한국통계학회논문집 제3권 2호
1996년 8월 pp. 291-298

축차확률비검정에서의 몬테칼로 주표본 연구[1)]

최 기 헌[2)], 김 용 철[3)]

요 약

통계학분야 그리고 또 다른 많은 분야에서 수치적 계산을 다루는 문제가 자주 발생한다. 적당한 컴퓨터 시간안에 상당한 정도의 정확성을 줄 수 있고 또한 보다 광범위하게 사용 가능한 유용한 알고리즘의 필요성을 느낀다. 이러한 문제에 가능한 하나의 몬테칼로 알고리즘인 주표본 알고리즘을 소개하였다. 그리고 특히 본 논문에서는 축차확률비검정의 오차확률을 계산하는 곳에 주표본 알고리즘을 적용하고 결과를 비교 분석하였다.

1. 서론

본 논문에서는 통계추론의 어떤 문제들은 축차적이므로 축차적인 관점들을 고려하지 않을 수 없다. 혹은 알려지지 않은 장애모수(nuisance parameter)가 존재하는 경우에는 고정표본 해가 존재하지 않으므로 축차적인 관점에서 해결 가능하다. 즉 축차검정을 하게된다. 또한 적분문제에서 컴퓨터를 이용한 수치적 기법은 컴퓨터의 대중화와 계산속도의 향상으로 대환영을 받았으나 통계에서 사용하기에는 문제점들이 있으므로 통계학자들이 개발한 몬테칼로 기법이 있다. 특히 복잡한 모양을 가진 함수의 다차원 적분의 경우에 자동적이고 효율적으로 수행하는 몬테칼로 주표본(Monte Carlo importance sampling) 알고리즘을 사용한다(Geweke, 1989). 따라서 축차확률비검정에서 오차확률을 직접 몬테칼로 알고리즘, 몬테칼로 주표본 알고리즘과 Wald의 근사값을 비교하여 몬테칼로 주표본 알고리즘 사용하는 것이 분산감소에 더욱더 효율적인 것을 보인다.

먼저 본 논문에서의 직접 몬테칼로(direct Monte Carlo)는 상대도수에 의하여 확률을 추정하는 것이다. 즉, 사건 A의 확률을 추정하는 것이다. $\alpha = P(A)$를 추정하기 위하여 A의 지시확률변수 I_A의 n개의 독립 실현치의 평균으로 사용한다. 즉,

$$\hat{\alpha} = \frac{1}{n}\sum I_A \,, \tag{1}$$

여기서 $P\{I_A=1\} = \alpha = 1 - P\{I_A=0\}$이다. 이 추정량은 불편추정량이고 분산은

1. 본 연구는 95년도 덕성여자대학교의 연구비 지원으로 이루어졌음.
2. (132-714) 서울시 도봉구 쌍문동 419, 덕성여자대학교 통계학과 조교수.
3. (449-714) 경기도 용인시 삼가동 117-6, 용인대학교 전산통계학과 전임강사.

$$\frac{1}{n} V(I_A) = \frac{1}{n} \alpha(1-\alpha) \tag{2}$$

이다. 만약 α가 적으면 정확한 추정을 위해 오히려 큰 n이 요구된다(3절 참조).

일반적으로 주표본 알고리즘을 이용하여

$$\theta = E\psi(X) \tag{3}$$

를 추정하는 것을 생각하자. 이때 $f(x) > 0$, $x \in \Omega$이고 이 범위에서 확률밀도함수라 하면

$$\theta = \int_\Omega \psi(x) f(x) dx \tag{4}$$

이다. 이때에 $x_1, x_2, \cdots, x_n$를 확률밀도함수 $f(x)$로부터의 확률표본이라하면 추정량은

$$\hat{\theta} = \sum \psi(x_i)/n \tag{5}$$

이다. 그러나 확률밀도함수 $f(x)$로부터의 확률변수의 생성이 용이하지 않을경우 주표본 알고리즘에서는 $f(x)$와 근사적으로 같고 확률변수의 생성이 보다 쉬운 다른 확률밀도함수 $g(x)$를 이용한다. 즉,

$$\begin{aligned} \theta &= \int_\Omega \psi(x) f(x)\, dx = \int_\Omega \psi(x) \frac{f(x)}{g(x)} g(x) dx \\ &= E_g\left(\psi(x) \frac{f(x)}{g(x)} \right) \end{aligned} \tag{6}$$

이다. 이때 $f(x)/g(x)$를 가중함수 $w(x)$라 하고 $x_1, x_2, \cdots, x_n$를 확률밀도함수 $g(x)$로부터의 확률표본이라하면 추정량은

$$\hat{\theta} = \sum \psi(x_i) w(x_i)/n \tag{7}$$

이다. 최근에는 가중함수의 평균이 확률 1을 갖고 1로 수렴한다는 사실을 이용하여 추정량을 가중평균으로서 추정한다.

Geweke (1989)에 의한 수치적으로 추정하기 위한 몬테칼로 주표본 알고리즘은 다음과 같다.

1단계: 확률밀도함수 $f(x)$와 모양이 비슷하면서도 다루기 쉬고 간단한 확률밀도함수 $g(x)$를 찾는다.

2단계: 컴퓨터를 사용하여 함수 $g(x)$를 확률밀도함수로 갖는 분포로부터 확률표본 $x_1, x_2, \cdots, x_n$를 생성시킨다.

3단계: 위에서 얻은 표본과 가중함수 $w_i = f(x_i)/g(x_i)$를 사용하여 가중평균

$$\hat{\theta} = \frac{\sum \psi(x_i) w_i}{\sum w_i} \tag{8}$$

을 구하고 이를 기대치에 대한 추정치로 한다. 또한 추정치의 분산은

$$V(\hat{\theta}) = \frac{\sum \{\psi(x_i) - \hat{\theta}\}^2 w_i^2}{(\sum w_i)^2} \tag{9}$$

이다.

따라서 몬테칼로 주표본 알고리즘은 $f(x) > g(x)$인 경우에는 더 가중을 많이 주고 반면에 $f(x) < g(x)$인 경우에는 덜 가중을 주므로 $g(x)$로 생성된 표본으로 정확히 기대치를 계산할 수 있다.

한편, 우도비 동등(likelihood ratio identity)으로부터 적당한 확률 Q에 대하여

$$P(A) = \int_A L\, dQ \tag{10}$$

와 같이 표현한다(Woodroofe, 1982). 여기서 L은 Q에 대한 P의 우도비를 나타낸다. 만약 P와 Q의 밀도함수를 p와 q라면 $L = p/q$이다. 그러면 α는 Q에 의하여 유도되는 분포로부터 $I_A L$의 n개 독립 실현치의 평균으로 추정된다. 즉

$$\hat{\alpha} = \frac{1}{n} \sum I_A L \tag{11}$$

이 추정량이면서 불편추정량이고 분산은

$$\frac{1}{n} V_Q(I_A L) = \frac{1}{n} \left(\int_A L^2 dQ - \alpha^2 \right) \tag{12}$$

이고 식 (2)보다 작다. 예를 들어 만약 $L \leq 1$이면

$$\int_A L^2 dQ \leq \int_A L dQ = P(A) = \alpha \tag{13}$$

이고 식 (12)은 식 (2)보다 크지 않다. 식 (12)으로부터 α의 추정량의 분산을 줄이기 위하여 적당한 Q의 선택은 L을 작게 만들고 A상에서 거의 상수이다. 만약 A가 통계검정에서 기각역이고 P가 귀무가설이면 대립가설에 의하여 적당한 Q를 찾으면 이 검정은 우도비검정으로 확장할 수 있다.

특히, 가중함수 $w = L$이라하면 가중평균을 이용한 주표본 알고리즘에서의 추정량은

$$\tilde{\alpha} = \frac{\sum I_A w_i}{\sum w_i} \tag{14}$$

이다.

본 논문에서는 축차확률비검정의 오차확률를 추정하는 것을 몬테칼로 주표본 알고리즘의 방법으로 설명할 것이다. Wald (1947)의 잘 알려진 근사값은 일반적인 조건하에서는 오차확률에 합당한 수치적 근사를 보였다 (Siegmund, 1985). 그러나 몬테칼로 주표본 알고리즘의 실제적인 응용은 수치적 근사가 존재하지 않거나 시뮬레이션의 정확도를 비교하기 위한 복잡한 문제에 더 주요하다. 3절에서 예를 보인다.

2. 오차확률의 근사

지금 $x_1, x_2, \cdots$는 아래 조건

$$-\infty < Ex_1 < 0 \quad \text{그리고} \quad P\{x_1 > 0\} > 0 \tag{15}$$

을 만족하는 서로 독립이고 동일한 분포를 따른다. 이때 $s_n = \sum x_k$라하고 $a \le 0 < b$에 대하여 정지시간(stopping time)을

$$T = \inf\{n: n \ge 1,\ s_n \notin (a, b)\} \tag{16}$$

라 하자. 본 논문의 주제는 $b \to \infty$에 따라 근사적으로

$$\alpha = P\{s_T \ge b\} \tag{17}$$

를 추정하기 위해 주표본의 몬테칼로 방법을 사용하는 것이다.

특히 x가 로그우도비인 경우에 T는 Wald의 축차확률비검정의 정지시간이고 일반적인 조건하에서 Wald는 α에 대해 근사값을 준다(Wald, 1947).

식 (16)의 특별한 경우로 $a = 0$는 분포함수 변화를 탐지하기 위한 Page형태의 품질관리 처리에서 일어난다(Page 1954). 또한 single-server queue에서도 일어나며 식 (10)은 single busy period동안 손님에 대한 최대대기시간이 b보다 클 확률을 나타낸다. $a = 0$일 때는 Wald의 근사를 사용할 수 없으므로 다른 방법을 찾아야 한다.

식 (16)와 (17)에 대한 Wald의 근사방법은 $(s_T - b)I\{s_T \ge b\}$와 $(s_T - a)I\{s_T \le a\}$를 무시하

는 것이다. 그래서 s_T는 단지 a와 b에서만 확률변수로 취급한다.

식 (17)을 추정하는 것은 쉬운 문제가 아니므로 이때 확률 P를 지수족(exponential family) $\{P_\theta\}$라 가정하자(Sigmund, 1985). 그러면 확률 P_θ하에서 $x_1, x_2, \cdots$는 서로 독립이고 동일한 밀도함수

$$P_\theta\{x_k \in dx\} = \exp(\theta x - \Psi(\theta))\, dH(x) \tag{18}$$

를 갖는다. 여기서 함수 Ψ는

$$\Psi(0) = \Psi'(0) = 0 \tag{19}$$

로 표준화되었고 $dH(x) = P_0\{x_k \in dx\}$이다. 예를 들어 x가 평균이 μ이고 분산이 1인 정규분포를 따르면 확률 P_θ하에서 그것은 평균이 θ이고 분산이 1이다. 따라서 주어진 식 (18)으로부터 $\Psi(\theta) = \theta^2/2$이고 확률밀도함수는 $dH(x) = (2\pi)^{-1/2}\exp(-x^2/2)dx$이다. 또한 함수 Ψ는

$$\Psi'(\theta) = E_\theta(x_1), \qquad \Psi''(\theta) = V_\theta(x_1) \tag{20}$$

이다. 따라서 식 (19)과 (20)으로부터

$$sgn\, E_\theta(x_1) = sgn\, \theta \tag{21}$$

임을 알수 있다. 또한 식 (15)와 (21)로부터 $\mu < 0$이다. 더욱이 식 (19)와 Ψ의 엄격한 볼록성(strictly convexity)에 의하여

$$\Psi(w) = \Psi(u) \tag{22}$$

를 만족하는 양의 w가 기껏해야 하나 존재한다. 여기서는 그런 w가 하나 존재하는 것을 가정한다.

식 (18), (19)와 식 (10)으로부터 임의의 θ에 대하여

$$a = P_u\{s_T \geq b\} = \int_{(s_T \geq b)} \exp\{-(u-\theta)s_T - T(\Psi(u) - \Psi(\theta))\}\, dP_w \tag{23}$$

이다. 또한 $\theta = w$이면 식 (22)로부터

$$a = P_u\{s_T \geq b\} = \int_{(s_T \geq b)} \exp\{-(w-u)s_T\}\, dP_w \tag{24}$$

이 된다. 따라서 식 (24)로부터 부등식

$$P_u\{s_T \geq b\} \leq \exp\{-(w-u)b\} P_w\{s_T \geq b\} \leq \exp\{-(w-u)b\} \tag{25}$$

는 Wald의 축차확률비검정의 기본를 이룬다. 식 (23)은 식 (10)의 특수한 경우이고

$$I_{(s_T \geq b)} \exp\{-(u-\theta)s_T - T(\Psi(u) - \Psi(\theta))\} \tag{26}$$

의 P_θ 실현값의 n개 독립의 평균으로 α를 추정한다. $\theta = w$인 몬테칼로 주표본 알고리즘인 경우에 이 추정량의 분산은 $1/n$ 곱하기

$$\begin{aligned} &\int_{(s_T \geq b)} \exp\{-2(w-u)s_T\} dP_w - \alpha^2 \\ &\quad \leq \exp\{-(w-u)b\} \int_{(s_T \geq b)} \exp\{-(w-u)s_T\} dP_w - \alpha^2 \\ &\quad = \exp\{-(w-u)b\}\alpha - \alpha^2 \end{aligned} \tag{27}$$

이다. 적당한 b값에 대하여 분산은 직접 몬테칼로(direct Monte Carlo)의 분산 $\alpha(1-\alpha)/n$보다 훨씬 적다. 일반적으로 $b \to \infty$함에 따라 식 (25)의 분산은 $O(\alpha^2)$이고 직접 몬테칼로에서는 $O(\alpha)$이다. 수치적인 예를 다음절에서 보일 것이다.

3. 수치적 예

P_θ 하에서 $x_1, x_2, \cdots$는 서로 독립이고 평균이 θ이고 분산이 1인 정규분포를 따른다 하자. 이때 $\Psi(\theta) = \theta^2/2$이고 $w = -u$이다. 그러면 소위 누적합 그림(cumulative sum chart)을 분석하는 품질관리의 중요한 문제는 부분합(partial sum)이 구간 a 아래로 가기 전에 b를 초과할 확률를 결정하는 것이다. 즉 $s_n = \sum_{i=1}^{n} x_i$라 하자. 그리고 a, b를 고정된 상수라 하면 정지시간은 식 (16)로 정의하자. 그러면 우리는 식 (17)를 추정하기를 원한다. 이때 $x_1, x_2, \cdots, x_T$는 시뮬레이트한 변수로서 각각은 평균이 $-u$이고 분산은 1인 정규분포를 따르고

$$I = \begin{cases} 1 & s_T \geq b \\ 0 & \text{다른 곳에서} \end{cases} \tag{28}$$

라하면 α의 추정치는

$$I \Pi_{i=1}^{T} \left[\frac{f_u(x_i)}{f_{-u}(x_i)} \right] \tag{29}$$

이고 여기서 f_c는 평균이 c이고 분산이 1인 정규확률밀도함수이다. 그런데

$$\frac{f_u(x)}{f_{-u}(x)} = \frac{\exp\left\{-\frac{(x-u)^2}{2}\right\}}{\exp\left\{-\frac{(x+u)^2}{2}\right\}} = \exp(2ux) \tag{30}$$

서로 독립인 실현값

$$I_{(s_T \geq b)} \exp\{2us_T\} \tag{31}$$

의 평균을 가지고 α의 시뮬레이션 값을 비교한 것이다. 첫째 대칭인 경우, 즉 $a = -b$인 경우에 서로 다른 b값과 $u = E(x_1)$에 대하여 표 1에 주어져 있다.

표 1

b	$-u$	$\hat{\alpha}$	$\hat{\hat{\alpha}}$	$\tilde{\alpha}$	σ	$\{\alpha(1-\alpha)\}^{1/2}$	상대효율	Wald
7	.5	0.0005	0.000509	0.000816	0.000231	0.0225	9500	0.000911
7	.25	0.0212	0.0219	0.0180	0.00633	0.1466	536	0.02931
7	.125	0.1234	0.1300	0.1273	0.539	0.3363	38	0.14805
7	.0	0.5073	0.4964	0.4964	0.4999	0.4999	1	0.50000
4	.5	0.01	0.0101	0.0137	0.0047	0.1004	454	0.01799
4	.25	0.0894	0.09289	0.0966	0.0366	0.2902	62	0.11920
4	.125	0.2437	0.2445	0.2526	0.1379	0.4298	9	0.26894
4	.0	0.4956	0.4979	0.4979	0.4999	0.4999	1	0.50000

둘째 대칭이 아닌 경우 표 2에 주어져 있다.

표 2

b	a	$-u$	$\hat{\alpha}$	$\hat{\hat{\alpha}}$	$\tilde{\alpha}$	σ	$\{\alpha(1-\alpha)\}^{1/2}$	상대효율	Wald
7	-4	.5	0.0004	0.0005019	0.0005310	0.0002345	0.02239	9118	0.018299
7	-4	.25	0.0203	0.02031	0.02027	0.008425	0.1410	280	0.13129
7	-4	.125	0.104	0.1066	0.1049	0.06992	0.3087	19	0.32471
7	-4	.0	0.3718	0.3801	0.3801	0.4854	0.4854	1	0.63636
7.5	-2.5	.5	0.0005	0.0002923	0.0002802	0.0001519	0.01709	12664	0.082043
7.5	-2.5	.25	0.0143	0.01384	0.01391	0.008043	0.1168	211	0.281664
7.5	-2.5	.125	0.0749	0.07569	0.07524	0.06684	0.2645	15	0.49370
7.5	-2.5	.0	0.2765	0.2779	0.2779	0.4479	0.4479	1	0.75000

첫째 표1과 표2로부터 Wald의 근사값이 α를 과대추정(overestimates)함을 알수 있다. 둘째는 표1과 표2에서의 $\hat{\alpha}$은 직접 몬테칼로 알고리즘을 이용한 추정치이며 식 (1)로부터 계산하였다. 또한 $\hat{\hat{\alpha}}$와 $\tilde{\alpha}$는 주표본 알고리즘을 사용한 추정치로서 각각은 식 (11)과 (14)를 이용하였으며 σ는 식 (31)의 표준편차값이다. 특히 상대효율은 추정치 $\hat{\hat{\alpha}}$의 직접 몬테칼로 알고리즘과 주표본 알고리즘의 분산 비를 가지고 비교하였다. 여기서 $\hat{\hat{\alpha}}$값이 작으면 보다 정확한 추정을 위해 상대적으로 큰 n을 필요함을 알수있고 α를 추정하는데 있어서 몬테칼로 주표본 알고리즘이 분산감소에 더 효율적인 것을 보여준다.

최근에는 가중평균을 이용한 몬테칼로 주표본 알고리즘 추정치 $\tilde{\alpha}$에 대하여 연구가 활발하고 추정치 $\tilde{\alpha}$가 강대수 법칙(Strong Law of Large number)에 의하여 확률 1을 갖고 실제값 α에 수렴함을 보여준다(Geweke, 1989). 그리고 여기서는 보여주지 않았지만 추정치 $\tilde{\alpha}$의 분산은 식 (9)를 사용하여 구할수 있으나 $\hat{\hat{\alpha}}$의 분산과 비교하여 별 차이가 없으므로 여기서는 제시하지 않았다.

참고문헌

[1] Anderson, T. W. (1960). A modification of the sequential probability ratio test to reduce the sample size, *Ann. Math. Statist.*, Vol. 31, 165-197.

[2] Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integrations, *Econometrica*, Vol. 24, 1317-1339.

[3] Ghosh, B. K. (1970). *Sequential Tests of Statistical Hypotheses.* Reading, Mass.: Addison-Wesley.

[4] Page, E.. S. (1954). Continuous inspection schemes, *Biometrika*, Vol. 37, 326-333.

[5] Siegmund, D. (1975). A note on the error probabilities and average sample number of the sequential probability ratio test, *J. Roy. Statist. Soc. Ser.* 37, 394-401.

[6] Siegmund, D. (1985). *Sequential Analysis,* Springer-Verlag, New York.

[7] Tanner M. A. (1991). *Tools for statistical inference,* Springer-Verlag, New York.

[8] Wald, A. (1947). *Sequential Analysis,* Wiley, New York.

[9] Woodroofe, M. (1982). *Non linear renewal theory in the sequential analysis,* Soc. Indust. Appl. Math., Philadelphia.