January, 2012

By David W. Smith

Retrosheet game files (not game logs) and Retrosheet discrepancy files

As noted on the Retrosheet home page:
(http://www.retrosheet.org/DiscrepanciesWithOfficialData), we began posting discrepancies between our data and the official Major League totals in the fall of 2011. These discrepancy files are now available for download on this page along with the game files described below.

There are three kinds of game files with Retrosheet data:

1. Full play by play with file names of the form YYYYTTT.EVL, where YYYY is the 4 digit year, TTT is the Retrosheet team abbreviation and L is the one letter league abbreviation. These are referred to as event files and for shorthand are designated as evx files. Accounts come from scorebooks and scorecards. Each file has the home games for one team for one season. A more detailed description of this file format may be found at: http://www.retrosheet.org/eventfile.htm.
2. Files with batting, pitching and fielding totals for each player used to create box scores on the Retrosheet site. The file names are: YYYY.EBL. These are referred to as box score event files and their shorthand notation is ebx files. Lineup information comes from newspaper box scores and player data from the official daily totals. Each file has data for one league for one season. Detailed description of the file format is available at: Box Score Event Files
3. Files with play by play accounts that have been deduced by reference to newspaper game stories and official totals. The file names are YYYYLL.EDL. These are referred to deduced event files and are identified as edx files. The format of the data is the same as the full event files, but the fielding credit for most plays is missing.

Details of deduced files and their use

The deduced files first appeared on the Retrosheet website in July of 2011. They were created to fill the gap of missing games in seasons for which we have full play by play for the majority, usually the large majority, of games. It has been frustrating to be limited in analysis by this small group of missing games. The box score event file format was created to address this issue. However, there are many basic questions, such as data for pitcher-batter matchups, that requires some form of play by play data and therefore the effort was launched to deduce the best possible account from the information that is available.

How is the deduction done? The primary sources for each game are the stories printed in newspapers from the cities of the two teams involved. In many cities, such as New York, Chicago, St. Louis, Philadelphia, Pittsburgh and Detroit there were two or even more different newspapers covering each game. Therefore, we almost always have two stories for a game and sometimes four or more. There is always redundancy in these accounts for the major events of the game, but there is also surprising variation of what each writer addressed with the result that combining all of the stories gives a very solid understanding of the major action of the game.

We are always more sure of some plays than others, but the bedrock offensive and pitching categories can generally be determined with a high degree of confidence. Assigning each base hit to the proper inning and the proper opposing pitcher is an important starting point. Walks and strikeouts are somewhat more difficult to match, but we do have the advantage of the "double entry" system of baseball in that a walk garnered by a batter must be charged to a pitcher so having multiple pitchers in a game makes the assignments easier. The official daily totals are essential for this analysis and we are fortunate to have that data from the microfilm we purchased from the Hall of Fame several years ago. The most prominent missing information relates to fielding credits and these accounts always have dozens of "unknown outs" where we are sure there was an out that was not a strikeout or a double play but we don't know anything else about the play.

How reliable are the resultant deduced accounts? After the deduction process is complete, each game is entered into our standard event file format and the totals are then checked against the official totals just as the full play by play accounts are checked. This is done with software I wrote for this purpose several years ago. Of course, having the daily game totals match the official data does not mean that we have assigned the events to the correct inning. Based on several hundred of these deductions that have been completed, I estimate that the assignment of hits is about 90% accurate, with walks and strikeouts probably in the 70% range. I like to refer to these accounts as "plausible."

It is important that users know unambiguously which games have been deduced and which are full play by play accounts. Therefore, the deduced games are in separate files with unique names. These are bundled with the eva and evn files for that year, but the user can then make a informed decision about including them or not in any analysis.

Where will games be found? The year 1950 is a major dividing line for the organization of our files. All games played before 1950 are in ebx files, even if they are also in evx files. Games played in 1950 or more recently follow one of the following four patterns:

1. In evx file only because there is a complete play by play account.
2. In evx <u>and</u> ebx
    a. because the play by play came from a source with "generic outs", meaning no fielding credit. However, the placement of hits, walks, etc is clear.
    b. In evx and ebx because the game was played before 1950 and all games are in ebx.
3. Only in ebx file because there is no play by play account of any type
4. In edx <u>and</u> ebx file.

No game will ever be in an edx file alone.

The creation of the deduced accounts is of great significance and allows users to complete analyses for full seasons that were not possible before. Of course, these accounts are always open to review, just as the regular event files are. I am confident that corrections will be made on a routine basis, which is one of the great benefits of the careful reading our site gets from so many of those who use Retrosheet data.