A Poisson Model For No-Hitters In Major League Baseball

by

Paul M. Sommers
David L. Campbell
Benjamin O. Hanna
Conor A. Lyons

September 2007

MIDDLEBURY COLLEGE ECONOMICS DISCUSSION PAPER NO. 07-17

DEPARTMENT OF ECONOMICS
MIDDLEBURY COLLEGE
MIDDLEBURY, VERMONT 05753

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

A POISSON MODEL FOR NO-HITTERS IN MAJOR LEAGUE BASEBALL

by

David L. Campbell
Benjamin O. Hanna
Conor A. Lyons
Paul M. Sommers

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

A POISSON MODEL FOR NO-HITTERS IN MAJOR LEAGUE BASEBALL

No-hit games are relatively rare events in Major League Baseball. Table 1 shows the number of (minimum nine-inning) no-hit games by a single pitcher in the major leagues each year from 1920 through 2006 (see http://sports.espn.go.com/mlbhist/alltime/nohitters).[1] The probability of a no-hit game is small and here assumed to be the same throughout a season. Moreover, the number of no-hit games occurring in a given season is assumed independent of the number of no-hit games in any other season. Under these assumptions, is the distribution of the annual number of no-hit games roughly Poisson?

The Poisson distribution is a discrete probability distribution which has the following formula:

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

where μ is the mean number of occurrences or expected value of the Poisson distribution.

In Table 2, a Poisson distribution is shown to be an appropriate model for the forty-nine no-hit games in the major leagues over the 40-year period 1920 through 1959. The expected value of the Poisson distribution (μ) is estimated by:

$$\bar{x} = \frac{\sum x_i O_i}{n} = 1.225$$

where n = 40.

How well does the observed frequency distribution conform to the Poisson distribution? The null hypothesis is that the observed or actual distribution can in fact be represented by the theoretical (Poisson) distribution and that the discrepancies between them are due to chance.

The value of the test statistic is

$$\chi^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} = .131$$

Since $\chi^2_{.05,3}$ = 7.815 (and exceeds the value of the test statistic), it follows that there is no reason to reject the null hypothesis.[2] In other words, the probability is greater than .05 that the observed discrepancies between the actual distribution and the Poisson distribution could be due to chance.

Baseball expanded from sixteen teams in 1960 to twenty-four teams by the end of the decade, thereby diluting the talent pool and making it arguably easier for dominant pitchers of the expansion era to throw no-hitters. Another look at Table 1 indeed reveals a jump in the number of no-hit games (minimum 9 innings by a single pitcher) during the 1960s (34 in the 1960s, 29 each in the '70s and '90s, 18 in the '50s, 13 each in the '40s and '80s, and 9 each in the '20s and '30s), with four no-hit gems recorded in the sixties by Dodgers pitcher Sandy Koufax alone. When Table 2 is revised for the extended period 1920-1969, $\bar{x}$ increases to 1.66 (where n = 50) and the value of the test statistic, χ^2, becomes 4.927. Since $\chi^2_{.05,5} = 11.07$, the theoretical (Poisson) distribution still provides a very good approximation to the empirical relative frequency distribution.[3]

When the period is further extended from 1920-1969 to 1920-2006, the value of the test statistic rises to 6.731, but is still less than the critical value of 7.815.[4]

We can conclude that the Poisson model provides an excellent fit to the data on no-hit games in Major League Baseball during the period 1920-1959, and a weaker but still surprisingly good fit for the longer periods 1920-1969 and 1920-2006.

Table 1. No-Hit Games (Minimum 9 Innings) by a Single Pitcher, 1920-2006

Year	Number of no-hit games pitched	Year	Number of no-hit games pitched	Year	Number of no-hit games pitched
1920	1	1949	0	1978	2
1921	0	1950	1	1979	1
1922	2	1951	4	1980	1
1923	2	1952	3	1981	3
1924	1	1953	1	1982	0
1925	1	1954	1	1983	3
1926	1	1955	1	1984	2
1927	0	1956	3	1985	0
1928	0	1957	1	1986	2
1929	1	1958	2	1987	1
1930	0	1959	1	1988	1
1931	2	1960	3	1989	0
1932	0	1961	1	1990	7
1933	0	1962	5	1991	5
1934	2	1963	3	1992	2
1935	1	1964	3	1993	3
1936	0	1965	4	1994	3
1937	1	1966	1	1995	1
1938	3	1967	3	1996	3
1939	0	1968	5	1997	1
1940	2	1969	6	1998	1
1941	1	1970	4	1999	3
1942	0	1971	3	2000	0
1943	0	1972	3	2001	3
1944	2	1973	5	2002	1
1945	1	1974	3	2003	1
1946	2	1975	2	2004	1
1947	3	1976	3	2005	0
1948	2	1977	3	2006	1

Table 2. No-Hitters in Major League Baseball, 1920-1959

Number of no-hit games	Observed number of seasons	Poisson probability	Expected number of seasons
(x_i)	(O_i)	(p_i)	($E_i = 40 \cdot p_i$)
0	11	.2938	11.7503
1	15	.3599	14.3941
2	9	.2204	8.8164
3	4	.0900	3.6000
4	1	.0276	1.1025

Footnotes

1. Between 1920 and 2006, there were ten no-hitters that were called before nine innings (one each in 1924, 1937, 1940, 1944, 1967, 1984, 1988, 1990 and two in 1959) and eleven additional no-hitters in which more than one pitcher combined to achieve the no-hit game (one each in 1956, 1967, 1975, 1976, 1990, 1995, 1997, 2003 and three in 1991).

2. 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.

3. The number of degrees of freedom is "5" because we have added two cells, one each for "5" and "6" no-hit games per season.

4. In carrying out this goodness-of-fit test, the interval (seasons with 4 no-hit games) was combined with seasons with 5, 6, or 7 no-hit games (which occurred for the first time in 1962) so that the theoretical or expected frequency in each and every class interval was close to being at least 5.