While it may or may not be true, as Alfred Lord Tenneson penned in his poem [*Locksley Hall*](https://www.poetryfoundation.org/poems/45362/locksley-hall), that “In the Spring a young man's fancy lightly turns to thoughts of love” it is definitely true that “In the Summer a sports fan’s fancy turns heavily to thoughts of baseball”. Long a staple of the warmer portion of the year, America’s pastime is as much an academic pursuit into obscure corners of statistical analysis as it is an athletic endeavor. Each season, thousands upon thousands of die-hards argue about thousands upon thousands of statistics using batting averages, on base percentage, earned run averages, slugging percentages and so on to defend their favorite players or to attack someone else’s favorites. And no testament could stand more clearly for the baseball enthusiasts love of the controversy stirred up by statistics than the movie [61\*](https://en.wikipedia.org/wiki/61*).

For those who aren’t used to seeing ‘special characters’ in their movie titles and who don’t know what significance the asterisk holds a brief explanation is in order. Babe Ruth, one of the best and most beloved baseball players, set many of baseball's records including a single season record of 60 home runs in 1927. As the Babe’s legend grew and the years from 1927 faded into memory without anyone mounting much of a challenge so did the belief that no one could ever break his record. In 1961, two players from the New York Yankees, Mickey Mantle and Roger Maris, hit home runs at a pace (yet another baseball statistic) to make onlookers believe that either of them might break a record dating back to before World War II and the Great Depression. Eventually, Roger Maris broke the record in the 162nd game of the season, but since Ruth reached his mark of 60 home runs when baseball only played 154 games, the Commissioner qualified the Maris record with an ‘\*’ as a means of protecting the mystique of Babe Ruth.

And for years afterwards, fans argued whether Maris really did break the record or not. Eventually, Maris’s record would also fall, first to Mark McGwire, who hit 70 ‘dingers’ (baseball lingo for home runs) in 1998 and then Barry Bonds, who hit 73 in 2001. What’s a baseball fan to do? Should McGwire’s and Bonds receive an ‘\*’? How do we judge?

At first glance, one might argue that the best way to do it would be to normalize for the number of games played. For example, Ruth hit 60 HRs (home runs – an abbreviation that will be used hereafter) in 154 games so his rate was 0.3896. Likewise, Bonds’s rate is 0.4506. And so, can we conclude that Bonds is clearly the better home run hitter. But not so fast the purist will say, Bond’s and McGwire hit during the [steroids era in baseball](https://en.wikipedia.org/wiki/History_of_baseball_in_the_United_States#The_steroid_era), when everyone, from the players on the field to the guy who shouts ‘Beer Here!’ in the stands, was juicing (okay… so maybe not everyone). Should we put stock in the purists argument or has Bonds really done something remarkable?

This is a standard problem in statistics when we try to compare two distinct populations be they be separated in time (1920s, 1960s, 1990s) or geographically (school children in the US v. those in Japan) or even more remotely separated, for example by biology. The standard solution is to use a Z-score, which normalizes the attributes of a member of a population to its population as whole. Once the normalization is done, we can then compare individuals in these different populations to each other.

As a concrete example, let’s compare [Kevin Hart](https://en.wikipedia.org/wiki/Kevin_Hart)’s height to that of Hiawatha the Giraffe. The internet lists Kevin Hart as being 5 feet, 2 inches tall and let us suppose that Hiawatha is 12 feet, 7 inches tall. If we are curious which of them would be able to fit through a door with a standard US height of 6 feet, 8 inches then a simple numerical comparison shows us that Hart will while Hiawatha won’t. However, this is rarely what we want. When we ask which is shorter or taller, it is often the case where we want to know if Hart is short as a man (he is) and if Hiawatha is short as a giraffe (she is as well). So, how do we compare?

Using the Z-score (see also the post [K-Means Data Clustering](http://aristotle2digital.blogwyrm.com/?p=635)), we convert the individual’s absolute height to a population-normalized height by the formula

\[ Z\_{height} = \frac{height - \mu\_{height}}{\sigma\_{height}} \; , \]

where $\mu\_{height}$ is the mean height for a member of the population and $\sigma\_{height}$ is the corresponding standard deviation about that mean.

For the US male, the appropriate statistics are a $\mu\_{height}$ of 5 feet, 9 inches and a $\sigma\_{height}$ of 2.5 inches. Giraffe height statistics are a little hard to come but according to the Denver Zoo, females range between 14 and 16 feet. Assuming this to be a 3-sigma bound, we can reasonably assume a $\mu\_{height}$ of 15 feet with a $\sigma\_{height}$ of 4 inches. A simple calculation then yields a $Z\_{height} = -2.8$ for Hart and $Z\_{height} = -7.25$ showing clearly that Hiawatha is shorter than Kevin Hart, substantially shorter.

Now, returning to baseball, we can follow a very interesting and insightful article by Randy Taylor and Steve Krevisky entitled [*Using Mathematics And Statistics To Analyze Who Are The Great Sluggers In Baseball*](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiOz-6p4vOAAxV-k2oFHb3YB8EQFnoECBkQAQ&url=https%3A%2F%2Fiase-web.org%2Fdocuments%2Fpapers%2Ficots7%2FC432.pdf%3F1402524967&usg=AOvVaw1jZcjitKxN7-gAQ_H9lCEM&opi=89978449)’. They looked at mix of hitters from both the American and National leagues of Major League Baseball, scattered over the time span from 1920-2002, and record the following statistics (which I’ve ordered by date).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Year** | **Hitter** | **League** | **HR** | **Mean HR** | **HR Standard Deviation** | **Z** |
| 1920 | Babe Ruth | AL | 54 | 4.85 | 7.27 | 6.76 |
| 1921 | Babe Ruth | AL | 59 | 6.05 | 8.87 | 5.97 |
| 1922 | Rogers Hornsby | NL | 42 | 6.31 | 7.18 | 4.97 |
| 1927 | Babe Ruth | AL | 60 | 5.42 | 10.05 | 5.43 |
| 1930 | Hack Wilson | NL | 56 | 10.86 | 11.2 | 4.03 |
| 1932 | Jimmie Foxx | AL | 58 | 8.59 | 10.32 | 4.79 |
| 1938 | Hank Greenberg | AL | 58 | 11.6 | 11.88 | 3.91 |
| 1949 | Ralph Kiner | NL | 54 | 10.87 | 9.78 | 4.41 |
| 1954 | Ted Kluszewski | NL | 49 | 14.01 | 11.7 | 2.99 |
| 1956 | Mickey Mantle | AL | 52 | 13.34 | 9.39 | 4.12 |
| 1961 | Roger Maris | AL | 61 | 15.01 | 12.34 | 3.73 |
| 1967 | Carl Yastrzemski Harmon Killebrew | AL | 44 | 11.87 | 8.99 | 3.57 |
| 1990 | Cecil Fielder | NL | 51 | 11.12 | 8.74 | 4.56 |
| 1998 | Mark McGwire | NL | 70 | 15.44 | 12.48 | 4.37 |
| 2001 | Barry Bonds | NL | 73 | 18.03 | 13.37 | 4.11 |
| 2002 | Alex Rodriguez | NL | 57 | 15.83 | 10.51 | 3.92 |

Interestingly, the best HR season, in terms of Z-score, is Babe Ruth’s 1920 season where he ‘only’ hit 54 HRs. That year, he stood over 6.5 standard deviations from the mean making it far more remarkable than Barry Bonds 73-HR season. Sadly, Roger Maris’s 61\*-HR season is one of the lower HR championships on the list.

To be clear, it is unlikely that any true, dyed-in-the-wool, baseball fan will be swayed by statistical arguments. Afterall, where is the passion and grit in submitting an opinion to the likes of cold, impartial logic. Nonetheless, the Z-score is not only in its own right but serves also as a launching pad for more sophisticated statistical measures.