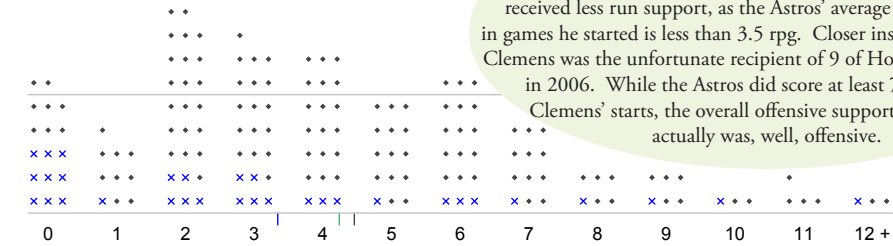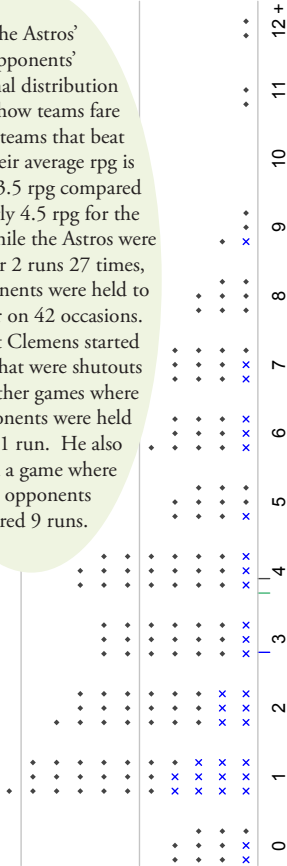# Bivariate Baseball Score Plots

Rafe Donahue, Tatsuki Koyama, Jeff Horner, Cole Beck
Vanderbilt University Medical Center, Department of Biostatistics

The information used here was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at "www.retrosheet.org".

The bivariate score plots present summary information for Major League Baseball teams' game scores.

Each game is represented as one mark in the joint score distribution grid and one mark in each marginal. Splits based on a variety of game parameters (starting pitcher, day/night, home/away, etc.) are available; different values of the split parameter are differentiated by color and shape. Games are shown collected into little groups (in this case, groups are size 3) so as to maintain a rational aspect ratio for the overall plot.

The marginal score distributions are shown on the top for the selected team and along the left side for that team's opponents; also shown via small tick marks on the runs scale are the overall mean runs per game (rpg), along with mean rpg for games meeting, and those not meeting, the split criterion.

Arbitrarily, games meeting the split criterion are placed at the bottom of each stack. Reference lines are drawn to improve one's ability to quickly count games in a column or row. If there are games with scores in excess of the arbitrary maximum (here, 12), a plus sign is added to denote the presence of such games.

The marginals are oriented along the top and left sides so as to facilitate comparison between the marginals. A simple twist of the head allows visual comparison of the two marginals without needing to reverse the positive direction mentally, as would be necessary if the marginals were shown protruding away from the center and located in the traditional bottom and left side positions.

The joint distribution is shown as collected marks in small squares. Victories for the selected team will be below the diagonal, losses above. One-run games will be just above and below the diagonal. Again, games are grouped for ease in counting; the squares are shaded in relation to the number of dots they contain. Thus, more ink means more data. This presents a layered presentation for the data; the overall distribution is visible from afar, while atomic-level datum details are available upon closer inspection. "Reward the viewer for mental and visual investment in the graphic."

Data for the plots were obtained from retrosheet.org. Programming was done using the R environment for statistical computing and graphics. An interactive website is available for examining score distributions of any team in the retrosheet database from 1876-2006 at http://data.vanderbilt.edu/rapache/bbplot/ and commentary and discussion are available at http://baseballplot.blogspot.com.

Abstract: Scores from baseball games are bivariate data, with each variate being one team's score. Information can be obtained from viewing the distribution of baseball scores from a bivariate point of view. We offer a baseball scores plot, with each game being a point in a two-dimensional grid, and an algorithm, database, and website to allow interested users access to the plot.

This example shows 163 games for the 2005 Houston Astros, with games started by Roger Clemens highlighted. The Astros finished the year 16 games over .500 but were 2 games under .500 in games which Clemens started.

Houston 2005

89 – 73 (0.549)
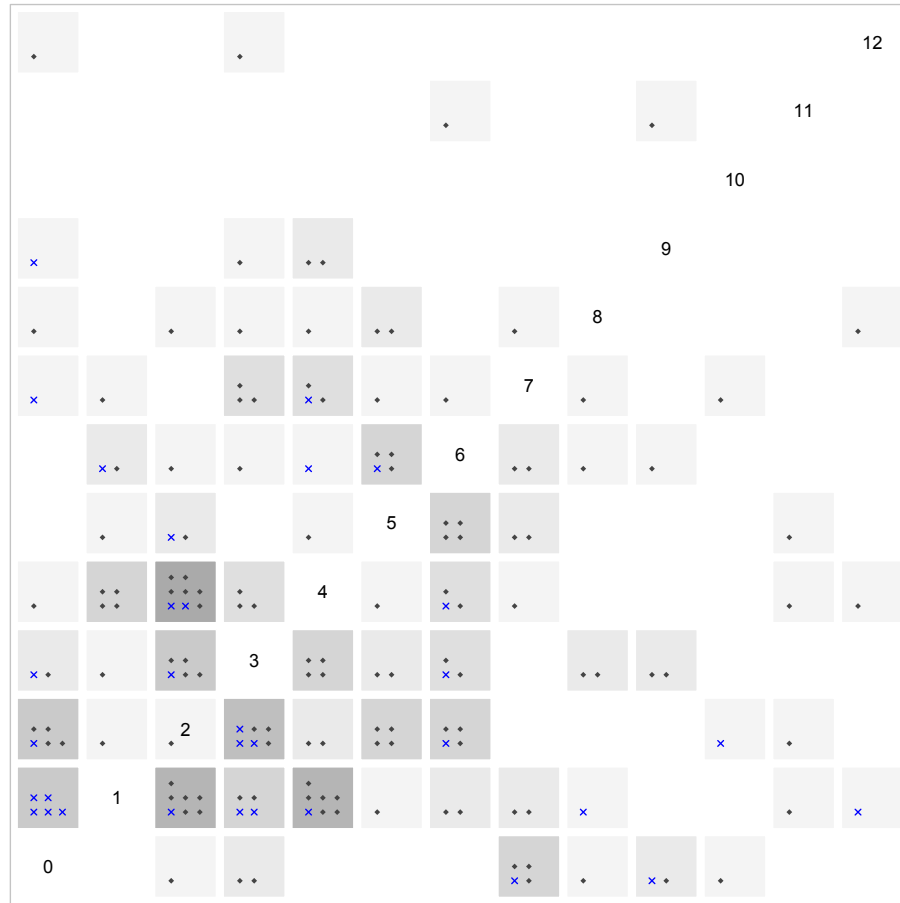
Starting Pitcher

Roger Clemens

15 – 17 (0.469)

The Astros' marginal score distribution shows typical numbers from a good team: an overall average of over 4 rpg. Clemens, however, appears to have received less run support, as the Astros' average offensive output in games he started is less than 3.5 rpg. Closer inspection reveals that Clemens was the unfortunate recipient of 9 of Houston's 17 shutouts in 2006. While the Astros did score at least 7 runs for 5 of Clemens' starts, the overall offensive support for Clemens actually was, well, offensive.

The joint distribution reveals details of Clemens' abysmal run support. The bottom-left corner of the distribution shows *five* games which Clemens started in which the Astros lost 1-0, a pitcher's nightmare. So, of the 11 games that Clemens started and the opponents were held to one run, 5 of those games failed to produce a single Houston run. In fact, Clemens was the only Astros pitcher to start a game in which the team lost 1-0.

The joint distribution reveals a rather ordinary overall record of 25-21 in one-run games, a measure often heralded as a mark of good teams.

The Astros' opponents' marginal distribution shows how teams fare against teams that beat them: their average rpg is just over 3.5 rpg compared with nearly 4.5 rpg for the Astros. While the Astros were held under 2 runs 27 times, their opponents were held to 1 or fewer on 42 occasions. Note that Clemens started 2 games that were shutouts and 11 other games where the opponents were held to just 1 run. He also started a game where the opponents scored 9 runs.

The keen eye will note a single game on the diagonal, a 2-2 tie. Prior to 2007, such games that were tied but suspended were kept on the books for purposes of individual statistics, but were replayed at the next available opportunity.

Opponents

Astros