

Decision of Sample Size on Successive Occasions

Hyeonah Park^{a,1} · Seongryong Na^b

^aDepartment of Statistics, Seoul National University

^bDepartment of Information and Statistics, Yonsei University

(Received February 7, 2014; Revised April 9, 2014; Accepted June 16, 2014)

Abstract

If the target error of an estimator at the present time is greater than the coefficient of variation(CV) of the estimator at the previous time, sample size at this point should be decreased. Various papers have researched sample size determination methods using the CV of an estimator at the previous time, variation of population size and target error of the estimator at this time in sampling on successive occasions. We research a new sample size determination method additionally using change of population CV. We compare the proposed method with existing ones in various simulation settings.

Keywords: Successive occasions, population size, coefficient of variation, sample size.

1. 서론

계속조사를 일반적으로 설명하면 한번 표본설계를 하면 일정한 기간동안 그 표본을 유지하는 조사로 표현할 수 있다. 그러므로 일정 기간이 지나면 표본설계를 다시 시행해야 하는 시점이 오는 데 그때 표본의 크기를 결정해야 한다. 표본의 크기 결정에는 조사비용, 추정량의 목표오차, 지난시점에서 추정량의 표본오차, 모집단의 크기 변동 등이 필요하게 된다. 만약 조사비용의 범위가 사전에 정해지는 경우 현 시점의 추정량의 목표오차가 지난 시점의 추정량의 표본오차보다 크다면 표본의 크기는 지난시점과 비교하여 줄어들어야 하며 또한 모집단의 크기가 지난시점보다 줄어든 다면 표본의 크기도 작아져야 한다. 지난 시점의 자료가 존재 할 때 표본의 크기에 대한 연구로는 Park (1989)과 Kim (2012)이 있다. Park (1989)은 지난 시점의 추정량의 변동계수(coefficient of variation; CV)와 현 시점의 추정량의 목표오차의 변동을 이용하여 표본의 크기를 구하는 것을 연구하였고 Kim (2012)은 현 시점의 추정량의 목표오차와 지난 시점의 추정량의 변동계수와 모집단의 크기 변동을 사용한 표본의 크기 구하는 공식을 제시하였다. 또한, Yoo와 Shin (2011)은 패널조사에서 비율과 총계 추정에 표본의 크기가 미치는 영향을 연구하였다.

본 논문에서는 모집단의 변동계수와 모집단의 크기의 변동과 추정량의 목표오차, 그리고 지난 시점의 추정량의 변동계수를 사용한 표본크기 문제를 연구하고자 한다. 기존에는 추정량의 산포변동과 모집단의 크기 변동만 반영이 되었다면 본 논문에서는 모집단의 산포변동도 추가로 반영하는 새로운 공식을 연구

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(NRF-2012R1A1A3003761).

¹Corresponding author: Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea. E-mail: hapk@daum.net

한다. 그리고 새로 제시된 표본크기공식에서 사용되는 모수들의 추정방법도 연구한다. 또한 시점별 정보를 사용한 Park (1989)과 Kim (2012)에 제시된 표본의 크기 공식과 비교 연구한다.

본 논문의 구성은 제 2절에서 여러 변동사항이 반영된 새로운 표본크기 공식을 유도하며 시점별 정보를 사용하는 기존의 표본크기 공식을 제시하여 차이점을 살펴본다. 그래서 제 3절에서는 여러 표본크기 공식들과 새로 제시된 표본크기 공식들을 다양한 환경에서 비교하는 모의실험을 실시한다. 여러 상황에서 의 모의 실험을 통하여 본 논문에서 제시되는 표본크기 공식의 실효성을 살펴본다.

2. 모집단의 변동을 반영한 표본크기 결정

계속조사에서 과거자료의 정보가 있을 때 표본의 크기를 결정하는 문제를 연구한 Park (1989)은 지난 시점 $t-1$ 의 추정량의 변동계수 CV_{t-1} 와 t 시점의 추정량의 목표오차 즉 목표 CV_t 의 비를 이용하였다. 즉 현 시점의 추정량의 목표오차가 지난 시점의 추정량의 변동계수보다 작으면 현시점의 표본크기는 지난 시점의 표본크기와 비교해서 증가되어야 하는 것에서 착안되었다. 임의의 t 시점에서의 표본크기를 n_t 라 하고 지난 시점의 표본의 크기는 n_{t-1} 라 할 때 표본크기 결정을 위한 공식은 다음과 같다.

$$n_t = n_{t-1} \left(\frac{CV_{t-1}}{CV_t} \right)^2, \quad (2.1)$$

여기서 지난 시점의 CV_{t-1} 는 추정되어서 사용된다. 위의 식 (2.1)은 추정량의 산포의 변동만을 고려한다. 만약 모집단의 크기가 t 시점에서 줄어들었다면 표본의 크기가 이전 시점과 비교해서 줄어들어야 하는 데 식 (2.1)의 표본의 크기 공식만 가지고는 구할 수 없게 된다. 그래서 Kim (2012)은 모집단의 크기 변동을 반영한 표본크기 공식을 제시하였다. 시점별 모분산, 모평균과 설계효과가 동일하다는 가정하에

$$n_t = n_{t-1} \left(\frac{CV_{t-1}}{CV_t} \right)^2 \left(1 + \frac{n_{t-1}}{N_{t-1}} \left(1 - \frac{N_{t-1}}{N_t} \left(\frac{CV_{t-1}}{CV_t} \right)^2 \right) \right) \quad (2.2)$$

을 제시하였다. 여기에서 N_t 와 N_{t-1} 은 각각 t 시점과 $t-1$ 시점의 모집단 크기를 나타낸다. 본 공식은 시점별 추정량의 변동계수의 변동과 모집단 크기 변동을 다 포함하고 있다.

본 논문에서는 $t-1$ 시점에서의 추정량의 변동계수와 t 시점에서의 추정량의 목표 변동계수의 변동과 더불어 모집단의 크기 변동과 모집단의 변동계수의 변동을 포함한 표본 크기 공식을 연구한다. t 시점에서의 모집단의 크기를 N_t 라 하고 i 번째 개체 관심변수를 y_{ti} 라 할 때 관심모수인 모평균은 $\mu_t = N_t^{-1} \sum_{i=1}^{N_t} y_{ti}$ 이다. t 시점에서 표본설계에 의한 일차 표본포함확률을 π_{ti} 라 하고 π_{tij} 을 결합표본포함 확률이라 할 때 모평균에 대한 Horvitz-Thompson(HT) 추정량 (Horvitz와 Thompson, 1952)은

$$\bar{y}_{HT(t)} = \frac{1}{N_t} \sum_{i=1}^{n_t} \frac{y_{ti}}{\pi_{ti}}$$

이며, 이 추정량에 대한 Sen-Yates-Grundy(SYG) 분산 공식 (Sen, 1953; Yates와 Grundy, 1953)은

$$V_t(\bar{y}_{HT(t)}) = -\frac{1}{2} \sum_{i=1}^{N_t} \sum_{j=1}^{N_t} (\pi_{tij} - \pi_{ti}\pi_{tj}) \left(\frac{y_{ti}}{\pi_{ti}} - \frac{y_{tj}}{\pi_{tj}} \right)^2$$

이다. 여기서 $CV_t(\bar{y}_{HT(t)}) = \sqrt{V_t(\bar{y}_{HT(t)})}/\mu_t$ 일때 $\bar{y}_{HT(t)}$ 의 정규근사이론을 사용하여 $100(1-\alpha)\%$ 오차한계를 d_t 이라로 설정한다면 $z_{\alpha/2}^{-1}d_t = CV_t(\bar{y}_{HT(t)})$ 가 된다. 만약 $\pi_{ti} = \pi_t$ 와 $\pi_{tij} = \pi_t'$ 라 가정한다면

$$\pi_t' \pi_t^{-2} = -\frac{N_t}{N_t-1} \left(\frac{CV_t(\bar{y}_{HT(t)})}{C_t} \right)^2 + 1 \quad (2.3)$$

이며 S_t^2 이 t 시점에서 모분산일 때 $C_t = \sqrt{S_t^2}/\mu_t$ 는 t 시점에서 관심변수에 대한 모집단 변동계수를 나타낸다. 여기서 포함확률이 모든 개체에 대해 동일하다는 가정을 만족하는 표집방법으로는 단순임의표집과 단순일단집락표집을 들 수 있다.

만약 단순임의표집을 가정한다면 $\pi_t = N_t^{-1}n_t$ 이고 $\pi'_t = [N_t(N_t - 1)]^{-1}n_t(n_t - 1)$ 이므로 식 (2.3)에서 표본크기의 공식을 유도할 수 있다 (Park, 1989).

$$n_t = \frac{(C_t/CV_t(\bar{y}_{HT(t)}))^2}{1 + N_t^{-1}(C_t/CV_t(\bar{y}_{HT(t)}))^2}. \quad (2.4)$$

$t - 1$ 시점에서의 표본크기 n_{t-1} 공식은 식 (2.4)에서 t 시점을 $t - 1$ 로 변경하여 얻어진다. 또한 단순임의표집에서 HT추정량은 표본평균이며 표준오차는 $\sqrt{V(\bar{y}_{HT(t)})} = \sqrt{(n_t^{-1} - N_t^{-1})S_t^2}$ 가 된다. 이 식의 양변을 μ_t 로 나누어서 표본의 크기를 구하면 식 (2.4)를 유도할 수 있다. 식 (2.4)와 n_{t-1} 표본공식의 비를 이용하여 과거시점의 자료를 이용한 표본크기 공식을 유도한다.

$$n_{t-1}^{-1}n_t = \left[\frac{(C_{t-1}/CV_{t-1}(\bar{y}_{HT(t-1)}))^2}{1 + N_{t-1}^{-1}(C_{t-1}/CV_{t-1}(\bar{y}_{HT(t-1)}))^2} \right]^{-1} \frac{(C_t/CV_t(\bar{y}_{HT(t)}))^2}{1 + N_t^{-1}(C_t/CV_t(\bar{y}_{HT(t)}))^2}.$$

결과적으로 t 시점에서 표본크기를 구하는 공식은 다음과 같이 나타내어진다.

$$n_t = n_{t-1} \frac{(CV_{t-1}(\bar{y}_{HT(t-1)})/C_{t-1})^2 + 1/N_{t-1}}{(CV_t(\bar{y}_{HT(t)})/C_t)^2 + 1/N_t}, \quad (2.5)$$

여기에서 C_{t-1} 는 $t - 1$ 시점에서 관심변수에 대한 모집단 변동계수를 나타내고 $CV_{t-1}(\bar{y}_{HT(t-1)})$ 는 $t - 1$ 시점에서의 추정량의 변동계수이며 $CV_t(\bar{y}_{HT(t)})$ 는 t 시점에서의 목표변동계수이다. 식 (2.5)은 시점별 모집단 변동계수들의 변동과 모집단 크기의 변동을 반영한 식이며 $t - 1$ 시점에서의 추정량의 변동계수와 t 시점의 목표변동계수의 증감도 반영한 식이다. 그러므로 식 (2.1)과 비교해 보면 모집단의 변동을 반영하였다는 것을 알 수 있고 식 (2.2)와 비교해 보면 모집단의 산포 변동을 추가로 반영하였다는 것을 알 수 있다. 일반적으로 표본크기 공식은 목표오차 외에 대부분 추정된 값을 사용하게 된다. 새로 제시된 표본크기 공식에서 모수에 대한 추정방법을 두 가지로 제시한다. 첫째, 실제 표본조사에서 관심변수에 대한 과거 표본 자료만을 얻을 수 있을 때는 C_{t-1} 와 C_t 는 관심 변수와 상관성이 높은 보조변수에 대한 t 시점과 $t - 1$ 시점의 모변동계수를 사용하여 계산되는 방법을 제시할 수 있다. 둘째, 만약 추가적으로 현 시점의 사전조사 자료가 존재한다면 C_{t-1} 은 관심변수에 대한 과거 표본조사 자료를 통해 추정되며 C_t 는 관심변수에 대한 사전조사 자료를 통해 추정되는 것을 사용하는 방법을 제시할 수 있다. 또한 $t - 1$ 시점에서의 추정량의 변동계수는 과거조사자료에 의해 추정되어진다.

Remark 2.1: 시점별 모집단의 크기 N_t, N_{t-1} 이 커져 모집단 크기의 역수가 무시할 정도로 작아지고 시점별 모집단의 변동계수가 $C_t = C_{t-1}$ 이라면 식 (2.5)는 식 (2.1)과 같아진다. 그리고 그러한 상황 하에서는 $t - 1$ 시점에서의 추출률이 거의 무시할 정도로 작아지기 때문에 식 (2.5)는 식 (2.2)와도 같아진다.

Remark 2.2: 만약, $C_{t-1} = C_t$ 이고 $CV_{t-1}(\bar{y}_{HT(t-1)}) = CV_t(\bar{y}_{HT(t)})$ 이면, 식 (2.5)는

$$n_t = n_{t-1} \frac{k^2 + 1/N_{t-1}}{k^2 + 1/N_t}$$

가 되고 여기서 $k = C_{t-1}^{-1}CV_{t-1}(\bar{y}_{HT(t-1)}) = C_t^{-1}CV_t(\bar{y}_{HT(t)})$ 이다. 위의 식을 보면 시점별 모집단 크기가 $N_{t-1} < N_t$ 이면 t 시점에서의 표본크기 n_t 는 $t - 1$ 시점에서의 표본크기 n_{t-1} 보다 커짐을 알 수 있다.

Remark 2.3: 시점별 모집단의 변동계수가 같다면 즉, $C_{t-1} = C_t = C$ 이라면 식 (2.5)는

$$n_t = n_{t-1} \frac{(CV_{t-1}(\bar{y}_{HT(t-1)}))^2 + C^2/N_{t-1}}{(CV_t(\bar{y}_{HT}))^2 + C^2/N_t}$$

이 된다. 위의 식을 보면 시점별 추정량의 변동계수들이 $CV_{t-1}(\bar{y}_{HT(t-1)}) > CV_t(\bar{y}_{HT(t)})$ 와 같은 관계를 가진다면 t 시점에서의 표본크기 n_t 는 $t-1$ 시점에서의 표본크기 n_{t-1} 보다 커져야 한다. 그러나 시점별 모집단의 크기가 $N_{t-1} > N_t$ 이라면 t 시점에서의 표본크기 n_t 의 증가량이 감소되는 것을 알 수 있다.

Remark 2.4: 시점별 표본크기 공식의 비를 이용하여 이전시점의 자료를 이용한 표본크기 공식을 유도하였다. 비를 이용한 방법을 층화임의표집에서 비례배정을 사용할 경우 총비용을 고정했을 때의 표본크기 공식에 적용한다. t 시점에서의 총 비용 공식을 $C'_t = c_0 + \sum_{h=1}^H n_{(t)h} c_{(t)h}$ 라 할 때 시점별 표본크기 공식의 비를 이용하여 지난 시점의 정보를 사용하여 표본의 크기를 구한다. 여기서 C'_t 는 t 시점에서의 총비용이고 c_0 는 시점별 같은 고정비용이며, $n_{(t)h}$ 는 t 시점의 h 층에서의 표본크기이고 $c_{(t)h}$ 는 t 시점의 h 층에서의 단위당 비용이다. $W_{(t)h}$ 는 t 시점의 모집단에서의 h 층의 상대적 크기라 하고 $t-1$ 시점에서도 동일한 규칙에 의해 비용 등을 정의하여 표본 크기 공식을 구해보면

$$n_t = n_{t-1} \left(\frac{C'_t - c_0}{C'_{t-1} - c_0} \right) \frac{\sum_{h=1}^H W_{(t-1)h} c_{(t-1)h}}{\sum_{h=1}^H W_{(t)h} c_{(t)h}}$$

이다. 이 표본크기를 구하는 공식을 보면 시점별 총비용이 $C'_{t-1} < C'_t$ 의 관계에 있으면 $n_t > n_{t-1}$ 이고 $C'_t = C'_{t-1}$ 일 때 모든 층에서 단위당 비용 $c_{(t-1)h} > c_{(t)h}$ 이면 t 시점에서의 표본의 크기가 전 시점의 표본크기보다 증가함을 알 수 있다.

3. 모의실험

모의실험을 통하여 시점간 추정량의 변동계수, 모집단의 크기, 모집단의 변동계수의 변동을 반영했을 때 변하는 시점간 표본의 변동을 살펴본다. 본 논문에서 제시되는 표본크기 공식 (2.5)과 Park (1989)과 Kim (2012)에 제시되어 있는 표본크기 공식을 여러 상황에서 비교 분석해 본다.

$t-1$ 시점의 모평균 추정량의 변동계수를 $C_{t-1} = 0.04$ 로 가정하고 표본크기 $n_{t-1} = 100$ 으로 가정하여 시점별 변동계수의 비 $CV_R = CV_{t-1}/CV_t$ 와 $t-1$ 시점의 표본추출률 $f_{t-1} = n_{t-1}/N_{t-1}$, 시점별 모집단의 크기 변동 비율 N_t/N_{t-1} , 모집단의 변동계수 변동 비율 C_{t-1}/C_t 에 따른 표본수의 비 n_t/n_{t-1} 를 여러 표본공식에 의해 구한 것을 비교한다. Table 3.1에서 Table 3.3에 제시되는 표본크기 공식은 4가지로써 다음과 같다. 'M1'은 본 논문에서 제시되는 표본크기 공식 (2.5)이고, 'M2'에 제시되는 표본 공식은 식 (2.2)이며 이것은 Kim (2012)에 제시되어 있고 'M3'에 제시되는 공식은 Park (1989)의 표본크기이며 이것의 공식은 식 (2.1)이다. 'M4'의 표본크기 공식은 Kim (2012)에 제시된 것이며 그 공식은 아래와 같다.

$$n_t = w \times n_{t-1} \left(\frac{CV_{t-1}}{CV_t} \right)^2 + (1-w) \times n_{t-1} \left(\frac{N_t}{N_{t-1}} \right), \quad (3.1)$$

여기서 w 는 가중치에 해당하며 본 모의실험에서는 $w = 1/2$ 를 가정한다. 그리고 Table 3.1은 모집단 변동계수의 변동 비율이 $C_{t-1}/C_t = 1$ 인 경우이며 Table 3.2는 $C_{t-1}/C_t = 0.938$ 인 경우이고 Table 3.3은 $C_{t-1}/C_t = 1.071$ 인 경우이다.

Table 3.1. Ratio of sample size n_t/n_{t-1} for $C_{t-1}/C_t = 1$

CV _R	f _{t-1}	N _t /N _{t-1}											
		0.5				1.0				1.5			
		M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
0.5	0.02	0.281	0.253	0.250	0.375	0.299	0.254	0.250	0.625	0.306	0.254	0.250	0.875
	0.01	0.266	0.251	0.250	0.375	0.275	0.252	0.250	0.625	0.279	0.252	0.250	0.875
	0.002	0.253	0.250	0.250	0.375	0.255	0.250	0.250	0.625	0.256	0.250	0.250	0.875
	0.001	0.252	0.250	0.250	0.375	0.253	0.250	0.250	0.625	0.253	0.250	0.250	0.875
	0.0003	0.251	0.250	0.250	0.375	0.251	0.250	0.250	0.625	0.251	0.250	0.250	0.875
0.8	0.02	0.603	0.636	0.640	0.570	0.695	0.645	0.640	0.820	0.732	0.647	0.640	1.070
	0.01	0.619	0.638	0.640	0.570	0.670	0.642	0.640	0.820	0.689	0.644	0.640	1.070
	0.002	0.635	0.640	0.640	0.570	0.646	0.640	0.640	0.820	0.650	0.641	0.640	1.070
	0.001	0.638	0.640	0.640	0.570	0.643	0.640	0.640	0.820	0.645	0.640	0.640	1.070
	0.0003	0.639	0.640	0.640	0.570	0.641	0.640	0.640	0.820	0.642	0.640	0.640	1.070
1	0.02	0.820	0.980	1.000	0.750	1.000	1.000	1.000	1.000	1.079	1.007	1.000	1.250
	0.01	0.890	0.990	1.000	0.750	1.000	1.000	1.000	1.000	1.043	1.003	1.000	1.250
	0.002	0.973	0.998	1.000	0.750	1.000	1.000	1.000	1.000	1.009	1.001	1.000	1.250
	0.001	0.986	0.999	1.000	0.750	1.000	1.000	1.000	1.000	1.005	1.000	1.000	1.250
	0.0003	0.995	1.000	1.000	0.750	1.000	1.000	1.000	1.000	1.002	1.000	1.000	1.250
1.2	0.02	1.019	1.386	1.440	0.970	1.313	1.427	1.440	1.220	1.453	1.441	1.440	1.470
	0.01	1.169	1.413	1.440	0.970	1.366	1.434	1.440	1.220	1.447	1.441	1.440	1.470
	0.002	1.370	1.435	1.440	0.970	1.423	1.439	1.440	1.220	1.442	1.440	1.440	1.470
	0.001	1.403	1.437	1.440	0.970	1.431	1.439	1.440	1.220	1.441	1.440	1.440	1.470
	0.0003	1.427	1.439	1.440	0.970	1.437	1.440	1.440	1.220	1.440	1.440	1.440	1.470
1.5	0.02	1.272	2.093	2.250	1.375	1.766	2.194	2.250	1.625	2.027	2.228	2.250	1.875
	0.01	1.572	2.171	2.250	1.375	1.950	2.222	2.250	1.625	2.119	2.239	2.250	1.875
	0.002	2.053	2.234	2.250	1.375	2.176	2.244	2.250	1.625	2.220	2.248	2.250	1.875
	0.001	2.146	2.242	2.250	1.375	2.212	2.247	2.250	1.625	2.235	2.249	2.250	1.875
	0.0003	2.214	2.247	2.250	1.375	2.237	2.249	2.250	1.625	2.245	2.250	2.250	1.875

Table 3.1에서 Table 3.3의 전반적인 성격을 살펴보면 몇가지로 정리할 수 있다. 첫째, CV_R의 값이 커지면 n_t/n_{t-1} 의 값이 커지고 그와 반대 현상이면 표본변동의 비의 값이 작아짐을 알 수 있다. 이것은 t 시점의 목표오차(추정량의 변동계수)가 전 시점의 추정량의 변동계수보다 작아지면 t 시점의 표본 크기는 늘려야 됨을 뜻한다. 또한 N_t/N_{t-1} 의 값이 작아지면 n_t/n_{t-1} 의 비율이 작아지고 N_t/N_{t-1} 의 값이 커지면 반대 현상이 일어난다. 모집단의 크기가 전 시점보다 증가하는 경우에는 표본크기는 늘어나야 됨을 뜻한다. 그러나 'M3' 표본크기 공식은 모집단의 변동이 반영되어 있는 공식이 아니므로 N_t/N_{t-1} 의 값의 변동에 상관없이 일정함을 알 수 있다

둘째, $t - 1$ 시점의 표본수와 비교하여 t 시점의 표본수 결정에 영향을 주는 것 중 추정량의 변동계수 변동 CV_R과 모집단의 크기 변동 N_t/N_{t-1} 을 봤을 때 CV_R의 영향력이 더 큼을 알 수 있다. 예를 들어 Table 3.1에서 CV_R = 0.5이고 $f_{t-1} = 0.02$ 일 때 $N_t/N_{t-1} = 1.5$ 의 M1에 대한 n_t/n_{t-1} 의 값과 $N_t/N_{t-1} = 0.5$ 의 M1에 대한 n_t/n_{t-1} 의 값의 차이를 보면 0.025인데 반해 $N_t/N_{t-1} = 0.5$ 이고 $f_{t-1} = 0.02$ 일 때 CV_R = 1.5의 M1에 대한 n_t/n_{t-1} 의 값과 CV_R = 0.5의 M1에 대한 n_t/n_{t-1} 의 값의 차이는 0.991이 된다.

셋째, M1과 M2 표본크기는 CV_R이 감소하고 N_t/N_{t-1} 이 증가하는 경우에는 추출률이 증가함에 따라 n_t/n_{t-1} 의 비가 감소하며 그와 반대되는 경우인 표본크기는 CV_R이 증가하고 N_t/N_{t-1} 이 감소하는 경

Table 3.2. Ratio of sample size n_t/n_{t-1} for $C_{t-1}/C_t = 0.938$

CV_R	f_{t-1}	N_t/N_{t-1}											
		0.5				1.0				1.5			
		M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
0.5	0.02	0.314	0.253	0.250	0.375	0.337	0.254	0.250	0.625	0.346	0.254	0.250	0.875
	0.01	0.300	0.251	0.250	0.375	0.312	0.252	0.250	0.625	0.316	0.252	0.250	0.875
	0.002	0.288	0.250	0.250	0.375	0.290	0.250	0.250	0.625	0.291	0.250	0.250	0.875
	0.001	0.286	0.250	0.250	0.375	0.287	0.250	0.250	0.625	0.288	0.250	0.250	0.875
	0.0003	0.285	0.250	0.250	0.375	0.285	0.250	0.250	0.625	0.286	0.250	0.250	0.875
0.8	0.02	0.662	0.636	0.640	0.570	0.774	0.645	0.640	0.820	0.821	0.647	0.640	1.070
	0.01	0.689	0.638	0.640	0.570	0.753	0.642	0.640	0.820	0.778	0.644	0.640	1.070
	0.002	0.719	0.640	0.640	0.570	0.734	0.640	0.640	0.820	0.739	0.641	0.640	1.070
	0.001	0.724	0.640	0.640	0.570	0.731	0.640	0.640	0.820	0.733	0.640	0.640	1.070
	0.0003	0.727	0.640	0.640	0.570	0.729	0.640	0.640	0.820	0.730	0.640	0.640	1.070
1	0.02	0.889	0.980	1.000	0.750	1.104	1.000	1.000	1.000	1.201	1.007	1.000	1.250
	0.01	0.983	0.990	1.000	0.750	1.119	1.000	1.000	1.000	1.173	1.003	1.000	1.250
	0.002	1.099	0.998	1.000	0.750	1.134	1.000	1.000	1.000	1.145	1.001	1.000	1.250
	0.001	1.118	0.999	1.000	0.750	1.136	1.000	1.000	1.000	1.142	1.000	1.000	1.250
	0.0003	1.131	1.000	1.000	0.750	1.137	1.000	1.000	1.000	1.139	1.000	1.000	1.250
1.2	0.02	1.092	1.386	1.440	0.970	1.437	1.427	1.440	1.220	1.606	1.441	1.440	1.470
	0.01	1.279	1.413	1.440	0.970	1.519	1.434	1.440	1.220	1.620	1.441	1.440	1.470
	0.002	1.542	1.435	1.440	0.970	1.610	1.439	1.440	1.220	1.634	1.440	1.440	1.470
	0.001	1.588	1.437	1.440	0.970	1.624	1.439	1.440	1.220	1.636	1.440	1.440	1.470
	0.0003	1.621	1.439	1.440	0.970	1.634	1.440	1.440	1.220	1.638	1.440	1.440	1.470
1.5	0.02	1.344	2.093	2.250	1.375	1.907	2.194	2.250	1.625	2.216	2.228	2.250	1.875
	0.01	1.698	2.171	2.250	1.375	2.147	2.222	2.250	1.625	2.355	2.239	2.250	1.875
	0.002	2.301	2.234	2.250	1.375	2.455	2.244	2.250	1.625	2.511	2.248	2.250	1.875
	0.001	2.422	2.242	2.250	1.375	2.506	2.247	2.250	1.625	2.535	2.249	2.250	1.875
	0.0003	2.512	2.247	2.250	1.375	2.542	2.249	2.250	1.625	2.552	2.250	2.250	1.875

우에는 추출률이 증가함에 따라 시점별 표본변동의 비 n_t/n_{t-1} 가 증가함을 알 수 있다. M3와 M4 표본 공식은 $t-1$ 시점의 추출률 변화에는 셀마다 일정한 값을 가짐을 알 수 있다. M3는 추출률과 관계 없는 것이기 때문에 셀마다 일정한 값을 가짐을 알 수 있다.

넷째, M1, M2, M4 표본공식은 $N_t/N_{t-1} = 1$ 인 경우를 기준으로 봤을 때 CV_R 이 감소하면 n_t 의 표본이 지난 시점의 표본보다 작게 뽑아도 됨을 알 수 있고 CV_R 이 증가하면 t 시점의 표본이 n_{t-1} 보다 많이 뽑혀야 됨을 알 수 있다. 그러나 $N_t/N_{t-1} > 1$ 인 경우에는 CV_R 이 감소할 때 n_t 의 표본의 감소 비율이 모집단의 크기가 시점별 같을 때와 비교하여 경감됨을 알 수 있다. 이것은 모집단의 변동비율의 영향이 표본크기 결정에 포함됨을 알 수 있다.

Table들의 각각의 성질을 살펴보면 다음과 같이 정리할 수 있다. 첫째, Table 3.1과 Table 3.2~Table 3.3의 차이점은 시점별 모집단의 변동계수의 변동 여부에 있으므로 Table 3.1에서는 $CV_R = 1$ 이고 $N_t/N_{t-1} = 1$ 인 경우 시점별 변동이 없는 경우에는 $n_t/n_{t-1} = 1$ 임을 알 수 있다. 그러나 Table 3.2와 Table 3.3에서는 $CV_R = 1$ 이고 $N_t/N_{t-1} = 1$ 인 경우 M2, M3, M4의 표본크기공식에서만 $n_t/n_{t-1} = 1$ 임을 알 수 있다. 그것은 M1은 시점별 모집단의 변동계수의 변동이 반영된 식으로써 Table 3.2에서는 $C_{t-1}/C_t < 1$ 이기 때문에 n_t/n_{t-1} 의 값이 1보다 큼을 알 수 있고 Table 3.3에서는 $C_{t-1}/C_t > 1$ 이기 때문에 n_t/n_{t-1} 의 값이 1보다 작아짐을 알 수 있다.

Table 3.3. Ratio of sample size n_t/n_{t-1} for $C_{t-1}/C_t = 1.071$

CV _R	f _{t-1}	N _t /N _{t-1}											
		0.5				1.0				1.5			
		M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
0.5	0.02	0.249	0.253	0.250	0.375	0.263	0.254	0.250	0.625	0.268	0.254	0.250	0.875
	0.01	0.234	0.251	0.250	0.375	0.241	0.252	0.250	0.625	0.243	0.252	0.250	0.875
	0.002	0.221	0.250	0.250	0.375	0.223	0.250	0.250	0.625	0.223	0.250	0.250	0.875
	0.001	0.219	0.250	0.250	0.375	0.220	0.250	0.250	0.625	0.220	0.250	0.250	0.875
	0.0003	0.218	0.250	0.250	0.375	0.219	0.250	0.250	0.625	0.219	0.250	0.250	0.875
0.8	0.02	0.544	0.636	0.640	0.570	0.617	0.645	0.640	0.820	0.647	0.647	0.640	1.070
	0.01	0.550	0.638	0.640	0.570	0.590	0.642	0.640	0.820	0.604	0.644	0.640	1.070
	0.002	0.556	0.640	0.640	0.570	0.564	0.640	0.640	0.820	0.567	0.641	0.640	1.070
	0.001	0.557	0.640	0.640	0.570	0.561	0.640	0.640	0.820	0.562	0.640	0.640	1.070
	0.0003	0.557	0.640	0.640	0.570	0.559	0.640	0.640	0.820	0.559	0.640	0.640	1.070
1	0.02	0.749	0.980	1.000	0.750	0.896	1.000	1.000	1.000	0.959	1.007	1.000	1.250
	0.01	0.798	0.990	1.000	0.750	0.885	1.000	1.000	1.000	0.919	1.003	1.000	1.250
	0.002	0.854	0.998	1.000	0.750	0.874	1.000	1.000	1.000	0.881	1.001	1.000	1.250
	0.001	0.862	0.999	1.000	0.750	0.873	1.000	1.000	1.000	0.876	1.000	1.000	1.250
	0.0003	0.868	1.000	1.000	0.750	0.872	1.000	1.000	1.000	0.873	1.000	1.000	1.250
1.2	0.02	0.942	1.386	1.440	0.970	1.188	1.427	1.440	1.220	1.301	1.441	1.440	1.470
	0.01	1.058	1.413	1.440	0.970	1.216	1.434	1.440	1.220	1.280	1.441	1.440	1.470
	0.002	1.205	1.435	1.440	0.970	1.246	1.439	1.440	1.220	1.260	1.440	1.440	1.470
	0.001	1.229	1.437	1.440	0.970	1.250	1.439	1.440	1.220	1.257	1.440	1.440	1.470
	0.0003	1.246	1.439	1.440	0.970	1.253	1.440	1.440	1.220	1.255	1.440	1.440	1.470
1.5	0.02	1.194	2.093	2.250	1.375	1.619	2.194	2.250	1.625	1.836	2.228	2.250	1.875
	0.01	1.441	2.171	2.250	1.375	1.753	2.222	2.250	1.625	1.889	2.239	2.250	1.875
	0.002	1.815	2.234	2.250	1.375	1.910	2.244	2.250	1.625	1.944	2.248	2.250	1.875
	0.001	1.884	2.242	2.250	1.375	1.934	2.247	2.250	1.625	1.952	2.249	2.250	1.875
	0.0003	1.934	2.247	2.250	1.375	1.951	2.249	2.250	1.625	1.957	2.250	2.250	1.875

둘째, Table 3.1을 살펴보면 시점별 모집단의 변동계수가 같고 시점별 모집단의 크기가 클 때(추출률이 작아질 때) M1 표본크기 공식은 M2, M3 표본크기공식과 같아지는 데 실제로 M1, M2, M3의 값을 살펴보면 추출률이 작아질 때 비슷해지는 경향이 있음을 알 수 있다.

셋째, Table 3.2를 살펴보면 t시점의 모집단의 변동계수가 크기 때문에 M1 표본크기 값은 Table 3.1의 M1값과 비교하여 t시점의 표본이 더 많이 뽑혀야 됨을 알 수 있으며 M2, M3, M4 표본크기 값은 Table 3.1과 Table 3.2가 동일함을 알 수 있다. 또한, Table 3.3은 t시점의 모집단의 변동계수가 작기 때문에 M1 표본크기 값은 Table 3.1의 M1값과 비교하여 t시점의 표본이 더 적게 뽑혀야 됨을 알 수 있으며 M2, M3, M4 표본크기 값은 Table 3.1과 Table 3.3이 동일함을 알 수 있다.

종합적으로 설명하면 M3는 이전시점의 추정량의 변동계수와 현 시점의 추정량의 목표오차의 변동만을 고려한 식이며 M2, M4는 시점별 추정량의 변동계수의 변동의 정보와 시점별 모집단의 크기 변동의 정보를 사용한 공식이고 본 논문에서 제공되는 M1 표본크기 공식은 이전시점의 추정량의 변동계수와 현 시점의 추정량의 목표오차간 변동정보와 시점별 모집단의 크기 변동의 정보와 시점별 모집단의 변동계수의 변동정보를 포함하여 제공되는 차이점이 있다. 그러므로 모집단이 크기변동과 산포변동이 발생하는 상황에서 이전시점의 추정량의 변동계수보다 현 시점의 추정량의 목표오차의 값이 변동이 생긴 경우에는 본 논문에서 제시되는 M1 표본크기 공식을 사용하는 것이 적당하다.

실제자료의 예제를 사용하기 위해 계속조사의 성격을 가지고 있는 2009년 국민여행실태조사자료 (Park 등, 2010)를 사용하여 여러 표본공식을 비교한다. 그래서 2009년 국민여행실태조사에서 6516명의 국내 여행 비용에 관한 표본자료를 과거시점의 모집단자료로 가정하며 다음 시점의 비용 자료를 균등분포에서 e_{t-1} 을 발생시켜 $y_t = 1000 + 1.2 * y_{t-1} + e_{t-1}$ 의 시계열 모형을 사용하여 가상의 모집단을 발생시키고 그중 516명의 자료를 삭제 시켜서 다음 시점의 모집단 자료가 줄어드는 상황을 가정한다. 여기서 과거시점의 비용자료에 대한 모집단의 변동계수는 2.704이며 다음 시점의 비용에 대한 모집단의 변동계수는 2.505로 계산된다. 과거 시점의 비용자료에서 단순임의표집을 사용하여 500개의 표본을 발생시켜 추정량의 변동계수를 0.094로 추정한다. 다음 시점의 목표변동계수를 0.085로 결정 할 때 본 논문에서 제안된 방법으로 표본크기를 구해보고 다른 표본크기결정방법들과 비교하면 식 (2.1)에 대한 표본크기는 611명이며 식 (2.2)에 대한 크기는 596명이며 식 (3.1)에 대한 표본은 536명이다. 본 논문에서 제시된 식 (2.5)에 대한 표본크기는 516명으로 계산된다. 추정량의 변동계수보다 목표변동계수를 작게 잡았을 때 과거 시점의 표본크기보다 모두 커짐을 알 수 있으며 Park (1989)가 제시한 표본크기가 가장 큰 것을 알 수 있다. 그에 반해 Kim (2012)이 제시한 표본크기공식과 본 논문에서 제시된 공식은 모집단의 크기가 줄어드는 것이 반영되어 Park (1989)과 비교하여 더 줄어듬을 알 수 있다. 또한 본 논문에서 제시된 표본크기는 모집단의 줄어드는 산표변동도 반영되어 더 작아짐을 알 수 있다. 그러므로 모집단의 크기와 산포의 변동을 동시에 고려되는 것은 본 논문에서 제시된 표본크기공식임을 알 수 있다. 그러나 여러가지 모의실험에서 제시된 표본크기 공식은 몇개의 모수들을 안다는 가정하에 실시한 것이다. 실제 표본조사에서는 표본크기 공식에 있는 모수들을 표본자료에 의해 추정되어 사용되어 진다. 향후 연구과제로 표본크기 공식에 추정된 값들을 사용했을 때의 변동을 살펴보는 연구를 진행할 수 있다.

References

- Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663–685.
- Kim, K. S. (2012). Sample size determination in repeated surveys with varying population sizes, *Survey Research*, **13**, 159–174.
- Park, H., Park, S., Park, J. and Jeon, J. (2010). Selection of stratification variables under a new sampling frame: A case study for the Korea National Tourism Survey, *Survey Research*, **11**, 103–114.
- Park, H. N. (1989). *Statistical Survey (2nd edition)*, Youngji Publishers, Seoul.
- Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities, *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.
- Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size, *Journal of the Royal Statistical Society: Series B*, **15**, 235–261.
- Yoo, Y. and Shin, K. L. (2011). A study on the decision of sample size for panel survey design, *The Korean Journal of Applied Statistics*, **24**, 25–34.

계속조사에서의 표본크기 결정

박현아^{a,1} · 나성룡^b

^a서울대학교 통계학과, ^b연세대학교 정보통계학과

(2014년 2월 7일 접수, 2014년 4월 9일 수정, 2014년 6월 16일 채택)

요약

일반적으로 현 시점에서 목표로 하고 있는 추정량의 산포가 지난시점의 추정량의 산포보다 크다면 지난시점과 비교하여 현 시점의 표본의 크기는 줄어드는 것이 타당하다. 계속조사에서 지난 시점의 추정량의 변동계수와 모집단의 크기 변동과 현 시점의 추정량의 목표오차를 이용하여 표본의 크기를 결정하는 것을 연구한 여러 논문들이 있다. 그런데 모집단은 크기의 변동과 산포의 변동이 있을 수 있으므로 본 연구에서는 지난 시점의 추정량의 변동계수와 모집단의 크기, 모집단의 산포 변동과 현 시점의 추정량의 목표오차를 반영하여 현 시점의 표본의 크기를 구하는 문제를 연구한다. 또한 모의실험을 통하여 기존 표본크기의 공식들과 비교분석한다.

주요용어: 계속조사, 모집단크기, 변동계수, 표본크기.

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2012R1A1A3003761).

¹교신저자: (151-742) 서울시 관악구 관악로 1, 서울대학교 통계학과. E-mail: hapk@daum.net