碩士學位請求論文

非復元 確率比例 抽出과 最大 엔트로피

指導敎授 金 益 贊

濟州大學校 敎育大學院

數學敎育專攻

金 英 熙

1996年 8月

非復元 確率比例 抽出과 最大 엔트로피

指導敎授 金 益 贊

이 論文을 敎育學 碩士學位 論文으로 提出함

1996年 6月 日

濟州大學校 敎育大學院 數學敎育專攻

提出者 金 英 熙

金英熙의 敎育學 碩士學位 論文을 認准함

1996年 7月 日

審査委員長 방 은 숙 印
審 査 委 員 [illegible] 印
審 査 委 員 金 益 贊 印

〈초 록〉

비복원 확률비례 추출과 최대 엔트로피

김 영 희

제주대학교 교육대학원 수학교육전공

지도교수 김 익 찬

본 연구에서는 놀라움의 개념을 양(量)으로 나타내려는 시도로서 엔트로피를 정의하여, 표본추출의 우연성을 최대화시키는 방안으로 엔트로피의 개념을 활용하였다. 즉 불균등 확률을 갖는 크기 N인 모집단에서 n개의 표본들로 이루어진 표본분포가 최대의 우연성을 가지게 하는 최대 엔트로피의 모형을 채택하고, 포함확률과 가중치 사이의 방정식을 세워 그 해법을 수치해석의 고정점 반복 절차에 의해 제시한다.

그리고 불균등 확률의 대표적 사례인 비복원 확률비례 추출 방법에서 해법을 적용한다.

목 차

I. 서 론

정보이론에서 내일 아침 해가 동쪽에서 뜬다라는 통보는 해가 항상 동쪽에서 뜨므로 정보를 가지고 있지 않다. 그러나 동쪽이 아니고 다른 방향에서 해가 뜬다면 이 통보는 많은 정보를 가지고 있다고 말할 수 있다. 확률적으로는 확실한 사상, 즉 확률이 1인 사상의 정보량은 0이며 발생확률이 0인 사상의 정보량은 무한대가 된다. 또한, 일기예보를 통보하는 기상대에서 내일의 일기예보를 <맑음>, <흐림>, <비가 옴>이라는 3개의 사상중에 하나를 택하여 예보를 한다고 가정할 때 내일 비가 온다는 예보는 비가 올 확률이 많다고 생각할 수 있고, 어느 정도의 정보가 들어 있으며 <맑음>, <흐림>, <비가 옴>은 서로 배반사건이고, 각각 통보에 대한 확률이 p_1, p_2, p_3일 때 각 통보에 대한 정보의 량을 규정할 수 있다. 사상이 여러개 존재하면 각정보량의 평균을 엔트로피(entropy)로 정의한다.

본 논문은 2장에서 놀라움의 개념에서 엔트로피를 정의하고

3장은 N개의 단위를 가진 유한 모집단에서 비복원으로 불균등 확률에 의해서 표본 추출을 하는 방법을 생각해 본다. 이 방법은 경품 추첨의 한 모형으로 Stern과 Cover[5]에 의해서 최초로 제안되었으며, 포함 확률 π_i가 주어진다는 조건하에서 엔트로피(entropy)를 최대화함으로서 특성화될 수 있다. 또는 일련의 가중치w_i들의 누승에 비례하는 선택된 표본의 확률 $p(x)$를 구함으로써 이 방법은 해결되어질 수 있다.

4장에서는 π_i가 주어졌을 때 w_i의 유일하고 π_i로부터 w_i를 계산하는 algorithm의 수렴성을 보일 것이다.

5장은 2계 순위의 포함 확률 π_{ij}가 $0 < \pi_{ij} < \pi_i\pi_j$의 조건을 만족한다는 사실을 입증한다.

마지막으로 6장에서 이 불균등 확률에 의한 표본 추출의 대표적 사례인 확률비례추출 방법을 적용하여 본 논문의 결과를 SAS의 Computer Program에 의하여 실례를 제시할 것이다.

Ⅱ. 엔트로피와 불확실성

어떤 실험 또는 관찰이 실시될 때 사상 E가 발생했다는 소식에 의해 야기되는 놀라움의 정도는 E가 발생할 확률에 좌우된다고 가정하는 것이 타당하다. 예를 들면, 한쌍의 주사위를 던지는 실험에서 E를 눈의 합이 짝수인 사상이라 할 때, E가 발생했다는 소리를 듣고 크게 놀라지 않을 것이다. 왜냐하면 E가 발생할 확률이 $\frac{1}{2}$ 이기 때문이다. 그러나 E가 눈의 합이 12인 사상이라 하면, E가 발생했다는 소리를 들으면 확실히 좀더 놀라게 될 것이다. 이 때는 E가 발생할 확률이 $\frac{1}{36}$이다.

이제 놀라움의 개념을 양으로 나타내려는 시도를 하려고한다. 이를 위해 사상 E가 발생했다는 것을 듣는 즉시 느끼는 놀라움은 단지 E의 확률에만 좌우된다고 가정하고, $P(E) = p$인 사상이 발생함으로 야기되는 놀라움을 $S(p)$로 나타내자. 먼저 $S(p)$가 만족할 합리적인 조건들을 결정한후, 이 공리들에 의해 $S(p)$의 함수적인 형태를 결정할 수 있다. 앞으로 항상 $S(p)$는 $p = 0$인 사상에 대해서는 정의되지 않고 $0 < p \leq 1$을 만족하는 p에 대해 정의 된다고 가정한다.

첫째는 확실히 발생할 사상에 대하여서는 놀라움이 없다는 직관적인 것이다.

공리 1 $S(1) = 0$

둘째는 발생할 가능성이 희박한 사상이 발생하면 발생할수록 더 놀란다는 것을 의미한다.

공리 2 $S(p)$는 p의 강한 감소함수(strictly decreasing function)이다.

즉, $p < q$ 이면 $S(p) > S(q)$이다.

세째는 p에서의 작은 변화에 대응하여 $S(p)$에서의 작은 변화가 직관적으로 기대된다는 사실의 수학적 서술이다.

공리 3 $S(p)$는 p의 연속함수이다.

마지막으로, 두개의 독립사상 E와 F의 확률이 각각 $P(E) = p, P(F) = q$라 하자. $P(EF) = pq$이므로 E와 F가 둘다 발생했다는 소식에 의해 야기되는 놀라움은 $S(pq)$이고, E가 발생한 후에 F가 발생했다고 가정하자. $S(p)$는 E의 발생에 의한 놀라움이므로, $S(pq)-S(p)$는 F가 또한 발생했음을 알았을 때 생긴 추가된 놀라움이다. 그러나 E와 F가 독립이기 때문에 추가된 놀라움은 단지$S(q)$이다. 이러한 사실로 마지막 공리를 제시할 수 있다.

공리 4 $S(pq) = S(p) + S(q), \quad 0 < p \leq 1, \quad 0 < q \leq 1$

이제 $S(p)$의 구조를 유도해 보자.

정리 1

만약 $S(\cdot)$가 공리1에서 공리4까지 만족한다면,

$$S(p) = -c \log_2 p \text{ , 여기서 } c\text{는 임의의 양의 정수이다.}$$

(증명) 공리 4로부터 $S(p^2) = S(p) + S(p) = 2S(p)$이고 귀납법에 의해

$$S(p^m) = mS(p) \tag{1}$$

또한, 임의의 정수 n에 대해 $S(p) = S(p^{\frac{1}{n}} \cdots p^{\frac{1}{n}}) = nS(p^{\frac{1}{n}})$이므로

$$S(p^{\frac{1}{n}}) = \frac{1}{n} S(p) \tag{2}$$

가 성립한다. 따라서 식 (1), (2)로부터

$$S(p^{\frac{m}{n}}) = mS(p^{\frac{1}{n}}) = \frac{m}{n}S(p)$$

이므로, x가 양의 유리수이면

$$S(p^x) = xS(p) \tag{3}$$

이다. 한편, S가 공리 3에 의해 연속함수이므로, 음이 아닌 모든 x에 대해 식(3)이 성립된다.

이제 임의의 $p(0 < p \leq 1)$에 대하여

$x = -\log_2 p$라 두면 $p = (\frac{1}{2})^x$

이 되고, 식(3)으로부터

$$S(p) = S((\tfrac{1}{2})^x) = xS(\tfrac{1}{2}) = -c\log_2 p$$

가 된다. 여기서 공리 1과 2이 의해서 $c = S(\frac{1}{2}) > S(1) = 0$.

c를 1로 두는 것이 일반적이며, 이경우에 놀라움은 비트(bit)의 단위로 표시된다.

$x_1, x_2, \cdots, x_n$중 하나를 취하며 각각의 확률이$p_1, p_2, \cdots, p_n$인 확률변수 x를 생각하자. 그리고 $\log_2 x$를 $\log x$로 표시하고 $\log_e x$를 $\ln x$로 표시한다. $-\log p_i$는 X가 x_i값을 갖는 경우 야기되는 놀라움이기 때문에, X의 값을 알게됨으로 받게될 놀라움의 기대정도는

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

가 된다.

이 값 $H(X)$는 정보이론에서 확률변수 X의 엔트로피(entropy)로 알려져 있다.

예) 한개의 동전을 던져 앞면이 나올 확률이 p이고 뒷면이 나올 확률이 $1-p$일 때 엔트로피는 $H(X) = -p\log p - (1-p)\log(1-p)$이다.

엔트로피의 최대치를 구하기 위하여 p_n을 다른 확률에 의존하는 종속변수로 보면 $p_n = 1-(p_1+p_2+\cdots+p_k+\cdots+p_{n-1})$이 되고 p_k와 p_n을 제외하고 다른 확률은 상수로 보고 p_k에 대하여 미분하면

$\frac{dH}{dp_k} = \frac{d}{dp_k}(-p_k\log p_k - p_n\log p_n)$이 되고

$\frac{d}{dx}\log u = \frac{1}{u}\log e\frac{du}{dx}$를 이용하면

$$\frac{dH}{dp_k} = -p_k\frac{1}{p_k}\log e - \log p_k + p_n\frac{1}{p_n}\log e + \log p_n$$
$$= \log\frac{p_n}{p_k}$$

이것은 $p_k = p_n$일 때 0이 되고 p_k는 임의로 선택하였으므로 $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$ 로서, 모든 p_i들이 같을 때 $H(X)$가 최대가 된다.

$H(X)$는 X의 값을 알게됨으로써 받는 놀라움의 평균량이므로, X의 값에 대해 존재하는 불확실성(uncertainty)의 양으로 해석될 수 있다. 사실, 정보이론에서 $H(X)$는 X의 값이 관측되었을 때 받아들여지는 정보(information)의 평균량으로 해석된다. 즉 X에 의한 평균 놀라움, X의 불확실성, 또는 X에 의해 나타난 정보의 평균량은 모두 약간 다른 관점에서 고찰한 동일한 개념이다. 따라서 본 논문에서 제시되는 엔트로피는 표본추출의 우연성(Randomness)을 최대화시키는 개념으로 활용되고 있음을 밝히고자 한다.

Ⅲ. 불균등 확률 추출과 최대 엔트로피 모형

N개의 단위를 가진 모집단에서 n개의 표본을 추출할 때 $\binom{N}{n}$가지의 가능한 갯수의 표본들이 추출될 확률이 $1/\binom{N}{n}$로 모두가 동일하지 않게 추출하는 표본 추출방법을 불균등 확률 추출이라 부르고 있다. 한 표본이 i번째 모집단 단위를 포함할 주변 확률 π_i는

$$0 < \pi_i < 1 \qquad (i = 1, 2, \cdots, N), \quad \sum_{i=1}^{N} \pi_i = n \tag{4}$$

이 되는 특별히 가중화된 표본 추출법을 정의하자.

먼저 임의 표본으로 $X = (X_1, \cdots, X_N)$로 표기하고 확률변수 X_i는 i번째 단위가 표본에 속하면 1로 속하지 않을 때는 0의 값을 갖는다고 하면 표본의 표본공간

$$D^n = \{x = (x_1, x_2, \cdots, x_N) : x_i = 0 \text{ or } 1,$$
$$x_1 + \cdots + x_N = n\}, \quad i = 1, 2, \cdots, N$$

라고 하고, 임의 벡터 X는 D^n안의 값을 취한다. 임의 벡터 $x \in D^n$에 대해서 이 추출방법에 있어서 $p(x) > 0$이고 $\sum p(x) = 1$를 만족하는 확률밀도함수를 $p(x)$라고 하자.

이제 표본이 i번째 단위를 포함하는 포함확률

$$\pi_i = E(X_i) = \sum_{x \in D^n} x_i p(x) \tag{5}$$

이며 π_i는 (4)식을 만족한다.

여기서 제안되는 이 특수한 일련의 표본 추출방법들은 다음 세가지중의 한가지 형태로

정의될 수 있으며 그들은 결국 동일한 것임이 밝혀진다.

방법 1. 모든 $i = 1, 2, \cdots, N$에 대해

$w_i > 0$되는 임의의 가중치 벡터$w = (w_1, w_2, \cdots, w_N)$를 선택하고

$$p(x) \propto \prod_{i=1}^{N} w_i^{x_i} \tag{6}$$

로 정의한다.

이 때 양의 상수의 승수(multiplier)에 의해 w_i를 재측정하는 것은 동일한 $p(x)$를 결정함은 물론 서로 다른 w에 대응하는 법(modulo)으로 재측도화된 $p(x)$가 서로 다르게 됨은 명백하다. 그러나 (5)식에 의해서 결정되어진 포함확률이 법(modulo)으로 재측도화된 가중치 w_i들과 1대1의 대응관계에 있을 것인지는 분명치 않다.

여기서 $w_i = e^{\theta_i}$으로 표기하면 (6)식은

$$p(x) \propto \exp(\sum_{i=1}^{N} \theta_i x_i) \quad , \theta = (\theta_1, \cdots, \theta_N)$$

로 쓸 수 있으며 이는 지수족으로 연결되어짐이 명백하다.

방법 2. (4)식을 만족시키는 임의의 포함확률 벡터 $\pi = (\pi_1, \cdots, \pi_N)$를 선택하자. 이제 불균등확률을 가지는 크기 N인 모집단에서 추출하는 n개의 표본들로 이루어진 표본분포가 최대의 우연성(Randomness)을 가지게 하는 방안으로서 엔트로피(entropy) $-\sum p(x) \log p(x)$를 최대화하기 위하여 (5)식을 충족시키는 $p(x)$를 선택한다.

만일 가중화된 벡터 w(또는 그에 대응하는 지수 벡터 θ)가 방법 1에 의해 정의된 $p(x)$가 방법 2에서 주어진 π와 짝지어지는 것과 같이 결정될 수 있다면 방법 1의 선택은

방법 2에서 제안된 유일한 최대 엔트로피(entropy) 방법임을 의미한다.

세번째 방법은 방법 1을 약간 변형한 형태이다.

방법 3. 임의의 확률벡터 $p = (p_1, p_2, \cdots, p_N)$ 단 $i = 1, 2, \cdots, N$에 대해 $0 < p_i < 1$인 p를 선택한다. 그리고 $Z = (Z_1, Z_2, \cdots, Z_N)$은 확률 $(p_1, p_2, \cdots, p_N)$을 가진 독립인 Bernoulli시행이라 하고 X의 표본분포를 $\sum Z_i = n$이 주어졌을 때의 Z의 조건부 분포가 되도록 정의한다. 그러면 방법 3이 방법 1과 동일한 표본추출 방법이 되기 위한 필요충분 조건은 w_i가 $p_i/(1-p_i)$에 비례한다는 것은 명백하다.

상기한 세가지 방법에 대한 기본 모형은

$$p(x) = \prod_{i=1}^{N} w_i^{x_i} / \sum_{y \in D^n} \prod_{i=1}^{N} w_i^{y_i} \propto \exp(\sum_{i=1}^{N} \theta_i x_i) \quad (x \in D^n) \tag{7}$$

가 된다. 단 w 또는 그와 동등한 θ는 (5)식에서 π에 의해 결정된다. 본 논문은 최대 엔트로피(entropy) 모형으로 (7)을 채택할 것이다.

Ⅳ. 가중치와 포함확률과의 관계

가중치 w와 π사이의 관계는 지수족에서의 자연대수와 평균값의 매개화 사이의 특수한 경우이다. 다음의 정리는 Brown[1]의 정리 3.6(p74)을 사용하면 증명될 수 있다.

정리 2 (4) 식을 만족하는 임의의 벡터 π에 대하여, (5)식을 충족시키는 최대 엔트로피(entropy) 모형에 대한 벡터w가 존재하며 그 w는 재측도화함에 있어 유일하다.

π로부터 w를 계산하기 위하여 (5)식을 다음의 일련의 방정식(8)의 형으로 재정리하고 (10)식에서 처럼 이들을 반복적으로 풀이한다. 본 논문에서 다음의 기호를 사용한다.

$S = \{1, 2, \cdots, N\}$, 대문자 A, B, C 등은 S의 부분집합 $A^c = S \backslash A$:S에서의 A의 여집합, $|A|$:A의 원소의 수이고

임의의 공집합이 아닌 집합 $C \subset S$와 $1 \leq k \leq |C|$에서

$$R(k, C) = \sum_{B \subset C, |B| = k} \left(\prod_{i \in B} w_i \right) \qquad (*)$$

로 두면 임의의 $k > |C|$에 대해는 $R(0, C) = 1$, $R(k, C) = 0$이 된다.

기호$(*)$에 따라서 다음과 같은 성질이 성립한다.

성질 1 임의의 집합 $C \subset S$가 $1 \leq k \leq |C|$일 때

$a) \sum_{j \in C} w_j R(k-1, C \backslash \{j\}) = kR(k, C)$

$b) \sum_{j \in C} R(k, C \backslash \{j\}) = (|C| - k) R(k, C)$

$c) \sum_{i=0}^{k} R(i, C) R(k-i, C^c) = R(k, C)$

성질 2 임의의 집합 $C \subset S$, $2 \leq k \leq |C|$에 대하여

$$R(k-2, C) R(k, C) < \{R(k-1, C)\}^2.$$

성질 3 X는 R^m의 볼록부분집합이고, $g = (g_1, \cdots, g_m)$는 $g : R^m \to R^m$.

두벡터$x, y \in X$는 주어지고 $l(x, y) = \{z | z = \lambda x + (1 - \lambda) y, 0 \leq \lambda \leq 1\}$.

만일 각각의 g_i는

$a)$ $l(x, y)$의 모든점에서 연속이다.

$b)$ $l(x, y)$의 모든 내점에서 미분가능하다. 라고 하면

$$\|g(x) - g(y)\| \leq \left\{ \sup_{z \in l(x, y)} \|g'(z)\| \right\} \|x - y\|,$$

여기서 $\|\cdot\|$는 l_∞노름(norm)을 나타낸다. 즉 벡터 x에 대하여, $\|x\| = max|x_i|$, 이고 행렬$A = (a_{ij})_{m\times n}$에 대하여는 $\|A\| = max(|a_{i1}| + \cdots + |a_{im}|)$.

이 기호를 사용하면 (5)식은 다음과 같이 표시될 수 있다.

$$\pi_i = \frac{w_i R(n-1, \{i\}^c)}{R(n,S)} \qquad (i = 1, 2, \cdots, N) \tag{8}$$

성질 1의 a)에 의해서 (8)식의 우변은 $\sum \pi_i = \sum_i \frac{w_i R(n-1,\{i\}^c)}{R(n,S)} = n$이 되고 이는 (4)식의 결과와 같다. 고정된 n에 대하여 (8)식의 N개의 관계중에는 $N-1$개의 선형 독립관계가 존재한다. 이제 일반성의 상실없이 우리는 $\pi_1 \le \pi_2 \le \cdots \le \pi_N$이라고 가정할 수 있으며 $w_N = \pi_N$이라 두자. 양변의 N개의 방정식으로 (8)식의 처음$N-1$개 방정식을 각각 나누고 각 방정식의 항을 재정리하면 (8)식에서 $\pi_N = \frac{w_N R(n-1,\{N\}^c)}{R(n,S)}$ 즉 $R(n,S) = R(n-1,\{N\}^c)$. 따라서

$$w_i = \frac{\pi_i R(n-1,\{N\}^c)}{R(n-1,\{i\}^c)} \qquad (i = 1, 2, \cdots, N-1, w_N = \pi_N) \tag{9}$$

(9)식은 두개의 미지수를 갖는 형태로 그 해법이 불가능한 것처럼 보이나 수치해석에서 자주사용되는 고정점 문제 해법과 같이 반복 절차를 사용함으로 해결될 수 있다. 다음 형태가 (9)의 방정식의 해법을 제시한다.

$$w_i^{(t+1)} = \left.\frac{\pi_i R(n-1,\{N\}^c)}{R(n-1,\{i\}^c)}\right|_{w=w^{(t)}} \quad (i = 1, \cdots, N-1), \quad w_N^{(t+1)} = w_N^{(t)} = \pi_N,$$

단, $w^{(t)} = (w_1^{(t)}, \cdots, w_N^{(t)})$ (10)

정리 3 $W = \{w : 0 < w_i \leq \pi_i, i = 1, \cdots, N-1; w_N = \pi_N\}$ 으로 정의하자.

a) (9) 식의 방정식들의 집합은 W안에서 유일한 해 w^*를 갖는다.

b) $w^{(0)} = \pi$로 시작하면 (10)식의 벡터 $w^{(t)}(t = 1, 2, \cdots)$들의 수열은 π_N에 의해 유계인 비를 가지며 단조적이고 기하적으로 w^*에 수렴한다.

보조정리 1. 임의의 $i, j \in S$에 대해서 다음의 성질이 성립한다.

$a) \pi_i = \pi_j \iff w_i = w_j$

$b) \pi_i > \pi_j \iff w_i > w_j$

$c) \pi_i > \pi_j \iff w_i/w_j > \pi_i/\pi_j$

$d)$모든 π_i와 어떤$c_1, c_2 \in (0,1)$에 대하여

$c_1 \leq \pi_n \leq c_2$ 이면 $N/n \to \infty$일 때 $w_i/w_j \to \pi_i/\pi_j$

(증명)

a) 정리 2에 의하여 $\tilde{w}_N = \pi_N$ 최대 엔트로피 모형에 대한

유일한$\tilde{w} = (\tilde{w}_1, \cdots, \tilde{w}_N)$가 존재한다. 보조정리 1에 의하여 $\tilde{w}_i/\tilde{w}_N \leq \pi_i/\pi_N$.

따라서 $\tilde{w}_i \leq \pi_i$ 이고 (9)은 W 안에서 바로 하나의 해를 갖는다.

b) $i \in S$에 대하여 $x_i = \log w_i$라하면

$g_i(x) = \log \pi_i + \log R(n-1, \{N\}^c) - \log R(n-1, \{i\}^c) \qquad i = 1, \cdots, N-1$

그리고 $g_N(x) = \log \pi_N$. 편의상 (9)의 로그변환을 하면 $x = g(x)$로 쓸 수 있다.

여기서 $x = (x_1, \cdots, x_N)$ 이고 $g(x) = (g_1(x), \cdots, g_N(x))$.

$X = \{x : \tilde{x}_i \leq x_i \leq \log \pi_i, i \in S\}$, $i \in S$에 대하여 $\tilde{x}_i = \log \tilde{w}_i$로 정의한다.

따라서 X와 g는 성질 3을 모두 만족하고 $g(X) \subseteq X$. 그리고 $i = N, j = N$일 때 $\frac{\partial g_i(x)}{\partial x_j} = 0$ 이다. 그러므로

$$\frac{\partial g_i(x)}{\partial x_j} = \begin{cases} \dfrac{w_i R(n-2, \{N,i\}^c)}{R(n-1, \{N\}^c)} & i = j \neq N; \\ \dfrac{w_j R(n-2, \{N,j\}^c)}{R(n-1, \{N\}^c)} - \dfrac{w_j R(n-2, \{i,j\}^c)}{R(n-1, \{i\}^c)} & i \neq j \neq N \end{cases}$$

성질 1과 2를 이용하여 다음 식을 보일 수 있다.

$$\sum_{j=1}^{N} \left| \frac{\partial g_i(x)}{\partial x_j} \right| = \frac{w_j R(n-2, \{i,N\}^c)}{R(n-1, \{i\}^c)} < \frac{w_N R(n-1, \{N\}^c)}{R(n,S)}.$$

성질 2를 이용하여 $w_N R(n-1, \{N\}^c)/R(n,S)$는 $i = 1, \cdots, N-1$에 대하여 w_i에서 감소하므로

$$\sup_{x \in X} \|g'(x)\| = \sup_{x \in X} \{ \max_{1 \leq i \leq N} \sum_{j=1}^{N} \left| \frac{\partial g_i(x)}{\partial x_j} \right| \} < \sup_{w \in W} \{ \frac{w_N R(n-1, \{N\}^c)}{R(n,S)} \}$$

$$= \left. \frac{w_N R(n-1, \{N\}^c)}{R(n,S)} \right|_{w=\tilde{w}} = \pi_N.$$

마지막 단계는 (8)의 해가 $\tilde{w}$이므로 (8)의 N번째 방정식에 의해서 얻어진다. 성질 3에 의해서

$$\|x^{(t+1)} - x^{(t)}\| = \|g(x^{(t)}) - g(x^{(t-1)})\| < \pi_N \|x^{(t)} - x^{(t-1)}\|. \qquad (**)$$

$(**)$에서 부등식은 고정점 절차가 수렴하기위한 충분조건이다. 따라서 $\tilde{x}$가 X안에서 유일한 고정점이므로 수열$x^{(t)}(t = 1, 2, \cdots)$는 $\tilde{x}$에 수렴한다. 그러므로 수열$w^{(t)}(t = 1, 2, \cdots)$는 단조 기하학적으로 $\tilde{w}$에 수렴하고, 수렴비는 π_N에 의해 유계이다. 미리 $\tilde{w}$를 알지 못하지만 출발점으로 $\pi \in W$를 선택할 수 있다.

V. 표본조사에서의 응용

표본조사에서의 가중화된 표본추출의 전형적인 목표는 N개의 단위를 갖는 유한 모집단의 총합 $Y = \sum y_i$를 추정하는 것이다. Y와 관련되는 추정량은 $X = (X_1, \cdots, X_N)$을 모집단에서 추출한 n개의 임의 표본이라할 때
$\hat{Y} = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} X_i$가 된다. 한편 $\hat{Y}$의 분산

$$V(\hat{Y}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + 2 \sum_{1 \le i \le j \le N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j \qquad (11)$$

로 주어진다. 여기서π_{ij}는 i와 j 두단위가 표본에 있게 될 2계 순위 포함확률을 의미한다. n이 고정될 때 (11)과 대응되는 표현방식으로

$$V(\hat{Y}) = \sum_{1 \le i \le j \le N} (\pi_i \pi_j - \pi_{ij}) (\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2 \qquad (12)$$

n이 고정되고 모든 $i \neq j$에 대해 만일 $\pi_{ij} \neq 0$이면 $V(\hat{Y})$의 추정량

$$V(\hat{Y}) = \sum_{1 \le i \le j \le N} \frac{\pi_i \pi_j}{\pi_{ij}} (\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2 X_i X_j \qquad (13)$$

이 된다.

이 일련의 표본추출에 주어지는 총합 Y의 분산 및 그 추정량들은 최대 엔트로피 모형에 의한 w_i로부터 직접적으로 손쉽게 구할 수 있는 장점이 있다. 한편 최대 엔트로피 모형은 다음과 같은 성질을 갖는다.

성질 1. 임의의 순위에 대한 포함확률은 π_i에 의해서 유일하게 결정되며 다음과 같은 형태로 표현할 수 있다.

$$\pi_{ij} = w_i w_j R(n-2, \{i,j\}^c)/R(n,S), \tag{14}$$

일반적으로 표본내에 $i_1, \cdots, i_k$단위가 포함될 k차 순위$(1 \le k \le n)$포함확률은

$$\pi_{i_1,\cdots,i_k} = (\prod_{i=1}^{k} w_{i_k}) \frac{R(n-k, \{i_1,\cdots,i_k\}^c)}{R(n,S)} \tag{15}$$

이다.

성질 2. 최대 엔트로피 모형에 대해서 $i \neq j$일 때

$$0 < \pi_{ij} < \pi_i \pi_j \tag{16}$$

가 성립한다.

본성질은 비복원 추출이 복원 추출 보다도 더 효율적임을, 다시말하면 보다 더 작은 분산치를 갖는 것임을 의미한다.

성질 3. (14)과 (15)식에 성질 2를 적용하면 모든 가능한 표본에 대하여

$$\pi_i \propto y_i \iff V(\hat{Y}) = 0 \iff v(\hat{Y}) = 0$$

이 성립한다.

성질 4.

$$\sum_{i=1}^{N} \pi_i = n, \quad \sum_{j \neq i}^{N} \pi_{ij} = (n-1)\pi_i, \quad \sum_{i}^{n} \sum_{j>i}^{N} \pi_{ij} = \frac{1}{2} n(n-1)$$

크기가 n인 표본에서 표본단위의 순위를 고려할 때 i단위는 n회의 순위를 취하므로 $\sum_{i=1}^{N}\pi_i = n$이 성립한다. 두번째 관계는 크기 n인 표본에서 i단위와 결합할 수 있는 다른 단위수는 $(n-1)$개 있기때문에 i단위를 고려할 때 성립한다. 세번째 관계는 두번째 관계에서

$$\sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi_{ij} = 2\sum_{i}^{N}\sum_{j>i}^{N}\pi_{ij} = (n-1)\sum_{i}^{\pi}\pi_i = n(n-1)$$

따라서 세번째 관계가 성립한다.

VI. 비복원 확률비례 추출과 최대 엔트로피

N집락 추출에 있어서 각 추출단위의 추출확률이 같지 않은 표본추출을 불균등 확률 추출이라고 하고 이경우 각 단위를 그 크기에 비례하는 확률로 추출하는 방법이 확률비례 추출(probability proportional to size : pps)이다. N개 단위 중에서 n개를 크기 m에 관하여 비복원 확률비례추출(pps wor)을 한다고 가정하자. 제1회 추출에서 U_i단위의 추출확률은 $p_i = \frac{M_i}{M_0}, i = 1, \cdots, N$ 이다. 단 여기서 $M_i = |v_i|$ 즉v_i단위의 크기이고 $M_0 = \sum_{i=1}^{N} M_i$이다. 제2회 추출에서 U_j단위의 추출확률은 제1회 추출에서 U_i가 추출된 조건부확률이므로

$$p_{j|i} = p_j/(1-p_i)$$

이다. 제3회 추출에서 U_k단위의 추출확률은 제1회, 제2회 추출에서 U_i, U_j가 추출된 조건부 확률이므로

$$p_{k|i,j} = p_k/(1-p_i-p_j), \quad i \neq j \neq k$$

와 같다.

크기 n인 표본을 단순 임의 비복원추출할 경우 U_i단위가 추출되는 확률은 매 회마다 동일한 값 $1/N$이다. 확률비례 비복원 추출의 경우는 매회마다 추출확률이 같지 않다. Yates와 Grundy [6]의 예를 들면

$N = 4, n = 2, p_i = M_i/M_0$는 다음과 같다고 하자.

단위U_i	1	2	3	4
상대크기 p_i	0.1	0.2	0.3	0.4

U_i이 제1회에 뽑히고 U_2가 제2회에 뽑히는 확률은

$$P_r(1,2) = P(1)P(2|1) = p_1(\frac{p_2}{p_2+p_3+p_4})$$
$$= (0.1)(\frac{0.2}{0.2+0.3+0.4}) = 0.022$$

U_1이 제1회에 뽑히는 확률은 표본 (U_1, U_2), (U_1, U_3), (U_1, U_4)가 뽑히는 확률의 합이므로 표2에서 0.10이 된다. 그리고 $n = 2$일 때 U_1이 제2회에 뽑히는 확률은 표본$(U_2, U_1), (U_3, U_1), (U_4, U_1)$이 뽑히는 확률이므로 0.135가 된다. 따라서 U_1이 표본에 포함되는 확률 π_1은 U_1이 제1회 또는 제2회에 추출되는 확률의 합이되므로 $\pi_1 = 0.100 + 0.135 = 0.235$이다.

표 1

단위 U_i	1	2	3	4	
π_i	0.235	0.441	0.609	0.715	$\sum^n \pi_i = 2$
$\bar{p}_i$	0.1175	0.2205	0.3045	0.3575	$\sum^N \bar{p}_i = 1$

마찬가지로 표1에서 π_2, π_3, π_4도 알 수 있다.여기서 $\bar{p}_i$는 $\bar{p}_i = \pi_i/n$ 라고 정의한다.

확률 π_i는 U_i단위가 표본에 포함되는 확률이므로 포함확률(inclusion probability)이며, U_i의 추출확률p_i를 최초확률(initial probability)라한다.

이제 π_i는 U_i단위가 표본에 포함되는 확률 π_{ij}는 U_i단위와 U_j단위가 다같이 표본에 포함되는 확률이라고 하자.

표 2 비복원추출확률 ($N=4, n=2$)

표본(S)	U_iU_j	pps wor	1	2	3	4	(단위)
1	1,2	(.1)(.2|.9)=.022	.022	.022			
2	1,3	(.1)(.3|.9)=.034	.034		.034		
3	1,4	(.1)(.4|.9)=.044	.044			.044	
4	2,1	(.2)(.1|.8)=.025	.025	.025			
5	2,3	(.2)(.3|.8)=.075		.075	.075		
6	2,4	(.2)(.4|.8)=.100		.100		.100	
7	3,1	(.3)(.1|.7)=.043	.043		.043		
8	3,2	(.3)(.2|.7)=.086		.086	.086		
9	3,4	(.3)(.4|.7)=.171			.171	.171	
10	4,1	(.4)(.1|.6)=.067	.067			.067	
11	4,2	(.4)(.2|.6)=.133		.133		.133	
12	4,3	(.4)(.3|.6)=.200			.200	.200	
		1.000					
		π_i	.235	.441	.609	.715	$\sum \pi_i = 2$
		p_i	.1175	.2205	.3045	.3575	$\sum p_i = 1$

표 2에서 $n=2$인 경우 성질4의 첫번째, 두번째 관계를 확인할 수 있다.

이제 π_i가 표2에서 처럼 0.235, 0.441, 0.609, 0.715로 각각 주어졌다고 가정

하자. 식(10)과 정리3에서 주어진 π로부터 w를 계산하기 위하여 $|\frac{w^{(t)}}{w^{(t-1)}} - 1| < 0.001$의 제한을 사용하면 $t = 9$단계에서의 w_i의 값은 각각 0.14727, 0.30939, 0.49508, 0.715가 됨을 Computer 프로그램에 의한 부록2로 확인할 수 있다. 이 결과를 다시 (14)식에 의해 2계순위 포함확률 π_{ij}를 구하면 표3과 같다.

표 3 2계순위 포함확률π_{ij}

i	j=2	j=3	j=4
1	0.04785	0.07658	0.11058
2		0.16087	0.23233
3			0.37178

표 3에서 $\sum_{i \neq j}^{N} \pi_{ij} = (n-1)\pi_i$가 성립되는 일례로서 $i = 1$인 경우 $\sum \pi_{ij} = \pi_{12} + \pi_{13} + \pi_{14} = 0.23491 = \pi_1$ 즉 최대 엔트로피 모형에 의하여 미확인되는 표본분포 확률밀도함수$p(x)$의 결과와 일치하고 있음을 확인할 수 있다.

부록1

(비복원 확률비례 추출에 대한 프로그램)

```
proc iml;
reset print;
p= { 0.235, 0.441, 0.609, 0.175 } ;
w=p;
i= { 0 1 1 1, 1 0 1 1, 1 1 0 1, 1 1 1 0 } ;
k=J(4,1,1);

q=w[1:3, ];
s=q[+, ];
d=w*diag(s);
n=w`*i;
w=d/n`;
oldw=w;
count=1;
print count, w;

do until(abs(m)<0.001);
oldw=w;
q=w[1:3, ];
s=q[+, ];
d=p*diag(s);
n=w`*i;
w=d/n`;
a=w/oldw-k;
a=a[1:3, ];
m=a[<>, ];
count=count+1;
print count, w;
end;
quit;
```

부록2

(부록1에 대한 program 실행 결과)
count2에서 count8까지는 지면관계로 생략하였음.

```
NOTE: AUTOEXEC processing completed.;

    1    proc iml;
IML Ready
    2    reset print;
    3    p= { 0.235,0.441,0.609,0.715 } ;

p          4 rows 1 columns
     0.235
     0.441
     0.609
     0.715
    4    w=p;

w          4 rows 1 columns
     0.235
     0.441
     0.609
     0.715
    5    i= { 0 1 1 1,1 0 1 1 ,1 1 0 1, 1 1 1 0 } ;
I            4 rows 4 columns
             0          1          1          1
             1          0          1          1
             1          1          0          1
             1          1          1          0
    6    k=j(4,1,1);
K            4 rows 1 columns
             1
             1
             1
             1
    7    q=w[1:3, ];
Q              3 rows 1 columns
      0.235
```

```
      0.441
      0.609
    8     s=q[+, ]
S         1 rows 1 columns
      1.285
    9     d=w*diag(s);
D         4 rows 1 columns
  0.301975
  0.566685
  0.782565
  0.918775
   10     n=w'*i;
N         1 rows 4 columns
           1.765      1.559      1.391      1.285
   11     w=d/n';
W         4 rows 1 columns
0.1710907
0.3634926
0.5625917
    0.715
   12     oldw=w;
OLDW      4 rows 1 columns
0.1710907
0.3634926
0.5625917
    0.715
   13     count=1;
COUNT     1 rows 1 columns
       1
14    print count,w;
                                  COUNT
                                      1
                                      W
                              0.1710907
                              0.3634926
                              0.5625917
```

```
                                0.715
     15    do until(abs(m)<0.001);
     16    oldw=w;
     17    q=w[1:3, ];
     18    s=q[+, ];
     19    d=p*diag(s);
     20    n=w'*i;
     21    w=d/n';
     22    a=w/oldw-k;
     23    a=a[1:3, ];
     24    m=a[<>, ]
     25    count=count+1;
     26    print count,w;
     27    end;
OLDW     4 rows 1 columns
0.1710907
0.3634926
0.5625917
    0.715
Q        3 rows 1 columns
0.1710907
0.3634926
0.5625917
S        1 rows 1 columns
1.0971749
D        4 rows 1 columns
0.2578361
0.4838541
0.6681795
0.7844801
N        1 rows 4 columns
1.6410843 1.4486823 1.2495833 1.0971749
W        4 rows 1 columns
0.1571133
 0.333996
0.5347219
```

```
    0.715
A         4 rows 1 columns
-0.081696
-0.081148
-0.049538
         0
A         3 rows 1 columns
-0.081696
-0.081148
-0.049538
M         1 rows 1 columns
-0.049538
COUNT     1 rows 1 columns
 2                          COUNT
                               2
                               W
                       0.1571133
                        0.333996
                       0.5347219
                           0.715

0.1474042
0.3097003
0.4957538
    0.715
A         4 rows 1 columns
-0.001531
-0.001688
-0.002312
        0
A         3 rows 1 columns
-0.001531
-0.001688
-0.002312
M         1 rows 1 columns
-0.001531
```

```
COUNT     1 rows 1 columns
   8
```

```
                                COUNT
                                    8
                                    W
                              0.1474042
                              0.3097003
                              0.4957538
                                  0.715
```

```
OLDW      4 rows 1 columns
0.1474042
0.3097003
0.4957538
    0.715
```

```
Q         3 rows 1 columns
0.1474042
0.3097003
0.4957538
```

```
S         1 rows 1 columns
0.9528583
```

```
D         4 rows 1 columns
0.2239217
0.4202105
0.5802907
0.6812937
```



```
N         1 rows 4 columns
1.5204541 1.358158 1.1721045 0.9528583
```

```
W         4 rows 1 columns
0.1472729
0.3093974
0.4950844
    0.715
```

```
A         4 rows 1 columns
-0.000891
-0.000978
 -0.00135
        0
```

```
A          3 rows 1 columns
-0.000891
-0.000978
 -0.00135
M          1 rows 1 columns
-0.000891
COUNT      1 rows 1 columns
 9                          COUNT
                               9
                               W
                       0.1472729
                       0.3093974
                       0.4950844
                           0.715
    28     quit;
Exiting IML.
NOTE: The PROCEDURE IML used 2.00 seconds.
```

참고문헌

1. Brown, L. D.(1986).*Fundermentals of Statistical Exponential Families (with Applications in Statistical Decision Theory). Hayward, CA:Institute of Mathematical Statistics.*

2. Hanif, M. & Brewer, K. R. W.(1980). Sampling with unequal probabilities without replacement: A review. Int. Statist. Rev. 48, 317-35.

3. Hansen,M.H.&Hurwitz,W.N.(1943). On the theory of sampling from a finite population.*Ann. Math. Statist. 33, 350-74.*

4. Horvitz,D.G. & Thomson,D.J.(1952) A generation of sampling without replacement from a finite universe. *J. Am. Statist. Assoc. 47,663-85.*

5. Stern,H.& Cover,T.M.(1989). Maximum entropy and the lottery.*J. Am. Statist. Assoc. 84,980-85.*

6. Yates,F & Grundy,P.M.(1953). Selection without replacement from within strata with probability proportional to size.*J.R.Statist. Soc. B 15,253-61*

7. Chen,X. and Dempster,A.P and LIu.J.S. Weighted finite population sampling to maximaze entropy.(1994) *Biometrica 81. 3. 457-69*

8. 박홍래 통계조사론(1989) 영지문화사

9. 한영열 정보이론(1985) 민음사

< Abstract >

A STUDY ON THE PROBABILITY PROPORTIONAL TO SIZE WITHOUT REPLACEMENT AND MAXIMAL ENTROPY

Kim, Yeong-Hee

Mathematics Education Major

Graduate School of Education, Cheju National University

Cheju, Korea

Supervised by professor Kim, Ik-Chan

In this paper, entropy is defined by the attempt to represent the conception of surprise as quantity and the conception of entropy is the model of maximal entropy in sample distribution which consists of the sampe of n distinct units from a population of N units with unequal probabilities is adopted, and an equation between inclusion probabilities and weights is set up, and the solution is presented by the fixed point iteration method of numerical analysis.

As a result, the solution is applied in the probability proportional to size without replacement recognized as the prominent case among unequal probabilities.

A thesis submitted to the Committee of the Graduate School of Educaion. Cheju National University in partial fulfillment of the requirements for the degree of Master of Education in August, 1996.

감사의 글

이 논문이 나오기까지 바쁘신 중에도 세심한 배려와 관심을 갖고 지도해 주신 김익찬 교수님과, 학문적으로나 살아가는 길에 대하여 아낌없이 많은 가르침과 격려를 주신 수학교육과, 수학과 모든 교수님들께 감사를 드립니다.

서로 의지하고 함께 강의를 받으며 고생했던 김정두 선생님, 양창길 선생님, 고영종 선생님, 김순찬 선생님께도 고마운 뜻을 전합니다.

그리고 학교 일과진행의 어려움 속에서도 과정을 마칠 수 있도록 배려해 주신 서귀포여자고등학교 교장선생님을 비롯한 모든 선생님들과, 주위에서 기도와 격려를 주신 모든 분들께도 감사를 드립니다.

끝으로, 따뜻한 마음으로 보살펴 주신 부모님, 많은 어려움도 불평없이 이겨내며 내조해 준 사랑하는 아내, 건강하고 착하게 자라는 온유, 대열이와 함께 이 조그마한 성취의 기쁨을 나누고자 합니다.

1996년 8월

김 영 희 드림