Official Baseball Statistics

December, 2024 by David W. Smith

History and Available Data

Professional baseball has a long and rich relationship to numbers. New ways to analyze the game with increasingly sophisticated metrics have greatly expanded in the computer age. Retrosheet is a historically focused organization which captures the events of games from earlier times and makes them available for examination by modern methods. This historical work requires careful validation in order to maximize the usefulness of the data. To put it most simply: the records for each player in each game are compared in detail for all official categories. Any differences between Retrosheet data and the official records are vigorously pursued. However, these detailed comparisons require the availability of reliable statistics. The further back in history the work of Retrosheet goes, the more we encounter issues with the official records that must be addressed.

There are many subtleties in this area, beginning with definitions. The most important is the meaning of "official". It literally means "of the office" (https://www.merriam-webster.com/dictionary/official).

It is often misused as a statement of accuracy or truth, but it is actually a statement of authority. This distinction is quite significant, as pointed out by the great sportswriter Leonard Koppett, recipient of the Spink Award from the Baseball Writers' Association of America (BBWAA) in 1992. Koppett wisely pointed out that without this authoritative endorsement there would be "dueling statistics" that would create confusion and undermine the credibility of the records. Therefore, "Official records" are those endorsed by Major League Baseball (MLB) as a corporate and administrative entity. Official records are occasionally changed when compelling evidence for new values becomes available, but the process to do so is intentionally very slow for just the reasons Koppett pointed out. Perhaps the most famous example of such a change is Hack Wilson's RBI total from 1930. It is now officially 191 after decades of being listed as 190. There have been many other adjustments to official records, including some major ones. The fluidity of official records was addressed nicely by Anthony Castrovince on the MLB site in the spring of 2024: https://www.mlb.com/news/baseball-record-books-changing-negro-leagues.

A brief summary of the way the records were created may be helpful. Before real-time electronic data became available late in the 20th century, the process was done entirely manually. The basic steps for most seasons were:

1. The Official Scorer of a game completed a form with the game's date and values for each batter and pitcher for all official categories. The number of these categories changed over the years, but typically there were 22 for batters and 17 for pitchers with some comments.
2. The reports were sent to the league office, on a weekly or daily basis.
3. The data from the reports were copied onto pages in large ledger books of approximately 18 inches by 30 inches. The ledgers were organized by team and the players alphabetically. The pages typically had 45 to 50 lines, one per game, so most players had multiple pages.

4. Categories were totaled by hand and various average calculated at the end of the season.

These massive ledgers are now housed in the library at the National Baseball Hall of Fame. Several years ago, the ledgers were microfilmed and Retrosheet was fortunate enough to obtain a full set of the dozens of reels. The film was scanned into pdf images, one per page, and then volunteers spent an enormous number of hours manually entering the data from the scans into spreadsheets. These files are the basis of the daily comparisons mentioned above.

The pdf scans are now available on the Retrosheet website. There are significant variations in the format of the pages from different years. Here is a detailed description of the structure of this archive.

The first Major League season was either 1871 or 1876, depending on whether the National Association (1871-1875) is included. The official MLB position is that the National Association was not a Major League. Unfortunately not all seasons have daily records and there are three distinct types of microfilmed pages.

1. Most common is the handwritten page as described above.
2. Computer printouts:
    a. AL beginning in 1973, NL in 1981
    b. Early seasons compiled by Information Concepts Incorporated (ICI)

The ICI pages require special comment. They were created for the first MacMillan Baseball Encyclopedia (Big Mac) in 1969 to fill gaps in existing records. These gaps included:

1. Federal League 1914-1915
2. American Association 1888-1891
3. National League 1891-1902
4. American League 1901-1904

There were no official records for these league-seasons. The ICI team scoured newspapers from many cities to construct game by game totals for the players on those teams. This Herculean effort provided totals otherwise not available, although it did not cover seasons prior to 1888, meaning several seasons of the American Association and National League as well as the Players League (1890) and Union Association (1884) are not represented in the microfilmed record.

One additional variable is which information is presented in each season. Individual batters are always present and individual pitchers are almost always in a separate file. For many seasons, there are daily team total files for batting, occasionally for pitching and much less often for fielding. In some seasons, batting and pitching data are intermingled. These differences are reflected in the file names which always contain our three letter team abbreviation. Here are some examples with the first being the most common.

1.

1955BROBat.pdf
1955BROPit.pdf
1955BROTeamBat.pdf

2.

1978BALBat.pdf
1978BALPit.pdf
1978BALTeamBat.pdf
1978BALTeamPit.pdf
1978BALTeamFld.pdf

3.

1920NY1BatPit.pdf. Batter and pitcher data intermingled.
1920NY1TeamBat.pdf

Within a given league-season, the naming convention is consistent.

Official Categories

A complicating factor is that the list of official categories changed over the years and sometimes varied between the AL and NL. For example, the 1903 NL files, which are the oldest of the handwritten pages, have 12 columns of numerical data for each batter for each game plus the date, opponent, position played and possible comments. Included were fielding data: putouts, assists and errors, but not double plays. There were eight categories for pitchers with wins and losses appearing in the comments.

On the other hand, the 1980 NL files, the last of the handwritten pages, have 22 columns of numerical data for each batter, including putouts, assists, errors, double plays and passed balls. There are also nine additional columns for fielding data of up to four positions! Needless to say, these are most often left blank. Separate columns were provided to indicate day or night game and home or away. Pitchers have 23 columns of numerical data plus separate ones for day/night and home/away. There are four more pitcher columns identifying the opposing pitcher with a decision, the score of the game, who relieved him (if any) and whom he relieved (if any).

Each of the categories added between 1903 and 1980 has its own story. There is a summary of many of these in an appendix to *Total Baseball* (Palmer and Thorn). The most recent summary was created by Dennis Bingham and Thomas Heitz, covering 1920-1992. Here are some details.

Caught Stealing

Although stolen bases were recorded from the earliest days, caught stealing appeared as an official category much later. The AL began recording them in 1918 and the NL in 1921. However the NL discontinued the inclusion of caught stealing from 1926 through 1950, reinstating them in 1951. The official definition of caught stealing was modified in 1979 so that it now includes pickoff plays in which the runner begins to advance to the next base and is out returning to the base from which he started. Prior to this, these plays were not recorded as caught stealing.

Grounding into Double Play

AL: First appeared in 1939 under the "Avg" column heading which had little sense on a daily basis. The until the 1973 computerization with the heading "GIDP" spread over two lines

NL: Records began in 1933 with handwritten change for two years to column heading of HDP or "Hit Into Double Play". In 1935, the printed form had a column headed "Hit into Double Play ".

RBI

Runs batted in entered the scoring rules in 1920. For several years the AL recorded these under the heading of RRF(Runs Responsible For) although the NL sheets had "Runs Batted In" printed on them. Variations in definition occurred. For example, in 1939 the provision was added that no RBI was to be awarded to a batter who grounded into a double play.

Strikeouts

Strikeouts were recorded for pitchers in the NL from the earliest records. However, strikeouts by batters in the NL did not appear in the official pages until 1910. In the AL there were no pages for pitchers at all until 1908 although batting data first appeared in 1905. Strikeouts for AL batters were not reported until 1913.

Batters faced by Pitchers (BFP)

This category is potentially confusing because the AL and NL differed significantly in what they reported, they changed over the years, and the headings on the columns of the official sheets were inconsistent. The two leagues differed as follows:

NL: Before 1915, the official sheets reported at bats against pitchers, not batters faced. Beginning in 1915, this was changed to BFP, but the column was still labeled "AB". It then changed to "Total AB" but was actually BFP.

AL: Until 1980, the AL reported at bats against, not BFP. They began recording BFP in 1980,

Earned Runs

This category was created by rule for the 1913 season in both AL and NL

Intentional Walks

Although many walks had been described as intentional in news accounts for decades, the rule creating them as an official category did not go into effect until 1954.

Wins and Losses for Pitchers

Perhaps surprisingly, there were no provisions in the rules for determining winning and losing pitchers until 1950. There was a general consensus among official scorers before that year although they had discretion to award these as they saw fit.

Saves

The save did not become an official category until 1969 and it has had four different formal definitions. The last of these came in 1975 and has continued in force to the present. The question for Retrosheet records is how to treat games played prior to 1969 since there is legitimate interest in extending this category backwards in time before the establishment of the rule. The answer has been to award a save to pitchers who a: did not start, but finished the game and b: did not receive credit for the win. This approach, which is decidedly unofficial, allows a consistent approach to games lacking play by play detail. As complete play by play is created for more games from over 100 years ago, this decision may be altered and different criteria applied.

Sacrifice Fly

This official category has perhaps the most complicated history as it has had several different versions since its inception. Here is a summary, crafted by Bill Deane.

- The SF Rule was in effect for 1908-30, 1939 and 1954 to date.
- 1908: Rule implemented, specifically requiring a runner to score with less than 2 outs.
- 1909: SF credited even if an error is made.
- 1920: SH and SF not distinguished in reporting.
- 1908 was the only year where the guide separated them,
- but published box scores noted them separately though 1919.
- 1926: Award SF for any advance, not just for scoring.
- 1931-1938: SF removed as official category.
- 1939: Rule from 1909-1925 reinstated.
- 1940-1953: SF again removed as official category.
- 1954: SF reinstated again, but only for fair balls.
- 1958: Rule advisory to credit SF even if a runner is forced on the play.
- 1961: Rule no longer required fair ball.

Throughout all years in all leagues, the batting order position of a player has never been part of the official record. Therefore, Retrosheet and other historical organizations must rely on published box scores in news sources for this information.