응용통계연구 제 6권 2호
1993년 9월 pp.341-353.

신종발견확률의 편의보정 비모수 최우추정량에 관한 연구[1)]

이 주 호[2)]

요 약

여러 개의 종으로 구성된 모집단에서 일정 크기의 표본을 추출하였을 경우, 다음 차례에 뽑힐 종이 새로운 종이 될 조건부확률의 추정량으로서 가장 널리 사용되어 온 것은 Good(1953)이 경험적 베이지안 접근법을 사용하여 제안한 비모수추정량이다. Clayton과 Frees(1987)는 Good의 추정량에 대한 대안으로서 비모수최우추정량을 제안하고, 시뮬레이션을 통해 모집단이 비교적 불균일할 경우 자신들이 제안한 추정량이 Good의 추정량보다 평균제곱오차가 작음을 보여 주었고, Lee(1989)는 모집단이 균등분포에 비교적 가깝지 않은 절단기하분포를 따를 때 이를 점근적으로 규명하였다. 그러나 비모수최우추정량은 상당한 편의를 지니고 있는데, 본 연구에서는 이 편의의 일부를 보정한 새로운 추정량이 대부분의 모집단분포 형태에 있어 비모수최우추정량보다 평균제곱오차가 작으며, 모집단이 균일분포에 아주 가까운 경우를 제외하고는 Good의 추정량보다도 평균제곱오차가 작음을 점근적으로 규명하고, 이를 소표본 시뮬레이션을 통하여 확인하였다.

1. 서 론

생태학을 비롯하여 언어학, 화폐학 등의 분야에서는 여러 개의 종(species)으로 구성된 모집단으로부터 일정 크기의 표본을 추출한 경우 다음번 관측에서 새로운 종을 발견할 확률의 추정에 대한 연구가 오래 전부터 그 빈도는 높지 않으나 지속적으로 진행되어 오고 있다. 이와같이 신종발견확률에 대해 관심을 보이는 이유는 이 확률이 조건부확률이어서 전통적 추정이론의 적용이 곤란하다는 점 외에도, 이 확률이 모집단의 종별 구성비를 요약하는 측도인 균일성지수(diversity index)의 일종으로 사용될 수 있을 뿐만 아니라 신종발견을 위한 축차탐색(sequential search)에 있어 표본관측의 종료시점을 결정하는 데 핵심적인 역할을 하기 때문이다(Rasmussen 과 Starr, 1979).

s개의 종으로 구성된 모집단의 종별 구성비가 각각 $p_1,\cdots,p_s$ ($p_1 + \cdots + p_s = 1$)일 때 이 모집단으로부터 추출한 크기 n인 확률표본내의 k번째 종의 개체수를 X_k라 하면, n+1번째 추출된 개체가 신종일 조건부확률 $U(n)$은 다음과 같이 표시할 수 있다.

$$U(n) = \sum_k p_k I(X_k=0)$$

(여기서 $I(A)$는 A가 진일 경우 1, 위일 경우 0의 값을 갖는 함수임). $U(n)$은 X_k들의 값에 의존하므로 이를 직접 추정할 경우에는 추정량 $\hat{U}(n)$의 평균제곱오차(mean squared error; MSE)를

$$MSE[\hat{U}(n)] = E[\ U(n) - \hat{U}(n)\ ]^2$$

으로 정의하고, 이를 추정량에 대한 평가기준으로 사용함이 바람직할 것이다 (Robbins, 1968).

1) 본 연구는 한국과학재단 연구과제 KOSEF-913-0105-010-1에 의하여 지원받았음
2) 305-764 대전직할시 유성구 궁동 충남대학교 자연과학대학 통계학과 조교수

한편 Starr(1979)는 $U(n)$ 대신에 비조건부확률인 $\theta_n = E[U(n)]$의 추정을 제안하였는데, θ_n의 추정에는 최우추정이나 최소분산불편추정 등의 이론이 적용될 수 있으나, θ_n의 추정량이 과연 $U(n)$을 효과적으로 추정할 수 있는 지에 대해서는 의문의 여지가 있다.

신종발견확률의 추정에 관한 연구는 Good(1953)이 경험적 베이지안 접근법을 사용한 비모수 추정량을 제시하므로써 비롯되었다. Good의 추정량 V_0는

$$V_0 = (1/n) \sum_k I(X_k=1)$$

로 표시되는데, 이는 결국 표본내 1개씩이 포함된 종의 수의 표본크기에 대한 비율을 의미한다. 그 이후 여러학자들의 연구는 몇가지의 다른 추정량 (Hill(1979), Chao(1981) 등)의 제시를 제외하고는 주로 Good의 추정량의 성격을 규명하거나 이를 일반화시키는 데에 집중되어 왔는데, 특히 Starr(1979)는 크기 n 이하의 표본에 근거한 θ_n의 불편추정량은 존재하지 않음을 보이고 V_0를 확장시킨 개념으로 크기 $n+b$인 표본에 근거한 θ_n의 불편추정량

$$V_b = \sum_{i=1}^{b} \binom{b-1}{i-1}\binom{n+b}{i}^{-1} \sum_k I(X_k=i)$$

를 제안하였으며, Clayton 과 Frees(1987)는 V_b가 θ_n의 최소분산불편추정량임을 증명하였다.

한편 Clayton 과 Frees(1987)는 Good의 추정량에 대한 대안으로서 θ_n의 비모수최우추정량인

$$\hat{\theta}_0 = \sum_k (X_k/n)(1-X_k/n)^n$$

을 제안하고 모집단의 구성이 비교적 불균일할 경우 $\hat{\theta}_0$이 Good의 추정량보다 MSE가 작음을 시뮬레이션을 통해 보여 주었으며, Lee(1989)는 모집단이 절단기하분포를 따를 경우 이를 이론적으로 규명하였다. 그러나 Lee의 소표본 시뮬레이션 결과는 비모수최우추정량의 편의를 보정할 경우 MSE를 더욱 줄일 수 있음을 시사하고 있는 바, 이를 이론적으로 규명하는 데 본 연구의 목적이 있다. 다음 절에서는 Good의 추정량과 편의보정 비모수최우추정량의 MSE를 이론적으로 비교하고, 3절에서는 이들을 절단기하분포를 따르는 모집단과 몇 가지의 실제 자료에 대하여 소표본 시뮬레이션을 통해 비교하고자 한다.

2. 점근적 비교

Esty(1983)는 일정조건하에서 $n \to \infty$ 일 때 V_0 가 정규분포로 수렴함을 보였으며, Lee(1989)는 유사한 조건하에서 $n \to \infty$ 일 때 $\hat{\theta}_0$ 도 또한 정규분포로 수렴함을 보였다. 그러나 V_0의 경우와는 달리 $\hat{\theta}_0$의 경우는 점근적 편의가 존재하며, 따라서 두 추정량의 효율을 비교하기 위하여 점근상대효율을 사용할 수 없다. 이 경우에 대안으로 사용될 수 있는 측도로서 극한위험효율(limiting risk efficiency; LRE)이라는 것이 있다 (Lehmann(1983)을 참조).

정의 1 : $g(\theta)$의 두 추정량 $\{\delta_{1n}\}$과 $\{\delta_{2n}\}$의 위험함수 $R(\theta, \delta_{in})$이 어떤 $r > 0$에 대하여 $n \to \infty$ 일 때 $n^r R(\theta, \delta_{in}) \to \tau_i^2 > 0$ (i = 1, 2)이면, $\{\delta_{2n}\}$의 $\{\delta_{1n}\}$에 대한 LRE는 $(\tau_1^2/\tau_2^2)^{1/r}$으로 정의된다.

특히 손실함수가 제곱오차손실함수이고 $r = 2$ 이면 LRE는 바로 두 추정량의 평균제곱오차 자승근(RMSE)의 비의 극한값이 된다. 이 절에서는 모집단이 절단기하분포를 따를 경우에 다음

에 정의되는 편의보정 비모수최우추정량 $\tilde{\theta}_0$의 V_0에 대한 RMSE 비의 극한값을 여러 가지의 모수값에 대해 구해 보고자 하는데, 먼저 다음과 같은 조건을 도입하기로 한다.

조건 1 : a) $m \to \infty$ 에 따라 $n_m \to \infty$.
b) $m \to \infty$ 에 따라 어떤 양의 수 C_1, C_2에 대하여

$$\sum_k n_m p_{km} \exp(-n_m p_{km}) \to C_1,$$
$$\sum_k n_m^2 p_{km}^2 \exp(-n_m p_{km}) \to C_2 \text{ 이고,}$$
$$\sum_k n_m^r p_{km}^{r+1} \exp(-n_m p_{km}) \to 0, \quad r = 2, 3.$$

다음 정리(Lee, 1989)는 조건 1이 만족될 경우 $\hat{\theta}_0$의 V_0에 대한 RMSE의 비가 수렴함을 보여 준다.

정리 1 : 조건 1이 성립할 때 $m \to \infty$ 에 따라 $\{ \hat{\theta}_{0m} \}$의 $\{V_{0m}\}$에 대한 제곱오차손실하에서의 LRE는 τ_1/τ_2가 된다. 여기서

$$\tau_1^2 = C_1 + C_2,$$
$$\tau_2^2 = (e^2-1)^{-1}C_1 + [(e^2-1)^{-2} + 3/4 - (e-1)^{-2}/4 + 2\, e^{-1}(2-e^{-1})^{-2}]\, C_2 + [1 - (e-1)^{-1}]\, C_1^2.$$

Lee는 $\hat{\theta}_0$의 점근적 분포를 유도하는 과정에서 $\hat{\theta}_0$이 점근적 편의

$$b_n = \sum_k p_k [\exp\{(e^{-1}-1)np_k - 1\} - \exp(-np_k)]$$

을 가짐을 보였는데, 이를

$$\hat{b}_n = \sum_k (X_k/n) [\exp\{(e^{-1}-1)X_k - 1\} - \exp(-X_k)]$$

으로 추정하여 보정하면, 편의보정 비모수최우추정량 $\tilde{\theta}_0 = \hat{\theta}_0 - \hat{b}_n$을 얻게 된다. 다음 정리는 조건 1이 만족될 경우 V_0와 $\tilde{\theta}_0$의 RMSE비에 대해서도 정리 1과 유사한 결과가 성립함을 보여 준다.

정리 2 : 조건 1이 성립할 때 $m \to \infty$ 에 따라 $\{ \tilde{\theta}_{0m}\}$의 $\{V_{0m}\}$에 대한 제곱오차손실하에서의 LRE는 τ_1/τ_3가 된다. 여기서

$$\tau_1^2 = C_1 + C_2,$$
$$\tau_3^2 = \{4(e^2-1)^{-1} - 4e^{-1}[\exp(2-e^{-1})-1]^{-1} + e^{-2}[\exp(2(1-e^{-1}))-1]^{-1}\}\, C_1$$
$$+ \{2(1+e^{-2})(e^2-1)^{-2} + 3/4 - (e-1)^{-2} + e^{-2}[\exp(2(1-e^{-1}))-1]^{-2}$$
$$+ 4e^{-1}(2-e^{-1})^{-2} + 4\exp(e^{-1}-3)[2-e^{-1}-\exp(e^{-1}-1)]^{-2} - 4e^{-1}[\exp(2-e^{-1})-1]^{-2}$$
$$- 1/4\, e^{-2}[\exp(1-e^{-1})-1]^{-2} - 2\exp(e^{-1}-2)[2-\exp(e^{-1}-1)]^{-2}\}\, C_2$$
$$+ \{e^{-1}[\exp(1-e^{-1})-1]^{-1} - 2(e-1)^{-1} + 1\}^2\, C_1^2.$$

여기서 과연 조건 1을 만족시키는 분포족이 있는가 하는 의문이 제기되는데 다음 명제(Lee, 1989)는 절단기하분포가 이 조건을 만족시킴을 보여 준다.

명제 1 : $p_k = [(1-p)p^{k-1}]/(1-p^s)$, $0 < p < 1$, $k = 1,\cdots,s$ 일 경우, $n \to \infty$, $s \to \infty$, $np^s \to 0$이면 조건 1이 만족된다.

절단기하분포는 $p \to 1$에 따라 $p_1 = \cdots = p_s = 1/s$ 인 균등분포로 수렴하고, $p \to 0$에 따라 $p_1 = 1$, $p_2 = \cdots = p_s = 0$인 퇴화분포로 수렴하며, p가 1에 가까울수록 모집단의 종별 구성비가 균일해지므로 다양한 형태의 모집단분포를 근사적으로 표시할 수 있는 장점을 지니고 있다 (종의 모집단분포에 대한 근사분포로서의 기하분포의 이론적 및 실증적 타당성에 대한 근거는 Pielou(1975), Engen(1975) 등을 참조). <표 1>은 정리 1 및 정리 2를 이용하여 절단기하분포의 모수 p가 0.5에서 0.9까지의 값을 취할 경우의 RMSE비의 극한값을 계산한 결과이다.

<표 1> 모집단이 모수 p 인 절단기하분포를 따를 경우의 RMSE 비의 극한값

p	$RMSE(V_0)/RMSE(\hat{\theta}_0)$	$RMSE(V_0)/RMSE(\tilde{\theta}_0)$
0.9	0.8480	1.0268
0.8	1.0245	1.1267
0.7	1.1137	1.1673
0.6	1.1686	1.1896
0.5	1.2062	1.2037

<표 1>에서 알 수 있듯이 $p \leq 0.8$인 경우 $\hat{\theta}_0$이 V_0 보다 RMSE가 작으며, p가 커질수록 그 상대적 차이도 커진다. 그러나 p가 0.8보다 클 경우에는 V_0가 $\hat{\theta}_0$보다 RMSE가 작은데 이와같이 p가 1로 접근할수록 $\hat{\theta}_0$의 RMSE가 상대적으로 커지는 이유는 $\hat{\theta}_0$의 편의가 증가함에 기인하는 것으로 보인다. 한편 새로운 추정량 $\tilde{\theta}_0$는 편의를 어느 정도 보정하므로써 $0.8 < p \leq 0.9$인 경우에도 V_0보다 RMSE가 작아지게 하였음을 알 수 있다.

3. 소표본 시뮬레이션 결과

이 절에서는 표본크기 n이 비교적 작을 경우 시뮬레이션을 통해, 먼저 모집단이 모수 p와 s인 절단기하분포를 따를 때 s의 값을 고정시켜 놓고 여러 p 값에 대해 V_0, $\hat{\theta}_0$ 및 $\tilde{\theta}_0$의 RMSE를 구해 이들의 비율이 2절에서 구한 이론적 극한값과 어느 정도 일치하는 가를 살펴 본 후, 세 가지의 비교적 큰 실제 표본집단에 대해 이들 세 추정량의 RMSE를 구해 비교해 보기로 한다. 각 경우의 시뮬레이션 횟수는 2,000번이며 난수는 합동혼합법(mixed congruential generator)을 사용하여 추출하였다.

<표 2>는 모집단이 $p = 0.9(0.1)0.5$, $s = 100$인 절단기하분포를 따를 때 $n = 10, 50, 250$인 경우에서의 V_0에 대한 $\hat{\theta}_0$ 및 $\tilde{\theta}_0$의 RMSE비를 각각 보여 주고 있다. 이 표에서 보면 $\tilde{\theta}_0$는 $\hat{\theta}_0$에 비해 상당히 편의가 작으며 고려된 모든 p 값에 대해서 V_0보다 RMSE가 작다. 한편 $p \leq 0.5$인 경우에는 $\tilde{\theta}_0$가 $\hat{\theta}_0$보다 다소 RMSE가 크나 이 범위에 해당하는 모집단의 분포는 극히 불균일성이 심한 경우로서 현실적으로 발생가능성이 희박하므로 문제가 되지 않을 것으로 보인다. 이 결과를 앞 절의 <표 1>과 비교하여 보면 시뮬레이션에 의한 RMSE비가 이론적 극한값과 매우 가까움을 알 수 있다.

〈표 2〉 모집단이 모수 p 및 s=100인 절단기하분포를 따를 경우의 RMSE 비
($RMSE_1 = RMSE(V_0)/RMSE(\hat{\theta}_0)$, $RMSE_2 = RMSE(V_0)/RMSE(\tilde{\theta}_0)$)

p	n	$E[U(n)]$	$E(V_0)$ $RMSE(V_0)$	$E(\hat{\theta}_0)$ $RMSE(\hat{\theta}_0)$	$E(\tilde{\theta}_0)$ $RMSE(\tilde{\theta}_0)$	$RMSE_1$ $RMSE_2$
0.9	10	0.610	0.6386 0.2347	0.2540 0.3725	0.3720 0.2732	0.6300 0.8589
	50	0.1848	0.1906 0.0791	0.1042 0.0978	0.1385 0.0771	0.8085 1.0255
	250	0.0379	0.0380 0.0171	0.0220 0.0204	0.0282 0.0169	0.8370 1.0100
0.8	10	0.3771	0.4161 0.2360	0.1842 0.2335	0.2639 0.1925	1.0109 1.2261
	50	0.0875	0.0902 0.0566	0.0507 0.0551	0.0656 0.0500	1.0282 1.1325
	250	0.0179	0.0178 0.0121	0.0103 0.0117	0.0133 0.0108	1.0320 1.1196
0.7	10	0.2532	0.2756 0.2123	0.1306 0.1815	0.1826 0.1691	1.1700 1.2553
	50	0.0560	0.0561 0.0456	0.0314 0.0417	0.0406 0.0399	1.0918 1.1429
	250	0.0111	0.0113 0.0094	0.0065 0.0083	0.0084 0.0081	1.1263 1.1635
0.6	10	0.1782	0.1944 0.1845	0.0950 0.1470	0.1300 0.1463	1.2553 1.2510
	50	0.0372	0.0390 0.0380	0.0222 0.0311	0.0287 0.0313	1.2194 1.2131
	250	0.0078	0.0078 0.0079	0.0045 0.0068	0.0058 0.0068	1.1670 1.1728
0.5	10	0.1299	0.1442 0.1527	0.0717 0.1194	0.0964 0.1216	1.2794 1.2558
	50	0.0285	0.0288 0.0317	0.0162 0.0262	0.0209 0.0265	1.2096 1.1982
	250	0.0057	0.0059 0.0067	0.0034 0.0055	0.0043 0.0056	1.2305 1.2054

비록 절단기하분포가 여러 종류의 모집단분포를 근사적으로 표현할 수 있다 하더라도 (Watterson, 1974), 앞에서 고려한 세 추정량의 RMSE를 실제의 모집단분포를 사용하여 비교할 수 있다면 바람직할 것이다. 그러나 실제의 모집단분포를 알 수는 없으므로 여기서는 표본크기 n이 비교적 클 경우의 경험적 분포(empirical distribution)를 사용하여 비교하기로 한다. 여기서 사용된 세 가지의 경험적 분포에 관한 자료가 <표 3>, <표 4>, 그리고 <표 5>에 나타나 있다. <표 3>의 자료는 바닷말의 모집단으로 부터 뽑은 크기 9,629인 표본에 포함된 113종의 개체수를 나타내며, <표 4>의 자료는 자작나무에 기생하는 균류에 서식하는 곤충류 모집단으로부터 뽑은 크기 1,501인 표본에 포함된 72종의 개체수를 나타낸다 (Pielou, 1975). 또한 <표 5>는 Mount Kenya의 특정지역에 서식하는 곤충류 모집단으로부터 뽑은 크기 1,043인 표본내에 포함된 32종의 개체수를 보여 준다 (Lewins와 Joanes, 1984). 각 표에서 f_r은 표본내 포함된 개체수가 r인 종의 수를 나타낸다.

<표 3> 바닷말 모집단의 표본내 개체수

r	f_r	r	f_r	r	f_r	r	f_r	r	f_r
1	16	9	6	24	1	67	1	272	2
2	10	10	9	26	1	75	1	408	1
3	5	11	3	32	1	82	1	640	1
4	10	17	6	33	4	110	1	2960	1
5	8	19	3	36	2	124	1	3032	1
6	1	20	4	37	1	147	1		
7	1	21	2	38	1	184	1		
8	2	23	1	62	1	192	1		

<표 4> 자작나무 균류에 서식하는 곤충류 모집단의 표본내 개체수

r	f_r	r	f_r	r	f_r	r	f_r	r	f_r
1	31	8	1	20	1	67	1	114	1
2	3	10	2	25	1	71	1	142	1
3	2	12	1	32	1	84	1	196	1
4	5	15	1	33	1	87	1		
5	1	16	1	40	1	91	1		
6	2	17	1	48	1	97	1		
7	3	18	1	49	1	98	1		

<표 5> Mount Kenya 지역에 서식하는 곤충류 모집단의 표본내 개체수

r	f_r	r	f_r	r	f_r	r	f_r	r	f_r
1	8	5	1	12	1	46	1	109	1
2	3	6	3	18	1	56	1	157	1
3	2	7	2	21	1	95	1	335	1
4	1	10	1	25	1	98	1		

이들 자료를 모집단분포로 가정하여 n = 50, 250의 크기로 각각 시뮬레이션한 결과가 <표 6>에 요약되어 있다. 편의상 <표 3>, <표 4>, <표 5>에 요약된 경험적 분포를 각각 모집단 A, B, C라 하면, 모집단 A가 상대적으로 가장 균일하고 모집단 C가 가장 불균일함을 알 수 있는데, 이에 따라 RMSE의 크기도 모집단 A에서는 V_0가, 모집단 C에서는 $\tilde{\theta}_0$가 각각 더 작게 나타나고 있으며, 중간적 형태인 모집단 B에서는 두 추정량의 RMSE가 서로 비슷한 크기를 보여 주고 있다. 한편 $\hat{\theta}_0$은 모집단 C에서도 V_0보다 RMSE가 크게 나타나 있는데, 그 이유는 아마도 모집단분포의 꼬리부분에 해당하는 희귀종의 구성비가 경험적 분포에서는 상대적으로 보다 균일하게 왜곡되어 나타나기 때문으로 추측된다.

<표 6> 모집단 A, B, C에 있어서의 RMSE 비
($RMSE_1 = RMSE(V_0)/RMSE(\hat{\theta}_0)$, $RMSE_2 = RMSE(V_0)/RMSE(\tilde{\theta}_0)$)

모집단	n	$E[U(n)]$	$E(V_0)$ $RMSE(V_0)$	$E(\hat{\theta}_0)$ $RMSE(\hat{\theta}_0)$	$E(\tilde{\theta}_0)$ $RMSE(\tilde{\theta}_0)$	$RMSE_1$ $RMSE_2$
A	50	0.1704	0.1722 0.0701	0.0762 0.1031	0.1075 0.0796	0.6802 0.8804
	250	0.0706	0.0705 0.0209	0.0338 0.0393	0.0464 0.0289	0.5322 0.7251
B	50	0.1668	0.1688 0.0767	0.0928 0.0915	0.1228 0.0743	0.8385 1.0316
	250	0.0392	0.0395 0.0172	0.0221 0.0213	0.0289 0.0173	0.8068 0.9948
C	50	0.0911	0.0913 0.0542	0.0475 0.0573	0.0628 0.0504	0.9471 1.0766
	250	0.0239	0.0241 0.0137	0.0136 0.0145	0.0179 0.0126	0.9447 1.0832

4. 결 론

신종발견확률에 대한 Good(1953)의 추정량은 표본크기가 비교적 클 경우 이 확률의 기대값에 대한 최소분산불편추정량에 매우 가까우므로 가장 선호되어 왔다. 그러나 Good의 추정량이 신종발견확률 자체에 대해서도 작은 MSE를 갖는다는 보장은 없으며, 또한 얼마간의 편의를 감수한다면 이 추정량보다 MSE가 작은 추정량을 찾을 가능성도 있다. Clayton과 Frees(1987)가 제안한 비모수최우추정량은 이러한 관점에서 Good의 추정량에 대한 하나의 유용한 대안이라 할 수 있다. 그러나 비모수최우추정량이 지닌 편의는 비교적 크며, 현실적으로 널리 존재할 수 있는 정도의 균일성을 지닌 모집단에 있어서 Good의 추정량보다 MSE가 큰 경우가 많다고 하겠다. 본 연구에서는 비모수최우추정량의 편의를 얼마간 보정하므로써 보다 광범위한 정도의 모집단 균일성에 대해 이 추정량이 Good의 추정량에 비해 MSE면에서 상대적으로 우월하다는 것을 이론적으로 규명하고 소표본 시뮬레이션을 통하여 이를 확인하였다. 실제의 모집단으로부터 일정 크기의 표본을 추출한 경우, 다음번 관측에서 신종을 발견할 확률을 추정하기 위해 Good의 추정량과 편의보정 비모수최우추정량중 어떤 것을 선택할 것인가는 표본으로부터 얻은 모집단의 균일성에 대한 정보를 활용하여 결정할 수 있을 것이다.

참 고 문 헌

[1] Chao, A. (1981), "On Estimating the Probability of Discovering a New Species," *The Annals of Statistics*, **9**, 1339-1342.

[2] Clayton, M. K. and Frees, E. W. (1987), "Nonparametric Estimation of the Probability of Discovering a New Species," *Journal of the American Statistical Association*, **82**, 305-311.

[3] Engen, S. (1975), "A Note on the Geometric Series as a Species Frequency Model," *Biometrika*, **62**, 697-699.

[4] Esty, W. W. (1983), "A Normal Limit Law for a Nonparametric Estimator of the Coverage of a Random Sample," *The Annals of Statistics*, **11**, 905-912.

[5] Good, I. J. (1953), "On the Population Frequencies of Species and the Estimation of Population Parameters," *Biometrika*, **40**, 237-264.

[6] Hill, B. M. (1979), "Posterior Moments of the Number of Species in a Finite Population and the Posterior Probability of Finding a New Species," *Journal of the American Statistical Association*, **74**, 668-673.

[7] Lee, J. (1989), On Asymptotics for the NPMLE of the Probability of Discovering a New Species and an Adaptive Stopping Rule in Two- Stage Searches, *Ph.D. Thesis*, Department of Statistics, University of Wisconsin, Madison.

[8] Lehmann, E. L. (1983), *Theory of Point Estimation*, Wiley, New York.

[9] Lewins, W. A. and Joanes, D. N. (1984), "Bayesian Estimation of the Number of Species," *Biometrics*, **40**, 323-328.

[10] Pielou, E. C. (1975), *Ecological Diversity*, Wiley, New York.

[11] Rasmussen, S. and Starr, N. (1979), "Optimal and Adaptive Stopping in the Search for the New Species," *Journal of the American Statistical Association*, **74**, 661-667.

[12] Robbins, H. (1968), "Estimating the Total Probability of the Unobserved Outcomes of an Experiment," *Annals of Mathematical Statistics*, **39**, 256-257.

[13] Starr, N. (1979), "Linear Estimation of the Probability of Discovering a New Species," *Annals of Statistics*, **7**, 644-652.

[14] Watterson, G. A. (1974), "Models for the Logarithmic Species Abundance Distribution," *Theory Pop. Biology*, **6**, 217-250.

부 록 : 정리 2 의 증명

증명에서는 편의상 아래첨자 m 은 생략하기로 한다. $m \to \infty$ 에 따라 $n^2E[\tilde{\theta}_0 - U(n)]^2 \to \tau_3^2$ 임을 보이면 되는데,

$$n^2E[\tilde{\theta}_0 - U(n)]^2 = n^2E[\hat{\theta}_0 - \hat{b}_n - U(n)]^2$$

$$= n^2E[\hat{\theta}_0 - U(n)]^2 + n^2E(\hat{b}_n^2) - 2n^2E[\hat{b}_n(\hat{\theta}_0 - U(n))]$$

이므로 $n^2E(\hat{b}_n^2)$과 $n^2E[\hat{b}_n(\hat{\theta}_0-U(n))]$의 극한값을 각각 구하면 주어진 결과를 얻을 수 있다. 먼저 $n^2E(\hat{b}_n^2)$의 극한값은

$$\begin{aligned} n^2E(\hat{b}_n^2) = & \sum_k E\{X_k^2[\exp((e^{-1}-1)X_k-1) - \exp(-X_k)]^2\} \\ & + \sum_{k\neq l} E\{X_kX_l[\exp((e^{-1}-1)X_k-1) - \exp(-X_k)\} \\ & \cdot\{\exp((e^{-1}-1)X_l-1) - \exp(-X_l)]\} \end{aligned}$$

이 되는데, 편의상 위의 등식 우측의 두 $\sum$기호 안에 있는 항들을 각각 A_k와 B_{kl}이라 하면,

$$\begin{aligned} A_k &= \sum_k j^2\{\exp[(e^{-1}-1)j-1]-e^{-j}\}^2 P(X_k=j) \\ &= \sum_k j^2\{\exp[(e^{-1}-1)j-1]-e^{-j}\}^2 \binom{n}{j} p_k^j (1-p_k)^{n-j} \\ &= n\sum_k \binom{n-1}{j-1} \{\exp[(e^{-1}-1)j-1]-e^{-j}\}^2 p_k^j (1-p_k)^{n-j} \\ &\quad + n(n-1)\sum_k \binom{n-2}{j-2} \{\exp[(e^{-1}-1)j-1]-e^{-j}\}^2 p_k^j (1-p_k)^{n-j} \\ &= n[\exp[2(e^{-1}-2)]p_k\{1-[1-\exp(2(e^{-1}-1))]p_k\}^{n-1} + e^{-2}p_k[1-(1-e^{-2})p_k]^{n-1} \\ &\quad - 2\exp(e^{-1}-3)p_k\{1-[1-\exp(e^{-1}-2)]p_k\}^{n-1}] + n(n-1)[\exp(4e^{-1}-6)p_k^2 \\ &\quad \cdot\{1-[1-\exp(2(e^{-1}-1))]p_k\}^{n-2} + e^{-4}p_k^2[1-(1-e^{-2})p_k]^{n-2} \\ &\quad - 2\exp(2e^{-1}-5)p_k^2[1-(1-e^{-2})p_k]^{n-2}] \end{aligned}$$

이 된다. 여기서 다음과 같은 보조정리(Lee, 1989)를 도입하기로 한다.

보조정리 A.1 : $m \to \infty$ 에 따라 $n_m \to \infty$ 이면 임의의 양수 α와 β에 대하여 다음이 성립한다 (여기서 ~는 극한값이 서로 같음을 의미함).

$$\begin{aligned} &\sum_k np_k(1-\alpha p_k)^n \sim \sum_k np_k\exp(-\alpha np_k), \\ &\sum_k n^2p_k^2(1-\alpha p_k)^n \sim \sum_k n^2p_k^2\exp(-\alpha np_k), \\ &\sum_{k\neq l} n^2p_kp_l(1-\alpha p_k-\beta p_l)^n \sim \sum_{k\neq l} n^2p_kp_l\exp(-\alpha np_k-\beta np_l). \end{aligned}$$

한편 조건 1로부터 임의의 양수 α와 β 및 $r = 0, 1, 2, 3$에 대하여 $m \to \infty$ 에 따라 다음과 같은 관계를 얻게 된다.

$$\sum_k np_k\exp(-\alpha np_k) = (1/\alpha)\sum_k \alpha np_k\exp(-\alpha np_k) \to C_1/\alpha, \tag{A.1}$$

$$\sum_k n^2p_k^2\exp(-\alpha np_k) = (1/\alpha^2)\sum_k (\alpha n)^2p_k^2\exp(-\alpha np_k) \to C_2/\alpha^2, \tag{A.2}$$

$$\begin{aligned} \sum_{k\neq l} n^{r+1}p_k^{r+1}p_l\exp(-\alpha np_k-\beta np_l) &\le \sum_k n^rp_k^{r+1}p_l\exp(-\alpha np_k)\sum_l np_l\exp(-\beta np_l) \\ &\to 0\cdot C_1/\beta = 0, \end{aligned} \tag{A.3}$$

$$\begin{aligned} &\sum_{k\neq l} n^{r+2}p_k^{r+1}p_l^2\exp(-\alpha np_k-\beta np_l) \\ &\quad \le \sum_k n^rp_k^{r+1}p_l\exp(-\alpha np_k)\sum_l n^2p_l^2\exp(-\beta np_l) \\ &\quad \to 0\cdot C_2/\beta^2 = 0. \end{aligned} \tag{A.4}$$

따라서 보조정리 A.1과 조건 1 및 (A.1), (A.2)를 이용하면 $m \to \infty$ 에 따라

$$\begin{aligned}\sum_k A_k \sim\ & \exp[2(e^{-1}-2)] \sum_k np_k \exp\{[\exp(2(e^{-1}-1))-1]np_k\} + e^{-2} \sum_k np_k \exp[(e^{-2}-1)np_k] \\ & - 2\exp(e^{-1}-3) \sum_k np_k \exp\{[\exp(e^{-1}-2))-1]np_k\} + \exp(4e^{-1}-6) \sum_k n^2 p_k^2 \\ & \cdot \exp\{[\exp(2(e^{-1}-1))-1]np_k\} + e^{-4} \sum_k n^2 p_k^2 \exp[(e^{-2}-1)np_k] \\ & - 2\ \exp(2e^{-1}-5) \sum_k n^2 p_k^2\ \exp\{[\exp(e^{-1}-2))-1]np_k\} \\ & \to \{e^{-2}[\exp(2(1-e^{-1}))-1]^{-1} + (e^2-1)^{-1} - 2e^{-1}[\exp(2-e^{-1})-1]^{-1}\} C_1 \\ & + \{e^{-2}[\exp(2(1-e^{-1}))-1]^{-2} + (e^2-1)^{-2} - 2e^{-1}[\exp(2-e^{-1})-1]^{-2}\} C_2 \end{aligned} \tag{A.5}$$

이 된다. 또한

$$\begin{aligned} B_{kl} =\ & \sum_{i,j} ij \{\exp[(e^{-1}-1)i-1] - e^{-i}\} \{\exp[(e^{-1}-1)j-1] - e^{-j}\} \binom{n}{i,j} p_k^i p_l^j \\ & \cdot (1-p_k-p_l)^{n-i-j} \\ =\ & n(n-1) \Big\{ e^{-2} \sum_{i,j} \binom{n-2}{i-1,j-1} [\exp(e^{-1}-1)p_k]^i [\exp(e^{-1}-1)p_l]^j (1-p_k-p_l)^{n-i-j} \\ & - e^{-1} \sum_{i,j} \binom{n-2}{i-1,j-1} [\exp(e^{-1}-1)p_k]^i (e^{-1}p_l)^j (1-p_k-p_l)^{n-i-j} \\ & - e^{-1} \sum_{i,j} \binom{n-2}{i-1,j-1} (e^{-1}p_k)^i [\exp(e^{-1}-1)p_l]^j (1-p_k-p_l)^{n-i-j} \\ & + \sum_{i,j} \binom{n-2}{i-1,j-1} (e^{-1}p_k)^i (e^{-1}p_l)^j (1-p_k-p_l)^{n-i-j} \Big\} \\ =\ & n(n-1)\{\exp[2(e^{-1}-2)]p_k p_l [1-(1-\exp(e^{-1}-1))(p_k+p_l)]^{n-2} - \exp(e^{-1}-3)p_k p_l \\ & \cdot [1-(1-\exp(e^{-1}-1))p_k - (1-e^{-1})p_l]^{n-2} - \exp(e^{-1}-3)p_k p_l \\ & \cdot [1-(1-e^{-1})p_k - (1-\exp(e^{-1}-1))p_l]^{n-2} + e^{-2}p_k p_l [1-(1-e^{-1})(p_k+p_l)]^{n-2}\} \end{aligned}$$

이므로 A_k의 경우와 마찬가지로 보조정리 A.1과 조건 1 및 (A.1), (A.2)를 이용하여 $m \to \infty$ 에 따라 다음이 성립함을 보일 수 있다.

$$\begin{aligned} \sum_{k \neq l} B_{kl} \to\ & \{e^{-1}\ [\exp(1-e^{-1})-1]^{-1} - (e-1)^{-1}\}^2\ C_1^2 - \{1/4\ e^{-2}[\exp(1-e^{-1})-1]^{-2} \\ & + 1/4(e-1)^{-2} - 2\ \exp(e^{-1}-3)\ [2-e^{-1}-\exp(e^{-1}-1)]^{-2}\}\ C_2 \end{aligned} \tag{A.6}$$

다음으로 $n^2\ E[\hat{b}_n(\hat{\theta}_0 - U(n))]$의 극한값은

$$\begin{aligned} n^2\ E[b_n(\hat{\theta}_0 - \hat{U}(n))] =\ & E[\sum_k X_k \{\exp[(e^{-1}-1)X_k - 1] - \exp(-X_k)\} \sum_l \{X_l(1-X_l/n)^n \\ & - n\ p_l\ I(X_l=0)\}] \\ =\ & \sum_k E[X_k^2(1-X_k/n)^n \{\exp[(e^{-1}-1)X_k - 1] - \exp(-X_k)\}] \\ & + \sum_{k \neq l} E[X_k X_l (1-X_l/n)^n \{\exp[(e^{-1}-1)X_k - 1] - \exp(-X_k)\}] \\ & - \sum_{k \neq l} n\ p_l\ E[X_k \{\exp[(e^{-1}-1)X_k - 1] - \exp(-X_k)\}\ I(X_l=0)] \end{aligned}$$

이 되는데, 여기서 편의상 $\sum$ 기호안에 있는 항들을 차례대로 각각 C_k, D_{kl}, E_{kl}이라 놓고, $\sum_k C_k$ 와 $\sum_{k \neq l} D_{kl}$을 구하기 위해서 다음의 보조정리(Lee, 1989)를 도입하기로한다.

보조정리 A.2 : $\lim_{n\to\infty} \sup_{0<x\le n} |n[e^{-x} - (1-x/n)^n] - 1/2\, x^2 e^{-x}| = 0$.

보조정리 A.2로부터 임의의 $\varepsilon > 0$에 대하여 n이 충분히 크면

$$1/2\, j^2 e^{-j} - \varepsilon \le n[e^{-j} - (1-j/n)^n] \le 1/2\, j^2 e^{-j} + \varepsilon$$

이므로 충분히 큰 m에 대하여

$$\begin{aligned} C_k &= \sum_j j^2 (1-j/n)^n \{\exp[(e^{-1}-1)j-1] - e^{-j}\} \binom{n}{j} p_k^j (1-p_k)^{n-j} \\ &\ge \sum_j j^2 [e^{-j} - j^2 e^{-j}/(2n) - \varepsilon/n] \{\exp[(e^{-1}-1)j-1] - e^{-j}\} \binom{n}{j} p_k^j (1-p_k)^{n-j} \end{aligned} \quad (A.7)$$

이 된다. 여기서

$$\begin{aligned} F_k &\equiv \sum_j j^2 e^{-j} \{\exp[(e^{-1}-1)\, j-1] - e^{-j}\} \binom{n}{j} p_k^j (1-p_k)^{n-j} \\ &= n \sum_j \binom{n-1}{j-1} \{\exp[(e^{-1}-1)\, j-1] - e^{-j}\} p_k^j (1-p_k)^{n-j} + n(n-1) \sum_j \binom{n-2}{j-2} \\ &\quad \cdot \{\exp[(e^{-1}-1)\, j-1] - e^{-j}\} p_k^j (1-p_k)^{n-j} \\ &= n \{\exp(e^{-1}-3)\, p_k [1 - (1 - \exp(e^{-1}-2) p_k]^{n-1} - e^{-2} p_k [1 - (1-e^{-2}) p_k]^{n-1}\} \\ &\quad + n(n-1) \{\exp(2e^{-1}-5)\, p_k^2 [1 - (1 - \exp(e^{-1}-2)\, p_k]^{n-2} - e^{-4} p_k^2 \\ &\quad \cdot [1 - (1-e^{-2}) p_k]^{n-2}\} \end{aligned}$$

이므로 보조정리 A.1과 조건 1, 그리고 (A.1), (A.2)를 이용하여 $m \to \infty$ 에 따라

$$\begin{aligned} \sum_k F_k \to \{e^{-1}[\exp(2-e^{-1})-1]^{-1} - (e^2-1)^{-1}\}\, C_1 + \{e^{-1}[\exp(2-e^{-1})-1]^{-2} \\ - e^{-2}(e^2-1)^{-2}\}\, C_2 \end{aligned} \quad (A.8)$$

임을 쉽게 보일 수 있다. 한편 (A.7)에서 F_k를 제외한 항을 $R_k(n)$이라 정의하면 (A.3)와 (A.4)를 이용하여 $m \to \infty$ 에 따라 $\sum_k R_k(n) \to 0$ 임을 보일 수 있다. 그런데 $C_k \le F_k$이므로 $\sum_k C_k$의 극한값은 결국 $\sum_k F_k$의 극한값과 같고 따라서 (A.8)에 있는 값이 된다. 또한

$$D_{kl} = \sum_{i,j} ij \{\exp[(e^{-1}-1)\, i-1] - e^{-1}\} (1-j/n)^n \binom{n}{i,j} p_k^i p_l^j (1-p_k-p_l)^{n-i-j}$$

이므로 보조정리 A.1 및 A.2, 조건 1, 그리고 (A.1) - (A.4)를 이용하여 $\sum_k C_k$의 극한값을 구한 방법과 유사한 방법으로 다음이 성립함을 보일 수 있다.

$$\begin{aligned} \sum_{k\ne l} D_{kl} \to \{e^{-1}(e-1)^{-1}[\exp(1-e^{-1})-1]^{-1} - (e-1)^{-2}\}\, C_1^2 - \{\exp(e^{-1}-3) \\ \cdot [2 - e^{-1} - \exp(e^{-1}-1)]^{-2} - 1/4\,(e-1)^{-2}\} C_2 \,. \end{aligned} \quad (A.9)$$

끝으로 $\sum_{k\neq l} E_{kl}$의 극한값은 보조정리 A.1, 조건 1, 그리고 (A.1), (A.2)로부터 $\sum_{k\neq l} B_{kl}$의 경우와 마찬가지로 다음과 같이 됨을 쉽게 보일 수 있다.

$$\sum_{k\neq l} E_{kl} \rightarrow \{e^{-1}[\exp(1-e^{-1})-1]^{-1}(e-1)^{-1}\} C_1^2 - \{\exp(e^{-1}-2) \cdot [2-\exp(e^{-1}-1)]^{-2} e^{-1}(2-e^{-1})^2\} C_2 . \quad \text{(A.10)}$$

이상에서 구한 (A.5), (A.6), (A.8) - (A.10)의 값을 τ_2^2에 더해 주면 정리 2에 주어진 τ_3^2 값을 얻을 수 있다. ■

On Asymptotics for a Bias-Corrected Version of the NPMLE of the Probability of Discovering a New Species[1)]

Jooho Lee[2)]

Abstract

As an estimator of the conditional probability of discovering a new species at the next observation after a sample of certain size is taken, the one proposed by Good(1953) has been most widely used. Recently, Clayton and Frees(1987) showed via simulation that their nonparametric maximum likelihood estimator(NPMLE) has smaller MSE than Good's estimator when the population is relatively nonuniform. Lee(1989) proved that their conjecture is asymptotically true for truncated geometric population distributions. One shortcoming of the NPMLE, however, is that it has a considerable amount of negative bias. In this study we proposed a bias-corrected version of the NPMLE and showed that the new estimator has a smaller asymptotic MSE than the NPMLE for virtually all realistic population distributions. We also showed that it has a smaller asymptotic MSE than Good's estimator except when the population is very uniform. A Monte Carlo simulation was performed for small sample sizes, and the result supports the asymptotic results.

1) This research was supported by the Korea Science & Engineering Foundation Grant KOSEF 913-0105-010-1.

2) Department of Statistics, Chungnam National University, Daejeon, 305-764