Journal of the Korean Data & Information Science Society
2016, **27**(6), 1477–1485

http://dx.doi.org/10.7465/jkdi.2016.27.6.1477
한국데이터정보과학회지

야구 피타고라스 승률의 수렴특성[†]

이장택[1]

[1]단국대학교 응용통계학과

접수 2016년 9월 5일, 수정 2016년 10월 1일, 게재확정 2016년 10월 11일

요 약

본 연구에서는 한국프로야구에서 팀의 득점과 실점을 가지고 시즌 승률을 예측하는 야구의 피타고라스 정리에 의한 기대승률의 수렴특성을 살펴보았다. 사용한 자료는 2005년부터 2014년까지의 한국프로야구 정규시즌 초부터 정규시즌 말까지의 팀대 팀 전체기록이며, 그 결과 야구 팀의 특징 중에서 팀의 순위와 경기진행률이 수렴특성에 영향을 주는 것으로 나타났다. 팀의 순위는 하위 팀들의 기대승률이 최종 기대승률에 빨리 수렴하였으며, 경기진행률은 20% 이하에는 최종 기대승률과 많은 차이를 보였으나 70% 이상부터는 통계적으로 최종 기대승률과 유의한 차이가 발생하지 않았다.

주요용어: 기대승률, 수렴특성, 피타고라스 정리, 한국프로야구.

1. 서론

야구에 대한 객관적인 지식을 찾고자 하는 움직임의 기원이 된 야구의 피타고라스 정리는 문헌 및 인터넷을 통하여 많은 스포츠 통계학자 및 스포츠팬들의 초미의 관심사가 되고 있다. James (1982)가 명명했던 야구의 피타고라스 정리는 야구의 승률과 득점 및 실점의 연관성을 잘 설명해 주는데, 그는 1980년대 초 메이저리그 팀들의 과거 성적을 정리하다가 특정 팀의 총득점과 총실점이 팀의 승률과 밀접한 관계가 있다는 것을 알게 되었다. 일반적으로 지수 γ를 사용한 야구의 피타고라스 정리에 의한 피타고라스 기대승률 (WP)은 식 (1.1)과 같이 시즌의 총득점 (RS)과 총실점 (RA)의 비선형함수로 정의된다.

$$WP = \frac{RS^{\gamma}}{RS^{\gamma} + RA^{\gamma}} \tag{1.1}$$

지수 γ의 값은 처음에는 James가 주장한 2를 사용하였으며 지수 값으로 2를 사용하면 식 (1.1)에서 분모는 총득점의 제곱과 총실점의 제곱의 합이므로 수학의 피타고라스 정리와 비슷한 모양이라서 야구의 피타고라스 정리라는 이름을 사용한다. 하지만 많은 사람들이 메이저리그의 누적된 자료를 다루다보니 지수 값을 1.83으로 낮추어 사용하는 것이 좀 더 바람직하다는 견해가 많다. 지수 γ는 일반적으로 많이 사용하는 추정량 선택기준인 평균제곱오차의 제곱근 (root mean square error)을 최소화하는 값으로 설정하는데, 한국프로야구의 경우에도 초기 자료를 이용하면 메이저리그와 다른 값이 나왔지만 프로야구 원년부터 2014년도까지의 데이터를 적합시켜 보면 미국의 경우와 거의 같은 값이 나타난다 (Lee, 2015). 야구의 피타고라스 정리는 한국프로야구에서도 적용이 매우 잘 되는데, Lee (2015)에 의하면 2005년부터 2014년까지 총 82개 팀들의 실제 승률과 피타고라스 정리에 의한 기대승률의 차이는 가장

[†] 이 연구는 2016학년도 단국대학교 대학연구비 지원으로 연구되었음.
[1] (448-701) 경기도 용인시 죽전동 126번지, 단국대학교 응용통계학과, 교수. E-mail: jtlee@dankook.ac.kr

큰 경우가 6.24%, 가장 작은 경우가 0.03%, 평균 1.95%로 이 정도의 오차는 일반적으로 설명할 수 없는 랜덤오차로 보기에 충분하다고 할 수 있겠다.

지금까지 수행된 야구의 피타고라스 정리에 관한 연구들은 크게 나누면 세 가지로 대별된다. 첫째는 피타고라스 정리에 사용되는 지수 γ의 최적 값을 구하는데 그 목적이 있는데, 메이저리그 데이터를 이용한 Davenport와 Woolner (1999), Cochran (2008) 및 한국프로야구 데이터를 이용한 Lee (2014b) 등이 이 범주에 속하는 연구들이다. 둘째는 야구의 피타고라스 정리가 야구에서 시작되었지만 다른 스포츠로 확장 적용이 가능하다는 연구들이다. 해외연구들은 수많은 결과들이 있으며, 국내연구로는 Lee와 Kim (2006a, 2006b)은 한국여자프로농구와 프로축구에서도 지수 값을 각각 10.8과 1.378을 사용하여 승률을 잘 추정할 수 있다고 밝혔다. 셋째로는 야구의 피타고라스 정리에 대한 이론적인 근거를 제공하는 연구들이다. 최초의 연구는 Miller (2006)에 의해서 몇 가지 가정과 와이블 분포를 이용하여 야구의 피타고라스 정리가 성립함을 이론적으로 보였으며 Dayaratna와 Miller (2013)는 하키에서도 이론적으로 성립가능하다는 사실을 증명하였다. 이밖에도 한국프로야구에 관한 최근 연구들을 소개하면 한국프로야구 타자들에 대한 세이버메트릭스 지수 값을 이용하여 선수들의 경기력과 연봉간의 패턴을 분석한 Seung과 Kang (2012), 한국프로야구에서 출루율 계수 추정을 다룬 Lee (2014a), 한국프로야구에 적당한 타자력 지수 모형과 지수를 제안한 Hong 등 (2016) 등이 있다.

야구의 피타고라스 정리는 한 시즌의 경기가 모두 끝난 다음 총득점과 총실점으로 승률을 예측한다. 하지만 이 사실은 단지 수학적 흥미일 뿐 실제로는 시즌 중간에 경기가 거듭될수록 앞으로 특정 팀의 승률이 어떻게 변해갈지에 더욱 많은 관심이 갈 것이다. 이런 관점에서 피타고라스 정리에 의한 기대승률의 수렴특성을 아무도 언급하지 않는 것은 어떻게 보면 놀라운 일이며 야구에서 피타고라스 정리가 어느 정도의 경기가 진행되어야 신뢰를 할 수 있는 지를 연구한 결과는 전혀 찾아 볼 수가 없다. 따라서 본 연구에서는 이와 같은 관점에 초점을 맞추어서 한국프로야구에서의 피타고라스 기대승률의 수렴특성을 살펴보았다. 본 논문은 다음과 같이 구성되어 있다. 2절에서는 승률과 기대승률의 정의, 분석데이터 및 통계분석에 대하여 언급하였으며, 3절에서는 기대승률 차이에 대한 기술통계량, 분산분석 및 회귀분석 결과를 소개하며 끝으로 4절에서는 본 연구의 결론에 대해 언급하였다.

2. 승률과 자료수집

표기의 간편성을 위하여 $W\%$는 승률, W는 승리한 경기 횟수, L은 패배한 경기 횟수, T는 무승부 경기 횟수를 각각 나타내면, 야구의 피타고라스 정리에서 James (1982)가 사용한 승률의 정의는 식 (2.1)과 같다.

$$W\% = \frac{W}{W+L} \tag{2.1}$$

하지만 한국프로야구에서 사용된 팀의 승률은 무승부제외 승률제, 무승부포함 승률제, 다승제가 있다 (Kim, 2011). 또한 한국프로야구에 대한 공식기록들은 모두 무승부인 경우도 포함해서 집계되었기 때문에 오랜 기간 동안의 데이터를 무승부를 제외하고 재집계하는 것은 거의 불가능하다. 따라서 본 연구에서는 승률의 정의로 1987시즌부터 1997시즌까지 사용한 무승부포함 승률제의 식 (2.2)를 사용하였는데,

$$W\% = \frac{W+0.5\times T}{W+T+L} \tag{2.2}$$

비록 식 (2.1)과 식 (2.2)가 약간의 차이가 있지만 거의 유사한 값을 제공하고, W^*와 L^*를 각각 $W^* = W+0.5\times T$, $L^* = L+0.5\times T$로 두면, 식 (2.2)는 식 (2.1)의 모양으로 기술할 수 있어서 야구의 피타고라스 정리를 적용할 수 있다. 또한 경기에서 승리할 비율인 승률 ($W\%$)과 패배할 비율인 패율 ($L\%$) 사이에는 식 (1.1)과 식 (2.2)를 사용하면 식 (2.3)이 성립하는 데,

$$\frac{W\%}{L\%} = \frac{W^*}{L^*} = \left(\frac{RS}{RA}\right)^{\gamma} \tag{2.3}$$

따라서 주어진 데이터와 최소제곱법을 이용하여 식 (2.4)와 같은 회귀모형을 고려하여 γ값을 추정할 수 있다.

$$\log(W^*/L^*) = \gamma \log(RS/RA) \tag{2.4}$$

본 연구에서 사용된 데이터는 2005년부터 2014년 사이에 있었던 한국프로야구 팀대 팀 경기결과 전체기록을 이용하였는데 총 5296개이며 출처는 롯데자이언츠 홈페이지 `http://www.giantsclub.com`이다. 식 (2.4)와 통계패키지 SAS 9.3 및 SPSS 21K를 이용하여 연도별 각 구단의 경기진행률에 대응되는 피타고라스 정리에 의한 시즌 중 기대승률을 구하고 최종 기대승률과의 차이에 대하여 여러 가지 통계 분석을 실시하였는데, 이 경우 특정 팀의 경기진행률 50%의 시즌 중 기대승률은 경기진행률 50%까지의 총득점과 총실점을 이용하여 식 (1.1)에 대입한 결과이다.

3. 분석결과

3.1. 기대승률 차이에 대한 기술통계량

Table 3.1은 퍼센트로 계산한 피타고라스 정리에 의한 시즌 중 기대승률값과 최종 기대승률 값의 차에 절대값을 취한 기대승률 차이 (winning percentage difference; WPD)에 대한 기술통계 값을 보여준다. 이 경우 예를 들어 3번의 경기를 치룬 결과에서 절대값을 취하지 않으면 기대승률 차가 양수, 음수, 음수가 되어서 기대승률 차의 평균이 0에 가까운 숫자로 나타날 수도 있기 때문이다. 피타고라스 정리의 기대승률 값은 모두 지수값을 프로야구 원년부터 2014년까지의 팀별 승률, 득점, 실점 데이터를 이용하여 추정한 최적지수값 1.834를 사용하여 계산하였다. 좀 더 구체적으로 WPD에 미치는 영향을 알아보기 위하여 WPD가 세 가지 인자에 의해 영향을 받을 것이라는 가정아래에서 연도 (year)를 2005년부터 2014년까지 10개, 팀의 순위 (rank)를 상, 중, 하의 3개, 팀의 경기진행률 (rate) 10개 그룹에 의한 WPD 값을 조사하였다. 여기서 팀의 순위는 각 연도 최종순위가 1위부터 3위는 상, 4위부터 6위는 중, 7위부터 9위는 하로 명명하였으며, 경기진행률은 해당경기 순번을 연간 총경기수로 나눈 값을 이용하였는데, 특정 팀이 12번째 경기를 하고 1년의 총 경기수가 120이면 경기진행률은 10%가 된다. 따라서 고려된 경기진행률 그룹은 모두 10개로 이름을 각각 G10부터 G100으로 명명하였는데, 예를 들면 G10은 경기진행률이 10% 이하, G20는 경기진행률이 10% 초과 20% 이하에 속하는 경기들의 결과 등등이다. Table 3.1은 연도별 WPD에 대한 평균, 표준편차, 왜도 및 첨도를 보여주는데, 첨도를 제외하고는 다른 통계량의 수치 값이 유사하다고 판단되어진다.

Table 3.1 Descriptive statistics for WPD by year

Year	Mean	Standard Deviation	Skewness	Kurtosis
2005	5.067	6.264	2.732	9.800
2006	3.711	6.143	3.848	19.296
2007	3.927	6.208	4.400	25.062
2008	3.939	5.904	3.410	14.356
2009	4.438	5.911	3.341	16.256
2010	4.120	6.096	2.972	10.991
2011	4.762	5.808	3.202	15.372
2012	4.181	6.079	3.426	14.428
2013	4.983	6.532	2.457	7.391
2014	4.138	5.073	3.529	21.974

Table 3.2는 팀의 순위에 따른 WPD에 대한 자료의 개수, 평균, 표준편차, 왜도 및 첨도를 보여준다. 팀의 순위가 하 (low)에 속하는 경우가 다른 2가지 경우보다 상대적으로 평균값이 작다.

Table 3.2 Descriptive statistics for WPD by rank

Rank	N	Mean	Standard Deviation	Skewness	Kurtosis
high	3876	4.344	5.838	3.207	14.916
middle	3876	4.616	6.217	3.186	13.462
low	2840	3.902	5.962	3.577	16.857

Table 3.3은 경기진행률에 따른 WPD에 대한 평균, 표준편차, 왜도 및 첨도를 보여준다. 당연한 귀결이지만 경기진행률이 커질수록 WPD의 값은 점점 작아지며, 표준편차도 확연하게 줄어드는데, 경기진행률에 따른 WPD의 변화는 20%까지는 큰 변화를 보이다가 20%를 넘기면서 완만하게 WPD 값이 줄어든다.

Table 3.3 Descriptive statistics for WPD by rate

Rate	Mean	Standard Deviation	Skewness	Kurtosis
G10	14.836	11.662	1.010	0.521
G20	7.327	5.448	0.749	-0.022
G30	5.663	4.120	0.866	0.551
G40	4.360	3.287	1.111	1.129
G50	3.398	2.799	1.252	1.651
G60	2.688	2.192	1.276	1.714
G70	2.125	1.721	1.062	0.776
G80	1.684	1.249	0.963	0.653
G90	1.230	0.990	1.259	2.252
G100	0.582	0.572	1.717	3.953

3.2. 기대승률 차이에 대한 분산분석

WPD에 미치는 영향을 알아보기 위하여 연도, 순위, 경기진행률과 같은 세 가지 요인의 영향을 분산분석을 통하여 살펴보았는데, 종속변수 WPD의 값은 유의성검정 결과에 미치는 표본수의 과대효과를 배제하기 위하여 해당그룹에 속하는 관측치들의 평균을 WPD 값으로 이용하였다. 예를 들면 연도가 2005년, 순위가 상, 경기진행률이 G30인 경우에 속한 관측치들의 평균을 세 가지 조합에 해당하는 관측치로 사용하였다. 따라서 제일 처음 고려한 모형은 세 가지 인자를 고려한 식 (3.1)과 같은 반복이 없는 삼원배치 분산분석모형이었다. 식 (3.1)에서 α_i는 연도 i, β_j는 순위 j, γ_k는 경기진행률 k, 그리고 Y_{ijk}는 i번째 연도, j번째 순위, k번째 경기진행률인 경우의 WPD를 나타내는 확률변수이다.

$$Y_{ijk} = \mu + \alpha_i + \beta_i + \gamma_k + \varepsilon_{ijk}, \quad i = 1, 2, \cdots, 10; \quad j = 1, 2, 3; \quad k = 1, 2, \cdots, 10 \tag{3.1}$$

삼원배치 분산분석을 수행한 결과, 순위와 경기진행률은 유의수준 1%에서 유의하였으나, 연도 인자가 유의확률이 0.122으로 유의수준 5%에서 유의하지 않았다. 따라서 주효과의 검정력을 높일 목적으로 유의하지 않은 주효과를 오차항으로 풀링한 후에 반복이 있는 이원배치 분산분석을 실시하였으며, 그 결과 순위와 경기진행률의 교호작용의 유의확률이 0.997으로 유의수준 5%에서 유의하지 않았기 때문에 역시 교호작용 효과를 오차항으로 풀링한 후 다시 이원배치 분산분석을 수행한 결과가 Table 3.4인데, 팀의 순위와 경기진행률은 각각 p값이 0.002, <0.001으로 유의수준 1%에서 유의하게 나타났다. 제안된 분산분석 모형은 잔차를 통하여 정규성과 등분산성을 만족하는 사실을 확인할 수 있었으며 그 결과 모형이 적합하다고 판단할 수 있었다.

Table 3.4 Results of two way ANOVA table

Source	Sum of Squares	df	Mean Square	F	Sig.
Order	44.530	2	22.265	6.153	0.002
Rate	4645.655	9	516.184	142.650	0.000
Error	1042.139	288	3.619		
C. Total	5732.324	299			

유의성이 입증된 2개의 주효과에 대하여 수준 간 차이를 다중 비교를 이용하여 살펴보았는데, 가장 보수적인 방법으로 알려진 Tukey 방법을 사용하였다. Table 3.5는 팀의 순위에 대한 다중비교 결과로서 WPD 평균이 순위가 하 (low)일 때 가장 낮았고 상 (high), 중 (middle) 순서이다. 하지만 그룹 상 (high)과 중 (middle)의 차이는 유의수준 5%에서 통계적으로 유의하지 않았다. Table 3.6은 경기진행률에 대한 다중비교 결과로서 예상대로 G100, G90 등의 순서지만 G10, G20는 확연히 다른 그룹들과 통계적으로 유의한 차이를 보였다. 즉 경기진행률이 20% 이하는 유의하게 최종 피타고라스 기대승률과 차이가 나며, G70부터 G100까지 한 개의 그룹으로 묶이는 것을 보아서 경기진행률이 70% 이상이 되면 최종피타고라스 기대승률과 통계적으로 유의한 차이가 발생하지 않음을 알 수 있다.

Table 3.5 Results of Tukey test on the factor: Rank

rank	subset	
	1	2
low	3.7306	
high		4.3979
middle		4.6421
Sig.	1.000	0.636

*Means for groups in homogeneous subsets are displayed at the.05 level

Table 3.6 Results of Tukey test on the factor: Rate

rate	subset						
	1	2	3	4	5	6	7
G10	0.571						
G09	1.200	1.200					
G08	1.610	1.610					
G07	2.007	2.007	2.007				
G06		2.566	2.566				
G05			3.272	3.272			
G04				4.218	4.218		
G03					5.469		
G02						7.065	
G01							14.585
Sig.	0.104	0.148	0.234	0.651	0.248	1.000	1.000

*Means for groups in homogeneous subsets are displayed at the.05 level

3.3. 기대승률 차이에 대한 회귀추정식

Figure 3.1은 Table 3.4에서 유의한 인자로 나타난 팀의 순위와 경기진행률에 따른 WPD의 평균값을 보여주는데, 분산분석의 결과처럼 순위에 따른 변화는 크지 않으나 경기진행률에 따른 변화는 진행률이 20%이하인 경우는 급격한 변화를 보이다가 그 이후로는 완만하게 WPD 평균값이 감소하는 사실을 확인할 수 있다.

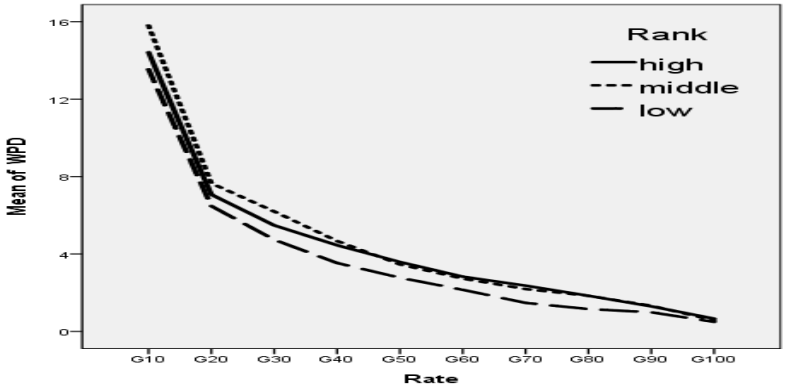

Figure 3.1 Scatterplot of mean WPD

따라서 팀의 순위와 경기진행률에 대한 가변수를 고려한 식 (3.2)와 같은 회귀모형을 고려할 수 있는데 여기서 변수 Y는 WPD, D_1은 팀의 순위가 상 또는 중이면 1, 하이면 0, D_2는 경기진행률이 20% 이하이면 1, 아니면 0인 가변수, X_3은 경기진행률을 각각 의미한다. 또한 간편성을 위해 사용한 X_4, X_5, X_6, X_7은 교호작용으로서 각각 $X_4 = D_1D_2$, $X_5 = D_1X_3$, $X_6 = D_2X_3$, $X_7 = D_1D_2X_3$을 의미한다.

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \epsilon \tag{3.2}$$

변수선택은 단계선택법을 이용하였으며 그 결과가 Table 3.7과 Table 3.8이다.

Table 3.7 Model summary for regression analysis

R Square	Adjusted R Square	Std. Error of the Estimate
0.814	0.811	1.90251

Table 3.8 Estimated regression model coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	VIF
	B	Std. Error	Beta			
Constant	06.339	0.401		+15.818	0.000	
$D1$	00.789	0.233	+0.085	+03.388	0.001	1.000
$D2$	15.241	0.860	+1.395	+17.721	0.000	9.810
$X3$	-0.065	0.005	-0.430	-12.200	0.000	1.964
$X6$	-0.687	0.049	-1.006	-13.895	0.000	8.298

회귀분석 분산분석표는 지면 관계상 생략되었지만 p값은 p <0.001로 회귀직선은 유의수준 1%에서 매우 유의한 것으로 나타났으며, Table 3.7에서 알 수 있듯이 결정계수도 81.4%로 높게 나타났다. Table 3.8은 추정된 회귀식, 표준화 회귀계수 및 VIF를 보여주는데, 단계선택법 결과 선택되어진 변수들은 모두 유의수준 1%에서 유의하였으며, 3개의 변수들에 대한 VIF의 값이 모두 10보다 작아서 다중공선성의 문제는 없는 것으로 나타났다. 또한 잔차를 통하여 회귀모형에서의 선형성, 독립성, 정규성, 등분산성과 같은 기본 가정들이 모두 성립함을 확인할 수 있었다. 그 결과, 추정된 회귀식은 식 (3.3)과 같이 기술할 수 있으며 표준화 회귀계수를 이용하면 경기진행률 D_2가 가장 영향력이 있는 것으로 나타났다.

$$\widehat{Y} = 6.339 + 0.789D_1 + 15.241D_2 - 0.065X_3 - 0.687X_6 \tag{3.3}$$

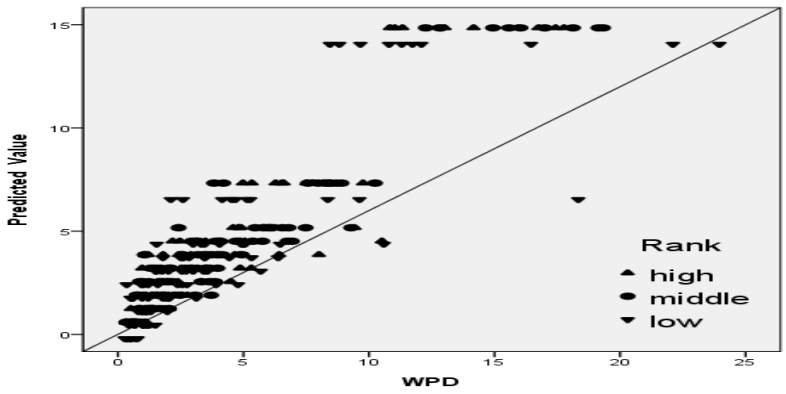

Figure 3.2 Scatterplot of predicted vs. actual values

Figure 3.2는 순위에 따른 변화를 고려하여 WPD와 회귀직선을 이용하여 추정된 WPD의 관계를 보여주는 산점도이다. 대체적으로 실제 WPD를 추정된 WPD로 잘 예측하고 있으나 WPD의 값이 증가할수록 잔차가 커짐을 알 수 있다.

4. 결론

피타고라스 기대승률은 야구에서 실제승률을 예측하는 가장 유명한 공식인데 팀의 득점과 실점을 가지고 시즌 승률을 예측한다. 야구팬들은 이 식을 이용해서 시즌 중간의 팀 순위를 가지고 시즌 후반의 팀 순위변동을 예측하기도 하며 어떤 팀이 실력보다 실제 승률이 높은지, 낮은지를 평가하기도 한다. 하지만 이 공식은 적은 경기의 득점과 실점이 아니라 한 시즌 전체의 득점과 실점을 사용하여야 더 신뢰도가 커지는데, 이와 같은 관점에서 시즌 중 경기진행률에 따른 피타고라스 기대승률의 신뢰도를 살펴보는 일은 매우 의미 있는 일이라고 할 수 있겠다.

본 연구에서는 최종 피타고라스 기대승률에 대한 시즌 중의 피타고라스 승률 수렴특성을 살펴보았다. 그 결과 팀의 순위가 낮고 경기진행률이 커질수록 전반적으로 최종 피타고라스 기대승률과의 차이가 적었다. 경기진행률은 20%까지는 최종 피타고라스 기대승률과의 괴리가 심하나 20%를 초과해서는 경기진행률이 커질수록 완만하게 최종 피타고라스 승률에 수렴하였으며, 경기의 진행률이 70%를 넘기면 최종 피타고라스 기대승률과 유의수준 5%에서 통계적으로 유의한 차이를 보이지 않았다. 향후 연구과제로는 본 연구에서의 고려된 연도, 팀의 순위 및 경기진행률 이외의 야구선수나 팀을 평가하는 타율, 장타율, 출루율, 수비율 등과 같은 야구통계량들을 사용하면 야구 피타고라스정리 수렴특성에 관한 좀 더 설득력 있는 결과를 창출할 수 있을 것으로 간주된다.

References

Cochran, J. J. (2008). The optimal value and potential alternatives of Bill James Pythagorean method of baseball. *STAtOR*, **2**, 2008.

Davenport, C. and Woolner, K. (1999). Revisiting the Pythagorean theorem: Putting Bill James' Pythagorean theorem to the test. *The Baseball Prospectus*, `http://www.baseballprospectus.com/article.php?articleid=342`.

Dayaratna, K. D. and Miller, S. J. (2013). The Pythagorean won-loss formulaand hockey: A statistical justification for using the classic baseball formula as an evaluative tool in hockey, `http://arxiv.org/ftp/arxiv/papers/1208/1208.1725.pdf`.

Hong, C. S.., Kim, J. Y. and Shin, D. S. (2016). Alternative hitting ability index for KBO. *Journal of the Korean Data & Information Science Society*, **27**, 677-687.

James, B. (1982). *The Bill James abstract*, Ballantine, New York.

Kim, H. J. (2011). Suggestion of a new method of computing percentage of victories for the Korean professional baseball. *The Korean Journal of Applied Statistics*, **24**, 1139-1148.

Lee, J. T. and Kim, Y. T. (2006a). A study on the estimation of winning percentage in Korean pro-baseball. *Journal of the Korean Data Analysis Society*, **8**, 857-869.

Lee, J. T. and Kim, Y. T. (2006b). Estimation of winning percentage in Korean pro-sports. *Journal of the Korean Data Analysis Society*, **8**, 2105-2116.

Lee, J. T. (2014a). Estimation of OBP coefficient in Korean professional baseball. *Journal of the Korean Data & Information Science Society*, **25**, 357-363.

Lee, J. T. (2014b). Estimation of exponent value for Pythagorean method in Korean pro-baseball.*Journal of the Korean Data & Information Science Society*, **25**, 493-499.

Lee, J. T. (2015). Measuring the accuracy of the Pythagorean theorem in Korean pro-baseball. *Journal of the Korean Data & Information Science Society*, **26**, 653-659.

Miller, S. J. (2006). A derivation of the pythagorean won-loss formula in baseball. *By the Numbers*, **16**, 40-48.

Seung, H. B. and Kang, K. H. (2012). A study on relationship between the performance of professional baseball players and annual salary. *Journal of the Korean Data & Information Science Society*, **23**, 285-298.

Journal of the Korean Data & Information Science Society
2016, **27**(6), 1477–1485

http://dx.doi.org/10.7465/jkdi.2016.27.6.1477
한국데이터정보과학회지

Convergence characteristics of Pythagorean winning percentage in baseball[†]

Jangtaek Lee[1]

[1]Department of Applied Statistics, Dankook University

Received 5 September 2016, revised 1 October 2016, accepted 11 October 2016

Abstract

The Pythagorean theorem for baseball based on the number of runs they scored and allowed has been noted that in many baseball leagues a good predictor of a team's end of season won-loss percentage. We study the convergence characteristics of the Pythagorean expectation formula during the baseball game season. The three way ANOVA based on main effects for year, rank, and baseball processing rate is conducted on the basis of using the historical data of Korean professional baseball clubs from season 2005 to 2014. We perform a regression analysis in order to predict the difference in winning percentage between teams. In conclusion, a difference in winning percentage is mainly associated with the ranking of teams and baseball processing rate.

Keywords: Convergence characteristics, Korean professional baseball, Pythagorean method, winning percentage.

[†] The present research was conducted by the research fund of Dankook University in 2016.
[1] Professor, Department of Applied Statistics, Dankook University, Gyeonggi-do 448-701, Korea. E-mail: jtlee@dankook.ac.kr