응용통계연구 (2010)
23(1), 189–196

유한 순서열의 임의화

허명회[1] · 이용구[2]

[1]고려대학교 통계학과, [2]중앙대학교 통계학과

(2009년 11월 접수, 2009년 12월 채택)

요 약

미국의 1970년 징병추출(draft lottery)은 유한 순서열 $(1, 2, \ldots, k)$의 물리적 임의화를 쉬운 일로 생각하였다가 사회적 물의가 빚어진 대표적인 사례이다. 본 소고는 숫자 1, 숫자 2,... 등의 순서로 쌓인 k장의 카드 뭉치를 물리적으로 임의화하는 데 있어 반복 시행(repeated trial)의 역할을 밝힌다. 부수적으로 독립시행 수 n, 성공의 확률이 θ인 이항분포 $B(n, \theta)$에서 성공 수가 짝수일 확률은 n이 커짐에 따라 0.5에 수렴하게 됨을 보인다.

주요용어: 추첨 상자, 물리적 임의화, 마코브 연쇄, 이항분포.

1. 배경과 목적

조재근 교수가 우리 글로 옮긴 라플라스(Pierre Simon Laplace, 1749–1827)의 <확률에 대한 철학적 시론>(1925, 제 5판)을 읽다가 다음 단락에 주목하게 되었다.

> "1부터 100까지 숫자를 순서대로 항아리에 넣고 잘 흔들어 섞은 다음 숫자 하나를 뽑는다고 하자. 숫자들이 잘 섞였다면 각 숫자가 뽑힐 확률은 모두 같을 것이다. 하지만 숫자를 항아리에 넣는 순서 때문에 숫자마다 뽑힐 확률에 차이가 생길 것 같다면, 이 숫자들을 뽑힌 순서에 따라 두 번째 항아리에 놓고 흔들어 섞으면 그 차이를 상당히 줄일 수 있다. 그 차이는 두 번째 항아리에서도 이미 피할 수 없는데, 세 번째, 네 번째 항아리를 계속 이용하면 줄일 수 있다." (조재근 역, 2009; 106쪽)

라플라스 책이 나온 지 근 150년 지나서 있은 1970년 미국의 징병 추첨에서 366장의 생일이 적힌 카드들이 뽑힌 순서가 처음 추첨 상자에 놓인 순서와 관련성을 보여 사회적 물의가 빚어진 바 있지 않은가? (Wikipedia, "Draft Lottery 1969", 2009/05/01 검색) 이 추첨에서 문제는 생일 카드들이 충분히 섞이지 않았기 때문인데 시대를 앞 선 라플라스의 예견에 놀라지 않을 수 없다.

<확률에 대한 철학적 시론>이 일반인들을 위해 "실상 수학을 쓰지 않고 확률과 통계학에 대해 설명한" 책이라서 그런지 (조재근 역, 2009), 라플라스는 앞의 인용문에 대한 수리적 논증을 붙이지 않았다. 만약 라플라스가 현 시점에서 위 부분을 보충한다면 어떻게 할 것인지 궁금하여 이 연구를 하게 되었다.

이 분야의 선두적 연구자인 Bayer와 Diaconis (1992)는 서양 게임 카드의 리플셔플(riffle shuffle)의 경우 충분히 섞기 위해서는 7회 정도의 반복이 필요하다는 연구결과를 발표한 바 있다.

2. 추첨 상자 모형

우리는 다음과 같은 추첨 상자(lottery box)를 고려할 것이다.

[1]교신저자: (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 통계학과, 교수. E-mail: stat420@korea.ac.kr

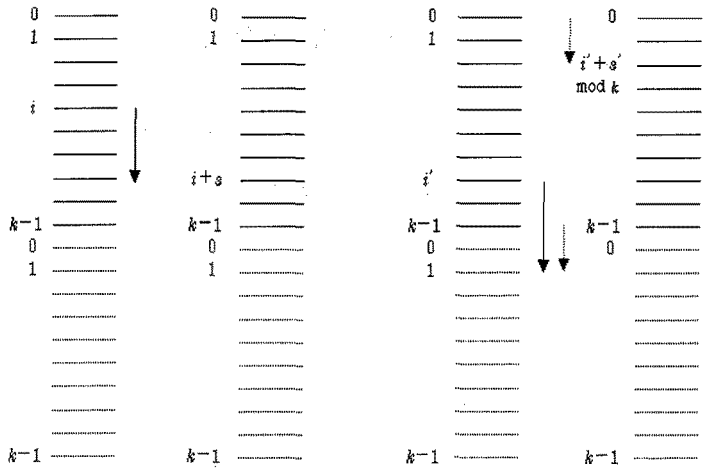

그림 **2.1.** 위치이동: (좌) $i \to i+s$, (우) $i' \to i'+s' \bmod k$

1) 1부터 k까지 숫자가 하나씩 적힌 카드들이 상자에 순서대로 쌓여 있는 상태에서 시작한다. 즉 1번 카드가 가장 위인 위치 0에 있고, 2번 카드가 그 아래인 위치 1에, 3번 카드가 그 아래에 있으며 k번 카드가 위치 $k-1$인 바닥에 있는 것으로 간주한다.
2) 추첨 상자에 손을 넣어 휘젓고 퍼내어 k장의 카드가 섞이게 한다.
3) 카드들을 위에서 아래로 일렬로 정렬한 다음 위에 있는 것부터 하나씩 꺼낸다.
4) 이런 과정을 거쳐 i번째 카드가 j번째 위치로 옮겨질 확률 $P(i,j)$에 대하여

$$P(i,j) = h_{j-i+k \bmod k}, \quad i,j = 0,1,\ldots,k-1$$

임을 가정한다(마코브 섞음, Markov mixing). 여기서

$$\sum_{l=0}^{k-1} h_l = 1, \quad h_l \geq 0 \quad (l = 0,1,\ldots,k-1)$$

이다. 즉, 위치이동(Shift) s를 $j-i \geq 0$이면 $j-i$로, 그렇지 않으면 $j+k-i$로 정의하고 $P(i,j)$를 s의 함수로 보자는 것이다. 그림 2.1을 보라.

이런 작업 후, 추첨 상자에 형성된 정렬에서 1번 카드의 위치를 Z_1이라고 하자. 그리고 이런 물리적 과정을 한 번 더 거쳐 만들어진 정렬에서 1번 카드의 위치를 Z_2라고 하자. 이런 식으로 n번의 물리적 임의화 작업 다음 1번 카드의 위치를 Z_n으로 표기한다. 그러면 $Z_1, Z_2, \ldots, Z_n$은 다음과 같이 기술될 수 있다: $X_1, X_2, \ldots, X_n$을 분포 H를 따르는 *i.i.d.* 확률변수라고 하자. 여기서 H는 j를 h_j의 확률로 취하는 이산형 분포를 지칭한다($j = 0,1,\ldots,k-1$). 그러면

$$Z_n = Z_{n-1} + X_n \bmod k, \quad n = 1,2,3,\ldots$$

으로 나타낼 수 있다. 여기서 $Z_0 = 0$. 따라서 Z_n은

$$Z_n = X_1 + \cdots + X_n \bmod k, \quad n = 1,2,3,\ldots$$

으로도 표현 가능한데 다음과 같은 확률적 행태를 보일 것이다.

정리 2.1 모든 h_j가 양이라고 하자($j = 0, 1, \ldots, k-1$). n이 크면, 근사적으로

$$Z_n \ (= X_1 + \cdots + X_n \mod k) \ \sim \ \text{Uniform } \{0, 1, \ldots, k-1\}$$

이다. 즉, n이 크면 Z_n은 $\{0, 1, \ldots, k-1\}$를 거의 같은 확률로 취한다:

$$\lim_{n\to\infty} P\{Z_n = j\} = \frac{1}{k}, \quad j = 0, 1, \ldots, k-1.$$

증명: $\{Z_n\}$이 마코브 연쇄(Markov chain)가 됨은 자명하다. 그리고 $i \to j$ 전이확률은

$$P\{Z_{n+1} = j \mid Z_n = i\} = h_{j-i+k \bmod k}$$

가 된다. 따라서 마코브 전이행렬(transition matrix)이 다음과 같이 표현된다.

$$\mathbf{P} = \begin{pmatrix} h_0 & h_1 & \cdots & h_{k-1} \\ h_{k-1} & h_0 & \cdots & h_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ h_1 & h_2 & \cdots & h_0 \end{pmatrix}.$$

$\mathbf{P}$의 모든 요소가 양이므로 이 마코브 과정은 비주기적(aperiodic)이고 재발생적(recurrent)이다. 따라서 k개의 상태 $0, 1, \ldots, k-1$은 정상(定常, stationary) 확률 π ($k \times 1$ 벡터)를 갖는다. π는

$$\pi^t = \pi^t \mathbf{P} \tag{2.1}$$

로부터 나오는데

$$\mathbf{e}_k^t = \frac{1}{k}(1, 1, \ldots, 1)$$

이 식 (2.1)을 만족하므로 $\{Z_n\}$의 정상분포는 바로 균일분포이다. □

정리 2.1에서 모든 h_j가 양이라고 가정하였지만, 꼭 그러해야만 하는 것은 아니다. 마코브 연쇄가 비주기적이고 재발생적이기만 하면 된다.

예로서 다음과 같은 절단 기하분포(truncated geometric distribution)를 생각하자.

$$P\{X_1 = j\} = c_k r^j, \quad j = 0, 1, \ldots, k-1,$$

여기서

$$c_k = \begin{cases} \dfrac{1-r}{1-r^k}, & \text{if } \ r \neq 1, \\ \dfrac{1}{k}, & \text{if } \ r = 1 \end{cases}$$

이다. r이 1에 가까울수록 Z_1 ($= X_1$)의 분포가 균일분포에 가깝고 r이 1에서 멀어질수록 균일분포로부터 멀어진다. $r \neq 1$인 경우, 1회의 마코브 섞음(Markov mixing)으로는 순서열의 임의화를 기대할 수 없다. 그러나 위의 정리는 마코브 섞음 과정을 반복함으로써 순서열의 임의화를 달성할 수 있음을 보여준다.

그렇다면 몇 번을 반복하여야 이 마코브 연쇄가 정상(定常, stationary) 상태에 이르게 될까? $k=10$인 경우, 마코브 섞음(Markov mixing)에 의하여

$$\max_{j=0,1,\ldots,k-1}\left|P\{Z_n=j\}-\frac{1}{k}\right|<10^{-2}$$

이 되는 최소의 n을 구해보면 다음과 같다.

- $r=0.9$인 경우 $n=2$.
- $r=0.8$인 경우 $n=3$.
- $r=0.7$인 경우 $n=5$.
- $r=0.6$인 경우 $n=7$.
- $r=0.5$인 경우 $n=11$.

즉, r이 작아서 Z_1의 분포가 균일분포로부터 먼 경우에서도 섞음(mixing) 횟수 n을 크게 하면 Z_n의 분포는 결국 균일분포에 수렴하게 된다.

3. 특수한 경우로서의 이항분포

$\theta=0.5$인 이항분포 $B(n,\theta)$에서는 모든 n에 대하여

$$\sum_{\text{even } x}\binom{n}{x}\theta^x(1-\theta)^{n-x}=\frac{1}{2} \tag{3.1}$$

이다 (여기서 'even x'는 0을 포함함). 왜냐하면 n이 1, 3, 5 등 홀수인 경우 분포가 $x=n/2$를 중심으로 대칭이므로 성공 횟수가 짝수 x일 확률과 홀수 $n-x$일 확률이 같으므로, n이 2, 4, 6 등 짝수인 경우에는

$$\sum_{\text{odd } x}\binom{n}{x}\frac{1}{2^n}=\sum_{\text{odd } x}\left\{\binom{n-1}{x-1}+\binom{n-1}{x}\right\}\frac{1}{2^n}=\frac{1}{2}\sum_{\text{all } j}\binom{n-1}{j}\frac{1}{2^{n-1}}=\frac{1}{2}$$

이기 때문이다.

그러면 $\theta\neq 0.5$인 이항분포 $B(n,\theta)$에서도 식 (3.1)이 성립할 것인가? 답은, 큰 n에 대해서는, 거의 그렇다는 것이다. 그 이유는 다음과 같다.

앞 절의 맥락에서 $k=2$인 경우를 고려하자. 즉, $X_1,X_2,\ldots,X_n$이 독립적인 베르누이 확률변수로서 성공의 확률이 θ라고 하자. 즉, $P\{X_1=1\}=\theta$, $P\{X_1=0\}=1-\theta$이다. 그러면 $S_n(=X_1+X_2+\cdots+X_n)$이 이항분포 $B(n,\theta)$를 따르므로 앞의 정리로부터

$$\lim_{n\to\infty}P\{S_n \bmod 2=j\}=\frac{1}{2},\quad j=0,1$$

임을 알 수 있다. 따라서 다음을 얻는다.

보조정리 3.1 모든 $0<\theta<1$에 대하여

$$\lim_{n\to\infty}\sum_{\text{even } x}\binom{n}{x}\theta^x(1-\theta)^{n-x}=\frac{1}{2}.$$

수치 예로서, 이항분포 $B(n, 0.8)$에서 ($\theta = 0.8$인 경우)

$$n = 4; \quad \sum_{\text{even } x} \binom{4}{x} \theta^x (1-\theta)^{4-x} = 0.5648,$$

$$n = 8; \quad \sum_{\text{even } x} \binom{8}{x} \theta^x (1-\theta)^{8-x} = 0.5084,$$

$$n = 16; \quad \sum_{\text{even } x} \binom{16}{x} \theta^x (1-\theta)^{16-x} = 0.5001$$

로, n이 커짐에 따라 성공 횟수가 짝수일 확률이 0.5에 가까워짐을 확인할 수 있다.

이것은 2장의 카드를 임의로 위치 이동하는 경우, 잔류와 이동의 확률이 0.5가 아니더라도 0과 1사이의 같은 확률로 위치 이동을 반복하면 $P\{Z_n = j\} \to 0.5$가 됨을 의미한다($j = 0, 1$). 여기서 $Z_n = S_n \bmod 2$이다.

4. 미국의 1970년 징병추첨

1969년 12월 1일에 실시된 미국의 1970년 징병 추첨에서 366일의 생일이 표시된 종이쪽지가 플라스틱 캡슐에 넣어져 1월 31일(= 31번 카드)부터 1월 1일(= 1번 카드)까지 그리고 2월 29일(= 60번 카드)부터 2월 1일(= 32번 카드)까지, ... , 마지막으로 12월 31일(= 366번 카드)부터 12월 1일(= 336번 카드)까지의 순서로 추첨 상자에 넣어져 쌓였다. 따라서 추첨 상자의 위쪽을 앞으로 정의하여 추첨 상자에서 생일 카드의 첫 순서는 다음과 같다.

$$336, 337, \ldots, 366, \ldots, 32, 33, \ldots, 60, 1, 2, \ldots, 31.$$

추첨 상자에 손을 넣어 휘젓고 퍼내어 생일 카드들이 섞이게 한 다음 추첨에 들어갔다. 첫 추첨 생일은 9월 14일(= 257번 카드)이었다. 다음은 4월 2일(= 93번 카드), ... 계속 추첨하여 마지막 추첨 생일을 6월 8일(= 160번 카드)이었다. 결국, 366장의 생일 카드들의 순서가 다음과 같이 정해졌다 (Finkelstein과 Levin, 2001).

$$257,\ 93, \ldots,\ 160.$$

그런데 2절의 표기에 맞추기 위해서는 입력순서에 따른 출현순서를 파악해야 할 필요가 있다. 예컨대 맨 앞의 카드(= 336번 카드)의 추첨번호 129, 다음 카드(= 337번 카드)의 추첨번호 328,..., 마지막 카드(= 31번 카드)의 추첨번호 211 등이다. 정리하면 입력순서(input.seq)와 출현순서(outcome.seq)가

$$\text{input.seq} = \quad 1, \quad 2, \ldots, 366 \,(\quad 0, \quad 1, \ldots, 365)$$

$$\text{outcome.seq} = 129, 328, \ldots, 211 \,(128, 327, \ldots, 210)$$

가 되고 (괄호 안은 2절과 표기를 맞추기 위하여 1을 뺀 '위치'), 이에 따라 이동 Shift가 다음과 같이 산출된다.

$$\text{Shift} = \text{outcome.seq} - \text{input.seq} + 366 \mod 366.$$

즉 Shift의 관측값은 다음과 같다.

$$128(= 128 - 0),\ 326(= 327 - 1), \ldots, 211(= 210 - 365 + 366).$$

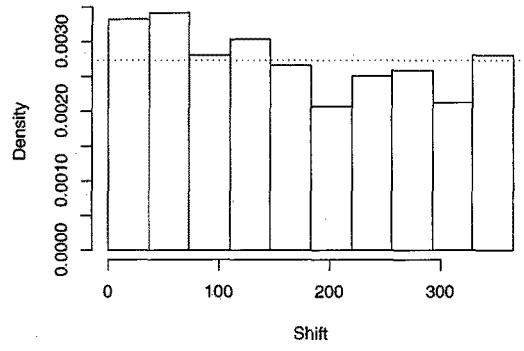

그림 **4.1.** 366개 Shift의 경험분포 (점선은 균일밀도)

그림 4.1은 Shift의 히스토그램이다. 균일분포보다는 작은 값들에 쏠려 있음을 볼 수 있다. 이것은 출현 순서가 입력순서와 연관되어 있음을 의미한다.

이런 비균질적인 섞음(mixing)을 반복하면 결과가 어떻게 될 것인가? 이를 묘사하기 위하여 다음과 같이 Shift의 확률함수 H를 추정하였다.

$$H: \ \hat{h}_j = \frac{\#\{\text{Shift} = j\}}{366}, \quad j = 0, 1, 2, \ldots, 365.$$

이런 섞음을 2회 하면 그 결과로 Shift에 대한 확률함수는

$$H^{[2]}: \ \hat{h}_j^{[2]} = \sum_{i=0}^{365} \hat{h}(i)\, \hat{h}(j - i + 366 \mod 366)$$

이 된다. 1970년 징병추첨 자료에서 모든 $j = 0, 1, 2, \ldots, 365$에 대하여 $\hat{h}_j^{[2]} > 0$이고 $\hat{h}_j^{[3]} > 0$이다. 따라서 이 마코브 연쇄는 비주기적(aperiodic)이고 재발생적(recurrent)이다. 이와 같은 식으로 $n\,(= 3, 4, \ldots)$회 섞음에서의 Shift에 대한 확률함수는

$$H^{[n]}: \ \hat{h}_j^{[n]} = \sum_{i=0}^{365} \hat{h}^{[n-1]}(i)\, \hat{h}(j - i + 366 \mod 366)$$

이다.

섞음 횟수 n이 커짐에 따라 균일분포에 수렴해 가는가를 다음과 같이 $H^{[n]}$의 분포와 균일 분포간의 최대차이를 살펴보았다. $n = 1, 2, 3, 4, \ldots$에 대하여

$$D_n = \max_{j=0,1,2,\ldots,365} \left| \hat{P}\{Z_n = j\} - \frac{1}{366} \right|$$

을 산출한 결과는 다음과 같다.

$$D_1 = 0.00820, \quad D_2 = 0.00051, \quad D_3 = 0.00005, \quad D_4 < 0.00001.$$

섞음 횟수 n이 커짐에 따라 균일분포에 수렴해감을 알 수 있다. 각 생일에 대한 균일 확률 1/366이 0.00273임을 감안하면 2회의 반복으로 상대적 편차가 20% 이내로 되고 3회의 반복을 하면 상대적 편차가 2% 이내로 된다.

5. 맺음말

1971년의 미국 징병추첨에서는 2개의 추첨상자가 사용되었는데 한 상자에서는 생일이, 다른 한 상자에서는 징병순서가 뽑혔다. 이런 방식으로 2회의 물리적 임의화를 일반인들에게 구현해 보인 것이다. 그러나 4절의 결과에 의하면 3회 정도의 물리적 임의화가 필요하므로 1971년의 추첨에서도 제도적 보완이 충분하지 않았다고 할 수 있다.

우리나라의 화투에서도 섞기가 중요한데 4절의 방법으로 필요한 반복수를 산출할 수 있을 것으로 생각한다.

참고문헌

라플라스 (1925). <확률에 대한 철학적 시론> (제 5판). 조재근 역. 지식을만드는지식.

Bayer, D. and Diaconis, P. (1992). Trailing the dovetail shuffle to its lair, *Annals of Applied Probability*, **2**, 294–313.

Finkelstein, M.O. and Levin, B. (2001). *Statistics for Lawyers* (Second Edition). Springer, New York. 262-265.

Randomizing Sequences of Finite Length

Myung-Hoe Huh[1] · Yonggoo Lee[2]

[1]Department of Statistics, Korea University; [2]Department of Statistics, ChungAng University

(Received November 2009; accepted December 2009)

Abstract

It is never an easy task to physically randomize the sequence of cards. For instance, US 1970 draft lottery resulted in a social turmoil since the outcome sequence of 366 birthday numbers showed a significant relationship with the input order (Wikipedia, "Draft Lottery 1969", Retrieved 2009/05/01).

We are motivated by Laplace's 1825 book titled Philosophical Essay on Probabilities that says

> "Suppose that the numbers $1, 2, \ldots, 100$ are placed, according to their natural ordering, in an urn, and suppose further that, after having shaken the urn, to shuffle the numbers, one draws one number. It is clear that if the shuffling has been properly done, each number will have the same chance of being drawn. But if we fear that there are small differences between them depending on the order in which the numbers were put into the urn, we can decrease these differences considerably by placing these numbers in a second urn in the order in which they are drawn from the first urn, and then shaking the second urn to shuffle the numbers. These differences, already imperceptible in the second urn, would be diminished more and more by using a third urn, a fourth urn, &c." (translated by Andrew I. Dale, 1995, Springer. pp. 35–36).

Laplace foresaw what would happen to us in 150 years later, and, even more, suggested the possible tool to handle the problem. But he did omit the detailed arguments for the solution. Thus we would like to write the supplement in modern terms for Laplace in this research note.

We formulate the problem with a lottery box model, to which Markov chain theory can be applied. By applying Markov chains repeatedly, one expects the uniform distribution on k states as stationary distribution. Additionally, we show that the probability of even-number of successes in binomial distribution with trials and the success probability θ approaches to 0.5, as n increases to infinity. Our theory is illustrated to the cases of truncated geometric distribution and the US 1970 draft lottery.

Keywords: Lottery box, physical randomization, Markov chain, binomial distribution.

[1]Corresponding author: Professor, Department of Statistics, Korea University. 5-1 Anam-Dong, Sungbuk-Gu, Seoul 136-701, Korea. E-mail: stat420@korea.ac.kr