 Due by 3:00pm, 30^{th} Oct 2020

CMPSC 301
Data Analytics
Fall 2020

Lab 5:
Analysis of the Pitcher-Batter Relationship in Baseball

Figure 1: Baseball analytics helps the coaches plan better strategies for winning games. Note: A single research question is often impossible to directly respond to the underlying phenomenon that makes up the pitcher-batter relationship in baseball. To determine a broader view of how to understand the relationship to make informed decisions about placing players in your game, you will often need to approach this central question by responding to several smaller questions which are still directly related.

Objectives

In baseball, it has been said that there is a *relationship* between the batter and the pitcher that may be understood to allow for one of the teams to gain a particular advantage. For example, it has often been speculated that a pitcher generally has an advantage when his handedness is the same as the batter's, and that the batter is likely to have the advantage when they are opposite. *Is there any truth in this speculation?* There are likely to exist other types of speculations that may give one player an advantage over another.

In this lab, we are going to use our skills in data analysis to argue that an advantage may exist for one player or team, over another player or team. This advantage, we argue emerges from a relationship between the pitcher and batter that occurs during the game.

To make our case, we will find a public baseball data set online and explore it to create plots of its variables to speculate that there *is* some relationship that exists between batters and pitchers. This relationship, we will argue, could eventually be used to form a strategy for one of the teams.

Remember, you may be the first person to ever see the results that you find in this lab!

Speculation From Research Questions and Data Analysis

Specifically, in the lab you are to propose and to answer exploratory research questions by studying two or more variables of a data set for baseball that you obtain yourself from online searches. Take care to propose well-though-out questions that may be *answered* by the data that you obtain.

Once you have devised your research questions which are in the scope of the data (i.e., can be *answered* from studying the data), you are to conduct an analysis and provide plots from rStudio to suggest that a relationship may actually exist. Notably, your plots are to argue that there is some *non-random event* happening that may be used to describe a pitcher-batter relationship, that could eventually be used to form a strategy in the sport. In your report, you will justify your plots, observations and thinking.

Reading Assignment

Please read the course chapters, slides and notes for reference in this work. There are other articles in the bibliography of this assignment that require your attention. In addition, take some time to gain experience with using Markdown to write your work. See *Mastering Markdown* `https://guides.github.com/features/mastering-markdown/` for more details about Markdown. In addition, you may view the department's own video ("Markdown Tidbits"), on how to produce Markdown content, on our YouTube channel at link; `https://www.youtube.com/playlist?list=PLsYZRXov75ZHSwWiCk0-jd1RcTuu_-zmD`.

GitHub Starter Link

https://classroom.github.com/a/Pq5AReUM

To use this link, please follow the steps below.

- Click on the link and accept the assignment.
- Once the importing task has completed, click on the created assignment link which will take you to your newly created GitHub repository for this lab.
- Clone this repository (bearing your name) and work on the lab locally.
- As you are working on your lab, you are to commit and push regularly. You can use the following commands to add a single file, you must be in the directory where the file is located (or add the path to the file in the command):
  - `git commit <nameOfFile> -m ''Your notes about commit here''`
  - `git push`

  Alternatively, you can use the following commands to add multiple files from your repository:
  - `git add -A`
  - `git commit -m ''Your notes about commit here''`
  - `git push`

A Relationship Between the Pitcher and Batter?

Imagine that that a coach for a major league baseball coach comes to you to ask how you could help his or her team achieve a better strategy for playing baseball. As mentioned in Figure 1, this strategy would have to come by way of an analysis of data of seasons of game-play...

In baseball, a pitcher throws a ball to a catcher who stands behind a batter. The batter attempts to hit the flying ball with a bat before the catcher can trap the ball. To prevent the batter from hitting the ball and scoring points for his or her team, the pitcher attempts to throw the ball in a way that complicates the batter's ability to negotiate a ball-hit. The pitcher and the batter in this scene are the actors that perform in the play of the pitcher-batter relationship. This relationship may be one of the most important factors to determine a winning outcome in a game and so the types of pitchers and batters are especially chosen by each team to play each other where each has some form of presumed advantage over the other. Here we will spend some time to investigate this scene using data obtained from public free online sources of baseball data.

Handedness

Work has been completed to compare the pitching biomechanics between left- and right-handed professional pitchers. In Solomite *et al.* [1] and Diffendaffer *et al.* [2] some of the kinematic and kinetic differences that exist between left- and right-handed pitchers were studied. In their work, both research groups they find that there were notable biomechanical differences stemming from elbow movements, in addition to foot and leg position, stance, shoulder movements and the position of other body parts, that contrasted between the pitches of left- and right-handed players.

On the other hand, left-handed batting in baseball has been shown to provide batters an advantage in winning batting titles over right-handed batters [3] and discussed in the *SportsRec* blog at `https://www.sportsrec.com/5950567/left-handed-batter-vs-a-right-handed-pitcher`. Interestingly, seven of the top 10, and 19 of the top 30, are left handed. Discussed in Walsh [4], most batters hit right-handed and the observation that a majority of these top-hitters (making the leaderboard ranks) hit left-handed is of great interest. This observation may be explained by the phenomenon that batters tend to hit better when facing an opposite-hand pitcher, however, further exploration needs to be completed in order to completely understand how this pitcher-batter *incompatibility* can be harnessed into a working strategy for baseball.

Analysis of a Relationship

The above discussion caters to a specific interest in baseball analytics where the batter-pitcher relationship has been reduced to factors of left- and right-handedness. There are likely to be many, many other types of investigations into baseball data which may explain other incompatibilities between the batter and pitcher and may help to inform strategies for placing batters in game against particular pitchers.

In this lab, your task is to investigate the **pitcher-batter relationship** using exploratory data analysis. In particular, you are to ask **deep, far-reaching**, questions of the data and to use your analysis skills to answer some of these questions. Note, that you will have to find this data and then examine it to be sure that your questions can actually be answered by the analysis of the data. Your analysis and results will then be used to inform the hypothetical coach (from above) to help him or her to consider the advantages of particular pitchers and batters, in effort to win games. Note: your work does not have to concentrate on the left- and right-handedness of players, and you are encouraged to pursue seemingly any type of investigation of your data to study the pitcher-batter relationship.

Links to Data Sources

You are to spend some time researching sources of data online which will be helpful to your investigation of the pitcher-batter relationship. Remember, if the questions cannot be answered from the data set, then either change your approach to answering your over-all research question, or change the data set that you are using for your investigation. Below are some suggestions of sources of data. Whether you find another source of data or use one of the below sets, **please be sure to cite the source in your work**.

- `http://www.seanlahman.com/baseball-archive/statistics/`
- `https://Baseball-Almanac.com`
- `https://sabr.org/how-to/statistical-databases-and-websites`
- `http://m.mlb.com/statcast/leaderboard#exit-velo,r,2019`
- `https://www.kaggle.com/`
- `https://www.baseball-reference.com/`
- And plenty more from online searches!

Ethical Reasoning

Your response to the below question is to be added to the end of your `writing/report.md` file submission.

In athletics, the use of steroids and other performance enhancing drugs to enrich a player's performance is largely prohibited throughout the world. The use of performance-enhancing drugs, as discussed in [5], is a controversial subject and can lead to a player's demise in sports. When players are found to have taken these agents, they are accused of cheating. In baseball, the subject of doping has been explored by Mallon in [6]. On the other hand, there is still data analysis in sports where teams may learn of informed strategies to gain wins during their season.

1. Discuss why performance-enhancing drugs is considered cheating in baseball, and in other sports as well. In clear and meaningful language, discuss how the application of data analytics to sports may or may not be considered cheating since its goal is to add some competitive advantage to a game.
2. Discuss two scenarios where ethics may be in question as a result of data analysis. Note: your response will amount to about half to a full page of written text.

Required Deliverables

Note: Please remember to include your name on everything you submit for the class.

1. **Report**; `writing/report.md`: Your report should not be more than two or three pages of text (not counting graphics), and is to have the following parts.
   (a) **Citations**: Please add a reference to your obtained data in your report. Please give a title, link and other information about the data as a reference.
   (b) **Research questions**: Your report of three or more (well-thought out) research questions which can be answered by the data that you have obtained. In this report, you are to describe your thinking behind the three or more questions and to explain why your obtained data is able to answer these questions.
   Note: If you are not convinced by your rigorous analysis that a direct relationship exists between a pitcher and batter in your data set, then in your report you could argue for the existence of a different kind of relationship, or for the existence of an *idea* in the data that may prove to be a relationship with further exploration and analysis of your ideas.
   (c) **Plots**: Show and explain your plots that you used to respond to your questions using at least two variables from the study of the pitcher-batter relationship. Explain why you choose those variables. You are also to explain how the plots are able to study some part of this phenomenon behind your research question(s). Note: there is markdown code included with your report document to help you add graphics to your work.
   (d) **Ethical Reasoning**: Your report is to have your response to the ethical reasoning questions, noted above.
2. **Dataset**; `data/*`: You are to store your versions of data in this directory to preserve it for a later analysis.
3. **Source code**; `src/analysis.r`: Your code that can be run to load the data files and to produce the plots of your work. Please add documentation to your code to help the instructor understand your thinking behind the code on a line-by-line basis.

When you have finished, please ensure that the GitHub web site has your pushed work by visiting your repository at the site. Please see the instructor if you have any questions about assignment submission.

References

[1] M. J. Solomito, J. V. Ferreira, and C. W. Nissen, "Biomechanical differences between left-and right-handed baseball pitchers," *Sports biomechanics*, vol. 16, no. 2, pp. 143–151, 2017.

[2] A. Z. Diffendaffer, G. S. Fleisig, B. Ivey, and K. T. Aune, "Kinematic and kinetic differences between left-and right-handed professional baseball pitchers," *Sports biomechanics*, vol. 18, no. 4, pp. 448–455, 2019.

[3] D. M. Brown, Z. A. Poucher, M. Myers, J. D. Graham, and J. Cairney, "Sinister right-handedness provides canadian-born major league baseball players with an offensive advantage: A further test of the hockey influence on batting hypothesis," *PloS one*, vol. 14, no. 8, 2019.

[4] J. Walsh, "The advantage of batting left-handed," *Retrieved October*, vol. 25, p. 2010, 2007.

[5] W. Andreff, "Doping: Which economic crime in sport?" in *An Economic Roadmap to the Dark Side of Sport*. Springer, 2019, pp. 55–90.

[6] A. Mallon, "The communication of cheating: A rhetorical analysis of the communication major league baseball players use when accused of taking performance enhancing drugs." 2017.