Where Have All the .400 Hitters Gone?

2016-05-28

Introduction

Going into the last day of the 1941 season, Ted Williams was sitting on a .39955 average and well clear of the 400 at-bats needed to qualify for the major league batting title. By sitting out, he would have been the first major league player to break the .400 barrier since Bill Terry in 1930.Instead, he played both games of a double header going 6 for 8 and finishing with a .406 average, good for 7th best in the modern era. While this made minor news, it would have been a much bigger deal if anyone could have guessed he would be the last player to do so in 74 years and counting.

All-in all, there were twelve .400 + seasons between 1903[^1] and 1941, and not one since:

Table 1: Twelve .400+ Batting Seasons since 1903

Name	Batting Average	Season
Rogers Hornsby	0.424	1924
Ty Cobb	0.420	1911
George Sisler	0.420	1922
Ty Cobb	0.409	1912
Shoeless Joe Jackson	0.408	1911
George Sisler	0.407	1920
Ted Williams	0.406	1941
Harry Heilmann	0.403	1923
Rogers Hornsby	0.403	1925
Ty Cobb	0.401	1922
Rogers Hornsby	0.401	1922
Bill Terry	0.401	1930

Baseball's finest hour has passed! There hasn't been a .400 hitter since 1941, and 9 of the ten top seasons happened before 1925! They just don't make 'em like Harry Heilmann anymore!

Obviously this argument doesn't hold up. Batting average is not a solitary statistic, but one composed of the skill of not only the batter, but also the pitcher and fielders against whom he his playing.

Here is what evolutionary biologist Stephen Gould has to say on the supposed "disappearance" of the .400 hitter:

> The overall batting average has been about .260 throughout the history of baseball. But the variation around that average has shrunk. It's at least plausible that variation declines because play improves. A batting average is a comparison between hitting and pitching. So if everybody's improving, as long as they improve at the same rate, the batting average will remain constant. But it gets to the point where everyone is so good that there's just not much variation anymore. Hitting .400 in baseball is a good example because there's a "right wall," if you will, of human limits. Given how our muscles work, there's just so much that the human body can do. There will always be a few individuals who, by dint of genetic gifts and obsessive commitment and training, will stand close to that right wall. That's where Ty Cobb was in 1911 and where Tony Gwynn is today. But there is this limiting wall. What has happened in baseball is that all aspects of play have improved enormously. Back in 1911, average play was so far inferior to where Ty Cobb was that his batting average could be measured as .420. Today, Tony Gwynn is just as good,

> maybe even closer to the wall than Cobb was. But the average player has improved so much that Gwynn's performance – equal to or better than Cobb's – is not measured as high.

In sum: It should be obvious that player performance has improved over time in all aspects – hitting, pitching, and fielding. The reason we don't see .400 hitters anymore is not that players have gotten worse, or even that pitchers have gotten better faster than hitters. It's that as players collectively approach the right limit of human achivement in all aspects of the game, variation shrinks. Cobb and Hornsby were able to put up such gaudy numbers because they were simply closer to this point before everyone else caught up.

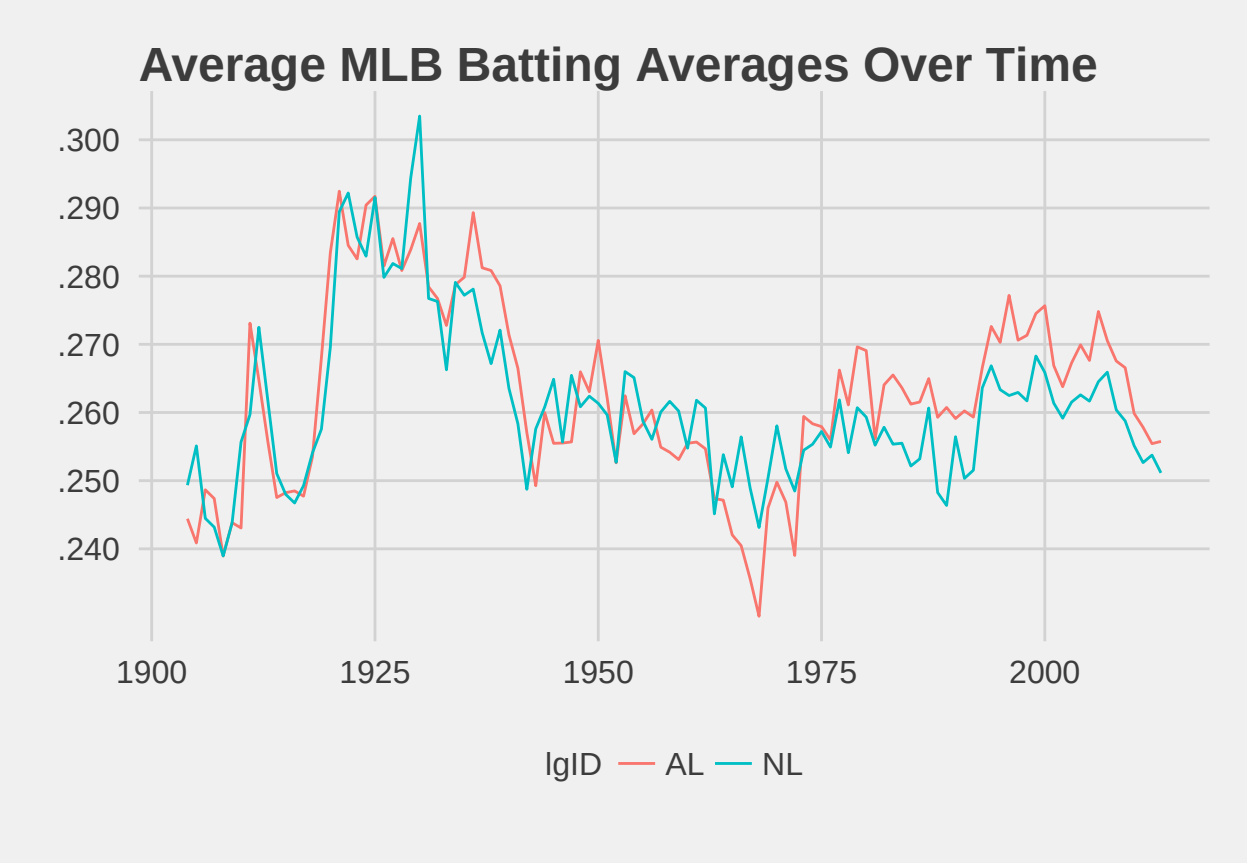
This brings me to my question: If we measure performance ***relative*** to one's time, how does the above table change? Who actually had the best, or perhaps more accurately ***transcendant*** seasons of all time?

Evaluating Gould's Hypothesis

First, I would like to test Gould's central argument:

> "The overall batting average has been about .260 throughout the history of baseball. But the variation around that average has shrunk."

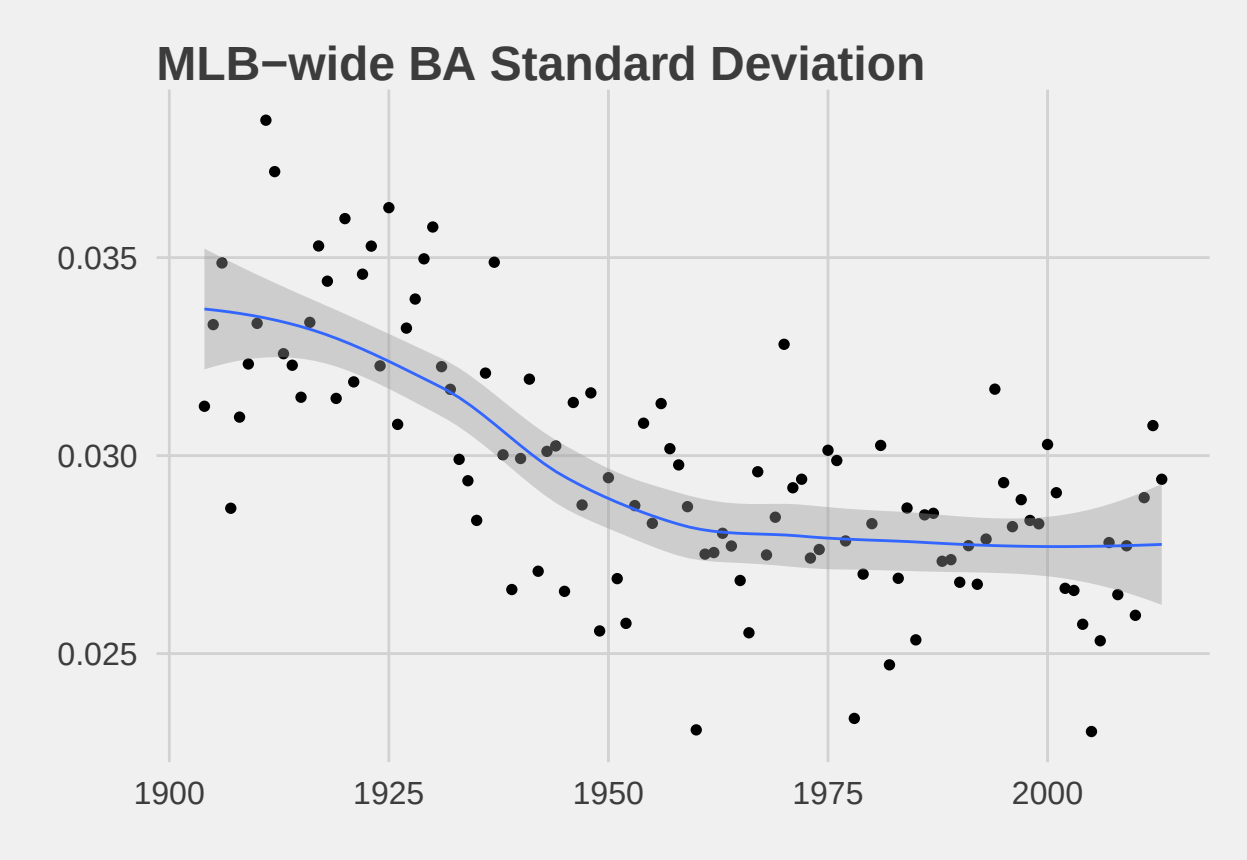
To test the first claim, I simply look at raw league-level batting averages by season.



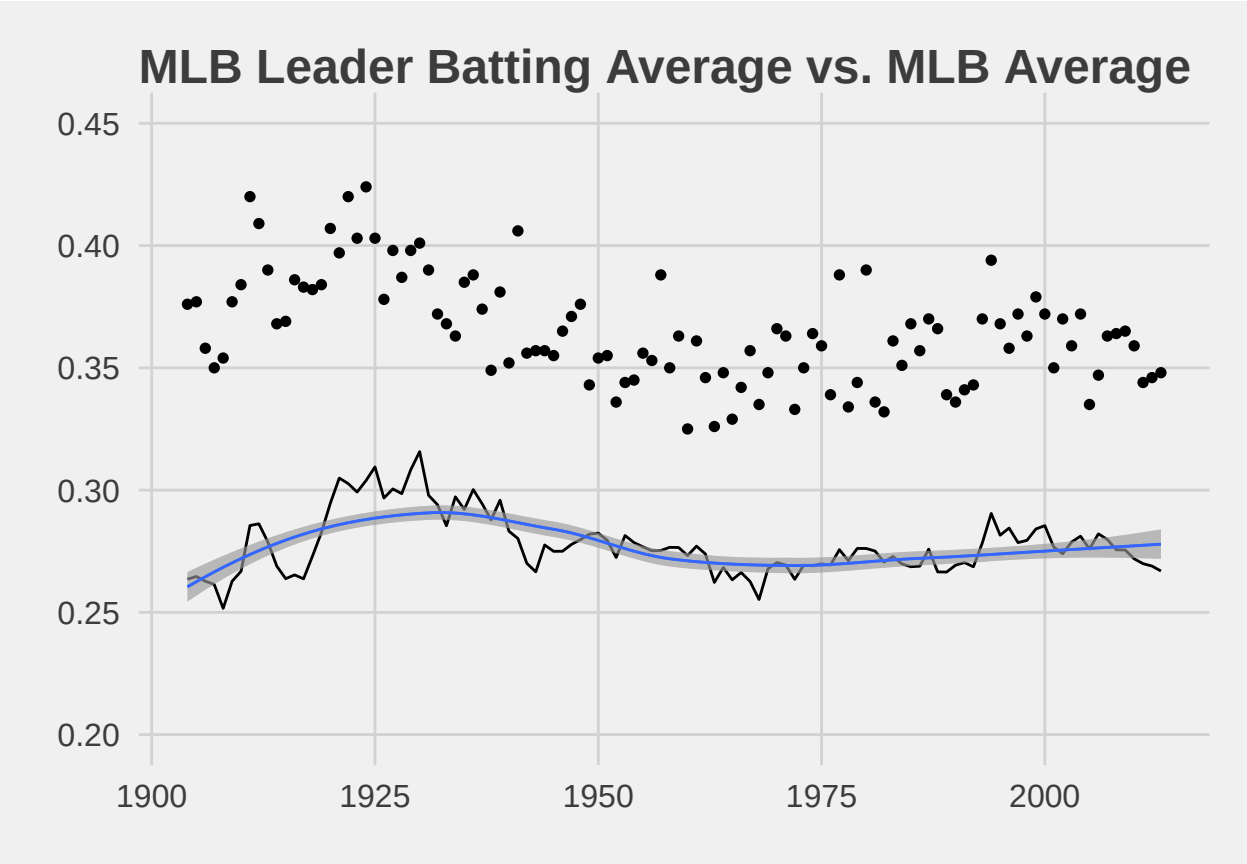
To a first approximation, Gould's claim appears to be a bit of an overstatement, in particular brushing assigned the sustained level of .280+ averaages in the 1920's which saw seven of our twelve post 1903 .400 seasons. Throw in Ty Cobb and Shoeless Joe Jackson's .400s in the 1911 and 1912 season spikes, and Bill Terry's .401 when the NL average peaks at .303, and at least some of their greatness does appear attributable to an increase in averages.

At the same time, Ted Williams hit his .406 when the NL+AL average sat at .262, a fairly pedestrian level historically. This is the first indication that his stature will only improve with further scrutiny.

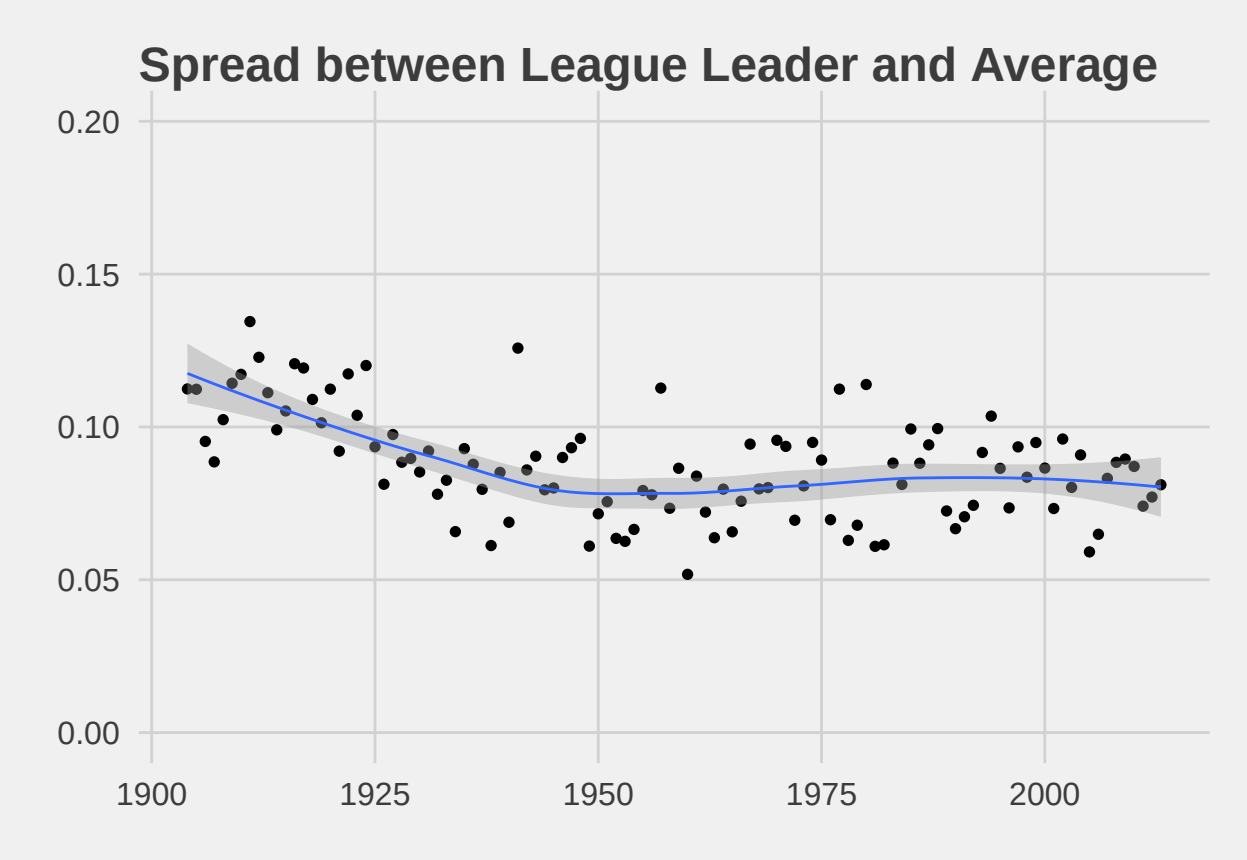
To test Gould's second claim, that variation has been decreasing, I've charted the standard deviation of batting average for players who had at least 400 plate appearances.



Gould appears unambiguously correct here, though the trend seems nearly complete by about 1950. Another way to visualize this is that, if true variance is decreasing, we should see a decrease in the difference between the league batting average leader and the league average over time as well.



Hard to say initially, but it does appear that league leader averages over the 1900-1950 time period fall faster than league wide averages. It's easier to see when we plot the spread (BA leader - league average).



Between 1900 and 1950, the difference between the typical league leader and the league average fell by about 50 points.

Redefining Greatness

Overall, it does appear that our early twentieth century greats (Cobb, Hornsby, Sisler) benefitted from both higher average environments as well as a high level of play variance. But what happens when we adjust for both of these? Who stood the farthest away from their peers, in terms of standard deviation, in a given season.

Table 2: Twelve Greatest Batting Seasons since 1903

Name	Standard Deviations Away	Season	Raw Average
Rod Carew	4.035657	1977	0.388
George Brett	4.027004	1980	0.390
Ted Williams	3.939779	1941	0.406
Wade Boggs	3.918909	1985	0.368
Ted Williams	3.735355	1957	0.388
Rogers Hornsby	3.722085	1924	0.424
Wade Boggs	3.639126	1988	0.366
Tris Speaker	3.617016	1916	0.386
Barry Bonds	3.604509	2002	0.370
Nap Lajoie	3.598441	1904	0.376
Ty Cobb	3.536917	1909	0.377
Ichiro Suzuki	3.529257	2004	0.372

The 400 season is out. The 3.5 standard deviation season is in!

****Biggest Winners****: Rod Carew and George Brett: The only players to break the 4.0 standard deviation barrier in baseball history. Well-known to baseball fans, but certainly not household names on the level of Ted Williams

Ted Williams: Not only is Ted williams right behind Carew and Brett, with his greatest season moving from 7th place to third, his less-known .388 season in 1957 (16 years later!!!) winds up placing 5th on our new list.

Wade Boggs: Wade Boggs ends up cracking the list not once but twice, with relatively pedestrian .366 and .368 averages.

****Biggest Losers****:

George Sisler, Ty Cobb, Rogers Hornsby

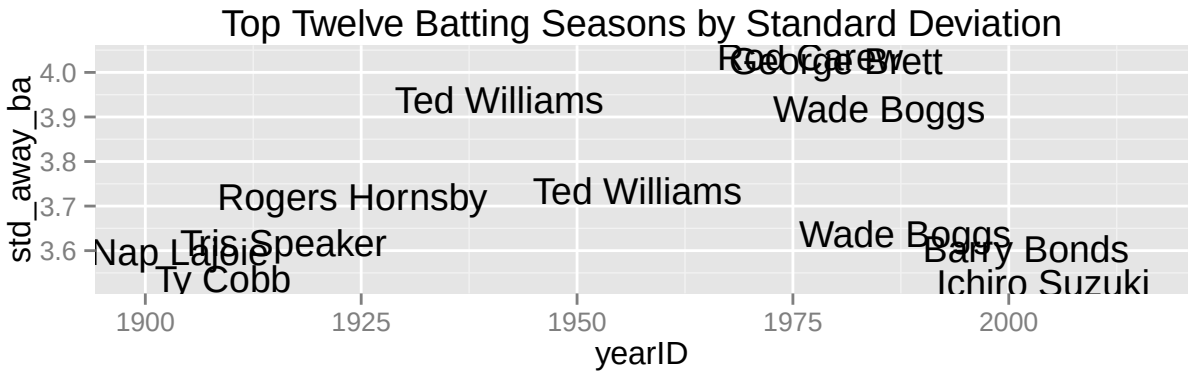
While Cobb and Hornsby still crack the list one time each, they composed 8 of the 12 seasons in our original list. Clearly players who were way ahead of their time, but also in an era in which it was **easier to be way ahead of your time**.

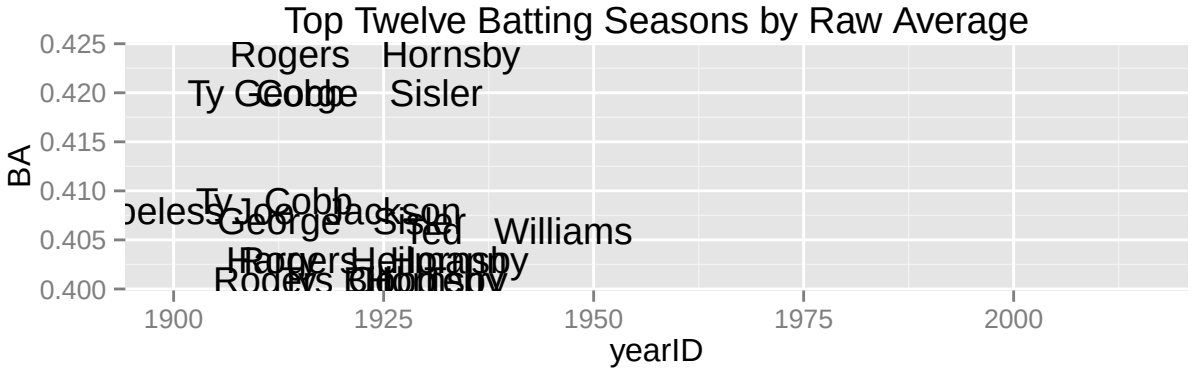
As a final sanity check, and one reason I like this method, compare the time distribution of the top 12 raw batting average seasons vs. the top 12 by standard deviation.

```
std_best <- ggplot(greatest_seasons_redux[1:12,],aes(x = yearID, y = std_away_ba, label = name)) + geom_

raw_best <- ggplot(greatest_seasons[1:12,],aes(x = yearID, y = BA, label = name)) + geom_text() + ggtitl

grid.arrange(std_best,raw_best)
```





By looking at standard deviations, we get a much more even spread of greatness over time,.. This seems more fare, and gives more justice to the fantastic seasons of Brett, Carew, Boggs, Bonds, and Ichiro.

[^1] 1903 is the first year that both the NL and AL adopted the fouls-as strikes rule. Nap Lajoie was able to take advantage of unlimited foul balls in 1901 to hit .426