CMPSC 301
Data Analytics
Spring 2020

Lab 4, Part2:
Group Work to Explore The Pitcher-Batter Relationship
in Baseball Using Data

Objectives

Part 1

The objectives of this group work are to gain practice in asking exploratory questions concerning the pitcher-batter relationship in baseball, and to use data analysis to respond to these questions and to extract new relationships among data variables. During the first part of the lab you will obtain the data, that you will pre-process as needed, and then will do exploratory data analysis. Then, during the second part of the lab, to be released next week, you will conduct a more in-depth statistical analysis of your data to respond to your questions.

Specifically, in the first part of the lab you are to answer your exploratory questions by studying two or more variables of your baseball data set. The data, which will be analyzed using R code, will be obtained from online sources. After you have your data, you may decide to wrangle your data, i.e., to transform the obtained data in to some suitable format for your work. Finally, you are to conduct the analysis and provide plots of the variables from your study of data that respond to your research question(s). Notably, your plots are to argue that there is some *non-random event* happening that may be used to explain the pitcher-batter relationship. In your report, you will justify your observations and thinking.

Part 2

In Part 2 of this lab, you will be gaining some experience in working with statistical tests (taken from class notes or from your own research online) that you will apply to your data sets from Part 1. These tests will help to put some numerical measurements on the observations described by your plots of Part 1. You will also respond to an ethical discussion question.

Reading Assignment

Please read Chapters assigned for this week's lessons which you will find in the class slides, in addition to reviewing your notes. Please 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 consult your notes on the talk given by Assistant Coach Dean Peterson of Allegheny Baseball who discussed motivations and methods for the study of this data.

Handed out: 2nd March 2020

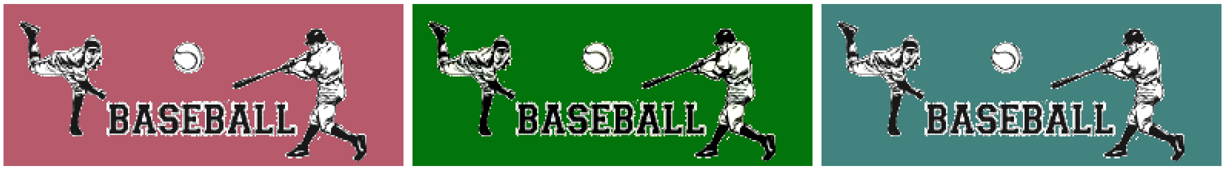

Figure 1: Baseball analytics helps the coaches plan better strategies for winning games. Note: A single research question is often impossible to directly respond to the underlying phenomenon that makes up the pitcher-batter relationship in baseball. To determine a broader view of how to understand the 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.

GitHub Starter Link for Groups

STOP! STOP!
Not everyone will be clicking this link at this time!
Only the team leader will be clicking the link to create the repository!!
`https://classroom.github.com/g/JxrCNON4`

Creating Your Repository

We will use a group assignment functionality of GitHub Classroom for this assignment. For group assignments **only one person will be creating the team while the other team member will join that team.** Please form a team of **no more than two people** and select one person to create the repository.

The selected person of the team should go into the link to the lab in the assignment sheet. Copy this link and paste it into your web browser. Now, you should accept the laboratory assignment and create a new team with a unique and descriptive team name (under “Or Create a new team”).

Now the other members of the team can click on the assignment link and select their team from the list under “Join an Existing Team”. When other team members join their group in GitHub Classroom, a team is created in our GitHub organization. Every team member will be able to push and pull to their teams repository.

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 locally
- As you are working on your lab, you are to commit and push regularly. The commands are the following.

```
– git add -A
– git commit -m ‘‘Your notes about commit here’’
– git push
```

Introduction

In Part 1 of this lab, you, working in a group, were to have come-up with some (*answerable*) research questions to help you explore and understand the *pitch-batter relationship*. Each member was to have his or her own contributed question and, together, these research questions were to support a larger question which may not be directly answerable without its smaller inquiries. The larger question was to have some interest in the Pitcher-Batter Relationship and the comprehension of this relationship would serve to inform of strategies that may help a team gain more wins during a season.

A “Larger-Question” In My Research?

So what about that, above-mentioned “larger question” idea from above? In analytics, research begins with an over-all, blanketing question that asks some general form of a research interest. Examples of such questions for this lab may be the following.

- *What types of incompatibilities exist between a pitcher and a batter which may help the batter to hit more home-runs?*
- *Can the pitcher’s throws be predicted by left- or right-handedness?*
- *Should we pair a pitcher who is likely to throw fast-balls with a batter who specializes at hitting fastballs? Who are the players with these skills?*

To investigate these lofty over-arching goals, smaller, conveniently manageable questions are used. The idea is that smaller questions can be used to explain parts of the whole and then, once the smaller questions have conclusions, then this knowledge can be combined to be used to explain and to draw informed conclusions about the larger research question.

Apply Numerical Statistics to Measure Plots

By this stage of the lab, you are to have determined an over-arching goal of your study, posed several smaller research questions, found and obtained relevant data, and have have amassed proof from plots of non-random events from at least two variables per plot. While it is important to have visual information to determine that a non-random signal is being made in the data, it is not a sufficient form of evidence of the signal. To secure this evidence, you must use statistical tools to measure the signal and in this part of the lab, you will be applying some these statistical analyses to the variables of your plots. Some of the statistical tests and concepts that we have covered in class are the following.

- Min, Max, Mean, Mode, Median, and Quantiles
- Correlation, Variance, Standard Deviation, t-Test

For instance, if your plot showed that there were at least two variables of your data that appeared to have a visual correlation, then in Part 2, you are to introduce a statistical correlation study to measure the actual amount of correlation that you observed. Is it suggested that you consult your recent slides and textbook for ideas about which statistical tests in addition may be helpful when studying correlation.

Several Statistical Tests May Be Necessary

Often, you will have to apply several different statistical tests simultaneously to explain a the observations in a plot. For example, one is often drawn to use mean (average) values to explain an observation in a plot. However, this mean-value is unlikely to be sufficient to show any relevance. To properly form your argument that there is some relevant observation in the plot, other tests will be also necessary which you can find in your `classDocs/` lessons or, by your own searches online.

Depending on the type of question(s) that you are asking, you may have to implement your own statistical tests. Such tests may concern; frequencies, proportions, range, linear models from smooth lines, and many others which may be discovered using online searches. **Please remember, you are to justify each test that you apply to your data: Please explain why a particular test was chosen and what the measurement allows us to learn about the original plots.**

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 [1], 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 [2]. On the other hand, there is still data analysis in sports where teams may learn of informed strategies to gain wins during their season.

With the help of your group, 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 is not considered cheating, although it does provide a competitive advantage. Discuss two scenarios where the unethical application of analyses may cause damage to the game, its players or to the tradition of the sport itself. Note: your response will amount to about half to a full page of written text.

Required Deliverables

1. From Part 1 (last week):
   (a) Report, `writing/report.md`: Your report of three or more (well-thought out) research questions which can be answered by the data you have obtained. In this report, you are 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. You are to explain how the plots are able to study some part of this phenomenon behind your research question.
   (b) Dataset, `data/*`: You are to store your versions of data in this directory to preserve it

for part2.

(c) 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 what it is doing on a line-by-line basis.

2. From Part 2 (this week)

   (a) Additional Source Code, `src/analysis.r`: You are to add to your code to include the statistical tests that you applied to at least two of the variables of each plot. Please label the Part 1 and Part 2 sections in your source code.

   (b) Additional parts in your Report, `writing/report.md`: Please be sure to add your discussion about the new statistical tests along with a short (and educated) justification about why the test was chosen to measure the non-random of a plot. Please label the Part 1 and Part 2 sections in your report.

   (c) Ethical question: See the question-in-blue above. Your response to this ethical question is to be added to the end of your report submission in `writing/report.md`.

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

[2] 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.