CHANCE

ISSN: 0933-2480 (Print) 1867-2280 (Online) Journal homepage: https://amstat.tandfonline.com/loi/ucha20

A Statistician Reads the Sports Pages

Phil Everson

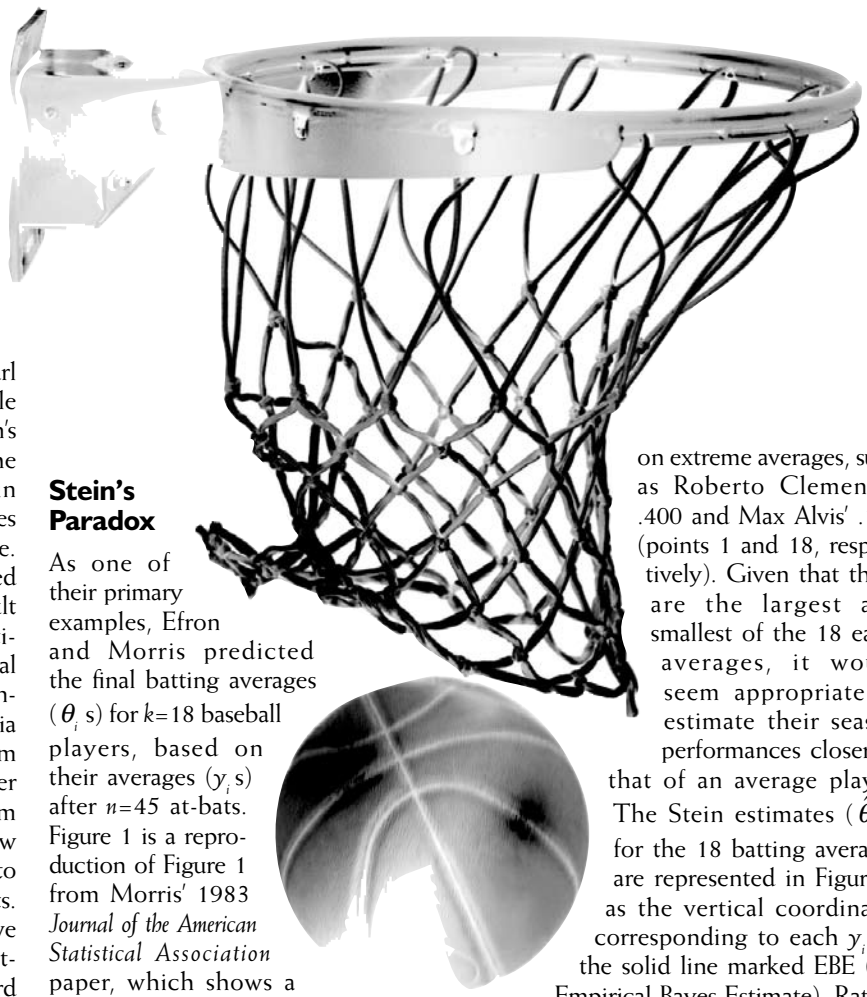
To cite this article: Phil Everson (2007) A Statistician Reads the Sports Pages, CHANCE, 20:3, 49-56, DOI: 10.1080/09332480.2007.10722861

To link to this article: https://doi.org/10.1080/09332480.2007.10722861

A Statistician Reads the Sports Pages

Phil Everson,
Column Editor

Stein's Paradox Revisited



Thirty years ago, Bradley Efron and Carl Morris published a beautifully readable *Scientific American* article titled "Stein's Paradox in Statistics." It describes the "James-Stein" estimate, introduced in 1961 by Charles Stein and Willard James and often referred to as the Stein estimate. The 1977 Efron-Morris article generated increased interest in this important result and demonstrated how the Stein estimate can be a powerful tool for statistical inference. After its publication, a then-colleague of Willard James at California State University, Long Beach named Jim Stein said he considered getting a rubber stamp that said, "I am James Stein, but I am neither James nor Stein; however, I know James and will pass your request on to him" and stamping it on reprint requests. This year alone, several colleagues have mentioned Stein's paradox as an interesting idea in statistics they had just heard about. To celebrate the 30th anniversary of the *Scientific American* article, and to help continue spreading the word about this fascinating result, I will review the Efron-Morris baseball example and give an illustration of Stein's paradox using data from the most recent NBA season. I regularly show similar examples in my mathematical statistics courses and make the connection between Stein's paradox, Bayesian hierarchical models, and regression to the mean.

Stein's Paradox

As one of their primary examples, Efron and Morris predicted the final batting averages (θ_i s) for k=18 baseball players, based on their averages (y_i s) after n=45 at-bats. Figure 1 is a reproduction of Figure 1 from Morris' 1983 *Journal of the American Statistical Association* paper, which shows a plot of the final player averages (excluding the first 45 at-bats) against the averages after 45 at-bats. Stein's paradox is that the early player averages $y_1,\ldots,y_{18}$ are not the best estimates of the final averages $\theta_1,\ldots,\theta_{18}$. Improvements are made by shrinking the y_i s together, in effect using the performances of the other players to modify each player's estimate $\hat{\theta}_i$.

Looking at Figure 1, this may not seem so paradoxical. Think, in particular, of making predictions based on extreme averages, such as Roberto Clemente's .400 and Max Alvis' .156 (points 1 and 18, respectively). Given that these are the largest and smallest of the 18 early averages, it would seem appropriate to estimate their season performances closer to that of an average player. The Stein estimates ($\hat{\theta}_i$ s) for the 18 batting averages are represented in Figure 1 as the vertical coordinates corresponding to each y_i on the solid line marked EBE (for Empirical Bayes Estimate). Rather than estimating each θ_i by y_i, as is represented by the line with slope 1.0 marked CLASSICAL, the $\hat{\theta}_i$ s fall on a line with a slope of about 0.2, and are roughly 80% closer to the average of the y_i s. Figure 1 resembles a linear regression example, although the problem here is to fit a line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ based only on the x s. The Stein estimate fits a line through the point $(\overline{x}, \overline{x})$, with slope $0 \le \hat{\beta} < 1$.

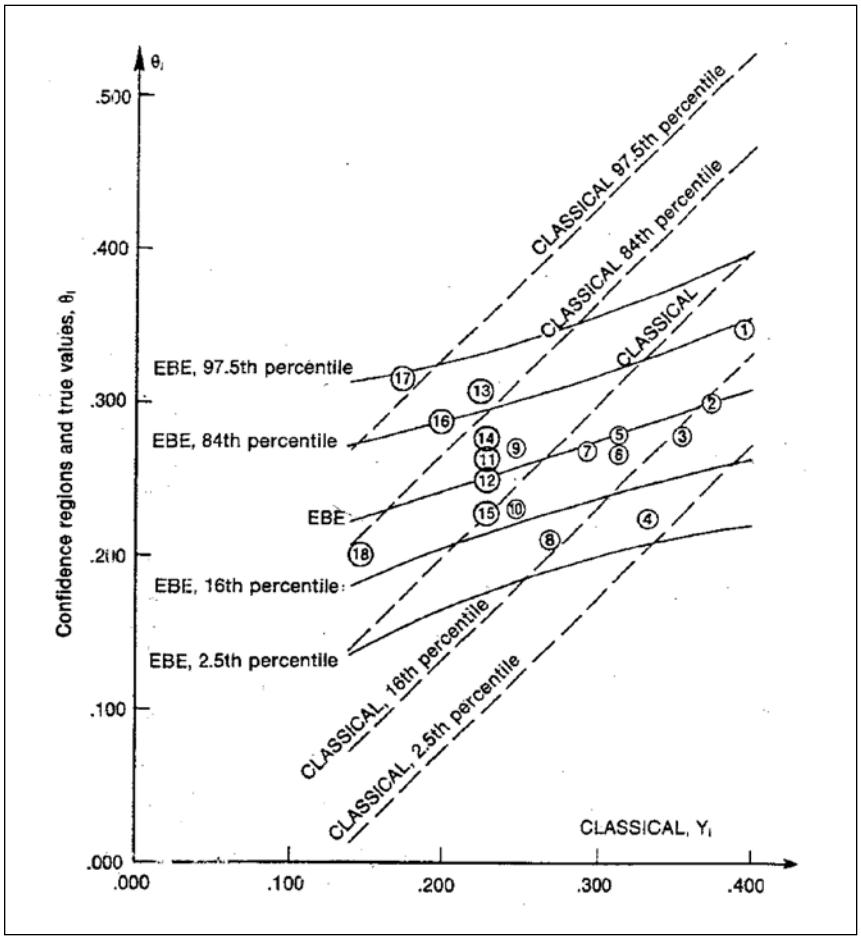


Figure 1. Eighteen points plotted at (Y_i, θ_i). CLASSICAL= $Y_i, Y_i \pm V^{1/2}$, *and* $Y_i \pm 1.96V^{1/2}$ *(dashed lines).* $EBE = \hat{\theta}_i, EBE \pm s_i$, *and* $EBE + 1.96\, s_i$ (solid curves)

Graph courtesy of *JASA*, March 1983

What makes Stein's result more paradoxical is his general statement that, when estimating a collection of k (possibly unrelated) means using a set of independent sample averages, the averages are inadmissible. That is, there is another estimate that does better for every possible set of means. The Stein estimate was shown to be such an estimate. Efron and Morris illustrate the paradoxical nature of this fact by including a nineteenth average in their collection: the percentage of a random sample of 45 foreign-made automobiles. Now, they consider estimating the 18 player means and the overall percent of foreign-made cars. Stein's theorem applies to the problem of estimating these k=19 values as well as it applies to the original k=18. So the Stein estimate of the percentage of foreign-made cars is influenced by the batting performances of Clemente and the others. One could be forgiven for thinking this is nonsensical.

Most of the confusion comes from thinking the Stein estimate improves on every average y_i as an estimate of its mean θ_i. It doesn't. For estimating a single mean, the average is optimal (and therefore admissible). Stein's problem involves estimating k>2 means simultaneously, using the sum of squared errors from the true means for the k estimates as a measure of the overall error ($SSE = \sum(\hat{\theta}_i - \theta_i)^2$). The Stein estimate tends to produce a smaller sum of squared error than the k averages (i.e., it has lower risk). In a way, this is no more paradoxical than regression to the mean. If one was to regress the final player averages on the averages after 45 at-bats, the least squares fitted values would be fewer standard deviations from the average $\bar{y}$ than the predictor values. That is, the $\hat{\theta}_i$ s would vary less about $\bar{y}$ than do the y_i s. This also would be true if we included the true proportion of foreign cars and the sample proportion for a sample of 45 cars. One could legitimately argue that the second regression model isn't appropriate without an indicator to signal that this last individual is fundamentally different from the other 18. But with a simple regression model, the fitted value for the percent of foreign cars would be influenced by the batting averages. The Stein estimate essentially attempts to fit this regression model without having observed the responses, but assuming that they will be centered about the same mean as the predictors.

Efron and Morris often say the Stein estimate "anticipates regression to the mean" and improves the estimates by "borrowing strength from the ensemble." In the context of the baseball players, the ensemble (the collection of 18 players with 45 at-bats on a specific date in 1970) provides information about the distribution of batting averages in Major League Baseball that year. If we observed only one player's average, it would be equally likely to be an over-estimate or an under-estimate of the player's true probability of getting a hit. But in the context of other similar players, we might reasonably judge a particular batting average is more likely to be an over-estimate or under-estimate. For example, Clemente's .400 was the largest of a sample of 18 batting averages. Without any outside knowledge that .400 is an unusually high average for a professional player, one could still judge that this was likely to be an over-estimate of Clemente's true probability. A classical confidence interval for Clemente's mean would be symmetric about .400, allowing for the possibility that he had a true probability larger than .400 and was in a batting slump for those first 45 at-bats. The Stein estimate predicts Clemente to bat closer to the overall average $\bar{y}$ =.265, but still to be above average. Similarly, the Stein estimate predicts Max Alvis' final average to be higher than .156, but still below average.

The overall average value $\bar{y}$ provides a context for the individual estimates. If a

National Football League (NFL) quarterback completed 40% of his first 45 pass attempts, the classical confidence interval would be the same as for a baseball player batting .400. But 40% would be far below the NFL average for completion percentages, so in this case, the Stein estimate for an observed value of 40% would be larger than 40%, but still below the NFL average of about 60%. The May 21, 2007, issue of *Sports Illustrated* reported that Jack Cust hit six homeruns in his first 26 at-bats (23%) with the Oakland Athletics this season, after hitting only five in his other 144 major league at-bats over the past six years. I am very confident that Cust won't hit 23 homeruns in his next 100 at-bats.

The Stein estimate makes sense because the y_is vary about their underlying θ_i s, so the overall variability in the y_is will typically be larger than the variability in the θ_i s. To anticipate this, each y_i is shrunk toward some common value, and the most obvious value to shrink toward is $\bar{y}$, the average of the y_is. The Stein estimate can be written as follows:

$$\hat{\theta}_i = \hat{B}\bar{y} + (1-\hat{B})y_i, \quad i=1,\ldots,k,$$

where y_i is the observed average for the ith individual, $\bar{y}$ is the average of the y_is, and $\hat{B}$ is a value between 0 and 1 estimated from the data. Stein actually defined the credibility $c=1-\hat{B}$, and Efron and Morris use this notation in their 1977 article. In later papers, Morris defines $\hat{B}$ as above, and I am more comfortable with this notation. Because $\bar{y}$ depends on observations that do not all vary about a particular mean θ_i, each $\hat{\theta}_i$ is a biased estimate for the corresponding θ_i. The idea is to introduce directed bias in the direction of the overall mean in order to reduce the total variability of the k estimates.

If $\hat{B}=1$, then all k of the means θ_i are estimated by the average value $\bar{y}$. This would happen if the y_is were very close to each other, suggesting they all may be estimating the same value. In the context of baseball, this would indicate the players were all of equal ability and that the variation in their batting averages might be explained easily by chance alone. Efron and Morris estimated $\hat{B}=0.788$ ($c=0.212$), so the observed averages were shrunk nearly 80% of the way toward $\bar{y}$ =.265. For example, the prediction for Clemente was 0.788(.265)+0.212(.400) =.294, and for Alvis 0.788(.265)+0.212 (.156)=.242. These players went on to bat .345 and .200, respectively. The SSE for the classical prediction is $\sum (Y_i - \theta_i)^2 = 0.077$. For the Stein estimate, it is $\sum(\hat{\theta}_i - \theta_i)^2 = 0.022$, a reduction of about 70% in the total prediction error.

A Normal Hierarchical Model for NBA Scoring

Stein did not assume a probability model for the underlying means $\theta_1,\ldots,\theta_k$, but by doing so, one can more easily see why Stein's formula for the shrinkage factor $\hat{B}$ makes sense. Consider the $k=30$ National Basketball Association (NBA) teams. Before November 21, 2006, each team had played about 10 games of the 2006–2007 NBA regular season. For now, I'll act as though it was exactly $n=10$ games each. The average points per game y_i for team i in its first 10 games is (perhaps) an unbiased estimate of the mean points per game in that team's remaining 72 games. The y_is ranged from 89.3 points per game for Miami to 108.0 points per game for Phoenix, with an average of $\bar{y}$ =98.4. Because 72 isn't excessively large, I make a correction in the interval estimates for a possible deviation between the average of these 72 scores and the true mean scoring rate θ_i for a particular team i, $i=1,\ldots,30$.

These are all NBA teams playing in the same season, so it seems likely that the team means (θ_is) would be similar, with only a few very high or low values. A normal model is a good description for scoring in basketball, and it also seems a reasonable description for how the team θ_is would vary. This suggests a two-level normal hierarchical model, summarized by what Morris calls "the big picture."

The distributions on the left side show the two-level structure, with the estimates Y_i varying about their respective θ_i s, and with the θ_i s also varying about some overall mean μ. The right side shows the marginal distributions of the Y_i s and the conditional (posterior) distributions of each θ_i given $Y_i=y_i$. The parameter V represents the variance of Y_i, the average points scored by team i in n games, with $n=10$ in the example. This model allows the team means θ_i to vary with variance A for different teams, but assumes a constant standard deviation σ for individual game outcomes, with $V=\sigma^2/n$ for all $k=30$ teams. The pooled estimate of σ is about $\hat{\sigma}=10.47$ points. The sample standard deviations s_i for the 30 teams ranged from 5.9 points to 15.5 points. This range is not that large for $k=30$ estimates based on equal standard deviations and $n=10$ observations each, so the assumption of equal σ s is not unreasonable. There are roughly 30(10–1)=270 degrees of freedom for estimating σ, and σ is treated as "known" in the Stein estimate. The variance for the average score in 10 games for the same team is then $V=10.47^2/10 \approx 11.0$, or $(3.31)^2$. Stein refers to my V as σ^2, but in this problem, I thought V would be less confusing as the variance for a sample average, with σ^2 the variance of individual game scores. Morris also uses V this way in the caption for Figure 1.

The parameters μ and A represent the mean and variance of the $k=30$ unknown θ_i s. With this explicit Normal model for the θ_i s, we can work out the marginal distribution for the sample averages Y_i displayed in the upper-right corner of "the big picture:"

$$\begin{aligned} Y_i &= \theta_i + \sqrt{V}Z_{y_i} \\ &= \mu + \sqrt{A}Z_{\theta_i} + \sqrt{V}Z_{y_i} \overset{\text{i.i.d.}}{\sim} N(\mu, V+A) \\ & Z_{y_i}, Z_{\theta_i} \overset{\text{i.i.d.}}{\sim} N(0,1). \end{aligned}$$

The Big Picture

$Y_i \mid \theta_i, \mu, A \overset{\text{indep}}{\sim} N(\theta_i, V).$	$Y_i \mid \mu, A \overset{\text{indep}}{\sim} N(\mu, V+A), \; i=1,\ldots,k.$
$\theta_i \mid \mu, A \overset{\text{indep}}{\sim} N(\mu, A).$	$\theta_i \mid \mu, A, y_i \overset{\text{indep}}{\sim} N(B\mu+(1-B)y_i, (1-B)V),$ $B=V/(V+A), \quad V=\sigma^2/n.$

The Y_i s and θ_i s are centered about the same mean μ. The θ_i s have variance A, and the sample averages Y_i have marginal variance $V+A$, reflecting the added variability in each estimate Y_i about θ_i. The marginal covariance between Y_i and θ_i is A, the variance of the component common to both. This is true even without Normal errors, which were not assumed for Stein's result. But with the Normal assumption, it is a standard probability problem to derive the posterior distribution of θ_i, given μ, A, and an observed value $Y_i=y_i$, displayed in the lower right corner of "the big picture." The Stein estimate mimics the formula for the posterior mean of θ_i, $E(\theta_i \mid y_i, B, \mu) = B\mu+(1-B)y_i$ by replacing μ and B with sample estimates $\bar{y}$ and $\hat{B}$. In this sense, it is an "empirical Bayes" estimate (EBE).

The true value of B depends on the variance A of the underlying θ_i s, and determines the weight put on the overall mean μ. If A is large (relative to V), then the θ_i s are very dispersed and B will be small, meaning most of the weight is put on the observed average $\bar{y}$. We could say there is relatively little information about θ_i in the other y_j s, compared to what we learn from y_i. If A is small, the θ_i s will be more similar, and the posterior mean has more weight on the overall mean μ.The conditional mean formula also reflects the formula for the mean in a linear regression model. The correlation between y_i and θ_i is $\rho = \sqrt{A/(V+A)} = \sqrt{1-B}$. So, the regression mean for predicting θ_i from y_i is $\beta_0 + \beta_1 y_i$ with

$$\beta_1 = \rho \frac{\sigma_{\theta_i}}{\sigma_{y_i}} = \sqrt{1-B}\,\frac{A^{1/2}}{(V+A)^{1/2}} = 1-B$$

and

$$\beta_0 = \mu_{\theta_i} - \beta_1 \mu_{y_i} = \mu - (1-B)\mu = B\mu.$$

The square of the correlation is $1-B = A/(V+A)$. The posterior variance of θ_i is $(1-B)V = BA$, which is smaller than A (the marginal variance of the θ_i s) by a proportion $1-B$. So, this agrees with the notion of R^2 as the proportion of variance explained by a regression.

The average of the Y_i s is unbiased for μ, and the sample variance of the Y_i s is unbiased for $V+A$. With V assumed known, an unbiased estimate for $B=V/(V+A)$ is

$$\hat{B} = \frac{(k-3)V}{\sum (Y_i - \bar{Y})^2},$$

and this is also Stein's formula for $\hat{B}$. The numerator has a $k-3$, rather than a $k-1$ to adjust for taking the inverse of the sum of squares. Stein's theorem, in fact, applies to any case with $k>2$. The factor $k-3$ includes an adjustment for estimating μ by $\bar{y}$ in the sum of squares, and necessitates $k>3$ to produce any shrinkage. With $k=3$, the sum of squares must be taken relative to some fixed value μ (e.g., $\mu=0$) that does not depend on the y_i s. Then, $k-3$ may be replaced by $k-2$ in the formula for $\hat{B}$. With $k=30$, the adjustment hardly matters. As Efron and Morris point out, the Stein estimate can be improved by setting $\hat{B} = 1$ whenever $\sum (y_i - \bar{y})^2 < (k-3)V$. Estimating $\hat{B} > 1$ would imply a negative variance for the θ_i s and give negative weight to each observed y_i. A small sum of squares suggests the θ_i s are not very different, and setting $\hat{B} = 1$ results in the common estimate $\bar{y}$.

For the NBA data, $\sum (y_i - \bar{y})^2 = 674.7$, and $\hat{B} = (30-3)(11.0)/674.7 = 0.44$. This corresponds to $\hat{A} = V(1-\hat{B})/\hat{B} = (11.0)(1-0.44)/0.44 = 14.0$, so the estimated standard deviation of the team θ_i s is $\sqrt{14.0} \approx 3.7$ points. If the team θ_i s were equal (H_o: $A=0$), then the marginal variance of each Y_i would be V, and $X^2 = \sum (y_i - \bar{y})^2 / V \sim \chi^2_{(k-1)}$. The observed value $X^2 = 674.7/11.0 = 61.3$ is unusually large for a chi-square variable with $k-1=29$ degrees of freedom ($p<0.001$). So, it would not be appropriate to estimate all the θ_i s by $\bar{y}$. The Stein estimate provides a compromise between that and using the y_i s as estimates. The shrinkage factor $\hat{B} = (k-3)/X^2$ reflects the degree of evidence for variability in the θ_i s. A larger value of X^2 indicates more evidence for $A>0$ and leads to a smaller $\hat{B}$ (meaning more weight is put on the observed y_i s) and a larger $\hat{A}$. If we had found $X^2 \le k-3$, then the χ^2 test for H_o: $A=0$ would not have been significant ($p>0.5$) and we would estimate $\hat{B}=1$ and $\hat{A}=0$.

Figure 2 shows the actual average point totals for each of the $k=30$ NBA teams in the games played on November 21, 2007, or later, plotted against the averages for games played that season before November 21. Unlike with the Efron-Morris baseball example, the team with the largest early average—the Phoenix Suns with 108.0 points per game (indicated by a 1 in Figure 2)—had an even more extreme late average of 110.5. I was disappointed when I first constructed this graph, hoping for a picture as dramatic as that in Figure 1 (which has happened in previous NBA seasons). But having decided at the beginning of the 2006–2007 season to use this example for my column, I didn't want to change it now just to dredge for something more impressive. And this is an important reminder that the Stein estimate will not necessarily give a better estimate for any particular observation—even for the most extreme values from a set of data.

Despite the digression from the mean by Phoenix, the overall pattern of variation in Figure 2 is similar to that in Figure 1. Of the five largest early averages, three of those teams did have lower averages for the remainder of the season. And the teams with the five lowest early averages all increased their averages for the remainder of the season. For example, Miami had the lowest early average at 89.3. The overall average is $\bar{y}=98.4$ and $\hat{B}=0.44$, so the Stein estimate for Miami is $(0.44)(98.4)+(1-0.44)(89.3) = 93.3$, which is much closer to Miami's actual final average of 95.3. And most importantly, the total sum of squared prediction errors (SSE) for the Stein estimate is about 33% smaller than the SSE using the $k=30$ team averages.

The darkest dashed line, with a slope of 1.0, represents using the observed averages to estimate the final averages. The darkest solid line represents the Stein estimate, with a slope of $1-\hat{B} = 0.56$. The least squares line has a slope of 0.50 and is plotted as the dashed line very close to the Stein estimate. The least squares line is chosen to minimize the SSE. The Stein estimate attempts to do the same thing, but without knowledge of the response values. The SSE using the y_i s is 512.8. The SSE for the Stein estimate is 346.0, which is about 33% smaller and only about 2% larger than the SSE for the least squares fit of 339.5. The Stein estimate is nearly the best linear estimate possible in this example, even though it did not make use of the late-averages to help estimate B and μ.

Interval Estimates

Figures 1 and 2 display interval estimates and point estimates for each team's final scoring average. The classical intervals (dashed lines) use the observed averages ±1.00 or ±1.96 prediction standard errors as 68% and 95% confidence intervals for the final averages. This is valid, assuming each final average will be distributed about the same mean θ_i as the corresponding observed average (which also is assumed in the hierarchical model). Considered individually, each estimate is equally likely to overestimate or underestimate its true mean, which justifies using symmetric intervals.

The Stein intervals (solid lines) are symmetric about the line representing the Stein estimate, which slopes more gradually than the classical estimate. The formula for the prediction standard deviation used for the Stein intervals takes some explanation, but what is important to notice is that they achieve roughly the same coverage as the classical estimates, but with narrower intervals. By combining information from all $k=30$ observations, the Stein estimates "borrow strength" to achieve greater precision.

The intervals for the Stein estimates are based on the joint posterior distribution of the θ_is given all the y_is. If the mean μ and shrinkage factor B were known, the posterior variance would simply be $(1-B)V$. We can easily adjust for the estimation of μ if we are willing to assume a flat prior density:
$p(\mu \mid A) \propto 1, \quad -\infty < \mu < \infty.$

I think of this as saying, "even though I may believe some values are more likely than others to be the true mean scoring rate in the NBA (e.g., positive values), let's assume *a priori* that all values of μ are equally likely and see what comes out when we condition on the observed data." Flat prior distributions are not always a good idea, but there is a long history of using them when estimating the mean of a Normal distribution. Extending the range of $p(\mu)$ to $\pm\infty$ is a mathematical convenience. The posterior density for μ is proportional to the prior density $p(\mu)$ multiplied by the likelihood function $L(\mu)$. The likelihood function is the joint density for all the y_is, treated as a function of μ, and it goes to 0 very quickly for values of μ far from $\bar{y}$. So, it does not matter to the posterior density whether $p(\mu)$ remains constant in the tails or tapers off to 0 in a way that would make it a proper density. Assuming a constant density $p(\mu)$ allows the data to determine which region of values is most relevant and yields a proper posterior density for the θ_is as long as $k>3$.

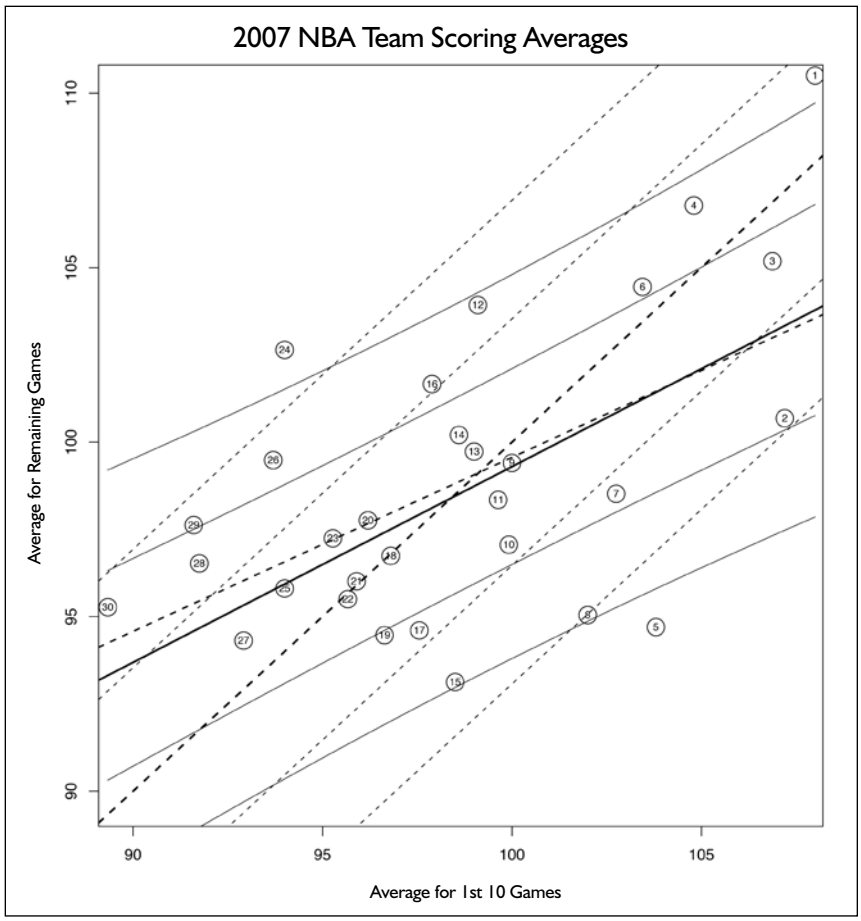


Figure 2. Each point corresponds to one of the $k=30$ NBA teams playing in the 2006–2007 season. The horizontal axis represents the average points y_i scored by each team in games played before November 21 (about $n=10$ games for each team), and the vertical axis represents the average points θ_i scored in the remaining games. The dashed lines have slope 1.0 and represent "classical" estimates and 68% and 95% confidence intervals for the θ_is. The solid lines represent the Stein estimate (the darkest solid line) and 68% and 95% prediction intervals for the θ_is. The dashed line close to the Stein estimate is the least squares fit of the θ_is on the y_is.

With this prior specification, the posterior distribution for μ is $\mu \mid B, \bar{y} \sim N(\bar{y}, V/(kB))$. The posterior variance is found by inverting the sum of the information (inverse variance) about μ from the y_is:

$$Var(\mu \mid A, \bar{y}) = \left(\sum_{i=1}^{k} (V+A)^{-1} \right)^{-1}$$
$$= \frac{V+A}{k} = \frac{V}{kB}; \quad B = \frac{V}{V+A}.$$

The posterior distribution for each θ_i is still Normal after integrating out μ. The adjusted mean and variance are found using the laws of total expectation and variance:

$$E(\theta_i \mid B, y) = E(E(\theta_i \mid B, \mu, y) \mid B, y)$$
$$= B\bar{y} + (1-B)y_i = y_i + (y_i - \bar{y})B.$$

$$Var(\theta_i \mid B, y) = E(\theta_i \mid B, \mu, y) \mid B, y)$$
$$+ Var(E(\theta_i \mid B, \mu, y) \mid B, y)$$
$$= E((1-B)V \mid B, y)$$
$$+ Var(B\mu + (1-B)\mu_i \mid B, y)$$
$$= (1-B)V + B^2 \frac{V}{kB}$$
$$= \left(1 - \left(\frac{k-1}{k}\right)B\right)V.$$

In his 1983 *JASA* paper, Morris makes a further adjustment for the uncertainty in the estimated shrinkage factor B. Let B^* and v^* represent the posterior mean and variance of B, given only the y_is.

Then, we can write:

$$E(\theta_i \mid y) = E(E(\theta_i \mid B, y) \mid y) = y_i + (y_i - \bar{y})B^*.$$

$$Var(\theta_i \mid y) = E(Var(\theta_i \mid B, y) \mid y) + Var(E(\theta_i \mid B, y) \mid = \left(1 - \left(\frac{k-1}{k}\right)B^*\right)V + v^*(y_i - \bar{y})^2.$$

The second term in the expression for the variance reflects the greater uncertainty in predictions for observations further from the mean. This explains the bowing in the Stein intervals in Figures 1 and 2 and mimics the regression confidence intervals for a mean response. Morris estimates B^* by $\hat{B}$ and v^* by $2\hat{B}^2/(k-3)$, which is an unbiased estimate for the sampling variance of $\hat{B}$. The prediction standard deviation is the square root of the estimated posterior variance, and the Stein (EBE) interval estimates in Figures 1 and 2 are made by taking the Stein estimates ±1.00 and ±1.96 estimated posterior standard deviations.

Assuming a constant standard deviation $\sigma = 10.47$ points for individual games for each team, the standard deviation for the average of $n = 10$ games is $V^{1/2} = 10.47/\sqrt{10} = 3.31$. To account for the uncertainty in the 72 remaining games, I compute

$$\sqrt{10.47^2/10 + 10.47^2/72} = 3.53$$

as the prediction standard deviation for the classical estimates. In comparison, the estimated posterior standard deviations for the Stein estimates (which also include the correction for 72 games) range from 2.51 for an observation right at $\bar{y} = 98.4$, to 3.03 for the most extreme observation (Phoenix). So, the classical intervals are from 17% to 41% wider than the Stein intervals.

Unequal Variances

In most practical situations, the variances V_i of k estimators will not all be equal. Even if the individuals in different groups have roughly the same (known) value of σ, different sample sizes n_i will produce different variances $V_i = \sigma^2/n_i$ for the Y_is. For example, the NBA teams had, in fact, played between 7 and 12 games before November 21, so $V_i = (10.47)^2/n_i$ ranges from $(3.02)^2$ to $(3.96)^2$. There is now no single shrinkage factor B for the k estimates. A larger n_i means a smaller V_i and, therefore, a smaller B_i, with more weight put on the observed value y_i. A smaller n_i means a larger V_i and B_i, and more weight on the estimated mean $\hat{\mu}$. This variable shrinkage may result in $\hat{\theta}_i$s that follow a different ordering than the Y_is, as is the case in the Efron-Morris example of estimating toxoplasmosis rates in El Salvador, described in the same 1977 *Scientific American* article.

With unequal V_is, the posterior distribution for θ_i is Normal with mean $B_i\mu + (1-B_i)y_i$ and variance $(1-B_i)V_i$, where $B_i = V_i/(V_i + A)$.

The elegant conditional structure of the Normal model makes iterative sampling techniques such as MCMC easy to implement for estimating μ and A in this problem. However, in a 1983 *JASA* article, Morris presents a simple algorithm for finding a method of moments estimate that gives a very good approximation and takes virtually no time to evaluate. It is based on the marginal distribution of the Y_is:

$$Y_i \mid \mu, A \overset{\text{indep}}{\sim} N(\mu, V_i + A), \quad i = 1, \ldots, k.$$

If we standardize each Y_i and sum the squares, we get a $\chi^2_{(k)}$ random variable. Fitting μ by its maximum likelihood estimate $\hat{\mu}$ reduces the degrees of freedom to $k-1$. The MLE $\hat{\mu}$ is the weighted average of the Y_is, using weights proportional to $w_i = (V_i + A)^{-1}$, the inverse variance. Averages based on a larger n_i have a smaller V_i, and therefore get more weight when estimating μ.

$$\sum_{i=1}^{k} \frac{(Y_i - \hat{\mu})^2}{V_i + A} = \sum_{i=1}^{k} w_i (Y_i - \hat{\mu})^2 \sim \chi^2_{(k-1)};$$

$$w_i = (V_i + A)^{-1}, \quad \hat{\mu} = \sum w_i Y_i / \sum w_i.$$

The algorithm works by equating the standardized sum of squares (computed for a given estimate of A) to its expected value, the degrees of freedom $k-1$:

$$\sum_{i=1}^{k} w_i (Y_i - \hat{\mu})^2 = k - 1 = \left(\frac{k-1}{k}\right)\sum_{i=1}^{k} w_i (V_i + A) = \left(\frac{k-1}{k}\right)\left(\sum_{i=1}^{k} w_i V_i + A\sum_{i=1}^{k} w_i\right).$$

Each $(Y_i - \hat{\mu})^2$ provides an unbiased estimate for $V_i + A$ when inflated by the factor $k/(k-1)$. Subtracting V_i from each and taking a weighted average gives a new estimate for A. The algorithm starts with an initial guess $A^{(0)} > 0$, and then updates from $A^{(j)}$ to $A^{(j+1)}$ at each iteration. For the current estimate $A^{(j)}$, the corresponding weights and mean estimate are $w_i^{(j)} = (V_i + A^{(j)})^{-1}, i = 1, \ldots, k$, and $\hat{\mu}^{(j)} = \sum (w_i y_i) / \sum w_i$. The updated estimate $A^{(j+1)}$ is then:

$$A^{(j+1)} = \frac{\sum\left(w_i^{(j)}\left(\left(\frac{k}{k-1}\right)(Y_i - \hat{\mu}^{(j)})^2 - V_i\right)\right)}{\sum w_i^{(j)}}.$$

The algorithm tends to converge very quickly. For example, starting with $A^{(0)} = 20$, we have convergence to five decimal places in four iterations:

j	$A_{(j)}$	$\hat{\mu}_{(j)}$
0	20.00000	98.41704
1	11.74735	98.41704
2	11.71905	98.41386
3	11.71892	98.41384
4	11.71892	98.41384
5	11.71892	98.41384

The shrinkage factor for each team i is now computed as $\hat{B}_i = V_i/(V_i + \hat{A})$, setting $\hat{A} = 11.71892 = (3.42)^2$. This produces shrinkage factors that range from $B_i = 0.44$ for teams with $n_i = 12$ games to $B_i = 0.57$ for the teams with $n_i = 7$ games. Replacing V and B by V_i and B_i in the formula for the posterior variance leads to adjusted interval estimates for predicting the true θ_is.

Discussion

As with any method, the Stein estimate and its generalizations perform as they should only if the underlying assumptions are met. The Stein estimate assumes the k averages are independent with a constant variance V. The interval estimates are based on the stronger assumptions of the Normal hierarchical model, and departures from these can lead to systematic errors in the estimates. For example, the mean μ for NBA scoring might change as the season goes on and teams have played more games together. Overall, the average for the 2,000 or so games after November 20, 2007, was less than 0.4 points per game higher than in

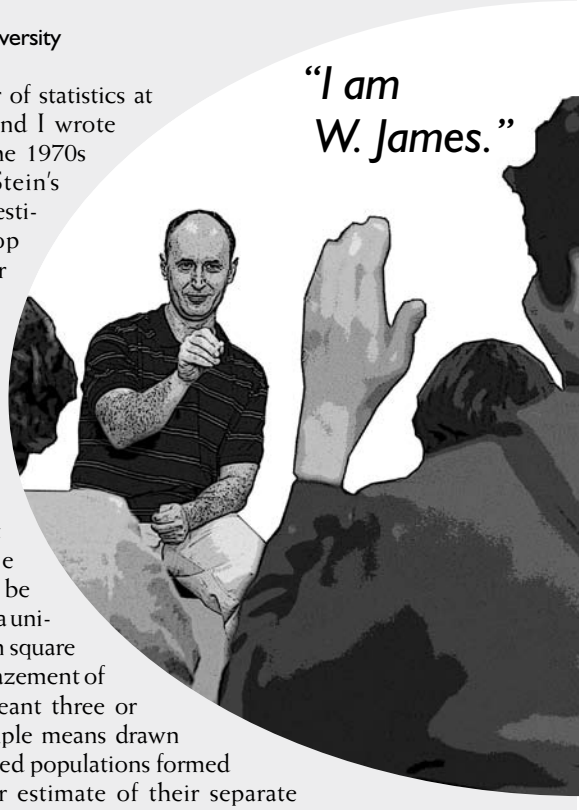
the first 300 games, and this difference is not statistically significant. On the other hand, part of Miami's unusually low scoring in their early games might be partly explained by injuries to star center Shaquille O'Neil during that time. And while the NBA season is balanced overall, the particular opponents each team faced in their first 10 or so games also might explain a lot of the variability in scoring, just as the particular pitchers faced by the 18 batters in their first 45 at-bats may have been important in the Efron-Morris baseball example. Also, there is a statistically significant difference of about 3.0 points in mean scoring between games played at home and games played on the road. Incorporating these factors into the probability model could improve predictions for any particular game or set of games.

The original Stein estimate is limited by the restrictive assumptions of a single mean for all observations and a constant and known variance V. Morris (1983) provides the generalizations needed to incorporate covariates and allow for different (but still known) V_is. Iterative sampling techniques such as MCMC can easily accommodate additional complications, such as unknown variances or more than two levels of variation. These models are important for research in education, health care, and many other fields. It is worth examining the simple model that yields the Stein estimate as a way of building up to these more complicated models and to help understand what happens when many parameters are estimated simultaneously.

It is a worthwhile exercise to simulate data from a Normal hierarchical model with whatever parameter values you choose. Then, compare the performance of the Stein estimate to that of an MCMC algorithm and see how the results differ, and whether one gives systematically better results. Remember the performances may depend on the particular choice of parameters. Quick approximations such as the Stein estimate are valuable for calibrating the results for a more complicated (and hopefully more accurate) inference procedure. For example, we might allow the standard deviation of game outcomes σ to vary by team, and estimate these from the teams' sample standard deviations. Accounting for this additional uncertainty will typically widen the interval estimates for the θ_is. But the posterior

Finding W. James of James-Stein Renown

Carl Morris, Harvard University



Brad Efron, a professor of statistics at Stanford University, and I wrote a series of papers in the 1970s to explain Charles Stein's pioneering shrinkage estimators and to develop extensions needed for wider applications. In 1961, the James-Stein estimator—so called because it appeared in Stein's famous 1961 paper with W. James—was Stein's newest and simplest example showing that independent sample mean estimates could be combined to guarantee a uniformly lower total mean square error (MSE). To the amazement of all statisticians, this meant three or more independent sample means drawn from normally distributed populations formed an inadmissible vector estimate of their separate population means.

That sensational paper had statisticians wondering and asking who W. James was. His first name was unavailable and he had left Fresno State, the institution listed on the paper. After trying for 16 years to identify James and to find out what "W" stood for—especially when exhorted by the *Scientific American* editor for our 1977 article about Stein's paradox—Efron and I felt we might never locate him.

Then W. James appeared—in dramatic fashion. It has been 30 years, yet the memory is vivid (occasional retelling has reinforced it).

James' Unveiling

Stein's estimator and its empirical Bayes underpinnings were exciting and new ideas when the Southern California Chapter of the American Statistical Association and its then president, Bob Newcomb, invited me to give an after-dinner talk at its meeting in Los Angeles on January 26, 1978. The ASA had circulated the title in advance: Estimating Many Parameters: The James-Stein Rule and Its Generalizations. An audience of about 50 was in high spirits after dinner. Lights were dimmed as I introduced my topic and identified Stein and his work. Then, I offered—ruefully—that statisticians didn't know who James was.

A middle-aged man at a rear table called out, "I do!" I could see him only dimly, but I still felt the chill and the premonition that surged through me during the eerie pause that preceded my asking, "Who?"

"I am."

For the next few moments, we conversed one-to-one across the room. Distracted by his appearance, I occasionally would mutter—even during the talk—my amazement that he had appeared. The statistics world finally

Continued on next page

knew his name. Willard D. James was on the California State University, Long Beach (CSU-LB) mathematics faculty at the time. As a mathematician whose statistics research had been limited to one summer for Stein, he had not kept track of the paper or its full impact. He told the audience he was embarrassed that the estimator Stein discovered was called the James-Stein estimator, and he asked that the "James" be removed to give Stein proper credit. Here are some highlights, mostly learned from our longer, private conversation later that night.

Only a remarkable coincidence brought my ASA talk to James' attention. The CSU-LB mathematics faculty included James Stein (who is still there). A colleague who spotted the ASA talk announcement with "James-Stein" in the title asked James Stein if that was his work. James Stein said no, but it was that of Willard James, who was down the hall. So Willard James learned of my ASA talk. And he came.

James, a 1957 PhD in mathematics from the University of Illinois and a research assistant in Stanford's mathematics department, met Stein at the statistics department's lunchtime board games. He told Stein in 1959 that he was looking for summer research support. James had the then unusual ability to program a computer, and Stein asked him to compute a numerical table of the MSE of his newest and simplest shrinkage estimator—the estimator Stein eventually presented at the Fourth Berkeley Statistics Symposium in 1961. Stein already knew this estimator offered substantial improvements in practice, but he needed its numerical MSE to convince disbelieving statisticians.

James accepted Stein's offer. However, disdaining tedious computer programming, James first sought a simpler mathematical expression and produced the elegant MSE formula in Section 2 of the James-Stein paper. James thought Stein's extraordinary generosity made James a coauthor, listed first for alphabetical order. To absolve James from responsibility for conclusions he had not been party to, Stein noted there that joint work ended at Section 2.

Many statisticians, before knowing the MSE of that new estimator, regarded the sample mean's inadmissibility as a mere mathematical curiosity. D. V. Lindley, speaking about Stein's inadmissibility result when Stein presented his 1962 *Journal of the Royal Statistical Society* discussion paper, said, "When I first heard of this suggestion several years ago, I must admit that I dismissed it as the work of one of these mathematical statisticians who are so entranced by the symbols that they lose touch with reality." Lindley went on to acknowledge the substantial MSE improvement of the new estimator over the sample mean and its potential practical importance.

James re-emphasized to me later that the estimator should be called Stein's estimator, not James-Stein. I said I understood, but that the double name helped to distinguish that estimator from others of Stein's. He said he still felt proud of his MSE contribution, and I affirmed its value.

I asked at night's end whether James ever wrote another statistics paper. He did not. After reminding him that this paper with Stein might be the most famous and influential statistics paper of its era, I warned, "If you ever write another statistics paper, your average surely will go down." I've always loved the irony of this prophecy. It is predicted by the James-Stein estimator itself, which works because it anticipates and accounts for regression-toward-the-mean.

Postscript

Recent detective work by *CHANCE* columnist Phil Everson revealed that James is retired, but still lives in Long Beach. That led, after writing this story, to my contacting him for the first time in 30 years. He clarified certain details that I've edited, but the story here is of my memory. He also confirmed that he never wrote another statistics paper.

mean estimates for the θ_is should not differ too much from the Stein estimates, with the V_is estimated and treated as known. If there are large differences in the results of the two methods, it may be because the differing assumptions changed the inference substantially. But often, it is due to coding errors that commonly arise with more complicated techniques, and which may go unnoticed without comparing to another method. No matter how well a model describes some real-life process, don't believe too strongly that you are, in fact, explaining the process. As George Box famously said: "All models are wrong, but some are useful."

Author Notes

Charles Stein is professor emeritus in statistics at Stanford University. Willard James retired from California State University, Long Beach in 1987 and still lives in Long Beach. Carl Morris is professor of statistics at Harvard University, where he was my thesis advisor (making me Stein's grandstudent). Just about everything I've written here is something I once heard Morris say, and I'm gradually figuring out what he meant. I am very grateful for the conversations we have had over the years, and particularly for those regarding this column. Bradley Efron is the Max H. Stein Professor of Statistics at Stanford University and was recently awarded the National Medal of Science. The 1977 Efron-Morris *Scientific American* article will be included in a volume of Efron's papers to celebrate his 70th birthday. The data for the Efron-Morris baseball example and my NBA example are available from *www.swarthmore.edu/NatSci/peverso1*.

Further Reading

James, W. and Stein, C. (1961). Estimation with Quadratic Loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, Vol.1.

Efron, B. and Morris, C.(1975). Data Analysis Using Stein's Estimator. *Journal of the American Statistical Association*. 70(350): 311–319.

Efron, B. and Morris, C. (1977). Stein's Paradox in Statistics. *Scientific America*. 236(5): 119–127.

Morris, C. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association*. 78(381): 47–55.

Column Editor: Phil Everson, *peverso1@swarthmore.edu*