# Do Different Nationalities Influence Certain Stats?

Ben Phillips, Cooper Shawver, Drake Mills, Hughston Preston

**Introduction/Lit Review**

The globalization of Major League Baseball has really increased in the past decade. Most major league players come from the United States but recently we've seen an increase in Asian and Latino players. In this project we wanted to analyze whether certain nationalities may have a correlation with statistics. In other words, we wanted to research certain statistics and see if players from certain nationalities are more likely to produce more or less of a statistic. For example, are players from the Dominican Republic more likely to hit more home runs? Or are players from Japan more likely to have a higher OBP?

Another aspect we wanted to research was player aging, and whether certain nationalities are more immune to aging than others. An article we found analyzed national and international players in a German handball league. They found that the international players are more likely to retire earlier (Schorer, 2009). Unfortunately, this is hard to correlate to Major League Baseball players of different nationalities because of the randomness of players longevity. Players who have ever been superstars in the league are more likely to have a longer career because their names are well known, and they've proven they can play at a high level. It's more difficult to try and figure out the longevity of the average international player because they usually come into the league at a slightly younger age so they have more time to develop so they could turn into a superstar, which makes it even more difficult to project their career. It's easy to say players like Fernando Tatis Jr and Ronald Acuna Jr. are going to play until their late thirties or forties but players like Yoan Moncada and Shohei Ohtani are hard to analyze because they're both young and have potential but they're coming off bad seasons.

Ever since the integration of Major League Baseball there has been an increase in African American players in the league until just recently. An article we found while researching

explores the globalization of the game and it was found that since WW2 the number of nonnative athletes in the league has increased which would lead you to assume that the game has done a good job expanding across the world. This is not the case though, most of the foreign players in the league since the 1990's has been from Latin America and especially the Dominican Republic. Most African, European, and Asian countries have shown little signs of participation in the league compared to Latin America (Chen, 2012).

The last thing we researched was how nationalities may impact the position that players play. Mark Armour and Daniel Levitt in their article "Baseball Demographics, 1947-2016" emphasize the impact of nationality on the position you play. They found that African American players are very underrepresented as pitchers and are more frequently position players. Furthermore, there are about 10 times more Latino pitchers than African American pitchers. African American players accounted for about 40% of all outfielders from 1967-1999 below dropping below 30%. Latino players have accounted for at least 20% of all outfielders in the league over the past 20 years.

In conclusion, the past 20 years has witnessed a decline in African American players but also an increase in players from Latin America. This decline can be witnessed for all positions, but part of this decline is due to teams having more roster spots being reserved for pitchers and catchers. Even though there is a decline in African American players there has been an increase for Asian and Latino players. To put it simply, the MLB has been more diverse than it has ever been. Once the datasets were completed, visualizations exploring significance could be generated and regressions could be run to prove that significance.

## Data/Methodology

Starting off, it was necessary to compile a large and comprehensive dataset that contained players from 1980-2020 and the necessary variables to conduct a proper analysis. This was done using the Sean Lahman database located in the baseballr package in the R coding software. In addition, fangraphs dashboards were scraped to acquire the stats that the Lahman database was missing. One of these stats which may not be known to the common fan is BsR. BsR is a fangraphs stat that compiles all different aspects of baserunning into one comprehensive stat. Essentially, it combines stolen bases, caught stealing, and grounding into double plays to create an all encompassing stat. In addition to BsR, the WAR stat was used. This stat is all encompassing and captures all aspects of the game. This includes offensive production, defense, and baserunning for batters, and for pitchers how their pitching impacts each game. Overall, many complex stats were used in order to get a complete analysis.

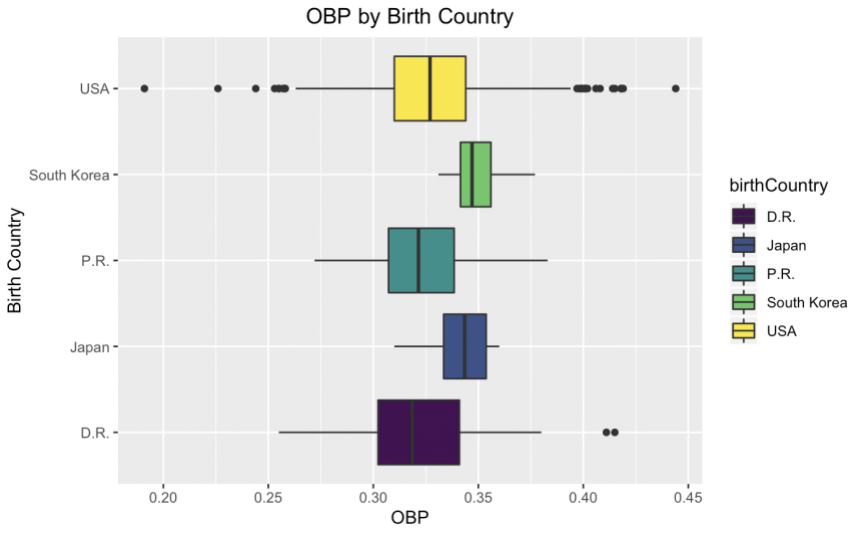
After compiling these two datasets, an inner join by name was used in order to combine them into one. Next, it was necessary to clean the data and compute career averages. Career averages would be better used for analysis since career totals rely too heavily on a player staying injury free and having a long MLB career. The initial dataset simply had career totals, so a line of code was run that would loop through the columns and divide the totals by the career length variable. For example, in order to get HRs per year of a player, it was necessary to do career HRs divided by the career length. This process produced a series of career average variables that could be utilized when conducting an analysis.

Next, it was necessary to generate dummy variables based on each player's birth country. These would become the essential independent variables for the analysis. This was done using the fast dummies package in R. This produced binary variables for each birth country found in

the dataset. Essentially, for each country, if the player was born in country x, they would get a 1, saying that the argument that they were born in country x is true. In addition, for every other country, that player will get a 0, which states that the argument is false. After generating the dummy variables, the dataset had to be cleaned of countries with sample sizes that were too small. This was because certain countries only had one player that was born there. The analysis would have been skewed if South African batters were found to be significant in increasing WAR if there was only one data point. This problem required certain countries to be filtered out of the dataset. The five countries that made the analysis were the USA, Puerto Rico, the Dominican Republic, Japan, and South Korea. This was done in order to include the USA, which was by far the largest sample size, as well as the two largest asian markets and the two largest latin american markets.

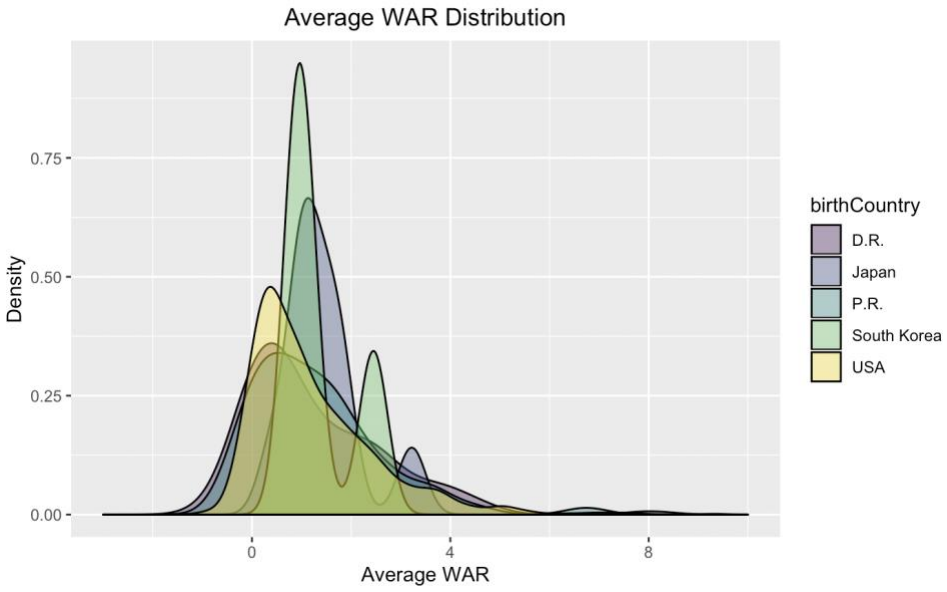
## Results/Discussion

Starting off, a variety of visualizations were created in order to investigate potential significance in the data. The first of which is a boxplot which investigates the OBP distributions of all of the players given their country. A boxplot is one of the best visualizations to use since it explores and compares entire distributions. This OBP boxplot demonstrates that both South Korean and Japanese players have higher OBPs across the entire distribution. This is because

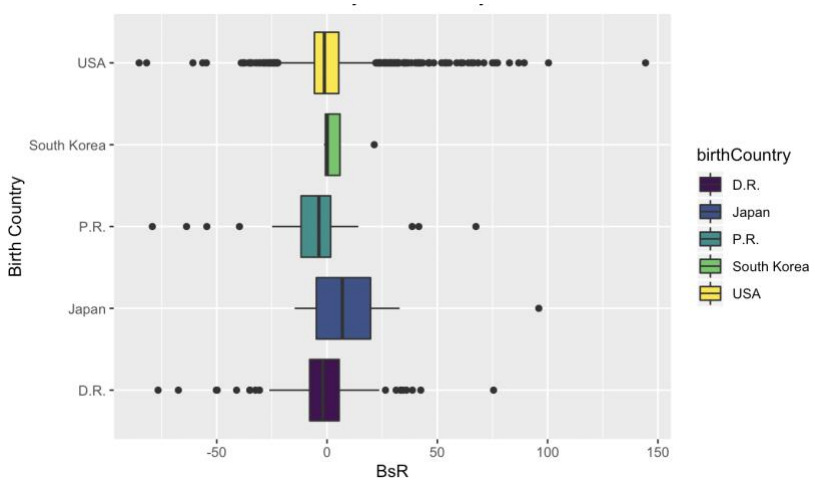


their 1st quartile, median, and 3rd quartile fall furthest to the right compared to the other countries. This demonstrates potential significance that Japanese and South Korean born players get on base at a higher rate.

The next generated visualization is a density plot which explores the distribution of hitter WAR given a player's birth country. The most noticeable peak on the graph is that the vast majority of South Korean hitters fall in the range 1-2 WAR on average. In addition, most of the peaks fall at that same point, with one outlier peak of Japanese players falling at around 3.5-4 WAR. Other than this outlier peak, however, there isn't much on the graph that would predict any type of significance in a linear regression model.

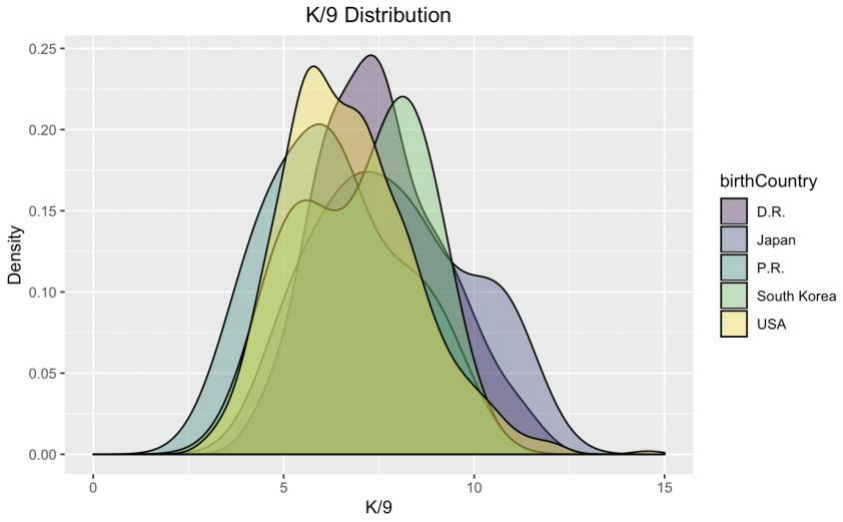



Moving on, the next visualization generated was a boxplot which investigates the distribution of BsR based on the birth country of the player. Looking at this visual, there is clear observation that Japanese born players have a higher average BsR than the rest of the countries included in the



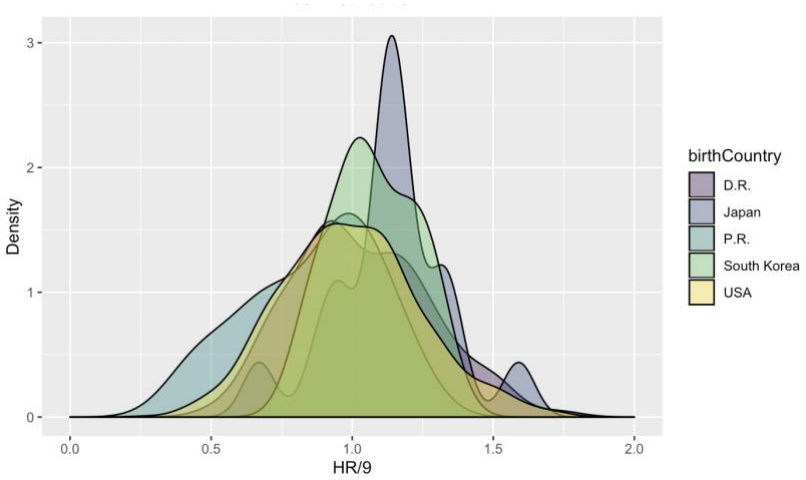
analysis. Their entire plot falls further to the right than the rest of the countries. What this means is if this visual was in the form of a density plot, there would be a clear peak of Japanese born players that fell far to the right of the rest of them. Upon further investigation, a regression would be able to determine whether there is actual significance of Japanese born players having a higher career BsR average.

The next two visualizations take a look at pitcher stats to see if there is any potential for significance. Starting off, a density plot was created to investigate the Strikeouts per 9 innings of all players given their birth countries. There appears to be potential significance where the USA peak falls to the left of the rest. In addition, South Korea and Puerto Rico have peaks which fall further to the right. This means there could be significance of those pitchers having a higher K/9 with significance.




The final visualization investigated is a density plot which illustrates the HR per 9 innings that pitchers give up given their birth country. It is apparent that both Japanese and Dominican pitchers have peaks which fall further to the right, meaning there could be significance that those nationalities tend to give up more HR/9. In addition, Puerto Rican pitchers may give up less HR/9 than the other nationalities based on the peak in their density curve which

falls further to the right of the rest of the curves. Overall, there are many potential opportunities for significance that need to be proven through linear regression models.




We ran multiple regressions in R with each stat we wanted to test as our dependent variable. The first statistics that we attempted to regress were pitcher WAR and hitter WAR. None of the five countries were close to significant and had small estimates. This shows that even with the statistics that will later show up as significant, all countries produce players that can be successful. We then tried other statistics for hitters that did not show any significance such as wRC+, wOBA, average stolen bases, average home runs, and career length. Other statistics we tried for pitchers that did not show any significance include ERA, home runs per fly ball, FIP, xFIP, and innings pitched. We finally found six statistics that did show significance. The first hitter statistic was on base percentage and the results are below.

| | Estimate | P-value | Significance Level |
|---|---|---|---|
| Intercept | 0.32 | 0 | 1% |
| D.R | 0.0022 | 0.49 | - |
| Japan | 0.0208 | 0.02 | 5% |
| P.R | 0.0027 | 0.48 | - |
| South Korea | 0.0301 | 0.03 | 5% |
| USA | 0.0076 | 0 | 1% |

This regression shows that Japan, South Korea, and the United States all showed significance at at least the 5% level. South Korean players had the largest estimate with significance. The interpretation of this is that we are 95% confident South Korean players get on base at a 3%

higher rate than the average player. The United States had less than a percentage difference than the intercept estimate, but was still significant at the 1% level.

The next statistic for hitters that we found significance in is walk percentage. The results of the regression are below.

| | Estimate | P-value | Significance Level |
|---|---|---|---|
| Intercept | 7.349 | 0 | 1% |
| D.R | 0.078 | 0.798 | - |
| Japan | 1.141 | 0.174 | - |
| P.R | 0.067 | 0.071 | - |
| South Korea | 3.951 | 0.003 | 1% |
| USA | 1.297 | 0 | 1% |

This regression tells us that players from South Korea and the United States walk at a higher rate. Both of their estimates are positive and significant at the 1% level. The South Korea statistic has an estimate of almost 4 percent higher than the intercept estimate.

The last hitter statistic that we found was significant was BsR. The results of this regression are below.

| | Estimate | P-value | Significance Level |
|---|---|---|---|
| Intercept | -4.025 | 0.002 | 1% |
| D.R | 2.626 | 0.198 | - |
| Japan | 18.475 | 0.001 | 1% |
| P.R | -0.895 | 0.719 | - |
| South Korea | 9.075 | 0.299 | - |
| USA | 5.296 | 0 | 1% |

This regression tells us that Japan and the United States both show significance at the 1% level. Their P-values are both very small which strengthens the significance. Japan has a very large estimate compared to the others which shows that Japanese baserunners run very efficiently compared to the rest of the league. The United States also has a positive estimate which shows that they are relatively efficient.

The first pitching statistic that we found significance in is walks per nine innings. The results of the regression are below.

| | Estimate | P-value | Significance Level |
|---|---|---|---|
| Intercept | 3.255 | 0 | 1% |
| D.R | 0.326 | 0.001 | 1% |
| Japan | -0.186 | 0.357 | - |
| P.R | 0.267 | 0.129 | - |
| South Korea | -0.001 | 0.998 | - |
| USA | 0.075 | 0.283 | - |

This regression shows that the Dominican Republic produces pitchers that walk more players per nine innings. This is significant at the 1% level and the interpretation of the estimate is that we are 99% confident that pitchers from the Dominican Republic walk 0.33 more hitters per nine innings than the average player.

The next pitching statistic that we had significance is strikeouts per nine innings. The results of the regression are below.

| | Estimate | P-value | Significance Level |
|---|---|---|---|
| Intercept | 7.093 | 0 | 1% |
| D.R | 0.468 | 0.038 | 5% |
| Japan | 0.812 | 0.076 | 10% |
| P.R | -0.78 | 0.051 | 10% |
| South Korea | -0.037 | 0.962 | - |
| USA | -0.364 | 0.022 | 5% |

This regression shows that the Dominican Republic, Japan, Puerto Rico, and the United States are all significant. The Dominican Republic has a positive estimate that is significant at the 5% level while the United States has a negative estimate that is significant at the 5% level. Japan has a relatively large positive estimate and is only significant at the 10% level. Puerto Rico has a negative estimate that is significant at the 10% level. The conclusions we can make from this regression are that pitchers from Puerto Rico and the United States strikeout batters at a lower rate and pitchers from The Dominican Republic and Japan strikeout batters at a higher rate.

The last pitching statistic we found that was significant is home runs per nine innings. The results of the regression are below.

| | Estimate | P-value | Significance Level |
|---|---|---|---|
| Intercept | 1.013 | 0 | 1% |
| D.R | 0.021 | 0.505 | - |
| Japan | 0.122 | 0.059 | 10% |
| P.R | -0.153 | 0.006 | 1% |
| South Korea | 0.063 | 0.573 | - |
| USA | -0.019 | 0.382 | - |

This regression tells us that Japan and Puerto Rico show significance. Japan has a positive estimate and is only significant at the 10% level and Puerto Rico has a negative estimate that is significant at the 1% level. This tells us that we can be confident that Puerto Rican pitchers give up home runs at a smaller rate and Japanese Pitchers surrender home runs at a higher rate.

These regressions answer our question of what on-field statistics might be predictable based on nationality. Certain statistics of pitchers and hitters can be influenced by how they were developed in their respective countries. This answer could be a direct sign of how young players are taken from Puerto Rico and the Dominican Republic compared to the older and more developed players taken from the Japanese and Korean leagues.

## Conclusion

As MLB player salaries rise, international markets in baseball offer the potential of cheap talent acquisition. However, it could be important to understand the trends of how baseball talent is developed and scouted in these countries in order to make effective decisions and projections as to how they may perform and fit in the MLB. In this project, we set out to find if player nationality impacted on-field performance in the MLB.

Using Fangraphs career data and Sean Lahman's baseball database, we used linear regression models to explore how national and international players have fared in the MLB in the

time span of 1980-2020. We ran multiple linear regression models on both hitting and pitching stats for our subsample of USA, Japan, South Korea, Dominican Republic, and Puerto Rico.

On the offensive side, we found that Japanese, South Korean, and American players have posted slightly higher OBP rates than the average player. We found that South Korean and American players posted considerably higher walk rates than the average player. Finally, we found that Japanese and American players have demonstrated higher BsR rates than the average player with Japanese players posting significantly higher rates. On the pitching side, we found that Dominican players tend to have higher walk rates. Japanese and Dominican pitchers have shown higher strikeout rates while Puerto Rican and American pitchers have posted lower strikeout rates than the average pitcher. Finally, we found that Japanese pitchers surrender home runs at a higher rate than the average player while Puerto Rican pitchers give them up at a lower rate.

The results of our study are quite interesting and open this topic up to further questions. The main result of our research is that players of different internationalites do in fact demonstrate different performances in a number of on field statistics. This is important because it suggests that on baseball's global scale, the game is being taught and played in several different styles. Our research suggests that American and Asian MLB players play the game with an emphasis for getting on base and efficiently moving along the base path. Meanwhile, we have evidence to conclude that Dominican pitchers pitch aggressively and therefore surrender higher walk rates than average but also achieve higher strikeout rates as a result. Meanwhile Purto Rican pitchers seem to be more contact driven by posting lower strikeout rates than average but also seeing higher home run rates, suggesting that hitters put the ball in play far more often on them with both negative and positive effects of that.

The other aspect to these findings is that they suggest that there are specific performative barriers of entry into the MLB for each country. It is possible that since Latin players are scouted at such a young age their play in the MLB is a reflection of that. When scouting a 16 year old, physical tools like arm strength, speed, pitch movement may be identified with more confidence than mental tools like a batter's plate discipline or contact ability. Hence, Dominican pitchers demonstrated higher strikeout rates but more random variation in other statistics. Meanwhile Asian markets have well developed leagues and are usually signed to MLB teams at far older ages. In the case of Japan, their NPB contracts often prohibit their players from signing MLB deals until their mid 20s. The effect of this is that mental skills like plate discipline and contact ability may develop and be scouted more confidently.

In conclusion, we found that players of different nationalities demonstrate differing on field performances in the MLB. It illuminates possible further areas of study into the effect of international scouting rules and regulations upon the play styles and barriers of entry for international players.

**Works Cited**

Admin. "Baseball Demographics, 1947-2016." *Society for American Baseball Research*, Admin /Wp-Content/Uploads/2020/02/sabr_logo.Png, 9 Apr. 2020, sabr.org/bioproj/topic/baseball-demographics-1947-2016/.

Ke Chen, Charles Gunter and Chunhua Zhang, How global is U.S. Major League Baseball? A historical and geographic perspective, https://www-jstor-org.libezproxy2.syr.edu/stable/23254353?Search=yes&resultItemClick=true&searchText=mlb&searchUri=%2Faction%2FdoBasicSearch%3FQuery%3Dmlb&ab_segments=0%2Fbasic_search_SYC-5462%2Ftest&refreqid=fastly-default%3Af77950623d50fa26a1462bd8d19f14b6&seq=1#metadata_info_tab_contents

Schorer, J., et al. "Influences of Competition Level, Gender, Player Nationality, Career Stage and Playing Position on Relative Age Effects." *Wiley Online Library*, John Wiley & Sons, Ltd, 8 July 2008, onlinelibrary.wiley.com/doi/full/10.1111/j.1600-0838.2008.00838.x?casa_token=t-D1V-INMcQAAAAA%3AoItBeu8xU-5AH5zfXG7PCSspF8wcgI1ZAibQ283_oOI6ABsnzawLhpRkzsH40Rkb-19yvUK0wexr.