한국통계학회논문집
2009, 16권, 4호, 595-603

층화확률화 응답기법에 대한 동적 최적배분

손창균[1,a], 홍기학[b], 이기성[c]

[a]한국 보건사회연구원, [b]동신대학교 컴퓨터학과, [c]우석대학교 아동복지학과

요약

통상적으로 표준적인 최적배분은 층별 조사비용을 고려하여 표본을 배분한다. 만일 조사단위당 비용이 서로 다를 경우 보다 현실적인 배분방법을 고려할 필요가 있다. 즉, 개별 조사단위의 특성에 따라 이익비용비를 최대로 하는 단위를 먼저 표본으로 고려하는 동적배분을 고려하였다. 이러한 관점에서 층별 표본수를 배분하고, 배분된 표본규모에 따라 임의로 표본을 선정하는 방식인 표준적인 최적배분과는 차이가 있다. 이 논문은 약물오용, 낙태, 알콜중독 등과 같은 민감한 특성을 조사하는 층화확률화 응답기법에 대해 각 층별로 표본을 배분할 경우 최적 동적배분을 고려하여 보다 현실적인 문제를 해결하고자 하였으며, 수치적 예제를 통해 동적배분 방법의 효과성을 증명하였다.

주요용어: 최적 동적배분, 확률화응답기법, 이익비용비.

1. 서론

사회 여러 분야의 표본조사에서 발생하는 오차에는 표본오차와 비표본오차가 있으며, 최근에 연구의 관심은 비표본오차를 줄이는 데 있다. 이러한 비표본오차는 응답자들이 민감하거나 개인적인 이해와 관계되는 질문을 받았을 경우 더욱 증가하게 된다. 예를 들어 음주운전, 낙태경험, 환각제사용, 동성연애 및 탈세여부 등과 같은 사회적으로나 개인적으로 매우 민감한 문제에 관한 조사에서 기존의 직접질문방식을 그대로 사용할 경우 응답자들이 응답을 회피하거나 거짓으로 응답하는 경향이 뚜렷이 나타나게 된다. 이는 응답자들이 민감한 질문에 응답함으로써 불이익을 받거나 사생활이 보장되지 않는다고 생각하기 때문이다. 이와 같은 문제점을 해결하고 사생활을 보장해주기 위한 대표적인 조사방법으로 간접질문 조사방법인 확률화응답기법(RRT)이 있다.

1965년 Warner (1965)에 의해 처음 제시된 확률화응답기법은 응답자의 신분이나 비밀을 노출시키지 않고 민감한 질문에 대한 정보를 이끌어 내기 위하여 응답자들에게 확률장치를 통한 간접응답을 하게 함으로써 그들의 익명성을 보장해 주면서 조사자가 얻고자하는 민감한 정보를 최대한 얻을 수 있도록 한 방법이다. 그 후 많은 학자들에 의해 확률화응답기법에 대한 연구가 활발하게 이루어져 왔다. 최근 들어 확률화응답기법들의 실용성을 높이기 위하여 응답자들을 추출하는 데 있어서 단순임의추출법 이외의 다양한 표본추출법들이 적용되고 있다. Hong 등 (1994)은 모집단이 여러 개의 층으로 구성되어 있을 때 각 층별로 민감한 정보를 얻을 수 있는 층화 확률화응답기법을 제안하였다. 또한 Kim과 Warde (2004)는 각 층에 대하여 서로 다른 확률장치를 사용하는 층화 확률화응답기법을 제안하였으며, 이때 각 층에 다양한 표본배분 방법을 적용하였다. 이들의 연구에서는 조사비용이 조사항목 또는 조사단위에 대해 변하지 않는 비용함수를 사용하였다. 그러나 조사 단위별 조사비용은 서로 다르며, 특히 연령별 또는 소득별로 층화된 모집단에 대해 확률화응답기법을 적용할 경우 각 층별로 응답을 얻기 위한 비용이 서로 다르게 나타날 수 있을 것이다.

[1] 교신저자: (122-705) 서울 은평구 진흥로 268번지 한국보건사회연구원 부연구위원. E-mail: chkson@kihasa.re.kr

이러한 관점에서 본 연구에서는 조사단위별로 다르게 소요되는 조사비용을 고려한 동적 표본배분(dynamic sample allocation)을 확률화응답기법에 적용하여 기존의 정적 표본배분방법에 의한 확률화응답기법의 추정량과 비교하여 조사비용대비 효율성에 대해 살펴보고자 한다.

2. 층화 확률화응답기법

유한모집단 $U = \{1, 2, \ldots, N\}$은 크기가 N_h인 L개의 층인 $U = \{U_1, \ldots, U_h, \ldots, U_L\}$로 구성되어 있다고 가정하자. 또한 $h = 1, \ldots, L$에 대해 단순임의복원(SRSWR)으로 크기가 n_h인 표본을 각 층에서 독립적으로 추출한다고 하자. 이때, 각 층의 크기 N_h는 모두 기지(known)라고 가정한다. Kim과 Warde (2004)기법에서 각 표본 층에 있는 응답자에게는 선택확률이 P_h인 민감한 질문이 적힌 카드 A와 선택확률 $1 - P_h$인 민감하지 않은 질문이 적힌 카드 A^c로 구성된 서로 다른 확률장치 R_h를 제공하게 되며, 응답자는 반드시 선택된 카드에 대해 "예" 또는 "아니오"로 응답하게 된다. 즉, 서로 다른 층에 속한 응답자는 서로 다른 확률을 갖는 확률장치를 사용하게 된다. 반면에 Hong 등 (1994)의 기법은 서로 다른 층에 속한 응답자들에게 동일한 확률장치를 사용하도록 하고 있으며, 이는 $P_h = P$인 경우에 해당한다. 이때 n_h는 h층의 표본단위들의 수이며, $n = \Sigma_{h=1}^{L} n_h$는 전체 표본의 크기가 된다. 이와 같은 가정하에서 응답자가 진실하게 응답한다고 하면, $h = 1, \ldots, L$층에 대해 "예"라고 응답할 비율은 다음과 같다.

$$Z_h = P_h\pi_h + (1 - P_h)(1 - \pi_h), \tag{2.1}$$

여기서 Z_h는 h층에 대해 "예"라고 응답할 비율이며, π_h는 h층에서 민감한 속성을 가진 응답자의 모비율이고, $P_h (0 < P_h < 1, P_h \neq 0.5)$는 민감한 질문이 적힌 카드 A를 선택할 확률이다.
그러면 h층에 대해, 모비율 π_h의 최우추정량(MLE)는 다음과 같다.

$$\hat{\pi}_h = \frac{\hat{Z}_h - (1 - P_h)}{2P_h - 1}, \tag{2.2}$$

여기서 $\hat{Z}_h$는 h층에 대해 "예"라고 응답한 표본비율이다.
각각의 $\hat{Z}_h$는 독립적으로 $B(n_h, Z_h)$인 이항분포를 따르므로, π의 최우추정량은 다음과 같다.

$$\hat{\pi}_{st} = \sum_{h=1}^{L} W_h\hat{\pi}_h = \sum_{h=1}^{L} W_h \left\{ \frac{\hat{Z}_h - (1 - P_h)}{2P_h - 1} \right\} \tag{2.3}$$

그리고 $\hat{\pi}_{st}$의 분산은 다음과 같다.

$$V(\hat{\pi}_{st}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left\{ \pi_h(1 - \pi_h) + \frac{P_h(1 - P_h)}{(2P_h - 1)^2} \right\}. \tag{2.4}$$

만일 표본이 단순임의비복원(SRSWOR)으로 추출된다면, $\hat{\pi}_{st}$의 분산은 다음과 같다.

$$V(\hat{\pi}_{st}) = \sum_{h=1}^{L} W_h^2 \left\{ \frac{\pi_h(1 - \pi_h)}{n_h}(1 - f_h) + \frac{P_h(1 - P_h)}{n_h(2P_h - 1)^2} \right\}, \tag{2.5}$$

여기서 $W_h = N_h/N$는 $h = 1, \ldots, L$층의 가중치이며, $f_h = n_h/N_h$는 층별 추출률이다.

3. 표본배분

이 절에서는 층화 확률화응답기법에서 최적의 표본수를 전통적인 정적 최적배분(static optimal allocation: SOA)의 경우와 동적 최적배분(dynamic optimal allocation: DOA)의 경우로 구분하여 각각 유도하고자 한다. 이를 바탕으로 두 배분방법 간의 효율성을 비교함으로서 조사현장에서 보다 실제적인 문제를 다루고자 한다.

3.1. 정적 최적배분(Static Optimal Allocation)

층화표본추출에서 각 층에 표본을 배분하는 방법 중 최적배분의 경우에는 조사비용을 고려하여 주어진 분산을 최소로 하는 층별 표본수를 결정하거나, 반대로 일정한 정도 하에서 비용을 최소로 하는 방법을 사용하게 된다. 이때, 각 항목별 조사비용은 고정으로 간주하고, 각 층별 조사비용만을 고려하여 분산을 최소로 하는 층별 표본수를 결정하게 되는데, 이러한 방법을 정적 최적배분방법이라 한다. 조사비용과 관련하여 Cochran (1977)은 비용함수에 여행비용(travel cost)을 고려하여 표본을 배분하는 방법을 다루기도 하였다.

층화추출설계를 적용한 직접조사방법 뿐만 아니라, 확률화응답기법과 같은 간접조사의 경우에도 조사비용을 고려하여 각 층별 표본수를 배분해야 하며, 이를 위해 비용함수를 다음과 같이 정의할 수 있다.

$$C = c_0 + \sum_{h=1}^{L} c_h n_h, \quad h = 1, \ldots, L, \tag{3.1}$$

여기서 C는 총 조사비용이고, c_0는 고정 비용, c_h는 각 층별 조사비용이다.

만일 조사단위당 조사비용과 더불어 여행비용 t_h가 현저히 많이 소요된다면 비용함수는 다음과 같다.

$$C = c_0 + \sum_{h=1}^{L} t_h \sqrt{n_h}, \quad h = 1, \ldots, L. \tag{3.2}$$

Kim과 Warde (2004)은 주어진 비용함수 식 (3.1)하에서 분산 식 (2.4)를 최소로 하는 최적배분된 표본수를 다음과 같이 도출하였다.

$$n_h = n \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}}, \tag{3.3}$$

여기서 $S_h^2 = \pi_h(1-\pi_h) + \{P_h(1-P_h)\}/(2P_h-1)^2$이며, Hong 등 (1994)의 기법에 대해서는 P_h 대신 p를 대입하여 구할 수 있다.

최적의 표본수 n_h를 다시 분산식 (2.4)에 대입하면 다음과 같이 최소분산을 구할 수 있다.

$$V_{\min}(\hat{\pi}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)\left(\sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{c_h}}\right). \tag{3.4}$$

비복원 추출의 경우 최소분산식은 다음과 같이 표현된다.

$$V_{\min}(\hat{\pi}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)\left(\sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{c_h}}\right) - \frac{1}{N}\sum_{h=1}^{L}\{\pi_h(1-\pi_h)\}. \tag{3.5}$$

만일 식 (3.2)와 같이 여행비용 t_h를 고려하면, 고정된 비용 하에서 층화추정량의 분산을 최소화하는 다음과 같은 최적 표본수를 얻을 수 있다.

$$n_h = n\frac{\left(W_h^2 S_h^2/t_h\right)^{\frac{2}{3}}}{\sum_{h=1}^{L}\left(W_h^2 S_h^2/t_h\right)^{\frac{2}{3}}}, \tag{3.6}$$

여기서 $S_h^2 = \pi_h(1-\pi_h) + \{P_h(1-P_h)\}/(2P_h-1)^2$이다.

최적의 표본수 n_h를 분산식 (2.4)에 대입하면 다음과 같이 최소분산을 구할 수 있다.

$$V_{\min}^{*}(\hat{\pi}_{st}) = \frac{1}{n}\sum_{h=1}^{L}(W_h S_h t_h)^{\frac{2}{3}}\sum_{h=1}^{L}\left(W_h^2\frac{S_h^2}{t_h}\right)^{\frac{2}{3}}. \tag{3.7}$$

3.2. 동적 최적배분(Dynamic Optimal Allocation)

3.1절에서는 전통적인 최적배분 방법을 확률화응답기법에 적용한 내용으로 기지의 비용함수 하에서 분산을 최소로 하는 최적 표본수를 결정하였다. 그러나 특히 조사 소요시간과 같은 시간비용은 실제 조사과정에서는 조사를 하기 전까지는 정확히 알 수 없으므로 대개의 경우 이전의 조사에서 소요된 내용을 토대로 단위당 조사비용 등을 산출하여 표본수를 결정하게 된다. 그러나 만일 조사당시 소요되는 시간비용 등을 고려한 표본수 결정방법을 적용할 수 있다면, 보다 현실적인 비용을 반영할 수 있으므로 이러한 관점에서 동적배분은 전통적인 표본배분방식에 비해 효과적인 방법이라 할 수 있다.

이를 위해 $\delta_{j,h}$를 표본의 h층에서 $j-1$번째 단위 대신 j번째 단위를 취함으로서 얻는 분산 감소분이라 하고, $c_{j,h}$를 이때 발생하는 비용의 추가분이라 하자. 그러면, 이익비용비(benefit-cost ratio) $\delta_{j,h}/c_{j,h}$는 모든 h층의 j단위에 대해 엄밀 감소함수(strictly decreasing function)라 가정하자. 먼저 각 층에 하나의 단위를 배분하는 문제를 고려하고, 가장 큰 $\delta_{j,h}/c_{j,h}$값을 가지는 단위를 우선 배분하는데, 이와 같은 방법이 동적으로 최적인 표본배분방법이 된다. 이러한 결과를 층화추출설계에 적용하기 위해 다음과 같이 정의하자.

L개의 층이 $h = 1,\ldots,L$로 구별되고, $j = 1,\ldots,\infty$를 지정된 층에 반복적인 과정으로 배분되는 단위들의 수라 하자. 이 경우 정적인 배분방법에서는 h층에 배분되는 표본수가 $(n_h > 1)$로 정해진다.

층화 확률화응답기법에서 비복원추출의 경우 추정량 $\hat{\pi}_{st}$에 대한 분산식 (2.5)는 다음과 같이 표현할 수 있다.

$$\begin{aligned} V(\hat{\pi}_{st}) &= \frac{1}{N^2}\sum_{h=1}^{L}N_h^2\frac{\pi_h(1-\pi_h)}{n_h} - \frac{1}{N^2}\sum_{h=1}^{L}N_h\pi_h(1-\pi_h) + \frac{K_h}{n_h} \\ &= \frac{1}{N^2}\sum_{h=1}^{L}N_h^2\frac{V_h^2}{n_h} - \frac{1}{N^2}\sum_{h=1}^{L}N_h V_h^2 + \frac{K_h}{n_h}, \end{aligned} \tag{3.8}$$

여기서 $K_h/n_h = P_h(1-P_h)/(2P_h-1)^2/n_h$이며 이는 확률화응답기법을 사용함으로서 발생하는 분산의 증가분으로 사전에 정해진 상수값이고, $V_h^2 = \pi_h(1-\pi_h)$이다. 식 (3.8)은 n_h에 대해 감소함수임을 직관적으로 알 수 있다.

전통적인 표본배분에서 비용함수는 식 (3.1)과 같이 정의되며, 이 비용함수는 경험적으로 예측 또는 추측한 비용을 나타내며, 실제 소요되는 비용은 아니다.

그러면 $d_h = N_h^2 V_h^2/N^2$이라 하면 h층에서 $j-1$번째 단위 대신 j번째 단위를 취함으로서 얻는 분산의 감

소분은 식 (3.8)로 부터 다음과 같다.

$$\delta_{j,h} = \left(\frac{d_h}{j-1} - \frac{d_h}{j}\right) + \left(\frac{K_h}{j-1} - \frac{K_h}{j}\right) = \frac{d_h + K_h}{j(j-1)} \tag{3.9}$$

따라서 $\delta_{j,h}/c_{j,h} = (d_h + K_h)/\{c_h j(j-1)\}$은 각각의 h층에 대해 j에서 엄밀 감소함수이기 때문에 동적인 최적표본계획이 존재하게 된다.
한편 식 (3.2)와 같이 여행비용을 고려한 최적 동적배분인 경우 비용함수는 $c_{j,h} = t_h(\sqrt{j} - \sqrt{j-1})$이므로, 이익비용비는 다음과 같다.

$$\frac{\delta_{j,h}}{c_{j,h}} = \frac{d_h + K_h}{t_h j(j-1)(\sqrt{j} - \sqrt{j-1})} \tag{3.10}$$

결과적으로 식 (3.10)은 다음과 같이 다시 표현된다.

$$\frac{\delta_{j,h}}{c_{j,h}} = \frac{(d_h + K_h)(\sqrt{j} + \sqrt{j-1})}{t_h j(j-1)} = \frac{d_h + K_h}{t_h}\left\{\frac{1}{j^{\frac{1}{2}}(j-1)} + \frac{1}{j(j-1)^{\frac{1}{2}}}\right\} \tag{3.11}$$

식 (3.11)은 j가 증가함에 따라 감소하게 된다. 즉, $\delta_{j,h}$와 $c_{j,h}$를 이용하여 동적 최적표본추출설계를 구할 수 있으며, 고정된 분산 또는 고정된 비용하에서 최적의 표본설계를 고려할 수 있다.

3.3. 정적배분과의 비교

정적배분(SOA)의 경우 3.2절의 식 (3.8)을 최소로 하는 층별 표본수를 구할 수 있다. 이때 다음과 같은 제한조건을 만족한다.

$$n_h \le N_h, \quad h = 1, \ldots, L. \tag{3.12}$$

통상적으로 정적배분(SOA)에서는 식 (3.12)의 제한조건은 무시할 수 있다. 분산을 최소로 하는 문제는 라그란쥐 승수를 이용하여 해를 구할 수 있다. 식 (3.8)을 다음과 같이 다시표현할 수 있다.

$$V(\hat{\pi}_{st}) = \sum_{h=1}^{L} \frac{d_h}{n_h} - T + \frac{K_h}{n_h}, \tag{3.13}$$

여기서 $T = 1/N^2 \Sigma_{h=1}^{L} N_h S_h^2$이며, K_h는 식 (3.8)에서 정의된 것과 같다.
그러면 라그란쥐함수는 비용함수의 식 (3.1)하에서 다음과 같이 정의할 수 있다.

$$L = \sum_{h=1}^{L} \frac{d_h}{n_h} - T + \frac{K_h}{n_h} + \lambda\left(\sum_{h=1}^{L} c_h n_h + c_0 - C\right)$$

n_h에 대해 1차 미분한 후 0으로 놓고 n_h에 대해 정리하면 다음과 같다.

$$n_h = \frac{1}{\sqrt{\lambda}}\sqrt{\frac{d_h + K_h}{c_h}} \tag{3.14}$$

라그란쥐 승수 λ는 식 (3.1)과 (3.14)의 관계로부터 구하여 정리하면 최종적인 층별 표본수는 다음과 같다.

$$n_h = \frac{(C - c_0)\sqrt{(d_h + K_h)/c_h}}{\sum_{h=1}^{L} \sqrt{c_h(d_h + K_h)}}. \tag{3.15}$$

표 1: 정적배분에 의한 표본수 배분($n = 20$)

	N_h	P_h	π_h	c_h	n_h	$V(\hat{\pi}_h)$	r_h^*
층1	100	0.6	0.10	100	8.955149	0.619156	0.000843
층2	200	0.7	0.15	200	6.158294	0.226631	0.000209
층3	300	0.8	0.02	300	4.886557	0.121681	8.5E–05

제한조건 식 (3.12)를 고려하여 표본배분할 경우에는 $n_h = N_h$에 대해 식 (3.15)는 나머지 층에 대해 재계산해야 한다. 이러한 사실은 나머지 층에 대해 식 (3.12)의 제한조건에 위배될 수 있으며, 식 (3.12)를 만족할 때 까지 반복적으로 층별 표본수를 구해야 한다.

동적배분(DOA)과 정적배분(SOA)의 관계를 살펴보기 위해 만일 식 (3.9)로 부터 $r = \delta_{j,h}/c_{j,h}$을 상수라 하면,

$$r = \frac{d_h + K_h}{c_h j(j-1)} = \frac{d_h + K_h}{c_h n_h(n_h - 1)} \approx \frac{d_h + K_h}{c_h n_h^2}$$

이 성립하고, 이를 n_h에 대해 정리하면

$$n_h \approx \frac{1}{\sqrt{r}} \sqrt{\frac{d_h + K_h}{c_h}}$$

로서, 정적배분(SOA)의 식 (3.14)와 점근적으로 같게 되어 정적배분(SOA)과 동일한 효과를 얻을 수 있다. 이는 이익비용비가 조사단위들 간에 동일할 경우 동적배분과 정적배분(SOA)이 같은 결과를 나타냄을 알 수 있다.

한편 n_h의 값이 정수가 아닌 경우에는 조건식 (3.12)를 만족하면서 분산식 (3.13)을 최소로 하는 새로운 n_h를 구하도록 해야 하며, 이 경우에는 n_h에 대한 비선형 방정식을 풀어야 한다. 결과적으로 층의 크기가 매우 작을 때, 표본 층의 크기가 정수이어야 한다는 제한조건에서는 동적배분과 정적배분(SOA)의 표본수에서 차이가 나게 된다.

4. 수치적 예제

확률화응답기법을 적용하기 위해 층별로 설문의 선택확률과 모비율, 조사비용 등에 대한 가정이 필요하다. 또한 동적배분의 경우 한번에 층별로 전체 표본수 n_h를 배분하는 최적 배분과는 달리 조사단위를 이익비용비의 크기 순으로 나열하여 각 층별로 이익비용비가 큰 단위를 우선 배분하여 각 층별로 정해진 표본수 n_h가 될 때까지 배분과정을 반복하게 된다.

본 절에서는 정적배분과 동적배분간의 비교를 위해 먼저 정적배분의 경우 정해지는 표본수를 가정하고, 그에 따라 층별로 각 단위별 이익비용비를 산출한 후 그에 따른 배분 효과를 살펴보기로 한다. 비교를 간단히 하기위해 층의 수는 3개($h = 1, 2, 3$)로 한정하고, 각 층별로 설문의 선택확률 P_h는 각각 0.6, 0.7, 0.8로 변화시키도록 하고, 층별 모비율 π_h는 각각 0.1, 0.15, 0.2로 한다. 모집단의 크기는 $N_1 = 100$, $N_2 = 200$, $N_3 = 300$, $\sum N_h = 600$으로 하였다. 층별 조사비용은 $c_1 = 100$, $c_2 = 200$, $c_3 = 300$으로 정하였다.

표본설계에 의해 전체 표본수 $n = 20$개의 표본을 추출한다고 가정할 때, 먼저 최적 배분에 의한 층별 표본수와 그에 따른 분산은 다음과 같다.

표 1로부터 r_h^*는 $j - 1$과 j 대신 전통적인 표본수 n_h를 대입하여 얻은 이익비용비를 나타내며, 고정표본배분 방법의 경우 모의실험 과정에서 단위별 이익비용비를 직접적으로 계산할 수 없기 때문에 층별 비용을 그대로 이용하여 이익비용비를 산출하였다. 또한 각 층에 표본을 배분할 때 비용이 가장 적

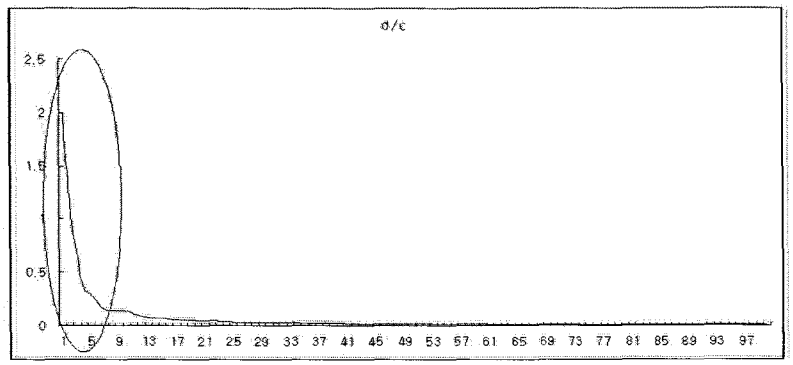

그림 1: *층1의 이익비용비에 따른 조사단위 선정*

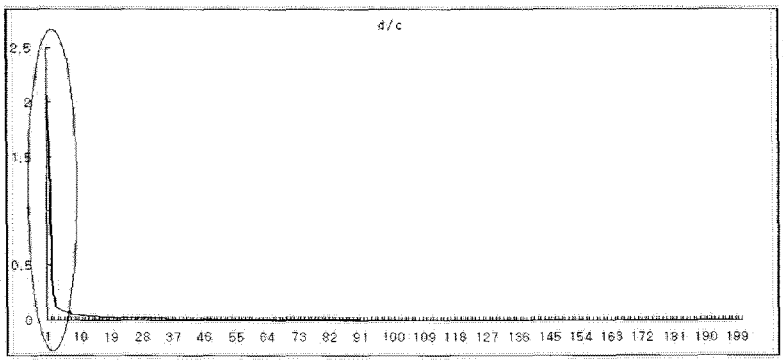

그림 2: *층2의 이익비용비에 따른 조사단위 선정*

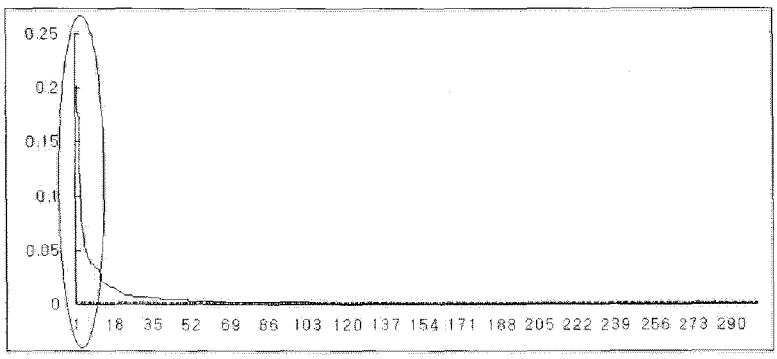

그림 3: *층3의 이익비용비에 따른 조사단위 선정*

게 드는 단위를 직접적으로 고려할 수 없기 때문에 실제 소요되는 비용을 고려할 수 없는 문제가 발생하여 층별 비용을 고려하여 산출하였다.

다음의 그림 1-3은 3개 층에 속한 모집단 단위들의 이익비용비를 산출하여 그림으로 표현한 것으로서 이들 단위 중에서 그래프의 좌측에 속한 타원안의 단위들을 우선 표본으로 선정하여 조사효율을 높일 수 있을 것이다.

표 2: 층별 동적배분을 위한 이익비용비 산출결과 및 조사단위 선정

층1($n_h = 9$)			층2($n_h = 6$)			층3($n_h = 5$)		
x	c_1	δ_1/c_1	x	c_2	δ_2/c_2	x	c_3	δ_3/c_3
39	0.00792	1.99475	457	0.00170	2.05544	904	0.0063	0.20214
169	0.01342	1.17734	363	0.00742	0.47047	982	0.0080	0.15944
151	0.02452	0.64424	360	0.02116	0.16500	858	0.0139	0.09168
117	0.04438	0.35590	332	0.03827	0.09123	644	0.0238	0.05361
46	0.05506	0.28689	218	0.04507	0.07747	685	0.0258	0.04950
161	0.07767	0.20337	470	0.04784	0.07298			
163	0.10685	0.14783						
98	0.11556	0.13669						
61	0.11615	0.13600						

앞의 그림 1-3으로부터 각 층별로 이익비용비의 크기 순서에 따라 표본으로 우선 배정된 단위들의 특성을 살펴보면 다음의 표로 요약된다. 이때 각 층별 표본크기는 전통적 표본배분 방식에 따라 산출된 층별 표본수이다.

표 2로부터 먼저 층1의 경우 $x = 39$인 단위의 이익비용비(δ_1/c_1)가 1.99로서 가장 크기 때문에 이 단위를 먼저 층1의 표본단위로 배정하게 된다. 다음으로 $x = 169$인 단위의 이익비용비가 1.177임으로 두 번째 표본단위로 배정하게 된다. 결과적으로 층1에 배분된 표본수 9개를 이익비용비의 크기 순에 따라 배분하게 된다. 같은 방법으로 층 2에 대해서도 6개의 표본을 이익비용비의 크기순서에 따라 배분할 수 있다. 이러한 방법으로 각 층에 대해 이익비용비가 큰 단위를 우선 배정함으로서 조사의 효율성을 증가시킬 수 있을 것이다.

5. 결론

본 논문에서는 확률화응답기법을 적용한 간접조사 방식에서 층별 표본배분 방법을 전통적인 표본배분 방식이 아닌 단위별 조사비용을 감안한 동적 표본배분방식에 따라 표본을 배분하여 조사하는 방법을 제안하였다. 전통적인 표본 배분방법은 층별 조사비용과 변동을 고려하여 표본을 배분하는 반면에 동적배분방법은 단위당 조사비용과 분산을 고려하여 표본을 배분하는 방식으로서, 예제로부터 각 단위당 이익비용비(benefit-cost ratio)를 고려하여 이들중 가장 큰 값을 가지는 단위를 표본으로 우선 배정하는 방법에 대해 살펴보았고, 이들 단위들을 표본으로 고려할 경우 다른 단위들보다 비용효과가 우월함을 보였다. 특히 확률화 응답기법과 같은 간접질문방식의 경우 직접질문방식에 비해 조사비용이 많이 소요되기 때문에 동적배분방법과 같은 표본배분방식을 고려하는 것도 효과적일 것으로 사료된다.

참고 문헌

Cochran, W. G. (1977). *Sampling Techniques*, 3rd Edition, Wiley, New York.

Hong, K., Yum, J. and Lee, H. (1994). A stratified randomized response technique, *The Korean Journal of Applied Statistics*, **7**, 141–147.

Kim, J. M. and Warde, W. D. (2004). A stratified Warner's randomized response model, *Journal of Statistical Planning and Inference*, **120**, 155–165.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association*, **60**, 63–69.

2009년 3월 접수; 2009년 6월 채택

An Dynamic Optimal Allocation for the Stratified Randomized Response Technique

Chang-Kyoon Son[1,a], Ki-Hak Hong[b], Gi-Sung Lee[c]

[a]Korea Institute for Health and Social Affairs
[b]Department of Computer Science, Dongshin University
[c]Department of Children Welfare, Woosuk University

Abstract

Typically the standard optimal allocation method distributes the sample for each stratum considering survey cost. In case of varying survey cost for each survey unit, we need to consider more practical allocation method. In other words, according to characteristics of an individual unit, we consider the optimal dynamic allocation method which first selects the survey unit having maximum value of benefit cost ratio. In terms of this, the proposed allocation method is different from standard optimal allocation method which allocate samples for each stratum and selects the random sample according to each size of sample. This paper is considered the dynamic optimal allocation method for the stratified randomized response technique which surveys for sensitive characteristic of survey units such as drug abuse, abortion, alcoholic. We prove the practical usefulness of proposed method using the numerical example.

Keywords: Dynamic optimal allocation, randomized response technique, benefit-cost ratio.

[1] Corresponding author: Research Fellow, Korea Institute for Health and Social Affairs, Seoul 122-705, Korea.
E-mail: chkson@kihasa.re.kr