Does a Pitcher Care About His Earned Run Average?

Kevin Ferris

2015-04-05

Introduction

Baseball fans, players, and analysts go to the park hoping to see their team win (Keri and Prospectus 2007). But they do not abandon their allegiances to their team the moment they leave the ballpark. They continue to discuss the game, team, and players in offices, bars, households, and blogs across the country.

These discussions vary from favorite players to in-game tactical decisions with heated arguments developing on even decisions in single at bats (cite some of the bunt discussions from Ned Yost in Game 7?). Another popular discussion topic is attempting to determine how much each individual player contributed to the team's success or failure. But assigning the credit or blame is difficult because all the players on the team were working together to achieve the win. So fans and analysts, create statistics which summarize the individual contributions from each player.

These statistics date back to the very beginning of baseball when (whatshisname) produced the first box score recording the number of hits, walks and runs batted in for each position player (**???**). For pitchers, he recorded the (innings pitched, ERA, etc?). Statistics have been continually introduced into baseball as fans and analysts try to find more precise ways of acknowledging a player's contribution. (State when errors, sacrifices, double plays, holds, etc were first recorded). In the 1970's, an analyst named Bill James started closely examining these statistics to see which were most important. This resulted in the rise of "sabermetrics" — using quantitative analysis to provide better statistics rather than just trying to record everything that happened in the game. This movement was championed by the early 2000s Oakland Athletics who became the most well known example of a team successfully incorporating analytics into their decision making (**???**). Because of this sabermetric revolution, it is fair to say that baseball fans, analysts, and players have never had access to the number or quality of statistics as we have access to today.

The potential problem with statistics that attempt to discern individual performance from a group result is that the statistics could incentives for the individuals to take actions that inflate their statistics at the potential expense of the group's performance. For example, it may be beneficial for an outfielder to attempt a diving catch when the team would prefer him to play it safe and not risk missing the catch. From the outfielder's point of view, his diving catch could be shown on ESPN's Web Gems segment or may otherwise catch the attention of people who would not remember him safely letting a ball drop. The fundamental objective of this paper is to examine whether one of the commonly used statistics presents an incentive for pitchers that is not beneficial from the team's point of view.

Explanation

One of the statistics that people have been using for decades to evaluate pitchers is Earned Runs Average (ERA) (cite?). A pitcher's job in baseball is to prevent the opposing team from scoring runs. ERA is used to measure the number of runners that the pitcher is responsible for (a pitcher is responsible for a runner if that batter gets a hit or a walk off of the pitcher) and scales it to account for the number of innings pitched. While ERA values have changed over time, an ERA of 3 has typically been the mark of a good pitcher, an ERA of 4 is average, and an ERA above 4.7 is poor.

The strange thing about this statistic is that only accounts for batters who scored *that a pitcher is responsible for* rather than all batters who the pitcher faces. A pitcher is deemed "not responsible" for a batter if, for example, a fielder makes an error and allows the runner to reach base. Because it is not the pitcher's fault that the runner reached base (due to the fielding error), the theory behind ERA is that the pitcher should not be penalized if that runner scores.

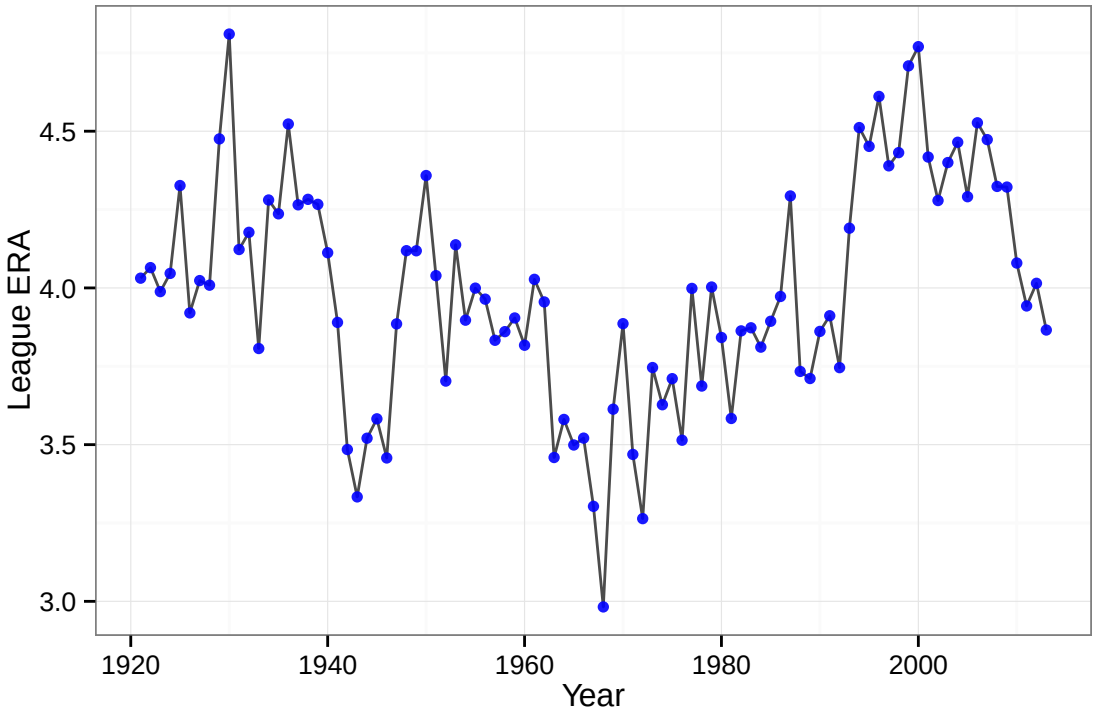


Figure 1: Changes in pitcher ERA for the entire MLB over time.

However, this creates a potential problem with the incentives present for the pitcher: because ERA is one of the primary tools used to evaluate pitchers, in the situations when ERA does not apply the incentives are not as strong. *A fair amount of research has been done to evaluate the effects that incentives have in sports (come back and discuss more thoroughly).*

We theorize that pitchers may pitch differently when ERA applies (when the pitcher is responsible for the runners) than when it does not. Specifically, we believe that when a pitcher is not responsible for the men on base, he will be more likely to pitch in "dangerous locations" — over the middle of the plate or far from the strike zone. Conversely, when a pitcher is responsible for the men on base, he will be somewhat more likely to pitch around the edges of the strike zone. *May have to come back and think about these expectations a bit more*

Maybe a graph like the one here? Now, the z-axis would just be pitch frequency.

Background

The difficulty of assigning credit or blame to individuals when success is measured on the group level is not unique to baseball. Firms measure their success by the company's profit line, but they must internally decide which employees helped or hindered the profit most of all.

Hopefully talk to Mark Anderson and find some examples of research on the effect of incentives at firms.

This paper will add to the literature by providing another example of how misaligned incentives can have unanticipated effects. It shows that these exist even in the high stakes of a professional sporting event. It also shows that they these influences can be seen even though every moment of a baseball game is heavily scrutinized by both professional and casual analysts.

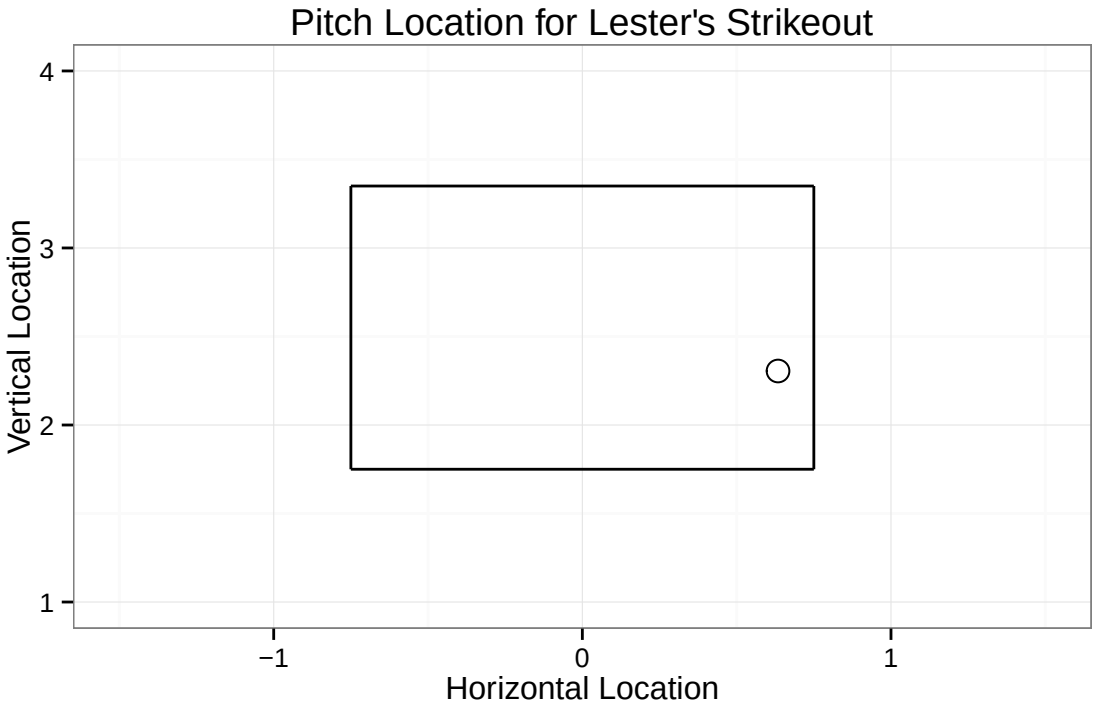


Figure 2: Location of final pitch from Lester to Lind. The black box denotes a typical strike zone. This plot is drawn from the catcher's point of view so Lind would be standing off to the right.

Data

Raw

Starting in the 2006 playoffs, Major League Baseball began using pitch tracking cameras to record information about each pitch thrown in a game. By 2008, these cameras were implemented in every ball park in the majors. These cameras track velocity, movement, release point, spin, and pitch location at the moment each pitch enters the strike zone. MLB's Gameday API provides these data (known as PITCHF/x data) for free in an XML format online. They also provide additional data such as balls, strikes, outs, and runners on base. For more information on the data provided by PITCHf/x, see (Fast 2010).

For example, on May 10, 2013 Jon Lester struck out Adam Lind with an 88 mph cut fastball to end the game. The pitch was located low and inside on the left-handed Lind. The video can be found here. Using the coordinates provided by the PITCHF/x data, the location of the pitch when it enters the strike zone can be plotted. While not shown here, the data also tell us that there were two balls, two strikes, to outs, and no men on base when this pitch occurred.

Using the `pitchRx` R package (C. Sievert and Sievert 2014), the regular season PITCHF/x data from 2014 were downloaded in April, 2015.

Cleaned

The question of interest for this project is whether baseball pitchers pitch differently when they are responsible for the runners on base. We began by extracting all the pitches thrown when runners were on base. Therefore, in our data, each observation is a single pitch that occurred when there was at least one runner on base. For

each observation, we recorded whether the pitcher was responsible for the men on base. In situations where there were multiple men on base and the pitcher was responsible for some but not all of them, the pitcher was deemed responsible.

Does this make sense?

A pitcher was deemed responsible if, according to the MLB Official Rules, the pitcher would not be held accountable if the runner had scored. There are several circumstances under which a pitcher would not be held accountable. They are briefly summarized below. For further explanation, see (http://mlb.mlb.com/mlb/official_info/official_rules/official_scorer_10.jsp).

1) batter reaches on a hit or otherwise after his time at bat is prolonged by a muffed foul fly (can't actually determine this from PITCHf/x)
2) batter reaches because of interference or obstruction
3) batter reaches because of any fielding error
4) the inning is prolonged because of a fielding error
5) a relief pitcher inherits runners on base

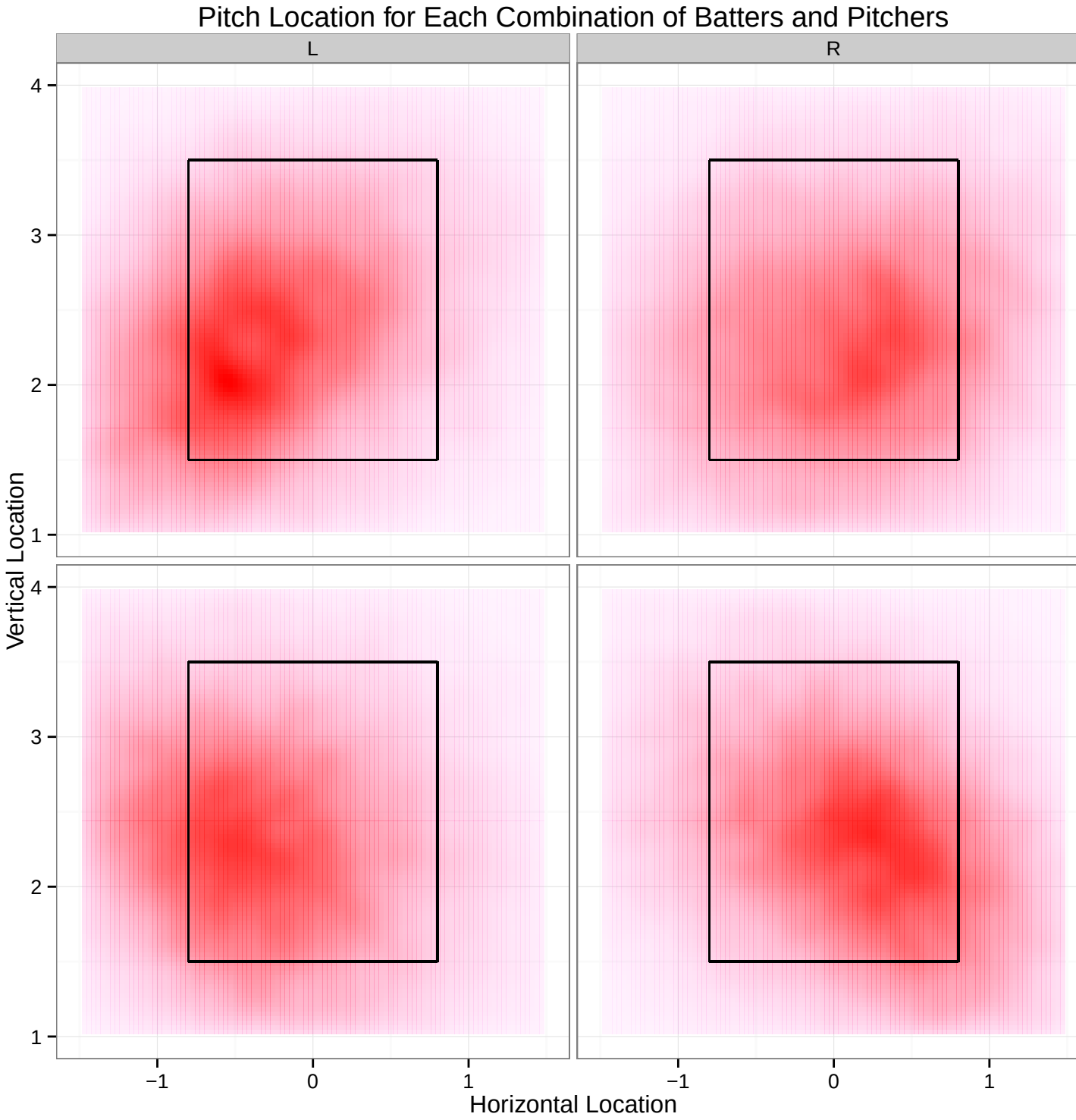
This resulted in 117766 pitches where the pitcher was responsible and 8806 pitches where the pitcher was not responsible. A small portion of the data set is presented below.

Source: local data frame [6 x 8]

```
gameday_link num p_throws stand id resp_pit    px   pz
```

1 gid_2014_03_30_lanmlb_sdnmlb_1 5 L R 38 resp 2.077 2.56 2 gid_2014_03_30_lanmlb_sdnmlb_1 5 L R 39 resp 1.100 2.85 3 gid_2014_03_30_lanmlb_sdnmlb_1 5 L R 41 resp 1.331 2.76 4 gid_2014_03_30_lanmlb_sdnmlb_1 5 L R 44 resp 0.712 2.74 5 gid_2014_03_30_lanmlb_sdnmlb_1 6 L R 51 resp 1.608 2.62 6 gid_2014_03_30_lanmlb_sdnmlb_1 6 L R 52 resp 0.629 3.10

In baseball, there are both left-handed and right-handed batters and left-handed and right-handed pitchers. Pitch locations differ for each of these handedness combinations (see plot below) so the data can be separated into four different pairs of handedness.



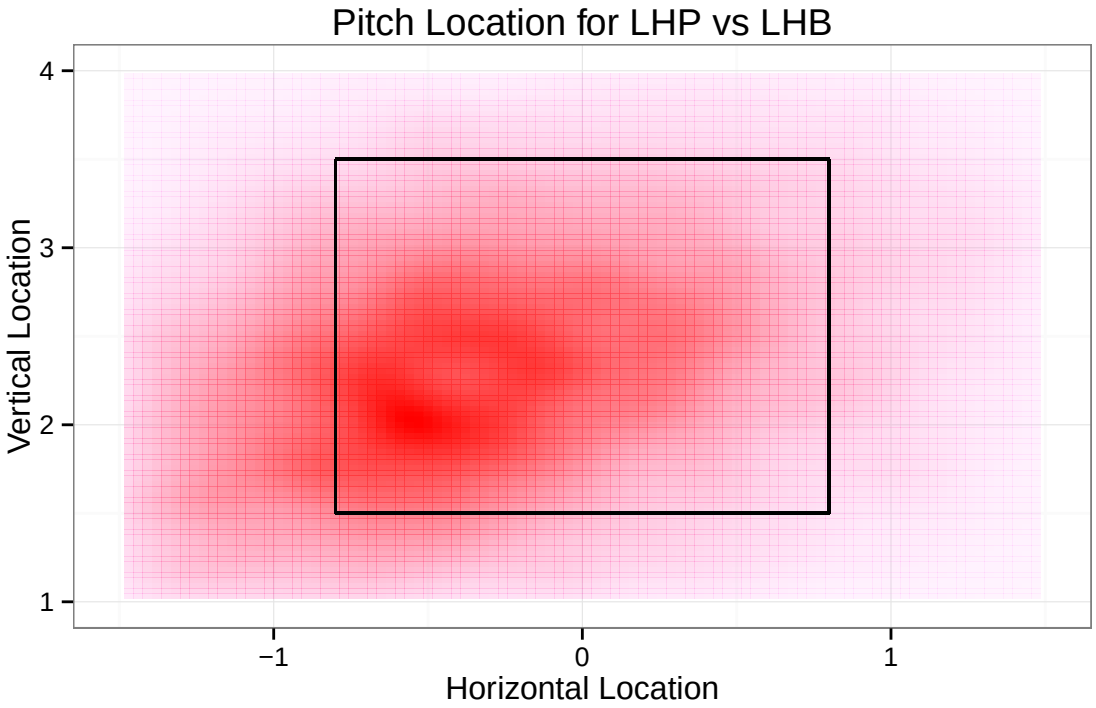
To simplify the analysis for this writing project, only pitches thrown by left-handed pitchers to left-handed batters were considered. This handedness pair was chosen for two reasons. It is the most infrequent combination. It, therefore, presented the smallest and easiest data to analyze. It was also the handedness pair that provide the most suggestive results during early exploration so most of the effort was put into understanding these data. Data were cleaned using the `dplyr` R package (Wickham and Francois 2014).

Methodology

How to Test the Idea

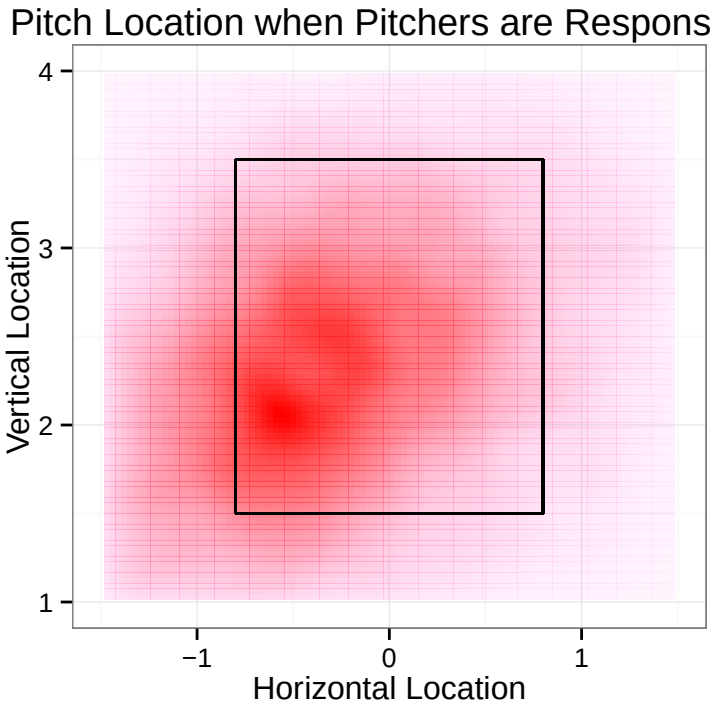
The plot below shows a 2d kernel density estimate of pitch location in 2014 for left handed pitchers vs left handed batters (Venables and Ripley 2002). This estimate is made by constructing a two dimensional $n \times n$ grid of the strike zone where n represents the number of grid points. At each grid point, the pitch density (i.e. the relative likelihood that a pitch is thrown in this location) is calculated for each of these grid points.

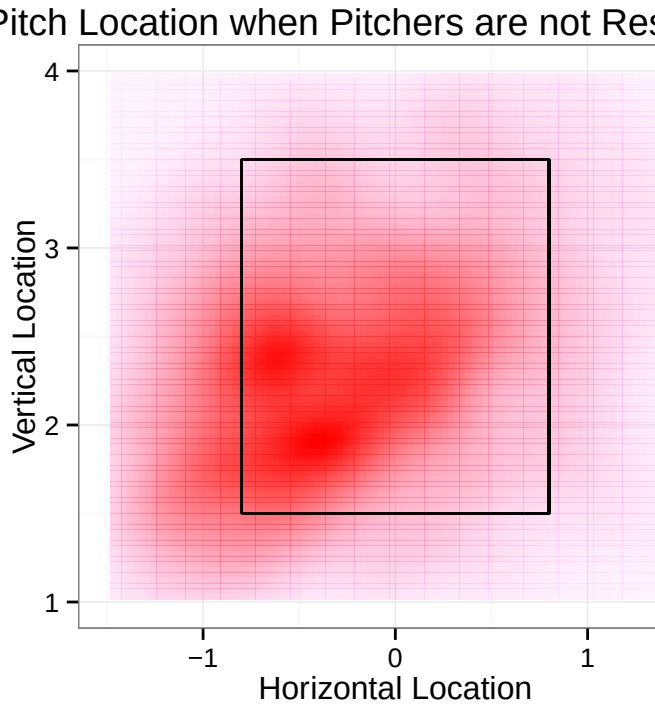
should I explain how densities are calculated?

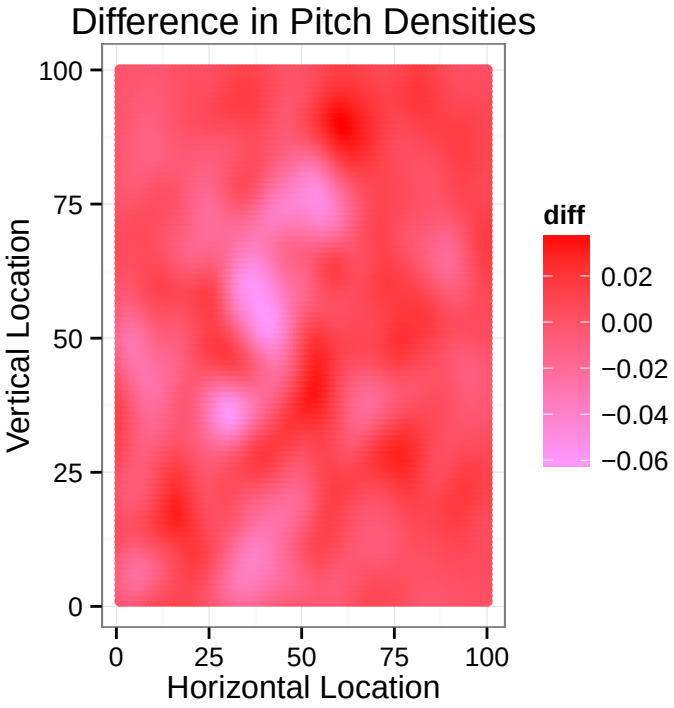


This plot is from the catcher's perspective so the batter stands on the right side of the plot. It appears that most pitches are thrown in the strikezone, down, and away from the batter.

To test the hypothesis that there is no difference in pitch location when pitchers are responsible vs when they are not, we need to see if pitch location is different when pitchers are responsible versus when they are not. An estimated density surface for both of these situations are plotted below along with the difference in the two densities.







In both settings, pitchers do tend to stay low and away from batters. However, it does appear that pitches are more concentrated when pitchers are not responsible. When pitchers are responsible, they spread out the location of their pitches a bit more. On the far right, the magnitude of the difference between the two density estimates is plotted. It appears that there are some differences in the estimates.

Permutation Testing

For this analysis, the null hypothesis is that the density surfaces in the two situations are the same, and the alternative is that they are different. According to that null hypothesis, pitch location is not related to whether or not pitchers are responsible for the men on base. Therefore, under the null hypothesis, the labels are arbitrary and have no meaning.

In this setting, permutation testing can be used to help test these hypotheses. Because the labels are arbitrarily assigned under the null, they can be permuted (i.e. randomly reassigned) to form a permuted dataset which is just as likely have occurred under the null hypothesis. By repeating this process many times, a distribution of permuted differences in densities can be constructed. This distribution represents the likely values of the density surface at each grid point under the null hypothesis. The observed difference in densities can be compared to this distribution of permuted differences in densities at each grid point. In these comparisons, the proportion of times a permuted difference is more extreme than the observed difference becomes the p-value. This creates a grid of $n \times n$ p-values.

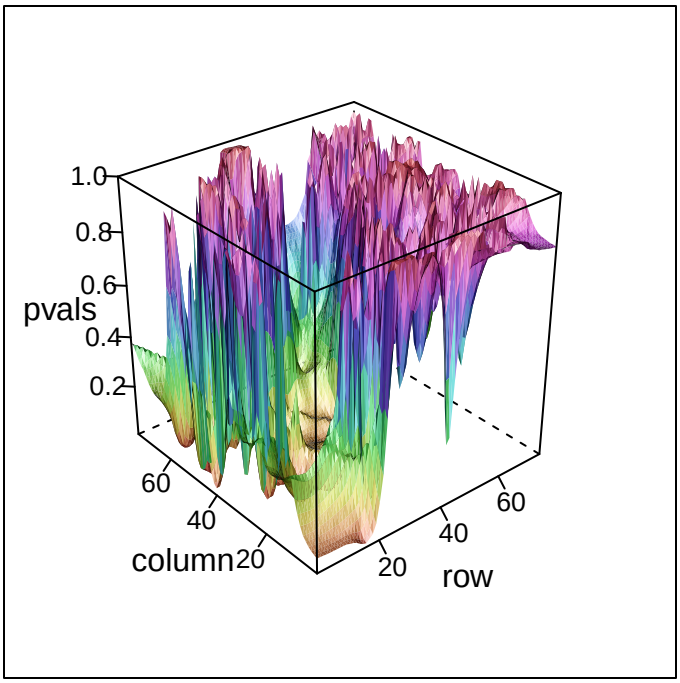
Adjusting for Multiple Testing

Analyzing $n \times n$ p-values creates a multiple testing scenario. Testing each p-value at a given significance level (α), sets the probability of rejecting the null hypothesis when it is true to be $(100 * \alpha)\%$. This controls the Type I Error Probability for any single test. So for each of the $n \times n$ p-values the Type I Error rate is controlled for that test. For all of the tests, however, the *family-wise error rate* (the probability of a Type I error for any of the $n \times n$ tests) is not controlled. A multiple correction procedure was implemented to hold the family-wise error rate constant.

More discussion! about the actual procedure we'll use(?)

Results

After running 1000 permutation on a grid of 100×100 points, a grid of 100×100 p-values was obtained. This grid is visualized below. Blues and purples suggest darker p-values while yellows and reds suggest smaller p-values. Larger rows are on the inner half of the plate, and larger columns are higher in the strikezone.

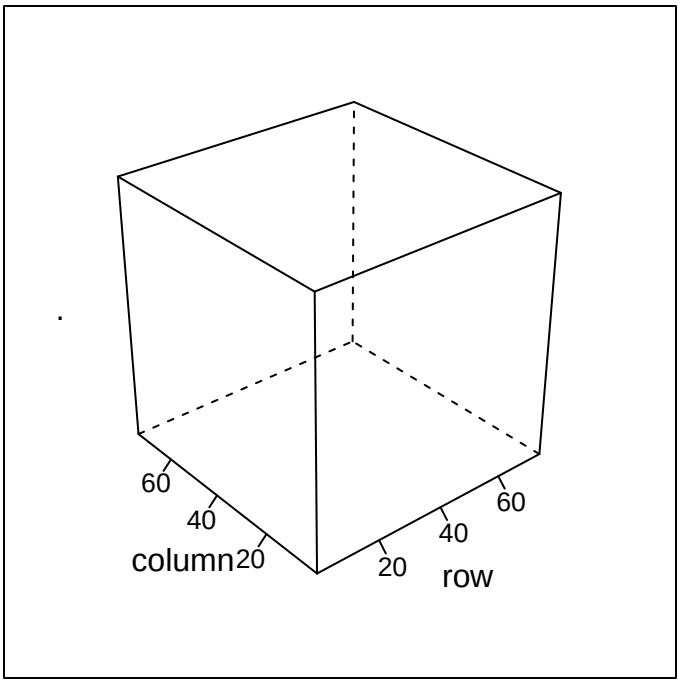


Most of the lower, inner half of the plate contains high p-values. There are a scattered few small p-values, but these are probably spurious. Moving up and away in the strikezone, there is an extremely interesting pattern present. There is a diagonal streak of small p-values which suggests that there might be a difference in the densities along this streak. Interestingly, there is a very steep drop off from the high p-values to the low p-values in this trough. I can't imagine why this might be occurring.

Continuing to move up and away in the strikezone, the p-values once again rise rapidly. Up and away in the strikezone the p-values are mostly very large once again. In the very upper left, there are some small p-values which might bear further investigation but for the most part this area is not very interesting.

Bonferonni Correction

As noted previously, with this many p-values some sort of correction must be used. To begin, a Bonferroni correction was implemented. This correction involves inflating the p-values by $2k$ where k is the number of tests that were run. Since a 100×100 grid was used, there were 10000 tests performed. So all of the p-values must be inflated by 20000. This results in the following adjusted grid of p-values.



Bonferroni is a very conservative correction method so we will be unable to detect any differences using it. In fact, there were not enough tests run to detect a single difference using these data because all the p-values were inflated up to one. More subtle methods of correction will have to be used to determine if there are meaningful differences in these data.

References

Fast, Mike. 2010. "What the Heck Is PITCHf/X." *The Hardball Times Annual*, 153–8.

Keri, Jonah, and Baseball Baseball Prospectus. 2007. *Baseball Between the Numbers: Why Everything You Know About the Game Is Wrong*. Basic Books.

Sievert, Carson, and Maintainer Carson Sievert. 2014. "Taming PITCHf/X Data with pitchRx and XML2R." *The R Journal*. http://journal.r-project.org/archive/accepted/sievert.pdf.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4.

Wickham, Hadley, and Romain Francois. 2014. *dplyr: dplyr: a Grammar of Data Manipulation*. http://CRAN.R-project.org/package=dplyr.