Journal of the Korean Data & Information Science Society
2014, **25**(6), 1231–1239

http://dx.doi.org/10.7465/jkdi.2014.25.6.1231
한국데이터정보과학회지

K-리그에서 축구 골의 분포

이장택[1]

[1]단국대학교 응용통계학과

접수 2014년 7월 25일, 수정 2014년 8월 16일, 게재확정 2014년 9월 2일

요 약

본 연구에서는 1983년부터 2012년까지의 한국프로축구 K-리그 전 경기 결과를 이용하여 홈 경기와 원정 경기에서의 골의 분포를 분석하였다. 고려된 확률분포는 포아송분포, 음이항분포, 극단치 분포 및 영과잉 포아송분포이며, 카이제곱분포를 이용한 적합도검정을 수행하였다. 그 결과 홈경기는 포아송분포, 원정경기는 영과잉 포아송분포가 골의 분포를 위한 최적 적합분포로 간주되며 홈경기와 원정경기 골의 수는 서로 약한 정도의 상관관계가 있는 것으로 나타났다.

주요용어: 골의 분포, 영과잉 포아송분포, 케이-리그, 포아송분포.

1. 머리말

세계에서 가장 많은 팬을 보유한 스포츠는 아마도 축구일 것이며 전 세계 수많은 인구가 축구문화에 심취해 있다. 대중성과 경제적 부가가치로 인하여 여러 분야에서 중요한 연구대상으로 간주되는 축구는 역동적인 축구경기의 특성상 통계적 분석의 가능성에 대한 의심이 많았지만, 분석결과들은 팀의 전술을 평가하고 발전시키는 수단이 되었다. 축구의 점수 분포에 대한 연구는 많은 학자들에 의해 50년 이상 진행되어지고 있는데, 처음에는 골을 넣을 수 있는 확률이 각 팀에만 종속되는 고정된 값으로 간주한 포아송분포를 이용하여 전체 및 각 팀별 골 득점의 분포를 모형화 하였다 (Moroney, 1956). 그런데 골의 분포가 포아송분포를 따른다는 사실은 골이 발생하는 것이 단위 시간의 길이에만 의존하기 때문에 특정 팀의 골의 기댓값은 모든 경기에서 같다고 할 수 있으나 상대 팀에 따라 결과도 달라질 수 있고, 두 팀이 비슷한 전력을 가지고 있다고 하더라도 1-0인 경우와 5-0인 경우는 선수들의 경기에 임하는 자세가 다를 수 있다고 할 수 있으므로 이런 경우는 음이항분포로 접근하는 것이 타당하다고 할 수 있다 (Reep 등, 1971; Maher, 1982). 한편 1999년부터 2001년 사이의 169개국 국내 장기리그 축구경기를 조사한 Greenhough 등 (2002)은 꼬리의 분포가 매우 두터워서 포아송분포와 음이항분포로는 설명력이 떨어지고 극단치분포가 타당한 경우가 있다고 주장하였다. 또한 축구의 골은 영의 값이 과잉관측 되는 경우가 많기 때문에 Germert (2010)은 영과잉 포아송분포를 이용하여 골의 분포를 설명하였다. 이와 같이 프리미어 리그나 분데스리가와 같은 해외유명 축구리그인 경우에는 선행연구들이 많이 있으나 비록 2002년 월드컵 4강까지 진출한 한국 축구이지만 한국을 대표하는 K-리그에 대한 체계적인 연구들은 많지 않다. 축구에 대한 국내연구들은 주로 스포츠 경영학, 스포츠 의학, 스포츠 심리학의 범주에 속하는 연구가 다수이며, 축구 경기 결과에 대해서 유럽 리그에서 득점과 실점을 이용한 승점 추정에 관한 연구 (Shin 등, 2009), 2010년 남아공 월드컵 축구의 최종 결과와 예측자료를 비교한 연구 (Hong 등, 2010), 축구대표팀의 선수들 간의 패스 정보를 이용한 사회네트워크분석 (Choi 등, 2011), 이변량 포아송모형을 K-리그 축구 경기 결과에 적용한 연구 (Kim, 2012) 등이 있다.

[1] (448-701) 경기도 용인시 죽전동 126번지, 단국대학교 응용통계학과, 교수. E-mail: jtlee@dankook.ac.kr

본 연구는 K-리그 축구경기에서의 골 분포를 살펴보고자 한다. 2절에서는 골 분포의 적합에 사용되는 여러 가지 확률분포를 기술하였으며 3절에서는 연구에 사용된 K-리그 데이터를 소개하고 홈경기와 원정경기의 골 분포를 여러 가지 확률분포에 적합시키고 골 분포의 상호 연관성을 살펴보았다. 끝으로 4절에서는 본 연구의 결론에 대해 언급하였다.

2. 골과 연관된 확률분포

이 절에서는 본 연구에서 고려된 축구의 골을 분석하기 위한 5가지 확률분포들을 설명한다.

2.1. 포아송분포

포아송분포 (Poisson distribution)는 축구에서 골의 분포를 설명하는 데 가장 많이 사용되는 분포로서 팀이 x골을 기록할 확률질량함수 $f_X(x)$는 다음과 같이 주어진다.

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!},\ x = 0, 1, 2, \cdots \tag{2.1}$$

위 식에서 모수 λ는 게임당 골의 평균이며 분산이 된다. 따라서 x골이 포아송분포를 따른다면 골의 평균과 분산이 유사한 값이 되어야하며, λ의 추정량으로 최대가능도 추정량인 표본평균을 일반적으로 사용한다.

2.2. 음이항분포

음이항분포 (negative binomial distribution; NBD)는 축구에서 골을 못 넣을 확률이 p일 때, r번째 골을 못 넣을 때까지의 골의 수 x에 대한 분포이며 확률질량함수 $f_X(x)$는 다음과 같다.

$$f_X(x) = \binom{r+x-1}{x} p^r (1-p)^x = \frac{\Gamma(x+r)}{x!\Gamma(r)} p^r (1-p)^x,\quad x = 0, 1, 2, \cdots \tag{2.2}$$

식 (2.2)에서 $\Gamma(x)$는 $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$로 정의되는 감마함수이며, 음이항분포의 모평균 μ와 모분산 σ^2은 다음과 같이 주어진다.

$$\mu = \frac{(1-p)}{p} \quad \sigma^2 = \frac{r(1-p)}{p^2}. \tag{2.3}$$

모수 r과 p의 최대가능도 추정량은 관측치 $x_1, x_2, \cdots, x_n$와 $f_X(x)$에 대해 식 (2.4) 와 같이 주어지는 로그가능도함수 $= \sum_{i=1}^n \ln f_X(x_i)$를 최대로 하는 값이다.

$$\ln L = \sum_{i=1}^n \ln\left(\Gamma(x_i + r)\right) - \sum_{i=1}^n \ln(x_i!) - n\ln(\Gamma(r)) + nr\ln(p) + \sum_{i=1}^n x_i \ln(1-p) \tag{2.4}$$

식 (2.4)를 각각 r과 p로 미분하여 0으로 두면 다음 식 (2.5)와 (2.6)을 얻게 된다.

$$\frac{\partial \ln L}{\partial p} = \frac{nr}{p} - \sum_{i=1}^n \frac{x_i}{1-p} = 0 \tag{2.5}$$

$$\frac{\partial \ln L}{\partial r} = \sum_{i=1}^n \Psi(x_i + r) - n\Psi(r) + n\ln(p) = 0 \tag{2.6}$$

위 식에서 $\Psi(x) = \Gamma'(x)/\Gamma(x)$로 정의되는 함수이다. 식 (2.5)로부터 p는 구체적으로 구할 수 있으나 식 (2.6)은 뉴톤-랍슨 (Newton-Raphson) 방법과 같은 수치기법을 사용하여 풀 수 있다.

2.3. 극단치분포

만일 축구에서의 x골의 분포가 두터운 꼬리를 가지고 있다면 로그 정규분포는 영의 값을 포함하지 않는 이유로 주로 극단치분포 (generalized extreme value distribution; GEV)를 고려한다. 극단치분포의 확률밀도함수 $f_X(x)$는 식 (2.7) 및 식 (2.8)과 같이 표시되는 데, $\xi \neq 0$인 경우에 $\xi > 0$이면 프레셰 (Frechet) 분포, $\xi < 0$이면 와이블 (Weibull) 분포, $\xi = 0$이면 검벨 (Gumbell) 분포가 된다. 이 경우 ξ는 형태모수, u는 임계점, σ는 척도모수를 각각 의미한다 (Woo와 Kim, 2009).

$$f_X(x) = \frac{1}{\sigma}\left(1 + \frac{\xi(x-\mu)}{\sigma}\right)^{-1-1/\xi} \exp\left[-\left(1 + \frac{\xi(x-\mu)}{\sigma}\right)^{-1/\xi}\right], \quad \xi \neq 0 \tag{2.7}$$

$$f_X(x) = \frac{1}{\sigma}\exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right) - \frac{x-\mu}{\sigma}\right], \quad \xi = 0 \tag{2.8}$$

극단치분포의 로그가능도함수는 관측치 $x_1, x_2, \cdots, x_n$와 $f_X(x)$에 대하여 $\xi \neq 0$인 경우는 식 (2.9)과 같이 주어지며, $\xi = 0$인 경우는 식 (2.10)과 같이 주어진다.

$$\ln L = -n\ln\sigma - \left(\frac{1+\xi}{\xi}\right)\sum_{i=1}^{n}\ln\left[1 + \frac{\xi(x_i-\mu)}{\sigma}\right] - \sum_{i=1}^{n}\left(1 + \frac{\xi(x_i-\mu)}{\sigma}\right)^{-1/\xi} \tag{2.9}$$

$$\ln L = -n\ln\sigma - \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right) - \sum_{i=1}^{n}\exp\left[-\left(\frac{x_i-\mu}{\sigma}\right)\right] \tag{2.10}$$

위 식들을 각각 모수들로 미분하여 0으로 둔 식들도 역시 비선형 방정식이기 때문에 반복알고리즘을 사용하여 최대가능도 추정량을 구할 수 있다.

2.4. 영과잉 포아송분포

축구에서 팀이 0골을 기록할 확률이 포아송분포 가정 아래에서보다 더 큰 경우에 사용되는 분포가 영과잉 포아송분포 (zero inflated Poisson distribution; ZIP)이다 (Singh, 1963; Kim, 2005). ZIP분포의 확률질량함수 $f_X(x)$는 다음 식 (2.11)과 같이 정의된다.

$$f_X(x) = \begin{cases} \pi + (1-\pi)e^{-\lambda}, & x = 0 \\ (1-\pi)\dfrac{e^{\lambda-}\lambda^x}{x!}, & x = 1, 2, 3, \cdots \end{cases} \tag{2.11}$$

ZIP 분포의 로그가능도함수는 식 (2.12)과 같이 주어진다. 이 경우 n_0는 0의 값을 갖는 표본의 개수이며, a_i는 x_i의 값이 0이면 0, 양수이면 1로 정의된다.

$$\ln L = n_0 \ln(1 - \pi + \pi e^{-\lambda}) + \sum_{i=1}^{n} a_i \ln \pi - \lambda \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} a_i x_i \ln \lambda - \sum_{i=1}^{n} a_i \ln x_i! \tag{2.12}$$

$\widehat{\lambda}$와 $\widehat{\pi}$가 각각 λ와 π의 최대가능도 추정량이라고 할 때, $\widehat{\lambda}$와 $\widehat{\pi}$는 다음 식 (2.13)과 같이 표현되며, 이 경우 $\bar{x}_0$는 양의 값을 갖는 관측치 만의 평균값이다. 비선형 방정식인 식 (2.13)은 일반적으로 앞의 경우들과 같이 수치기법을 이용하여 구할 수 있다.

$$\hat{\pi} = \frac{n - n_0}{n(1 - e^{-\widehat{\lambda}})}, \quad \bar{x}_0 = \frac{\widehat{\lambda}}{1 - e^{\widehat{\lambda}}} \tag{2.13}$$

3. 데이터와 분석결과

3.1. 사용된 데이터

본 연구에서 사용된 데이터는 한국프로축구연맹이 주관하는 한국 프로축구리그인 K-리그 결과만을 포함하였으며, 챔피온스 리그는 제외하였다. 데이터셋은 1983년부터 2012년까지의 모든 K-리그 결과로 구성되어졌는데, 총 4,218 경기의 결과로써 모든 경기는 연장전을 제외한 정규시간만의 결과이며, 모든 자료의 출처는 K-리그 공식사이트 (http://www.kleague.com)이다. Table 3.1은 원정팀의 실점인 홈팀의 득점 (GS)과 홈팀의 실점인 원정팀의 득점 (GA)의 평균, 중앙값, 분산, 왜도 및 첨도를 연대별과 전체로 나타낸 결과이다. 1983년부터 2012년까지 30년간 홈팀 득점 (GS)은 평균 1.33, 분산 1.28이며, 원정팀 득점 (GA)은 평균 1.18, 분산 1.21으로 홈경기가 원정경기보다 평균과 분산이 크게 나타났다. 홈팀 득점과 원정팀 득점의 평균은 연대별로 거의 유사하며 중앙값은 모두 같다. 이상을 살펴보면 K-리그 경기 결과는 연대별로 큰 차이가 없으며, 경기당 골의 빈도가 많지 않음을 알 수 있다.

Table 3.1 Statistics for home goals and away goals

variable	year	mean	median	variance	skewness	kurtosis
GS	1983-1989	1.26	1.00	1.34	0.82	0.20
	1990-1999	1.34	1.00	1.23	0.65	0.01
	2000-2012	1.35	1.00	1.29	0.89	0.99
	1983-2012	1.33	1.00	1.28	0.82	0.65
GA	1983-1989	1.16	1.00	1.20	1.20	2.36
	1990-1999	1.22	1.00	1.25	0.84	0.44
	2000-2012	1.16	1.00	1.19	0.99	1.33
	1983-2012	1.18	1.00	1.21	0.98	1.21

Table 3.2 Frequencies for the number of home goals and away goals

goal	GS		GA	
	frequency	percent	frequency	percent
0	1089	25.8	1325	31.4
1	1485	35.2	1496	35.5
2	1024	24.3	903	21.4
3	445	10.6	364	8.6
4	130	3.1	94	2.2
5	37	0.9	29	0.7
6	6	0.1	4	0.1
7	2	0.0	2	0.0
8	0	0.0	1	0.0

Table 3.2는 30년간의 홈팀 득점 (GS)과 원정팀 득점 (GA)의 빈도표를 보여준다. 최빈값은 GS와 GA 모두 1이지만 그 다음으로 많은 경우는 GS와 GA 모두 0임을 알 수 있는 데, 이 사실로부터 경기당 발생하는 득점이 적은 것을 확인할 수 있다. Figure 3.1은 홈경기와 원정경기의 골 분포의 히스토그램으로 2가지 경우 모두 다른 나라의 경우와 마찬가지로 포아송분포와 유사한 패턴이다. 포아송분포는 데이터가 일정기간 동안 주어진 사건의 발생 횟수를 나타내고 서로 독립적으로 발생하는 경우에 값이 크지 않으면서 과잉산포 (overdispersion)화되어 있지 않은 경우에 가장 적합하다. 그런데 프리미어 리그나 분데스리그와 같은 해외리그는 모두 평균보다 분산이 더 큰 과잉산포 문제가 발생하기 때문에 포아송모형의 타당성이 결여되어 있지만, K-리그의 경우에 30년간 292건의 팀별 평균과 분산의 비율의 결과를 나타내는 Table 3.3에서 알 수 있듯이 홈팀 및 원정팀의 득점은 과잉산포의 문제가 없기 때문에 사건 간의 독립성이 어느 정도 전제되어지면 포아송모형으로 접근할 수 있다고 간주된다.

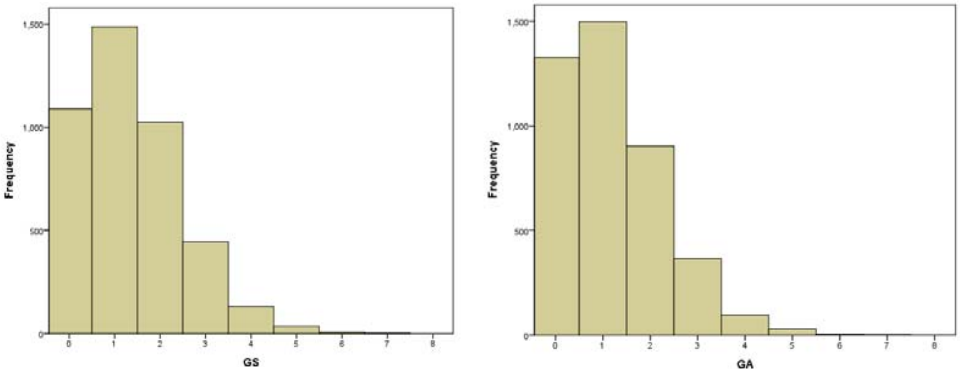

Figure 3.1 Histogram of home goals and away goals in K-league

Table 3.3 Frequencies for the number of overdispersion

variable	variance $\leq$ mean	variance $>$ mean
GS	188 (64.3%)	104 (35.7%)
GA	185 (63.4%)	107 (36.6%)

3.2. 여러 가지 지수 값에 대한 통계모형의 비교

Table 3.4 Expected frequencies for the number of home goals

Home Goals	Observed	Poisson	NBD	GEV	Gumble	ZIP
0	1089	1112.1	1121.9	1042.9	1022.3	1111.9
1	1485	1482.6	1476.0	1663.3	1638.4	1482.2
2	1024	988.2	980.6	978.7	970.5	987.9
3	445	439.1	438.6	370.5	386.3	439.1
4	130	146.3	148.6	116.9	134.3	145.9
5	37	39.1	40.7	33.6	44.6	38.8
6	6	8.7	9.4	9.1	14.6	8.4
7	2	1.7	1.9	2.3	4.8	1.3

홈경기와 원정경기의 골 분포에 대한 5가지 확률모형의 타당성은 카이제곱 적합도 검정을 통하여 수행되어졌으며, 고려된 다섯 가지 모형의 모수들은 RStudio (Ver. 0.97.320)를 사용하여 추정하였다. Table 3.4와 Table 3.5는 홈경기 및 원정경기 골의 실제 빈도수와 5가지 확률모형에 의한 기대빈도수를 각각 나타내는데, 2개의 극단치 분포를 제외하고는 비슷한 적합결과를 보여준다. Table 3.6은 홈 및 원정경기인 경우에 5가지 확률모형에 필요한 모수들의 최대가능도 추정치를 나타낸다. 그리고 Table 3.7는 서로 다른 모수의 개수를 고려한 5가지 모형에 대한 카이제곱 통계량의 값을 자유도로 나눈 통계량 $\widetilde{\chi}^2$과 유의확률을 나타낸다.

Table 3.7을 살펴보면 홈경기는 포아송분포, 원정경기는 ZIP 분포가 가장 적당하다고 간주된다. 음이항분포는 적합도의 순위에서 홈경기는 3위, 원정경기는 2위로 나타났으며 유의수준 15%에서도 모두 실제 자료에 적합한 것으로 나타났지만 일반적으로 알려져 있는 음이항분포가 포아송분포보다 훨씬 더 바람직하다는 결과는 K-리그에는 성립하지 않는다고 할 수 있다. 하지만 K-리그에서 Greenhough 등 (2002)의 결과와 달리 2개의 극단치분포 GEV와 Gumble 분포는 모두 적합하지 않다고 나타났으며, 포아송분포, 음이항분포, 영과잉 포아송분포를 비교하면 3개의 분포 모두 홈경기인 경우에 골의 수가 0인 경우는 과대추정하나 1이상 3이하인 경우는 과소추정하는 경향이 있으며, 원정경기는 골의 수가 1개는 과대추정, 골의 수가 0, 2, 3개인 경우는 대부분 과소추정하는 경향이 있는 것으로 나타났다.

Table 3.5 Expected frequencies for the number of away goals

Away Goals	Observed	Poisson	NBD	GEV	Gumble	ZIP
0	1325	1301.1	1320.1	1327.9	1222.7	1325.0
1	1496	1530.3	1514.7	1658.2	1686.1	1499.4
2	903	899.9	887.5	737.6	864.1	897.0
3	364	352.8	353.9	282.1	306.3	357.8
4	94	103.7	108.0	113.3	96.7	107.0
5	29	24.4	26.9	49.2	29.5	25.6
6	4	4.8	5.7	23.1	8.9	5.1
7	2	0.8	1.1	11.6	2.7	0.9
8	1	0.1	0.2	6.1	0.8	0.1

Table 3.6 Maximum likelihood parameter estimates

Home / Away	Poisson	NBD	GEV	Gumble	ZIP
Home	$\lambda = 1.333$	$r = 100$ $p = 0.987$	$\mu = 0.795$ $\sigma = 0.878$ $\xi = 0.031$	$\mu = 0.810$ $\sigma = 0.890$	$p = 0.001$ $\lambda = 1.333$
Away	$\lambda = 1.1761$	$r = 46.932$ $p = 0.976$	$\mu = 0.611$ $\sigma = 0.771$ $\xi = 0.153$	$\mu = 0.678$ $\sigma = 0.831$	$p = 0.017$ $\lambda = 1.197$

Table 3.7 $\widetilde{\chi}^2$values per degree of freedom with p-value

Home / Away		Poisson	NBD	GEV	Gumble	ZIP
Home	$\widetilde{\chi}^2$	0.91	1.59	10.28	7.74	1.01
	p-value	0.483	0.159	0.000	0.000	0.407
Away	$\widetilde{\chi}^2$	1.04	1.00	23.28	7.60	0.68
	p-value	0.402	0.422	0.000	0.000	0.670

각 경기의 승리, 무승부, 패배에 대한 확률은 골의 수에 대한 확률분포를 이용하여 추정할 수 있는데, 이 경우 K-리그의 득점과 실점이 서로 독립이라는 가정이 성립하게 되면 훨씬 수월하게 모형화를 시도할 수 있다. 하지만 득점과 실점이 서로 독립이라는 가정은 강한 가정이라고 할 수 있는데, 왜냐하면 축구경기에서 득점과 실점은 서로 연관이 있다고 생각되는 경우가 많이 있다고 알려져 있기 때문이다. Table 3.8은 GS와 GA에 대한 분할표이며, 피어슨 카이제곱검정 결과는 유의확률 0.000으로 득점과 실점은 서로 연관성이 있다고 나타났다.

Table 3.8 Crosstabulation of goals scored by home and away teams

Home Team	Away Team					Total
	0	1	2	3	4+	
0	448	350	189	81	21	1089
1	470	535	342	98	40	1485
2	261	394	218	119	32	1024
3	102	159	116	42	26	445
4+	44	58	38	24	11	175
Total	1325	1496	0903	0364	0130	4218

하지만 득점과 실점의 스페어만 순서상관계수는 0.143, 람다계수 값이 0.022로 거의 무시할 수 있을 정도로 나타났기 때문에 카이제곱 검정의 결과가 표본 수의 영향과 무관한지를 살펴보기 위하여 연도별로 스페어만 순서상관계수의 p-값을 작성하고 유의수준 5%와 1%에서 유의한 해당연도를 조사한 결과가 Table 3.9와 같다.

Table 3.9 Proportions of significant years for home and away teams

Significance level	N	Frequency	Percent	Mean ± SD
5%	30	12	40.0	0.012 ± 0.013
1%	30	7	23.3	0.003 ± 0.003

그 결과 대체적으로 득점과 실점은 서로 약한 정도의 상관정도가 있다고 할 수 있는데, 이 사실을 좀 더 명확하게 하기 위하여 Hasselblad (1994)에 기술되어있는 여러 개의 p-값을 통합하는 방법 중의 하나인 역카이제곱법을 사용하였다. 역카이제곱법에 의하면 서로 독립인 m개의 분할표에 의해 얻어진 p-값들을 p_i $(i = 1, 2, \cdots, m)$라 할 때, 다음 식 (3.1)이 성립하며, 기각역은 $U > \chi^2(2m; \alpha)$이 되는데,

$$U = -2\sum_{i=1}^{m} \ln p_i \ \sim \ \chi^2(2m) \tag{3.1}$$

K-리그의 경우, 데이터를 이용하여 검정통계량 U의 값을 구하면 164.54이며, 자유도가 60인 카이제곱 분포를 이용하여 계산한 p-값은 0.001보다 작아진다. 따라서 약한 정도의 상관관계가 존재한다고 결론지을 수 있다.

4. 결론

본 연구에서는 K-리그인 경우에 홈경기와 원정경기의 골 분포에 가장 적당한 확률분포를 찾아보았다. 여러 가지 선행 연구에서 사용된 5가지 확률 분포를 고려하였으며, 그 결과 홈경기와 원정경기 골의 주변확률분포는 홈경기는 포아송분포, 원정경기는 ZIP분포가 가장 적당하다고 판명되어졌으며, 홈경기와 원정경기의 골 수는 상관관계는 약하지만 존재하는 것으로 나타났다. 또한 음이항분포를 이용하여 득점 및 실점을 모형화하는 경우에도 타당하다고 나타났지만, 모수의 개수가 포아송분포보다 많고 적합도검정에서도 장점이 없는 것으로 나타났으며, 수많은 국가에서의 국내리그의 득점 및 실점을 설명하기 위한 극단치 분포는 K-리그의 경우에는 필요하지 않은 것으로 나타났다. 한편 K-리그 골의 분포가 연대별로 큰 차이가 없기 때문에 특정 연대에 따라 모형을 따로 구축할 필요는 없다고 간주되며, 본 연구의 후속편으로 향후연구관제로 K-리그에서 득점과 실점 사이의 독립성을 간주한 모형과 상관정도를 고려한 모형을 서로 비교하고 팀들 간의 승부와 승점 및 승률에 대한 예측을 위한 모형화를 시도해볼 예정이다.

References

Choi, S. B., Kang, C. W., Cho, H. J. and Kang, B. Y. (2011). Social network analysis for a soccer game. *Journal of the Korean Data & Information Science Society*, **22**, 1053-1063.

Gemert, D. (2010). *Modelling the scores of premier league football matches*, Master's Thesis, University of Amsterdam, Amsterdam, Netherlands.

Greenhough J., Birch P. C., Chapman S. C. and Rowlands G. (2002). Football goal distributions and extremal statistics. *Physica A*, **316**, 615-624.

Hasselblad, V. (1994). Meta-analysis in environmental statistics. *Handbook of Statistics*, **12**, 691-716.

Hong, C. S., Jung, M. S. and Lee, J. H. (2010). Prediction model analysis of 2010 South Africa World Cup. *Journal of the Korean Data & Information Science Society*, **21**, 1137-1146.

Kim, K. M. (2005). A study on the zero-inflated poisson regression model. *Journal of the Korean Data Analysis Society*, **7**, 497-505.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, **36**, 109-118.

Moroney M. J. (1956). *Facts from figures*, 3rd edition, Penguin, London.

Reep C., Pollard R. and Benjamin B. (1971). Skill and chance in ball games. *Journal of the Royal Statistical Society A*, **134**, 623-629.

Shin, S. K., Cho, Y. J. and Cho, Y. S. (2009). A study on points per game using scored goal per game and lossed goal per game in the union of European football professional league. *Journal of the Korean Data & Information Science Society*, **20**, 837-844.

Singh, S. N. (1963). A note on inflated poisson distribution. *Journal of the Indian Statistical Association*, **1**, 140-144.

Woo, J. Y. and Kim, M. S. (2009). Comparison study of parameter estimation methods for some extreme value distributions. *The Korean Communications in Statistics*, **16**, 463-477.

Journal of the Korean Data & Information Science Society
2014, **25**(6), 1231–1239

http://dx.doi.org/10.7465/jkdi.2014.25.6.1231
한국데이터정보과학회지

Soccer goal distributions in K-league

Jang Taek Lee[1]

[1]Department of Applied Statistics, Dankook University

Received 25 July 2014, revised 16 August 2014, accepted 2 September 2014

Abstract

In this paper we analyse the distributions of the number of goals scored by home teams and away teams in K-league soccer outcomes between 1983 and 2012. Real soccer data is explained in K-league using statistical distributions such that Poisson, negative binomial, extreme value and zero inflated Poisson. How close the goals of home and away fits the different distributions are tested by performing chi-square goodness of fit tests. According to these tests, the Poisson distribution gives the best fit to the home goals data. But it is best to model the away goals data on zero inflated Poisson distribution. Also, there is some weak evidence of the dependence for home and away goals.

Keywords: Goal distributions, K-league, Poisson, zero inflated Poisson.

[1] Professor, Department of Applied Statistics, Dankook University, Yongin 448-701, Korea.
E-mail: jtlee@dankook.ac.kr