Journal of Statistical Software

March 2018, Volume 83, Book Review 3. *doi: 10.18637/jss.v083.b03*

Reviewer: Tim Downie
Beuth University of Applied Sciences, Berlin

Visualizing Baseball

Jim Albert
Chapman & Hall/CRC, Boca Raton, 2017.
ISBN 978-1-498-78275-3. 142 pp. USD 29.95.
https://www.crcpress.com/9781498782753

Visualizing Baseball explores the ability to analyze baseball statistics through graphical methods. It follows two previous books by Jim Albert on the subject of baseball statistics: *Curve Ball* (Albert and Bennett 2001) and *Analyzing Baseball Data with R* (Marchi and Albert 2013). It is a little surprising that this subject appears in the *ASA-CRC* Series on *Statistical Reasoning in Science and Society* as the subject matter is extremely domain based. The reader is required to have a considerable background knowledge of US Major League Baseball and only a basic knowledge of descriptive statistics. Correspondingly, this book will appeal to baseball fans with an interest in data analysis, but not to statisticians and data scientists interested in data visualization, who only occasionally watch the game. As an example, the statistic *on base percentage (OBP)* is defined using a formula involving five other unexplained abbreviations. In cases where a metric is explained using words, the description can be imprecise, albeit clear enough for followers of baseball statistics (e.g., the statistic *plate appearances* is not exactly what one might at first expect). One can argue that this approach is appropriate for a book of this type, but it undoubtedly reduces the potential readership. For those who are in the target audience, the book is interesting, informative and easy to read.

The primary aim of the book is well executed. In almost all cases the graphics are well thought out, and quickly communicate characteristics in the data, that would be difficult to convey using tables or summary statistics. The quality of the diagrams and the printing is very good. Each diagram comes with a clear explanation and many of the results are demonstrated using well known players. Most of the graphics are produced using the commands in the in R package **ggplot2** (Wickham 2009). The R code for all the graphics and the markdown generated HTML files are available via the author's GitHub site (Albert 2017). All the diagrams are restricted to two colors, black and purple. It is unclear if this was restriction made by the publisher or was the author's choice, in order to promote simplicity over prettiness. In most cases using only two colors is appropriate, but a few diagrams would be improved by using more colors. For example, a scatter plot with axes *proportion of swinging strikes* and *proportion of first pitch strikes*, displays a third dimension *wins above replacement* proportional to the size of each point. I find this third dimension is difficult to perceive; it would be easier to visualize the third variable using a standard graduated color scheme instead.

A particularly strong point throughout is that the data is as up to date as possible for a text book. In most cases the cut off date is the end of the 2016 season, but in some cases data up to the halfway point in the 2017 MLB season is analyzed. The data also come from a variety of sources and some of the datasets are very detailed, such as *PITCHf/x* data (type of pitch thrown and location of pitch), and the *StatCast System* (angle and speed of the ball as it leaves the bat).

Visualizing Baseball consists of nine short, self contained, chapters, each one investigating two or three aspects of that chapter's subject. In order to illustrate the type of topics covered in the book, I briefly discuss three of these topics, which I found particularly interesting.

Runs expectancy is an appealing concept, which is discussed in detail in *Curve Ball* (Albert and Bennett 2001). Runs expectancy is defined as how many runs one can expect in the remainder of the inning, given the current inning status in terms of base runners and outs. These expected values are displayed graphically for all the inning statuses with runs expectancy plotted on the y-axis. This concept is then extended to the *value of a plate appearance*, which is the difference between runs expectancy after and before the completed plate appearance. The value of some hypothetical plate appearances are illustrated by adding an arrow to the runs expectancy scatter plot. The arrow leads from the before inning status and points to the after inning status. The impact of a strikeout when there is a runner on first and no outs is visually conveyed by a downward arrow corresponding to a drop in about 0.4 expected runs. There follows two further examples in which runs are scored in the plate appearance. In these cases some run expectancy has been converted into actual runs on the scoreboard, so the value of runs expectancy decreases. This is property is explicitly mentioned by the author, but the diagrams are nevertheless visually misleading, as they suggest a bad outcome. In these examples the arrow also points downwards, often by a larger amount than in the previous strikeout example. This visual anomaly could corrected by plotting "expected *total* runs in inning" instead of "expected runs in the remainder of the inning".

The eighth chapter, *Probability and Modeling*, contains a section called *A Baseball Season.* This investigates how likely it is that a poor or average team wins the world series and suggests that this has become more likely in the last 50 years, due to the change in the playoff structure. The analysis is based purely on a simulation study and the assignment of poor team, great team etc. is itself simulated via a normally distributed *talent score.* The simulation results are interesting, but, unlike all of the other sections, is not based on any data; some readers could easily overlook this. Assigning talent scores to real baseball teams, simulating their progress according to different season structures and comparing these results with the actual playoff teams or world series winners would be an interesting extension to this section.

The final chapter deals with *Streakiness and Clutch Play.* Streakiness corresponds to a long uninterrupted sequence of an outcome, such as how many of plate appearances a batter has had since his last hit. A "clutch situation" means this plate appearance will have a relatively big influence on the final result of the game. Some batters have a reputation for being "clutch hitters", that is they perform particularly well in the important game situations. In analyzing streakiness and clutch play, the challenge is to differentiate between genuine player skill and mere randomness, something which fans and journalists regularly underestimate. An ingenious method of calculating the clutch measure for each player is presented, based on the principle that a clutch neutral player's performance is independent from the game situation but a good clutch hitter's performance will be better than his average performance, when batting in a clutch situation. By randomly permuting the batter's game situations the extent of this dependency on game situation can be measured. The *clutch measure* is computed for each player in the 2016 season, followed by a discussion of the players with very high clutch measures. It would be easy to over-interpret these results, as the clutch measure used here is a measure of historical performance but not of a player's clutch potential. The potential problem is, that players with very high or low clutch measures will only be found in players who frequently bat in a clutch situation, which in turn is dependent on team strategies and other player skills. An example is a batter in the fourth spot in the line-up will have more clutch opportunities than a lead off batter playing in the National League. A lead off batter might be a very good clutch batter but will have fewer opportunities to push up his clutch measure.

In summary, this is a well presented and interesting book, demonstrating how good graphical presentations can efficiently increase the understanding of a dataset and its domain field. The required background knowledge of baseball is high, which would make the book unaccessible to many statisticians, who would otherwise find the data visualization aspects interesting.

References

Albert J (2017). "Visualizing Baseball." URL https://bayesball.github.io/VB/.

Albert J, Bennett J (2001). *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game*. Springer-Verlag.

Marchi M, Albert J (2013). *Analyzing Baseball Data with R*. Taylor & Francis.

Wickham H (2009). ***ggplot2**: Elegant Graphics for Data Analysis*. Springer-Verlag, New York. URL http://ggplot2.org/.

Reviewer:

Tim Downie
Beuth University of Applied Sciences
Department II
D-13353, Berlin, Germany
Email: tim.downie@beuth-hochschule.de

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
March 2018, Volume 83, Book Review 3 *Published:* 2018-02-22
doi:10.18637/jss.v083.b03