Supporting data analysis for “Challenging nostalgia and performance metrics in baseball”

Daniel J. Eck
Department of Statistics, University of Illinois Urbana-Champaign
email: `dje13@illinois.edu`

Abstract

We show that the great baseball players that started their careers before 1950 are overrepresented among rankings of baseball’s all time greatest players. The year 1950 coincides with the decennial US Census that is closest to when Major League Baseball (MLB) was integrated in 1947. We also show that performance metrics used to compare players have substantial era biases that favor players who started their careers before 1950. In showing that the these players are overrepresented, no individual statistics or era adjusted metrics are used. Instead, we argue that the eras in which players played are fundamentally different and are not comparable. In particular, there were significantly fewer MLB eligible players available at and before 1950. As a consequence of this and other differences across eras, we argue that popular opinion, performance metrics, and expert opinion over include players that started their careers before 1950 in their rankings of baseball’s all time greatest players.

Creating this document

The purpose of this document is to make transparent the analysis in Eck (2019). All calculations and tables that appear in Eck (2019) are displayed in this document. This document doubles as a technical report that can be reproduced by anyone with a computer who has the file

`population.csv.`

This file is available at the author’s GitHub page along with this technical report. See here:

`https://github.com/DEck13/baseball_research/tree/master/Challenging_nostalgia`

The columns age20 and age25 corresponding to counts of males aged 20-24 and 25-29 respectively.

1 Introduction

It is easy to be blown away by the accomplishments of great old time baseball players when you look at their raw or advanced baseball statistics. These players produced mind-boggling numbers. For example, see Babe Ruth’s batting average and pitching numbers, Ty Cobb’s 1911 season, Walter Johnson’s 1913 season, Tris Speaker’s 1916 season, Rogers Hornsby’s 1925 season, and Lou Gehrig’s 1931 season. The statistical feats achieved by these players (and others) far surpass the statistics that recent and current players produce.

At first glance it seems that players from the old eras were vastly superior to the players in more modern eras, but is this true? In this paper, we investigate whether baseball players from earlier eras of professional baseball are overrepresented among the game’s all-time greatest players according to popular opinion, performance metrics, and expert opinion.

We define baseball players from “earlier eras” to be those that started their MLB careers in the 1950 season or before. This year is chosen because it coincides with the decennial US Census and is close to 1947, the year in which baseball became integrated.

This article does not compare baseball players via their statistical accomplishments. Such measures exhibit era biases that are confounded with actual performance. Consider the single season homerun record as an example. Before Babe Ruth, the single season homerun record was 27 by Ned Williamson in 1884. Babe Ruth broke this record in 1919 when he hit 29 homeruns. He subsequently destroyed his own record in the following 1920 season when he hit 54 homeruns. The runner up in 1920 finished the season with a grand total of 15 homeruns.

At that point, homerun hitting was not an integral part of a batter's approach. This has changed. Now, we often see multiple batters reach at least 30-40 homeruns within one season and a 50 homerun season is not a rare occurrence. In the 1920s, Babe Ruth stood head and shoulders above his peers due to a combination of his innate talent and circumstance. His approach was quickly emulated and became widely adopted. However, Ruth's accomplishments as a homerun hitter would not stand out nearly as much if he played today and put up similar homerun totals.

The example of homeruns hit by Babe Ruth and the impact they had relative to his peers represents a case where adjustment towards a peer-derived baseline fails across eras. No one reasonably expects 1920 Babe Ruth to hit more than three times the amount of homeruns hit by the second best homerun hitter if the 1920 version of Babe Ruth played today. This is far from an isolated case.

Several statistical approaches are currently used to compare baseball players across eras. These include wins above replacement as calculated by baseball reference (bWAR), wins above replacement as calculated by fangraphs (fWAR), adjusted OPS+, adjusted ERA+, era-adjusted detrending (Petersen et al., 2011), computing normal scores as in Jim Albert's work on a Baseball Statistics Course in the Journal of Statistics Education, and era bridging (Berry et al., 1999). A number of these are touted to be season adjusted and the remainder are widely understood to have the same effect.

In one way or another, all of these statistical approaches compare the accomplishments of players within one season to a baseline that is computed from statistical data within that same season. This method of player comparison ignores talent discrepancies that exist across seasons as noted by Stephen J. Gould in numerous lectures and papers. Currently, there is no definitive quantitative or qualitative basis for comparing these baselines, which are used to form intra-season player comparisons, across seasons. These methods therefore fail to properly compare players across eras of baseball despite the claim that they are season adjusted.

Worse still, these approaches exhibit a favorable bias towards baseball players who played in earlier seasons (Schmidt and Berri, 2005). We explore this bias from two separate theoretical perspectives underlying how baseball players from different eras would actually compete against each other.

The first perspective is that players would teleport across eras to compete against each other. From this perspective, the players from earlier eras are at a competitive disadvantage because, on average, baseball players have gotten better as time has progressed. Specifically, it is widely acknowledged that fastball velocity, pitch repertoire, training methods, and management strategies have all improved over time. We do not find the teleportation perspective to be of much interest for these reasons.

The second perspective is that a player from one era could adapt naturally to the game conditions of another era if they grew up in that time. This line of thinking is challenging to current statistical methodology because adjustment to a peer-derived baseline no longer makes sense. Even in light of these challenges with the second perspective, we find that the players from earlier eras are overrepresented among baseball's all time greats. We justify our findings through the consideration of population dynamics which have changed drastically over time.

2 Data

The MLB eligible population is not well-defined. As a proxy, we can say that the MLB eligible population is the decennial count of males aged 20-29 that are living in the United States (US) and Canada. Baseball was segregated on racial grounds until 1947. As a result, African American and Hispanic American population counts in the US and Canada are added to our dataset starting in 1960. The year 1960 is chosen because the integration of the MLB was slow as noted in Armour's work on the integration of baseball in the Society for American Baseball Research.

Players from Latin, Central and South American countries and the Caribbean islands were also targets of discrimination. We have added data from these countries to the MLB eligible population starting in 1960: Aruba, the Bahamas, Colombia, Cuba, the Dominican Republic, Honduras, Jamaica, Mexico, Netherlands Antilles, Nicaragua, Panama, Peru, Puerto Rico, the United States Virgin Islands, and Venezuela.

In the mid- to late 1990s, the MLB and minors saw an influx of Asian baseball players from Japan, South Korea, Taiwan, and the Philippines. We have added the populations of these countries to the MLB eligible population starting in 2000.

In 2010, the MLB established a national training center in Brazil as noted in Loré's work on the popularity of baseball in Brazil in the Culture Trip. Therefore, we have included the Brazilian population of 20-24 year old men into our MLB eligible population starting in 2010. We estimate that the 2011-2015 MLB eligible population is half of the MLB eligible population counted in the 2010 decennial Censuses. We expect that this underestimates the actual 2011-2015 MLB eligible population since we have observed a constant increase in the overall MLB eligible population as time increases.

We now load in the population dataset and begin our analysis. The code below creates our MLB eligible population.

```
# load in UN dataset and remove irrelevant variables
library(dplyr)
library(stringr)
library(knitr)
options(warn=-1)
population_data <- read.csv("population.csv", header = TRUE)[, -1] %>%
  mutate(age20 = age20 / 1e3, age25 = age25 / 1e3) %>%
  filter(year > 1950) %>% mutate(region = as.factor(region))

# restrict attention to countries of interest
countries <- c("Aruba", "Bahamas", "Brazil", "Colombia", "Cuba",
  "Dominican Republic", "Honduras", "Jamaica", "Japan", "Mexico",
  "Netherlands Antilles", "Nicaragua", "Panama", "Philippines", "Peru",
  "Puerto Rico", "Republic of Korea", "Taiwan",
  "Venezuela (Bolivarian Republic of)", "United States Virgin Islands",
  "Canada", "United States of America")

# trim population dataset
foo <- population_data[population_data$region %in% countries, ]
foo$region <- droplevels(foo$region)
bar <- split(foo, f = as.factor(foo$region))

# highlight countries that will be included starting in 2000
asian_countries <- c("Japan", "Philippines", "Republic of Korea", "Taiwan")

# extract relevant age 20-29 totals
baz <- lapply(bar, function(xx){
  region <- unique(xx$region)
```

```
  # only include age 20-24 males from Brazil staring in 2010
  if(region == "Brazil"){
    xx <- (xx %>% filter(year >= 2010))
    xx$age25 <- 0
    xx <- xx %>% mutate(total = age20 + age25) %>%
      dplyr::select("year", "total")
  }
  # filter out Asian countries
  else if(region %in% asian_countries){
    xx <- xx %>% mutate(total = age20 + age25) %>%
      dplyr::select("year", "total") %>% filter(year >= 2000)
  }
  else{
    xx <- xx %>% mutate(total = age20 + age25) %>%
      dplyr::select("year", "total")
  }
})
```

We now add in the Taiwan population figures from the CIA World Factbook. The CIA World Factbook does not provide population data for men aged 20-29 for 2000 or 2010. It does provide estimated population data for men aged 15-24 for 2018. We will use this estimate as the estimate of the Taiwan male population aged 20-29 for 2000 or 2010.

```
#approximation from CIA World Factbook at
#https://www.cia.gov/library/publications/the-world-factbook/geos/print_tw.html
baz$'Taiwan' <- data.frame(year = c(2000, 2010), total = c(1.5, 1.5))
```

We now add in the US and Canadian population data for the Census years from 1880-1950. Note that the Canadian Census is recorded one year after the US Census. A precise Canadian Census exists for males aged 20-30 for the years 1881, 1891, and 1901. We use this information to estimate the proportion of Canadian males aged 20-30 from 1911-1951 for which an accurate count of Canadian males aged 20-30 was not easy to find.

```
### Canadian population before 1960
#https://www65.statcan.gc.ca/acyb02/1907/acyb02_1907001701a-eng.htm
Can1881 <- 0.21 + 0.17
#https://www65.statcan.gc.ca/acyb02/1907/acyb02_1907001701a-eng.htm
Can1891 <- 0.24 + 0.19
#https://www65.statcan.gc.ca/acyb02/1907/acyb02_1907001701a-eng.htm
Can1901 <- 0.26 + 0.22
#https://www65.statcan.gc.ca/acyb02/1947/acyb02_19470113009-eng.htm
CanM1881 <- 2.19
#https://www65.statcan.gc.ca/acyb02/1947/acyb02_19470113009-eng.htm
CanM1891 <- 2.16
#https://www65.statcan.gc.ca/acyb02/1947/acyb02_19470113009-eng.htm
CanM1901 <- 2.75
propCan2030 <- (Can1881 + Can1891 + Can1901) /
```

```
  (CanM1881 + CanM1891 + CanM1901)
#https://www65.statcan.gc.ca/acyb02/1947/acyb02_19470113009-eng.htm
Can1911 <- 3.82 * propCan2030
#https://www65.statcan.gc.ca/acyb02/1947/acyb02_19470113009-eng.htm
Can1921 <- 4.53 * propCan2030
#https://www65.statcan.gc.ca/acyb02/1947/acyb02_19470113009-eng.htm
Can1931 <- 5.37 * propCan2030
#https://www65.statcan.gc.ca/acyb02/1947/acyb02_19470113009-eng.htm
Can1941 <- 5.90 * propCan2030
#https://www65.statcan.gc.ca/acyb02/1967/acyb02_19670194010-eng.htm
Can1951 <- 7.09 * propCan2030

### US population before 1960
#https://www2.census.gov/library/publications/decennial/1880/vol-01-population/188
US1880 <- 2.22 + 1.84
#https://www2.census.gov/library/publications/decennial/1900/volume-2/volume-2-p5.p
US1900 <- 2.73 + 2.37
US1890 <- mean(c(US1880,US1900))
#https://www2.census.gov/library/publications/decennial/1910/volume-1/volume-1-p6.p
US1910 <- 4.07 + 3.79
#https://www2.census.gov/library/publications/decennial/1920/volume-2/41084484v2ch
US1920 <- 4.02 + 4.09
#https://www2.census.gov/library/publications/decennial/1930/population-volume-2/1
US1930 <- 4.69 + 4.25
#https://www2.census.gov/library/publications/decennial/1940/population-volume-4/3
US1940 <- 1.08 + 1.06 + 1.01 + 1.00 + 1.01 +
  1.01 + 0.99 + 0.98 + 0.97 + 0.95
#https://www2.census.gov/library/publications/decennial/1950/population-volume-2/2
US1950 <- 5.00 + 5.30
```

The final dataset with all countries' populations is created.

```
# get post 1960 population totals
qux <- do.call(rbind, baz)
pops_year_post1950 <- unlist(lapply(split(qux, f = as.factor(qux$year)),
  function(xx) sum(xx$total)))

# build MLB eligible population
pops <- as.numeric(c(US1880 + Can1881,
  US1890 + Can1891,
  US1900 + Can1901,
  US1910 + Can1911,
  US1920 + Can1921,
  US1930 + Can1931,
  US1940 + Can1941,
  US1950 + Can1951,
  pops_year_post1950))
pops <- c(pops, pops[14]/2)
```

```r
# proportion of total population in current year and previous years
pop_prop <- cumsum(pops) / sum(pops)

# build the dataset
year <- c(1880 + 0:13 * 10, 2015)
dat <- cbind(year, pops, pop_prop)
dat[, 2] <- round(dat[, 2], 2)
```

The MLB eligible population is displayed in Table 1. The cumulative proportion means that at each era, the population of the previous eras is also included. As an example of how to interpret this dataset, consider the year 1950. There were 11.59 million males aged 20-29. The proportion of the historical MLB eligible population that existed at or before 1950 is 0.178.

```r
# the dataset
library(xtable)
colnames(dat)[c(2:3)] <- c("population", "cumulative population proportion")
print(xtable(as.data.frame(dat), digits = c(0,0,3,3),
  align = c("l","c","c","c"),
  caption = "Eligible MLB population throughout the years. The first
    column indicates the year, the second column indicates the
    estimated MLB eligible population size (in millions), and the
    third column indicates the proportion of the MLB eligible
    population in row x that was eligbile at or before row x."))
```

	year	population	cumulative population proportion
1	1880	4.440	0.012
2	1890	5.010	0.026
3	1900	5.580	0.041
4	1910	8.550	0.064
5	1920	8.930	0.089
6	1930	9.920	0.116
7	1940	11.130	0.146
8	1950	11.590	0.178
9	1960	19.580	0.231
10	1970	26.100	0.303
11	1980	36.370	0.402
12	1990	40.680	0.513
13	2000	64.060	0.688
14	2010	76.190	0.896
15	2015	38.100	1.000

Table 1: Eligible MLB population throughout the years. The first column indicates the year, the second column indicates the estimated MLB eligible population size (in millions), and the third column indicates the proportion of the MLB eligible population in row x that was eligbile at or before row x.

3 The greats

To determine which players are the all-time greatest players, we consult four lists which reflect popular opinion, performance metrics, and expert opinion that purport to determine the greatest players. The first list is compiled by Ranker, which is constructed entirely from popular opinion as determined by up and down votes. The second and third lists rank players by highest career WAR as calculated by baseball reference and fangraphs, respectively. These three lists were compiled in 2018. The fourth list is a ranking from ESPN and is based on expert opinion and statistics.

The rankings for all four lists are given in Table 2. As an example of the information contained in Table 2 consider the greatest players of all time according to ESPN displayed in the fourth column. We see that 5 players that started their careers before 1950 are in the top 10 all time and 11 players that started their careers before 1950 are in the top 25 all time. When the MLB eligible population is considered, it appears that the players from the earlier eras are overrepresented in this particular list.

4 Statistical evidence

We now provide evidence that the top 10 and top 25 lists displayed in Table 2 overrepresent players who started their careers before 1950. We require two assumptions for the validity of our calculations which we will explore in detail in the next Section. These assumptions are:

- First, we assume that innate talent is uniformly distributed over the MLB eligible population over the different eras.
- Second, we assume that the outside competition to the MLB available by other sports leagues after 1950 is offset by the increased salary incentives received by MLB players.

With these assumptions in mind we calculate the probability that at least x people from each top 10 and top 25 list in Table 2 started their career before 1950 using the proportion depicted in Table 1. Consider the bWAR list for example. According to bWAR, we see that 6 of the top 10 players started their careers before 1950. From Table 1 we see that the proportion of the MLB eligible population that played at or before 1950 was approximately 0.178. We then calculate the probability that one would expect to observe 6 or more individuals in a top 10 list from that time period where the chance of observing each individual is about 0.178. We calculate this probability using the Binomial distribution, details are included in the Appendix. This calculation and all other calculations for each top 10 and top 25 list depicted in Table 2 are conducted using R statistical software below. The results are provided in Table 3.

```
options(scipen=999)
# proportion of MLB eligible population before who lived before 1950
p <- dat[8, 3]

# count of great players who started their careers before 1950
ranker10 <- 7; ranker25 <- 15
bWAR10 <- 6;   bWAR25 <- 15
fWAR10 <- 6;   fWAR25 <- 12
ESPN10 <- 5;   ESPN25 <- 11

# binomial calculations for top 10 lists
pranker10 <- pbinom(ranker10 - 1, p = p, size = 10, lower = FALSE)
```

rank	Ranker	bWAR	fWAR	ESPN
1	**Babe Ruth**	**Babe Ruth**	**Babe Ruth**	**Babe Ruth**
2	**Ty Cobb**	**Cy Young**	Barry Bonds	Willie Mays
3	**Lou Gehrig**	**Walter Johnson**	Willie Mays	Barry Bonds
4	**Ted Williams**	Barry Bonds	**Ty Cobb**	**Ted Williams**
5	**Stan Musial**	Willie Mays	**Honus Wagner**	Hank Aaron
6	Willie Mays	**Ty Cobb**	Hank Aaron	**Ty Cobb**
7	Hank Aaron	Hank Aaron	Roger Clemens	Roger Clemens
8	Mickey Mantle	Roger Clemens	**Cy Young**	**Stan Musial**
9	**Rogers Hornsby**	**Tris Speaker**	**Tris Speaker**	Mickey Mantle
10	**Honus Wagner**	**Honus Wagner**	**Ted Williams**	**Honus Wagner**
11	**Cy Young**	**Stan Musial**	**Rogers Hornsby**	**Lou Gehrig**
12	**Walter Johnson**	**Rogers Hornsby**	**Stan Musial**	**Walter Johnson**
13	**Joe Dimaggio**	**Eddie Collins**	**Eddie Collins**	Greg Maddux
14	Sandy Koufax	**Ted Williams**	**Walter Johsnon**	Rickey Henderson
15	Ken Griffey Jr.	**Pete Alexander**	Greg Maddux	**Rogers Hornsby**
16	**Jimmie Foxx**	Alex Rodriguez	**Lou Gehrig**	Mike Schmidt
17	**Tris Speaker**	**Kid Nichols**	Alex Rodriguez	**Cy Young**
18	**Joe Jackson**	**Lou Gehrig**	Mickey Mantle	Joe Morgan
19	Mike Schmidt	Rickey Henderson	Randy Johnson	**Joe Dimaggio**
20	Nolan Ryan	Mickey Mantle	**Mel Ott**	Frank Robinson
21	**Christy Mathewson**	Tom Seaver	Nolan Ryan	Randy Johnson
22	Roberto Clemente	**Mel Ott**	Mike Schmidt	Tom Seaver
23	Albert Pujols	**Nap Lajoie**	Rickey Henderson	Alex Rodriguez
24	**Cap Anson**	Frank Robinson	Frank Robinson	**Tris Speaker**
25	Greg Maddux	Mike Schmidt	Burt Blyleven	Steve Carlton
pre-1950 in top 10	7 / 10	6 / 10	6 / 10	5 / 10
pre-1950 in top 25	15 / 25	15 / 25	12 / 25	11 / 25

Table 2: Lists of the top 25 greatest baseball players to ever play in the MLB according to Ranker.com (1st column), bWAR (2nd column), fWAR (3rd column), and ESPN (4th column). Players that started their career before 1950 are indicated in bold. The last two rows count the number of players that started their careers before 1950 in each of the top 10 and top 25 lists respectively.

```
pbWAR10   <- pbinom(bWAR10 - 1, p = p, size = 10, lower = FALSE)
pfWAR10   <- pbinom(fWAR10 - 1, p = p, size = 10, lower = FALSE)
pESPN10   <- pbinom(ESPN10 - 1, p = p, size = 10, lower = FALSE)

# binomial calculations for top 25 lists
pranker25 <- pbinom(ranker25 - 1, p = p, size = 25, lower = FALSE)
pbWAR25   <- pbinom(bWAR25 - 1, p = p, size = 25, lower = FALSE)
pfWAR25   <- pbinom(fWAR25 - 1, p = p, size = 25, lower = FALSE)
pESPN25   <- pbinom(ESPN25 - 1, p = p, size = 25, lower = FALSE)
```

As an example of how to interpret the results of Table 3, continue with bWAR's top 10 list. Table 3 shows that the probability of observing 6 or more players that started their careers at or before 1950 of

	Ranker	bWAR	fWAR	ESPN
probability of extreme event in top 10 list	4.08×10^{-4}	0.00345	0.00345	0.0203
probability of extreme event in top 25 list	3.01×10^{-6}	3.01×10^{-6}	5.18×10^{-4}	0.00214
chance of extreme event in top 10 list	1 in 2450	1 in 290	1 in 290	1 in 49
chance of extreme event in top 25 list	1 in 3.32594×10^5	1 in 3.32594×10^5	1 in 1932	1 in 467

Table 3: The probability and chance (1 in 1/probability) of each extreme event calculation corresponding to the four lists in Table 2.

the top 10 all time players, based on population dynamics, is about 0.00345 (a chance of 1 in 290). The same interpretation applies to remainder of Table 3. The results provided in Table 3 present overwhelming evidence that players who started their careers before 1950 are overrepresented in top 10 and top 25 lists from the perspectives of fans, analytic assessment of performance, and experts' rankings.

5 Assumptions and Sensitivity Analysis

The results in Table 3 are valid under the two assumptions provided in the previous Section. In the first of these assumptions we specify that innate talent is evenly dispersed across eras. We do not fully believe that the first assumption holds because the distribution of innate talent has improved over time as the MLB eligible population has expanded as noted by Stephen J. Gould, Christina Kahrl at ESPN, and in Martin B. Schmidt and David J. Berri's work on concentration of baseball talent in the Journal of Sports Economics. This suggests that the probabilities displayed in Table 3 are conservative.

With respect to the second assumption, we note that the National Basketball Association (NBA) and the National Football League (NFL) started in 1946 and 1920 respectively with both sports greatly rising in popularity since the inception of their respective professional leagues. Soccer and hockey have also risen in popularity in the United States.

That being said, it is widely known that MLB salaries have substantially increased. For example, the 1967 census lists the median US household income as $7,200. The minimum MLB salary at that time was $6,000 as noted by the LA Times sports writer Bill Shaiken in a piece titled "A look at how Major League Baseball salaries have grown by more than 20,000% the last 50 years."

In short, baseball players made far less than they do today relative to the general US population and it is unlikely that one could consider playing professional baseball to be a lucrative career in the earlier eras. These figures offer evidence that while other professional leagues may have drawn from the MLB eligible talent pool, salary incentives have led to an increase in the overall quality of MLB players.

Although we cannot confirm this theory with absolute certainty, the second assumption suffers modest violations at worst. To account for this possibility, we consider applying a sensitivity analysis to the findings in Table 3. We weight the decennial populations displayed in Table 1 to reflect the overall interest that the US population has had in baseball over time irrespective of salary increases based on Gallup polling data.

The four weighting regimes that we consider are given in Table 4 below. These regimes serve as proxies for the proportion of the MLB eligible population thought to strive towards a career in professional baseball. In an effort to be conservative, we have deliberately placed greater weight on the time periods before 1940 for each weighting regime because no polling data is available. We do not expect the MLB eligible population before 1940 to be as high as our weighting regimes suggest because of relatively modest baseball attendance figures in early eras of baseball, non existence of the radio prior to 1920, the dead-ball era, and low compensation.

David W. Moore and Joseph Carroll's Gallup article entitled "Baseball Fan Numbers Steady, But De-

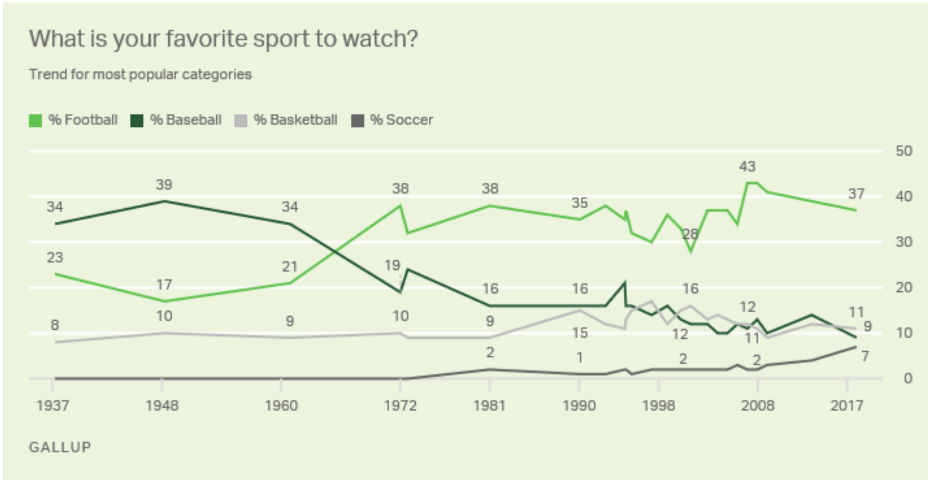

Figure 1: Gallup polling data on the topic of the favorite sports of Americans over time.

cline May Be Pending" shows that interest in baseball has remained steady since 1937, at approximately 40%. Consistent with this benchmark, the first and second weighting regimes (w1 and w2) conservatively place 0.50 and 0.60 weights, repectively, on fan interest prior to 1940. The third weighting regime (w3), constructed from the Gallup polling data in Figure 1, reflects the proportion of the US population who listed baseball as their favorite sport.

The appropriateness of this regime is intuitively questionable because some people play baseball even if it is not their favorite sport and the weight placed on pre-1940 years is very high.

The fourth weighting regime (w4) is the average of w2 and w3.

These weights are obtained from survey data from the US because similar data is unavailable from other countries. We applied these same weights to all of the other countries, even though interest in baseball in these other countries is thought to either be on par with or much greater than the US. Therefore our weighting regimes address, and in fact, overcompensate for any potential shortcomings of no weighting.

The weights are now added to our analysis, and we construct four new datasets with reweighted populations corresponding to each of the four weighting schemes outlined in Table 4.

```
# compute w1 based on information from
# Gallup on baseball fans (50% fans before 1940)
w1 <- c(rep(0.5, 6), rep(0.4, 9))

# construct rewighted dataset with respect to w1
data.w1 <- dat
data.w1[, 2] <- data.w1[, 2] * w1
data.w1[, 3] <- cumsum(data.w1[, 2]) / sum(data.w1[, 2])

# compute w2 based on information from
# Gallup on baseball fans (60% fans before 1940)
w2 <- c(rep(0.6, 6), rep(0.4, 9))

# construct rewighted dataset with respect to w2
data.w2 <- dat
```

```
data.w2[, 2] <- data.w2[, 2] * w2
data.w2[, 3] <- cumsum(data.w2[, 2]) / sum(data.w2[, 2])

# compute w3 based on information from
# Gallup on favorite sport
w3 <- c(rep(0.4, 6), 0.35, 0.38, 0.34, 0.28, 0.16, 0.16, 0.13, 0.12, 0.10)

# construct rewighted dataset with respect to w3
data.w3 <- dat
data.w3[, 2] <- data.w3[, 2] * w3
data.w3[, 3] <- cumsum(data.w3[, 2]) / sum(data.w3[, 2])

# construct rewighted dataset with respect to w4
w4 <- (w2 + w3) / 2
data.w4 <- dat
data.w4[, 2] <- data.w4[, 2] * w4
data.w4[, 3] <- cumsum(data.w4[, 2]) / sum(data.w4[, 2])
```

Table 4 displays the weighting regimes used in our analysis, and it is constructed with the code below.

```
tab <- cbind(w1,w2,w3,w4)
rownames(tab) <- year
print(xtable(t(tab), #digits = c(0,0,2,2,2,2),
  digits = cbind(rep(0, 4), matrix(2, ncol = 15, nrow = 4)),
  caption = "Weighting regimes"), size = "footnotesize",
  #hline.after = c(-1,0,nrow(data)))
  hline.after = c(-1,0,4))
```

	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010	2015
w1	0.50	0.50	0.50	0.50	0.50	0.50	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
w2	0.60	0.60	0.60	0.60	0.60	0.60	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
w3	0.40	0.40	0.40	0.40	0.40	0.40	0.35	0.38	0.34	0.28	0.16	0.16	0.13	0.12	0.10
w4	0.50	0.50	0.50	0.50	0.50	0.50	0.38	0.39	0.37	0.34	0.28	0.28	0.27	0.26	0.25

Table 4: Weighting regimes

Table 5 shows the effect of these weighting regimes as applied to the results in Table 3. The conclusions from weighting populations with respect to w1, w2, and w4 in Table 5 are largely consistent with those in Table 3. The third weighting regime presents some conflicting conclusions. When weighting populations with respect to w3, popular opinion and bWAR still overrepresent players who started their careers before 1950. However, the same is not so for fWAR and ESPN. The overall finding of this sensitivity analysis is that conservatively weighting populations with respect to fan interest in baseball yields the conclusion as the analysis in Section 4: it is very unlikely that the pre-1950s time period could have produced so many historically great baseball players.

We now perform the same Binomial distribution calculations in the previous Section with respect to the reweighted populations and each top 10 and top 25 list depicted in Table 2.

```r
## Recompute extreme probabilities consistent with w1
p.w1 <- data.w1[8, 3]
pranker10.w1 <- pbinom(ranker10 - 1, p = p.w1, size = 10, lower = FALSE)
pbWAR10.w1   <- pbinom(bWAR10 - 1, p = p.w1, size = 10, lower = FALSE)
pfWAR10.w1   <- pbinom(fWAR10 - 1, p = p.w1, size = 10, lower = FALSE)
pESPN10.w1   <- pbinom(ESPN10 - 1, p = p.w1, size = 10, lower = FALSE)
pranker25.w1 <- pbinom(ranker25 - 1, p = p.w1, size = 25, lower = FALSE)
pbWAR25.w1   <- pbinom(bWAR25 - 1, p = p.w1, size = 25, lower = FALSE)
pfWAR25.w1   <- pbinom(fWAR25 - 1, p = p.w1, size = 25, lower = FALSE)
pESPN25.w1   <- pbinom(ESPN25 - 1, p = p.w1, size = 25, lower = FALSE)

## Recompute extreme probabilities consistent with w2
p.w2 <- data.w2[8, 3]
pranker10.w2 <- pbinom(ranker10 - 1, p = p.w2, size = 10, lower = FALSE)
pbWAR10.w2   <- pbinom(bWAR10 - 1, p = p.w2, size = 10, lower = FALSE)
pfWAR10.w2   <- pbinom(fWAR10 - 1, p = p.w2, size = 10, lower = FALSE)
pESPN10.w2   <- pbinom(ESPN10 - 1, p = p.w2, size = 10, lower = FALSE)
pranker25.w2 <- pbinom(ranker25 - 1, p = p.w2, size = 25, lower = FALSE)
pbWAR25.w2   <- pbinom(bWAR25 - 1, p = p.w2, size = 25, lower = FALSE)
pfWAR25.w2   <- pbinom(fWAR25 - 1, p = p.w2, size = 25, lower = FALSE)
pESPN25.w2   <- pbinom(ESPN25 - 1, p = p.w2, size = 25, lower = FALSE)

## Recompute extreme probabilities consistent with w3
p.w3 <- data.w3[8, 3]
pranker10.w3 <- pbinom(ranker10 - 1, p = p.w3, size = 10, lower = FALSE)
pbWAR10.w3   <- pbinom(bWAR10 - 1, p = p.w3, size = 10, lower = FALSE)
pfWAR10.w3   <- pbinom(fWAR10 - 1, p = p.w3, size = 10, lower = FALSE)
pESPN10.w3   <- pbinom(ESPN10 - 1, p = p.w3, size = 10, lower = FALSE)
pranker25.w3 <- pbinom(ranker25 - 1, p = p.w3, size = 25, lower = FALSE)
pbWAR25.w3   <- pbinom(bWAR25 - 1, p = p.w3, size = 25, lower = FALSE)
pfWAR25.w3   <- pbinom(fWAR25 - 1, p = p.w3, size = 25, lower = FALSE)
pESPN25.w3   <- pbinom(ESPN25 - 1, p = p.w3, size = 25, lower = FALSE)

## Recompute extreme probabilities consistent with w4
p.w4 <- data.w4[8, 3]
pranker10.w4 <- pbinom(ranker10 - 1, p = p.w4, size = 10, lower = FALSE)
pbWAR10.w4   <- pbinom(bWAR10 - 1, p = p.w4, size = 10, lower = FALSE)
pfWAR10.w4   <- pbinom(fWAR10 - 1, p = p.w4, size = 10, lower = FALSE)
pESPN10.w4   <- pbinom(ESPN10 - 1, p = p.w4, size = 10, lower = FALSE)
pranker25.w4 <- pbinom(ranker25 - 1, p = p.w4, size = 25, lower = FALSE)
pbWAR25.w4   <- pbinom(bWAR25 - 1, p = p.w4, size = 25, lower = FALSE)
pfWAR25.w4   <- pbinom(fWAR25 - 1, p = p.w4, size = 25, lower = FALSE)
pESPN25.w4   <- pbinom(ESPN25 - 1, p = p.w4, size = 25, lower = FALSE)
```

weight		Ranker	bWAR	fWAR	ESPN
w1	probability of extreme event in top 10 list	8.93×10^{-4}	0.00654	0.00654	0.0335
	probability of extreme event in top 25 list	1.45×10^{-5}	1.45×10^{-5}	0.00161	0.00579
	chance of extreme event in top 10 list	1 in 1120	1 in 153	1 in 153	1 in 30
	chance of extreme event in top 25 list	1 in 6.9039×10^4	1 in 6.9039×10^4	1 in 619	1 in 173
w2	probability of extreme event in top 10 list	0.00172	0.0111	0.0111	0.0504
	probability of extreme event in top 25 list	5.3×10^{-5}	5.3×10^{-5}	0.00406	0.0129
	chance of extreme event in top 10 list	1 in 583	1 in 90	1 in 90	1 in 20
	chance of extreme event in top 25 list	1 in 1.8883×10^4	1 in 1.8883×10^4	1 in 247	1 in 78
w3	probability of extreme event in top 10 list	0.0247	0.0913	0.0913	0.242
	probability of extreme event in top 25 list	0.0085	0.0085	0.119	0.219
	chance of extreme event in top 10 list	1 in 40	1 in 11	1 in 11	1 in 4.1
	chance of extreme event in top 25 list	1 in 118	1 in 118	1 in 8.4	1 in 4.6
w4	probability of extreme event in top 10 list	0.00474	0.0251	0.0251	0.0937
	probability of extreme event in top 25 list	3.84×10^{-4}	3.84×10^{-4}	0.016	0.0416
	chance of extreme event in top 10 list	1 in 211	1 in 40	1 in 40	1 in 10.7
	chance of extreme event in top 25 list	1 in 2602	1 in 2602	1 in 63	1 in 24

Table 5: The probability and chance (1 in 1/probability, rounded) of each extreme event calculation corresponding to the four lists in Table 2 after the MLB eligible population in Table 1 is weighted according to the four conservative weighting regimes.

6 Additional comparison methods

6.1 Versus your peers methods

Several methods are used to compare players across eras by computing a baseline achievement threshold within one season and then comparing players to that baseline. These methods then rank players by how far they stood above their peers; the greatest players were better than their peers by the largest amount. This approach can exhibit major biases in player comparisons, as evidenced by career bWAR and fWAR. Adjusted OPS+ is a worse offender than bWAR or fWAR. Adjusted ERA+ is right in line with ESPN rankings.

6.2 PPS detrending

The methodology of Petersen et al. (2011) (PPS) detrends player statistics by normalizing achievements to seasonal averages, which PPS claim accounts for changes in relative player ability resulting from both exogenous and endogenous factors, such as talent dilution from expansion, equipment and training improvements, and performance enhancing drug usage. However, PPS misunderstand the effect of talent dilution from expansion and ignores reality.

The talent pool was more diluted in the earlier eras of baseball than now because of a small relative eligible population size and the exclusion of entire populations of people on racial grounds. See Table 6 for the specifics. PPS's position with respect to equipment and training improvements is likewise not without fault because the same improvements are equally available to every competitor. Finally, PPS does not account for increases in salary compensation enjoyed by MLB players in modern eras, and their methodology fails to address segregation prior to 1947.

year	eligible pop.	number of teams	roster size	eligible pop. per roster spot
1880	4.44	8	15	37
1890	5.01	8	15	41.7
1900	5.58	8	15	46.5
1910	8.55	16	25	21.4
1920	8.93	16	25	22.3
1930	9.92	16	25	24.8
1940	11.13	16	25	27.8
1950	11.59	16	25	29
1960	19.58	16	25	48.9
1970	26.1	24	25	43.5
1980	36.37	26	25	56
1990	40.68	26	25	62.6
2000	64.06	30	25	85.4
2010	76.19	30	25	101.6

Table 6: Relative talent dilution when considering the MLB eligible population sizes at select time periods. Eligible population totals are in millions in column 2 and are in thousands in column 5.

The mathematics of PPS detrending is also questionable in the context of comparing baseball players across eras. PPS notes that the evolutionary nature of competition results in a non-stationary rate of success. They then detrend player statistics by normalizing achievements to seasonal averages. The normalization goes as follows: Suppose a player hits 40 homeruns in a given season and that the league average prowess for homerun hitting in that season is 10 homeruns. If the historical average prowess for homerun hitting is 5 homeruns then our player's detrended homerun count in that particular season is $40 \times (5/10) = 20$. In general, the detrending formula is $Y \times (\text{historic prowess}/\text{league prowess})$ where Y is individual prowess for a particular player in a given season. We see PPS detrending as an inflationary metric of relative prowesses and not a detrending metric. Fundamentally different approaches for detrending are advocated in authoritative textbooks such as Introduction to Time Series and Forecasting, by Peter J. Brockwell and Richard A. Davis. Table 2 in PPS displays the top 25 career detrended homerun totals. Their top 10 list contains 7 players who started their careers before 1950 (the same as Ranker), and their top 25 list contains 12 players who started their careers before 1950 (the same as fWAR). It is clear that having higher prowess relative to your peers, hitting more runs in this case, is not indicative of a player's prowess with respect to peers from fundamentally different eras.

6.3 Era bridging

Berry et al. (1999) claim that their era bridging technique accounts for talent discrepancies across eras. However, they do not explicitly parameterize this in their hierarchical models. They state that "globalization has been less pronounced in the MLB (relative to other sports)... Baseball has remained fairly stable within the United States, where it has been an important part of the culture for more than a century" (Berry et al., 1999). This rationale ignores segregation, increases in the MLB eligible population relative to available roster spots, and increases in the average overall talent of that population. Therefore, there methodology does not fully address the characteristics of a changing talent pool.

In Berry et al. (1999, panel (c) of Figure 7) we see that their model predicts that a .300 hitter in 1996 will have a lower than .300 average for several seasons from 1900-1920. This conflicts with the well-established notion that the talent of baseball players has improved over time. In Berry et al. (1999, Table 9) we see that 6 of the 10 best hitters for average started their career before 1950 and 10 of the 25 best hitters for batting average started their careers before 1950. Their paper was published in 1999 so we recompute the chances of these events where the MLB eligible population ends at 1999. These calculations are performed with the code below.

```
# population data through 1999
dat2 <- dat[1:13, ]
dat2[13, 2] <- 0.9 * dat[13, 2]
dat2[, 3] <- cumsum(dat2[, 2]) / sum(dat2[, 2])

# proportion of MLB eligible population who started their career
# before 1950
p2 <- dat2[8, 3]

# Binomial probabilities for Berry's top 10 and top 25 lists
bridge10 <- 6; bridge25 <- 10
pbridge10 <- pbinom(bridge10 - 1, p = p2, size = 10, lower = FALSE)
pbridge25 <- pbinom(bridge25 - 1, p = p2, size = 25, lower = FALSE)
```

We calculate the chance that one would expect to observe 6 or more individuals in a top 10 list who started their careers before 1950 as 1 in 38. We calculate the chance that one would expect to observe 10 or more individuals in a top 25 list who started their careers before 1950 as 1 in 9.99. These chances are not as extreme as those in Table 3, but they still correspond to events that are unlikely. However, it is clear that if we were to reweight the MLB eligible population as we did in Section 5, then the calculations in Berry et al. (1999) would be in line with expectations.

The above calculation that appears in the published Chance paper needs some revision. Berry et al. (1999) ran their models on seasonal data collected over the years 1901-1996. We now rerun our analysis over that frame.

```
# population data over 1901-1996
dat2 <- dat[3:13, ]
dat2[1, 2] <- 0.9 * dat2[1, 2]
dat2[11, 2] <- 0.6 * dat2[11, 2]
dat2[, 3] <- cumsum(dat2[, 2]) / sum(dat2[, 2])

# proportion of MLB eligible population who started their career
# before 1950
p2 <- dat2[6, 3]

# Binomial probabilities for Berry's top 10 and top 25 lists
bridge10 <- 6; bridge25 <- 10
pbridge10 <- pbinom(bridge10 - 1, p = p2, size = 10, lower = FALSE)
pbridge25 <- pbinom(bridge25 - 1, p = p2, size = 25, lower = FALSE)
```

We now calculate the chance that one would expect to observe 6 or more individuals in a top 10 list who started their careers before 1950 as 1 in 46. We calculate the chance that one would expect to observe 10 or more individuals in a top 25 list who started their careers before 1950 as 1 in 12.52. These chances are slightly more extreme than those originally calculated above. However, it is clear that if we were to reweight the MLB eligible population as we did in Section 5, then the calculations in Berry et al. (1999) would be in line with expectations.

7 Conclusions

The MLB players from the early eras of baseball receive significant attention and praise as a result of their statistical achievements and their mythical lore. We find that these players are collectively overrepresented in rankings of the greatest players in the history of the MLB, and that popular performance metrics such as WAR fail to properly compare players across eras. Superior statistical accomplishments achieved by players that started their careers before 1950 are a reflection of our inability to properly compare talent across eras. It is highly unlikely that athletes from such a scarcely populated era of available baseball talent could represent top 10 and top 25 lists so abundantly.

As a general discussion of greatness, the conclusions in this article have broader implications than just rankings of athletes. Who are the greatest actors and actresses, artists, musicians, scientists, revolutionaries, or leaders who have ever lived? Do our perceptions change when we focus beyond nostalgia? Do our perceptions change when we recognize and properly account for gender and racial discrimination that has existed throughout human history into the present day?

Acknowledgements

This work would not be what it is today without the many conversations that the author had with family, friends, and peers. I am very grateful to Jim Albert, Peter M. Aronow, James Burrell, Xiaoxuan Cai, R. Dennis Cook, Forrest W. Crawford, Steven A. Culpepper, Andrew Depuy, Evan Eck, Fred Eck, Joe Eck, Kim Eck, Marcus A. Eck, Michael Eck, Phil Eck, Shirley Eck, Wes Eck, Margret Erlendsdottir, Soheil Eshghi, Charles J. Geyer, Jon Hill, Ed Kaplan, Zehong (Richard) Li, Adam Maidman, Aaron Molstad, Olga Morozova, Oliver Om, Kerry Purcell, Ken Ressel, Erick Ruuttila, Jesse Ruuttila, Anne Schuh, Bill Schuh, Yushuf Sharker, Ben Sherwood, Stephanna Szotkowski, Dootika Vats, Brandon Whited, Jarret Zook, the editorial board at Chance, and two anonymous referees at Chance for helpful comments and discussions (some of which went very long). I would also like to think the r/baseball and r/Sabermetrics reddit boards for some useful discussions. This work was partially supported by NIH grants NICHD DP2 HD091799-01.

Appendix: Mathematical Details

We revisit the calculation discussed in the Statistical Evidence section. The chance that a MLB eligible player whose career started before 1950 appears in the top 10 occurs with probability $p = 0.178$ (up to three significant figures). We use the binomial distribution to calculate the probability of at least 6 MLB eligible players whose career started before 1950 appear in the top 10. More formally, we are interested in calculating $P(X \geq 6)$ where X is the number of players whose careers started before 1950 appearing in the top 10. So we have

$$P(X \geq 6) = P(X = 6) + \cdots + P(X = 10).$$

The binomial distribution is used to calculate $P(X = 6)$, …, $P(X = 10)$ where

$$P(X = x) = \binom{10}{x} p^x (1-p)^{10-x}$$

for $x = 6, 7, 8, 9, 10$. For example, when $x = 6$ we have

$$P(X = 6) = \binom{10}{6} p^6 (1-p)^4.$$

Specifying $x = 6$ is equivalent to 6 MLB eligible players whose careers started before 1950 appearing in the top 10. We do not care about the order that players appear in the top 10. The number $\binom{10}{6}$ corrects for this. The number $\binom{10}{6}$ is the number of possible orderings of 6 "successes" and 4 "failures" that can occur. In our context a success is a MLB player whose career began before 1950 and a failure is an MLB player whose career began after 1950. Any particular ordering of the top 10 with 6 MLB players whose careers started before 1950 appearing in the list occurs with probability $p^6(1-p)^4$. Therefore the probability that 6 MLB players who started their careers before 1950 appear in the top 10 is given by

$$P(X = 6) = \binom{10}{6} p^6 (1-p)^4 = 210 \times 0.1779072^6 \times (1 - 0.1779072)^4 = 0.0030413.$$

Similar calculations to this one are made to calculate all of the probabilities that appear in Tables 3 and 5.

References

Berry, S. M., Reese, C. S., and Larkey, P. D. (1999), "Bridging Different Eras in Sports," *Journal of the American Statistical Association*, **94**, 447, 661–676.

Eck, D. J. (2019). "Challenging nostalgia and performance metrics in baseball." Accepted at *Chance*.

Petersen, A. M., Penner, O., Stanley, H. E. (2011), "Methods for detrending success metrics to account for inflationary and deflationary factors," *The European Physical Journal B*, **79**, 67–78.

Schmidt, M. B. and Berri, D. J. (2005), "Concentration of Playing Talent: Evolution in Major League Baseball," *Journal of Sports Economics*, **6**, 412–419.