ADVANCES IN MATHEMATICS 16, 259–277 (1975)

Biased Versus Unbiased Estimation

BRADLEY EFRON

Stanford University, Stanford, California 94305

DEDICATED TO STAN ULAM

Statisticians have begun to realize that certain deliberately induced biases can dramatically improve estimation properties when there are several parameters to be estimated. This represents a radical departure from the tradition of unbiased estimation which has dominated statistical thinking since the work of Gauss. We briefly describe the new methods and give three examples of their practical application.

1. INTRODUCTION

Two young statisticians were interviewing for the same job. "Suppose," said the would-be employer, "that you observed n independent normally distributed random variables with mean θ and variance 1, say

$$x_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta, 1) \qquad i = 1, 2, \ldots, n,$$

and on the basis of $\mathbf{x} \equiv (x_1, x_2, \ldots, x_n)$ I asked you to estimate the unknown parameter θ. What would you do?"

The first statistician, a quick fellow, answered "I would use the unbiased estimation rule

$$\delta^0(\mathbf{x}) \equiv \bar{x} \equiv \sum_{i=1}^{n} x_i/n.$$

It has minimum variance among all unbiased estimators (those satisfying $E_\theta \delta(\mathbf{x}) = \theta$ for all θ, "E_θ" indicating expectation when θ is the parameter value), and likewise among all translation invariant estimators ($\delta(\mathbf{x} + (c, c, c, \ldots, c)) = \delta(\mathbf{x}) + c$). Moreover it is minimax (minimizes the maximum expected squared error), admissable (no competing estimation rule has smaller expected squared error for all values of θ), and it is the maximum likelihood estimator (choosing $\theta = \bar{x}$ maximizes, among all values of θ, the probability of obtaining the value of $\mathbf{x}$ actually observed)."

Naturally the employer was impressed, but a sense of fairness compelled him to give the second statistician his chance. "Well, sir," he responded after an embarrassing silence, "I think I might try

$$\delta^1(\mathbf{x}) \equiv \bar{x} - \Delta(\bar{x})$$

where $\Delta(-x) = -\Delta(x)$ and for $x \geqslant 0$,

$$\Delta(x) \equiv \frac{1}{2\sqrt{n}} \min(\sqrt{n}\, x, \Phi(-\sqrt{n}\, x)),$$

where of course

$$\Phi(t) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-(1/2)s^2}\, ds$$

represents the normal distribution function as usual."

After the laughter had died down the employer asked him how he could justify such a bizarre recommendation. The second statistician admitted that δ^1 had none of the nice properties of δ^0. It wasn't unbiased, wasn't invariant, wasn't minimax, and wasn't even admissable. "On the other hand my guess will be closer than my competitor's to the true value of θ more than half of the time, no matter what θ is!" After an easy computation, which the reader may want to do for himself, it turned out that indeed

$$\text{Prob}_\theta\{|\, \delta^1(\mathbf{x}) - \theta\, | < |\, \delta^0(x) - \theta\, |\} > \tfrac{1}{2}$$

for all θ. He got the job of course. As the employer put it, "Why should I settle for second best?"

Now it is possible to give a good argument that the first statistician deserved the job, and that δ^0 is really a better estimator than δ^1. Nevertheless our shaggy statistician story has a serious point: in more complicated situations involving the estimation of several parameters at the same time, statisticians have begun to realize that biased estimation rules have definite advantages over the usual unbiased estimators. This represents a radical departure from the tradition of unbiased estimation which has dominated statistical thinking since Gauss' development of the least squares method. A brief description of the new theory is given in Section 2 followed in Section 3 by examples of its application to three data analysis problems.

My purpose in writing this article is to whet the interest of non-

statisticians in the use of these new estimators. An equivalent of the Surgeon-General's warning may be in order: these methods are not perfectly understood yet, and are still the subject of heated controversy among statisticians (see the discussion following [7]). Most of the material presented here is abstracted from a series of articles by Carl Morris and myself [3–8]. The unsatisfied reader may wish to read [6] for more theoretical background and [8] for a fuller description of data analysis procedures.

2. The James–Stein Estimator

Suppose we wish to estimate several parameters $\theta_1, \theta_2, \ldots, \theta_k$, and for each one we observe n independent random variables,

$$x_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, \sigma^2) \qquad \begin{matrix} i = 1, 2, \ldots, k, \\ j = 1, 2, \ldots, n. \end{matrix} \tag{2.1}$$

Here σ^2 is the common variance of the x_{ij}, which for convenience of presentation we assume known although this isn't necessary for the theory which follows. A sufficient statistic (one which contains all the information) for θ_i is

$$y_i \equiv \bar{x}_i \equiv \sum_{j=1}^{n} x_{ij}/n, \tag{2.2}$$

the more exact statement being that the vector $\mathbf{y} = (y_1, y_2, \ldots, y_k)$ is sufficient for the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. The y_i are independently normally distributed with mean θ_i and variance $D \equiv \sigma^2/n$,

$$y_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, D) \qquad i = 1, 2, \ldots, k. \tag{2.3}$$

From the vector $\mathbf{y}$ we wish to infer the value of $\boldsymbol{\theta}$. It is customary to do so by means of an estimation rule

$$\boldsymbol{\delta}(\mathbf{y}) \equiv (\delta_1(\mathbf{y}), \delta_2(\mathbf{y}), \ldots, \delta_k(\mathbf{y})),$$

$\delta_i(\mathbf{y})$ being the estimator of θ_i.

Model (2.3) is the simplest case of Gauss' "linear model"

$$\mathbf{y} = \boldsymbol{\theta}\mathbf{M} + \boldsymbol{\epsilon} \tag{2.4}$$

where $\mathbf{y}$ is a $1 \times r$ vector of observed variables, $\boldsymbol{\theta}$ a $1 \times k$ vector of unknown parameters, and $\mathbf{M}$ a known $k \times r$ "structure matrix" which we assume to be of rank k with $k \leqslant r$ to avoid some nasty details. (For (2.3), $r = k$ and $\mathbf{M} = \mathbf{I}$, the $k \times k$ identity matrix.) The noise vector $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_r)$ is assumed to have independent normal components with mean 0 and common variance, that being D in (2.3).

Gauss suggested the "least squares estimator" for this situation,

$$\boldsymbol{\delta}^0(\mathbf{y}) \equiv \mathbf{y}\mathbf{M}^T(\mathbf{M}\mathbf{M}^T)^{-1},$$

which is also "maximum likelihood" and "minimum variance unbiased" for $\boldsymbol{\theta}$. (Among unbiased estimators, $E_{\boldsymbol{\theta}}\boldsymbol{\delta}(\mathbf{y}) = \boldsymbol{\theta}$ for all $\boldsymbol{\theta}$, $\boldsymbol{\delta}^0$ has minimum variance component by component, and also for any linear combination of the components.)

The linear model is used extensively in all the sciences, $\boldsymbol{\delta}^0$ usually being the estimator of choice. It turns out that (2.4) can be reduced to (2.3) by suitable linear transformations, so that the simpler structure (2.3) actually has all the statistical content of the full linear model. The least squares estimator $\boldsymbol{\delta}^0$ is simply

$$\boldsymbol{\delta}^0(\mathbf{y}) = \mathbf{y} \tag{2.5}$$

in this case. This says the obvious: y_i is the best unbiased estimator of θ_i in the model (2.3).

In 1960[1] James and Stein [9] suggested another estimator for $\boldsymbol{\theta}$,

$$\boldsymbol{\delta}^1(\mathbf{y}) \equiv \left[1 - \frac{(k-2)D}{S}\right]\mathbf{y}, \tag{2.6}$$

where

$$S \equiv \|\mathbf{y}\|^2 \equiv \sum_{i=1}^{k} y_i^2. \tag{2.7}$$

(*Here and henceforth we assume* $k \geqslant 3$.) At first sight $\boldsymbol{\delta}^1$ appears ridiculous since the estimate of θ_i,

$$\delta_i^1 = \left[1 - \frac{(k-2)D}{S}\right] y_i,$$

depends not only on y_i but on the seemingly irrelevant values of y_j, $j \neq i$. Ah, but beware!

[1] Stein suggested a similar estimator in his 1955 paper [11].

Suppose one measures the performance of an estimator δ by its expected sum of squared error risk,

$$R(\theta, \delta) \equiv E_\theta L(\theta, \delta(\mathbf{y})) \equiv E_\theta \sum_{i=1}^{k} (\delta_i(\mathbf{y}) - \theta_i)^2 \equiv E_\theta \| \delta - \theta \|^2. \tag{2.8}$$

(In the usual parlance $R(\theta, \delta)$ is the *risk function* of δ under the *loss function* $L(\theta, \delta) = \| \delta - \theta \|^2$.) For $\delta^0 = \mathbf{y}$ we have from (2.3) that

$$R(\theta, \delta^0) = kD$$

for all θ. We know that δ^0 minimizes $R(\theta, \delta)$ among all unbiased estimators δ. However James and Stein showed that

$$R(\theta, \delta^1) < kD$$

for all θ, so that in terms of total squared error risk δ^1 is uniformly preferable to δ^0!

The advantage enjoyed by δ^1 is by no means trivial. Figure 1 compares

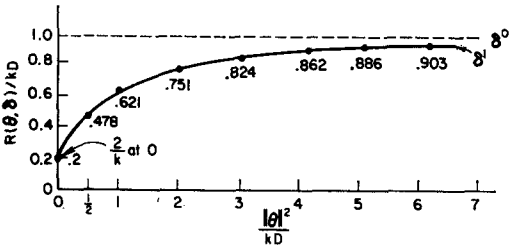

FIG. 1. Risk of the James–Stein rule δ^1 compared to that of the unbiased estimator δ^0, for $k = 10$.

$R(\theta, \delta^1)$ and $R(\theta, \delta^0)$, for $k = 10$, as a function of $\| \theta \|^2/kD$. We see that the most favorable case is $\theta = 0$ for which $R(\theta, \delta^1)/R(\theta, \delta^0) = 2/10$. (For dimension k, $R(0, \delta^1)/R(0, \delta^0) = 2/k$.)

Why haven't statisticians and users of statistics rushed to embrace this considerable improvement over the method of least squares? (They haven't-it's barely been used at all.) Several reasons can be given.

(i) *Blind stupid prejudice.* Unbiased estimators have been used on literally millions of real problems, with generally satisfactory results. The new estimators haven't. Their theoretical superiority has yet to be tested in the rigors of wide-spread application. Prejudice of this sort isn't really blind or stupid, just conservative.

(ii) *Counter-intuition.* We would expect that changing the origin and scale of our measurements would change any reasonable estimator in the obvious way,

$$\boldsymbol{\delta}(b\mathbf{y} + (c, c, c, \ldots, c)) = b\boldsymbol{\delta}(\mathbf{y}) + (c, c, \ldots, c)$$

for any numbers b and c. $\boldsymbol{\delta}^1$ does not have this *invariance* property. As mentioned before $\boldsymbol{\delta}^1$ is also statistically counter-intuitive in that it uses data other than y_i to estimate θ_i. If the different θ_i refer to obviously disjoint problems (e.g., θ_1 is the speed of light, θ_2 is the price of tea in China, θ_3 is the efficacy of new treatment for psoriasis, etc.) combining the data can produce a definitely uncomfortable feeling in the statistician.

(iii) *Bias.* The very name "unbiased" suggests the appeal of this concept to notions of scientific objectivity. All values of the parameter vector $\boldsymbol{\theta}$ are treated with an equal hand by $\boldsymbol{\delta}^0$. Not so with $\boldsymbol{\delta}^1$, which biases estimators toward the origin, the more so the closer $\mathbf{y}$ itself is to the origin. As a matter of fact (2.6) shows that for $S < (k-2)$, $\boldsymbol{\delta}^1(\mathbf{y})$ is actually pulled past the origin in the direction opposite $\mathbf{y}$. It can be shown that if we legislate out this behaviour by refusing to go past the origin, that is we use the "plus-rule" estimator

$$\boldsymbol{\delta}^{1+}(\mathbf{y}) \equiv \left[1 - \frac{(k-2)D}{S}\right]_+ \mathbf{y}, \tag{2.9}$$

where $[x]_+ \equiv \max(0, x)$ as usual, then $R(\boldsymbol{\theta}, \boldsymbol{\delta}^{1+}) < R(\boldsymbol{\theta}, \boldsymbol{\delta}^1)$ for all θ, giving a reduction in risk for all $\boldsymbol{\theta}$. (The plus-rule and its variations are actually the proposed competitors to $\boldsymbol{\delta}^0$ in Section 3, but of course unless $S < k-2$ you can't tell the difference between $\boldsymbol{\delta}^{1+}$ and $\boldsymbol{\delta}^1$.)

There is no need to give $\mathbf{0}$ the preferred position in the definition of $\boldsymbol{\delta}^1$ or $\boldsymbol{\delta}^{1+}$. For any vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)$ define

$$\delta_i^{1+} \equiv \mu_i + \left[1 - \frac{(k-2)D}{S}\right]_+ (y_i - \mu_i), \qquad i = 1, 2, \ldots, k, \tag{2.10}$$

where now

$$S \equiv \sum_{i=1}^{k} (y_i - \mu_i)^2 = \| \mathbf{y} - \boldsymbol{\mu} \|^2.$$

This rule also uniformly dominates $\boldsymbol{\delta}^0$ no matter what the choice of $\boldsymbol{\mu}$. By choosing $\boldsymbol{\mu}$ in different ways we can make $\boldsymbol{\delta}^{1+}(\mathbf{y})$ take on any value

we want within a sphere of radius $(k-2)^{1/2}$ centered at $\mathbf{y}$. Even assuming we play fair and choose μ before observing $\mathbf{y}$ we are still biasing our results toward μ, a value of θ that may have vested interest for the statistician.

(iv) *Sum of squared-error loss.* What if we change the loss function from $L(\theta, \delta) = \| \delta - \theta \|^2$ to, say, $L(\theta, \delta) = \sum_{i=1}^{k} | \delta_i - \theta_i |$. Is δ^1 still preferable to δ^0? The answer is probably yes, though the mathematics speak less clearly than for $\| \theta - \delta \|^2$. Roughly speaking, present indications are that for any "ensemble" loss function, that is one that sums errors over all the coordinates, δ^1 or some variation of it will dominate δ^0. Even the assumption of normal distributions doesn't seem to be very important, see [2], [6].

Another example of an ensemble loss function is the multi-parameter analogue of that employed by the statistician who got the job in Section 1. It is possible to show that $\text{Prob}_\theta\{\| \delta^1 - \theta \| < \| \delta^0 - \theta \|\}$ is greater than $\frac{1}{2}$ for all values of θ. An explicit formula, which won't be derived here, is

$$\text{Prob}_\theta\{\| \delta^1 - \theta \| < \| \delta^0 - \theta \|\} = \text{Prob}\left\{\chi_k^2\left(\frac{\| \theta \|^2}{4D}\right) \leqslant \frac{\| \theta \|^2}{4D} + \frac{k-2}{2}\right\}, \quad (2.11)$$

where $\chi_k^2(\sum_1^k \sigma_i^2)$ is the distribution of $\sum_1^k w_i^2$, w_i being independent normal random variables with mean σ_i and variance 1. For k reasonably large (say $\geqslant 8$) and $\| \theta \|^2/kD$ reasonably small (say $\leqslant 3$), (2.11) gives quite high probabilities of $\| \delta^1 - \theta \|$ being less than $\| \delta^0 - \theta \|$, on the order of 90%.

It is actually the "sum" in "sum of squared errors" that is the crucial assumption. If we really aren't interested in $\theta_2, \theta_3, \ldots, \theta_k$, just θ_1, then $L_1(\theta, \delta) = (\delta_1 - \theta_1)^2$ is a more reasonable loss function than $\| \delta - \theta \|^2$. It is *not* true that $E_\theta(\delta_1^1 - \theta_1)^2 < E_\theta(\delta_1^0 - \theta_1)^2$ for all θ. As a matter of fact $E_\theta(\delta_1^1 - \theta_1)^2 / E_\theta(\delta_1^0 - \theta_1)^2$ can be as large as about $\sqrt{k}$ for certain configurations of $\theta_1, \theta_2, \ldots, \theta_k$, (namely $\theta_1 = \sqrt{k}$, $\theta_2 = \theta_3 = \cdots = \theta_k = 0$).

There are the beginnings of a paradox here: we expect $| \delta_i^1 - \theta_i |$ to be smaller than $| \delta_i^0 - \theta_i |$ for a majority of the coordinates θ_i. (For example if $\| \theta \|^2/kD$ is near 1 and no one of the θ_i is enormously large then for large k it is possible to show that δ_i^1 will be closer for about 65% of the coordinates.) On the other hand for any one coordinate we can't guarantee we are doing better, even in expectation, and we may do considerably worse.

Morris and I have confronted those criticisms, at least the last 3, in our long series of papers. Some of our answers will be clear from the examples of Section 3. The truth is that δ^1 is less automatic than δ^0, does involve more judgmental factors in its use, and can lead to greater disasters if misused. Nevertheless, as the examples show, *not* using δ^1 or some close cousin can make one a very inefficient statistician, a disaster in its own right.

3. Three Examples of Biased Estimation in Practice

(i) 18 *Baseball Players.* Table I, column 1, shows the batting averages of 18 major league players after their first 45 times at bat in

TABLE I

1970 Batting Avergages for 18 Major League Players

i	Unbiased estimate $\delta_i^0 = y_i$	Parameter value θ_i	James–Stein est. δ_i^1		At bats, remainder of 1970
1	0.400	0.346	0.293	(0.334)[a]	367
2	0.378	0.298	0.289	(0.312)[a]	426
3	0.356	0.276	0.284	(0.290)[a]	521
4	0.333	0.221	0.279		276
5	0.311	0.273	0.275		418
6	0.311	0.270	0.275		467
7	0.289	0.263	0.270		586
8	0.267	0.210	0.265		138
9	0.244	0.269	0.261		510
10	0.244	0.230	0.261		200
11	0.222	0.264	0.256		277
12	0.222	0.256	0.256		270
13	0.222	0.304	0.256		434
14	0.222	0.264	0.256		538
15	0.222	0.226	0.256		186
16	0.200	0.285	0.251		558
17	0.178	0.319	0.247	(0.243)[a]	405
18	0.156	0.200	0.242	(0.221)[a]	70

Player number	Batting average after 45 at bats $\bar{y} = 0.265$	Batting average remainder of season	$\bar{y} + [1 - (k-3)D/S](\bar{y}_i - \bar{y})$ $= 0.265 + 0.212(y_i - \bar{y})$	Average remainder at bats $= 369.3$

[a] Limited translation estimate. All other values agree with δ_i^1.

the 1970 season. These can be considered as unbiased estimates y_i of θ_i, the true probability of Player i getting a hit. Column 2 gives a much better estimate of θ_i, the batting average for Player i during the remainder of the season, based on an average of about 370 more at bats. We consider these to be the acutal θ_i, though a more careful analysis would include the sampling error in these numbers.

We now apply the James–Stein estimator in form (2.10), choosing all the μ_i equal to $\bar{y} \equiv \sum_1^k y_i/18$. Letting the data choose the μ_i in this way effectively removes one dimension ("one degree of freedom" in statistical jargon) from the problem, leading to the estimator

$$\delta_i^1 = \bar{y} + \left[1 - \frac{(k-3)D}{S}\right](y_i - \bar{y}),$$
$$S \equiv \sum_1^k (y_i - \bar{y})^2. \tag{3.1}$$

Notice that the lost dimension has changed $k-2$ to $k-3$. For this problem D is unknown (actually depending on θ_i) but from the properties of the binomial distribution we can estimate it by

$$D = \frac{\bar{y}(1-\bar{y})}{45} = 0.004332.$$

The resulting estimation rule, $\delta_i^1 = 0.265 + 0.212(y_i - \bar{y})$, is given in column 3. The losses are

$$\sum_1^{18} (\delta_i^0 - \theta_i)^2/D = 17.68, \qquad \sum_1^{18} (\delta_i^1 - \theta_i)^2/D = 5.05 \tag{3.2}$$

so δ_i^1 outperforms δ_i^0 by a factor of 3.50.

It is shown in [4] that if we follow the estimation rule δ_i^1 as closely as possible *subject to the constraint that no estimated value be more than one standard deviation D away from the unbiased estimate* δ_i^0, then (a) we still get most of the sum of squared errors risk reduction associated with δ_i^1; and (b) the maximum possible risk for individual components is reduced substantially. This "limited translation rule," an attempt to have our cake and eat it too vis-a-vis objection (iv) of Section 2, is also given in Table I. Notice that it does appear to protect Player 1 (Roberto Clemente!) from over shrinking toward the common mean.

A more careful treatment of this data is given in [8], but it is reassuring

to see δ^1 working as predicted in a situation where no attempt has made to exactly satisfy the model (2.1).

(ii) *Ten Reaction Time Experiments.*[2] Each of ten subjects was asked to perform a certain task under seven different conditions. Let x_{ij} indicate the (natural) log reaction time of subject i under condition j, $i = 1, 2,..., 10$, $j = 1, 2,..., 7$. The two way analysis of variance model ("ANOVA")

$$x_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}\,, \qquad \sum_1^{10} \alpha_i = \sum_1^n \beta_j = 0, \tag{3.3}$$

was used to analyze the data. Here μ is the overall mean, α_i the main effect for Subject i, β_j the main effect for Condition j, and ϵ_{ij} the random noise, assumed to be independent normal with mean 0 and variance σ^2,

$$\epsilon_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2) \qquad i = 1, 2,..., 10, \qquad j = 1, 2,..., 7. \tag{3.4}$$

For example a value of $\alpha_1 = 0.12$ means that Subject 1 would average 0.12 greater on the log scale than all 10 subjects together, in the absence of the noise from the ϵ_{ij}—that is he would be about 12% slower than the group average. In the discussion below we are only interested in estimating the patient main effects α_i from the data x_{ij}.

An unbiased estimate of α_i is

$$y_i \equiv \sum_{j=1}^{7} x_{ij}/7 - \sum_{i=1}^{10}\sum_{j=1}^{7} x_{ij}/70 \tag{3.5}$$

for which a simple calculation gives

$$y_i \sim \mathcal{N}(\alpha_i\,, D), \qquad D = (9/70)\,\sigma^2. \tag{3.6}$$

The y_i are not independent since they, like the α_i, must sum to zero. The applicable version of the James–Stein estimator is

$$\delta_i^1 = \left[1 - \frac{(k-3)D}{S}\right] y_i\,, \qquad S = \sum_1^k y_i^2, \qquad k = 10. \tag{3.7}$$

This is really the same rule as (3.1) except here we are not interested in putting the overall mean back into the estimates.

[2] I am grateful to Dr. R. Angel of the Stanford Medical School for allowing me to abstract this data from a larger experiment he is conducting.

σ^2, hence D, is unknown to us, but can be estimated in the usual way from model (3.3). This gives an unbiased estimate $\hat{\sigma}^2$ independent of y_i, and we can take $\hat{D} = (9/70)\hat{\sigma}^2$. (For those familiar with ANOVA, $\hat{\sigma}^2$ is biased on a chi-square variable with 58 degrees of freedom. Actually it can be shown that a slightly biased estimator for D is preferable, but we shall ignore this small improvement.)

This whole experiment was repeated 10 times, yielding a total of 700 observations—10 Subjects, 7 conditions, 10 experiments. We can use these repetitions as a check on how a given estimator performed in any given experiment. Each column of Table 2 refers to an experiment analyzed separately from the others. The upper number in each box is $\delta_i^0 = y_i$, the unbiased estimate of α_i for that experiment. For example experiment 1 has $y_3 = -.16$, indicating that Subject 1 reacted about 16% faster than the average of the 10 subjects for that experiment. The lower number in each box is δ_i^1 as given in (3.7), with $\hat{D}$ substituted for D. $\hat{D}$ is given at the bottom of each column, along with the shrinkage factor $[1 - (k-3)\hat{D}/S]$ in (3.7).

For each i we can average the values of δ_i^0 over the 10 experiments to obtain a much more accurate unbiased estimate of α_i. We take these to be the true α_i values even though they still have some sampling variability in them. They are listed as the top numbers in the "combined" column. The last column of Table 2 compares δ_i^0 with δ_i^1 over the 10 experiments. If $(\delta_{ie}^0, \delta_{ie}^1)$ represent the two estimates for α_i in experiment e, $e = 1, 2, \ldots, 10$. Then the two numbers given are

$$\sum_{e=1}^{10} (\delta_{ie}^0 - \alpha_i)^2 \quad \text{and} \quad \sum_{e=1}^{10} (\delta_{ie}^1 - \alpha_i)^2. \tag{3.8}$$

The first of these is greater than the second for eight out of the ten Subjects. Overall

$$\sum_{i=1}^{10}\sum_{e=1}^{10} (\delta_{ie}^0 - \alpha_i)^2 = 1.025, \qquad \sum_{i=1}^{10}\sum_{e=1}^{10} (\delta_{ie}^1 - \alpha_i)^2 = 0.759 \tag{3.9}$$

indicating that δ_i^1 was about 25% more accurate than δ_i^0 over all 10 experiments.

Notice that the two Subjects who do worse under δ_i^1 then δ_i^0 are the slowest (Subject 6) at the fastest (Subject 8, tied with Subject 3). This underscores the point that δ_i^1 can have high risk on the more unusual components θ_i. The limited translation modification mentioned previously can be used effectively here.

TABLE II

Ten Reaction-Time Experiments. Upper Number is Unbiased Estimate—
Lower Number is James–Stein Estimate on that Column's Data

Subject		Experiment → 1	2	3	4	5	6	7	8	9	10	Combined	$\frac{\sum_e (\delta^0_{ie} - \alpha_i)^2}{\sum_e (\delta^1_{ie} - \alpha_i)^2}$
	1	−0.03	0.06	−0.08	−0.04	−0.05	0.15	0.05	0.03	−0.10	−0.06	−0.00	0.0545
		−0.01	0	−0.05	−0.02	−0.03	0.13	0.04	0.01	−0.07	−0.02	−0.00	0.0278
	2	0.08	0.24	0.17	0.14	−0.06	0.37	0.20	0.11	0.26	−0.07	0.14	0.1704
		0.03	0	0.10	0.08	−0.04	0.32	0.14	0.04	0.18	−0.02	0.13	0.1389
	3	−0.16	0.07	−0.32	−0.30	−0.28	−0.18	−0.19	−0.25	−0.08	−0.09	−0.18	0.1300
		−0.06	0	−0.18	−0.18	−0.19	−0.15	−0.13	−0.10	−0.06	−0.03	−0.17	0.0936
	4	−0.04	0.02	0.02	0.04	−0.05	−0.05	−0.16	−0.02	0.13	−0.12	−0.03	0.0615
		−0.01	0	0.01	0.02	−0.03	−0.04	−0.11	−0.01	0.09	−0.04	−0.03	0.0256
	5	0.08	0.01	0.03	−0.00	0.14	0.06	−0.22	−0.00	−0.23	−0.10	−0.02	0.1307
		0.03	0	0.02	−0.00	0.10	0.05	−0.15	−0.00	−0.16	−0.03	−0.02	0.0612

Subject													
	6	0.25	0.15	0.17	0.35	0.21	0.24	0.29	0.02	0.28	0.22	0.22	0.0718
		0.09	0	0.10	0.21	0.14	0.21	0.20	0.01	0.19	0.07	0.20	0.1542
	7	0.09	−0.10	−0.06	0.09	0.30	0.09	0.09	0.11	0.07	0.29	0.10	0.1431
		0.02	0	−0.03	0.05	0.20	0.08	0.06	0.04	0.05	0.09	0.09	0.0525
	8	−0.13	−0.14	−0.16	−0.23	−0.23	−0.34	−0.18	−0.08	−0.15	−0.10	−0.18	0.0524
		−0.05	0	−0.09	−0.14	−0.16	−0.29	−0.13	−0.03	−0.10	−0.03	−0.17	0.1214
	9	−0.09	−0.13	0.17	−0.03	0.04	−0.20	0.03	−0.08	−0.17	0.05	−0.04	0.1183
		−0.02	0	0.10	−0.02	0.03	−0.17	0.02	−0.03	−0.12	0.02	−0.04	0.0572
	10	−0.04	−0.16	0.07	−0.02	−0.03	−0.14	0.09	0.17	−0.00	−0.02	−0.01	0.0898
		−0.01	0.00	0.04	−0.01	−0.02	−0.12	0.06	0.07	−0.00	−0.01	−0.01	0.0266
	$\hat{D}$	0.011	0.020	0.012	0.014	0.011	0.007	0.010	0.010	0.011	0.016	0.00122	1.025 / 0.759
	$\left[1-\frac{7\hat{D}}{S}\right]$	0.35	−0.02	0.57	0.60	0.68	0.86	0.70	0.40	0.69	0.31	0.929	Total
	$F_{9,54}$	1.25	0.82	1.93	2.09[a]	2.64[a]	6.01[b]	2.74[b]	1.38	2.66[a]	1.21	3.42[b]	

[a] Sig at 0.05 level.

[b] Sig at 0.01 level.

In experiment 2 the shrinkage factor comes out negative, and we use the plus-rule δ_i^{1+} to estimate all $\alpha_i = 0$. One biased estimation procedure that has been widely used is to run the usual F test for the hypothesis "all $\alpha_i = 0$," estimating α_i by 0 if the test accepts the hypothesis, and by δ_i^0 if the test rejects the hypothesis. This amounts to estimating α_i by

$$[1 - I(F)]\, y_i ,$$

where $I(F)$ equals 1 or 0 as F is less or greater than some conventional value, usually the 95th percentile of F under the hypothesis all $\alpha_i = 0$. $\boldsymbol{\delta}^{1+}$ is a smoother version of this same idea, being expressable as

$$\delta_i^{1+} = \left[1 - \frac{8.33}{10}\frac{1}{F}\right]_+ y_i \tag{3.10}$$

in the case at hand. It is shown in [10] that $\boldsymbol{\delta}^{1+}$ uniformly dominates the F-test procedure in terms of sum of squared error risk. The F value for each experiment is listed as "$F_{9,54}$" (indicating the proper degrees of freedom) in Table II.

Actually the 10 experiments were run under somewhat different conditions. The widely varying values of F in Table II suggest that the α_i themselves may have changed from experiment to experiment. The reader may wish to propose an estimate for α_{ie}, the ith subject main effect in experiment e, from the data in Table II. Hint: in each row we can shrink the values δ_{ie}^0, $e = 1, 2, \ldots, 10$, toward their mean value α_i.

The lower numbers in the combined column were obtained by applying $\boldsymbol{\delta}^1$ to the α_i. There isn't much shrinkage effect because the α_i are individually quite accurate themselves.

(iii) *The Sunspot Data.* Figure 2 is a graph of the log spectrogram for the number of sunspots occurring annually from 1947–1924. The plotted point at each frequency is an unbiased estimate of the log power at the frequency.[3] The successive frequencies plotted are separated from each other by approximately 0.0057 cycles/year (0.0057 = 1/176 years). Assuming the sunspot generating mechanism is stationary over time, there are theoretical reasons for believing that the plotted estimates will be independent of each other with mean say θ_i, the log

[3] Those numbers are obtained as $\log S_p + 0.573$ where S_p is the Schuster's spectrogram for the sunspot data taken from Table A.3.2 of [1]. Adding 0.573 compensates for the bias induced by taking logarithms.

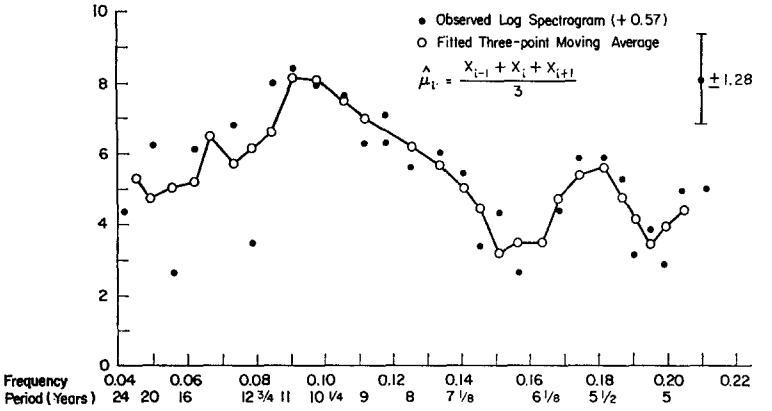

FIG. 2. Log Spectrogram for Schuster's sunspot data.

power at the ith frequency, and variance $\sigma^2 = (1.28)^2$. The sampling distribution is nonnormal (actually being the log of an exponential random variable), but this will not affect the analysis which follows.

We would like to estimate the "spectrum," that is the θ_i values. Spectral estimation is one area of statistics where biased estimates have traditionally played a preferred role. Smoothing the sample spectrogram by some type of moving average process is usually recommended. If successive θ_i are not too different from one another smoothing reduces sampling variance while inducing only mild bias, thus reducing the expected mean square error. Figure 2 shows the results of a three point moving average. If x_i is the observed value of the spectrogram at frequency i, the smoothed value is defined to be

$$\mu_i \equiv \frac{x_{i-1} + x_i + x_{i+1}}{3}. \tag{3.11}$$

The residuals

$$y_i \equiv x_i - \mu_i \tag{3.12}$$

have variance

$$D = \tfrac{1}{3}\sigma^2 = 1.09$$

and standard deviation $\sqrt{1.09} = 1.05$. Looking at Fig. 2 with this in mind it is clear that we have oversmoothed the spectrogram, at least near the low end of the frequency scale. If we are serious about estimating the ϕ_i, rather than just getting an idea of the spectrum's general

appearance, we can use Stein-type ideas to compromise between the two extremes of complete smoothing and no smoothing.

A digression is necessary here. Suppose $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is a parameter vector and $\mathbf{y} = (y_1, \ldots, y_k)$ an observed vector such that y_i has mean θ_i and variance some known value D. Define

$$A \equiv \sum_1^k \theta_i^2/k. \tag{3.13}$$

It is easy to show that among all estimates of the form $\delta_i = (1 - b)y_i$, the choice of the constant b which minimizes the expected total squared error is $B \equiv D/(A + D)$. (This result does not depend on normality or even independence of the y_i.) In practice we will not know the value of A, so we cannot use the optimum "linear shrinking rule" $\delta_i = (1 - B)y_i$. However, $E_{\boldsymbol{\theta}} \| y \|^2 = k(A + D)$, so $\| \mathbf{y} \|^2/k$ is an unbiased estimator of $A + D$. This suggests the estimator $\hat{B} \equiv kD/\| \mathbf{y} \|^2$ for B, and therefore the estimation rule

$$\delta_i \equiv \left(1 - \frac{kD}{\sum_1^k y_i^2}\right) y_i, \tag{3.14}$$

which should behave like the optimum linear estimator if $\hat{B}$ is near B. Actually the theory is quite forgiving, and it can be shown that using even a rough guess of B is quite likely to improve on the unbiased estimator, which uses $b = 0$, [6].

The similarity of (3.14) with (2.6) is one of the best arguments for robustness of $\boldsymbol{\delta}^1$ to changes of the underlying assumptions. The choice of the constant $k - 2$ instead of k in (2.6) can be shown to produce a uniform improvement, but this is not the case for the plus-rule version (2.10), see [6]. We won't worry about this fine point any more.

Returning to the sunspot data, let

$$\mu_i \equiv \frac{\phi_{i-1} + \phi_i + \phi_{i-1}}{3}, \qquad \theta_i \equiv \phi_i - \mu_i \tag{3.15}$$

so

$$\phi_i = \theta_i + \mu_i.$$

The $\hat{\mu}_i$ are unbiased estimators for the μ_i, the y_i unbiased estimators for the θ_i. The unbiased, completely unsmoothed, estimator of ϕ_i is $x_i = \hat{\mu}_i + y_i$, which effectively estimates $\hat{\theta}_i$ by y_i. The completely smoothed estimator estimates θ_i by zero and hence ϕ_i by $\hat{\mu}_i$. We can

use (3.14) to obtain estimates of $\hat{\theta}_i$ between 0 and y_i; these can then be employed to improve the estimates of the ϕ_i.

Table III shows the data and the calculations. We are attempting to estimate ϕ_i for the 28 frequencies shown in column 2, beginning with 0.0417 and ending with 0.2051. Column 4 gives the value of x_i for these 28 values, plus the two end values necessary to calculate $\hat{\mu}_1$ and

TABLE III

Compromise Estimates of the Log Power Spectrum, Sunspot Data

i	Freq.	Period	x_i	$\hat{\mu}_i$	$y_i=\hat{\theta}_i$	$\hat{\delta}_i=(1-\hat{B})y_i$	$\hat{\phi}=\hat{\mu}_i+\hat{\delta}_i$
	0.0417	24	4.4				
1	0.0455	22	5.3	5.3	0	0	5.3
2	0.0500	20	6.3	4.8	1.5	0.4	5.2
3	0.0556	18	2.8	5.1	−2.3	−0.7	4.4
4	0.0625	16	6.2	5.2	1.0	0.3	5.5
5	0.0667	15	6.6	6.6	0	0	6.6
6	0.0727	$13\frac{3}{4}$	6.9	5.7	1.2	0.3	6.0
7	0.0784	$12\frac{3}{4}$	3.6	6.2	−2.6	−0.8	5.4
8	0.0843	$11\frac{6}{7}$	8.1	6.7	1.4	0.4	7.1
9	0.0909	11	8.5	8.3	0.2	0.1	8.4
10	0.0976	$10\frac{1}{4}$	8.2	8.2	0	0	8.2
11	0.1053	$9\frac{1}{2}$	7.8	7.4	0.4	0.1	7.5
12	0.1111	9	6.4	7.1	−0.7	−0.2	6.9
13	0.1176	$8\frac{1}{2}$	7.2	6.4	0.8	0.2	6.6
14	0.1250	8	5.7	6.3	−0.6	−0.2	6.1
15	0.1333	7.5	6.1	5.8	0.3	0	5.8
16	0.1403	$7\frac{1}{8}$	5.6	5.1	0.5	0	5.1
17	0.1455	$6\frac{7}{8}$	3.5	4.5	−1.0	0	4.5
18	0.1509	$6\frac{5}{8}$	4.5	3.3	0.8	0	3.3
19	0.1568	$6\frac{3}{8}$	2.5	3.6	−1.1	0	3.6
20	0.1633	$6\frac{1}{8}$	3.7	3.6	0.1	0	3.6
21	0.1686	5.93	4.6	4.8	−0.2	0	4.8
22	0.1739	$5\frac{3}{4}$	6.0	5.5	0.5	0	5.5
23	0.1818	$5\frac{1}{2}$	6.0	5.8	0.2	0	5.8
24	0.1860	$5\frac{3}{8}$	5.4	4.9	0.5	0	4.9
25	0.1905	$5\frac{1}{4}$	3.3	4.2	−0.9	0	4.2
26	0.1951	$5\frac{1}{8}$	4.0	3.5	0.5	0	3.5
27	0.2000	5	3.1	4.1	−1.0	0	4.1
28	0.2051	$4\frac{7}{8}$	5.1	4.5	0.6	0	4.5
	0.2105	$4\frac{3}{4}$	5.2	↑	↑	↑	↑
				$\hat{\mu}_i=(x_{i-1}+x_i+x_{i+1})/3$	$y_i=x_i-\hat{\mu}_i$	$\hat{B}=kD/\|\|y\|\|^2$	Compromise Estimate

$\hat{\mu}_{28}$. The values of $\hat{\mu}_i$ and y_i are given in columns 5 and 6. The θ_i have been estimated from the y_i by the rule (3.14). This has been done separately for the first fourteen coordinates, $i = 1, 2, \ldots, 14$, and for the last fourteen coordinates, $i = 15, 16, \ldots, 28$. The first 14 coordinates give $\hat{B} = 0.71$, hence $\hat{\delta}_i = 0.29y_i$. The second 14 coordinates give an estimate $\hat{B}$ greater than 1. This would lead to a rule which shrinks past the origin, so as before we replace this by $\hat{\delta}_i = 0$. These values are listed in column 7 as $\hat{\delta}_i$. Finally, column 8 gives $\hat{\phi}_i = \hat{\mu}_i + \hat{\delta}_i$, our presumably improved estimates of the ϕ_i.

In this case we have no way of checking whether they really are improvements. Almost certainly they improve on the unbiased estimates x_i, but in this case they are so similar to the completely smoothed estimates it is doubtful whether they are much better than them, except on the obviously badly fitted coordinates $i = 3$ and $i = 7$. (For these two the limited translation rule should also be invoked.)

Writing ϕ_i as $\theta_i + \mu_i$ illustrates a technique for extending the usefullness of the James–Stein estimator. The usual unbiased estimators are used on part of the problem, the μ_i here, and then δ^1 is used to mop up what is left over, the θ_i here. The hope is that the θ_i, being residuals from a smooth model, will be small in magnitude and hence estimated very efficiently by δ^1.

It is usually unwise to pool too many estimation problems together. In [7] we recommend 10–12 as the best value for k. Here we have split the 28 coordinates into 2 groups of 14 in the most obvious way, but more subtle and potentially more advantageous methods of separating and recombining estimation problems are possible, see [5] and [7]. Overpooling relates to objection (ii) of Section 2, the combination of unrelated problems. The usual penalty for doing so is to reduce $\boldsymbol{\delta}^1$ to $\boldsymbol{\delta}^0$, S getting so large it nullifies the $(k - 2)D/S$ term in (2.6).

References

1. T. W. Anderson, "The Statistical Analysis of Time Series," John Wiley, New York, 1971.
2. L. Brown, On the admissability of invariant estimators of one or more location parameters, *Ann. Math. Statist.* **37** (1966), 1087–1136.
3. B. Efron and C. Morris, Limiting the risk of Bayes and empirical Bayes estimators—Part I: the Bayes case, *J. Amer. Statist. Assoc.* **66** (1971), 807–815.
4. B. Efron and C. Morris, Limiting the risk of Bayes and empirical Bayes estimators—Part II: the empirical Bayes case, *J. Amer. Statist. Assoc.* **67** (1972), 130–139.

5. B. EFRON AND C. MORRIS, Empirical Bayes on vector observations—an extension of Stein's method, *Biometrika* **59** (1972), 335–347.
6. B. EFRON AND C. MORRIS, Stein's estimation rule and its competitors—an empirical Bayes approach, *J. Amer. Statist. Assoc.* **68** (1973), 117–130.
7. B. EFRON AND C. MORRIS, Combining possibly related estimation problems, (with discussion), *J. Royal Statist. Soc., B* **35** (1973), 379–421.
8. B. EFRON AND C. MORRIS, Data analysis using Stein's estimator and its generalizations, RAND report R-1394–OED.
9. W. JAMES AND C. STEIN, Estimation with quadratic loss, *in* "Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability," pp. 361–379, University of California Press, CA, 1961.
10. S. SCLOVE, C. MORRIS, AND R. RADHAKRISHNAN, Non-optimality of preliminary test estimators for the mean of a multivariate normal distribution, *Ann. Math. Statist.* **43** (1972), 1481–1490.
11. C. STEIN, Inadmissability of the usual estimator for the mean of a multivariate normal distribution, *in* "Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability," pp. 197–206, University of California Press, CA, 1955.