Lecture 1: A brief history of data visualization

Michael Lopez, 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, $c = 1 - \frac{(k-3)\sigma^2}{\sum(y-\bar{y})^2}$
  - k is number of unknown means
  - σ^2 is variance of individual observations
  - $\sum(y - \bar{y})^2$ reflects variance from mean to mean

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, $c = 1 - \frac{(k-3)\sigma^2}{\sum(\hat{p}-\bar{\hat{p}})^2}$
  - k is number of shooters
  - σ^2 is variance of individual shooter given certain number of attempts
  - $\sum(\hat{p} - \bar{\hat{p}})^2$ reflects variance of mean from shooter to shooter

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, $c = 1 - \frac{(k-3)\sigma^2}{\sum(\hat{p}-\bar{\hat{p}})^2}$
  - k is number of shooters
  - σ^2 is variance of individual shooter given certain number of attempts
  - $\sum(\hat{p} - \bar{\hat{p}})^2$ reflects variance of mean from shooter to shooter
- Plug in $c = 1$: $\hat{p}_{JS} = \hat{p}$
- Plug in $c = 0$: $\hat{p}_{JS} = \bar{\hat{p}}$

James-Stein estimator, translated

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

- $\bar{\hat{p}}$ is average of each players shooting percentage
- $\hat{p}$ is a single players observation
- c is a shrinking factor, $c = 1 - \frac{(k-3)\sigma^2}{\sum(\hat{p}-\bar{\hat{p}})^2}$
  - k is number of shooters
  - σ^2 is variance of individual shooter given certain number of attempts
  - $\sum(\hat{p} - \bar{\hat{p}})^2$ reflects variance of mean from shooter to shooter
- Increases in $\sum(\hat{p} - \bar{\hat{p}})^2$: $c \sim 1$, large amount of player specific information
- Increases in σ^2: $c \sim 0$, less amount of player specific information

James-Stein estimator, implemented

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

```
first <- filter(nhl.data, Season==20122013)
first.players <- first %>%
  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)
```

```
## Source: local data frame [6 x 5]
## Groups: Name [6]
##
##                Name Position Goals Shots        ShP
##               (chr)    (chr) (int) (int)      (dbl)
## 1     Jason.Chimera        L     4   101 0.03960396
## 2     Johan.Franzen       RL     8   105 0.07619048
## 3 Brendan.Gallagher        R    13   103 0.12621359
## 4       Taylor.Hall        L    12   106 0.11320755
## 5     Jarome.Iginla        R    10   103 0.09708738
## 6      David.Krejci        C    17   106 0.16037736
```

12 forwards, each with between 100-106 shots

James-Stein estimator, implemented

```
k <- nrow(first.players)
k
```

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

```
p.bar <- mean(first.players$ShP)
p.bar
```

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

```
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
```