Predicting Save Percentage for NHL Goalies with James-Stein Estimation

Abstract

In sports, measuring a parameter that is a gauge of a player's overall performance is often a challenge. We will examine whether James-Stein Estimators can provide a better guess for the season save percentage of a goalie than the sample mean. We used data from the first ten games of 2016-2017 NHL season to predict season-end save percentage for goalies in the 2016-2017 season. For 15/20 goalies in our data the James-Stein estimate for season save percentage was better than the sample mean, as judged by mean-squared error. The mean-squared error for the sample mean was on average 2.16 times greater than the mean-squared error of the Stein estimates.

I Introduction

This paper is a look into James-Stein Estimation, and how it applies to hockey. James-Stein Estimation suggests that an individual sample statistic weighted with the population statistic is a better estimator of an individual's true parameter than just the individual's sample statistic.

Consider independent RVs $(X_1, X_2, ..., X_n)$, and their means $(\mu_{X_1}, \mu_{X_2}, ..., \mu_{X_n})$. The James-Stein (J-S) Estimator works well when the variance between μ_{X_i}'s is less than the variance of point estimates for $(\mu_{X_1}, \mu_{X_2}, ..., \mu_{X_n})$. That is, the J-S Estimator is a good estimator for the μ_{X_i}'s when between player variance is less than the variance of individual players.

In sports, measuring a parameter that is a gauge of a player's overall performance is often a challenge. For example, if we wish to know the "true batting average" of a baseball player at a given point in time, by the time we have enough trials to make a reasonable guess, the player's "true batting average" has likely changed. The player's average could have been affected by a countless number of circumstances, including an injury to the player, improvement or regression in the player's ability, or a change in the opponent's ability. Even though attaining a player's "true batting average" at any given point is impossible, James-Stein Estimators can provide a better guess for the true parameter than the sample mean (the player's batting average at that time).

We will examine the James-Stein Estimator in relation to save percentage of NHL goalies. We will first look at the the performance of NHL goalies approximately 5-10 games into the season and estimate their performance for the season as a whole.

II Formula

The general formula [2] of the James-Stein estimator, is

$$\hat{\theta}_i^{JS} = \hat{\theta} + c\left(x_i - \hat{\theta}\right),$$

for a normal distribution $\hat{\theta} \sim N(\theta, V)$ where in our case, θ is the goalie's actual season save percentage, c is the shrinkage estimation, V is the variance of between the season save percentages of goalies, and x_i is the goalie's save percentage from early in the season. The equation to calculate c, the shrinkage factor, is

$$c = 1 - (N-3)/S, \text{ when } N > 3 \text{ , where}$$

$$S = \sum_{i=1}^{N} (x_i - \bar{x})^2 \text{ and } N = \text{number of goalies in the sample.}$$

III Application to Hockey

Motivation

We will conduct our study of James-Stein Estimation by first considering a given point in an NHL season. We will use a group of goalies who have roughly the same number of shots against them. For the save percentage of those goalies, we will compare how well the arithmetic average (sample mean) compares to J-S Estimators in predicting the goalie's overall season performance. We chose this particular set-up because Efron and Morris's 1977 article "Stein's Paradox in Statistics" employed a similar method. They looked at the batting averages of 18 players who had batted exactly 45 times in the 1970 season, and predicted their season batting average using Stein's estimator and the arithmetic average. In their study the mean squared error of the MLE was approximately 3.5 times larger on average than the mean squared error of the Stein Estimator [1] . We hope to employ this same tactic to estimate overall season performance for goalies in the NHL.

We will look into the differences in the mean squared error between J-S Estimators and the MLE. This will be to explore the benefits of using the James-Stein over other estimators in certain situations.

Dataset

Our data for this study came from the website **offsidereview.com**. For our early-season data, we looked at the save percentages of goalies' with 150 - 350 saves a month into the 2016-2017 NHL season. Of those goalies, we only considered those who played most of the season (over 50 games). The resulting sample was composed of 20 NHL goalies, see *Table 1* to see the sample.

Methodology

Consider the MLE for each player, $\hat{p}_i$, i.e. the save percentage of a given goalie at some point in the NHL season (the sample mean). Then we claim that $n\hat{p}_i \sim Bin(n, P_i)$, where n is the number of shots against the goalie and P_i is the season save percentage of the goalie. By the Central Limit Theorem, we can approximate the distribution by $\hat{p}_i \dot{\sim} N(P_i, \sigma^2)$, where σ^2 is the variance from the binomial best estimated by:

$$(1) \quad \sigma^2 = \bar{p}(1-\bar{p})/n$$

Applying the James-Stein Estimator, with p_i's as x_i's and p_{season}, the 2016-2017 season save percentage of goalies, as our parameter of interest, we have:

$$\hat{p}_i^{JS} = \bar{p} + c\left(\hat{p}_i - \bar{p}\right), \qquad \bar{p} = \text{grand average of averages 5-10 games into the season}$$

$$\bar{p} = \frac{1}{20}\sum_{i=1}^{20}\hat{p}_i, \text{ and} \qquad c = \text{shrinkage factor}$$

$$c = 1 - \frac{17\sigma^2}{\sum_{i=1}^{20}(\hat{p}-\bar{p})^2}. \qquad \hat{p}_i = \text{individual early-season save percentage}$$

$\sigma^2 =$ the variance of an individual goalie given a certain number of saves estimated by (1)

Because we had varying values of n (shots against) for each player, we took the average of the 20 values of n and got $\bar{n} = 232.7$. $\bar{p}$ was computed to be 0.9187, so $\sigma^2 = 0.00032$. After calculating $\sum_{i=1}^{20}(\hat{p}-\bar{p})^2$, we get that $c = 0.41493$. See *Figure 3* in the Appendix for the associated R code.

Results

The average mean-squared error (MSE) for the James-Stein Estimator was 0.007 compared to the average MSE of the sample mean (the MLE), which was 0.0150. On average, the MSE of the MLE was 2.16 times larger. As seen in *Table 2* or *Figure 2*, for 15 of the 20 players in the sample, the James-Stein Estimate had a smaller MSE than the MLE. I.e. for 15/20 players in our sample the James-Stein Estimator was better. See *Figure 1* for an intuitive visualization of the J-S Estimator.

IV Conclusion

The James-Stein Estimator outperformed the MLE by a factor of 2.16. That is, the mean-squared error for the MLE was on average 2.16 times larger than the mean-squared error the for the Stein-Estimator.

The James-Stein Estimator relies on the fact that the variance between players' season save percentages is expected to be lower than the variance between early-season save percentages. As seen in *Figure 3*, the J-S Estimator reduces the between $\hat{p}$ variability in the sample, squeezing the point estimates towards $\bar{p}$. Note that in *Figure 1* the p_{season} values are roughly centered around $\bar{p}$ which is represented by the black line.

One of the greatest weaknesses of our James-Stein Estimator is that it does not factor in prior information. Consider two goalies, through 300 shots against both have a save percentage of 0.90. Goalie A has a career save percentage of 0.92, whereas Goalie B has a career save percentage of 0.85. Consider a league-wide average save percentage of 0.87 . With our James-Stein Estimator, Goalie A and Goalie B's season averages will be estimated by some common value in the range $R = (0.87, 0.90)$, between the league-wide save percentage and their identical individual save percentages. However based off of what we know about these players, we know that is is not sensible to predict these two players will perform similarly for the remainder of the season, given their past career performance. A sensible solution to this issue is to use some Bayesian Estimator in conjunction with a Stein Estimator that pulls our estimator for a player towards their career average.

Despite this shortfall of the James-Stein Estimator, in our data the Estimator worked very well. The J-S estimator consistently provided better estimates than the MLE (sample mean).

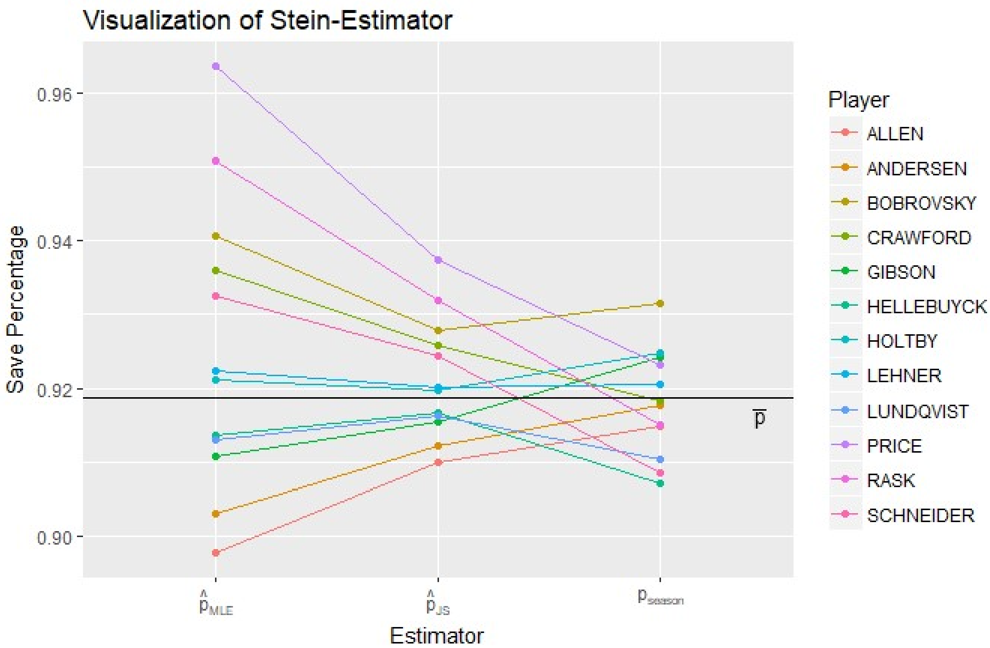

Figure 1: Comparison of sample mean ($\hat{p}_{MLE}$), J-S estimator ($\hat{p}_{JS}$), and season save percentage (p_{season})
Note: $\bar{p}$ is the grand average of save percentages 5-10 games into the season
Note 2: For *Figure 1* a random sample of twelve players was taken from the twenty players for the graph. We could not show all twenty players at once, due to the amount of clutter in the resulting visualization.

V Appendix

Sample GA refers to the goals against the goalie in the early season sample. *Sample SA* refers to the shots against the goalie in the early season sample. $\hat{p}_{MLE}$ is 1 - *Sample GA*/*Sample SA*, i.e. the goalie's save percentage 5-10 games into the season. p_{season} is the goalie's end of season save percentage. And *Games Played* refers to the total number of games played by the goalie during the season.

	Player	Sample GA	Sample SA	$\hat{p}_{MLE}$	p_{season}	Games Played
1	BRADEN HOLTBY	18	228	0.9211	0.9249	63
2	CAM TALBOT	25	341	0.9267	0.9193	73
3	CAM WARD	20	169	0.8817	0.9053	61
4	CAREY PRICE	7	193	0.9637	0.9231	62
5	CONNOR HELLEBUYCK	14	162	0.9136	0.9072	56
6	COREY CRAWFORD	18	281	0.9359	0.9183	55
7	CORY SCHNEIDER	17	252	0.9325	0.9085	60
8	DEVAN DUBNYK	12	231	0.9481	0.9235	65
9	FREDERIK ANDERSEN	29	299	0.9030	0.9176	66
10	HENRIK LUNDQVIST	20	230	0.9130	0.9103	57
11	JAKE ALLEN	22	215	0.8977	0.9148	61
12	JOHN GIBSON	24	269	0.9108	0.9242	52
13	MARTIN JONES	23	254	0.9094	0.9119	65
14	PEKKA RINNE	21	248	0.9153	0.9180	61
15	PETER BUDAJ	19	193	0.9016	0.9168	53
16	PETR MRAZEK	23	250	0.9080	0.9008	50
17	ROBIN LEHNER	16	206	0.9223	0.9205	59
18	SERGEI BOBROVSKY	16	270	0.9407	0.9315	63
19	STEVE MASON	22	180	0.8778	0.9081	58
20	TUUKKA RASK	9	183	0.9508	0.9150	65

Table 1: Early season and end of season save percentage for NHL Goalies in the 2016-2017 season

	Player	$\hat{p}_{MLE}$	$\hat{p}_{JS}$	p_{season}	MLE MSE	JS MSE
1	BRADEN HOLTBY	0.9211	0.9197	0.9249	0.0038	0.0052
2	CAM TALBOT	0.9267	0.9220	0.9193	0.0074	0.0027
3	CAM WARD	0.8817	0.9033	0.9053	0.0236	0.0020
4	CAREY PRICE	0.9637	0.9374	0.9231	0.0406	0.0143
5	CONNOR HELLEBUYCK	0.9136	0.9166	0.9072	0.0064	0.0094
6	COREY CRAWFORD	0.9359	0.9258	0.9183	0.0176	0.0075
7	CORY SCHNEIDER	0.9325	0.9244	0.9085	0.0240	0.0159
8	DEVAN DUBNYK	0.9481	0.9309	0.9235	0.0246	0.0074
9	FREDERIK ANDERSEN	0.9030	0.9122	0.9176	0.0146	0.0054
10	HENRIK LUNDQVIST	0.9130	0.9163	0.9103	0.0027	0.0060
11	JAKE ALLEN	0.8977	0.9100	0.9148	0.0171	0.0048
12	JOHN GIBSON	0.9108	0.9154	0.9242	0.0134	0.0088
13	MARTIN JONES	0.9094	0.9148	0.9119	0.0025	0.0029
14	PEKKA RINNE	0.9153	0.9173	0.9180	0.0027	0.0007
15	PETER BUDAJ	0.9016	0.9116	0.9168	0.0152	0.0052
16	PETR MRAZEK	0.9080	0.9143	0.9008	0.0072	0.0135
17	ROBIN LEHNER	0.9223	0.9202	0.9205	0.0018	0.0003
18	SERGEI BOBROVSKY	0.9407	0.9278	0.9315	0.0092	0.0037
19	STEVE MASON	0.8778	0.9017	0.9081	0.0303	0.0064
20	TUUKKA RASK	0.9508	0.9320	0.9150	0.0358	0.0170

Table 2: Comparison of sample mean ($\hat{p}_{MLE}$), J-S estimator ($\hat{p}_{JS}$), and season save percentage (p_{season}) and the MSE of the sample mean and J-S estimates

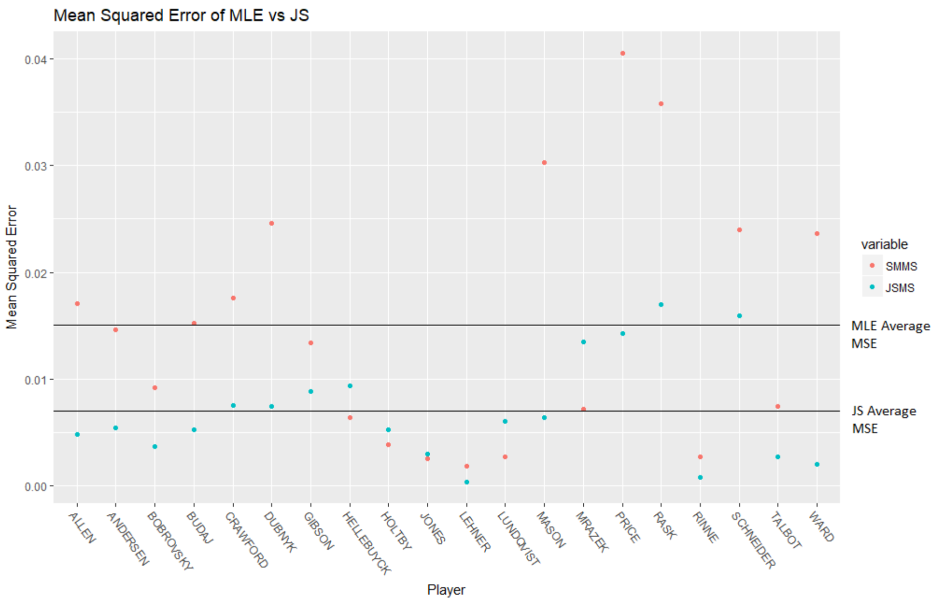

Figure 2: Mean-Squared Error of MLE (SMMS) vs JS (JSMS) by Player

```
### ----------------------------- ###
###         DATA ANALYSIS         ###
### ----------------------------- ###

### ------------------------------------- ###

splitStats <- read.csv("GoalieSplit.csv")
seasonStats <- read.csv("GoalieSeason.csv")

# Keep only the players who have between 150-350 shots against in the first split,
# and played at least 50 games in the season
subSplit <- subset(splitStats, SA >= 150)
subSplit <- subset(subSplit, SA <= 350)
seasonSplit <- subset(seasonStats, GP >= 50)

# Merge the split and season data and remove unnecessary columns
mergedData <- merge(subSplit,seasonSplit, by = "Player")
cutDownData <- subset(mergedData, select=c("Player", "GA.x", "SA.x",
                                           "GA.y", "SA.y", "Sv..x", "Sv..y", "GP.y"))
cutDownData$SvP.x <- cutDownData$Sv..x/100
cutDownData$SvP.y <- cutDownData$Sv..y/100
sum(cutDownData$SA.x)

# Caulculate the shrinkage c for the JS Estimator
k <- nrow(cutDownData) # number of unknown means
pbar <- mean(cutDownData$SvP.x)#total average of averages
n <- mean(cutDownData$SA.x) #average number of shots against
phat <- cutDownData$SvP.x # Sample means, the MLEs

c <- 1 - (k-3)*(pbar*(1 - pbar)/n)/sum((phat - pbar)^2) # apply the shrinkage formula

# Calculate our MSE values for JS Estimator and the MLE (SvP.x)
meanSq <- function(x, y){sqrt(mean((x-y)^2))}
cutDownData$JS <- pbar + c*(phat - pbar) # create a column for JS estimates
cutDownData$SMMS <- mapply(meanSq,cutDownData$SvP.x, cutDownData$SvP.y)
cutDownData$JSMS <- mapply(meanSq,cutDownData$JS, cutDownData$SvP.y)
meanSq(cutDownData$SvP.x, cutDownData$SvP.y) ## total MLE MSE
meanSq(cutDownData$JS, cutDownData$SvP.y)## total JS MSE
```

Figure 3: R Code for James-Stein Estimates

VI References

1. Efron, Bradley, and Carl Morris. "Stein's Paradox in Statistics." Scientific American, May 1977, 119-27.

2. Efron, Bradley, and Trevor J. Hastie. Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. New York: Cambridge University Press, 2014, 83-97.

3. Lopez, Michael, Professor. "Lecture 8: Steins Paradox and Hockey Shooting Statistics." Lecture, Skidmore College. Accessed April 3, 2018.