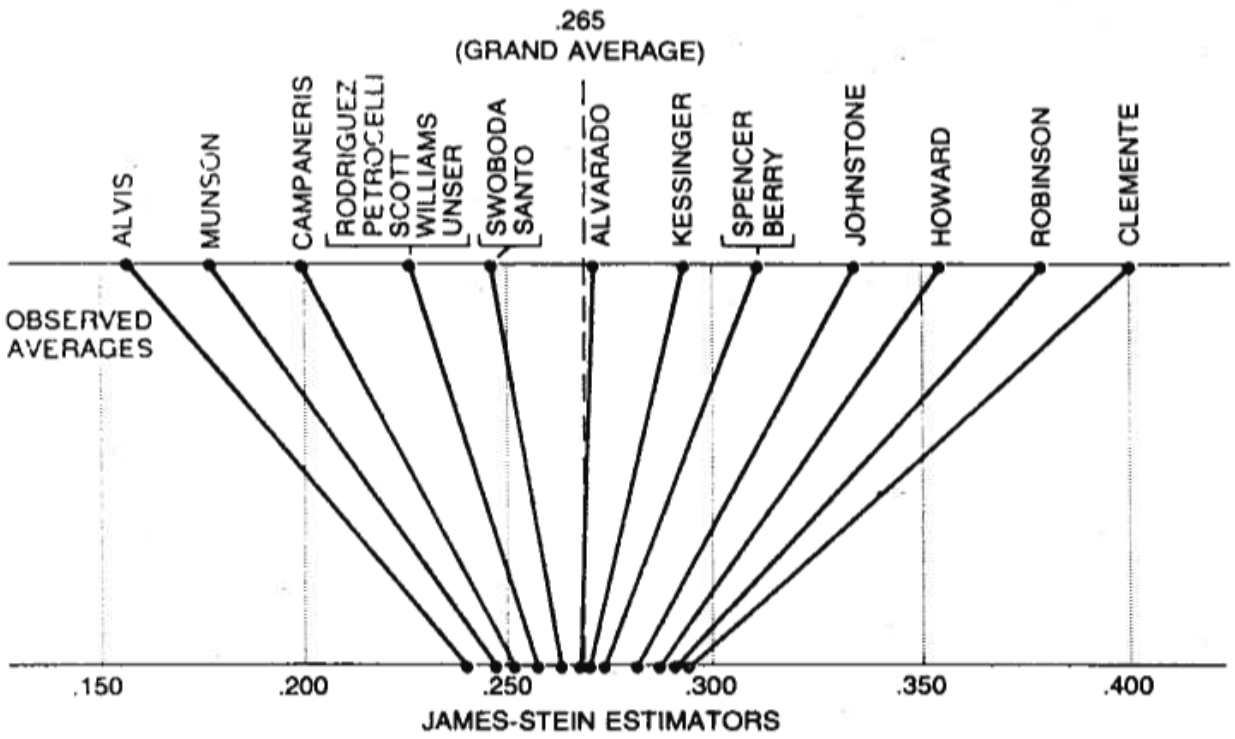
STATS 50: Mathematics of Sport **Spring 2019**

Week 4 – Regression to the mean

Lecturer: Maxime Cauchois

Warning: *these notes may contain factual errors*

1 The James-Stein phenomenon



JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.

Intuitively, if one wants to estimate several completely independent quantities at the same time, it does not really make sense to combine all your data to form your estimates. For instance, if we wanted to estimate Kevin Durant's real free throw percentage, as well as this class's passing rate and the number of goals scored by Leo Messi next season, it is not clear at all why we should consider these quantities altogether.

However, it turns out that James-Stein phenomenon proves us wrong: it is better to aggregate all your estimates!

Table 1.1: Batting averages $z_i = \hat{\mu}_i^{(\mathrm{MLE})}$ for 18 major league players early in the 1970 season; μ_i values are averages over the remainder of the season. The James–Stein estimates $\hat{\mu}_i^{(\mathrm{JS})}$ (1.35) based on the z_i values provide much more accurate overall predictions for the μ_i values. (By coincidence, $\hat{\mu}_i$ and μ_i both average 0.265; the average of $\hat{\mu}_i^{(\mathrm{JS})}$ must equal that of $\hat{\mu}_i^{(\mathrm{MLE})}$.)

Name	hits/AB	$\hat{\mu}_i^{(\mathrm{MLE})}$	μ_i	$\hat{\mu}_i^{(\mathrm{JS})}$
Clemente	18/45	.400	**.346**	.294
F Robinson	17/45	.378	**.298**	.289
F Howard	16/45	.356	**.276**	.285
Johnstone	15/45	.333	**.222**	.280
Berry	14/45	.311	**.273**	.275
Spencer	14/45	.311	**.270**	.275
Kessinger	13/45	.289	**.263**	.270
L Alvarado	12/45	.267	**.210**	.266
Santo	11/45	.244	**.269**	.261
Swoboda	11/45	.244	**.230**	.261
Unser	10/45	.222	**.264**	.256
Williams	10/45	.222	**.256**	.256
Scott	10/45	.222	**.303**	.256
Petrocelli	10/45	.222	**.264**	.256
E Rodriguez	10/45	.222	**.226**	.256
Campaneris	9/45	.200	**.286**	.252
Munson	8/45	.178	**.316**	.247
Alvis	7/45	.156	**.200**	.242
Grand Average		.265	**.265**	.265

2 The James-Stein Estimator

Suppose that we want to estimate n different quantities $\mu_1, ..., \mu_n$, and that we have n estimates $Z_1, ..., Z_n$ of those quantities, which all have the same variance σ^2.

We can then form the following James-Stein estimate based on this:

$$\hat{\mu}^{(JS)} = \left(1 - \frac{(n-2)\sigma^2}{\|Z\|_2^2}\right) Z \tag{1}$$

We see that we actually shrink our estimate towards 0. Here, in our baseball case, the JS estimate described in the seminal paper [1] by Bradley Efron actually regresses all the batting averages towards the "average of the averages", as it is more likely that each player who has performed better than the others actually regresses to a lower mean.

Indeed, in the case where you have prior information that the mean in your population population average is likely to be around some ν, you can actually get a better estimate by shrinking towards this value rather than towards 0, which in this case would be a somehow artificial value to consider. The JS estimate then becomes:

$$\hat{\mu}_{\nu}^{(JS)} = \left(1 - \frac{(n-2)\sigma^2}{\|Z-\nu\|_2^2}\right)(Z-\nu) + \nu \tag{2}$$

3 Bayesian interpretation

Suppose that each player's level (i.e. batting average, shooting percentage, speed, ...) can be evaluated with a unique number μ_i, which was drawn (independently from the others) from a normal distribution with the same mean ν and variance τ^2. In other terms, one can imagine that each μ_i is the measure of each player's talent, and obviously some players are going to get more abilities than others through hard work, good coaching and also sheer luck!

Now, if we want to estimate each player's talent, we are usually going to observe a more or less unbiased (but noisy) estimate Z_i of his abilities: typically, it is going to be the batting average over the first month of the season, or the shooting percentage in the last ten games... Here, we are to assume that each Z_i are drawn independently from the others with mean μ_i and variance σ^2. In other terms, there is some part of randomness which makes you perform better on certain moments of the season, but your performance obviously also depends on your talent.

The interesting part is the following: we actually observe every Z_i, and we can have a pretty decent estimate of ν, which is the average talent of all the players. Now, given that we observe all the Z_i, we can prove using the Bayes formula that:

$$\mu_i | Z_i \sim \mathcal{N}\left(\frac{\nu/\tau^2 + Z_i/\sigma^2}{1/\tau^2 + 1/\sigma^2}, \frac{\tau^2\sigma^2}{\tau^2+\sigma^2}\right)$$

In other terms, once you observed all the Z_i, the best estimate that you can come with for μ_i seems to be an actual weighted average of the all players mean and of your individual observation. In particular, for those players which have performed better than average ($Z_i > \nu$), you will in fact shrink your estimate towards ν!

4 Regression to the mean with unequal variances

Very often in sports, one does not get to observe as many events for each player, which results in different variances. Indeed, a player which has attempted 100 shots will have a less noisy estimate of the actual true ability than the one with just 10 attempts.

More generally, in the Bayesian framework described above, suppose again that the ability of each player was drawn from some normal distribution:

$$\mu_i \overset{i.i.d}{\sim} \mathcal{N}(\mu, \tau^2)$$

Then, we suppose that the player i made N_i independent attempts with the same mean μ_i of success, resulting in an observed average Z_i such that, given the player i's true ability μ_i:

$$Z_i | \mu_i \sim Binomial(N_i, \mu_i)/N_i \simeq \mathcal{N}(\mu_i, \sigma_i^2)$$

where $\sigma_i^2 = \mu_i(1-\mu_i)/N_i$, and where we used the Central Limit Theorem to approximate a Binomial distribution with a Normal distribution. In other terms, the number of made shots is a normal variable centered around the player i's true ability μ_i and with a variance proportional to the inverse of the number of shots attempted.

By Bayes formula, we know that:

$$\mu_i | Z_i \sim \mathcal{N}\left(\frac{\mu/\tau^2 + Z_i/\sigma_i^2}{1/\tau^2 + 1/\sigma_i^2}, \frac{\tau^2\sigma_i^2}{\tau^2 + \sigma_i^2}\right)$$

which implies that ideally:

$$\hat{\mu}_i = \frac{\mu/\tau^2 + Z_i/\sigma_i^2}{1/\tau^2 + 1/\sigma_i^2}$$

Now obviously, there are here a lot of unknown parameters, including μ, σ_i and τ^2.

First, μ can be easily estimated with the overall average across all players, since it is just the probability that a player chosen at random actually makes a shot:

$$\hat{\mu} = \frac{\sum_{i=1}^N N_i Z_i}{\sum_{i=1}^N N_i} = \frac{\text{Shots made}}{\text{Shots attempted}}$$

Then, assuming that all μ_i should be rather peaked around μ, we can estimate each σ_i^2:

$$\hat{\sigma}_i^2 = \frac{\hat{\mu}(1-\hat{\mu})}{N_i}$$

i.e we take σ_i^2 to be just a linear function of $1/N_i$.

The hardest parameter to estimate is τ^2, but here is a method to approach its true value accurately and rigorously: we can remark that, marginally, each Z_i has the following distribution:

$$Z_i \sim \mathcal{N}(\mu, \tau^2 + \sigma_i^2)$$

That is, if we don't know the player at all (or if we sample it at random), its average number of made shots will be normal, with mean μ the overall population mean and variance $\tau^2 + \sigma_i^2$, that is we add the variance due to inherent variability in our population (τ^2) as well as the "luck" variance (σ_i^2) specific to the player and depending on his/her number of attempts.

But this precisely means that $(Z_i - \mu)^2 - \sigma_i^2$ is an unbiased estimator of τ^2, for any i! In particular, any estimator of the form:

$$\hat{\tau}_w^2 = \sum_{i=1}^N w_i((Z_i - \mu)^2 - \sigma_i^2)$$

will be unbiased for τ^2 as long as $\sum_{i=1}^N w_i = 1$.

The last question that remains lies in the choice of w. At first sight, it could make sense to just use $w_i = 1/N$ for any i, that is grant equal weight to each player for estimating τ^2. However, a normal variable $X \sim \mathcal{N}(\mu, \sigma^2)$ verifies the following identity:

$$\text{Var}((X-\mu)^2) = 2\sigma^4$$

Therefore the variance of $\hat{\tau}_w^2$ can be computed as follows:

$$\begin{aligned}
\text{Var}(\hat{\tau}_w^2) &= \mathbb{E}((\hat{\tau}_w^2 - \tau^2)^2) \\
&= 2\sum_{i=1}^N w_i^2 \text{Var}((Z_i - \mu)^2) \\
&= 2\sum_{i=1}^N w_i^2 (\tau^2 + \sigma_i^2)^2
\end{aligned}$$

Because we want $\hat{\tau}_w^2$ to be as close as possible to τ, we want to minimize its variance. It turns out that the latter expression is minimal for:

$$w_i \propto \frac{1}{(\tau^2 + \sigma_i^2)^2}$$

Therefore, our estimate of τ^2 should be:

$$\hat{\tau}^2 = \left\{ \sum_{i=1}^{N} \frac{(Z_i - \mu)^2 - \sigma_i^2}{1/(\tau^2 + \sigma_i^2)^2} \right\} / \left\{ \sum_{i=1}^{N} 1/(\tau^2 + \sigma_i^2)^2 \right\}$$

But, if you look at it very carefully, you'll notice that our estimator $\hat{\tau}^2$ depends itself on μ and σ_i^2. This is not a problem since those two quantities have been previously estimated and can be replaced by their respective estimators. More annoyingly, our $\hat{tau}^2$ intricately depends on τ^2, which is precisely the parameter it is trying to estimate!

For this reason, we use an iterative method and find $\hat{\tau}^2$ that solves:

$$\hat{\tau}^2 = \left\{ \sum_{i=1}^{N} \frac{(Z_i - \hat{\mu})^2 - \hat{\sigma}_i^2}{1/(\hat{\tau}^2 + \hat{\sigma}_i^2)^2} \right\} / \left\{ \sum_{i=1}^{N} 1/(\hat{\tau}^2 + \hat{\sigma}_i^2)^2 \right\} \tag{3}$$

This τ^2 is precisely the one we computed during the R session using an iterative approach where we would update recursively $\hat{\tau}$ until it satisfies (3).

Remark In the case of equally sampled players (where each one attempted the same number of shots), all w_i can be taken equal to $1/N$

5 Some examples drawn from sports

5.1 Computing the JS Estimate in the Baseball case

In the baseball example drawn by Efron, we first μ by $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} Z_i$: if there are enough players, the law of large numbers will ensure that $\bar{\mu}$ is close to μ. It is around .265 in our case, and we call it the *Grand Average*.

Then we see that we are to estimate each real batting average by:

$$\hat{\mu}_i^{(JS)} = \bar{\mu} + c(Z_i - \bar{\mu})$$

where $0 < c \simeq .212 < 1$ is a number which depends on all our data (and determines the level of shrinkage).

5.2 Estimating real 3pt percentage

After a month into the regular season, we have access to each player 3PT percentage so far, and our goal is to estimate the 3PT percentage for the whole regular season. Here, we assume that all players have attempted a similar number of shots.

Note that in our James-Stein estimator (2), there are two unknown parameters: ν and σ^2. In order to give accurate predictions, we will have to find decent estimates of those two quantities.

We know that we can replace σ^2 by the following sample variance $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \bar{\mu}(1 - \bar{\mu})/N$$

where N is the number of attempts for each player. This is simply due to the fact that, given a player's ability μ_i, its number of made shots will be a Binomial variable with parameters N and μ_i: each shot will be a $0-1$ variable, equal to 1 with probability μ_i.

As for ν, we know that it should be close to a reasonable season average value for all players: we choose to use 42%, which can seem a bit arbitrary but can be thought as an estimation gained from prior information over the past seasons. We could also have chosen to use the average of all players during the first month of the season.

In this example, the JS estimate actually overperforms the naive estimate by a factor of 17 in terms of mean squared error! In the figure 1, we displayed both the James-Stein estimator described above, but also the linear regression curve. Of course, the latter gives a more accurate prediction of the real 3PT percentage, but note that it was allowed to look at the data to compute the fit! On the other hand, our JS estimate only uses our values from the first month of the season and yields predictions for the entire season, from these observations only.

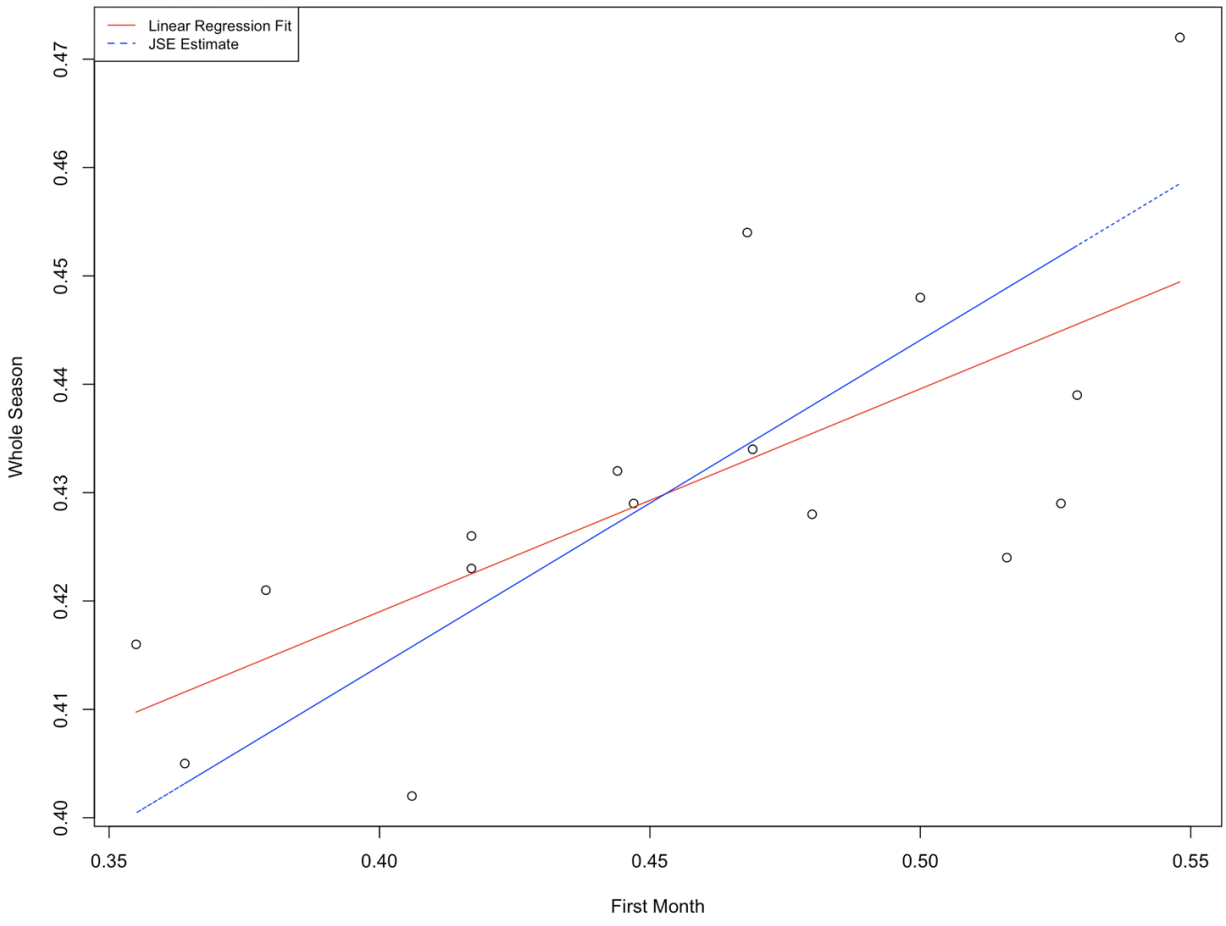


Figure 1: James-Stein Estimator for prediction of NBA real 3PT percentage.

6 The winner's curse

The underlying reason why these estimates are better than naive ones is the following: suppose that all players have more or less the same level. Then it means that those who performed better during the first month were actually just lucky: you expect them to perform worse in the next part of the season and to regress to the mean, which is the reason why you'd better predict a lower estimate for the rest of the season.

For instance, if one player had made say 50 3-pointers in the first 10 games of the season, and you want to predict the number of shots he will make in total during the regular season, you are probably not going to say 410, even though this would be the "unbiased"estimate (at 5 3 pointers per game). Obviously there can be some outliers: those are players who are actually more talented than others!

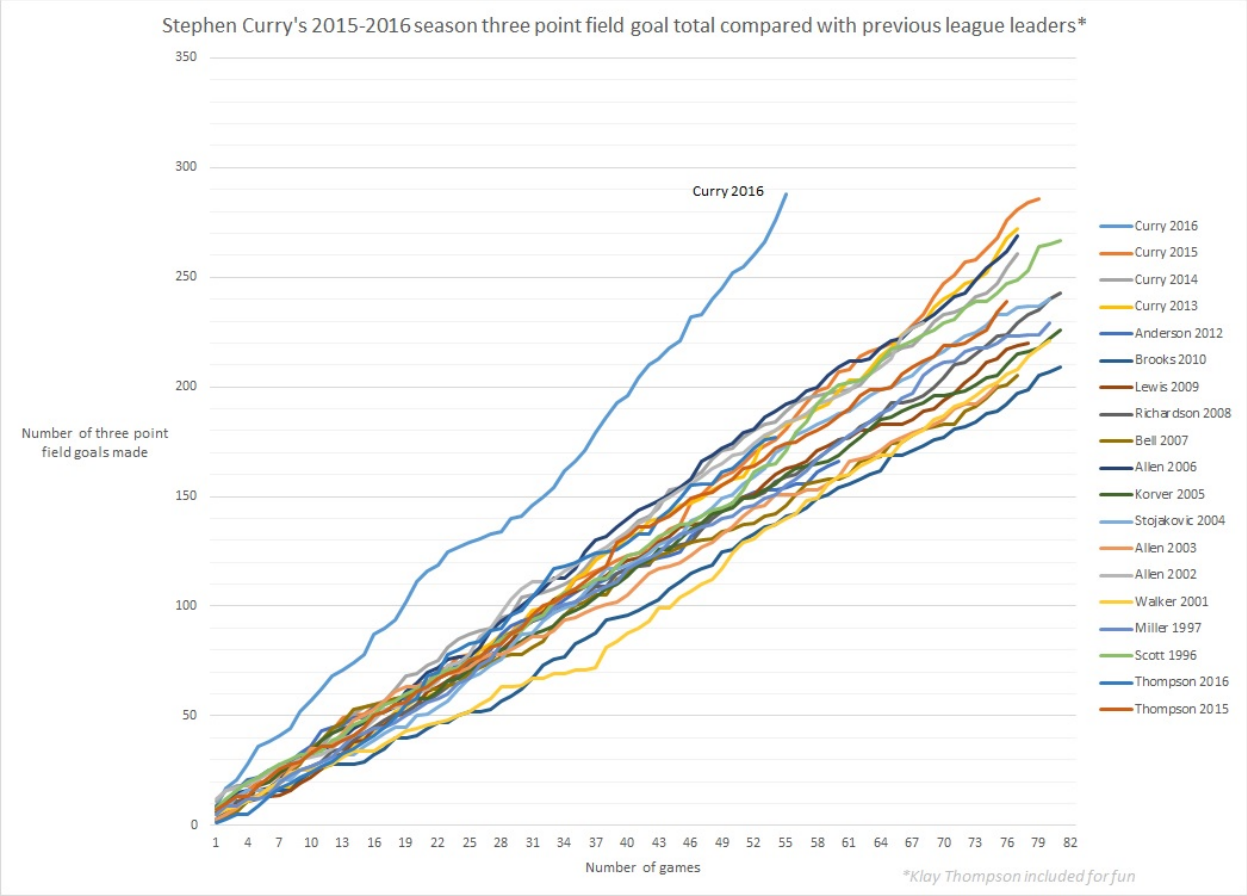


Figure 2: Number of 3 pointers in function of the number of games played

The winner's curse can be explained very easily mathematically speaking. Suppose that we are given $Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. A priori, all the Z_i come from the same distribution, but consider the player who has achieved $\max_j Z_j$, and you can very easily verify that:

$$\mathbb{E}\left(\max_{j\in[n]} Z_j\right) > 0 = \max_j \mathbb{E}(Z_j)$$

i.e. you expect that the player who comes first actually performed better than his/her level! Because

he was the first among all (during the competition or the first month or whatever), you actually expect him/her to have exceeded the expectations!

It also means that for the next game or the next competition, you actually expect this same player to show poorer performance, because he or she just got lucky this time!

We call this kind of phenomenon **selection bias**, and it is very much studied in statistics nowadays.

References

[1] Efron, Bradley & Morris, Carl. (1977). Stein's Paradox in Statistics. Scientific American - SCI AMER. 236. 119-127. 10.1038/scientificamerican0577-119.