Home Runs by Position

also, Batting Average Analysis Versus Time

Eric Carlsen

14 January 2016

1 Introduction

In this project I analyzed the major league baseball (MLB) data to investigate two questions. The first question I wanted to answer is if there is a statistically significant difference in home run hitting ability by position. The second question was to investigate a hypothesis as to why there are no longer any players batting over .400.

Regarding the first question, I chose to compare two positions that, at first blush, require different levels of athleticism. The shortstop position requires speed and agility and fields a significant amount of infield hits. The first baseman, on the other hand, appears to not move over as large of an area and thus may not require as much speed. In theory this would allow a slower, but perhaps more powerful, player at first base. Conversely, a large frame and musculature might slow down shortstops. In theory, that might indicate less ability to hit home runs. I decided to test this theory by quantifying home runs for each position and comparing the two.

I also wanted to perform an analysis of batting averages over time as a step in answering another question: why do we no longer see .400 batters? It might be difficult to determine the causes of that phenomena, but I intend to specifically observe if it's a decrease in the mean averages, or a decrease of batting average standard deviations (or both) that is causing this effect.

2 Data

2.1 Data Sets

The data sets required for this analysis were the Batting.csv file, the Appearances.csv, and the Teams.csv file from Lahman's Baseball Database.

2.2 Data Wrangling

Some work was required to get the data into the proper format for analysis. In analyzing the home runs by position data I used the entire data set.

I had to determine the positions for each player identified in the appearances DataFrame (which came from importing the Appearances.csv file). A position itself isn't given, just the number of times a player appeared at each position. I created a new POS column that was set to 'Shortstop' if the number of occurrences at that position was greater than the number of positions than at first base. If there were more appearances at first base then I set POS to 'First Basemen'. For all other ratios (include equal, and only zeros in both fields) the position was set to 'Other'. I then merged this appearances DataFrame with the batting DataFrame.

For the batting average data I added a column to the pandas DataFrame object for the batting statistics that corresponded to the batting average. This was given by the equation

$$AV = \frac{H}{AB}, \tag{1}$$

where AV is the batting average, H is the number of hits, and AB is the number of at bats.

For the batting average data I restricted the histogram and statistics to only take into account results for players that had at least 400 attempts at bat. This reduced the outliers of players with few attempts at bat, who often had either .000 average or an unsustainably high average. The number was chosen based on the minimum appearances required to lead a rate stat given in [1]. There are a complicated set of rules there, that have been applied at various periods of time. One of the rules to lead a rate stat at one point, however, was that a player needed to appear at bat at least 400 times in a year. This rule is a good first order approximation to the rules given in [1] and should be sufficient to determine general trends in the data.

3 Exploration

3.1 Home Runs by Position

Once I had the position information merged into the batting information I was ready to test my hypothesis. The null hypothesis is that shortstops can hit the same or more home runs than first basemen. In other words,

$$H_0 : \mu_D \leq 0, \tag{2}$$

where $\mu_D = \mu_{firstbase} - \mu_{shortstop}$.

The alternative hypothesis is that first basemen can hit more home runs than shortstops:

$$H_1 : \mu_D > 0 \tag{3}$$

Since there are different numbers of players in each category this is an independent samples t-test. Given that I wish to test if first basemen have higher home run hitting averages, and I don't care about lower values, I employed a one-sided t-test. The statistical measures from each sample are shown in Table 1. The histograms of home runs by each position is shown in Figure 1.

Given the above counts the degrees of freedom is computed as

$$DOF = (n_1 - 1) + (n_2 - 1) = (23552 - 1) + (16536 - 1) = 40086. \tag{4}$$

The difference between the means of the first basemen and shortstop players is

$$\Delta_{mean} = 3.16, \tag{5}$$

	First Basemen	Shortstop
count	23,552	16,536
mean	6.22	3.07
std	8.94	5.41
min	0	0
max	70	57

Table 1: Home Runs by Position from 2005-2014

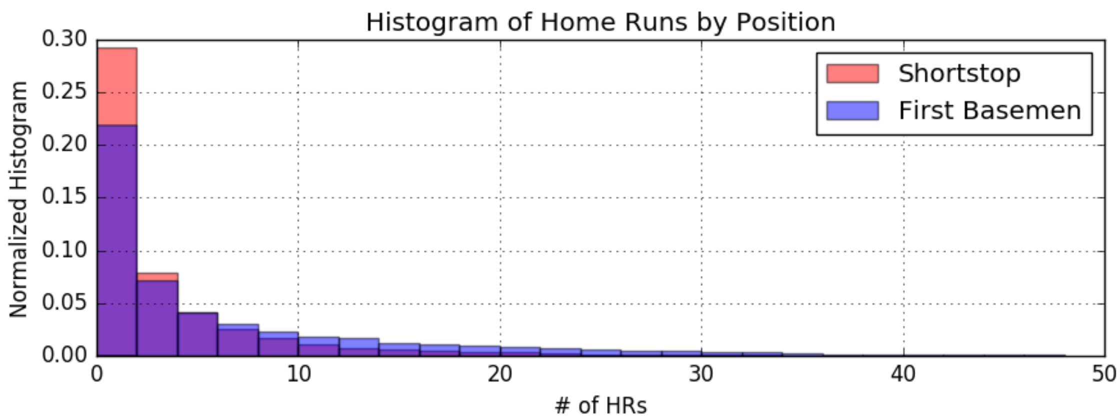

Figure 1: Histogram of Home Runs by Position

home runs. The standard error of the mean is given by

$$SEM = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{8.94^2}{23552} + \frac{5.41^2}{16536}} = 0.0719. \tag{6}$$

The t-statistic would would then be

$$t = \frac{\Delta_{mean}}{SEM} = \frac{3.16}{0.0719} = 43.9. \tag{7}$$

The critical value of the student's t distribution can be computed from tables for the given degrees of freedom, or using the built-in scipy.stats student's t functions. Using the scipy functions returns a critical value of 1.64 for 40,086 degrees of freedom with $\alpha = 0.05$ and a one-sided t-test. Given that the t-statistic of 43.9 is much greater than the critical value that means **we can reject the null hypothesis**. The p-value for the given t-statistic is exceedingly low, effectively zero for any significant amount of digits.

To compute Cohen's D we first need to calculate the standard units

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = 7.68. \tag{8}$$

Then we can compute Cohen's D

$$d = \frac{\Delta_{mean}}{s} = \frac{3.16}{7.68} = 0.411. \tag{9}$$

3.2 Batting Average over Time

We must first establish the maximum batting average over time. Figure 2 shows these values. Remember that this is restricting the players to at least 400 at bat attempts. As the plot shows, the last player to bat over .400 was Ted Williams in 1941 [2].

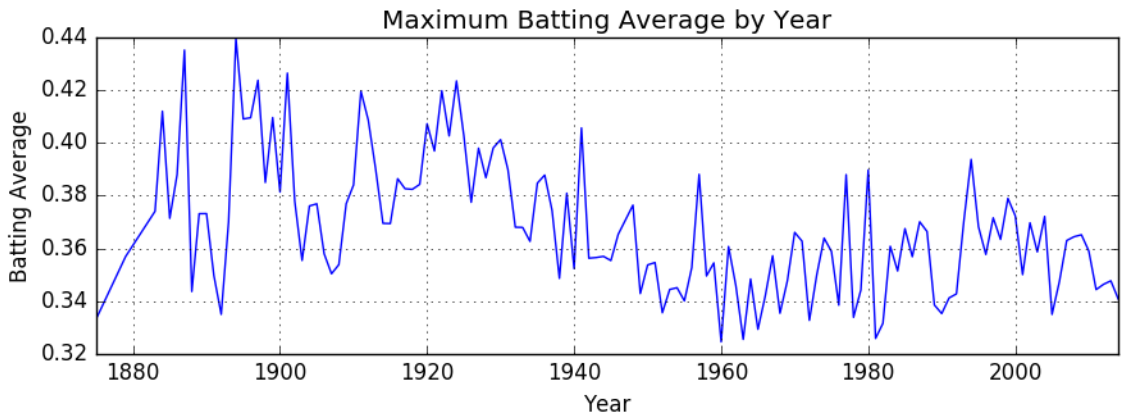

Figure 2: Maximum Batting Averages per Year

Next I computed the standard deviation and mean of the batting average of the players per year. This is shown in Figure 3. The first subplot shows the standard deviation values alone. The bottom subplot shows the mean values, with error bars indicating the standard deviations. Although the plot of the means includes error bars for the standard deviation, it is difficult to get a sense for the trend of the size of the standard deviation from these error bars. That's why I included the standard deviation values in its own subplot as well. The mean values showed some variation, but not an overtly obvious downward trend. The standard deviations, however, mostly exhibit a downward trend over time. Looking at plots in this way, however, is somewhat of a qualitative analysis and not strongly conclusive. However, it would seem that the decrease in batters with averages above .400 is mostly driven by a tighter clustering of batting averages, not a decrease in mean values.

4 Conclusions

In this report I investigated two questions. The first question was if there was a statistically significant difference in the home runs hit by shortstops and first basemen. I applied a one-tailed, two sample, independent t-test to determine that there was a statistically significant difference, below the $\alpha = 0.05$ level.

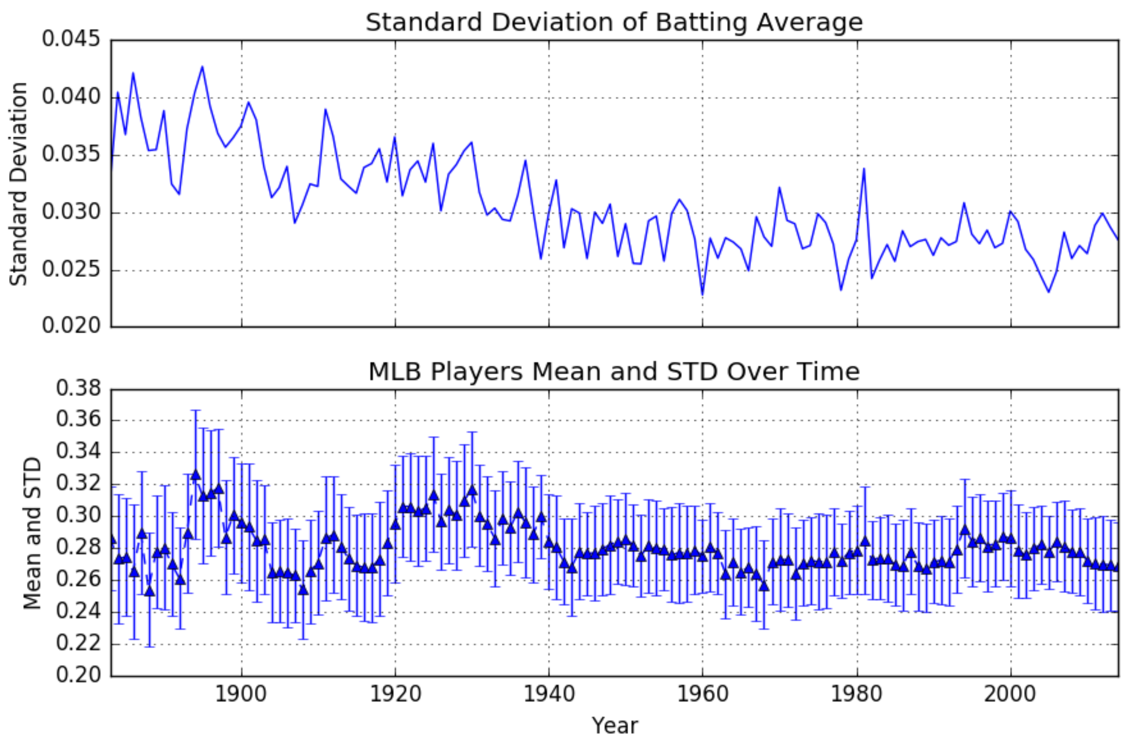

Figure 3: Measure of batting average standard deviation over time.

The second question was an attempt to address why there are no longer hitters batting .400. Observing the plots of the mean batting averages and batting average standard deviations points to a decrease in standard deviations as being more responsible than a change in mean values for the decrease.

References

[1] baseball reference.com. Leaderboard glossary. `http://www.baseball-reference.com/about/leader_glossary.shtml#min_req`, January 2016.

[2] Wikipedia. List of major league baseball players with a .400 batting average in a season. `https://en.wikipedia.org/wiki/List_of_Major_League_Baseball_players_with_a_.400_batting_average_in_a_season`, August 2015.