A NEW DEVELOPMENT IN MULTIVARIATE STATISTICAL ANALYSIS*

T. W. Anderson**

1. Introduction

When I accepted the invitation to address the Japan Statistical Society, I planned to survey the developments in multivariate analysis since the publication of An Introduction to Multivariate Statistical Analysis. However, preparation of the intended survey soon convinced me that this field of statistics had advanced too rapidly to permit a comprehensive outline to be presented in a single lecture. Restriction to even the most significant mathematical advances would be taxing; the general distribution theory of covariance matrices and their latent roots, based on zonal polynomials and hypergeometric functions with matrix arguments, requires considerable background.

My topic is one new development that has great importance for statistics generally, as well as for multivariate analysis. It is "noninvariant estimation" or "biased estimation" due primarily to Charles Stein (Stein [8] and James and Stein [5]). The implications of this approach have been explored by a number of statisticians, including Bradley Efron and Carl Morris; my exposition draws heavily on two of their papers ([2] and [4]).

2. Inadmissibility of the Sample Mean

Suppose $x_{ia}$ is an observation from $N(u_i, \tau^2)$, $a=1, \ldots, N$, $i=1, \ldots, p$, and that all observations are independent. In vector notation $x_a=(x_{1a}, \ldots, x_{pa})'$ is an observation from $N(p, \tau^2I)$, where $p=(\mu_1, \ldots, \mu_p)'$. The sample mean vector

$$\bar{x} = \frac{1}{N} \sum_{a=1}^{N} x_a$$

is a sufficient statistic for $\mu$ when the variance $\tau^2$ is known. The usual or "natural" estimator of $\mu$ is $\bar{x}$, which has the distribution $N[\mu, (\tau^2/N)I]$. The sample mean is unbiased; that is, $E\bar{x} = \mu$. For the $i$-th component $\bar{x}_i$ is the minimum variance unbiased estimator of $\mu_i$; that is, the variance of any other unbiased estimator is at least equal to $\tau^2/N$. The property of "invariance" is that if an arbitrary vector $\nu$ is added to $x_a$, $a=1, \ldots, N$, and to $\mu$, then the estimating procedure is such that the error of estimator

$$\bar{x} - (\mu + \nu)$$

is independent of $\nu$. It seems natural that the statistical error should not depend on the choice of the origin. Another property of the natural estimator is that it is

---

** Stanford University.
maximum likelihood.

For convenience in the remainder of this lecture, we shall define \( y = \bar{x} \) and \( \sigma^2 = r^2/N \). Then \( y \) has the distribution \( N(\mu, \sigma^2 I) \), and we treat the estimation of \( \mu \) with \( \sigma^2 \) known.

A measure of the goodness of an estimator \( \hat{\mu} \) is the sum of the mean square errors of the estimators of the components, namely

\[
\sum_{i=1}^{p} \mathcal{E}(\hat{\mu}_i - \mu_i)^2 = \mathcal{E}(\hat{\mu} - \mu)'(\hat{\mu} - \mu).
\]

In the case of the estimator \( \hat{\mu} = y \) the mean square error is \( p\sigma^2 \). The surprising result of Stein [8] was that there exist estimators with smaller total mean square error than \( y \), at least for large \( p \). In terms of the \( \text{loss function} \) \( (3) \), \( y \) is an inadmissible estimator. An improved estimator is possible by letting the estimator of \( \mu_i \) depend on more than the \( i \)-th component of \( y \); the expected value will depend on \( \mu \). Since \( y_1, \cdots, y_p \) are independently distributed and only the distribution of \( y_i \) depends on \( \mu_i \), it seems that irrelevant information is being used. Stein "explained" the phenomenon by arguing that the sample distance squared from the origin \( y'y \) overestimates the population distance squared \( \mu'\mu \) and hence that the estimator \( y \) could be improved by bringing it nearer to the origin.*

3. The James-Stein Estimator

James and Stein [5] showed that the estimator

\[
\hat{\mu}_{JS} = \left[ 1 - \frac{(p-2)\sigma^2}{S} \right] y,
\]

where \( S = y'y \), for \( p \geq 3 \) is an improvement over \( y \); that is

\[
\mathcal{E}(\hat{\mu}_{JS} - \mu)'(\hat{\mu}_{JS} - \mu) < p\sigma^2
\]

for all \( \mu \). The effect of this procedure is to "shrink" the estimator towards the

<table>
<thead>
<tr>
<th>( \mu'\mu )</th>
<th>( \mathcal{E}(\hat{\mu}<em>{JS} - \mu)'(\hat{\mu}</em>{JS} - \mu) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 10\sigma^2 )</td>
<td>( 10\sigma^2 )</td>
</tr>
<tr>
<td>0.0</td>
<td>.200</td>
</tr>
<tr>
<td>0.5</td>
<td>.478</td>
</tr>
<tr>
<td>1.0</td>
<td>.621</td>
</tr>
<tr>
<td>2.0</td>
<td>.751</td>
</tr>
<tr>
<td>3.0</td>
<td>.824</td>
</tr>
<tr>
<td>4.0</td>
<td>.862</td>
</tr>
<tr>
<td>5.0</td>
<td>.886</td>
</tr>
<tr>
<td>6.0</td>
<td>.903</td>
</tr>
</tbody>
</table>

* \( y'y = \mu'\mu + p\sigma^2 \), \( \text{Var} (y'y) = 4p'\mu + 2p\sigma^2 \).
origin \[S \geq (p-2)s^4\]. The further \(y\) is from the origin (the larger \(S\) is) the less \(y\) is shrunk. Clearly, this procedure is good if \(\mu\) is close to the origin, but it is also better than \(y\) if \(\mu\) is not near the origin.

Is the improvement appreciable? Table 1 for \(p=10\) illustrates that the James-Stein estimator can be considerably better than the natural estimator. For example, when the distance of \(\mu\) to the origin is \(\sqrt{10} = 3.162\) standard deviations, the average mean square error of the James-Stein estimator is 62.1\% of that of the natural estimator.

The origin is an arbitrary point towards which to shrink \(y\). Shrinking towards a vector \(\nu\) is done by the estimator

\[
\nu + \left[1 - \frac{(p-2)s^4}{(y-\nu)\cdot(y-\nu)}\right](y-\nu).
\]

4. Modification in Reduced Dimensions

In many cases it may be reasonable to shrink towards a point that is at least partly determined by the observations. In particular, the shift may be made towards the equiangular line. If \(\bar{y} = \frac{1}{p} \sum_{i=1}^{p} y_i / p\) and \(T = \sum_{i=1}^{p} (y_i - \bar{y})^2\), an estimator of \(\mu_i\) is

\[
\bar{y} + \left[1 - \frac{(p-3)s^4}{T}\right](y_i - \bar{y}).
\]

An example may be taken from a game popular in Japan and the United

<table>
<thead>
<tr>
<th>i</th>
<th>Name</th>
<th>Spring Batting Average</th>
<th>Season Batting Average</th>
<th>James-Stein Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Roberto Clemente</td>
<td>0.400</td>
<td>0.346</td>
<td>0.293</td>
</tr>
<tr>
<td>2</td>
<td>Frank Robinson</td>
<td>0.378</td>
<td>0.298</td>
<td>0.289</td>
</tr>
<tr>
<td>3</td>
<td>Frank Howard</td>
<td>0.356</td>
<td>0.276</td>
<td>0.284</td>
</tr>
<tr>
<td>4</td>
<td>Jay Johnstone</td>
<td>0.333</td>
<td>0.221</td>
<td>0.279</td>
</tr>
<tr>
<td>5</td>
<td>Ken Berry</td>
<td>0.311</td>
<td>0.273</td>
<td>0.275</td>
</tr>
<tr>
<td>6</td>
<td>Jim Spencer</td>
<td>0.311</td>
<td>0.270</td>
<td>0.275</td>
</tr>
<tr>
<td>7</td>
<td>Don Kessinger</td>
<td>0.289</td>
<td>0.263</td>
<td>0.270</td>
</tr>
<tr>
<td>8</td>
<td>Luis Alvarado</td>
<td>0.267</td>
<td>0.210</td>
<td>0.265</td>
</tr>
<tr>
<td>9</td>
<td>Ron Santo</td>
<td>0.244</td>
<td>0.269</td>
<td>0.261</td>
</tr>
<tr>
<td>10</td>
<td>Ron Swoboda</td>
<td>0.244</td>
<td>0.230</td>
<td>0.261</td>
</tr>
<tr>
<td>11</td>
<td>Del Unser</td>
<td>0.222</td>
<td>0.264</td>
<td>0.256</td>
</tr>
<tr>
<td>12</td>
<td>Billy Williams</td>
<td>0.222</td>
<td>0.256</td>
<td>0.256</td>
</tr>
<tr>
<td>13</td>
<td>George Scott</td>
<td>0.222</td>
<td>0.304</td>
<td>0.256</td>
</tr>
<tr>
<td>14</td>
<td>Rico Petrocelli</td>
<td>0.222</td>
<td>0.264</td>
<td>0.256</td>
</tr>
<tr>
<td>15</td>
<td>Ellie Rodriguez</td>
<td>0.222</td>
<td>0.226</td>
<td>0.256</td>
</tr>
<tr>
<td>16</td>
<td>Bert Campaneris</td>
<td>0.200</td>
<td>0.285</td>
<td>0.251</td>
</tr>
<tr>
<td>17</td>
<td>Thurman Munson</td>
<td>0.178</td>
<td>0.319</td>
<td>0.247</td>
</tr>
<tr>
<td>18</td>
<td>Max Alvis</td>
<td>0.156</td>
<td>0.200</td>
<td>0.242</td>
</tr>
</tbody>
</table>
Efron and Morris [4] treated batting averages of American baseball players who had exactly 45 times at bat on a certain spring day. The observation \( y_i \) is the batting average of the \( i \)-th player on that day, \( i=1, \cdots, 18 \). The "true batting average" of each player is unknown, but for purposes of the illustration, we let \( \mu_i \) be the batting average for the entire season of the \( i \)-th player. The variance \( \sigma^2 \) is unknown, but can be estimated as \( \bar{y}(1-\bar{y})/45 \), where \( \bar{y}=.265 \). The estimate for the \( i \)-th player is

\[
0.265 + 0.212(y_i - 0.265).
\]

The figures are given in Table 2. For convenience the numbering of the players is in descending order of the springtime batting average.

It will be seen that \( y_i \) tends to be more extreme than \( \mu_i \), and \( \hat{\mu}_i \) tends to be closer to \( \mu_i \) than \( y_i \). The sum of the squared errors of the "natural" estimate and of the modified estimate are

\[
\sum_{i=1}^{18} (y_i - \mu_i)^2 / \sigma^2 = 17.7, \quad \sum_{i=1}^{18} (y_i - \hat{\mu}_i)^2 / \sigma^2 = 5.0.
\]

The latter is a considerable reduction of the former.

### 5. Comments

The baseball example suggests that there are elements of a Bayesian approach. In fact, if \( \mu \) is given a prior normal distribution with mean \( \nu \) and covariance proportional to \( I \), the Bayes estimator is (6) with \((p-2)\sigma^2\) replaced by a constant depending on \( \sigma^2 \) and the variance of the prior distribution. The estimator can also be developed in terms of an "empirical Bayes" procedure (Robbins [7]) or as a solution to a compound decision problem (Robbins [6]).

The estimator (4) can be improved by setting it equal to 0 if the factor multiplying \( y \) is negative. This estimator is not admissible, but it seems reasonable to believe that further possible improvement is very small.

In the discussion to the paper by Efron and Morris [3], Stein gave a more general approach to the problem. The estimator (4) is then a special case.

Another estimation problem in multivariate analysis is estimating the covariance matrix \( \Sigma \) on the basis of a sample of \( N \) from \( N(\mu, \Sigma) \). The usual or natural estimator is the sample covariance matrix \( S \). In unpublished work Stein has shown that \( S \) is inadmissible in terms of a loss function that is invariant with respect to linear transformations. Analogous to shrinking \( y \) towards \((\bar{y}, \cdots, \bar{y})'\) as an estimator of \( \mu \), distortion of \( S \) towards \((\text{tr} \, S/p)I \) improves the estimation of \( \Sigma \).

The full effect of this approach on multivariate analysis and indeed on statistics in general has yet to be worked out. That it will have an important impact is evident.
REFERENCES


