SEAM method built from statcast data

Daniel J. Eck

Background

This lecture builds on the statcast and R Shiny lectures.

We will introduce a method for estimated batted-ball distributions for individual batter-pitcher matchups.

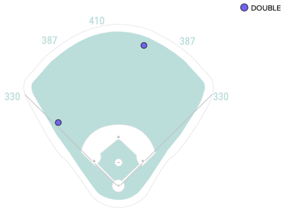
Here is the application:

https://seam.stat.illinois.edu/

This is joint work with your professor and David Dalpiaz and our former students Julia Wapner and Charles Young.

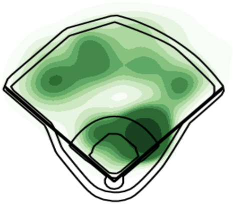
Julia Wapner and Charles Young currently work for baseball teams.

Verlander vs Votto

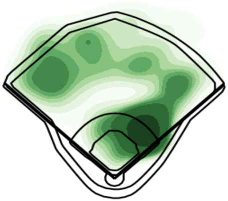


Statcast

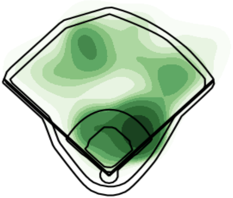
Justin Verlander vs. Joey Votto



Full SEAM



Empirical Batter



Empirical Pitcher

Nonparametric estimation

We will let $\mathbf{x}_p$ and $\mathbf{x}_b$ be pitching and batting characteristics, and let $\mathbf{x} = (\mathbf{x}_p', \mathbf{x}_b')'$.

We will let $f(\mathbf{y}|\mathbf{x})$ be a batted-ball density function where y are 2-dimensional coordinates of batted-balls.

We will estimate $f(\mathbf{y}|\mathbf{x})$ with a bivariate nonparametric Gaussian kernel density estimator

$$\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}) = \frac{1}{n h_{y_1} h_{y_2}} \sum_{i=1}^{n} \phi\left(\frac{y_1 - y_{1i}}{h_{y_1}}\right) \phi\left(\frac{y_2 - y_{2i}}{h_{y_2}}\right),$$

where the sample of batted-ball locations $(y_{1i}, y_{2i}) \in \mathcal{Y}$, $i = 1, \ldots, n$ are conditioned on the characteristics of the players in the matchup.

Similarity scores

The estimator $\hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x})$ is often not feasible for $f(\mathbf{y}|\mathbf{x})$ in practice since there is not enough individual matchup data.

To address this challenge we will incorporate batted-ball data from other matchups involving either the batter or pitcher under study.

Batted-ball locations from such matchups are weighted by their similarity to the players in the matchup under study.

Similarity will be assessed via similarity scores.

Similarity scores

We will suppose that the pitcher in the matchup under study throws n_{type} different types of pitches.

We will let $\mathbf{x}_{p,t}$ be the pitcher covariates for pitch type $t = 1, \ldots, n_{\text{type}}$.

Similarly, let $\mathbf{x}_{b,t}$ be the batter covariates when facing pitch type $t = 1, \ldots, n_{\text{type}}$.

The similarity scores between two pitchers for pitch type t will be of the Mahalanobis distance between covariate vectors, ie

$$s(\mathbf{x}_{p,j_1,t}, \mathbf{x}_{p,j_2,t}) = \exp(-\|\mathbf{x}_{p,j_1,t} - \mathbf{x}_{p,j_2,t}\|_{\mathbf{V}_{p,t}}),$$

where

$$\|\mathbf{x}_{p,j_1,t} - \mathbf{x}_{p,j_2,t}\|_{\mathbf{V}_{p,t}} = \left((\mathbf{x}_{p,j_1,t} - \mathbf{x}_{p,j_2,t})'\mathbf{V}_{p,t}(\mathbf{x}_{p,j_1,t} - \mathbf{x}_{p,j_2,t})\right)^{1/d_p},$$

d_p is the dimension of $\mathbf{x}_{p,j_1,t}$ and $\mathbf{V}_{p,t}$ is a diagonal weight matrix.

Pitch-level density

Let ρ_t is the proportion of time that a ball in play yielded by the pitcher in the matchup under study corresponds to pitch type t.

We will let

$$f(y|\mathbf{x}) = \sum_t \rho_t f(y|\mathbf{x}_t),$$

where $\mathbf{x}_t = (\mathbf{x}'_{p,t}, \mathbf{x}'_{b,t})'$.

SEAM density estimator

For pitcher $j = 1, ..., J$ and pitch-type t, we define weights as

$$w_{p,j,t} = \frac{s_{p,j,t}}{\sum_{l=1}^{J} s_{p,l,t}}.$$

For pitch type t, the spray chart density for a batter facing the synthetic pitcher is

$$f_{\mathrm{sp},t}(\mathbf{y}|\mathbf{x}_{b,t}) = \sum_{j=1}^{J} w_{p,j,t} f(\mathbf{y}|\mathbf{x}_{p,j,t}, \mathbf{x}_{b,t}). \tag{1}$$

The spray chart density for a batter facing the synthetic pitcher is then

$$f_{\mathrm{sp}}(\mathbf{y}|\mathbf{x}) = \sum_{t=1}^{n_{\mathrm{type}}} \rho_t f_{\mathrm{sp},t}(\mathbf{y}|\mathbf{x}_{b,t}). \tag{2}$$

We then estimate (1) with

$$\hat{f}_{\mathrm{sp},t}(\mathbf{y}|\mathbf{x}_{b,t}) = \sum_{j=1}^{J} w_{p,j,t}\hat{f}_{\mathbf{h}_{p,j,t}}(\mathbf{y}|\mathbf{x}_{p,j,t}, \mathbf{x}_{b,t}), \tag{3}$$

where $\mathbf{h}_{p,j,t}$ is a bandwidth parameter.

We estimate the densities in (2) with

$$\hat{f}_{\mathrm{sp}}(\mathbf{y}|\mathbf{x}) = \sum_{t=1}^{n_{\mathrm{type}}} \rho_t \hat{f}_{\mathrm{sp},t}(\mathbf{y}|\mathbf{x}_{b,t}). \tag{4}$$

The synthetic batter density $\hat{f}_{\mathrm{sb}}(\mathbf{y}|\mathbf{x})$ will be estimated similarly.

Our implementation will estimate $f(\mathbf{y}|\mathbf{x})$ with

$$\hat{f}_{\mathsf{SEAM}}(\mathbf{y}|\mathbf{x}) = \lambda \hat{f}_{\mathbf{h}}(\mathbf{y}|\mathbf{x}) + \lambda_p \hat{f}_{\mathsf{sp}}(\mathbf{y}|\mathbf{x}) + \lambda_b \hat{f}_{\mathsf{sb}}(\mathbf{y}|\mathbf{x}), \tag{5}$$

where

$$\lambda + \lambda_b + \lambda_p = 1.$$

Values of $(\lambda, \lambda_b, \lambda_p)$ are chosen to balance the natural bias that exists in our synthetic player construction and the inherent estimation variation.

Our implementation will estimate λ, λ_p, and λ_b as

$$\lambda = \frac{\sqrt{n}}{\sqrt{n} + \sqrt{n_p} + \sqrt{n_b}},$$

$$\lambda_p = \frac{\sqrt{n_p}}{\sqrt{n} + \sqrt{n_p} + \sqrt{n_b}},$$

$$\lambda_b = \frac{\sqrt{n_b}}{\sqrt{n} + \sqrt{n_p} + \sqrt{n_b}},$$

where

$$n_p = \sum_t \rho_t \sum_{j=1}^{J} s^2_{p,j,t} n_{p,j,t},$$

$$n_b = \sum_t \rho_t \sum_{k=1}^{K} s^2_{b,k,t} n_{b,k,t},$$

and $n_{b,k,t}$ and $n_{b,k,t}$ denote matchup sample sizes.

Considerations

Batter characteristics include: exit velocity, launch angle, and pull%, middle%, oppo%

Pitcher characteristics include: velocity, spin rate, movement, release angles, and release point.

Our defaults favor stuff and launch conditions over release information and batted-ball locations.

Pre-processing

- Pitches classified as Eephus, Knuckleball, and Screwball are removed since these pitch types are rare.
- Pitches classified as Knuckle-Curve are renamed to Curveball.
- Pitches classified as Forkball are renamed to Splitter.
- Pitch launch angles are calculated using rudimentary kinematics:
  - $launch_h = \arctan(\frac{vx_r}{vy_r})$
  - $launch_v = \arctan\left(\frac{vz_r}{\sqrt{vx_r^2 + vy_r^2}}\right)$

  where vx_r, vy_r, vz_r are, respectively, the x, y, z components of release velocity.
- Batted ball locations are adjusted to make home plate the origin $(0, 0)$.
- Spray angle is calculated from the x and y coordinates of the batted ball locations (what we previously called y_1 and y_2).
- Data was limited to regular season batter-pitcher matchups.
- Sacrifice hits and sacrifice flies are removed from consideration.

Pitchers and batters are aggregated on a season, handedness, and pitch type basis.

Validation

Why go through all this?

	Marginal coverage			Conditional coverage success rate		
level	0.5	0.75	0.9	0.5	0.75	0.90
SEAM	0.559	0.811	0.953	0.579 (0.035)	0.641 (0.034)	0.779 (0.030)
batter	0.547	0.793	0.938	0.513 (0.036)	0.574 (0.035)	0.662 (0.034)
pitcher	0.541	0.788	0.935	0.549 (0.036)	0.595 (0.035)	0.713 (0.032)

Table 1: Marginal and conditional coverage properties for SEAM, the empirical batter spray distribution, and the empirical pitcher spray distribution (standard errors in parentheses). The conditional coverage success rate is the proportion of individual batter-pitcher conditional coverages that is at or above the nominal level.

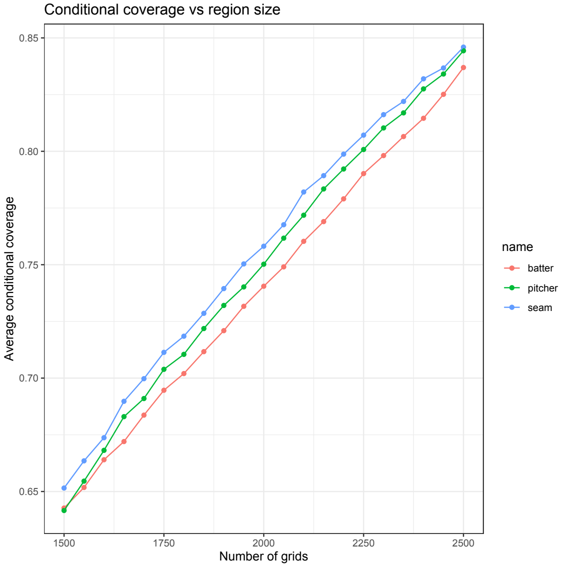


Figure 1: Average conditional coverage of fixed-size confidence regions for SEAM, the empirical batter spray distribution, and the empirical pitcher spray distribution. The fixed-size regions are constructed from the n_g most probable grids according to each method.

There were 129937 baseballs put into play in 2021.

Thus, an increase in coverage by 1% roughly corresponds to an increase of fielding 40 outs per team, and a difference of 40 outs corresponds to about one and half baseball games.

It is reasonable to expect that any baseball team that implements our SEAM method instead of using batter or pitcher specific spray charts could win an additional game over the course of the season.

MLB organizations have repeatedly demonstrated a willingness to spend millions of dollars to obtain a comparable advantage via free agency.

An additional win could be all that is needed to have a chance at competing for a championship.