# Are MLB Baseballs Juiced?

Matt Kalin and Thomas W. Ilvento

07/13/2017

## Introduction

The home run is arguably the most popular and decisive occurrence in the game of baseball. It always scores at least one run, and it is simply entertaining to see a ball hit over the fence. Home run rates in Major League Baseball peaked at the height of the steroid era in 2000, and had declined steadily over the following decade and a half. However, the past two years have seen the MLB home run rate shoot back up and even eclipse its steroid era maximum. An article (https://fivethirtyeight.com/features/are-juiced-balls-the-new-steroids/) written by Ben Lindbergh and Rob Arthur from FiveThirtyEight hypothesized that MLB's baseballs had undergone a change around the All-Star break in July 2015, and that the newer balls are bouncier and thus are hit farther than the older balls.

## Acquire Data

ESPN provides team batting data (http://www.espn.com/mlb/stats/team/_/stat/batting/split/40) that can be filtered for every month of every season since 2000. It shows totals for each team as well as the NL, AL, and MLB averages. We looked at every month since April 2013 and entered the MLB average at-bats, home runs, and fly balls into an excel spreadsheet.

From this data, we want to determine the following:

- Monthly home run per at-bat percentage
- Monthly home run per fly ball percentage
- Monthly fly ball per at-bat percentage

One issue with the data given is that it is an average for all teams, and not the total for the MLB. This does not matter when determining the home run rates, but it will become an issue when determining if the difference in rates is statistically significant. We multiplied by 30 (the number of MLB teams) to turn the averages into sums.

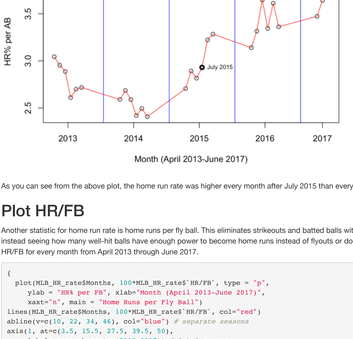
```
library(readxl)
MLB_HR_rate <- read_excel("~/Desktop/Summer 2017/MLB HR rate.xlsx")
MLB_HR_rate$Num=1:27 # the number in the time series, to make it easier for plotting
head(MLB_HR_rate)
```

```
## # A tibble: 6 x 9
##    Year Month        AB    HR    FB `HR/AB` Months `HR/FB`    Num
##   <dbl> <chr>     <dbl> <dbl> <dbl>   <dbl>  <dbl>   <dbl> <int>
## 1  2013 April       887    27   376  0.0304      1  0.0718     1
## 2  2013 May         948    28   403  0.0295      2  0.0695     2
## 3  2013 June        935    27   399  0.0289      3  0.0677     3
## 4  2013 July        881    23   371  0.0261      4  0.0620     4
## 5  2013 August      963    26   413  0.0270      5  0.0630     5
## 6  2013 September   920    25   390  0.0272      6  0.0641     6
```

## Plot HR/AB

A good statistic for capturing home run rate is HR/AB, which is the percentage of at-bats that result in home runs. Here I plot the MLB average HR/AB for every month from April 2013 through June 2017.

```
{
  plot(MLB_HR_rate$Months, 100*MLB_HR_rate$`HR/AB`, type = "p",
  ylab = "HR% per AB", xlab="Month (April 2013-June 2017)",
  xaxt="n", main = "Home Runs per At-Bat")
lines(MLB_HR_rate$Months,100*MLB_HR_rate$`HR/AB`, col="red")
abline(v=c(10, 22, 34, 46), col="blue") # separate seasons
axis(1, at=c(3.5, 15.5, 27.5, 39.5, 50),
     labels = as.character(2013:2017)) # label seasons
points(28, 100*MLB_HR_rate[16, "HR/AB"], lwd=3)
text(28, 100*MLB_HR_rate[16, "HR/AB"],
     labels = "July 2015", pos = 4, cex = 0.7) # label July 2015 on the plot
}
```

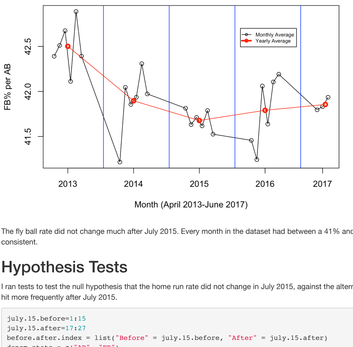



As you can see from the above plot, the home run rate was higher every month after July 2015 than every month before July 2015 in the dataset.

## Plot HR/FB

Another statistic for home run rate is home runs per fly ball. This eliminates strikeouts and batted balls with no shot of going over the fence, instead seeing how many well-hit balls have enough power to become home runs instead of flyouts or doubles. Here I plot the MLB average HR/FB for every month from April 2013 through June 2017.

```
{
  plot(MLB_HR_rate$Months, 100*MLB_HR_rate$`HR/FB`, type = "p",
     ylab = "HR% per FB", xlab="Month (April 2013-June 2017)",
     xaxt="n", main = "Home Runs per Fly Ball")
lines(MLB_HR_rate$Months, 100*MLB_HR_rate$`HR/FB`, col="red")
abline(v=c(10, 22, 34, 46), col="blue") # separate seasons
axis(1, at=c(3.5, 15.5, 27.5, 39.5, 50),
     labels = as.character(2013:2017)) # label seasons
points(28, 100*MLB_HR_rate[16, "HR/FB"], lwd=3)
text(28, 100*MLB_HR_rate[16, "HR/FB"],
     labels = "July 2015", pos = 4, cex = 0.7)
year.hrfb.avg=NA
for(i in 1:5){
  year=i+2012
  year.hrfb.avg[i]=sum(MLB_HR_rate[which(
    MLB_HR_rate$Year==year), "HR"])/sum(
      MLB_HR_rate[which(MLB_HR_rate$Year==year), "FB"])
}
}
```




Just like HR/AB, the monthly home run per fly ball rates are higher after July 2015 compared to before.

## Plot FB/AB

The two plots above look very similar. I then wondered if the rate at which fly balls were hit changed at all during the time period in this dataset.

```
{
  fb.ab=MLB_HR_rate$FB/MLB_HR_rate$AB
  plot(MLB_HR_rate$Months, 100*fb.ab, type = "p",
       ylab = "FB% per AB", xlab="Month (April 2013-June 2017)",
       xaxt="n", main = "Fly Balls per At Bat")
  lines(MLB_HR_rate$Months, 100*fb.ab, col="black")
  abline(v=c(10, 22, 34, 46), col="blue") # separate seasons
  axis(1, at=c(3.5, 15.5, 27.5, 39.5, 50),
       labels = as.character(2013:2017)) # label seasons
  year.fbab.avg=NA
  for(i in 1:5){
    year=i+2012
    year.fbab.avg[i]=sum(MLB_HR_rate[which(
      MLB_HR_rate$Year==year), "FB"])/sum(
        MLB_HR_rate[which(MLB_HR_rate$Year==year), "AB"])

  } # yearly averages
  points(c(3.5, 15.5, 27.5, 39.5, 50.5), 100*year.fbab.avg, col="red", lwd=3)
  lines(c(3.5, 15.5, 27.5, 39.5, 50.5), 100*year.fbab.avg, col="red")

  legend(35, 42.7, c("Monthly Average", "Yearly Average"), lty=NULL,
         pch=1, cex = 0.6, col = c("black", "red"), lwd = c(1, 3))
}
```




The fly ball rate did not change much after July 2015. Every month in the dataset had between a 41% and 43% fly ball rate, which is fairly consistent.

## Hypothesis Tests

I ran tests to test the null hypothesis that the home run rate did not change in July 2015, against the alternative hypothesis that home runs were hit more frequently after July 2015.

```
july.15.before=1:115
july.15.after=117:27
before.after.index = list("Before" = july.15.before, "After" = july.15.after)
denom.stats = c("AB", "FB")
for (d in denom.stats) {
  hr.vals = numeric()
  denom.vals = numeric()
  for (i in 1:2) {
    # calculate stats for before and after July 2015
    hr.vals[i] = sum(MLB_HR_rate[before.after.index[[i]], "HR"])
    denom.vals[i] = sum(MLB_HR_rate[before.after.index[[i]], d])
  }
  print(paste0("Testing difference in proportion: HR/", d))
  print(prop.test(hr.vals, denom.vals))
}
```

```
## [1] "Testing difference in proportion: HR/AB"
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  hr.vals out of denom.vals
## X-squared = 11.095, df = 1, p-value = 0.0008654
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.012163113 -0.003006944
## sample estimates:
##      prop 1      prop 2
## 0.02703495 0.03461998
##
## [1] "Testing difference in proportion: HR/FB"
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  hr.vals out of denom.vals
## X-squared = 12.341, df = 1, p-value = 0.0004431
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.029327984 -0.007971477
## sample estimates:
##      prop 1      prop 2
## 0.06420741 0.08285714
```

Both of these tests show a statistically significant change in both the HR/AB and HR/FB rates in July 2015. Were the balls juiced during the all-star break? Maybe. But correlation does not mean causation. Home runs are happening more frequently since then, but we don't know exactly why. Perhaps we will never know the reason.