that the city with the highest apparent incidence (according to the measured rates y) is ranked 12th according to the James-Stein estimators.

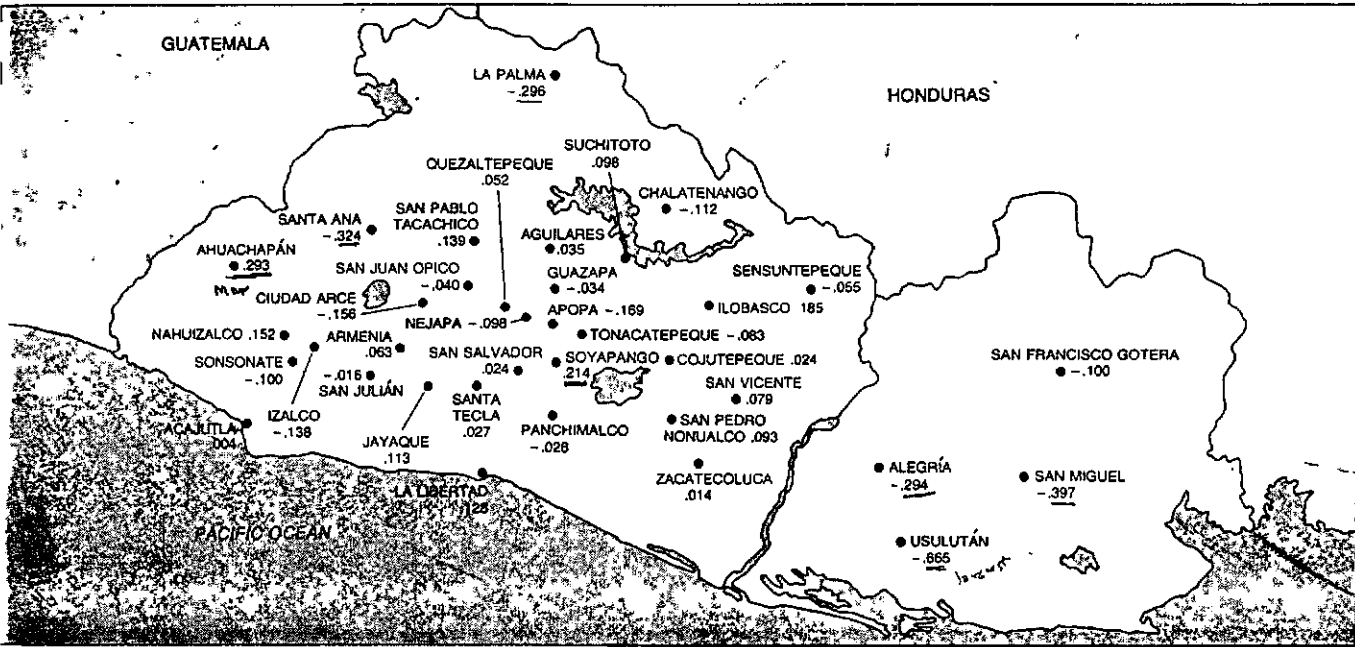
The estimate is drastically reduced, because the sample was very small in that city. This information might be useful if there were funds for only one hospital.

Suppose an epidemiologist wants to investigate the correlation of the true incidence in each city with attributes such

as rainfall, temperature, elevation or population? Once again the James-Stein estimators are preferred; a rough calculation shows that they would give a closer approximation in about 70 percent of the cases.

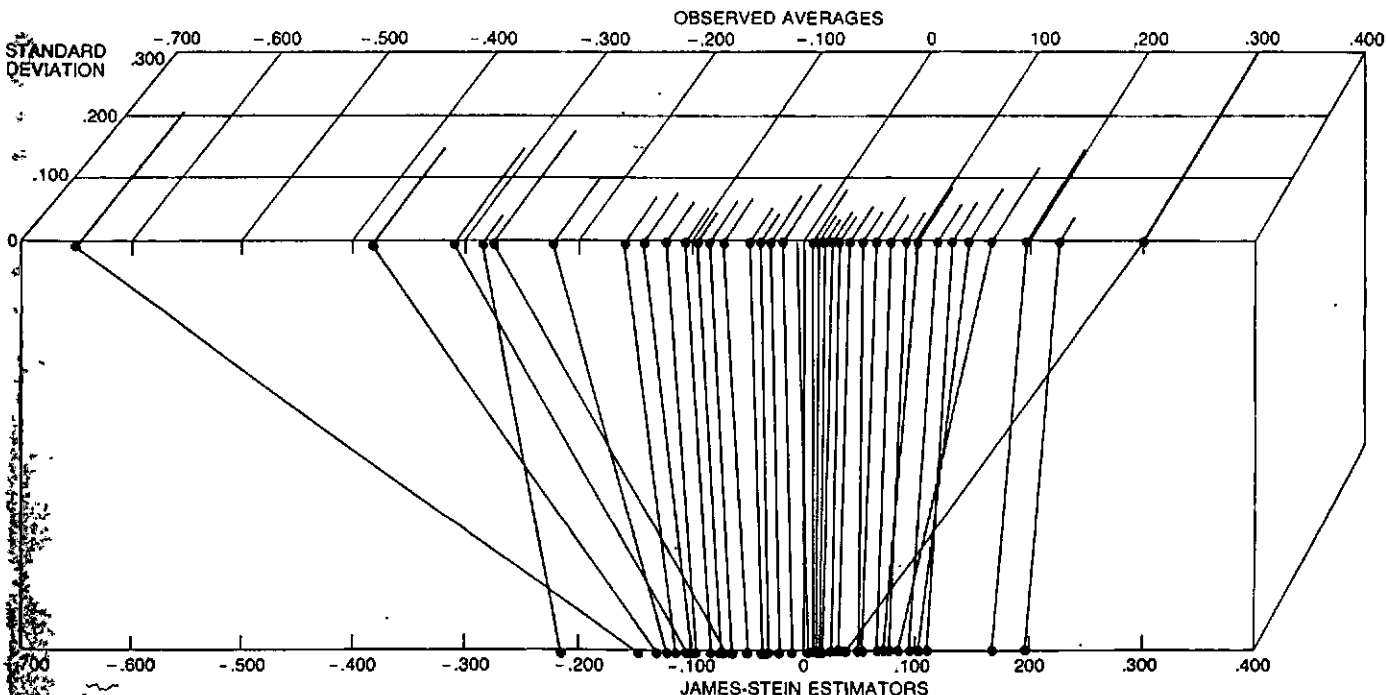
There is one purpose for which the

measured incidence may well be superior to the James-Stein estimator: when a single city is considered in isolation. As we have seen, the James-Stein method gives better estimates for a majority of cities, and it reduces the total error of estimation for the sum of all cities. It



INCIDENCE OF TOXOPLASMOSIS, a disease of the blood, was surveyed in 36 cities in the Central American country El Salvador. The measured incidence in each city can be regarded as an estimator of the true incidence, which is unobservable. The measured incidence has a normal distribution whose standard deviation is determined by

the number of people surveyed in that city. The measured rates are expressed in terms of deviation from the national incidence (the average of the rates observed in all the cities). Thus zero denotes exactly the national rate, and a city with a measured incidence of $-.040$ would have an observed rate 4 percent lower than the country as a whole.



SHRINKING of the observed toxoplasmosis rates to yield a set of James-Stein estimators substantially alters the apparent distribution of the disease. The shrinking factor is not the same for all the cities but instead depends on the standard deviation of the rate measured in that city. A large standard deviation implies that a measurement is based on a small sample and is subject to large random fluctuations;

that measurement is therefore compressed more than the others are. In the El Salvador data the most extreme observations tend to be correlated with the largest standard deviations, again suggesting the unreliability of those measurements. Compared with the observed rates, the James-Stein estimators can be proved to have a smaller total error of estimation. They also provide a more accurate ranking of the cities.

cannot be demonstrated, however, that Stein's method is superior for any particular city; in fact, the James-Stein prediction can be substantially worse.

Estimating the true mean for an isolated city by Stein's method creates serious errors when that mean has an atypical value. The rationale of the method is to reduce the overall risk by assuming that the true means are more similar to one another than the observed data. That assumption can degrade the estimation of a genuinely atypical mean. Now we see why imported cars should not be included in the same calculations with the 18 baseball players. There is a substantial probability that the automobiles will be atypical.

Suppose we ignore this hazard and lump together all 19 problems; we can then calculate the total expected squared error as a function of the true percentage of imported cars. It turns out that the risk for both the baseball players and the automobiles is reduced only if the percentage of imported cars happens to lie in the same range as the esti-

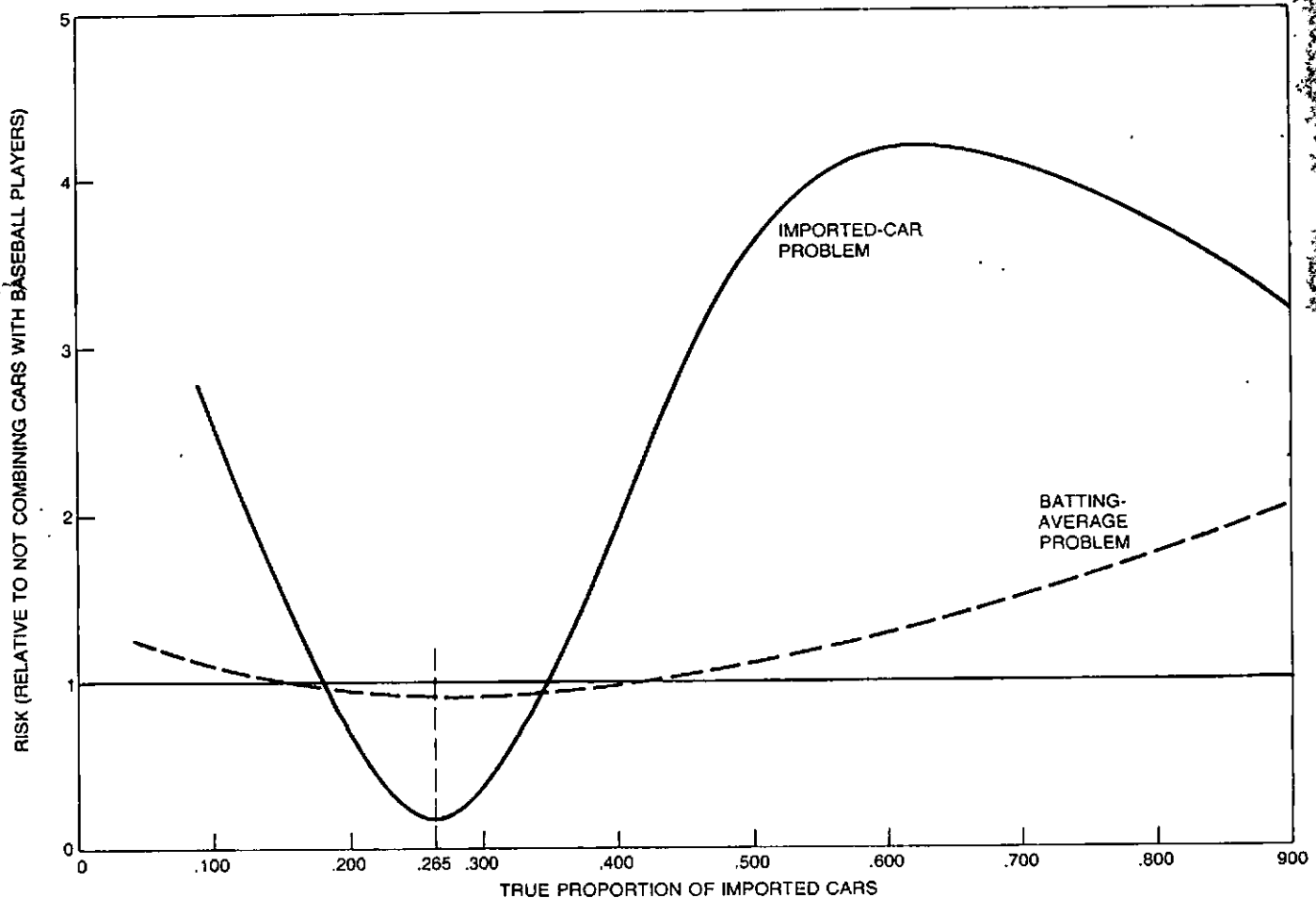
ated batting averages; otherwise the risk of error for both kinds of problem is increased.

The question of whether or not a particular mean is "typical" is a subtle one whose implications are not yet fully understood. Returning to the problem of toxoplasmosis in El Salvador, let us single out for attention the city of Alegría, which has the fifth-smallest measured incidence of the disease: .294. It is one of four cities included in the survey that are east of the Rio Lempa; all four have distinctly negative values of measured incidence y . It is plausible to suppose that this is no coincidence and that the rate of toxoplasmosis east of the Lempa is genuinely lower. A James-Stein estimator that consolidates information from the entire country therefore may be less than optimal in these cities. We have developed techniques for taking advantage of extra information of this kind, but the theory underlying those techniques remains rudimentary.

An astute follower of baseball might be aware that just as each player's batting ability can be represented by a

Gaussian curve, so too the true batting abilities of all major-league players have an approximately normal distribution. This distribution has a mean of .270 and a standard deviation of .015. With this valuable extra information, which statisticians call a "prior distribution," it is possible to construct a superior estimate of each player's true batting ability. This new estimator, which we shall give the label Z , is defined by the equation $Z = m + C(y - m)$. Here y is again the observed batting average of the player, but \bar{y} , the grand average, has been replaced by m , the mean of the prior distribution, which is known to have the value .270. In addition there is a different shrinking factor, C , which depends in a simple way on the standard deviation of the prior distribution (equal to .015).

This procedure is not a refinement of Stein's method; on the contrary, it predates Stein's method by 200 years. It is the mathematical expression of a theorem published (posthumously) in 1763 by the Reverend Thomas Bayes.



UNRELATED PROBLEMS can be lumped together for analysis by Stein's method, but only at the risk of increasing error. To the 18 batting averages computed earlier, for example, one might add a 19th number representing the proportion of imported cars observed in Chicago. New James-Stein estimators could then be calculated for both the baseball players and the automobiles, based on the grand

average of all 19 numbers. Nothing in the statement of Stein's theorem prohibits such a procedure, but the evident illogic of it has justifiably been criticized. In fact, including the unrelated data can reduce the risk function only if the proportion of imported cars happens to be near the mean batting average of .265; otherwise the expected error of estimation for both the cars and the baseball players is increased.

He was able to show that this estimator minimizes the expected squared error associated with the randomness in both the observed averages (y) and in the true means (θ).

The formula for the James-Stein estimator is strikingly similar to that of Bayes's equation. Indeed, as the number of means being averaged grows very large, the two equations become identical. The two shrinking factors c and C converge on the same value, and the grand average \bar{y} becomes equal to the mean m precisely when all players are included in the calculation. The James-Stein procedure, however, has one important advantage over Bayes's method. The James-Stein estimator can be employed without knowledge of the prior distribution; indeed, one need not even suppose the means being estimated are normally distributed. On the other hand, ignorance has a price, which must be paid in reduced accuracy of estimation. We have shown that the James-Stein method increases the risk function by an amount proportional to $3/k$, where k is again the number of means being estimated. The additional risk is therefore negligible when k is greater than 15 or 20, and it is tolerable for k as small as 9.

In this historical context the James-Stein estimator can be regarded as an "empirical Bayes rule," a term coined by Herbert E. Robbins of Columbia University. In work begun in about 1951 Robbins demonstrated that it is possible to achieve the same minimum risk associated with Bayes's rule without knowledge of the prior distribution, as long as the number of means being estimated is very large. Robbins' theory was immediately recognized as a fundamental breakthrough; Stein's result, which is closely related, has been much slower in gaining acceptance.

The James-Stein estimator is not the only one that is known to be better than the sample averages. Indeed, the James-Stein estimator is itself inadmissible! Its failure lies in the fact that the shrinking factor c can assume negative values, and it then pulls the means away from the grand average rather than toward it. When that happens, simply replacing c with zero produces a better estimator. This estimator in turn is also inadmissible, but no uniformly better estimator has yet been found.

The search for new estimators continues. Recent efforts have been concentrated on achieving results like those obtained with Stein's method for problems involving distributions other than the normal distribution. Several lines of work, including Stein's and Robbins' and more formal Bayesian methods seem to be converging on a powerful general theory of parameter estimation.

How to cheat a kid.



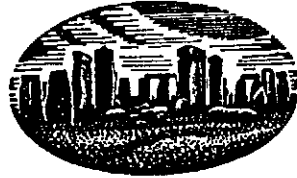
Thousands upon thousands of youngsters

are being cheated out of quality physical education programs every year because too few parents and school officials understand the difference between physical education and "gym" of bygone days. There's a new, enlightened physical education in many of our schools today. Physical education that touches and benefits every single boy and girl . . . develops individual confidence and self-esteem for a lifetime of sport and activity. Don't let your child miss the opportunity! Write for a free folder, "What Every Parent Should Know About The New Physical Education." If it's not in your child's school already, we'll tell you how to get it there.

PEPE

Physical Education Public Information
American Alliance for Health,
Physical Education, and Recreation
1201 16th St., N.W., Wash., D.C. 20036

See Robbins' SE



ANTIQUITY

A Periodical Review of Archaeology
edited by Glyn Daniel

Founded in 1927 by O. G. S. Crawford, ANTIQUITY has appeared regularly ever since and won acclaim the world over as the most authoritative journal in its field. While written by specialists, the articles, notes and reviews are popular in character and indispensable to all interested in the development of man and his past.

Professor Glyn Daniel, Faculty of Archaeology and Anthropology in the University of Cambridge, and Fellow of St John's College, has been Editor since 1956.

The annual subscription, postage included, is \$25. Subscription forms and bankers' orders are available on request from

ANTIQUITY PUBLICATIONS LIMITED
Heffers Printers Ltd, King's Hedges Road,
Cambridge, England CB4 2PQ

Save on Calculators

Hewlett-Packard

Model	Your Cost	Model	Your Cost
HP 21	\$ 64.00	HP 55 (was \$335.00)	149.00
HP 22	100.00	HP 67	369.00
HP 25	116.00	HP 80	236.00
HP 25 C	168.00	HP 91 Scient. Printer	249.00
HP 27	140.00	HP 97	629.00

Free Reserve Power Pack with purchase of HP-21, -22, -25, -25C and -27 if bought before May 31. We are an H-P franchised dealer. We carry all accessories at a discount.

Texas Instruments

Model	Your Cost	Model	Your Cost
SR 52	\$177.00		
PC 100	147.00		
SR 52/PC 100	322.00		
Combo Sale	79.00	Model	
SR 56	52.00	TI 5050M	88.00
SR 51-2	52.00	TI 5100	47.90
SR 40	31.00	Little Professor	16.00
TI 30SP	20.00	TI Digital Watch 501	16.00
TI 2550-3	29.00	Money Manager	20.00
Bus. Analyst	31.00	Libraries for SR 52	23.00 up
TI 5040	108.00	TI accessories available	



Specials

Model	Your Cost	Model	Your Cost
Norelco #95	\$149.00	Canon #P 1010 Elec. Printer	89.00
Norelco #185	107.00	Sharp #EL 1052 Elec. Printer	85.00
Norelco #88	255.00	3M Dry Photocopier 051	139.00
Norelco #186	265.00	SCM Elec. Typs. with case #2200	229.00
Norelco #97	309.00	Craig Elec. Notebook #2625	149.00
Norelco #98	419.00	Casio CQ1 Comp. Alarm Clock	44.00
Norelco Cassette	2.95	Chrono-Alarm-Calc. Digital Watches—All kinds!	

Also SCM - Olivetti - Rockwell - Victor - APF - Lloyds - Unirax - Amara - 3M
Litton - Sharp - Craig - Canon - Panasonic - Sony - Sanyo - and many more.

Prices FOB L.A. - Goods subject to availability - Please request our famous catalog. We will beat any deal if the competition has the goods on hand - Add \$3.00 for shipping handheld calculators - CA residents, add 6% sales tax.

OLYMPIC SALES COMPANY, INC.

216 South Oxford Ave. - P.O. Box 74545
Los Angeles, CA 90004 - (213) 381-3911 - Telex 67 3477