慶北大學校 理學碩士學位論文

로또 데이터에서의 항등성 검정과 추정

大學院 統計學科

朴 美 貞

2004年 12月

慶北大學校 大學院

로또 데이터에서의 항등성 검정과 추정

이 論文을 理學碩士 學位論文으로 提出함

大學院 統計學科

朴 美 貞

指導敎授 孫 重 權

朴美貞의 理學碩士 學位論文을 認准함

2004年 12月

委員長 김 달 호 인

손 중 권 인

최 진 갑 인

慶北大學校 大學院委員會

차 례

1. 머리말

요즘 인기를 끌고 있는 로또 복권의 효시는 1530년대 이탈리아 제노바 공화국 정부가 구매자에게 90개의 숫자 가운데 5개를 고르게 해 당첨자를 뽑는 □□5/90 게임□□으로 알려져 있다. 제노바 공화국은 해마다 90명의 국회의원 가운데 추첨을 통해 5명을 상원의원으로 뽑았는데 □□5/90 게임□□은 이를 본뜬 것이다. 로또(Lotto)라는 단어는 □□운명□□(Lot)에서 유래했으며, 복권은 고대 로마의 아우구스투스(BC 63~AD 14) 황제가 연회(宴會) 입장 티켓에 일련번호를 표시해 참가자들에게 나눠주고 추첨을 해서 노예, 유람선, 저택 등을 준 것이 시작이라고 한다. 한국에는 1945년 일제가 군비마련을 위해□□승찰□□(勝札)이란 이름으로 처음 발행했고 해방 이후인 1947년에는 대한올림픽위원회(KOC)가 올림픽 후원권을 발행했다.

지금 한국에서 선풍적인 인기를 끌고 있는 국민은행발행의 로또 복권은 2002년 12월 시작됐으며 기존 전통 추첨식 복권에 현대적인 정보통신 기술을 결합하여 신속하고 다양해지는 고객의 요구에 부응하기 위해 발전된 흥미롭고 기술 집약적인 게임형 복권이라고 할 수 있다. 한국에서 발매되는 로또는 1부터 45까지의 숫자 중에 자신이 원하는 6개의 숫자를 임의로 고르는 '6/45' 방식이다. 5등(5,000원)을 제외한 1등에서 4등 당첨금은 확정되어 있지 않고 총 당첨금은 판매금액의 50% 이내로 판매금액에 따라 당첨금액이 달라진다. 1등은 총 당첨금 중 5등 금액을 제외한 60%, 2등은 총 당첨금 중 5등 금액을 제외한 10%, 3등은 총 당첨금 중 5등 금액을 제외한 10%, 4등은 총 당첨금 중 5등 금액을 제외한 20%를 상금으로 가지게 되며 5등은 5000원을 상금으로 받게 된다. 6개 숫자가 모두 맞아야 하는 1등 당첨확률은 814만 5,060분의 1이다. 로또는 기존 가판점에서 판매하는 추첨식 종이복권 대신 통신전용망과 단말기를 사용하고, 이미 정해진 번호를 사는 대신 고객이 직접 번호를 고르는 것이 특징이다. 그리고 당첨자가 없으면 당첨금이 이월된다. 1회분 발행 복권 수에 제한이 없고 참여자가 많을수록 당첨금이 늘어나는 점이 기존 복권과는 확연히 구분된다. 이런 로또는 엄청난 상금으로 인해 사람들로 하여금 관심의 대상으로 되고 있으며, 이에 따라 많은 학자들에 의해 로또의 여러 가지 연구가 행해지고 있다.

로또에 대한 연구는 Stern(1987)이 캐나다 로또 (6/49) 데이터로 최대동질성(Maximum Entropy)에 대해 연구했으며 그 논문에서는 몬테칼로 방법(Monte Carlo methods)을 이용해서 로또 전략에 대해 제시했다. 또 Joe(1993)가 로또 데

이터로 항등성 테스트를 하기 위한 방법을 제시했으며, Johnson(1993)이 미국 데이터로 인기 있는 숫자(Hot number)의 추정과 항등성 테스트(uniformity test)를 했다.

본 논문에서는 한국 로또의 소개와 최근까지의 데이터로 여러 가지 항등성 검정(uniformity test)을 해 보고 Johnson(1993)이 제안한 모델에 따라 각 번호의 확률을 추정해 보았다.

제 2장에서는 한국의 로또를 소개하고 한국 로또데이터로 항등성 검정과 Johnson(1993)이 제안한 모델로 각 번호에 대한 확률과 점근분산을 알아보았다.

2. 로또 데이터의 분석

로또 데이터를 분석하기 위해 다음과 같이 가정한다.

1) 1에서 K개까지의 공(ball)이 있다

2) 각 회차에 선택되는 공(ball)의 개수는 m개($1 \le m \le K$)이다.

3) 각 공(ball)의 확률을 p_k로 둔다. 여기서 $\sum_{k=1}^{K} p_k = 1$이 된다.

4) 매 회 차마다 선택된 공(ball)은 독립이다.

2.1 항등성 검정

최근 로또의 엄청난 상금으로 인해 로또 애호가들이 늘고 있고 로또에 대한 많은 가설이 나오고 있다. 이 절에서는 로또의 많은 가설 중에서 로또 번호에 대한 항등성을 검정해 보도록 하겠다.

첫째 각 번호에 대한 항등성 검정으로, 카이제곱 적합도 검정의 검정 통계량은

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \qquad (2.1)$$

여기서 $E_i = n/K$ 이다.

이 통계량의 귀무가설은 $H_0 : p_k = 1/K \ (k = 1, 2, \cdots, K)$이다.

둘째, Joe(1993)가 제안한 선택된 숫자의 집합에 대한 항등성 검정으로 점근 카이제곱 적합도 검정의 검정 통계량은

$$\chi^2 = \frac{K-1}{K-m} \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \qquad (2.2)$$

여기서 $E_i = nm/K$ 이다.

이 통계량의 귀무가설은 $H_0 : p_k = m/K \ (k = 1, 2, \cdots, K)$이다.

위에서 제시한 검정 통계량으로 한국 로또의 99회까지 데이터로 항등성 검정을 해 보았다. 한국 로또의 99회까지 1등 데이터는 표<2-1>과 같고, 항등성 검정의 결과는 표<2-2>와 같다.

표<2-1> 한국 로또 데이터

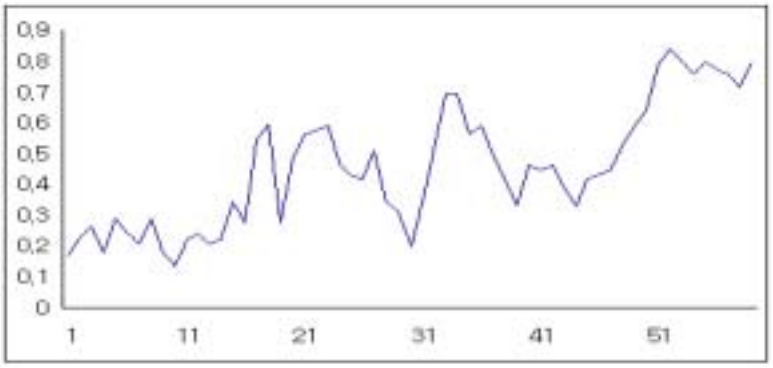
회차	1등 숫자	회차	1등 숫자	회차	1등 숫자
1회	10,23,29,33,37,40	34회	09,26,35,37,40,42	67회	03,07,10,15,36,38
2회	09,13,21,25,32,42	35회	02,03,11,26,37,43	68회	10,12,15,16,26,39
3회	11,16,19,21,27,31	36회	01,10,23,26,28,40	69회	05,08,14,15,19,39
4회	14,27,30,31,40,42	37회	07,20,30,33,35,37	70회	05,1922,25,28,43
5회	16,24,29,40,41,42	38회	16,17,22,30,37,43	71회	05,09,12,16,29,41
6회	14,15,26,27,40,42	39회	06,07,13,15,21,43	72회	02,04,11,17,26,27
7회	02,09,16,25,26,40	40회	07,13,18,19,25,26	73회	03,12,18,32,40,43
8회	08,19,25,34,37,39	41회	13,20,23,35,38,43	74회	06,15,17,18,35,40
9회	02,04,,16,17,36,39	42회	17,18,19,21,23,32	75회	02,05,24,32,34,44
10회	09,25,30,33,41,44	43회	06,31,35,38,39,44	76회	01,03,15,22,25,37
11회	01,07,36,37,41,42	44회	03,11,21,30,38,45	77회	02,18,29,32,43,44
12회	02,11,21,25,39,45	45회	01,10,20,27,33,35	78회	10,13,25,29,33,35
13회	22,23,25,37,38,42	46회	08,13,15,23,31,38	79회	03,12,24,27,30,32
14회	02,06,12,31,33,40	47회	14,17,26,31,36,45	80회	17,18,24,25,30,32
15회	03,04,16,30,31,37	48회	06,10,18,26,37,38	81회	05,07,11,13,20,33
16회	06,07,24,37,38,40	49회	04,07,16,1933,40	82회	01,02,03,14,27,42
17회	03,04,09,17,32,37	50회	02,10,12,15,22,44	83회	06,10,15,17,19,34
18회	03,12,13,19,32,35	51회	02,03,11,16,26,44	84회	16,23,27,34,42,45
19회	06,30,38,39,40,43	52회	02,04,15,16,20,29	85회	06,08,13,23,31,36
20회	10,14,18,20,23,30	53회	07,08,14,32,33,39	86회	02,12,37,39,41,45
21회	06,12,17,18,33,42	54회	01,08,21,27,36,39	87회	04,12,16,23,34,43
22회	04,05,06,08,17,39	55회	17,21,31,37,40,44	88회	01,17,20,24,30,41
23회	05,13,17,18,33,42	56회	10,14,30,31,33,37	89회	04,26,28,29,33,40
24회	07,08,27,29,36,43	57회	07,10,16,2529,44	90회	17,20,29,35,38,44
25회	02,04,21,26,43,44	58회	10,24,25,33,40,44	91회	01,21,24,26,29,42
26회	04,05,07,18,20,25	59회	06,29,36,39,41,45	92회	03,14,24,33,35,36
27회	01,20,26,28,37,43	60회	02,08,25,36,39,42	93회	06,22,24,36,38,44
28회	09,18,23,25,35,37	61회	14,15,19,30,38,43	94회	05,32,34,40,41,45
29회	01,05,13,34,39,40	62회	03,08,15,27,29,35	95회	08,17,27,31,34,43
30회	08,17,20,35,36,44	63회	03,2023,36,38,40	96회	01,03,08,21,22,31
31회	07,09,18,23,28,35	64회	14,15,18,21,26,36	97회	06,07,14,15,20,36
32회	06,14,19,25,34,44	65회	04,25,33,36,40,43	98회	06,09,16,23,34,32
33회	04,07,32,33,40,42	66회	02,03,07,17,22,24	99회	01,03,10,27,29,37

표<2-2> 항등성 검정 결과

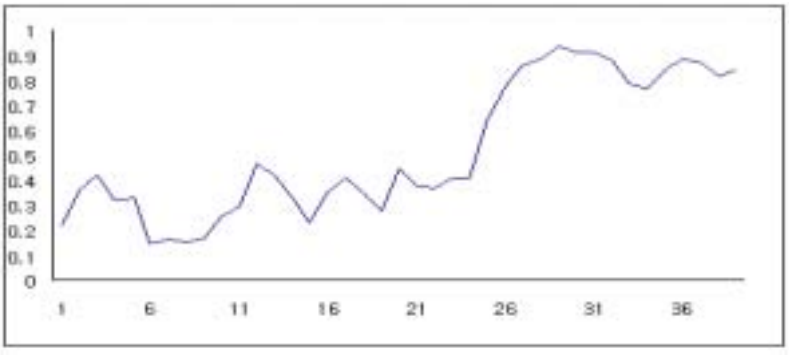
	자유도	통계량	p 값
카이제곱 적합도 검정	44	36.6061	0.7779
점근 카이제곱 적합도 검정	44	41.2992	0.5880

두 검정 모두 귀무가설을 채택하여 각 번호들이 동일하게 채택된다고 할 수 있으며 선택된 집합에 대해서도 동일하다고 할 수 있다.

그렇다면 앞에서 제시한 카이제곱 적합도 검정을 1회부터 40회까지, 2회부터 41회까지 ,…, 각40회로 묶고 1회부터 60회까지, 2회부터 61회까지 ,…, 각60회로 묶어서 카이제곱 적합도 검정을 해 보았다. 그리고 각 경우의 p값을 그래프로 그려보면 다음 그래프<2-1>과 그래프<2-2>와 같다.



그래프<2-1> 40회 단위별 항등성 검정을 위한 p값



그래프<2-2> 60회 단위별 항등성 검정을 위한 p값

두 그래프를 보면 60회로 묶었을 때의 p값의 변화가 40회로 묶었을 때의 p값의 변화가 더 자연스럽게 증가하고 있는 것을 알 수 있고 두 그래프 모두 점차적으로 p값이 증가하는 경향을 보이고 있다. 즉 회수가 거듭 할수록 각 번호는 동일하게 채택됨을 말해주고 있다.

그 밖에도 번호를 7개씩 7구간으로 나누었을 때에 각 구간마다 차이가 있는지, 색에 따른 차이가 있는지 알아보았다.

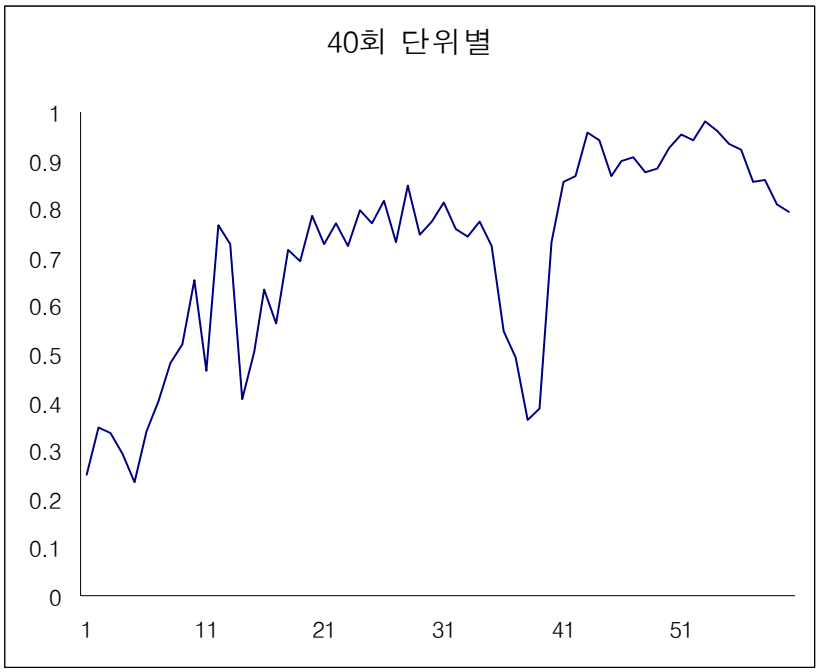
표<2-3> 7개씩 나누었을 때 항등성 분석

구간	빈도	확률	통계량	p값
1~7	100	16.83502	0.107534	0.999975
8~14	77	12.96296		
15~21	97	16.32997		
22~28	88	14.81481		
29~35	92	15.48822		
36~42	104	17.50842		
42~45	36	6.060606		
합 계	594	100		

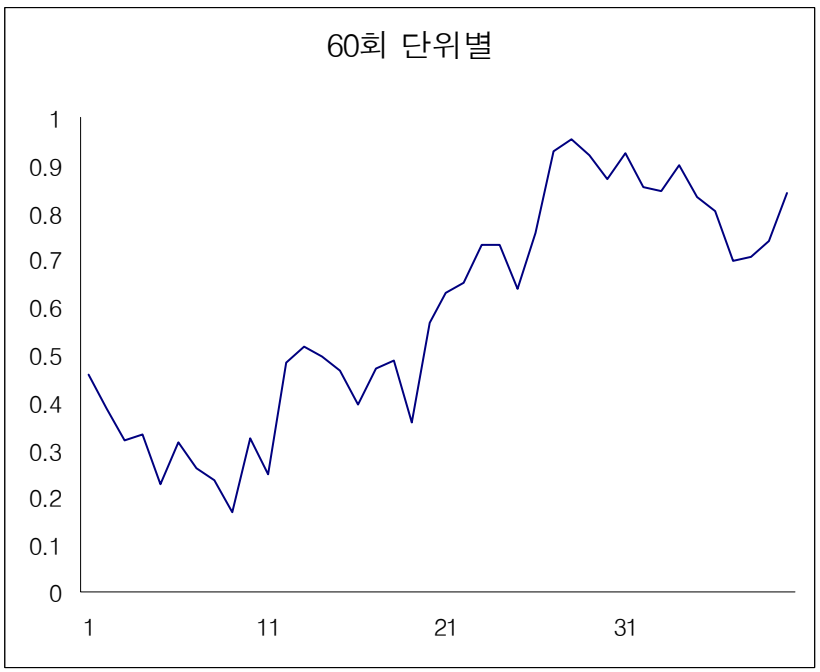
표<2-4> 색에 따른 항등성 분석

색	빈도	확률	통계량	p값
빨강	136	22.89562	0.440896	0.978994
주황	126	21.21212		
노랑	127	21.38047		
파랑	148	24.91582		
초록	57	9.59596		
합계	594	100		

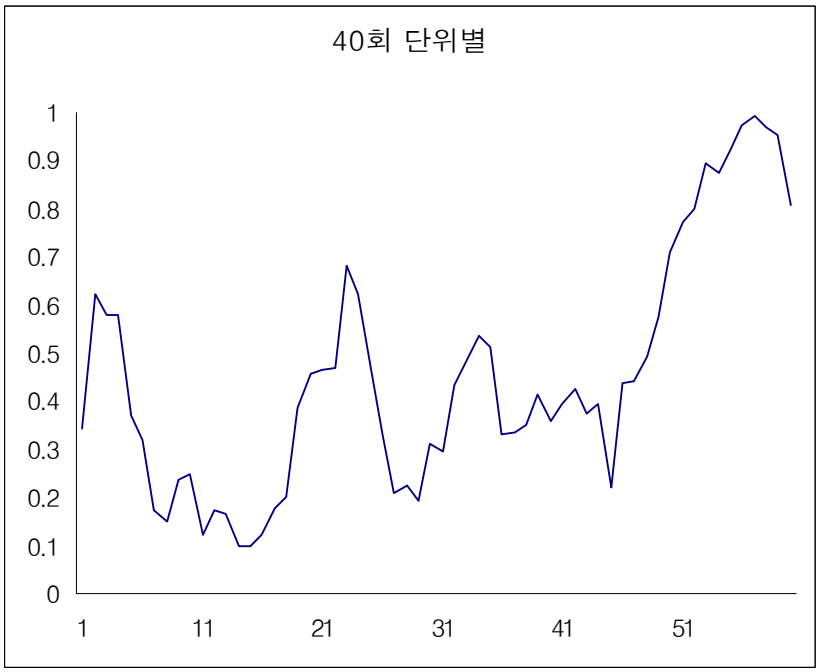
두 결과 각 공이 동일하게 선택된다는 것을 지지하고 있음을 알 수 있다.
7개씩 7구간으로 나누었을 때 1회부터 40회, 2회부터 41회 ,…, 각 40회마다 p값의 변화량은 그래프<2-3>와 같고 또 1회부터 60회, 2회부터 61회 60회마다 p값의 변화량은 그래프<2-4>와 같다. 이와 같은 방법으로 색에 따른 각 40회마다의 p값의 변화량은 그래프<2-5>와 같고 60회마다의 p값의 변화량은 그래프<2-6>와 같다.



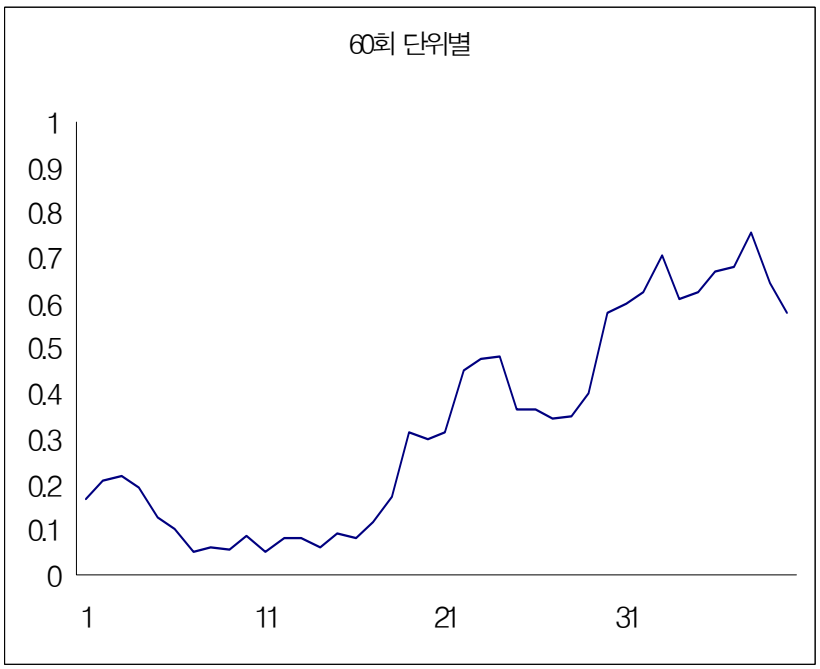
그래프<2-3> 7개씩 구간별 항등성 검정을 위한 p값



그래프<2-4> 7개씩 구간별 항등성 검정을 위한 p값



그래프<2-5> 색에 따른 항등성 검정을 위한 p값



그래프<2-6> 색에 따른 항등성 검정을 위한 p값

숫자를 7개씩 7구간으로 나눈 그래프의 변화와 색에 따른 그래프의 변화 모두 p값이 증가하는 경향을 보이고 있으며, 그래프<2-3>에 38회 근처에서 p값이 떨어지는 경향을 보이는데 이것은 어느 한 구간에서 선택된 공이 많아서 p값이 작게 나타난 것으로 추측된다.

2.2 각 공의 선택될 확률의 추정

이 절에서 m개의 공(ball)의 집합을 $X_i = (s_{i1}, s_{i2}, \cdots, s_{im})$라고 두고 각 볼(ball)이 선택될 확률인 모수 p 를 추정해 보자. 먼저 로또 데이터에 대한 함수 몇 가지를 소개하겠다. 앞으로 쓰게 될 각 회 차(X_i)에 1등 번호($s_{i1} < s_{i2} < \cdots < s_{ir}$)는 오름차순으로 정렬한다.

첫 번째로 소개할 함수는 각 회 차의 정렬된 1등 데이터를 r개씩 나누었을 때 빈도수를 세는 함수이다.

그 함수를 식으로 표현하면

$$n_r(s_1, s_2, \cdots, s_r) = \sum_{i=1}^{n} \prod_{j=1}^{r} I[X_{ij} = s_J] \tag{2.3}$$

여기서 $I[A]$는 지시함수로 만약 A 가 참이면 1 그 외에는 0 이다.

그 다음으로 소개할 함수는 모든 회 차의 각 숫자의 빈도수를 세는 함수이다.

그 함수를 식으로 표현하면

$$n_0(s) = \sum_{i=1}^{n} \sum_{j=1}^{r} I[X_{ij} = s_J] \tag{2.4}$$

이다.

이제 결합 확률 분포함수에 대해 알아보자. 한 회 차의 결합 확률 분포함수를 식으로 나타내면 다음과 같다.

$$P[X_i = (x_{i1}, x_{i2}, \cdots, x_{im}) = f_m(x_{i1}, x_{i2}, \cdots, x_{im} | p)]$$

$$= P[X_{i1} = x_{i1}] \prod_{j=2}^{m} P[X_{ij} = x_{ij}, X_{i1} = x_{i1}, \cdots, X_{i(j-1)} = x_{i(j-1)}]$$

$$= p_{x_{i1}} \frac{p_{x_{i2}}}{(1-p_{x_{i1}})} \frac{p_{x_{i3}}}{(1-p_{x_{i1}}-p_{x_{i2}})} \cdots \times \frac{p_{x_{im}}}{(1-p_{x_{i1}}-\cdots-p_{x_{i(m-1)}})} \quad (2.5)$$

각 회 차에 뽑혀지는 공은 독립이므로 결합확률 분포함수를 다음과 같이 표현 할 수 있다.

$$P[X = x] = \prod_{i=1}^{n} f_m(x_{i1,} x_{i2,} \cdots, x_{im} | p) \quad (2.6)$$

식 (2.6)의우도함수를 식으로 표현하면

$$\begin{aligned} L_n(p) = & \sum_{s_o=1}^{K} n_0(s_0) \ln(p_{s_0}) - \sum_{s_1=1}^{K} n_1(s_1) \\ & \ln(1-p_{s_0}) - \sum_{1 \le s_1} \sum_{s_2 \le k} n_2(s_1, s_2) \ln(1-p_{s_1}-p_{s_2}) \cdots \\ & - \sum_{1 \le s_1} \sum_{s_2 < \cdots < s} \sum_{m-1 \le k} n_2(s_1, s_2) \ln(1-p_{s_1}-\cdots-p_{s_{m-1}}) \\ = & \sum_{s_o=1}^{K} n_0(s_0) \ln(p_{s_0}) - \sum_{r=1}^{m-1} \sum_{s \in R_r} n_{r(s)} \ln(1-p_{s_1}-\cdots-p_{s_r}) \end{aligned} \quad (2.7)$$

최우 추정치를 구하기 위해서 로그우도함수에 로짓 모수 $\theta = (\theta_1, \theta_2, \cdots, \theta_{k-1})$를 쓰면 식을 간단하게 표현 할 수 있다.

$$\begin{aligned}
\mathrm{Ln}(\theta) &= \sum_{s_o=1}^{n} n_0(s_0)\ln\left(\frac{e^{\theta_{s_0}}}{1+\sum_{j=1}^{n-1} e^{\theta_j}}\right) \\
&\quad - \sum_{r=1}^{m-1}\sum_{s\in R_r} n_{r(s)}\ln\left(1-\frac{e^{\theta_{s_1}}}{1+\sum_{j=1}^{k-1} e^{\theta_j}}-\cdots-\frac{e^{\theta_{s_r}}}{1+\sum_{j=1}^{k-1} e^{\theta_j}}\right) \\
&= \sum_{s_o=1}^{k} n_0(s_0)\theta_{s_0} - n\ln\left(1+\sum_{j=1}^{k-1} e^{\theta_j}\right) \\
&\quad - \sum_{r=1}^{m-1}\sum_{s\in R_r} n_{r(s)}\ln\left(1+\sum_{j=1}^{k-1} e^{\theta_j}-e^{\theta_{s_0}}-e^{\theta_{s_1}}-\cdots-e^{\theta_{s_r}}\right)
\end{aligned} \tag{2.8}$$

로그우도함수 식의 최우 추정치(MLE)를 구하기 위해 Newton -Raphson 방법을 사용한다.

이 모델에서 Newton-Raphson 방법에 따른 형식은

$$-\frac{\partial \mathrm{Ln}(\theta^n)}{\partial\theta} = \frac{\partial^2 \mathrm{Ln}(\theta^n)}{\partial\theta\partial\theta^T}\ u^n \tag{2.9}$$

여기서 $u^n = \theta^{n+1} - \theta^n$ 이고, $\frac{\partial^2 \mathrm{Ln}(\theta^n)}{\partial\theta\partial\theta^T}$ 은 헤시안 행렬 (Hessian matrix)이다.

Newton-Raphson 방법에 의해 $\partial L_n(\theta)/\partial\theta = 0$ 로 풀어보면

$$\frac{\partial \mathrm{Ln}(\theta)}{\partial\theta_i} = n_0(i) - np_i - \sum_{r=1}^{m-1}\sum_{s\in R_r}\left(\frac{p_i}{1-p_{s_1}-\cdots--p_{s_r}}\right) \tag{2.10}$$

(K-1) ×(K-1)인 헤시안 행렬(Hessian matrix)의 대각원소는

$$\begin{aligned}
\frac{\partial^2 \mathrm{Ln}(\theta)}{\partial\ \theta_i^2} &= np_i(1-p_i) + \sum_{r=1}^{m-1}\sum_{s\in R_r} n_r(s)\times \\
&\quad \left(\frac{p_i}{1-p_{s_1}-\cdots--p_{s_r}} - \left(\frac{p_i}{1-p_{s_1}-\cdots--p_{s_r}}\right)^2\right)
\end{aligned} \tag{2.11}$$

또 비대각원소는

$$\frac{\partial^2 \mathrm{Ln}(\theta)}{\partial \theta_i \partial \theta_j} = -n p_i p_j + \sum_{r=1}^{m-1} \sum_{s \in R_r} n_r(s) \left(\frac{p_i p_j}{(1 - p_{s_1} - \cdots - - p_{s_r})^2} \right) \quad (2.12)$$

이다.

여기서 Newton-Raphson 방법에 의해 유일한 해를 구할 수 있다.

또 정칙조건하에서 최대우도방정식의 해는 $\widehat{\theta}_n$은 점근평균이 θ 이고 점근분산이 $1/nE(\frac{\partial}{\partial \theta} \ln f(X, \theta)^2)$인 정규분포를 근사적으로 따른다는 성질을 가진다.

따라서 우리가 관심이 있는 모수 p에 대한 분산은

$$I_n^{-1}(\widehat{p}) = \left(\frac{\partial p(\theta)}{\partial \theta^T} \right) I_n^{-1}(\theta) \left(\frac{\partial p(\theta)}{\partial \theta^T} \right) \quad (2.13)$$

여기서 $I_n^{-1}(\theta) = \left(- \frac{\partial^2 \mathrm{Ln}(\widehat{\theta})}{\partial \theta \partial \theta^T} \right)^{-1}$ 이고

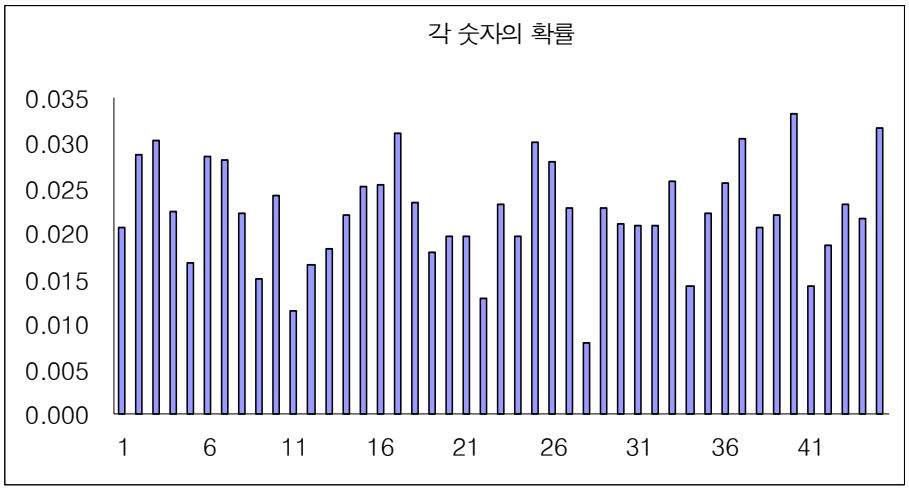
$$\frac{\partial p(\theta)}{\partial \theta^T} = \begin{bmatrix} \frac{e^{\theta_1}\left(1 + \sum_{j=2}^{k-1} e^{\theta_j}\right)}{\left(1 + \sum_{j=1}^{k-1} e^{\theta_j}\right)^2} & \cdots & \cdots & \frac{-e^{\theta_1} e^{\theta_{k-1}}}{\left(1 + \sum_{j=2}^{k-1} e^{\theta_j}\right)^2} \\ \vdots & & \ddots & \vdots \\ \frac{-e^{\theta_{k-1}} e^{\theta_1}}{\left(1 + \sum_{j=2}^{k-1} e^{\theta_j}\right)^2} & \cdots & \cdots & \frac{e^{\theta_1}\left(1 + \sum_{j=1}^{k-2} e^{\theta_j}\right)}{\left(1 + \sum_{j=1}^{k-1} e^{\theta_j}\right)^2} \\ \frac{-e^{\theta_1}}{\left(1 + \sum_{j=1}^{k-1} e^{\theta_j}\right)^2} & \cdots & \cdots & \frac{-e^{\theta_{k-1}}}{\left(1 + \sum_{j=1}^{k-1} e^{\theta_j}\right)^2} \end{bmatrix} \quad (2.14)$$

이다.

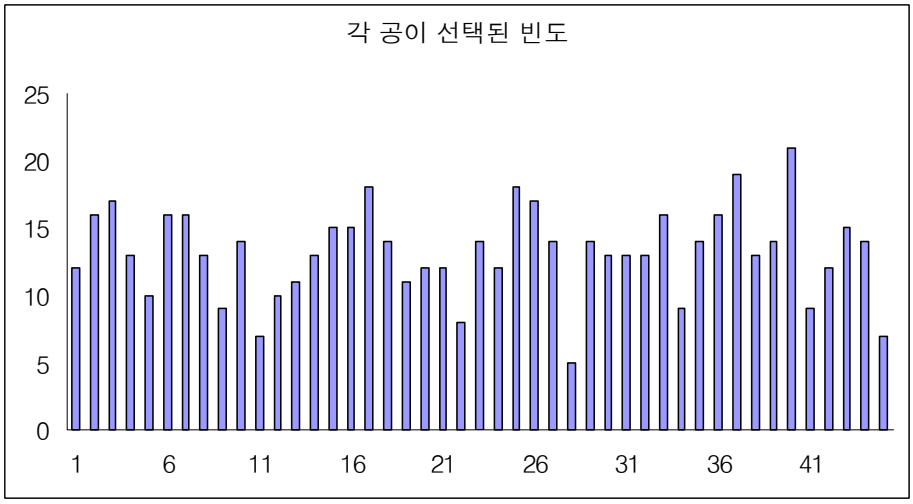
앞에서 제시한 Richard Johnson의 모델을 따라 한국 로또의 각 번호에 대한 확률 p 와 그 숫자의 점근 분산을 알아보겠다. 데이터는 표<2-1>와 같이 99회까지 데이터를 사용했고, 포트란 프로그램으로 데이터를 정리한 후 식 (2.10), 식 (2.11), 식 (2.12)에 따라 각 번호에 대한 확률 p 값을 추정했고 식 (2.14)에 따라 점근 분산을 추정했다. 여기서 초기값은 1로 두고 Newton-Raphson 방법 의해 MLE를 추정했다. 그 결과는 다음 표<2-5>와 같다.

표<2-5> 각 숫자에 대한 확률과 점근 분산

i	$\hat{p}_i$	$\hat{\sigma}(\hat{p}_i)$	i	$\hat{p}_i$	$\hat{\sigma}(\hat{p}_i)$
1	0.02068	0.00003508	24	0.01966	0.00003173
2	0.02865	0.00005018	25	0.03001	0.00004890
3	0.03034	0.00005289	26	0.02792	0.00004489
4	0.02239	0.00003789	27	0.02290	0.00003681
5	0.01678	0.00002779	28	0.00784	0.00001222
6	0.02854	0.00004980	29	0.02273	0.00003628
7	0.02806	0.00004816	30	0.02106	0.00003359
8	0.02214	0.00003706	31	0.02088	0.00003301
9	0.01491	0.00002444	32	0.02076	0.00003263
10	0.02417	0.00004094	33	0.02580	0.00004081
11	0.01131	0.00001812	34	0.01424	0.00002229
12	0.01652	0.00002696	35	0.02228	0.00003486
13	0.01824	0.00002983	36	0.02566	0.00004036
14	0.02210	0.00003693	37	0.03041	0.00004758
15	0.02515	0.00004135	38	0.02057	0.00003207
16	0.02532	0.00004192	39	0.02204	0.00003412
17	0.03111	0.00005249	40	0.03325	0.00005134
18	0.02337	0.00003833	41	0.01406	0.00002175
19	0.01788	0.00002868	42	0.01858	0.00002838
20	0.01962	0.00003160	43	0.02327	0.00003548
21	0.01966	0.00003172	44	0.02164	0.00003292
22	0.01281	0.00002033	45	0.03158	0.00031133
23	0.02311	0.00003747			



그래프<2-7> 각 숫자의 확률



그래프<2-8> 99회까지 데이터에 대한 각 숫자의 빈도

각 번호의 확률은 보면 최고값은 40번 공으로 0.03325이고, 최저값은 9번 공으로 0.00784이며, 두 값의 차이는 0.02541이다. 또 Johnson이 제안한 모델에 의해 추정된 각 숫자에서 확률이 높은 수의 집합과 확률이 낮은 수의 집합을 구해보면 표<2-6>과 같고 확률이 높은 집합과 낮은 집합은 약 300배의 차이가 나는 것을 볼 수 있다. 또 99회까지 데이터에 따른 빈도수를 살펴보면 숫자 40이 21회

로 가장 많고 다음으로는 숫자 37이 19회 17이 18회 나타났으며, 빈도가 낮은 숫자를 보면 숫자 28이 5회로 가장 적게 나타났으며, 그 다음으로는 숫자 45와 숫자 11이 7회 나타났다. 빈도가 많은 것과 적은 수를 정리하면 표<2-7>와 같다.

표<2-6> 확률이 높은 수의 집합과 확률이 수의 낮은 집합

X	P(6)
{40,45,17,37,03,25}	6.51×10^{-7}
{09,34,41,22,11,28}	2.44×10^{-9}

표<2-7> 빈도수가 많은 숫자와 빈도수가 적은 숫자

빈도수가 높은 수	{40,37,17,03,26,02,06,07,33,36}
빈도수가 낮은 수	{28,45,11,22,41,34,09}

표<2-6>에서 보는 바와 같이 확률이 높은 집합은 일반적인 집합에 비해 약 2배가 높으며 확률이 낮은 집합은 200배 낮은 것을 볼 수 있다. 표<2-6>과 표<2-7>를 비교해 보면 확률이 낮은 집합과 빈도가 적은 숫자는 비슷한데 비해 확률이 높은 집합과 빈도가 많은 숫자는 차이가 있는 것을 볼 수 있다. 이 차이는 추정 오차에 대한 것 일수도 있고, 혹은 한국 로또가 항등성을 따른다는 가정에 의한 것 일수도 있다.

3. 맺음말

본 논문에서는 한국 로또에서 선택된 번호에 대해 분석해 보았다. 많은 국민들의 관심으로 어떤 경우 1등에 당첨된 인원이 지나치게 많거나 전혀 당첨자가 없는 경우가 나옴으로써 랜덤성에 대한 이해 부족으로 로또 추첨에 대한 공정성에 대해 의심을 하기도 했다. 따라서 2장에는 지난 99회차 까지 선택된 공의 번호에 대한 항등성을 검정해 보았다. 그 결과 로또에서의 번호추출에 있어 조작이 의심되지 않는 항등성이 만족됨을 알았다. 아울러 각 번호의 공이 선택될 확률을 최우추정법을 사용하여 구해 보았다. 이를 근거로 선택될 공의 확률이 가장 높은 조합과 가장 낮은 조합을 구해 보았으며 실제 얻어진 경우와 비교해 보면 낮은 경우는 어느 정도 맞아들어 가지만 높은 경우는 실제와 차이를 보이기도 했다. 전체적으로는 항등성을 통한 공정함에 대한 결론을 내릴 수가 있었다. 로또의 번호 선정에 대해 온갖 방법이 다 동원되기도 하고, 자동으로 선정하는가 혹은 본인의 선정에 따르는가에 대한 관심도 기대하지만 랜덤성을 기초하여 어떤 방법이든 동일하다는 결론을 내릴 수 있었다.

참고문헌

1. Joe, H. (1993). Tests of Uniformity for Sets of Lotto Numbers, *Statistics and Probability Letters*, 16, 181-188

2. John, H. (1997). The Statistics of the National Lottery, *Journal of the American Statistical Association*, 160, 187-206

3. Johnson, R. L. and J. Klotz (1993). Estimatng Hot Numbers Testing Uniformity for the Lottery, *Journal of the American Statistical Association*, 88, 662-668

4. Rockafellar, R. T. (1970). Convex Analysis, Princeton, NT: Princeton University Press.

5. Stern, H, and Cover, T. M. (1989). Maximum Entropy and the Lottery, *Journal of the American Statistical Association*, 84, 980-985

Testing Uniformity and Estimating for the Lottery[1)]

Park Mi Jung

Department of Statistics
Graduate School, Kyungpook National University
Daegu, Korea
(Supervised by Professor Joong Kweon Sohn)

(Abstract)

We consider testing uniformity for selected ball and sets of lotto numbers. We estimated the probabilities of selected balls in the lottery. The probabilities were estimated by using the model Johnson suggested.

A thesis submitted to the committe of Gradute School of Kyungpook National University in partial fulfilment od the requirements for the degree of Master of Science in December, 2004