Home Runs by Position

also, Batting Average Analysis Versus Time

Eric Carlsen

20 December 2015

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? Many years ago I was acquainted with the theory that it's due to the decreased gap between the best and worst players in the league. This stems from the number of players in the league relative to the overall population. As the number of players becomes a smaller percentage of the population that means there is less variation of abilities in both pitching and hitting.

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. I also downloaded Census data to determine the population of the United States over time [1].

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 first chose to use a sample of the data, restricting it to the results from the

last ten years.

I then had to determine the positions for each player identifier in the appearances DataFrame. A position itself isn't given, just the number of times they 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 (from the last 10 years).

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. Note that this was the full data set, not just the batting data from the previous 10 years.

For the batting average data I restricted the histogram and statistics to only take into account results for players that had at least 25 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.

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. I'll focus on the difference between the shortstop and first baseman positions and ignore the others.

Given the above counts the degrees of freedom is computed as

$$DOF = (n_1 - 1) + (n_2 - 1) = (2519 - 1) + (1738 - 1) = 4255. \tag{4}$$

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

$$\Delta_{mean} = 3.92, \tag{5}$$

	First Basemen	Shortstop	Other
count	2519	1738	9601
mean	9.11	5.20	1.69
std	10.2	7.30	5.01
min	0	0	0
max	58	54	45

Table 1: Home Runs by Position from 2005-2014

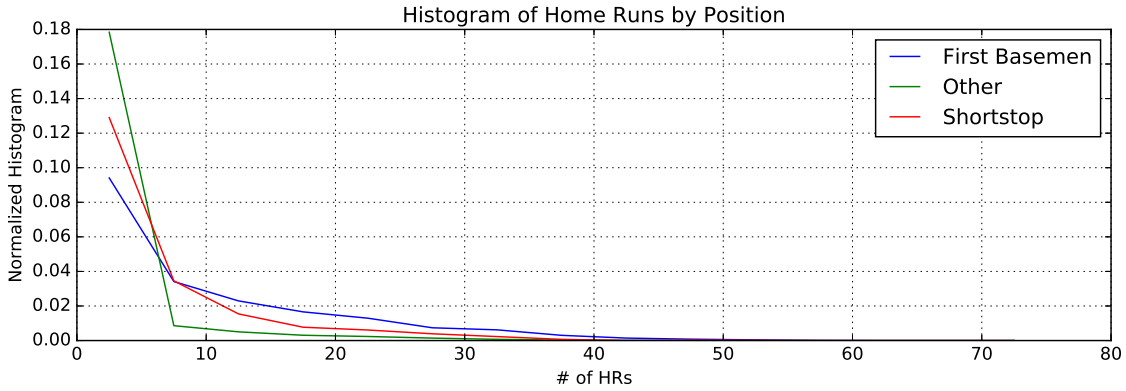

Figure 1: Histogram of Home Runs by Position

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

$$SEM = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{10.19^2}{2519} + \frac{7.296^2}{1738}} = 0.268. \tag{6}$$

The t-statistic would would then be

$$t = \frac{\Delta_{mean}}{SEM} = \frac{3.92}{0.268} = 14.63. \tag{7}$$

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 1.65 for 4255 degrees of freedom with $\alpha = 0.05$ and a one-sided t-test. Given that the t-statistic of 14.63 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, approximately 1.68×10^{-47}.

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

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = 9.12. \tag{8}$$

Then we can compute Cohen's D

$$d = \frac{\Delta_{mean}}{s} = \frac{3.92}{9.12} = 0.429. \tag{9}$$

3.2 Batting Average over Time

The theory behind the question of why we don't see .400 batters anymore has to do with the variance of the professional baseball players over time. As the number of players in the league has become a smaller percentage of the overall population that means the abilities are in the tails of the distributions. For example, see Figure 2. The x-axis of this distribution is some unquantified baseball playing skill level. In the past, when the proportion of the population that was MLB players was bigger, that would correspond to the portion of the tail of the curve in blue. Now that the proportion is smaller, that corresponds to the portion in red. If we assume that, due to both the limited number of players and the theoretical limits of human performance, there is some maximum skill level that's more or less the same, and isn't infinity, then we can see that the delta between that maximum on the x-axis and the x-axis value where the red and the blue starts will create a wider range of skill levels for the blue section. This will hold true for both batters and pitchers. The hypothesis is that this closer grouping of skill levels means that it's harder to have outliers, and performance will have a lower standard deviation.

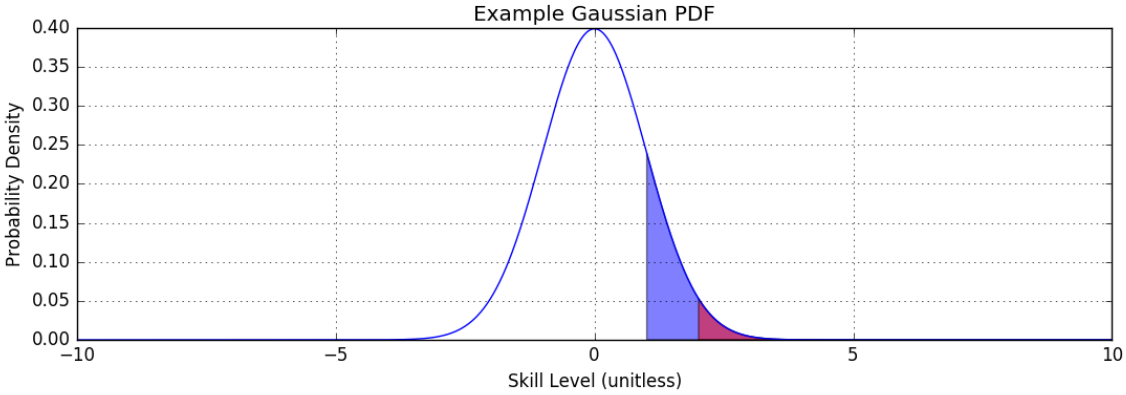

Figure 2: Example Gaussian Distribution

The way I've decided to measure this is to observe the standard deviation of the batting average for quantiles of the years over the span of the existence of the league. But first, it's important to determine if it's even a valid assumption that the number of MLB players (as a percentage of the population) has been shrinking. Figure 3 shows the data for the number of MLB players, the growth of the US population, and the ratio between the two. It's worth noting that the number of MLB players is an estimate from the number of teams in the league. I assumed that there were 25 players per team, as per current roster limits. The ratio subplot shows a downward trend in the overall percentage of the total population who are MLB players. Although I was surprised to find the general increase in teams in recent years seems to attempt to at least grow in some proportion to the overall population. Note, however, that the first subplot in Figure 3 has two different y-axes, with different scales. So, although it may appear that they directly track each other, it's really just a function of each plot scaling their y-axis to fill the available space. However, after 1920 the number of players monotonically increases, as does the US population.

Next I split the span of years of the data (from 1871 to 2014) into quantiles. Figure 4 shows a

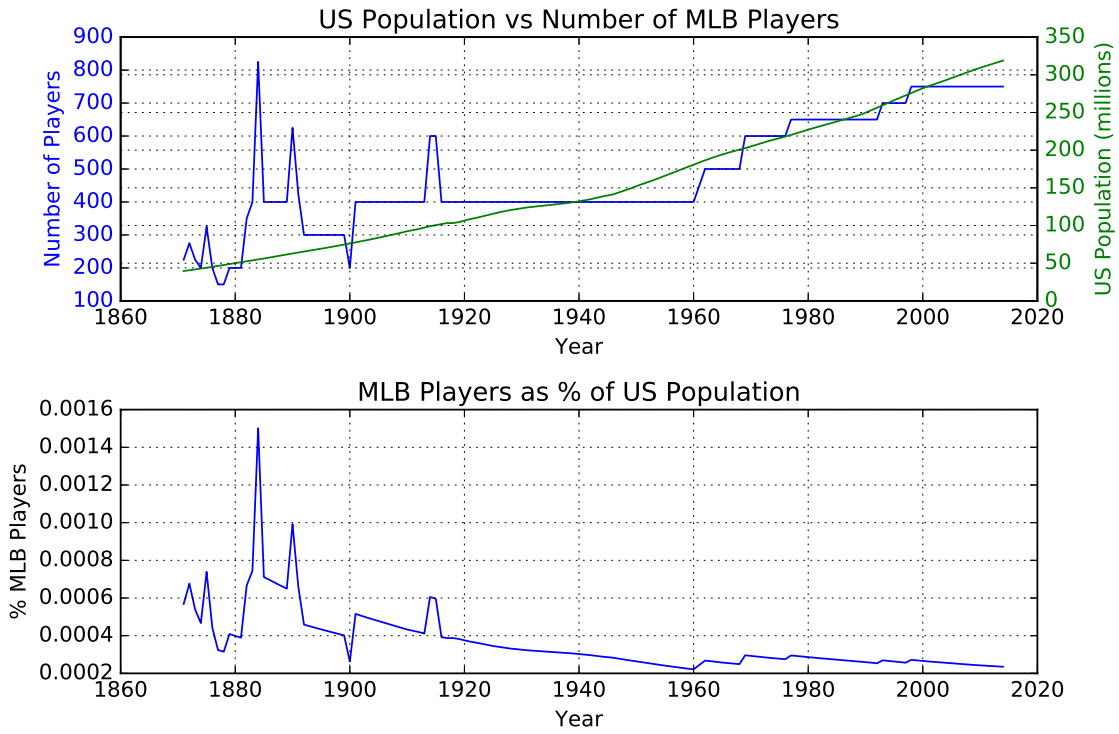

Figure 3: Plots indicating the percentage of the total US population that are MLB players.

histogram of the batting average data, with one plot per quantile. In this case, for visualization, the data was split into only three quantiles. However, I also measured the standard deviation and mean of each quantile, as shown in Figure 5. For this plot twelve quantiles were used, to provide more granularity on the results. The mean values showed some variation, but not an overtly obvious overall trend. The standard deviations appeared to mostly follow a downward trend, but the it wasn't strongly linear. Perhaps some type of regression analysis would be useful to apply to the standard deviation over time.

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.

The second question was an attempt to address why there are no longer hitters batting .400. The hypothesis was that the abilities of MLB players (both batters and pitchers) are now more tightly clustered, leaving less likelihood for outliers. This relies on the underlying assumption that there are fewer MLB players now than in times past (as a percentage of the total population). While

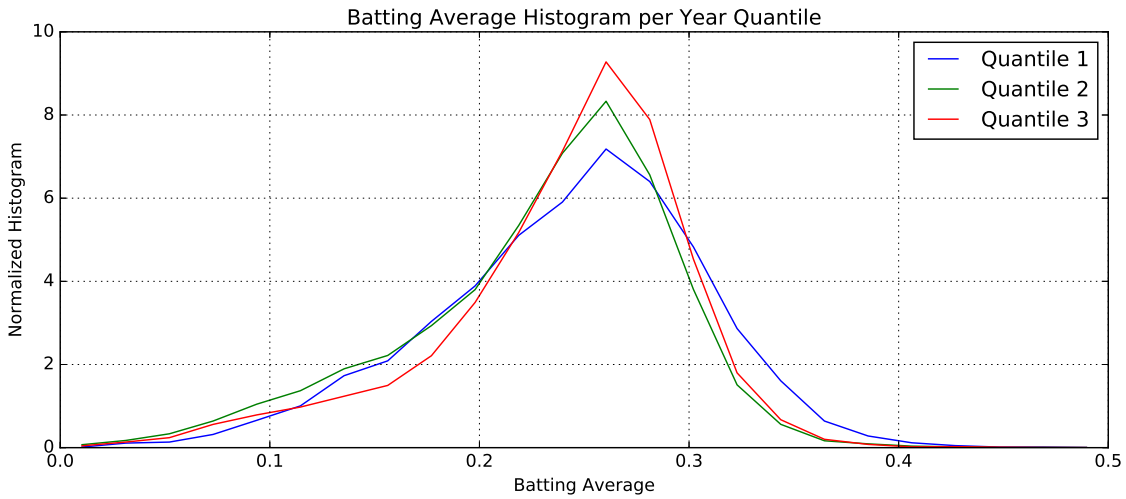

Figure 4: Histogram of batting averages separated into year quantiles.

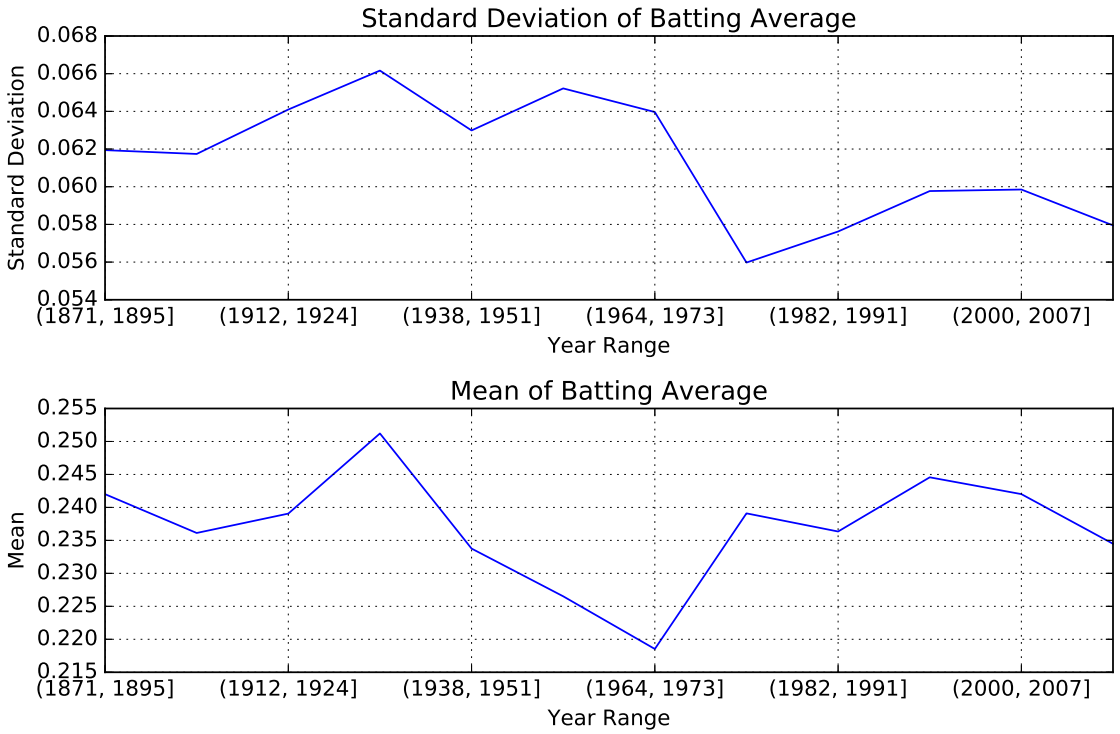

Figure 5: Measure of batting average standard deviation over time.

there might be some data supporting the conclusion, the results of the standard deviation spread of batters over time does not strongly support this conclusion. Therefore, the results of this hypothesis are inconclusive with the analysis provided so far, and further confirming or dis-confirming evidence should be sought after.

References

[1] U.S. Census Bureau. 2010 census. U.S. Department of Commerce, February 2015.