The Relative Age Effect in Major League Baseball:

Is it Still Present Today?

Due: December 11, 2016

Vishal Ahir, Earl Shaw, Jessica Wheeler

**Abstract:** The *History of Baseball* database [1] was originally created by Sean Lahman in 1994 in an effort to make baseball statistics freely available to the general public. The database is available on [www.kaggle.com](http://www.kaggle.com) with full detail availabe at: http://seanlahman.com. The database encompasses a variety of data on Major League Baseball (MLB) players from 1871-2015 such as debut date, birth date, birthplace, and more. Additionally, pitching, batting, and fielding statistics are provided along with player salary. For the purposes of this project, we narrowed down the data to the past 3 years (2013-2015) and used MySQL Workbench to create a database schema.

Our research focused on batting and pitching averages, as well as salary. We also wanted to establish whether the relative age effect, commonly known as RAE, was present. Various studies on RAE claim that in a league sport, those that are the eldest due to age cut-offs tend to perform better and are more likely to play professionally and earn higher wages [2.] We wanted to see if RAE was present in the players who had the highest batting or pitching averages and salaries.

**1. Introduction**

**1.1 The Dataset**

The *History of Baseball* database brings together individual datasets that include information on players, teams, pitching, batting, fielding, hall of fame, all-star, awards, salary, team managers, and even ballparks. We decided to narrow our focus and use tables that contained salary, player, batting, and pitching information from 2013-2015. Our goals were to elicit information about the best batting and pitching statistics and see how those results correlate to a player’s salary and if there was any relative age effect present.

**1.2 Related Work**

In the 140-year history of baseball, there have been numerous statistical analyses that provide insight into each baseball game. Since 2007, the Pitchf/x and Hitf/x systems implemented in the MLB have been able to capture more information on hitting and pitching aspects than ever before [3]. Studies utilizing this dataset have explored whether closers are over-paid [1], and which factors affect the attendance of baseball games, that is, whether it’s home runs or wins, losses or errors [1]. Others have looked more into the historical aspects of the data such as who were the world series winners since 1900 [1].

Considering past articles and studies on the relative age effect were primarily historical in their approach, with results based on data as recent as 2005, our analysis represents an updated and relevant view into whether or not the effect remains, due to our focus on the prior three years of data, i.e. 2013-2015. These studies suggested that more successful baseball players are all born in August. Whether or not this occurs recently in Major League Baseball is our aim.

**2. Database Schema and Organization**

Our analysis focused on the following files from the database:

* *Player.csv* (master table; primary key: *player\_id*)
* *Salary.csv* (supplementary-salary stats; )
* *Batting.csv* (supplementary-batting stats)
* *Pitching.csv* (supplementary-pitching stats)

We used *MySQL Workbench* to load and store the files and to conduct table joins and queries for our analyses and conclusions. We filtered only a subset of the last 3 years of data for the files *Batting.csv* and *Pitching.csv,* which contained the “year” variable needed for limiting our data.

In examining the batting statistics of players over the 2013-2015 timeframe, the files of interest were: *player.csv*, *batting.csv*, and *salary.csv*. The *player.csv* file represented the primary table with 18,846 observations on all players from 1871-2015. This table contained the primary key, *player\_id*, useful for joining to other files contained within the database. The *batting.csv* file initially contained 101,332 observations on a variety of batting statistics such as “at bats”, “hits”, “home runs”, etc. for MLB batters from 1871-2015. As mentioned prior, only years 2013-2015 were retained for this analysis, which effectively removed ~96% of the original battingdata set, leaving 4,330 observations for analysis. Additionally, the *batting.csv* data set consisted of numerous duplicate records due to players playing across multiple years and possibly multiple teams. It contained the foreign key, *player\_id*, which was used to inner join to the *player.csv* file, thereby keeping only those records pertaining to the appropriate time frame. Further, the *salary.csv* file initially contained 25,575 observations, but only those records pertaining to batters from the years 2013-2015 were retained. This table was left joined to the batting table using the following fields: *player\_id*, *year*, and *team\_id*, so the salary associated with each player represented the appropriate player, year of batting, and relevant team. After all joins, the final result set contained 4,330 observations.

The pitching data set was filtered for the last 3 years based on the “year” field as well. The 3 year record count for pitching was 2,281. The pitching dataset also contained the field “player\_id” which was used to join to the *player.csv* file in order to get player specific details like *name*, *birth* *date*, etc. We were able to get a record to record match for all the players in *pitching.csv* from *player.csv*.

**3. Relative Age Effect**

Players born in August have a 50-60% better chance of playing in the major leagues than those born in July [4]. This finding comes as a result of the phenomenon called the Relative Age Effect. In 1985, the first major study documenting RAE was published and concluded NHL players of the early 1980s were four times more likely to be born in the first three months of the year compared to the last three months [4]. A larger study conducted in 2005 focused on European youth soccer and found that the relative age effect can be observed in almost every European country [4].

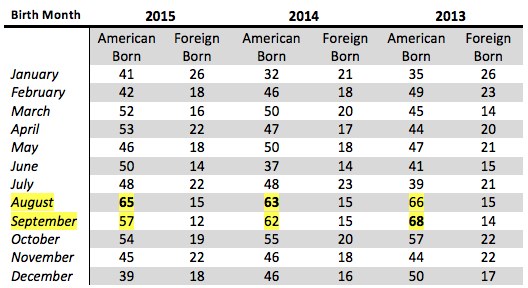
Age cut-off dates seem to be the culprit of the relative age effect and it starts with sports leagues at a very young age. Up until 2005, the age cut off date of any American baseball youth league was July 31 [4]. Therefore, a child born on August 1st is automatically the eldest of the group. Twelve full months of development makes a tremendous difference for an 11 to 12 year old. Coaches confused maturity with ability and tended to favor the eldest of the group [5]. Due to this, the eldest were given more opportunities to advance their skills, were chosen to play in all-star teams which offered them more practice and continual skill development [9]. Compared to their younger counterparts, the eldest of the group were given numerous opportunities to improve and develop with each passing year.

RAE was still present in 2005 when the last published study analyzed its effects and found that 503 American born major leaguers were born in August versus 313 born in July with a steady decrease of amount of players born from August to July [4]. The study did not find significant patterns in foreign born major leaguers where age cut-off dates are not in effect. Another significant event occurred in 2005 wherein USA baseball and all its affiliated leagues shifted the “league determination date” from July 31 to April 30 to complement the baseball season calendar. Thus, any future effects of RAE would make the majority of major leaguer’s birthdays occur in May instead of August.

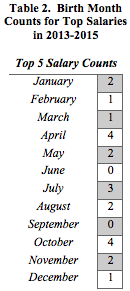
**3.1 Is RAE Still Present?**

The tables *Player* and *Salary* were joined to elicit the counts of American born versus Foreign born MLB players, by birth month. The results in *Table 1* show that most MLB players are born in August and September. August is still quite pronounced and has the highest number of players born in that month for 2014 and 2015 (bold). In 2013, September held a slight edge over August in count, which represents an insignificant difference. Overall, our results show that RAE can still be seen and may be diluted then shifted as a result of the change in league determination date.

**Table 1. *Counts of American and Foreign Born Major Leaguers by Month, from 2013 to 2015.***



**4. Salary**

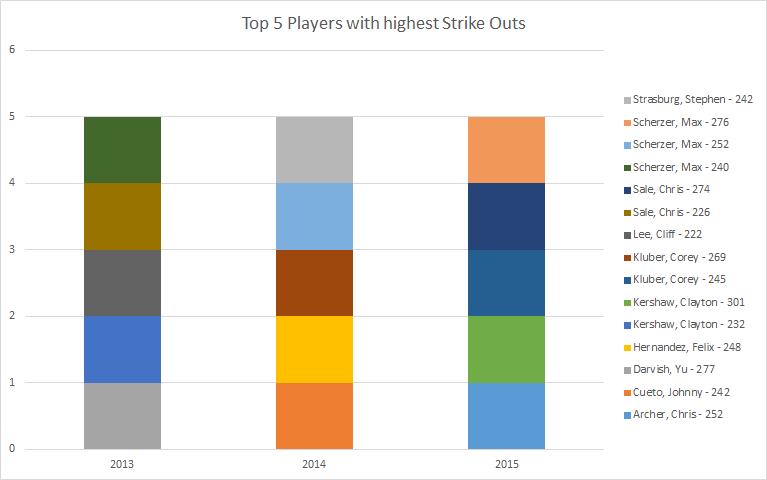


The top 5 salaries from 2013 to 2015 ranged from $24 million to $29 million. There were a few ties in highest salary ranks and counts for each month generated. *Table 2* shows the top 5 salary counts by month for the years 2013-2015. Both April and October have the highest counts of births. Although the counts may be too low to detect a significant difference, in this case, there is no relative age effect that can be seen when looking at players who get paid the most.

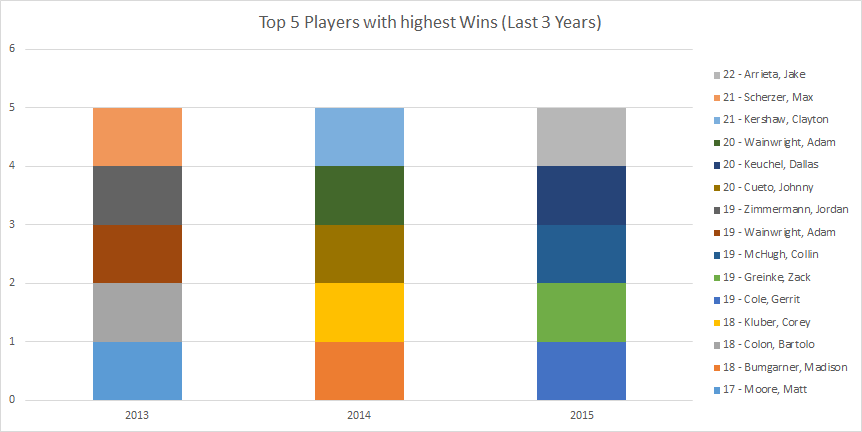
**5. Pitching**

The *pitching.csv* and *player.csv* files*,* were joined on player\_id in order to get the birth month, first name, and last name for every player. We then ranked all players to reveal which players achieved the most Wins, Strikeouts, and gave away the least Runs. This was done individually for all 3 years.

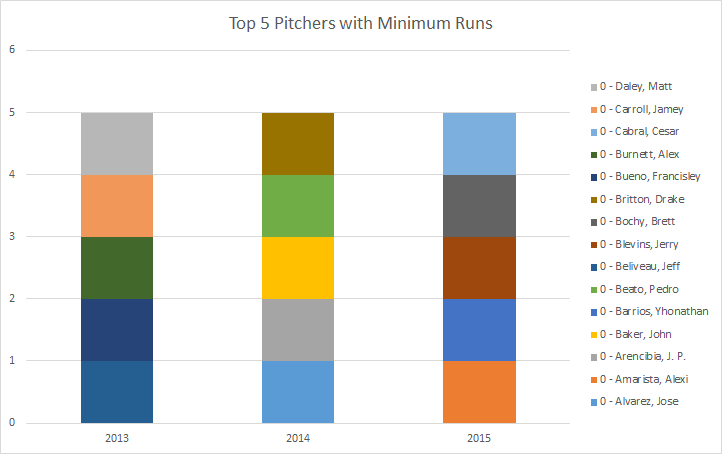
*Figure 1, Figure 2*, and *Figure 3* below show the top 5 players for the years 2013 to 2015 in *Wins*, *Strikeouts*, and *Runs*, respectively.



**Figure 1. *Top 5 Players with highest Strikeouts in 2013 to 2015***

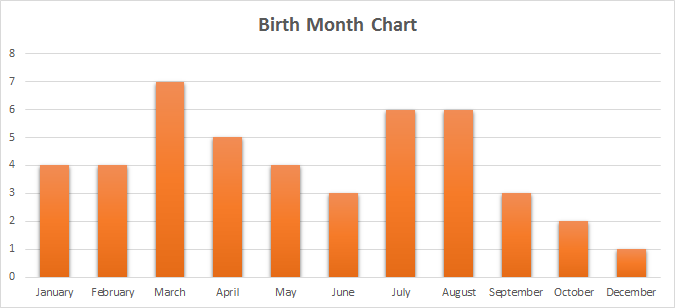


**Figure 2. *Top 5 Players with highest Wins in 2013 to 2015***



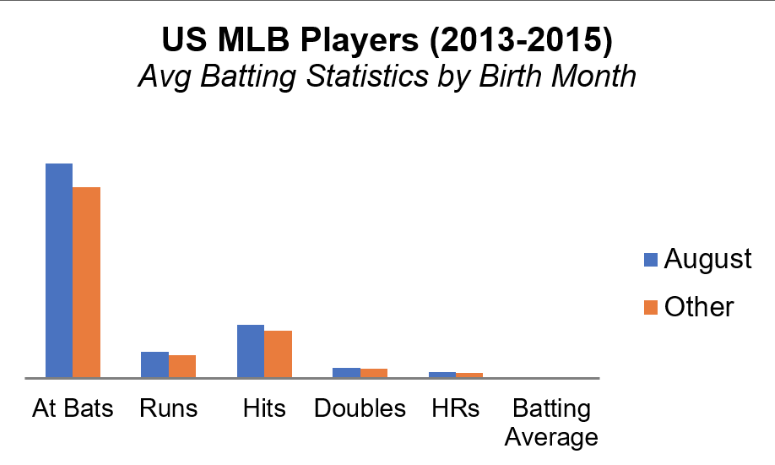
**Figure 3. *Top 5 Players with minimum Runs in 2013 to 2015***

The top 5 players for the last three years in the parameters mentioned above totaled 45 players. In order to see if RAE is in effect when it comes to the top 5 pitching statistics, *Figure 4* below shows the counts of these players with each respective birth month. We see that March stands out as the highest number of players born in that month, followed by July and August. Given these results, we can conclude that the relative age effect is not particularly seen when it comes to the players with the top 5 pitching statistics for 2013-2015.

**Figure 4. *Birth Month Counts for Top Pitchers in 2013-2015***  


**6. Batting**

Considering the initial hypothesis that US Major League Baseball players born in August are more likely to play professionally and become more successful, an initial query against our database was run to determine average batting statistics for US players born in August compared to those born in all other months to determine the degree of variance between the two groups. As a note, batting average was determined using players with at least 20 at bats to eliminate possible outliers. For example, a player with only one at bat who registered a hit would yield a batting average of 1.000, thus skewing the average. As can be seen in the graph below, the variance across the birth month groups is minimal, suggesting the hypothesized relative age effect is muted given recent data.

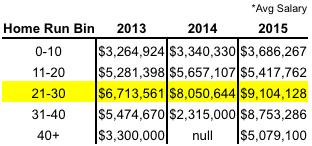


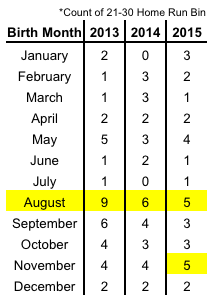
**Figure 5: *Average Batting Statistics by Birth Month***

Additionally, when exploring the top five batting averages for US players by year, the birth month with the highest frequency was April, with August in a tie for third with several other months, which again suggests the relative age effect doesn’t appear to be as dominant as acknowledged prior.

Finally, we explored salary as it pertains to batting statistics. In particular, we created 5 bins to represent the total number of home runs recorded for each US player by year. They are as follows: ‘0-10’, ‘11-20’, ‘21-30’, ‘31-40’, and ‘40+’. The average salary was highest for the ‘21-30’ home run bin for all three years of analysis, as can be seen in *Table 3* below.

**Table 3: *Average Salary by Home Run Bin***



Moreover, when exploring the frequency distribution of birth month for the ‘21-30’ home run bin across each year of analysis, as provided in *Table 4*, it was highest for August in 2013 and 2014, while there was a tie with November for 2015. Also of note is the declining frequency of players with an August birth month in the 21-30 home run bin across 2013-2015, suggesting any perceived effect and/or correlation may be fading. So, in regards to performance and salary for US players born in August, there appears to be a certain degree of correlation, which would lend credence to the presence of a relative age effect; but given the declining trend and other results to the contrary, additional analyses beyond the scope of this study are warranted to quantify the effect with any measure of statistical certainty. 

**7. Conclusion**

In all, considering the relative age effect is weak to non-existent regarding data on salary and pitching for MLB players from 2013-2015 and has a relatively weak effect on batting performance, it appears the effect serves as a stronger indicator for determining a player’s chances of making it to the major leagues more so than in determining a player’s performance once drafted. And with updated cutoff dates for youth sport leagues in effect, prior and current research into the relative age effect will need to be revisited at a later date to determine if it remains or is an outdated fantasy serving to misguide coaches, parents, and players towards an expensive and frustrating pursuit of athletics glory.

### **References**

[1] Lahman, Sean. "The History of Baseball." Kaggle. N.p., n.d. Web. Sections: ReadMe, Are Closers Overpaid, Attendance in the Modern Era, Worst World Series Winners.<https://www.kaggle.com/seanlahman/the-history-of-baseball>.

[2] "Relative Age Effect." Wikipedia. Wikimedia Foundation, 2016. Web. 08 Dec. 2016. <https://en.wikipedia.org/wiki/Relative\_age\_effect#The\_relative\_age\_effect\_in\_sports>.

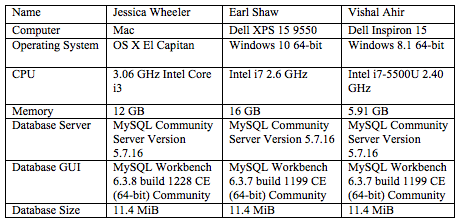
[3] Woodie, Alex. "Today's Baseball Analytics Make Moneyball Look Like Child's Play." *Datanami*. N.p., 23 Oct. 2014. Web. 08 Dec. 2016. <https://www.datanami.com/2014/10/24/todays-baseball-analytics-make-moneyball-look-like-childs-play/>.

[4] Spira, Greg. "The Boys of Late Summer." *Slate Magazine*. N.p., 16 Apr. 2008. Web. 08 Dec. 2016. <http://www.slate.com/articles/sports/sports\_nut/2008/04/the\_boys\_of\_late\_summer.html>.

[5] Bradstreet, Tyler. "The “Relative Age Effect” in Sports." *IN FOCUS*. N.p., 15 Aug. 2013. Web. 08 Dec. 2016. <https://attainingmentalfocus.wordpress.com/2013/02/28/the-relative-age-effect-in-sports/>.

**Appendix**

**Table 5. *Configurations***



**GitHub Repository:**

<https://github.com/jessicawheeler/BaseballAndRAE>