Are the baseball umpires biased?

Wenxin Song
ISBD, Renmin University of China

May 11, 2023

1 Introduction

Since its birth, at least 1858 when the box score had been invented, baseball has been a data-driven sport and one of the largest this kind of sports. Data and associated analytics has been applied to almost every sub-fields of baseball, especially professional baseball. In fact, 'baseball statistics' even has its own name, sabermetrics(originally SABRmetrics, SABR means Society for American Baseball Research).

My focus is on this question: will the baseball umpires give different judgments about whether the pitching is strike or ball according to different situations?

In the baseball game, whether the pitch is strike or ball is quite important, but it is almost up to the umpires's personal judgment and their bias phenomenon have been well observed by many baseball fans and analysts. This work answers the question partly, from a statistical point of view.

2 Methods

2.1 Dataset

In the beginning, baseball data only include the so-called box score, recording every players' performance in the whole game. Of course, this way is beneficial for newspapers' publication, at the cost of information. Gradually, baseball data develop various levels: individual level, team level, season level, and this work focuses on another level of data, pitch level data.

Pitch-level data for every pitch thrown during the 2015-2018 MLB regular seasons. Pitch level data consist of three parts. The first is the trajectory of the ball, including the speed, the spin rate, the coordinates and the types. The second is the situation, including the outs and the occupations of the bases. The third is the judgment, made by the umpires according to his/her own eyes, including many possibilities. In this work, I just consider two possibilities:the called strike and ball, i.e. when the batter did not bat and the umpire only need to judge whether the pitch fall in or out of the strike zone.

For simplicity, we can view the called strike and ball as a reflection for the distance between the pitch and the center.

In short, I investigate how the judgments are affected by other variables. The population are scraped from http://gd2.mlb.com/components/game/mlb/,downloaded in kaggle.com.

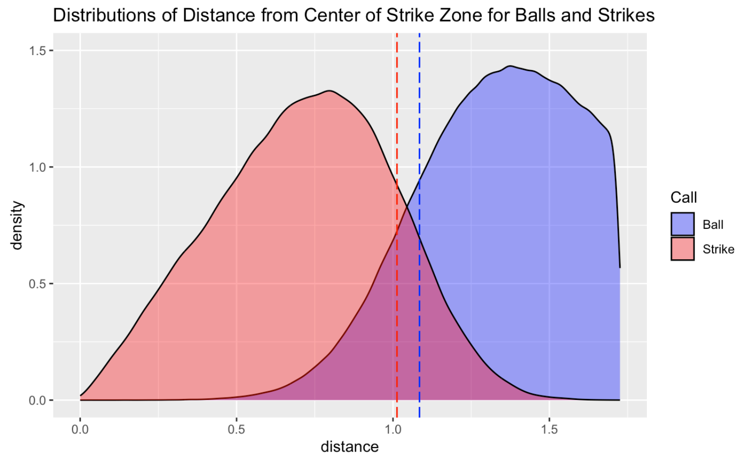
The population size is 2867154.

2.2 Sampling

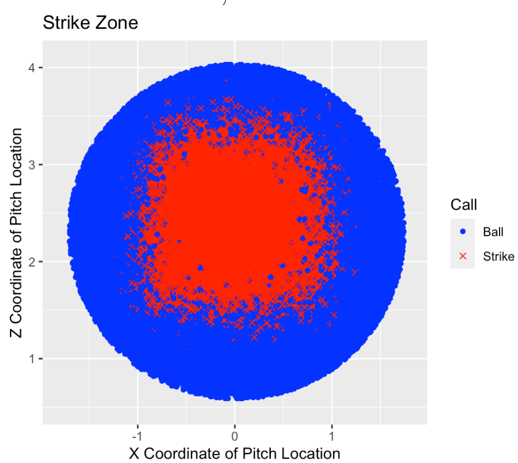
First, I dismiss the too far pitches, which should not affect the umpires' judgment.

Second, I give new variables reflecting the current situation.

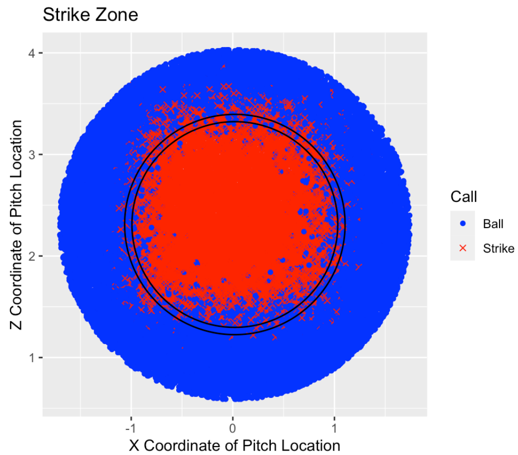
After sampling, the sample size is 907899. Their distance distribution towards the center follows,



From the catcher's view, the distribution of the coordinate follows,

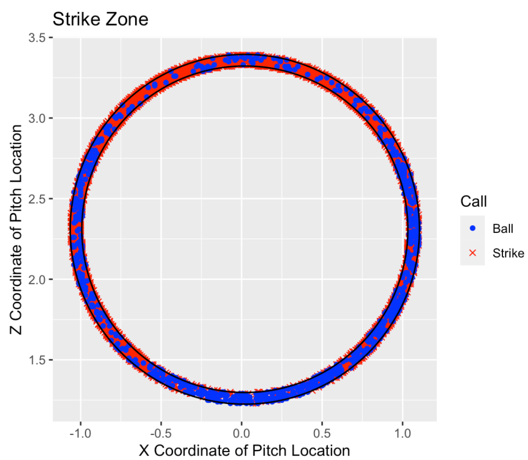


Last but not least, the umpires hardly made mistakes when the baseball was obvious in or out of the strike zone. So I choose the sample inside the ”fuzzy zone”, whose definition is a little arbitrary but quite understandable.



2.3 Decision tree

So I just choose the data inside the annulus. The sample size is 54560.



The radius of the annulus is decided by decision tree method, predicting whether the pitch is strike or ball based on the distance between the pitch and the center.

2.4 Generalized Linear Regression(GLM)

Generalized linear regression is quite common. In this work, the model follows.

$$g(call_i) = \log \frac{\mathrm{E}call_i}{1 - \mathrm{E}call_i} = \beta_0 + \beta_1 * full_count + \beta_2 * Three1 + \beta_3 * Three0 +$$
$$\beta_4 * Two.Outs + \beta_5 * One.Out + \beta_6 * Score.pos + \beta_7 * Bases.loaded + \beta_8 * On.First \tag{1}$$

where g is the link function(in practice we use binomial distribution to get a logistic regression analysis), calls mean whether the umpires give a strike judgment and other variables are specific situations, which will be explained in the following.

In short, these variables can be divided into three parts: counts, outs and occupations.

Counts means the current number of strikes and balls. If a batter gets three strikes, he will be "strike-out", losing his this bat. If a batter gets four balls, he will just walk on the first base, quite worthy. In this work, I only consider situations where the batter gets three balls and zero/one/two strikes.

Outs means the current number of the outs of the batting team. If this team has three outs, it will be the fielding team. In this work, I consider two outs and one out, giving the zero out the base status.

Occupations means the current situation whether the bases are occupied by the batting team, i.e. whether there are base-runners on the bases. Score.pos means whether the second base and the third base are both occupied. Bases.loaded means whether the first, the second and the third bases are all occupied. On.First means whether only the first base is occupied.

3 Results

3.1 Coefficient of the GLM regression

	Estimate	Std. Error	z value	Pr(>|z|)	Percent.Change.Odds	LB95	UB95
Intercept	-1.60	0.01	-288.59	0.00	-79.78%	-80.00%	-79.56%
full.count	-3.51	0.12	-29.64	0.00	-97.00%	-97.62%	-96.22%
Three1	-0.95	0.04	-23.94	0.00	-61.16%	-64.06%	-58.04%
Three0	1.06	0.02	64.04	0.00	187.27%	178.14%	196.70%
Two.Outs	0.24	0.01	28.98	0.00	26.66%	24.65%	28.70%
One.Out	0.13	0.01	15.58	0.00	13.33%	11.56%	15.13%
Score.Pos	-18.09	29.99	-0.60	0.55	-100.00%	-100.00%	47175860
Bases.loaded	-0.01	99.53	-0.00	1.00	-1.06%	-100.00%	51763570
On.First	-18.07	32.04	-0.56	0.57	-100.00%	-100.00%	26348037

The column "Percent.Change.Odds" is equal the E*call*.

3.2 Explanation

3.2.1 Counts

Let us concentrate on the estimate of the variables "full.count", "Three1" and "Two.Outs".

"full.count" decreased the call-strike odds by 97% when compared by the non three ball situation, "Three1" decreased it by 61% while "Three0" increased it by 187.27%.

This has a baseball explanation. When there has been three balls and no strike, the umpires tend to give no chance to the batter, because he may view the batter as weak.

3.2.2 Outs

Let us concentrate on the estimate of the variables ”Two.Outs” and ”One.Out”.

Data shows that the umpire tend to give called strike more possibly when there is one out or two outs and the possibility is a little bigger when the outs is two.

3.2.3 Occupation

Let us concentrate on the standard error of the variables ”Score.Pos”, ”Bases.loaded” and ”On.First”.

They are so big that the p-value is quite big, maybe reflecting the sample size is small, so the estimate is not convincing.

4 Conclusion and discussion

From the analysis of the result we can see that there are bias of the umpires' judgment. So I will accordingly give some advice towards the batter, especially the pitch, which is my position during my Beijing Baseball Tournament.

First, we batters should remember the counts and outs, not only for better deciding whether batting or not, but also for reacting more according to the umpires' bias. For example, when there are three balls and no strike, we can not give up the pitch in the fuzzy zone because the umpire has had his ”prior” knowledge. Or when there have been outs, we will face more possibility of striking.

Second, as the pitcher, we can adjust our pitch strategy according to the count and outs, giving more challenging pitches and less ER(Earned Run).

Moreover, this work does not discuss how the MLB data can be adopted by our Beijing Baseball Tournament, nor discuss whether different umpires have different biases. I leave these to future.

SUPPLEMENTARY MATERIAL

Code: Rmd file.

Baseball data set: Since it is so big, I only list the website below. https://www.kaggle.com/datasets/psc
pitch-data-20152018

5 Reference

None