한국통계학회논문집 제4권 2호
1997년 8월 pp. 403 - 409

집락 표본추출에 있어서 이-단계 표본 추출

신 민 웅[1] , 이 주 영[2]

요 약

일단-집락 추출을 할 때에 예비표본으로 부터 얻은 정보를 활용하여 추가표본을 추출한다. 특히, 예비표본의 크기(예비표본의 집락의 수) n_1 과 추가표본의 크기 n_2 를 모두 변수로 간주하여 베이즈 위험을 최소로 하는 n_1 과 n_2 의 크기를 결정한다.

1. 서론

이 논문은 일단 집락추출을 할 때에 예비표본을 추출하고, 이 표본으로부터 얻은 정보를 이용하여 추가표본을 추출하는 문제를 다룬다. 예비표본으로 n_1 개의 집락을 추출하고, 추가로 n_2 개의 집락을 추출하는데, n_1 과 n_2 를 모두 변수로 간주하여 베이즈 위험을 최소로하는 n_1 과 n_2 를 결정한다. 즉, 이-단계 표본추출을 하여 사전 정보를 이용하고자 한다.

먼저, 단순 랜덤 표본추출의 경우에는 Miller와 Freund(1977)가 이항분포에서 모수 p 를 추정하는데 있어서, 주어진 예비표본(n_1)으로부터 p 의 추정치 $\hat{p}_1$ 를 구한 후, 이 추정치를 이용하여 추가표본의 크기(n_2)를 정하였다. Cohen과 Sackrowitz(1984a)는 지수족 분포의 평균에 대한 이-단계 베이지안 추정(double sample Bayes estimation)을 하였다. 그 과정은 예비표본의 크기가 n_1 인 자료 X_1 으로부터 $n_2(X_1)$ 을 결정한다. 그들은 더욱이 n_1 과 n_2 를 확률변수로 보아 최적의 값을 구하였다. Josep(1995)은 베이지안 접근방법으로 사후밀도 함수(HPD)의 관점에서 표본의 크기를 결정하였다. 손실함수는 잘못된 추정으로부터 손실과 표본추출비용으로 이루어진다. 추정으로부터의 손실은 제곱오차를 사용하고, 베이즈 과정에서는 공액 사전분포를 적용한다. 우리는 단순랜덤 추출 대신에 일단 집락 표본추출의 경우로 표본추출 시에 집락의 크기를 결정하는 문제를 다루었다.

본 논문은 2절에서는 이항분포를 하는 경우에 대하여 일단 집락 추출 시에 예비표본의 크기와 추가표본의 크기를 구하였다. 3절에서는 모의실험을 통하여 최적의 집락의 수 n_1과 n_2를 구하였다.

1) (449-850) 경기도 용인군 모현면 왕산리 산 89, 한국 외국어 대학교 자연과학대학 통계학과 교수
2) (402-751) 인천 광역시 남구 용현동 253, 인하 대학교 이과대학 통계학과 박사과정

2. 베이즈 추정에 의한 이-단계 표본 추출

확률변수 X가 다음 분포를 갖는다고 하자.

$$dP_\theta(x) = e^{x\cdot\theta - M(\theta)}d\mu(x)$$

단, $M(\theta) = \log\int e^{\theta x}d\mu(x)$ 이고, μ는 R위에서의 σ-유한측도이다. 그리고

$$\theta\in\Theta=\{\theta \mid M(\theta) < \infty\}$$

이다. 그러면 Diaconis 와 Ylvisaker(1979)에 의하면

$$\begin{aligned} E(X\mid\theta) &= \partial M(\theta)/\partial\theta \\ &= M_1(\theta) \end{aligned} \tag{2.1}$$

그리고

$$E_\theta(X-M_1(\theta))(X-M_1(\theta)) = M_{11}(\theta) \tag{2.2}$$

우리는 Θ위에서의 공액 사전 확률 분포를 다음과 같다고 하자.

$$\pi_{n_0,x_0}(\theta) = \beta(n_0,x_0)\exp(n_0x_0\cdot\theta - n_0M(\theta))d\theta,\ n_0\in R^1, x_0\in R^d \tag{2.3}$$

단, n_0와 x_0는 Diaconis 와 Ylvisaker(1979)의 정리1의 조건을 만족하는 상수, d는 차원을 나타내고, $\beta(n_0,x_0)$는 (2.3)식의 적분 값이 1이 되도록 한다. 그러면 같은 참고 문헌으로부터 사후분포는 $\pi_{n_0+n,(n_0x_0+n\overline{X})/(n_0+n)}$ 이다. 또한

$$E\left(\frac{\partial M(\theta)}{\partial\theta}\mid X_1,X_2,\cdots,X_n\right) = (n_0x_0+n\overline{X})/(n_0+n) \tag{2.4}$$

이며, $\overline{X}$는 표본 평균이다.

이제 집락 추출에서 c_1은 부차단위에 대한 조사비용이고, c_2는 집락들 사이에서 일어나는 비용이라고 하자. 손실함수는 행동 $a=(r,n_1,n_2)$에 대하여

$$L_1(\omega,a) = (r-\omega)^2 + c_1M(n_1+n_2) + c_2\sqrt{n_1+n_2} \tag{2.5}$$

이라고 하자. 이때 $\omega = M(\theta)$, r은 ω의 추정치이고, $n = n_1+n_2$이다. 여기서 M은 집락의 크기로 모든 집락은 같은 크기를 갖는다고 가정한다.

이 논문에서의 n_1M개의 원소들과 n_2M개의 원소들은 단순 랜덤 표본이라고 가정한다. (이러한 예는 품질관리에서 표본 검사의 경우나 지역 표본 추출의 경우에 일어날 수 있다.) X_{ij}를 i번째 집락 안에 j번째 부차단위의 관찰 값이라고 하자. i번째 집락의 j번째 원소 X_{ij}, $i=1,2,\cdots$, $j=1,2,\cdots,M$는 모수 p를 갖는 베르누이 확률변수이다.

이제,

$$Y_1=\sum_{i=1}^{n_1}\sum_{j=1}^{M}X_{ij}\ ,\qquad Y_2=\sum_{i=n_1+1}^{n_2}\sum_{j=1}^{M}X_{ij}$$

이고,

$$Z_1=E_\theta[M_{11}(\theta)\mid Y_1],\qquad Z_2=E_\theta[M_{11}(\theta)\mid Y_1,Y_2]$$

이라고 놓자.

우리의 목적은 손실함수(2.5)에 대하여 베이즈 위험을 최소로 하는 n_1과 n_2를 구하는 것이다. 여기서 n_1은 n_1개의 집락들로 예비표본이고, n_2는 추가표본을 의미한다. 베이즈 위험은 다음과 같다.

$$\begin{aligned}E_{Y_1,Y_2}E_\theta[E(M(\theta)\mid Y_1,Y_2)-M_1(\theta)\mid Y_1,Y_2]^2\\+E_{Y_1,Y_2}E_\theta[c_1M(n_1+n_2)+c_2\sqrt{n_1+n_2}]\end{aligned}\tag{2.6}$$

이 베이즈 위험을 최소로 하는 n_1과 n_2는 Cohen과 Sackrowitz(1984a)와 유사하게 구하면 다음과 같다. 즉, 베이즈 위험 (2.6)을 변형하면 이항분포를 하는 경우에는 다음과 같다.

$$\begin{aligned}\sum[\frac{Z_1}{n_1M+n_2M+n_0}+c_1(n_1+n_2)M\\+c_2\sqrt{n_1+n_2}\ ]\binom{n_1M}{y_1}p^{y_1}(1-p)^{n_1M-y_1}\end{aligned}\tag{2.7}$$

그런데, 계산을 용이하게 하기 위하여 비용 $c_2\sqrt{n_1+n_2}$을 무시할 수 있는 경우를 생각한다. n_1과 y_1을 상수로 놓고, n_2에 대하여 미분하면

$$n_2M=[(z_1/c_1)^{1/2}-n_1M-n_0)]^+\tag{2.8}$$

이다.

여기서, $[x]^+$는 $x>0$ 이면 x보다 작은 최대의 정수이고, 그렇지 않으면 0을 의미한다. n_2를 (2.7) 에 대입하여 다음 식을 최소로 하는 n_1을 구한다.

$$\begin{aligned}\sum_B\left[\left(\frac{Z_1}{n_1M+n_0}\right)+n_1Mc_1\right]\binom{n_1M}{y_1}p^{y_1}(1-p)^{n_1M-y_1}\\+\sum_{B'}[2(Z_1c_1)^{1/2}-n_0c_1]\binom{n_1M}{y_1}p^{y_1}(1-p)^{n_1M-y_1}\end{aligned}\tag{2.9}$$

여기서 $B=\{y_1:Z_1\le c_1(n_1M+n_2M)\}$이고, $B'=\{y_1:Z_1>c_1(n_1M+n_2M)\}$이다.

3. 모의실험

확률변수 X_{ij} , $i=1,2,\cdots$, $j=1,2,\cdots,M$ 이 모수 p를 갖는 베르누이 분포를 한다고 하자. 자연모수는

$$\theta = \log(p/(1-p))$$

이고,

$$M_{11}(\theta) = e^{\theta}/(1+e^{\theta}) = p(1-p)$$

이다.

또한

$$Z_1 = E_{\theta}(M_{11}(\theta) \mid Y_1) = (y_1 M + n_0 x_0)\frac{n_1 M - y_1 M + n_0(1-x_0)}{(n_1 M + n_0 + 1)(n_1 M + n_0)} \tag{3.1}$$

이-단계 표본 추출의 값을 평가하기 위하여, 일양사전분포, 즉 $n_0 x_0 = 1$, $x_0 = 1/2$에 대하여 표1에 예비표본에 대한 최적의 크기 n_1을 구하였다.

또한 베이즈 위험을 작게 하는 최적의 표본의 수를 조사하기 위하여 (2.8)과 (2.9)로부터 모의실험을 실행하였다. 이-단계 표본추출을 할 때에 베이즈 위험을 최소로 하는 n_1 과 n_2 를 구하였다.

그리고, 이-단계표본추출과 일회추출($n_2 = 0$)과의 베이즈 위험을 비교하였다. 그 결과 이-단계 표본추출을 할 때가 베이즈 위험이 더 작았다.

표1. c_1의 변화에 따른 최적 표본수와 베이즈위험비교

p	c_1	n_1	n_2	2-sample	1-sample	c_1	n_1	n_2	2-sample	1-sample
0.05		8	2	.0026256	0.002637		25	10	.0041884	0.004270
0.10		10	2	.0032689	0.003295		30	10	.0050244	0.005155
0.15		12	2	.0037818	0.003796		35	10	.0057208	0.005849
0.20		12	3	.0041860	0.004271		40	10	.0062944	0.006391
0.25		4	12	.0044633	0.008847	0.00064	20	30	.0067060	0.009567
0.30	0.000028	4	12	.0046694	0.009615		20	35	.0070149	0.010335
0.35		4	13	.0048208	0.010212		20	35	.0072455	0.010932
0.40		4	13	.0049287	0.010639		20	35	.0074102	0.011359
0.45		4	14	.0049914	0.010895		20	40	.0075073	0.011615
0.50		4	14	.0050118	0.010980		20	40	.0075376	0.011700
0.05		35	10	.0030202	0.003039		25	10	.0043284	0.004370
0.10		45	10	.0037235	0.003735		30	10	.0051844	0.005275
0.15		50	10	.0042866	0.004336		35	10	.0059008	0.005989
0.20		55	10	.0047382	0.004799		35	10	.0064860	0.006732
0.25		20	50	.0050521	0.009007		20	30	.0069060	0.009647
0.30	0.000036	20	50	.0052868	0.009775	0.000068	20	30	.0072310	0.010415
0.35		20	55	.0054578	0.010372		20	35	.0074655	0.011012
0.40		20	55	.0055797	0.010799		20	35	.0076302	0.011439
0.45		20	60	.0056528	0.011055		20	35	.0077290	0.011695
0.50		20	60	.0056757	0.011140		20	35	.0077619	0.011780
0.05		30	10	.0035507	0.003580		25	10	.0044684	0.004470
0.10		35	10	.0043255	0.004431		30	10	.0053444	0.005395
0.15		40	10	.0049545	0.005083		30	10	.0060800	0.006360
0.20		45	10	.0054613	0.005582		35	10	.0066660	0.006872
0.25		20	40	.0058205	0.009247		20	30	.0071060	0.009727
0.30	0.000048	20	40	.0060931	0.010015	0.000072	20	30	.0074310	0.010495
0.35		20	45	.0062894	0.010612		20	30	.0076837	0.011092
0.40		20	45	.0064295	0.011039		20	35	.0078502	0.011519
0.45		20	45	.0065136	0.011295		20	35	.0079490	0.011775
0.50		20	45	.0065416	0.011380		20	35	.0079819	0.011860

위의 결과를 그래프로 나타내면 다음과 같다.

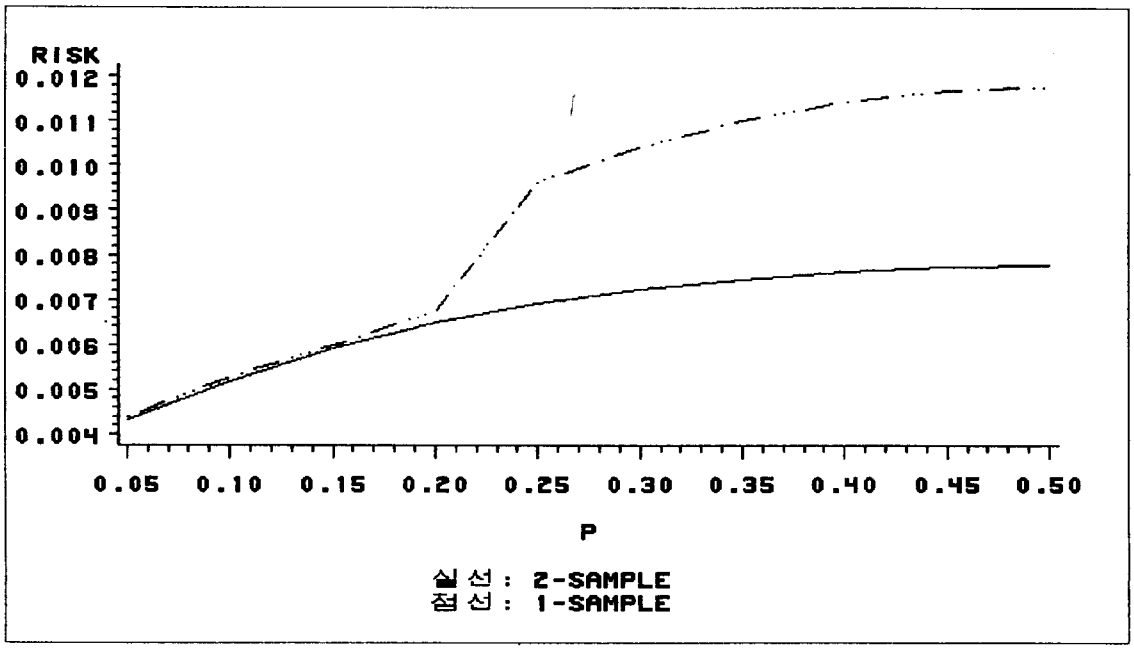

그림1. $c_1 = 0.000068$ 일때의 2-sample 대 1-sample 베이즈 위험

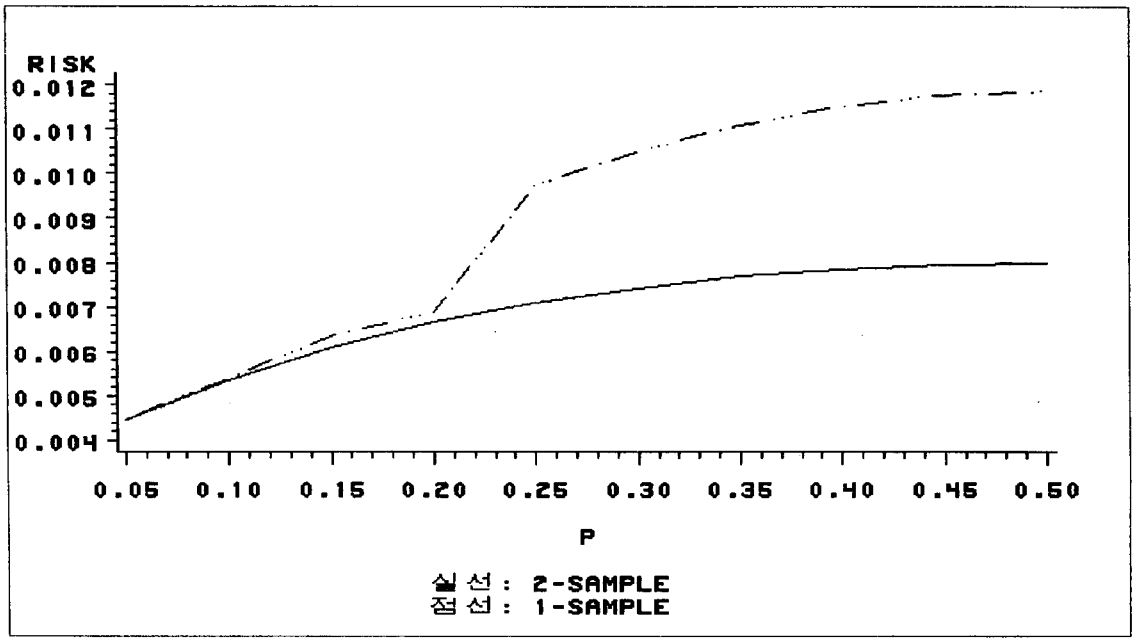

그림2. $c_1 = 0.000072$ 일때의 2-sample 대 1-sample 베이즈 위험

4. 결론

이 논문은 집락표본 추출시 모든 집락의 크기가 같을 경우에 베이즈 위험을 최소로 하는 표본의 수 n_1, n_2 를 구하였다. 그리고, 이-단계 기법을 적용한 경우의 위험이 일회 추출한 경우보다 베이즈 위험이 작아짐을 보였다. 우리는 집락의 조사단위를 M로 고정 시켰는데 각 집락의 단위가 M_i 개로 다른 경우에 대한 연구가 요구된다. 나아가 이-단계 표본추출(two-stage sampling)의 경우로 확장 시켜 연구해야한다.

참 고 문 헌

[1] Cochran (1977). Sampling Technique, John Willy & Sons.

[2] Cohen, A and Sakrowitz, H, B. (1984a). Decision theory results for vector risks with applications, *The Annals of Statistics,* 12, 1035-1049

[3] Cohen, A and Sakrowitz, H, B. (1984b). Results in double sample estimation for the binomial distribution, *The Annals of Statistics,* 12, 1109-1116.

[4] Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors to for exponential families, *The Annals of Statistics,* 7, 269-281.

[5] Freguson, T. S. (1967). Mathematical Statistics, Academic, New York.

[6] Miller , I. and Freund, J. E. (1977). Probability and Statistical for Engineers, Second Ed. Prentice-Hall, Englewood Cliffs, New Jessey.

[7] Joseph, L. Wolfson, D. B and du Berger, R (1995). Sample size calculations for binomial proportion via highest posterior density intervals. *Statistician,* 44, 143-154.

[8] Lehmann, E, L. (1959). Testing Statistical Hypotheses. Willy, New York.

[9] Stein, C. (1945). A two sample test for a linear hypothesis whose power is independent of the variance, *The Annals of Statistics,* 16, 243-258.