The Korean Communications
in Statistics Vol. 11, No. 2, 2004
pp. 413-418

Are There Hot Numbers in the Lotto Korean Lottery?[1)]

Ji-Hyun Kim[2)]

Abstract

Statistically illiterate people seem to believe that there are some strategies for choosing winning numbers in lottery. One seemingly plausible strategy is to select the hot numbers which most frequently appeared in the past. In this article we investigate the existence of hot numbers in the Korean national lottery called Lotto. A numerical method is proposed to estimate the exact sampling distribution of test statistic for checking the existence of hot numbers among 45 possible numbers of choice.

Keywords : sampling without replacement, extreme values, exact distribution, uniformity test.

1. 서 론

로또복권의 열풍과 함께 당첨확률을 높일 수 있는 전략에 대해 일반인들의 관심이 높다. 만약 45개의 번호 중에서 매주 독립적으로 6개의 서로 다른 번호를 같은 확률로 추출한다면, 남들이 선택하지 않는 번호를 의도적으로 선택함으로써 당첨금액을 높일 수는 있겠지만 당첨확률을 높일 수 있는 전략은 있을 수 없다. 그럼에도 불구하고 성공사례가 소개되기도 하고 전략을 소개하는 책자도 서점에서 볼 수 있다. 심지어 자동적으로 확률이 높은 당첨번호를 생성시킨다는 소프트웨어까지 나오고 있다. 당첨확률을 높일 수 있다고 하는 전략 중에서 그나마 그럴듯해 보이는 것은 지금까지 나온 당첨번호를 분석해서 확률이 유난히 높은 번호, 이른바 "족보 번호"들을 선택한다는 것이다. 신문보도에 의하면 2004년 2월 28일 제65회 로또복권 1등 당첨자는 실제로 이 전략을 써서 당첨되었다고 한다. 그렇다면 과연 이러한 번호들이 존재하는 것일까?

TV에서 보면 45개의 공이 든 기구에 바람을 불어넣어 좁은 구멍으로 공이 하나씩 나오도록 하는 추출방법을 쓰고 있다. 공을 제조하는 기계적인 과정에서 선택이 잘 되는 공이 만들어질 수도 있고, 공을 넣고 섞는 과정에 문제가 있을 수도 있으며(Johnson & Klotz (1993) 5절 참조), 미처 생각하지 못한 다른 이유로 인해 특별한 행운의 번호가 있을 수 있다. 반대로 각 번호의 선택확률이 같다고 하더라도 실제 선택횟수에는 차이가 있기 마련이고 따라서 특별히 선택 확률이 높은 번호가 존재하지 않음에도 불구하고 제일 많이 나타난 번호는 특별한 번호인 것처럼 보일 수도 있다. 문제는 '우연에 의한 것이라고 보기 어려울 만큼 빈도가 높은가'이다. 본 연구에서는 과연 행운의 번호가 존재하는

1) This work was supported by the Soongsil University Research Fund.
2) Professor, Dept. of Statistics, Soongsil University, Dongjak-Ku Sangdo-Dong, Seoul 156-743, KOREA.
E-mail: jhkim@stat.soongsil.ac.kr
The author thanks two referees for their helpful comments.

것인지를 통계적으로 검정하고자 한다.

2. 검정 방법

로또복권에서는 매주 1회 1부터 45까지의 번호 중에서 6개의 당첨번호를 비복원추출로 결정한다. 이 때 제일 처음 추출하는 번호의 확률분포를 $p=(p_1,\cdots,p_{45})$라고 하자. 즉 처음 추출하는 번호가 i일 확률은 p_i이며, 두 번째 추출하는 번호가 j일 확률은

$$\frac{p_j}{1-p_i},\ j\neq i$$

이다. 마찬가지로 6번째 추출번호의 확률은 처음 다섯 개의 번호를 제외한 나머지 번호들 중에서 처음 확률분포에 비례한 값을 갖는다고 하자. 그리고 매회 독립적으로 6개의 번호를 앞에서 설명한 대로 비복원추출한다고 가정하자. 로또복권 주관자는 $p_0=(1/45,\cdots,1/45)$을 목표로 할 것이며, 우리가 검정하고 싶은 '특별히 선택 확률이 높은 번호가 존재하지 않는다'는 영가설(null hypothesis)도 $H_0: p=p_0$ 로 표현할 수 있다.

$(X_1,\cdots,X_{45})$를 n회 동안 매회 6개씩 선택된 45개 번호의 빈도를 각각 나타낸다고 하면 가장 많이 선택된 번호의 빈도는 $\max_{1\le i\le 45}X_i$ 이다. 이 값의 크기로 영가설을 기각할 수 있는지를 판단할 수 있을 것이다. 영가설 하에서 $X_{(45)}=\max_{1\le i\le 45}X_i$ 의 임계값(critical value)을 정하기 위한 준거분포(reference distribution)는 어떤 분포가 되어야 할까를 생각해보자. 먼저 확률벡터 $(X_1,\cdots,X_{45})$에 대한 분포로, 시행횟수가 $6n$이고 가능한 45개 번호의 출현 확률이 $p_0=(1/45,\cdots,1/45)$ 로서 균일한 다항분포를 생각해볼 수 있겠으나 정확한 분포는 아니다. 왜냐하면 영가설 하에서 n번의 시행은 독립이지만 매회 선택되는 6개의 번호는 서로 중복되지 않아야 하기 때문에 독립적이지 않으며 같은 확률 p_0가 아니기 때문이다.

$(X_1,\cdots,X_{45})$의 정확한 분포함수는 비복원추출로 인해 매우 복잡하므로, 영가설 하에서 $X_{(45)}=\max_{1\le i\le 45}X_i$ 의 분포를 $(X_1,\cdots,X_{45})$의 정확한 분포함수로부터 유도하기는 어렵다. 만약 $(X_1,\cdots,X_{45})$의 분포를 시행횟수가 $6n$이고 45개 셀의 확률이 같은 다항분포로 가정할 수 있다면, David(1981)의 연습문제 5.3.7에서 언급된 Kozelka(1956)의 결과를 이용하여 $X_{(45)}$의 상위 α 분위수를 근사적으로 계산할 수 있다. $N=6n$이라고 할 때, X_i를 표준화한 변수를

$$Z_i=\frac{(X_i-\frac{N}{45})}{\sqrt{44N/45^2}} \tag{2.1}$$

라고 정의하자. 그러면 Φ를 표준정규분포의 누적분포함수라고 할 때

$$P(Z_{(45)}\ge z^*)\approx 45\,(1-\Phi(z^*)) \tag{2.2}$$

임을 Kozelka(1956)는 증명하였다. 따라서 $Z_{(45)}$의 분포에 대한 근사적 상위 α 분위수는

$$\Phi^{-1}(1-\frac{\alpha}{45}) \tag{2.3}$$

이 된다. 다항분포가 성립할 때 (2.2)나 (2.3)에 의한 근사가 보다 정확한 근사가 되기 위해서는

$$45 \times 44 \times [P(Z_i \geq z^*)]^2$$

가 0에 가까워야 하고, 동시에 Z_i의 근사적 분포로 정규분포를 쓸 수 있을 만큼 N이 충분히 커야 한다.

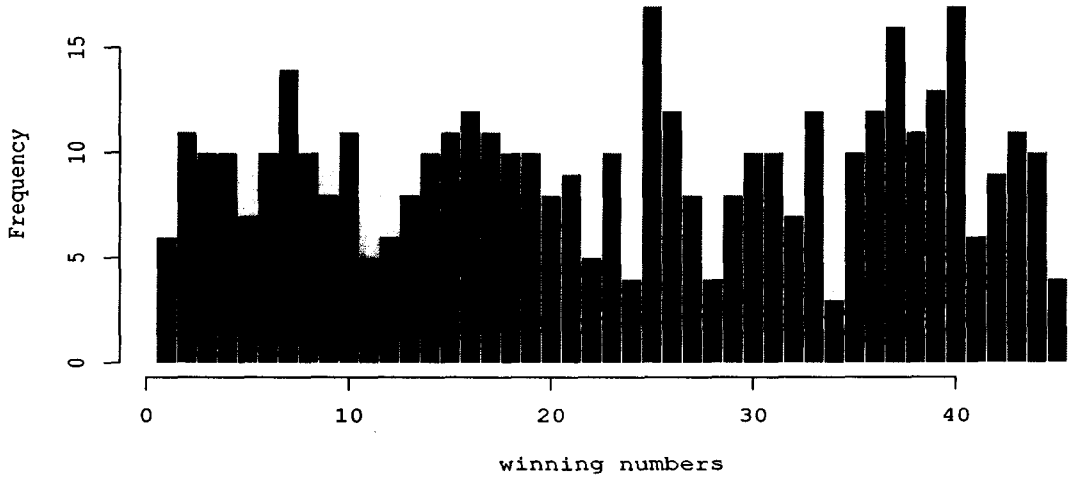
비복원추출로 인해 다항분포를 가정할 수 없으므로 앞의 결과를 이용하는 대신 컴퓨터를 이용한 수치적 방법으로 검정통계량의 표집분포(sampling distribution)와 임계값을 추정하고자 한다. 영가설 하의 실험을 반복하여 재현한 결과로 $X_{(45)}$의 표집분포를 추정하는 방법을 기술하면 다음과 같다.

(1) 1부터 45까지의 번호 중에서 비복원으로 6개의 번호를 랜덤하게 추출한다.
(2) 위 과정을 n번 반복하여 $(x_1, \cdots, x_{45})$와 $x_{(45)} = \max\ x_i$ 등을 계산한다.
(3) 위 (1)-(2) 단계를 B번 반복하여 구한 $x_{(45)}^{(1)}, \cdots, x_{(45)}^{(B)}$ 값으로 $X_{(45)}$의 정확한 분포를 추정할 수 있으며 이로부터 p값이나 분위수를 추정할 수 있다.

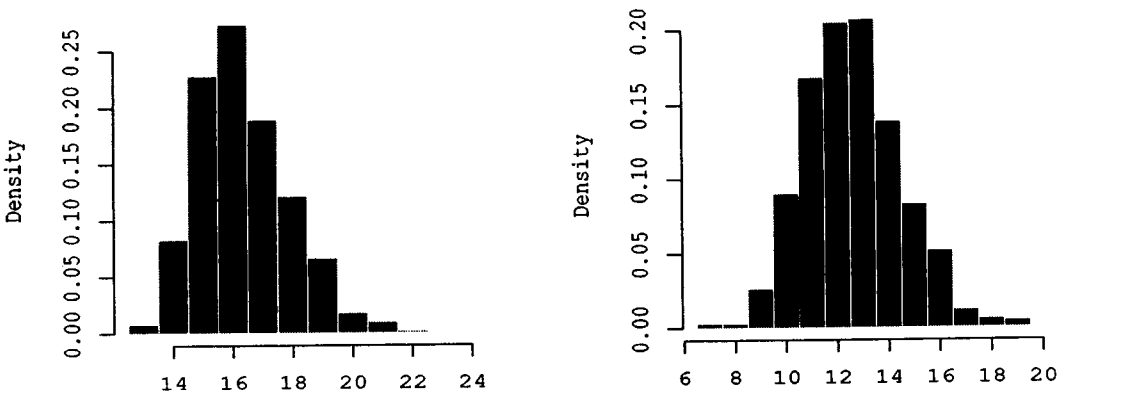
위 과정을 R로 구현하였으며 부록에 R 코드를 첨부하였다.

3. 실제 자료의 분석

2004년 4월 중순까지 71회에 걸쳐 실시된 로또복권의 당첨번호를 이용하여 행운의 번호가 존재하는지를 검정해 보자. 먼저 71회에 걸친 1등 당첨번호 426개의 빈도분포를 <그림 1>에서 히스토그램으로 나타내었다. 번호 25와 40이 17번으로 제일 많이 추출되었으며 번호 34가 3번으로 제일 적게 추출되었다. 영가설 하에서 $n = 71$ 인 실험을 $B = 1000$ 만큼 반복하여 추정한 $X_{(45)}$의 표집분포는 <그림 2> (a)와 같다.



<그림 1> 71회 동안 추출된 426개 당첨 번호의 빈도분포



(a) $X_{(45)}$의 표집분포 (b) $X_{(45)} - X_{(1)}$의 표집분포

<그림 2> 검정통계량의 표집분포: B=1,000

이 표집분포에서 p값은

$$\hat{P}(X_{(45)} \geq 17) = 0.408$$

로 추정되고, 검정통계량 $X_{(45)}$의 이산성을 고려한 중간 p값(mid p-value)은

$$\hat{P}(X_{(45)} \geq 18) + \frac{1}{2}\hat{P}(X_{(45)} \geq 17) = 0.296$$

으로 추정된다. 그리고 상위 0.05 분위수는

$$\hat{P}(X_{(45)} \geq 19) = 0.098, \quad \hat{P}(X_{(45)} \geq 20) = 0.032$$

이므로, 19로 추정된다. 위 결과가 얼마나 안정적인가를 알아보기 위해 같은 작업을 20번 반복하여 평균과 표준오차를 구해보았다. p값의 평균과 표준오차는 0.405와 0.0035이었고, 중간 p값의 평균과 표준오차는 0.301과 0.0028이었다. 그리고 상위 0.05 분위수는 예외 없이 19로서 안정된 결과를 보였다.

한편, 다항분포를 가정하여 구한 근사적 상위 0.05 분위수는 (2.1)과 (2.3)에 의해 18.8로 계산되는데, $X_{(45)}$가 정수값을 가짐을 고려할 때 19가 되어 다항분포를 가정하지 않고 수치적 방법으로 구한 결과와 일치한다. 그리고 $P(X_{(45)} \geq 17)$의 근사값은 (2.1)과 (2.2)에 의해 0.299이다.

검정통계량으로 $X_{(45)}$ 대신 $X_{(45)} - X_{(1)}$을 고려할 수도 있다. <그림 2> (b)는 이 검정통계량의 표집분포를 나타낸다. 표집분포로부터 구한 p값은 0.298이며, 중간 p값은 0.229이다. 검정통계량으로 $X_{(45)}$을 사용했을 때의 결과와 차이를 보이는 이유는 검정통계량의 이산성 때문으로 판단된다. 하지만 영가설을 기각할 만한 충분한 증거가 없다는 점에서는 두 통계량의 결과가 같다. 그리고 $X_{(45)}$와 $X_{(45)} - X_{(1)}$는 n이 커짐에 따라 같이 커지므로 식 (2.1)과 같이 표준화하여 분석할 수 있겠으나, n의 값이 주어지면 상수를 빼고 상수를 나누어 준 값이 되므로 표준화하지 않았을 때의 검정결과와 차이

가 없다. 오히려 표준화하지 않았을 때 더 직접적으로 해석할 수 있다는 장점이 있다.

4. 결론 및 토의

71회까지 진행된 로또복권의 당첨번호를 살펴보면 제일 많이 나온 번호의 빈도가 17, 제일 적게 나온 번호의 빈도가 3으로서 언뜻 행운의 번호가 존재하는 것처럼 보인다. 하지만 분석 결과 모든 번호가 동일한 확률을 가질 때에도 이 정도의 불균형은 충분히 있을 수 있다는 결론을 얻었다. 따라서 과거 당첨번호의 빈도를 근거로 번호를 선택하는 전략은 과학적 근거가 없다고 판단된다.

영가설을 검정하기 위해 45개의 당첨번호의 빈도 $(X_1, \cdots, X_{45})$를 이용하는 대신 다른 방법도 가능하다. 6개씩 n회 추출하는 과정에서, 매번 추출되는 번호를 Y라 할 때, 총 $6n$개의 관측값으로부터 Y의 분포가 균일분포(uniform distribution)를 따르는지를 검정할 수도 있다. Johnson & Klotz(1993)는 비복원추출을 감안하여 $\boldsymbol{p}$의 최대가능도추정량(MLE)을 얻는 방법을 제시하였으며, 이를 이용하여 $H_0 : \boldsymbol{p} = \boldsymbol{p}_0$를 검정하기 위한 방법으로 가능도비검정(likelihood ratio test)을 제안하였다. 본 연구에서 제안한 수치적 방법은 Johnson & Klotz(1993)의 방법과 달리 횟수 n이 크지 않아도 적용할 수 있으며 직접적으로 해석하기도 쉽다고 생각한다. 실행시간도 펜티엄4를 장착한 개인용 컴퓨터에서 $n = 71, B = 1000$의 실험에 불과 수초밖에 걸리지 않는다. Johnson & Klotz(1993) 방법과의 검정력 비교에 대한 연구는 본 연구의 주제가 아니므로 수행하지 않았다.

참고문헌

[1] David, H. A. (1981). *Order Statistics*, second edition, John Wiley & Sons, New York.

[2] Johnson, R. and Klotz, J. (1993). Estimating hot numbers and testing uniformity for the lottery, *Journal of the American Statistical Association*, Vol. 88, 662-668.

[3] Kozelka, R. M. (1956). Approximate upper percentile points for extreme values in multinomial sampling, *Annals of Mathematical Statistics*, Vol. 27, 507-512.

[2004년 4월 접수, 2004년 7월 채택]

부록: 두 검정통계량 $X_{(45)}$와 $X_{(45)} - X_{(1)}$의 표집분포와 p값 등을 추정하기 위한 R 코드

```
lottery <- function (B=1000, n.week=71, x.45=17, d=14, prob=0.95, hist.plot=TRUE) {
#   n.week회 동안 관측된 당첨번호 자료로부터 구한 X_(45)의 값을 x.45로 지정하고,
#   D = X_(45) - X_(1)의 값을 d로 지정해야 함
    X.45 <- numeric()
    D <- numeric()
    for (b in 1:B) {
        X <- rep(0,45) # 45개 당첨번호의 빈도를 나타내는 (X_1,..., X_45)를 초기화
        num <- numeric() # 6개 당첨번호를 초기화
        for (j in 1:n.week) {
            num <- sample(1:45,size=6)
            X[num] <- X[num] + 1
        }
        X.45[b] <- max(X)
        D[b] <- X.45[b] - min(X)
    }
    if (hist.plot) {
        oldpar <- par(mfrow=c(1,2))
        hist(X.45,breaks=((min(X.45):(max(X.45) + 1)) - 0.5),freq=FALSE,
                xlab="frequency of the hottest numbers",
                col="red",border="white",main=NULL)
        hist(D,breaks=((min(D):(max(D) + 1)) - 0.5),freq=FALSE,
                xlab="freq. of the hottest number - freq. of the coldest",
                col="red",border="white",main=NULL)
        par(oldpar)
    }
    pval1 <- length(X.45[X.45 >= x.45])/B # p값 계산
    midpval1 <- pval1 - 0.5*(length(X.45[X.45 == x.45])/B) # 중간 p값 계산
    q1 <- quantile(X.45,prob) # 상위 1-prob 분위수 계산
    pval2 <- length(D[D >= d])/B
    midpval2 <- pval2 - 0.5*(length(D[D == d])/B)
    q2 <- quantile(D,prob)
    return(list(pval.X.45 = pval1, midpval.X.45 = midpval1, q.X.45 = q1,
                pval.D = pval2, midpval.D = midpval2, q.D = q2))
}
lottery()
```