Spring 2018

James-Stein Estimation in Baseball

Isaac Matejin
iam15@zips.uakron.edu

University of Akron

Honors Senor Thesis in Statistics

James-Stein Estimation in Baseball

Author:

Isaac Matejin

Advisor:

Dr. Richard Einsporn

Submitted to the University of Akron's

Honors College in Partial Fulfillment of the

Degree of Bachelors of Science

March 23, 2018

Abstract

Isaac Matejin

Major: Statistics-Actuarial Sciences

Project Sponsor: Dr. Richard Einsporn

Number of Project Credits: 3

James-Stein Estimation in Baseball

The purpose of this essay will be to examine the effectiveness of the James-Stein estimator, and its application to baseball and the saber metrics movement. This report will utilize James-Stein theorem and average prediction error to predict the batting averages, slugging percentage, and OPS with only the first 50 at bats of the 2017 season. Also, this report will provide graphs, charts, and data to prove that the James-Stein estimator is an improvement of the MLE in terms of expected loss.

1 Introduction

From the development of the box score to sabermetrics, statistics and sports have gone hand in hand. The analytics movement along with modern technology has allowed for statistical theories and formulas to be applied to baseball. The projection of performance is useful to teams and fans alike as it sheds light into the outcome of a team's performance.

In the 1900's if one was to estimate an unknown parameter θ, one would turn to a maximum likelihood estimator, or MLE. In a 1977 edition of Statistical Science, John Aldrich explained how R.A. Fisher developed the procedure of finding the value of one or more parameters for a given statistic which makes a known likelihood distribution a maximum (Aldrich 1977). But in 1956, Charles Stein proved to the statistical world that for $k \geq 3$, the maximum likelihood estimator is inadmissible. Stein later developed an estimator for $k \geq 3$ that was applied to baseball in the 1970's, around the same time Billy James developed his sabermetric formula.

My research consists of reviewing prior methods and applying them in an area that usually isn't examined. I will be using the classic Stein Estimator example. Efron and Morris, in 1975, provided he classic example consists of taking a player's batting average in the first 50 at bats to project the season totals using the original estimator. I will then be applying the classic Stein example of player's batting averages and then to slugging percentage. When researching I found that there was no application of the Stein's estimator to slugging percentage. I will also be looking at the application of Stein's estimator for on-base plus slugging percentage as it is a hybrid/combination of the batting average and slugging percentage. Before delving into the experiment, a good place to start is to thoroughly examine the variables and data.

2 Variables and Data

Baseball data is the most complete data among sports due to the stop and start aspect along with a success or failure Binomial aspect. My data comes from a database called Baseball-Reference.com. It contains box scores and season statistics for player's dating back to 1914. The data was processed in Microsoft Excel and my shrinking constant was computed in R. The code used in this project is provided in Appendix 4.

The data being used are the season totals of batting average, slugging percentage, and on-base plus slugging percentage, or OPS, for right fielders for the 2017 season. I then also utilized the first 50 at bats during their 2017 campaign. An average baseball player has 550 at bats in a given season. I chose right fielders due to the variability of the data to challenge the estimator's effectiveness, giving me a total of 30 data points to estimate. I choose to use the top 30 plate appearances for right fielders because of the variation in the data which will test the effectiveness and ability of the Stein estimator. For example, Giancarlo Stanton of the Miami Marlins started the season slow but led the National League in slugging percentage by the season's end. Another example is Aaron Judge of the New York Yankees, whose numbers lowered after the All-Star break. In choosing players who play right field, I believe that the data will test the Stein estimator as well as possibly correctly projecting the season totals with only the first 50 at bats. Due to the nature of baseball, the data are dependent because different players use shared information against certain opponents. For example, Aaron Judge might pick up a technique to use against Corey Kluber when watching film of Bryce Harper vs Corey Kluber. Although this phenomenon is a commonplace in sports, I do not believe that it will affect my data or my experiment.

Section 3 of my project details the methodology and theory behind my experiment. Its subsections explain the different layers within Stein's estimators, from the mathematical proofs from Stein (1961) to simplified explanations as seen in Efron and Morris (1977). These different parts will fit together by the end of the section; and in Section 4, I will talk about the results of my experiment before concluding on which statistic Stein's estimator predicted the most effectively.

3 Methodology and Prior Findings

Charles Stein changed how estimators were looked at with his proof of his estimator as an improvement over the traditional estimators on the maximum likelihood estimator in terms of total square error risk. In 1975, Bradley Efron and Carl Morris applied Stein's estimator to baseball batting averages to use a shrink each X_i toward the $\bar{X}$ by a shrinking factor c. In this section, I will present the work of both Stein and Efron & Morris to build a foundation for my experiment.

3.1 Stein (1956)

More than 60 years ago, Stein (1956) published his breakthrough paper "Inadmissibility of the usual estimator for the mean of a Multivariate Normal Distribution." Stein (1956) describes the multivariate problem and the gives a practical, geometric argument intended to convince that the usual estimator should be inadmissible if the dimension is sufficiently large. If we observe the real random variable $X_1, \ldots, X_n$ independently and normally unknown vector of means $\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$ and a variance equal to 1, it is customary to estimate θ by X itself. Stein (1956) considered the problem of estimating θ with the loss function L given by

$$L(\theta, d) = (\theta - d)^2 = \sum(\theta_i - d_i)^2 \tag{1}$$

where d is a vector of estimators. For n≥3, Stein (1956) proposes the estimator θ_1(X) given by

$$\theta_1(\mathrm{X}) = \left(1 - \frac{b}{a+X^2}\right)\mathrm{X} \tag{2}$$

With a,b>0. Stein (1956) proved that for a sufficiently small b and large a, θ_1 has an improvement over the risk of usual estimator. An improvement of risk means the expected loss is improved upon. The complete proof can be found in [3].

3.2 Stein (1961)

Stein followed up his discovery of the inadmissibility of the usual estimator for n≥3 in 1956 with another discovery in 1961. When an estimator is deemed inadmissible, this only means that there is loosely another estimator that is uniformly better in terms of risk. Stein (1961) published "Estimation with Quadratic Loss." Stein, along with his graduate assistant Willard James, continued on the work Stein gave the statistical world in 1956. Using the loss function in (1), and the usual estimator θ_0, defined by

$$\theta_0(X) = X, \tag{3}$$

and its risk is

$$\rho(\theta, \theta_0) = E\left[L\big(\theta, \theta_0(X)\big)\right] = E(X-\theta)'(X-\theta) = p. \tag{4}$$

With p≥3,

$$E\left\|\left(1-\frac{P-2}{\|X\|^2}\right)X-\theta\ \right\|^2 = p - E(\frac{(p-2)^2)}{p-2+2K} < p, \tag{5}$$

where K has a Poisson Distribution with mean $\frac{\|\theta\|^2}{2}$. Thus, the estimator θ defined by

$$\theta(X) = (1-\frac{(p-2)}{\|X\|^2})X \tag{6}$$

has a smaller risk than θ_0 for all θ. In conclusion, Stein (1961) proved that for p≥3, the use of his estimator for the risk makes a uniform improvement on the maximum likelihood estimator (MLE) in terms of total square error risk. The complete proof and findings of Stein (1961) can be found in [4].

3.3 Efron and Morris (1975)

Building on Stein's work, Efron and Morris incorporated Stein's formula into the baseball world. Efron and Morris (1975) published "Data Analysis Using Stein's Estimator and its Generalizations." The following excerpt is from the paper and is crucial in setting up my experiment.

> To review the James-Stein estimator in the simplest setting, suppose that for given θi
>
> $$X_i|\theta_i \sim N(\theta_i, 1),\ i = 1, \dots, k \geq 3, \tag{7}$$
>
> meaning the {Xi} are independent and normally distributed with mean $EX_i \equiv\ \theta_i$ and variance Varθ_i(Xi) = 1. The unknown vector of means $\theta_i \equiv (\theta_1, \dots, \theta_k)$ is to be estimated with loss being the sum of squared component errors
>
> $$L\big(\theta, \hat{\theta}\big) \equiv \sum(\theta_i - d_i)^2, \tag{8}$$
>
> where $\hat{\theta} \equiv (\hat{\theta}_1, \dots, \hat{\theta}_k)$ is the estimate of θ. The MLE, which is also the sample mean, $\delta^0(X) \equiv X \equiv (X_1, \dots, X_k)$ has constant risk k,
>
> $$R(\theta, \delta^0)\ \equiv\ E_\theta \textstyle\sum_{i=1}^{k}(X_i - \mu_i)^2 = k, \tag{9}$$
>
> E_θ indicating expectation over the distribution (7). James and Stein [4] introduced the estimator $\delta^1(X) = (\delta_1^1(X), \dots, \delta_k^1(X))$ for k ≥3,

$$\delta_1^1(X) \equiv \mu_i + \left(1 - \frac{k-2}{S}\right)(X_i - \mu_i), i = 1, \ldots, k \qquad (10)$$

with $\mu_i \equiv (\mu_i, \ldots, \mu_k)'$ any initial guess at θ and $S \equiv \sum(X_i - \mu_i)^2$. This estimator has risk

$$R(\theta, \delta^1) \equiv E_\theta \sum_{i=0}^{k} (\delta_i^1(X) - \theta_i)^2 \qquad (11)$$

$$\leq k - \frac{(k-2)^2}{k-2+\sum(\theta_i - \mu_i)^2} < k, \qquad (12)$$

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

The complete application of Stein's estimator in baseball can be found in [5].

4 Experiment Results

4.1 Estimator for Batting Average

The application of James-Stein theorem to baseball indicates that the J-S $\hat{\mu}$ will be an improvement over the MLE $\hat{\mu}$. If we consider the batting averages of the 30 Major League players that were observed over the 2017 season, my goal is to predict Total BA from the first 50 at bats. The full data can be seen in Appendix 1. If we consider the outcome of an at bat, it can be classified into two outcomes; success or failure. Therefore, we can model each player's first 50 at bats with a binomial distribution,

$$x_i \sim Binomial(50, \theta_i). \qquad (13)$$

Here θ_i is the true average, or how a given player would perform over an infinite number of at bats. In essence every column of the data in Appendix 1 is its own binomial distribution, as each column has a success or failure component.

Since our n in the Binomial distribution is greater than 30, we can use a Normal Approximation to (1),

$$x_i \sim Normal(\theta_i, \sigma_0^2), \qquad (14)$$

where σ_0^2 is the binomial variance,

$$\sigma_0^2 = \bar{p}(1 - \bar{p})/50, \qquad (15)$$

with $\bar{x} = .263$ as the average of the averages. Letting $y_i = x_i/\sigma_0^2$, and applying it to (10), gives the estimator

$$JS = \bar{x} + [1 - \frac{(K-3)\sigma_0^2}{S}](x_i - \bar{x}). \qquad (15)$$

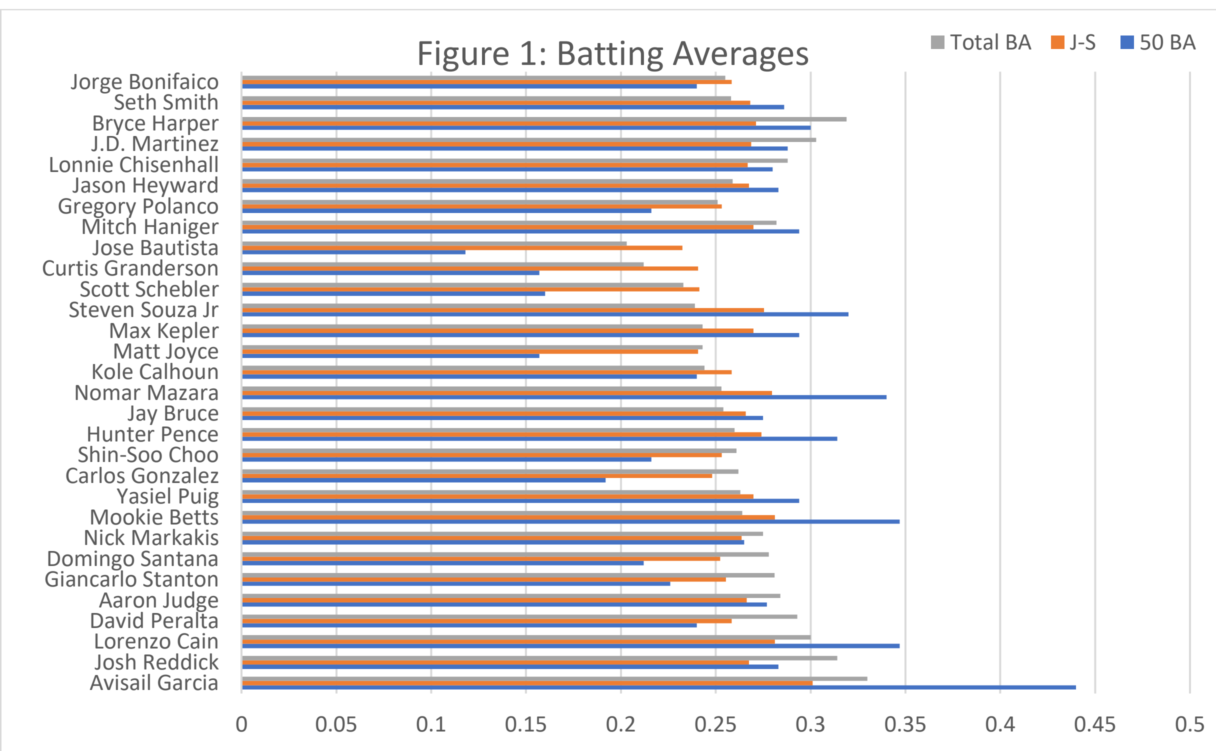
Setting our shrinking factor equal to $1 - \frac{(K-3)\sigma_0^2}{S}$, we get a linear function

$$\text{JS} = \bar{x} + c(x_i - \bar{x}). \qquad (16)$$

The formula in (15) is used in Appendix 1 to calculate an estimate for Major Leaguer's batting averages. To show that the J-S $\hat{\mu}$ will be an improvement over the MLE $\hat{\mu}$, we can look at the average prediction error of both the J-S $\hat{\mu}$ and MLE $\hat{\mu}$ compared to the player's average for the 2017 season. In (6) below, we can see that the J-S $\hat{\mu}$ will be an improvement over the MLE $\hat{\mu}$, and therefore is the better estimator for predicting player's batting average for the 2017 season based on the first 50 at bats.

$$\sum_{i=1}^{30} \frac{|50\ BA_i - Total\ BA_i|}{30} = .045 \text{ while } \sum_{i=1}^{30} \frac{|JS_i - Total\ BA_i|}{30} = .02. \qquad (17)$$

Figure one below is a horizontal bar chart comparing the batting averages for the 30 Major League players, their batting averages for the first 50 at bats, and the James- Stein estimate of their batting averages. Looking at the bar chart, we can see that the estimator seems like a good estimator as it almost correctly estimates the averages for both Jorge Bonifacio and Matt Joyce. One interesting aspect of the averages is that the estimator worked better for Avisail Garcia, whose batting average was .440 (the highest among right fielders) after 50 at bats, then Bryce Harper, whose batting average was .300. This can be a result of personal aspects that a model or estimator can not determine. Bryce Harper is a perennial all-star player who was able to bat .319 despite battling a hamstring injury the second half of the season. In Efron and Morris (1977), they had the same problem with Roberto Clemente, a hall of fame player in the 1960' and 1970's. Overall, I believe my James-Stein estimator compares nicely to Efron and Morris' as both have a smaller average prediction error. I will be discussing some solutions in my data with the player's slugging percentage. The next section will look at trying to come up with a decent estimator for slugging percentage, as there is no James-Stein estimator on a common and useful statistic in baseball.



4.2 Estimator for Slugging Percentage

In baseball, slugging percentage is a measure of a hitter's productivity at the plate. It is calculated as total bases divided by at bats, through the following formula, where AB is the number of at-bats for a given player, and 1B, 2B, 3B, and HR are the number of singles, doubles, triples, and home runs, respectively:

$$\text{SLG} = \frac{(1B)+(2*2B)+(3*3B)+(4*HR)}{AB} \tag{18}$$

Unlike batting average, slugging percentage gives more weight to extra-base hits such as doubles and home runs, relative to singles. To give prospective of this counting statistic, the slugging percentage average was .417 for the 2017 MLB season. The highest slugging percentage possible is 4.000, but only achievable if a player hits a homerun with every at bat. In the MLB, a great slugging percentage is .600 while .400 is a bad slugging percentage. I will be taking the slugging percentage of the first 50 at bats for the same the baseball players and use the James-Stein estimator to predict the respective slugging average for the 2017 MLB season.

We can now look back at the James-Stein estimator given below,

$$JS = \bar{x} + [1 - \frac{(K-3)\sigma_0^2}{S}]\,(x_i - \bar{x}).$$

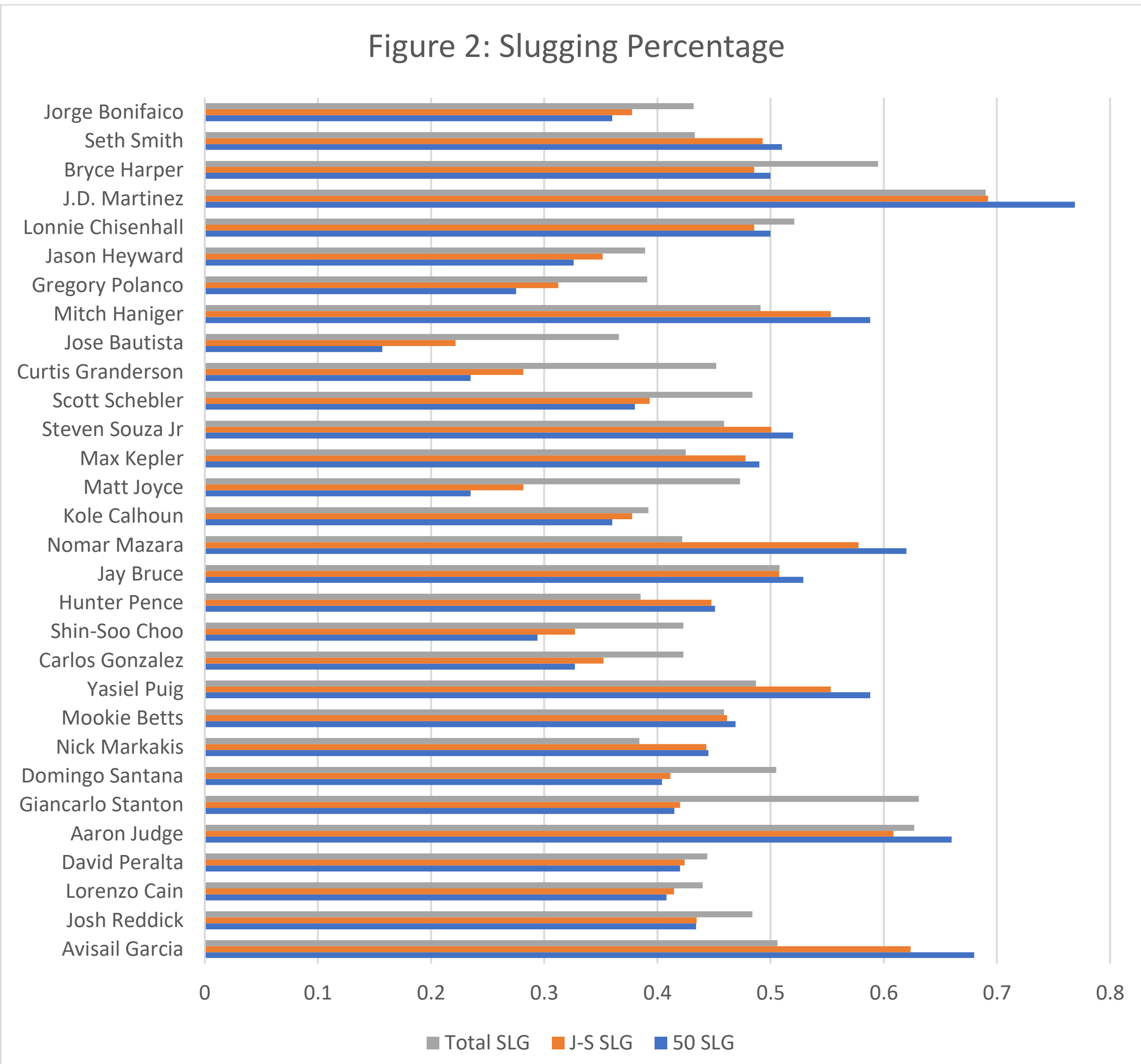
Setting our shrinking factor equal to $1 - \frac{(K-3)\sigma_0^2}{S}$, we get a linear function

$$\text{JS} = \bar{x} + c(x_i - \bar{x}).$$

Plugging the slugging percentage values into the formula above gives us the slugging percentage after 50 at bats, the slugging percentage estimate through the James-Stein formula, and the slugging percentage for the entirety of the 2017 season which can be found in Appendix 2.To show that the J-S $\hat{\mu}$ will be an improvement over the MLE $\hat{\mu}$, we can look at the average prediction errors of both the J-S $\hat{\mu}$ and MLE $\hat{\mu}$ and compare it to the player's average for the 2017 season. In (#) below, we can see that the J-S $\hat{\mu}$ will be an improvement over the MLE $\hat{\mu}$, and therefore is the better estimator for predicting player's slugging percentage for the 2017 season based on the first 50 at bats.

$$\sum_{i=1}^{30} \frac{|50\ BA_i - Total\ SLG_i|}{30} = .095 \text{ while } \sum_{i=1}^{30} \frac{|JS_i - Total\ SLG_i|}{30} = .073. \tag{19}$$

Looking above, we can see that the James-Stein estimator is an improvement over the mean of the MLE. Although the James-Stein estimator for the slugging percentage is an improvement in terms of estimated loss, it doesn't have quite the improvement, in terms of estimated loss, that the James-Stein estimator for the batting averages had from the previous section. Looking at the shrinking factor in Appendix 2, we can see that c=.769. With a shrinking factor of .769, the James Stein only shrinks each $x_i - \bar{x}_i$ by about 25% of the difference. Looking at Figure 2 below, we can see that while the James-Stein estimator is adequate for the players who have an average slugging percentage; the estimator for the extremes are not that adequate. For example, Avisail Garcia the estimator is .118 points off the total slugging percentage. This is probably because Garcia had a great April at the plate but tapered off as the year went on. But in another instance, the estimator is correct in its estimate of J.D. Martinez. Martinez was an outlier and had a slugging percentage of .769 after 50 at bats. Instead of regressing to the mean of the slugging percentages, Martinez kept the torrid pace up and finished with the 2017 season with a slugging percentage of .690. The James-Stein formula gave an estimator if .692 for the 2017 season, which is only .002 points off. Possible reasoning for Martinez's success is that he was he played in a hitter's paradise, Chase Field in Arizona, and that he was in a contract year. This phenomenon is proved in an article published in the Fall 2014 Baseball Research Journal by Heather M. O'Neill [8]. O'Neill proved that players have career years in contract years due to the incentives of a lucrative contract. In conclusion, the James-Stein estimator is effective but not as effective as the estimator for batting average. In the next subsection I will be looking at the on-base plus slugging percentage, or OPS, for my data.



4.3 Estimator for On-Base Plus Slugging Percentage

OPS is the sum of a players on-base percentage and slugging percentage to get a more complete baseball statistic. It's meant to combine how well a hitter can reach base, with how well he can hit for average and for power. As a result, OPS is widely considered one of the best evaluative tools for hitters. To give perspective, if a hitter has an OPS over 1.000, that is considered phenomenal. Likewise, an OPS of under .700 is not good as it indicates a poor ability to reach base and hit for power. Even though OPS gives weight to homeruns and doubles, like the slugging percentage, it can still be modeled by a binomial distribution. Even though the data isn't really a binomial distribution, we can still use the Normal Approximation. Using the Normal Approximation, as our sample size is greater than 30, from the previous section and applying it to (10), gives the estimator

$$JS = \bar{x} + [1 - \frac{(K-3)\sigma_0^2}{S}]\,(x_i - \bar{x}).$$

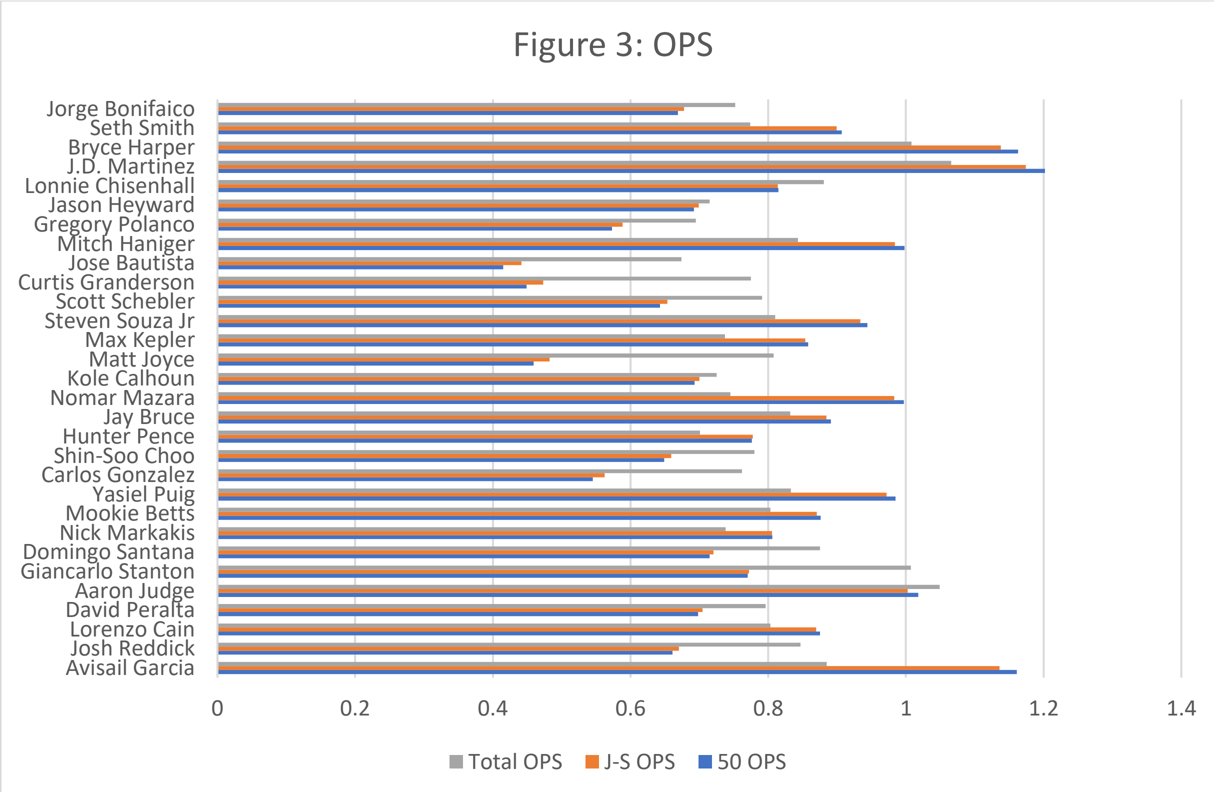
Setting our shrinking factor equal to $1 - \frac{(K-3)\sigma_0^2}{S}$, we get a linear function

$$\text{JS} = \bar{x} + c(x_i - \bar{x}).$$

Plugging OPS into the formula above gives us the Stein estimator for OPS which can be found in relation to the outcome in Appendix 4. Looking at the average prediction errors of both the J-S $\hat{\mu}$ and MLE $\hat{\mu}$ and compared to the player's average, J-S $\hat{\mu}$ should give us an improvement over the MLE.

$$\sum_{i=1}^{30} \frac{|50\,BA_i - Total\ OPS_i|}{30} = .1443 \text{ while } \sum_{i=1}^{30} \frac{|JS_i - Total\ OPS_i|}{30} = .1338. \qquad (20)$$

Looking below, we can see that the James-Stein estimator is an improvement over the mean of the MLE. Although the OPS James-Stein estimator is a good estimator, but it is not as good of an estimator as the James-Stein estimator for the batting averages from the sub-section 1. Looking below at Figure 3, we can see that the James-Stein estimator only partially shrinks the 50 OPS column values. Looking at our shrinking factor, c= .931, we can see that the estimator only shrinks the $(x_i - \bar{x})$ by about 10 percent. In Figure 3, if the observation for the 50 at bats is within .100 of the Total OPS for that player, then the Stein estimator is adequate. Looking at Kole Calhoun and Shin Soo Choo's OPS in Appendix 3, they both have an OPS after 50 at bats close to their season total. This allows for a shrinking factor of .931, as an improvement of 10 percent puts their Stein estimator within reasonable range of their actual outcome. If we look at Avasail Garcia and Matt Joyce, their OPS after 50 at bats were extremes, 1.161 and .451 respectively. Naturally, both of their OPS's regress to the mean. But with a shrinking factor that only shrinks by 10 percent, Figure 3 below shows that Stein's estimator is not adequate for Garcia and Joyce. This also indicates that the Stein estimator might not be an improvement over the MLE. In conclusion, although our test shows that the J-S $\hat{\mu}$ is an improvement over the MLE $\hat{\mu}$, it is not a practical estimator.



5 Discussion

My results are an interesting look into in what circumstances the James-Stein estimator works with baseball data. It is apparent that by looking at the average prediction error, the James-Stein estimator reduced the expected loss the greatest when it was used in the batting average data. Although the James-Stein estimator reduces each $x_i$to the overall average, I think that if the James-Stein estimator took the first 50 at bats of the 2017 season and regressed it to the career average of each individual player; the estimator would better predict the season averages of a player's batting average, slugging percentage, and OPS.

Another interesting application of James-Stein estimator would be to expand into different sports. With the growing number of three-point attempts per season in the NBA and given the range of averages of three-point percentages being like that of batting averages, an analysis using James-Stein estimator may prove useful.

Regardless of the practical use of my results in this paper, the James-Stein estimator gives a new way to look at predicting a baseball player's batting average, slugging percentage, and OPS. While there are clearly some interesting results from a baseball standpoint, the simplicity of the estimator provides an opportunity for the James-Stein estimator to be used in all facets of our world, given how much data is at our fingertips.

6 Appendices

6.1 Appendix 1

Player	50 BA	J-S	Total BA	DIFF
Avisail Garcia	0.44	0.301	0.33	0.029
Josh Reddick	0.283	0.268	0.314	0.046
Lorenzo Cain	0.347	0.281	0.3	0.019
David Peralta	0.24	0.258	0.293	0.035
Aaron Judge	0.277	0.266	0.284	0.018
Giancarlo Stanton	0.226	0.255	0.281	0.026
Domingo Santana	0.212	0.252	0.278	0.026
Nick Markakis	0.265	0.264	0.275	0.011
Mookie Betts	0.347	0.281	0.264	-0.017
Yasiel Puig	0.294	0.270	0.263	-0.007
Carlos Gonzalez	0.192	0.248	0.262	0.014
Shin-Soo Choo	0.216	0.253	0.261	0.008
Hunter Pence	0.314	0.274	0.26	-0.014
Jay Bruce	0.275	0.266	0.254	-0.012
Nomar Mazara	0.34	0.280	0.253	-0.027
Kole Calhoun	0.24	0.258	0.244	-0.014
Matt Joyce	0.157	0.241	0.243	0.002
Max Kepler	0.294	0.270	0.243	-0.027
Steven Souza Jr	0.32	0.275	0.239	-0.036
Scott Schebler	0.16	0.241	0.233	-0.008
Curtis Granderson	0.157	0.241	0.212	-0.029
Jose Bautista	0.118	0.232	0.203	-0.029
Mitch Haniger	0.294	0.270	0.282	0.012
Gregory Polanco	0.216	0.253	0.251	-0.002
Jason Heyward	0.283	0.268	0.259	-0.009
Lonnie Chisenhall	0.28	0.267	0.288	0.021
J.D. Martinez	0.288	0.269	0.303	0.034
Bryce Harper	0.3	0.271	0.319	0.048
Seth Smith	0.286	0.268	0.258	-0.010
Jorge Bonifaico	0.24	0.258	0.255	-0.003

x bar = 0.263

c = 0.213

6.2 Appendix 2

Player	50 SLG	J-S SLG	Total SLG
Avisail Garcia	0.68	0.624	0.506
Josh Reddick	0.434	0.435	0.484
Lorenzo Cain	0.408	0.415	0.44
David Peralta	0.42	0.424	0.444
Aaron Judge	0.66	0.609	0.627
Giancarlo Stanton	0.415	0.420	0.631
Domingo Santana	0.404	0.412	0.505
Nick Markakis	0.445	0.443	0.384
Mookie Betts	0.469	0.462	0.459
Yasiel Puig	0.588	0.553	0.487
Carlos Gonzalez	0.327	0.352	0.423
Shin-Soo Choo	0.294	0.327	0.423
Hunter Pence	0.451	0.448	0.385
Jay Bruce	0.529	0.508	0.508
Nomar Mazara	0.62	0.578	0.422
Kole Calhoun	0.36	0.378	0.392
Matt Joyce	0.235	0.282	0.473
Max Kepler	0.49	0.478	0.425
Steven Souza Jr	0.52	0.501	0.459
Scott Schebler	0.38	0.393	0.484
Curtis Granderson	0.235	0.282	0.452
Jose Bautista	0.157	0.222	0.366
Mitch Haniger	0.588	0.553	0.491
Gregory Polanco	0.275	0.312	0.391
Jason Heyward	0.326	0.352	0.389
Lonnie Chisenhall	0.5	0.485	0.521
J.D. Martinez	0.769	0.692	0.69
Bryce Harper	0.5	0.485	0.595
Seth Smith	0.51	0.493	0.433
Jorge Bonifaico	0.36	0.378	0.432

x bar = 0.436862069
c = 0.7692028

6.3 Appendix 3

Player	50 OPS	J-S OPS	Total OPS
Avisail Garcia	1.161	1.136	0.885
Josh Reddick	0.661	0.670	0.847
Lorenzo Cain	0.875	0.870	0.803
David Peralta	0.698	0.705	0.796
Aaron Judge	1.018	1.003	1.049
Giancarlo Stanton	0.77	0.772	1.007
Domingo Santana	0.715	0.721	0.875
Nick Markakis	0.806	0.805	0.738
Mookie Betts	0.876	0.871	0.803
Yasiel Puig	0.985	0.972	0.833
Carlos Gonzalez	0.545	0.562	0.762
Shin-Soo Choo	0.649	0.659	0.78
Hunter Pence	0.776	0.777	0.701
Jay Bruce	0.891	0.885	0.832
Nomar Mazara	0.997	0.983	0.745
Kole Calhoun	0.693	0.700	0.725
Matt Joyce	0.459	0.482	0.808
Max Kepler	0.858	0.854	0.737
Steven Souza Jr	0.944	0.934	0.81
Scott Schebler	0.643	0.654	0.791
Curtis Granderson	0.449	0.473	0.775
Jose Bautista	0.415	0.441	0.674
Mitch Haniger	0.998	0.984	0.843
Gregory Polanco	0.573	0.588	0.695
Jason Heyward	0.692	0.699	0.715
Lonnie Chisenhall	0.815	0.814	0.881
J.D. Martinez	1.202	1.174	1.066
Bryce Harper	1.163	1.138	1.008
Seth Smith	0.907	0.899	0.774
Jorge Bonifaico	0.669	0.678	0.752
c =	0.931		
x bar =	0.797		

6.4 Appendix 4: R Code

Batting Average Shrinking Factor:

```
y <- c(.44, .283, .347, .24, .277, .226, .212, .265, .347, .294, .192, .216, .314,
.275, .34, .24, .157, .294, .32, .16, .157, .118, .294, .216, .283, .28, .288,
.3, .286, .24)
s2 <- mean(y)*(1-mean(y))/50
c <- 1 - 27*s2/(29*var(y))
c
```

Slugging Percentage Shrinking Factor:

```
y <- c(.68, .434, .408, .42, .66, .415, .404, .445, .469, .588, .327, .294, .451,
.529, .62, .36, .235, .49, .52, .38, .235, .157, .588, .275, .326, .5, .769, .5,
.51, .36)
s2 <- mean(y)*(1-mean(y))/50
c <- 1 - 27*s2/(29*var(y))
c
```

OPS Shrinking Factor:

```
y <- c(1.161, .661, .875, .698, 1.018, .77, .715, .806, .876, .985, .545, .649, .776
.891, .997, .693, .459, .858, .944, .643, .449, .415, .998, .573, .692, .815, 1.202, 1.163, .907, .669)
s2 <- mean(y)*(1-mean(y))/50
c <- 1 - 27*s2/(29*var(y))
c
```

1 References

Aldrich, John. “R.A. Fisher and the Making of Maximum Likelihood 1912-1922.” Statistical Science, vol. 12, no. 3, 1997, pp. 162–176., doi:10.1214/ss/1030037906.

Bradley Efron & Carl Morris (1975) Data Analysis Using Stein's Estimator and its Generalizations, Journal of the American Statistical Association, 70:350, 311-319, DOI: 10.1080/01621459.1975.10479864

Efron, Bradley, and Carl Morris. “Steins Paradox in Statistics.” Scientific American, vol. 256, no. 5, 1977, pp. 119–127., doi:10.1007/978-0-387-75692-9_7.

James, W.; Stein, Charles. Estimation with Quadratic Loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, 361--379, University of California Press, Berkeley, Calif., 1961. https://projecteuclid.org/euclid.bsmsp/1200512173

“MLB Player Batting Stats - 2017.” ESPN, ESPN Internet Ventures, www.espn.com/mlb/stats/batting/_/position/rf.

O'Neill, Heather M. “Do Hitters Boost Their Performance During Their Contract Years?” Do Hitters Boost Their Performance During Their Contract Years? | Society for American Baseball Research, Fall 2014 Baseball Research Journal, sabr.org/research/do-hitters-boost-their-performance-during-their-contract-years.

Stein, Charles. “Inadmissibility of the Usual Estimator for the Variance of a Normal Distribution with Unknown Mean.” Annals of the Institute of Statistical Mathematics, vol. 16, no. 1, 1964, pp. 155–160., doi:10.1007/bf02868569.

“2017 Major League Baseball Batting Leaders.” Baseball-Reference.com, www.baseball-reference.com/leagues/MLB/2017-batting-leaders.shtml.