Lecture 8: Steins paradox and hockey shooting statistics

Skidmore College

Goals

- Stein's Paradox
- Shooting Percentages in hockey
- Tools: Bayesian statistics, likelihood estimation, bias/variance

Set-up:

We are NHL general managers after the 2012-2013 season. Who are we going to sign? Assume all else is equal (same contract, same stats), here are two players in the 2012-13 season.

Player	Goals
David Krejci	17
Evgeni Malkin	7

Set-up:

We are NHL general managers after the 2012-2013 season. Who are we going to sign?

Player	Goals	Shots	Shooting %
David Krejci	17	106	16.0%
Evgeni Malkin	7	101	6.9%

Why does this information matter?

Set-up:

We are NHL general managers after the 2012-2013 season. Who are we going to sign?

Player	Goals	Shots	Shooting %
David Krejci (C)	17	106	16.0%
Evgeni Malkin (C)	7	101	6.9%

Information we want:

- What shooting percentages can we expect for Krejci and Malkin going forward?

Statistical definitions:

- Bias vs. Unbiased, Bias/Variance trade-off, James-Stein estimator

Interlude:

Let's say we are interested in the overall fraction of the Skidmore students that will support a football team, p_0. In a completely randomized survey of 100 students, 22% of the Skidmore campus supports the adoption of a football team.

- Our sample statistic, $\hat{p} = 0.22$, is **unbiased** for p_0 because $E[\hat{p}] = p_0$.
- That is, our best guess as to the true fraction of the Skidmore students that support a football team is 22%. If we had one guess, that's it.
- *Note*: $\hat{p} = 0.22$ is biased for p_0 if $E[\hat{p}] \neq p_0$

Back to hockey

Player	Goals	Shots	Shooting %
David Krejci (C)	17	106	16.0%
Evgeni Malkin (C)	7	101	6.9%

- Let p_K and p_M are the true probabilities that a Krejci or Malkin shot will score a goal, respectively
- What are our estimates of p_K and p_M?
  - $\hat{p}_K = 0.160$ is unbiased for p_K ($E[\hat{p}_K] = p_K$)
  - $\hat{p}_M = 0.069$ is unbiased for p_M ($E[\hat{p}_M] = p_M$)
- *Note*: $\hat{p}_M$ and $\hat{p}_K$ are called maximum likelihood estimators

Back to hockey

Player	Goals	Shots	Shooting %
David Krejci (C)	17	106	16.0%
Evgeni Malkin (C)	7	101	6.9%

What other information could we use?

- League-wide shooting percentage for forwards is 10.6%
- How do we incorporate this information?

James-Stein estimator

Via Efron & Morris, $z = \bar{y} + c(y - \bar{y})$,

- $\bar{y}$ is grand average of averages
- y is average of a single data set
- c is a shrinking factor, $c = \frac{N/0.25}{N/0.25 + 1/\sigma^2}$
  - N is number of observations we have on a player
  - σ^2 is variance of observations from one player to the next

James-Stein estimator, translated

Via Efron & Morris, $\hat{p}_{JS} = \bar{\hat{p}} + c * (\hat{p} - \bar{\hat{p}})$,

- $\bar{\hat{p}}$ is average of each players shooting percentage
- $\hat{p}$ is a single players observation
- c is a shrinking factor, $c = \frac{N/0.25}{N/0.25 + 1/\sigma^2}$
  - k is number of shooters
  - σ^2 is variance of individual shooter given certain number of attempts
- Plug in $c = 1$:
- Plug in $c = 0$:

James-Stein estimator, translated

Via Efron & Morris, $\hat{p}_{JS} = \bar{\hat{p}} + c * (\hat{p} - \bar{\hat{p}})$,

- $\bar{\hat{p}}$ is average of each players shooting percentage
- $\hat{p}$ is a single players observation
- c is a shrinking factor, $c = \frac{N/0.25}{N/0.25 + 1/\sigma^2}$
  - k is number of shooters
  - σ^2 is variance of individual shooter given certain number of attempts
- What happens as σ^2 goes up/down?

James-Stein estimator, implemented

- Initial data: shooting statistics from the 2012-2013 season

```
## # A tibble: 2 x 11
##   Name        Position Team  Games Season   Age Salary Goals Assists Shots  <
##   <chr>       <chr>    <chr> <dbl>  <dbl> <dbl>  <dbl> <dbl>   <dbl> <dbl>  <
## 1 Justin.Ab~ RL       DET      61 2.01e7    22   0.71     4       4    77 0.
## 2 Justin.Ab~ RL       DET      85 2.01e7    23   0.75     7      11   128 0.
```

```r
first_season <- nhl_data %>% filter(Season == 20122013)
first_players <- first_season %>%
  group_by(Name) %>%
  filter(Shots <= 106, Shots >= 100, Position !="D") %>%
  select(Name, Position, Goals, Shots, ShP)
dim(first_players)
```

```
## [1] 12  5
```

James-Stein estimator, implemented

```r
head(first_players)
```

```
## # A tibble: 6 x 5
## # Groups:   Name [6]
##   Name              Position Goals Shots    ShP
##   <chr>             <chr>    <dbl> <dbl>  <dbl>
## 1 Jason.Chimera     L            4   101 0.0396
## 2 Johan.Franzen     RL           8   105 0.0762
## 3 Brendan.Gallagher R           13   103 0.126
## 4 Taylor.Hall       L           12   106 0.113
## 5 Jarome.Iginla     R           10   103 0.0971
## 6 David.Krejci      C           17   106 0.160
```

12 forwards, each with between 100-106 shots

James-Stein estimator, implemented

```r
p_bar <- mean(first_players$ShP)
p_bar
```

```
## [1] 0.1057114
```

```r
p_hat <- first_players$ShP
p_hat
```

```
##  [1] 0.03960396 0.07619048 0.12621359 0.11320755 0.09708738 0.16037736
##  [7] 0.06930693 0.13725490 0.08571429 0.19417476 0.11000000 0.05940594
```

James-Stein estimator, implemented

```
N <- first_players$Shots
N
```

```
##  [1] 101 105 103 106 103 106 101 102 105 103 100 101
```

```
sigma_sq <- sd(p_hat)^2 ##Rough approximation
sigma_sq
```

```
## [1] 0.001953588
```

James-Stein estimator, implemented

```
c <- (N/0.25)/(N/0.25 + 1/sigma_sq)
c
```

```
##  [1] 0.4411065 0.4507024 0.4459460 0.4530502 0.4459460 0.4530502 0.4411065
##  [8] 0.4435368 0.4507024 0.4459460 0.4386548 0.4411065
```

- Hockey shrinking factor after 100-105 shots: $c = 0.45$
- How to interpret c?

James-Stein estimator, implemented

Calculating the MLE and James-Stein estimates

```
first_players$Shp_MLE <- first_players$ShP
first_players$Shp_JS <- p_bar + c*(p_hat - p_bar)
head(first_players)
```

```
## # A tibble: 6 x 7
## # Groups:   Name [6]
##   Name              Position Goals Shots    ShP Shp_MLE Shp_JS
##   <chr>             <chr>    <dbl> <dbl>  <dbl>   <dbl>  <dbl>
## 1 Jason.Chimera     L            4   101 0.0396  0.0396 0.0766
## 2 Johan.Franzen     RL           8   105 0.0762  0.0762 0.0924
## 3 Brendan.Gallagher R           13   103 0.126   0.126  0.115
## 4 Taylor.Hall       L           12   106 0.113   0.113  0.109
## 5 Jarome.Iginla     R           10   103 0.0971  0.0971 0.102
## 6 David.Krejci      C           17   106 0.160   0.160  0.130
```

James-Stein estimator, implemented

How to judge estimation accuracy?

- Let's compare to career shooting percentage through March, 2016
- Each player with at least 200 shots
- In principle, a player's career % represents something closer to the truth (his true %)

Comparing the estimates

Mean absolute error

```r
first_players1[1:3,]
```

```
## # A tibble: 3 x 5
## # Groups:   Name [3]
##   Name                 ShP Shp_MLE Shp_JS Shp_Career
##   <chr>              <dbl>   <dbl>  <dbl>      <dbl>
## 1 Jason.Chimera      0.04    0.04   0.077      0.076
## 2 Johan.Franzen      0.076   0.076  0.092      0.083
## 3 Brendan.Gallagher 0.126   0.126  0.115      0.095
```

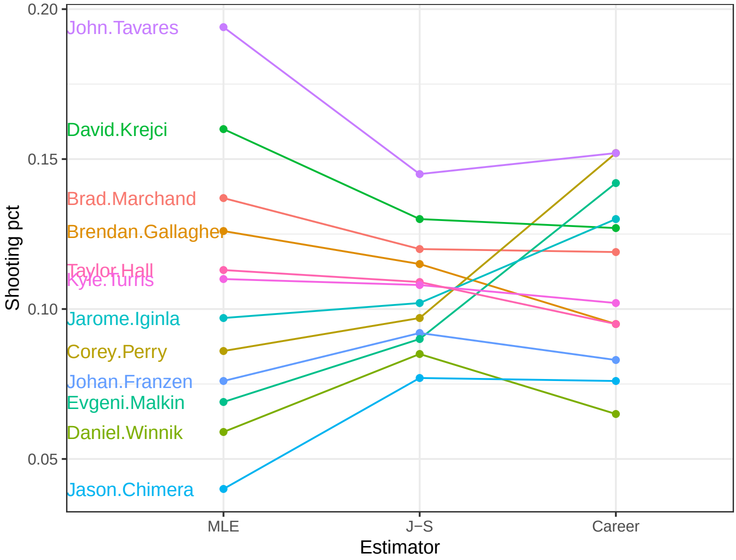
Comparing the estimates

```
first_players1 %>%
  ungroup() %>%
  mutate(abs_error_mle = abs(Shp_MLE - Shp_Career),
         abs_error_js = abs(Shp_JS - Shp_Career)) %>%
  summarise(mae_mle = mean(abs_error_mle),
            mae_js = mean(abs_error_js))
```

```
## # A tibble: 1 x 2
##   mae_mle mae_js
##     <dbl>  <dbl>
## 1  0.0309  0.018
```

How'd we do? How to interpret these numbers?

Visualizing the J-S estimator



Summary:

1. **Stein's Paradox**: Circumstances in which there are estimators better than the arithmetic average

- `better` defined by accuracy (RMSE - plot this?)
- `better` estimators use combination of individual ones ($k \geq 3$)
- `better` than any method that handles the parameters separately.

2. Bias/Variance trade-off: $\hat{p}_{JS}$ versus $\hat{p}$

Summary:

4. Can be tweaked for different sample sizes.
5. Next step: intervals for future performance
6. Links to Bayesian statistics + empirical Bayes (link)