A POISSON MODEL FOR "HITTING FOR THE CYCLE" IN MAJOR LEAGUE BASEBALL

by

Paul M. Sommers
Christopher J. Teves
Joseph T. Burchenal
And
Daniel M. Haluska

March 2010

MIDDLEBURY COLLEGE ECONOMICS DISCUSSION PAPER NO. 10-11

DEPARTMENT OF ECONOMICS
MIDDLEBURY COLLEGE
MIDDLEBURY, VERMONT 05753

http://www.middlebury.edu/~econ

A POISSON MODEL FOR "HITTING FOR THE CYCLE" IN MAJOR LEAGUE BASEBALL

by

Christopher J. Teves
Joseph T. Burchenal
Daniel M. Haluska
Paul M. Sommers

Department of Economics
Middlebury College
Middlebury, Vermont 05753
psommers@middlebury.edu

A POISSON MODEL FOR "HITTING FOR THE CYCLE" IN MAJOR LEAGUE BASEBALL

In a recent article in this journal [1], Campbell *et al.* showed that the Poisson probability distribution provides an excellent fit to the data on no-hit games in Major League Baseball, especially during the period 1920-1959.

Hitting for the cycle (that is, when a batter hits a single, double, triple and home run in the same game) is another rare event in Major League Baseball.[1,2] And, here too, the Poisson probability distribution given by

$$p(X = x) = \frac{e^{-\mu}\mu^{x}}{x!}, \quad x = 0, 1, 2, \ldots$$

where x denotes the number of ballplayers who "hit for the cycle" (hereafter a "cycle") in a given season provides a remarkably good fit.

The number of cycles by league from 1901 through 2007 are shown in Table 1 (see http://mlb.mlb.com/mlb/history/rare_feats/index.jsp?feature=hit_for_cycle). Bob Meusel of the New York Yankees hit for the cycle three times in his career (1921, 1922, and 1928). So too did Babe Herman, twice in 1931 as a Brooklyn Dodger and once in 1933 as a Chicago Cub. George Brett's two career cycles (1979 and 1990) were the farthest apart (eleven years and 57 days). In eighteen different seasons between 1901 and 2007 (nine times before 1920), not one player hit for the cycle. The average number of cycles per season in both leagues combined jumped from 1.16 between 1901 and 1919 (during the so-called "deadball era") to 2.17 between 1920 and 1960 (the year before the American League and two years before the National League expanded their season to 162 games) [the p-value on the difference between the two means is .022]. Not surprisingly, the average number of cycles per season increased from 2.17 between 1920 and 1960 (154-game seasons) to 2.66 between 1961 and 2007 (162-game seasons), but this difference between the two respective means is not statistically discernible (p = .174).[3]

In Table 2, a Poisson model is shown to be an appropriate model for the 103 (111) times a player hit for the cycle in the American (National) League over the 88-year period 1920 through 2007. The expected value of the Poisson distribution is estimated by:

$$\bar{x} = \frac{\sum x_i O_i}{n}$$

where n = 88 ($\bar{x}_{AL}$ = 1.17045 and $\bar{x}_{NL}$ = 1.26136).

Figure 1 shows how well the observed frequency distribution in each league (1920-2007) conforms to the Poisson distribution. The conformity of the observed (O_i) and expected (E_i) frequencies can be measured using a chi-squared goodness-of-fit test where

$$\chi^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i}$$

The test statistic is 3.86 for the American League and 3.03 for the National League. Since the critical value is $\chi^2_{.05,3} = 7.815$ (and hence exceeds the value of the test statistic in each case), it follows that the observed or actual distribution can be represented by the theoretical (Poisson) distribution.[4] When the results for the two leagues are combined, x_i (the number of cycles per season) ranges from "0" to a high of "8" (in 1933, with four cycles that year in each league and three of the four in the American League on one team – the Philadelphia Athletics). The test statistic is now $\chi^2 = 8.51$ and $\chi^2_{.05,6} = 12.59$, meaning that the observed and expected frequencies are still quite compatible.

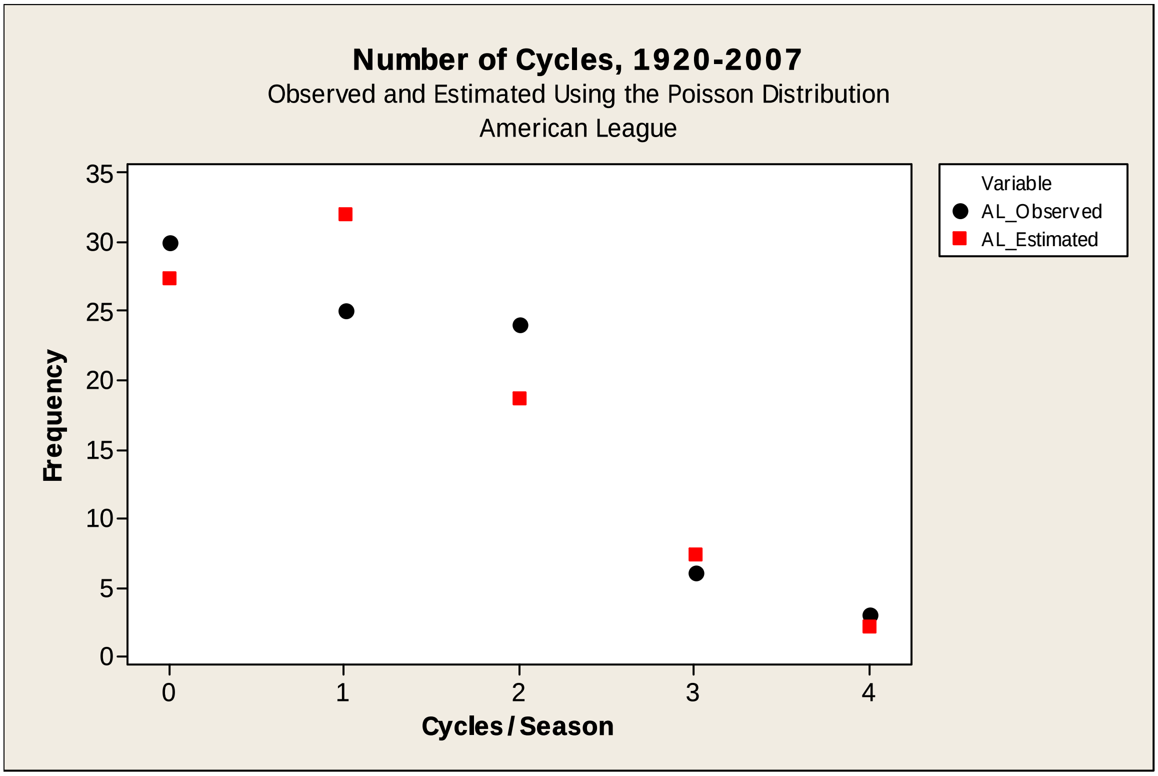
Table 1. Hitting for the Cycle,
Number of Cycles by League, 1901-2007

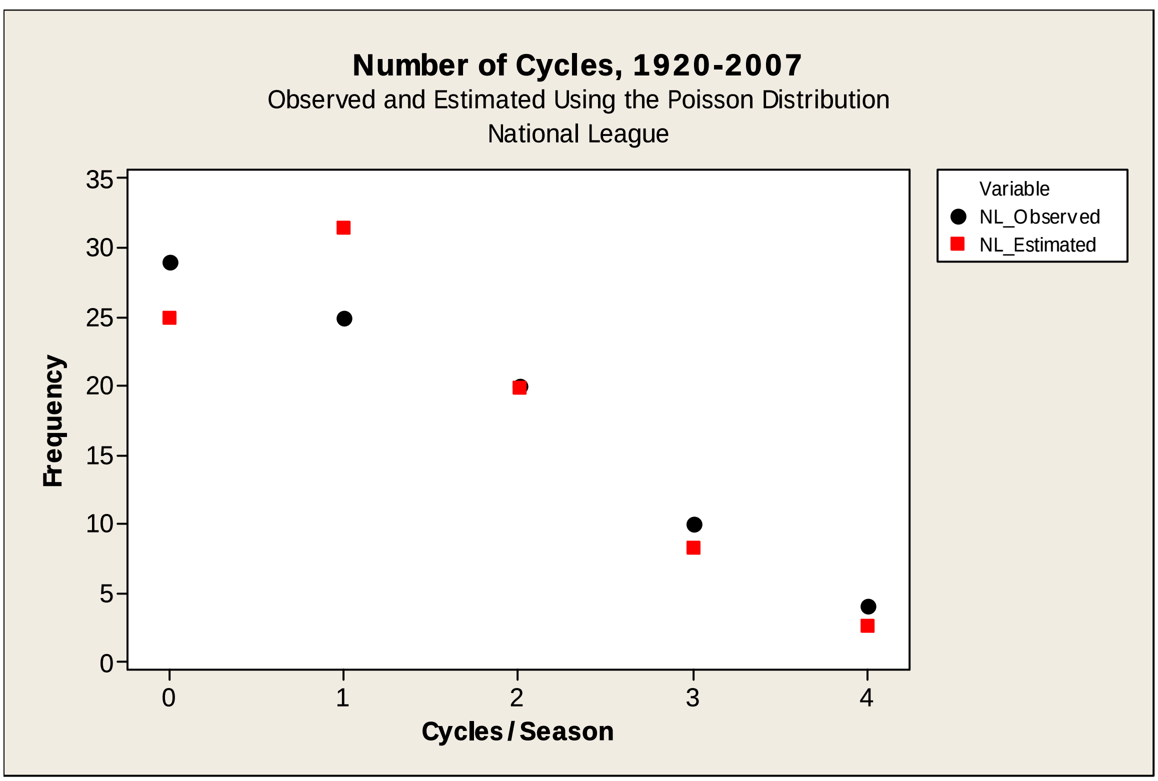
Year	Number of Cycles AL	Number of Cycles NL	Year	Number of Cycles AL	Number of Cycles NL	Year	Number of Cycles AL	Number of Cycles NL
1901	2	1	1937	3	0	1973	0	1
1902	0	0	1938	1	0	1974	0	1
1903	2	1	1939	2	1	1975	0	1
1904	0	2	1940	3	3	1976	2	4
1905	0	0	1941	1	0	1977	2	1
1906	0	0	1942	0	0	1978	2	1
1907	0	1	1943	1	0	1979	4	0
1908	1	0	1944	2	1	1980	2	3
1909	0	0	1945	0	2	1981	0	0
1910	1	2	1946	2	0	1982	1	0
1911	1	1	1947	2	0	1983	0	0
1912	2	3	1948	1	2	1984	3	1
1913	0	0	1949	0	3	1985	2	2
1914	0	0	1950	3	2	1986	2	0
1915	0	1	1951	0	1	1987	0	4
1916	0	0	1952	1	0	1988	1	2
1917	0	0	1953	0	0	1989	1	3
1918	0	1	1954	0	1	1990	1	0
1919	0	0	1955	0	0	1991	1	3
1920	2	1	1956	0	0	1992	0	1
1921	2	2	1957	1	1	1993	2	1
1922	2	2	1958	0	0	1994	1	0
1923	0	1	1959	0	1	1995	1	2
1924	2	0	1960	1	1	1996	1	2
1925	1	2	1961	0	1	1997	1	1
1926	1	0	1962	1	0	1998	1	2
1927	0	2	1963	0	2	1999	1	2
1928	1	1	1964	2	2	2000	2	3
1929	2	1	1965	1	0	2001	4	1
1930	0	3	1966	0	2	2002	0	2
1931	0	3	1967	0	0	2003	2	2
1932	2	0	1968	1	0	2004	2	4
1933	4	4	1969	0	0	2005	0	3
1934	3	0	1970	2	3	2006	3	2
1935	0	1	1971	1	0	2007	2	1
1936	0	1	1972	2	2			

Table 2. The Number of Cycles
by League, 1920-2007

	Number of cycles (x_i)	Observed number of seasons (O_i)	Poisson probability (p_i)	Expected number of seasons ($E_i = 88 \times p_i$)
American League				
	0	30	.3102	27.2999
	1	25	.3631	31.9533
	2	24	.2124	18.6999
	3	6	.0829	7.2958
	4	3	.0243	2.1349
National League				
	0	29	.2833	24.9275
	1	25	.3573	31.4427
	2	20	.2253	19.8303
	3	10	.0947	8.3378
	4	4	.0298	2.6292

Figure 1


Number of Cycles, 1920-2007
Observed and Estimated Using the Poisson Distribution
American League
Variable
AL_Observed
AL_Estimated
Frequency
0
5
10
15
20
25
30
35
0
1
2
3
4
Cycles / Season


Number of Cycles, 1920-2007
Observed and Estimated Using the Poisson Distribution
National League
Variable
NL_Observed
NL_Estimated
Frequency
0
5
10
15
20
25
30
35
0
1
2
3
4
Cycles / Season

Reference

1. D.L. Campbell, B.O. Hanna, C.A. Lyons, and P.M. Sommers, "A Poisson Model for No-Hitters in Major League Baseball," *Journal of Recreational Mathematics*, Vol. 34(1), pp. 49-52.

Footnotes

1. We assume that the number of cycles per season is independent of the number of cycles in any other season and that the probability of hitting for the cycle is constant from game to game throughout the season.

2. Between 1901 and 2007, fourteen different players hit "natural cycles" – that is, they hit a single, double, triple, and a home run, in that order. No player accomplished this rare feat more than once.

3. A 162-game season is only 5.2 percent longer than a 154-game season. Yet, between the two periods — 1920 and 1960 (154-game seasons) and 1961 and 2007 (162-game seasons) — the average number of cycles per season increased from 2.17 to 2.66, a 22.6 percent increase.

4. The number of degrees of freedom is one less than the number of values of $(O_i - E_i)^2/E_i$ that are summed up; that is, $5 - 1 = 4$. But, it is important to note that χ^2 would have one less degree of freedom (3, not 4), because we estimated an additional parameter of the theoretical distribution (namely, μ) from the actual distribution.