The Korean Journal of Applied Statistics (2018) DOI: https://doi.org/10.5351/KJAS.2018.31.5.587
31(5), 587–597

Sample size using response rate on repeated surveys

Hyeonah Park[a] · Seongryong Na[a,1]

[a]Department of Information and Statistics, Yonsei University

(Received May 31, 2018; Revised July 16, 2018; Accepted July 16, 2018)

Abstract

Procedures, such as sampling technique, survey method, and questionnaire preparation, are required in order to obtain sample data in accordance with the purpose of a survey. An important procedure is the decision of the sample size formula. The sample size formula is determined by setting the target error and total cost according to the sampling method. In this paper, we propose a sample size formula using population changes over time, estimation error of the previous time and response rate of past data when the target error and the expected response rate are given in the simple random sampling. In actual research, we use estimators that apply complex weights in addition to design-based weights. Therefore, we induce a sample size formula for estimators using design-based weights and nonresponse adjustment coefficients, that can be a formula that reflects differences in response rates when survey methods are changed over time. In addition, we use simulations to compare the proposed formula with the existing sample size formula.

Keywords: repeated survey, sample size, response rate, population size, target error

1. 서론

여러 추출방법과 추정방법들에 의해 유도되는 표본크기 공식들 중에서 현 설계에 맞는 적절한 방안을 선택하여 표본크기를 결정한다. 대부분의 표본크기는 목표오차와 총비용 등에 의해 주어지는데 정밀한 조사의 목적을 가진다면 목표오차를 작게 잡아 표본 크기를 크게 산출하며 비용이 작게 정해져 있다면 크기를 줄이는 방향으로 결정한다. 일반적으로 총비용은 일정부분 미리 결정된다고 한다면 목표오차의 기준에 따라 표본의 크기가 정해지는데 여러 시점에서의 계속조사에서는 과거 추정오차의 정보의 도움을 받을 수 있다. 즉 과거 추정오차의 결과가 효율적이지 않으면 현 시점에서 목표오차를 작게 잡아 표본크기를 과거시점에 비해 크게 정할 수 있다.

이와 같은 연구로 Park (1989)은 과거 시점의 표본크기에 그 시점의 추정량의 상대표준오차와 목표오차의 비의 함수를 이용하여 현재 시점의 표본 크기를 구하는 것을 제안하였다. 이에 대한 연구의 확장으로 Kim (2012)은 계속조사에서 상대표준오차의 시점별 변동과 모집단의 크기 변동을 반영한 표본크기 공식을 제안하였고 Park과 Na (2014)에서 시점별 모집단의 크기 및 산포의 변동과 과거 시점의 상대표준오차에 따른 현재 시점의 목표오차를 고려한 표본크기 공식을 제시하였다. 그리고 추정에 표본크기 결정이 미치는 영향성에 대한 연구로 Han과 Lee (2015)에서는 실제 사업체 자료를 사용하여 Park (1989)의 표본크기 공식의 적합성을 살펴보았으며 Yoo와 Shin (2011)은 장애인고용실태조사인 패널자

[1]Corresponding author: Department of Information and Statistics, Yonsei University, 1 Yonseidae-gil, Wonju, Gangwon-do 26493, Korea. E-mail: nasr@yonsei.ac.kr

료를 사용하여 추이확률을 이용한 부모집단의 비율과 총계추정에서 표본크기가 미치는 영향을 연구하였다. 또한 배정에 관한 연구로 Lee와 Park (2015)의 연구에서는 층화임의추출에서 응답률을 반영한 배정공식을 제안하기도 하였다.

지금까지 계속조사의 표본크기 공식들의 산출을 위해 사용되는 추정량의 형태는 설계 가중치만 사용한 것이었는데 실제 조사에서는 무응답의 보정과 사후가중 등의 절차가 반영된 복잡한 가중치를 사용하여 추정량의 값을 계산한다. 그리고 실제로 표본크기를 결정하기 위해서는 추출방법, 추정방법, 추정량의 정확도 기준, 표본조사를 통해 얻고자 하는 조사결과, 조사방법, 조사원들의 업무량, 전체적인 조사비용 등이 복합적으로 고려되어야 한다. 그러나 현재의 표본크기 공식들은 위의 과정을 반영하지 않았기 때문에 본 논문에서 단순임의추출에서 설계가중치와 무응답 보정 가중치를 사용한 추정량을 가지고 표본크기 공식을 유도한다. 또한 계속조사에서 과거 정보를 사용할 수 있는 경우 제안되는 표본크기 공식은 과거 시점의 상대표준오차와 목표오차, 시점별 모집단의 크기 및 산포의 변동, 과거 시점의 응답률과 현 시점의 예상 응답률의 의미를 담게 된다. 그리고 시점별 조사방법이 다를 경우 응답률의 차이를 보일 수 있는데 본 연구는 그 의미를 내포할 수 있는 표본크기 공식이며 기존 Park과 Na (2014)의 연구의 확장으로 볼 수 있다. 추가적으로 본 연구는 비추정량에서의 표본크기 공식의 내용을 포함한다.

이와 같은 논문의 내용을 두 부분으로 나누어서 정리하는데, 첫째 단순임의추출에서 모평균 추정을 위한 설계가중치와 무응답 보정을 사용한 선형추정량 또는 비추정량에서 목표오차가 결정될 때 표본크기 공식의 이론적 내용을 유도하며 계속조사에서 시점별 변동을 반영한 공식으로 확장하고 이론적 의미들을 입증한다. 둘째 제안된 공식의 의미를 살펴보기 위해 상대표준오차의 변동, 모집단의 크기 및 산포의 변동, 응답률의 변동을 반영하는 다양한 모의 환경에서 기존의 표본크기 공식들과 비교하여 제안된 공식 사용의 타당성을 실험적으로 살펴본다.

2. 시점별 정보와 응답률을 반영한 표본크기 공식

크기가 N인 모집단에서 관심변수 자료를 y_i라 하고 보조변수의 자료를 x_i라 할 때 주요 추계 정보 중 모평균(population mean) $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$와 모비(population ratio) $r = (\sum_{i=1}^{N} x_i)^{-1}\sum_{i=1}^{N} y_i$에 대한 추정을 위해서 가중치 부분에 표본설계효과와 무응답 및 사후층화보정 등이 이루어져야 한다. 먼저 모비 추정을 위해 단순임의표본추출(simple random sampling)에서 표본설계효과와 무응답 보정이 된 추정량의 형태는 설계효과에 의한 표본포함확률(inclusion probability) π_i와 무응답 보정을 위한 응답확률(response probability) ϕ_i를 사용하여

$$\hat{r} = \left(\sum_{i=1}^{N} \pi_i^{-1}\phi_i^{-1} I_i R_i x_i\right)^{-1} \sum_{i=1}^{N} \pi_i^{-1}\phi_i^{-1} I_i R_i y_i \tag{2.1}$$

이고 I_i는 i번째 개체가 표본에 포함되는 여부를 나타내는 지시변수(indicator variable)이고 R_i는 i번째 개체의 응답여부를 나타내는 지시변수로써

$$I_i = \begin{cases} 1, & \text{if } i \in \mathrm{S}, \\ 0, & \text{otherwise}, \end{cases} \qquad R_i = \begin{cases} 1, & \text{if } i \in S_r, \\ 0, & \text{otherwise} \end{cases}$$

식으로 표현되고 S와 S_r은 각각 표본으로 구성된 집합과 표본에서 응답한 집합이다. 그리고 $x_i = 1$인 경우 식 (2.1)은 Hájek 추정량으로 모평균 추정량의 일종이 된다 (Kim과 Kim, 2007).

표본크기 공식을 유도하기 위해 추정량의 분산을 계산하고자 먼저 테일러(Taylor)의 정리를 사용하여

식 (2.1)를

$$\hat{r} = r + \left(\sum_{i=1}^{N} x_i\right)^{-1} \sum_{i=1}^{N} \pi_i^{-1} \phi_i^{-1} t_i R_i (y_i - r x_i).$$

형태로 전개할 수 있고 추정량 $\hat{r}$의 분산은 표본에서의 응답구조에 대한 조건부 평균 E_R 및 분산 V_R 계산을 실시한 후 표본설계에 의한 평균 E_S 및 분산 V_S를 계산함으로써 다음과 같이 유도된다.

$$\begin{aligned} V(\hat{r}) &= V_S\left[E_R(\hat{r})\right] + E_S\left[V_R(\hat{r})\right] \\ &= \frac{1}{(\sum_{i=1}^{N} x_i)^2} \left[\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{(y_i - r x_i)}{\pi_i} \frac{(y_j - r x_j)}{\pi_j} (\pi_{ij} - \pi_i \pi_j) + \sum_{i=1}^{N} \frac{(y_i - r x_i)^2 (\phi_i^{-1} - 1)}{\pi_i}\right]. \end{aligned}$$

단순임의표본추출에서 추정량의 분산 $V(\hat{r})$을 일정한 값 V'로 고정한 경우 표본크기 공식을 다음과 같이 유도할 수 있다.

$$n = \frac{S_d^2 + N^{-1} \sum_{i=1}^{N} (y_i - r x_i)^2 (\phi_i^{-1} - 1)}{\mu_x^2 V' + N^{-1} S_d^2} \tag{2.2}$$

단, S_d^2은 $y_i - r x_i$의 모분산으로써 $S_d^2 = (N-1)^{-1} \sum_{i=1}^{N} (y_i - r x_i)^2$이고 $\mu_x = N^{-1} \sum_{i=1}^{N} x_i$는 보조변수의 모평균이다. 응답확률이 일정한 값 $\phi_i = \phi$라고 가정하고 모집단 크기가 $N \simeq N - 1$이라 가정한다면 식 (2.2)는

$$n = \frac{S_d^2}{\phi\left(\mu_x^2 V' + N^{-1} S_d^2\right)} \tag{2.3}$$

이 되고 추가적으로 $x_i = 1$로 설정하면 $S^2 = (N-1)^{-1} \sum_{i=1}^{N} (y_i - \mu_y)^2$을 사용하여

$$n = \frac{S^2}{\phi(V' + N^{-1} S^2)} \tag{2.4}$$

이 된다. 이것은 식 (2.1)이 모평균 추정량이 되는 것에 대한 분산을 고정했을 때의 표본크기 공식이라 할 수 있다. 이 표본크기 공식은 목표오차에 의해 결정된 표본크기에 일정한 응답확률 즉 응답률이 반영된 것이라 할 수 있다. 또한 식 (2.4)의 양변에 μ_y를 사용하면 상대표준오차를 사용한 식으로 재구성될 수 있다.

$$n = \frac{CV^2}{\phi(cv'^2 + N^{-1} CV^2)}. \tag{2.5}$$

단, $CV = \sqrt{S^2}/\mu_y$이며 $CV(\hat{r}) = \sqrt{V(\hat{r})}/\mu_y$인데 $CV(\hat{r}) = cv'$로 고정한다. Park (1989)에서 단순임의추출하에 모집단의 변동계수와 추정량의 상대표준오차에 따라 필요한 표본크기를 나타내는 표들을 제시하고 있는데 그 결과를 살펴보면 모집단의 변동계수가 클수록 표본크기가 커지고 목표정도가 현저히 작아질 때 표본크기 증가율이 커짐을 알 수 있는데 이와 같은 현상이 반영된 표본크기 공식이라 할 수 있다.

또한 계속조사에서 과거의 조사결과를 기반으로 하여 t 시점의 표본크기를 결정하기 위해 식 (2.5)의 표본크기공식을 사용하여 과거시점인 $t-1$ 시점의 표본크기 공식과 t 시점의 공식을 나타낸다. 그리고 두 시점에 대한 표본크기 공식의 비를 이용한다.

$$n_{t-1}^{-1} n_t = \left[\frac{CV_{t-1}^2}{\phi_{t-1}\left(cv_{t-1}'^2 + N_{t-1}^{-1} CV_{t-1}^2\right)}\right]^{-1} \frac{CV_t^2}{\phi_t\left(cv_t'^2 + N_t^{-1} CV_t^2\right)}$$

결과적으로 계속조사에서 과거 시점인 $t-1$ 시점의 정보를 사용하여 현재 시점인 t 시점의 표본크기 공식을 유도할 수 있다. 그러므로 현재시점의 표본크기 n_t는

$$n_t = n_{t-1}\frac{\phi_{t-1}\left(CV_{t-1}^{-2}cv_{t-1}'^2 + N_{t-1}^{-1}\right)}{\phi_t\left(CV_t^{-2}cv_t'^2 + N_t^{-1}\right)} \tag{2.6}$$

이며 ϕ_{t-1}, CV_{t-1}, N_{t-1}, cv_{t-1}'는 각각 과거 시점 $t-1$의 응답률, 모변동계수, 모집단크기, 추정량의 상대표준오차이며 ϕ_t, CV_t, N_t, cv_t'는 각각 현 시점 t의 목표 응답률, 모변동계수, 모집단크기, 목표오차이다. 제안된 표본크기 공식을 살펴보면 과거 시점의 상대표준오차와 현 시점의 목표오차의 변동과 두 시점의 모집단의 크기 및 변동계수의 차이가 반영되었으며 추가적으로 과거 시점의 응답률과 목표 응답률의 변화도 반영된 표본크기 공식임을 알 수 있다.

Remark 2.1: 비추정량 $\hat{r}$을 사용하여 모평균 추정량으로

$$\bar{y}_r = \hat{r}\mu_x$$

와 같은 추정량을 제시할 수 있고 이 추정량을 사용하여 일정 응답확률을 가정하고 모집단크기가 $N \simeq N-1$임을 가정할 때 일정 분산 $V(\bar{y}_r) = V''$하에 표본크기 공식을 유도하면 식 (2.3)과 비슷하게 다음과 같이 표현할 수 있다.

$$n = \frac{S_d^2}{\phi(V'' + N^{-1}S_d^2)}.$$

위의 표본크기의 공식은 목표오차 V''의 정도에 따라 변동이 발생하며 응답률을 반영하여 적당한 표본크기를 확보할 수 있다.

Remark 2.2: 두 시점의 모집단 크기 N_{t-1}, N_t가 충분히 커서 무시할 수 있고 시점별 응답률 ϕ_{t-1}, ϕ_t가 비슷하다는 가정을 한다. 그리고 상대표준오차(변동계수) 자체가 세월이 가도 그다지 변하지 않는 성질을 이용하면 (Sung, 2012) 시점별 모집단의 상대표준오차가 $CV_{t-1} = CV_t$임을 가정할 수 있다. 결과적으로 식 (2.6)은

$$n_t = n_{t-1}\left(\frac{cv_{t-1}}{cv_t}\right)^2 \tag{2.7}$$

식으로 추정량의 형태와 상대표준오차는 다르지만 Park (1989)이 제시한 계속조사에서 과거자료의 정보를 사용하여 표본크기를 결정하는 공식으로 도출될 수 있다. 즉 과거 상대표준오차보다 현 시점의 목표오차가 작다면 표본크기가 커지게 되는 공식이다.

Remark 2.3: 식 (2.6)은 $\phi_{t-1} = \phi_t$이면 추정량의 형태와 상대표준오차는 다르지만 Park과 Na (2014)의 논문에서 제시된 표본크기 공식과 유사하게 된다.

$$n_t = n_{t-1}\frac{\left(CV_{t-1}^{-2}cv_{t-1}'^2 + N_{t-1}^{-1}\right)}{\left(CV_t^{-2}cv_t'^2 + N_t^{-1}\right)} \tag{2.8}$$

면접조사, 전화조사, 우편조사와 같은 여러가지 조사방법들은 응답률에서 현저한 차이를 보인다. 즉 조사방법과 응답률간에 연관관계가 존재한다고 할 수 있으므로 현 조사에서 조사방법 중 응답률을 높일 수 있는 방법을 선택한다면 가령 예전에는 전화조사를 하였는데 면접조사로 바꾸려는 계획을 가지고 있다면 $[(\phi_t)^{-1}\phi_{t-1}] < 1$가 되므로 식 (2.8)에 의해 계산되는 표본크기보다 더 작은 표본크기로 동일한 효율

을 얻을 수 있게 된다. 이와 반대로 면접조사에서 우편조사로 변경 계획이 있다면 응답률이 작아질 것을 대비해야 하므로 $[(\phi_t)^{-1}\phi_{t-1}] > 1$가 되어 실제로 표본크기가 식 (2.8)의 크기보다 크게 계산된다.

만약 $\phi_{t-1} = \phi_t$이고 $cv'_{t-1} = cv'_t = cv'$라고 가정한다면 식 (2.6)이

$$n_t = n_{t-1}\frac{\left(CV_{t-1}^{-2}cv'^2 + N_{t-1}^{-1}\right)}{\left(CV_t^{-2}cv'^2 + N_t^{-1}\right)}$$

이 되고 표본크기 공식의 의미를 살펴보면 모집단의 변동계수가 현 시점에서 작아지고 모집단의 크기가 작아지면 현 시점의 표본크기는 과거 시점보다 작아지게 되는 것이며 의미상으로 변동이 작고 크기가 작아진 모집단에 대해서는 표본크기를 작게 해도 된다는 것을 반영한 공식이라 할 수 있다.

또한 N_t와 N_{t-1}이 충분히 크고 $CV_t = CV_{t-1}$이면 식 (2.6)은

$$n_t = n_{t-1}\left(\frac{\phi_{t-1}cv_{t-1}^2}{\phi_t cv_t^2}\right)$$

이 되어서 추정량의 형태에 따른 상대표준오차는 다르지만 Park (1989)이 제시한 표본크기 공식에 두 시점의 응답률의 비가 곱해진 형태가 된다. 현 시점의 상대표준오차를 지난 시점보다 작게 잡는다면 표본크기는 과거보다 증가해야 하지만 목표 응답률을 지난 시점보다 크게 설정한다면 표본크기의 증가를 둔화시킬 수 있다.

Remark 2.4: $n_{t-1}\phi_{t-1}$을 과거 시점의 평균적 응답표본크기라 한다면 식 (2.6)은 모집단 및 오차의 변동을 반영하여 즉

$$\frac{\left(CV_{t-1}^{-2}cv'^2_{t-1} + N_{t-1}^{-1}\right)}{\left(CV_t^{-2}cv'^2_t + N_t^{-1}\right)}$$

의 식을 과거 시점의 평균적 응답표본크기에 곱하여 현 시점의 표본크기를 구하고 최종적으로 목표 응답률 ϕ_t을 나누어 줌으로써 현 시점의 표본크기를 결정하는 내용을 담고 있다.

Remark 2.5: 식 (2.2)에서 식 (2.5)까지의 표본크기 공식은 시점에 관계없는 공식으로 사용될 수 있다. 식 (2.2)와 (2.3)은 비추정량에서 고려할 수 있는 표본크기 공식인데 응답률(응답확률)이 다르다고 생각한다면 식 (2.2)를 사용할 수 있고 일정한 응답률을 고려한다면 식 (2.3)을 적용할 수 있다. 또한 식 (2.4)와 식 (2.5)는 모평균 추정량에서의 표본크기 공식으로 응답률과 추정량의 분산 또는 응답률과 상대표준오차를 조절하면서 크기를 결정하는 내용을 담고 있다.

3. 표본크기 공식의 비교

표본추출에 의한 설계가중치와 응답확률 보정계수가 반영된 모평균 추정량에 대해서 두 시점의 정보를 반영한 표본크기 공식의 의미를 살펴보기 위해 모의실험을 실시한다. 두 시점의 모집단의 크기 및 산포의 변동, 과거시점의 상대표준오차와 목표오차의 차이, 기존 표본의 응답률과 목표 응답률의 변화에 따라 본 논문에서 제안되는 식 (2.6)의 표본크기 의미를 살펴본다. 또한 기존의 표본크기 공식으로 Park (1989)이 제안한 식 (2.7)과 Park과 Na (2014)에 제시된 식 (2.8)을 사용한다. 본 논문에서 제안된 식 (2.6)의 표본크기 공식을 기존 표본크기 공식들과 여러 상황에서 비교 분석함으로써 제안된 공식의 유용성을 판단한다.

모의실험을 위하여 이전 시점인 $t-1$ 시점에서 상대표준오차를 $cv'_{t-1} = 0.04$로 설정하고 표본크기를 $n_{t-1} = 100$으로 가정하여 두 시점의 여러 변동의 비를 가지고 두 시점의 표본크기의 비인 n_t/n_{t-1}의 변

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

Res	CV_R	f_{t-1}	N_R								
			0.8			1.0			1.2		
			E1	E2	E3	E1	E2	E3	E1	E2	E3
0.8	0.7	0.01	0.412	0.515	0.490	0.418	0.523	0.490	0.423	0.529	0.490
		0.001	0.394	0.493	0.490	0.395	0.493	0.490	0.395	0.494	0.490
		0.0003	0.393	0.491	0.490	0.393	0.491	0.490	0.393	0.491	0.490
	1.0	0.01	0.776	0.970	1.000	0.800	1.000	1.000	0.817	1.021	1.000
		0.001	0.797	0.997	1.000	0.800	1.000	1.000	0.802	1.002	1.000
		0.0003	0.799	0.999	1.000	0.800	1.000	1.000	0.801	1.001	1.000
	1.3	0.01	1.189	1.486	1.690	1.246	1.558	1.690	1.287	1.609	1.690
		0.001	1.331	1.664	1.690	1.339	1.674	1.690	1.344	1.680	1.690
		0.0003	1.345	1.681	1.690	1.348	1.685	1.690	1.349	1.687	1.690
1.0	0.7	0.01	0.515	0.515	0.490	0.523	0.523	0.490	0.529	0.529	0.490
		0.001	0.493	0.493	0.490	0.493	0.493	0.490	0.494	0.494	0.490
		0.0003	0.491	0.491	0.490	0.491	0.491	0.490	0.491	0.491	0.490
	1.0	0.01	0.970	0.970	1.000	1.000	1.000	1.000	1.021	1.021	1.000
		0.001	0.997	0.997	1.000	1.000	1.000	1.000	1.002	1.002	1.000
		0.0003	0.999	0.999	1.000	1.000	1.000	1.000	1.001	1.001	1.000
	1.3	0.01	1.486	1.486	1.690	1.558	1.558	1.690	1.609	1.609	1.690
		0.001	1.664	1.664	1.690	1.674	1.674	1.690	1.680	1.680	1.690
		0.0003	1.681	1.681	1.690	1.685	1.685	1.690	1.687	1.687	1.690
1.2	0.7	0.01	0.618	0.515	0.490	0.627	0.523	0.490	0.634	0.529	0.490
		0.001	0.591	0.493	0.490	0.592	0.493	0.490	0.593	0.494	0.490
		0.0003	0.589	0.491	0.490	0.589	0.491	0.490	0.590	0.491	0.490
	1.0	0.01	1.164	0.970	1.000	1.200	1.000	1.000	1.225	1.021	1.000
		0.001	1.196	0.997	1.000	1.200	1.000	1.000	1.203	1.002	1.000
		0.0003	1.199	0.999	1.000	1.200	1.000	1.000	1.201	1.001	1.000
	1.3	0.01	1.783	1.486	1.690	1.869	1.558	1.690	1.931	1.609	1.690
		0.001	1.997	1.664	1.690	2.009	1.674	1.690	2.017	1.680	1.690
		0.0003	2.018	1.681	1.690	2.021	1.685	1.690	2.024	1.687	1.690

화량을 살펴본다. 첫째, 두 시점의 모집단의 변동계수의 비가 $CV_{t-1}/CV_t = 1$인 경우로 변동이 같은 상황과 $CV_{t-1}/CV_t = 0.9$인 경우로 t 시점에서 모집단의 산포가 큰 경우를 살펴본다. 둘째, 두 시점의 상대표준오차의 변동량 $CV_R = cv'_{t-1}/cv'_t$ 값이 0.7, 1, 1.3인 상황을 고려한다. 셋째 모집단 크기의 변화량 $N_R = N_t/N_{t-1}$은 0.8, 1, 1.2인 경우를 설정한다. 넷째, 두 시점의 응답률 ϕ_{t-1}, ϕ_t의 변화량으로 Res $= \phi_{t-1}/\phi_t = 1$인 경우와 응답률이 다른 상황인 Res 값이 0.8, 1.2인 경우를 살펴본다. Table 3.1과 Table 3.2에서 $t-1$ 시점의 추출률 $f_{t-1} = n_{t-1}/N_{t-1}$의 값은 0.01, 0.001, 0.0003을 사용하고 비교되는 표본크기 공식은 본 논문에서 제안된 식 (2.6)의 E1, Park과 Na (2014)의 논문에서 제시된 식 (2.8)의 E2, 그리고 Park (1989)에서 제안된 식 (2.7)의 E3이다.

Table 3.1과 Table 3.2의 내용을 살펴보겠다. 첫째, Res $= 1$이고 $CV_R = 1$이며 $N_R = 1$인 경우 Table 3.1의 경우는 $n_t/n_{t-1} = 1$이며 Table 3.2의 경우는 E1, E2에서는 표본크기의 변화량이 $n_t/n_{t-1} > 1$이고 E3는 Table 3.1과 같다. 그 이유는 Table 3.1은 두 시점의 모집단의 산포의 변동이 없으며 Table 3.2는 t 시점의 모집단의 변동이 늘어나기 때문에 n_t/n_{t-1}의 값에 그 변화가 반영되어야 하는데 E1과 E2는 변화량이 반영되었으며 E3는 반영되지 않음을 알 수 있다.

Table 3.2. Ratio of sample size n_t/n_{t-1} for $CV_{t-1}/CV_t = 0.9$

Res	CV_R	f_{t-1}	N_R								
			0.8			1.0			1.2		
			E1	E2	E3	E1	E2	E3	E1	E2	E3
0.8	0.7	0.01	0.496	0.620	0.490	0.504	0.630	0.490	0.510	0.637	0.490
		0.001	0.485	0.607	0.490	0.486	0.608	0.490	0.487	0.608	0.490
		0.0003	0.484	0.605	0.490	0.485	0.606	0.490	0.485	0.606	0.490
	1.0	0.01	0.936	1.170	1.000	0.965	1.206	1.000	0.985	1.231	1.000
		0.001	0.982	1.227	1.000	0.985	1.231	1.000	0.987	1.234	1.000
		0.0003	0.986	1.232	1.000	0.987	1.233	1.000	0.988	1.234	1.000
	1.3	0.01	1.433	1.792	1.690	1.502	1.878	1.690	1.552	1.940	1.690
		0.001	1.639	2.049	1.690	1.649	2.061	1.690	1.655	2.069	1.690
		0.0003	1.659	2.074	1.690	1.662	2.078	1.690	1.664	2.081	1.690
1.0	0.7	0.01	0.620	0.620	0.490	0.630	0.630	0.490	0.637	0.637	0.490
		0.001	0.607	0.607	0.490	0.608	0.608	0.490	0.608	0.608	0.490
		0.0003	0.605	0.605	0.490	0.606	0.606	0.490	0.606	0.606	0.490
	1.0	0.01	1.170	1.170	1.000	1.206	1.206	1.000	1.231	1.231	1.000
		0.001	1.227	1.227	1.000	1.231	1.231	1.000	1.234	1.234	1.000
		0.0003	1.232	1.232	1.000	1.233	1.233	1.000	1.234	1.234	1.000
	1.3	0.01	1.792	1.792	1.690	1.878	1.878	1.690	1.940	1.940	1.690
		0.001	2.049	2.049	1.690	2.061	2.061	1.690	2.069	2.069	1.690
		0.0003	2.074	2.074	1.690	2.078	2.078	1.690	2.081	2.081	1.690
1.2	0.7	0.01	0.744	0.620	0.490	0.756	0.630	0.490	0.765	0.637	0.490
		0.001	0.728	0.607	0.490	0.729	0.608	0.490	0.730	0.608	0.490
		0.0003	0.727	0.605	0.490	0.727	0.606	0.490	0.727	0.606	0.490
	1.0	0.01	1.404	1.170	1.000	1.447	1.206	1.000	1.477	1.231	1.000
		0.001	1.472	1.227	1.000	1.478	1.231	1.000	1.481	1.234	1.000
		0.0003	1.478	1.232	1.000	1.480	1.233	1.000	1.481	1.234	1.000
	1.3	0.01	2.150	1.792	1.690	2.253	1.878	1.690	2.328	1.940	1.690
		0.001	2.459	2.049	1.690	2.473	2.061	1.690	2.483	2.069	1.690
		0.0003	2.489	2.074	1.690	2.493	2.078	1.690	2.497	2.081	1.690

둘째, Table 3.1에서 각 칸을 살펴보면 추출률 f_{t-1}이 작아질수록 E2와 E3의 n_t/n_{t-1}은 거의 같은 값을 가지며 E1은 응답률의 비인 Res가 같은 경우에 그와 같은 현상이 일어나며 응답률의 비가 다른 경우에는 E2, E3의 n_t/n_{t-1}와 다른 값이 생성된다. 그것은 표본크기 식에서도 나타나듯이 모집단의 산포의 변동이 없는 경우 모집단의 크기가 커질 때 E2와 E3는 같게 되며 E1이 다르게 나온 것은 두 시점의 응답률의 변동이 반영된 결과라 할 수 있다.

셋째, Table 3.1과 Table 3.2를 살펴보면 응답률의 변동 Res의 각각의 값마다 CV_R의 값이 커질수록 E1, E2, E3에서 n_t/n_{t-1}의 값이 커지는 현상을 보인다. 이것은 t 시점의 목표오차가 $t-1$ 시점의 상대표준오차보다 작아질 때 CV_R의 값이 커지는 것이며 이와 같은 경우에는 t 시점의 표본크기가 커져야 하므로 표본크기의 비가 커지게 된다. E1, E2, E3의 표본크기 공식은 이와 같은 현상을 반영하는 것으로써 기본적으로 목표오차를 조절함으로써 표본크기를 결정하는 성질을 설명한다고 할 수 있다. Table 3.1과 Table 3.2의 차이점은 E1, E2의 표본크기에서 Res의 각각의 값마다 CV_R의 증가량에 따라 n_t/n_{t-1}의 값의 크기가 Table 3.1보다 Table 3.2에서 더 커지는 현상이 나타나는데 이것은 t 시점의 모집단의 변동계수가 이전 시점보다 커지기 때문에 더 많은 표본크기가 필요해지는 것이 반영된 것이라 할

수 있다. 한편 Table 3.1과 Table 3.2에서 E3의 값이 변동이 나타나지 않으며 이 표본크기는 모집단의 산포의 변동이 반영되지 않는 공식이기 때문이다.

넷째, Table 3.1과 Table 3.2에서 N_R의 변화량에 관계없이 n_t/n_{t-1}의 값이 일정하게 나오는 표본크기 공식은 E3이며 이 공식에는 모집단의 크기가 반영되어 있지 않기 때문이다. Table 3.1과 Table 3.2를 살펴보면 N_R이 커질 수록 E1과 E2에서 n_t/n_{t-1}의 값이 커지는 현상을 보이며 Table 3.1보다 Table 3.2의 n_t/n_{t-1}의 값이 큰 것은 모집단의 변동의 반영이라 할 수 있다. 하지만 추출률이 1%, 0.1%, 0.03%인 실험에서 N_R의 변화량에 따른 n_t/n_{t-1}의 값의 변화의 양은 CV_R의 변동에 따른 n_t/n_{t-1}의 값의 변화량보다 작게 나타나며 추출률이 커질수록 N_R의 변화량에 따른 n_t/n_{t-1}의 변화가 작아짐을 알 수 있다. 그것은 식 (2.6)을 살펴보면 모집단의 크기가 커짐으로써 표본크기 공식에 미치는 영향이 작아지기 때문이다.

다섯째, Table 3.1과 Table 3.2에서 응답률의 변화량인 Res의 증감을 살펴보면 E1은 값의 변동이 있으나 E2, E3는 n_t/n_{t-1}가 변화하지 않음을 발견할 수 있다. 그리고 Res = 1인 경우에는 E1의 값과 E2의 값이 일치하는 현상이 나타나며 이것은 본 논문에서 제안한 표본크기 공식은 응답률의 변동이 없으면 E2의 표본크기 공식과 일치하게 되는 것을 반영한 것이다. 그리고 Table 3.1과 Table 3.2에서 E1의 표본크기 공식은 Res가 커질수록 n_t/n_{t-1}의 값이 커지는 현상이 나타나며 응답률이 작아지면 목표 정확도를 위하여 표본크기를 크게해야 함을 의미한다.

여섯째, 대략적으로 Table 3.1과 Table 3.2에서 E1과 E2는 Res의 증감에 상관없이 CV_R이 감소하고 N_R이 증가할수록 f_{t-1}의 추출율이 작아지면 n_t/n_{t-1}가 감소하는 현상이 있으며 이와 반대로 CV_R이 증가하고 N_R이 감소할수록 n_t/n_{t-1}가 증가하는 현상을 나타내고 있다.

마지막으로 두 Table의 차이점은 모집단의 산포의 변동이 t 시점에서 커짐으로 전반적으로 표본크기의 비인 n_t/n_{t-1}가 Table 3.1보다 Table 3.2에서 증가한다는 것인데 $CV_{t-1}/CV_t > 1$의 경우는 t 시점의 모집단 산포변동이 작아진다는 것으로 표본크기의 비가 Table 3.1의 값보다 작아지게 될 것이다. 그리고 이것 이외의 다른 수치의 변동에 관한 내용은 Table 3.1과 Table 3.2의 차이점에서 설명한 것과 비슷한 현상을 보이게 된다. 그러므로 $CV_{t-1}/CV_t > 1$의 경우는 본 논문에서 살펴보지 않고 두 모집단의 산포의 변동이 없는 경우와 산포의 변동이 있는 경우만을 언급하였다.

또한 응답률의 변화에 따른 표본크기의 변동을 그림으로 표현한다. Figure 3.1과 Figure 3.2에서 시점간 표본크기의 비인 n_t/n_{t-1}을 n_r로 표시하며 공통적으로 모집단크기의 변동은 $N_R = 1$인 경우와 추출률이 $f_{t-1} = 0.001$인 경우를 분석한다. Figure 3.1과 Figure 3.2의 각각은 $CV_{t-1}/CV_t = 1$과 $CV_{t-1}/CV_t = 0.9$인 것으로 모집단의 산포의 변동이 없는 것과 있는 것을 나타낸다. 두 그림에서 Res의 값에 따라 CV_R의 증가에 따른 표본크기의 비 n_t/n_{t-1}의 값의 증가 여부를 보기 위해 산점도를 표본크기 공식 E1, E2, E3에 대해 분석한다. 모집단 크기의 변동에 따른 n_t/n_{t-1}의 변화량은 추정오차의 변동인 CV_R의 변화에 따른 n_t/n_{t-1}의 변화량보다 작기 때문에 살펴보지 않는다.

Figure 3.1과 Figure 3.2에서 E1, E2, E3에 공통적으로 나타나는 특징은 CV_R이 증가함에 따라 n_t/n_{t-1}도 증가하는 경향을 보인다. 이것은 목표오차를 이전 시점의 추정오차보다 작게 결정할수록 표본크기가 커져야 하는 것을 반영하고 있다.

Figure 3.1와 Figure 3.2의 차이점은 E1, E2의 n_t/n_{t-1}값이 Figure 3.2에서 더 큼을 알 수 있는데 그것은 모집단의 산포가 이전 시점보다 증가하기 때문에 표본을 더 많이 확보해야 함에서 나타나는 성질이다. 두 그림에서 E3는 CV_{t-1}/CV_t와 Res의 값에 상관없이 일정한 값을 가지는데 E3의 표본크기 공식은 이전 시점의 상대표준오차와 현 시점의 목표오차만을 반영하고 있기 때문이다. 응답률의 비인 Res에 따라 변동이 발생하는 것은 E1으로써 Res값이 커짐에 따라 n_t/n_{t-1}값이 커지는 현상이 있다. 이와 같

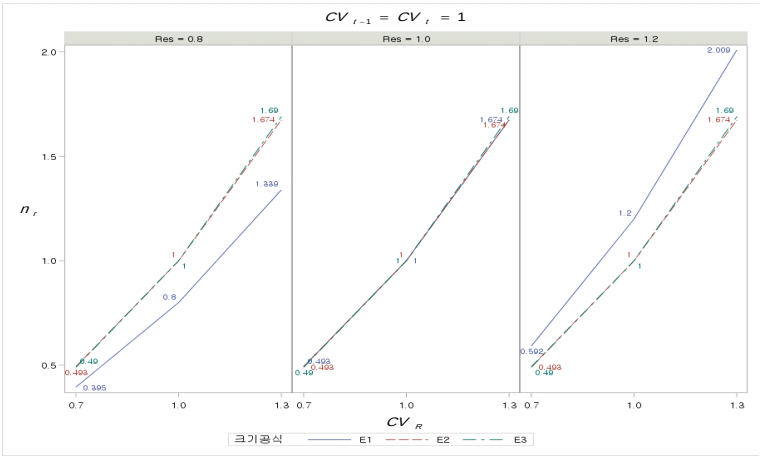

Figure 3.1. Scatter plot of CV_R and n_t/n_{t-1} under Res ($CV_{t-1}/CV_t = 1$).

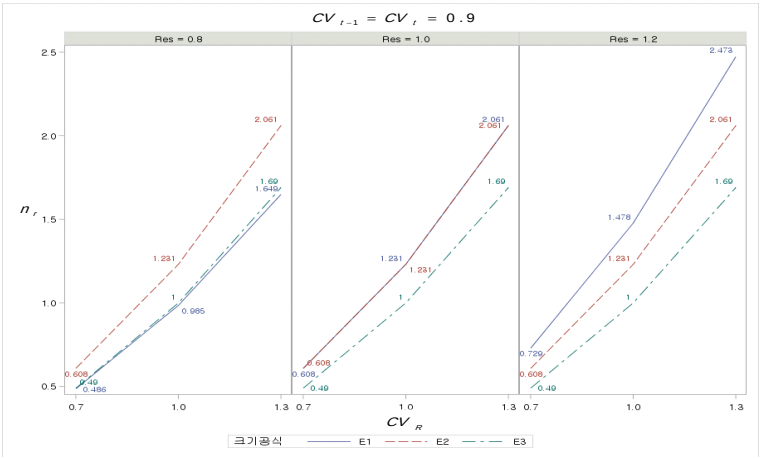

Figure 3.2. Scatter plot of CV_R and n_t/n_{t-1} under Res ($CV_{t-1}/CV_t = 0.9$).

은 현상은 Res값이 커진다는 것은 현 시점의 응답률을 작게 잡는 경우로써 현 시점의 표본크기를 많이 확보하는 것을 반영하고 있다. $N_R = 1$인 경우에 $CV_{t-1}/CV_t = 1$이면 E2와 E3가 거의 같게 나오지만 $CV_{t-1}/CV_t = 0.9$이면 모집단의 산포가 반영되어 E3보다 E2에서 n_t/n_{t-1}값이 더 크게 나온다. 이것은 현 시점의 모집단의 산포가 커지는 것을 반영하는 것이 E2의 표본크기 공식임을 알 수 있다. Figure 3.1에서는 Res = 1일 때 E1, E2, E3의 n_t/n_{t-1}값이 거의 같게 나오는데 그것은 시점별 모집단의 크기, 응답률 및 모집단의 산포가 각각 같으면 세 가지의 표본크기 공식이 거의 같게 된다는 것을 반영한다.

4. 결론

종합적으로 설명하면 CV_R의 변동에 n_t/n_{t-1}가 반응하는 표본크기 공식은 E1, E2, E3이며 추가적으로 N_R에 n_t/n_{t-1}가 움직이는 크기 공식은 E1과 E2이다. CV_R, N_R과 Res의 변동에 n_t/n_{t-1}값의 변동이

나타나는 공식은 E1이며 본 연구에서 제안한 표본크기 공식으로써 추정오차 및 모집단의 변동, 그리고 응답률의 변화의 내용도 포함할 수 있는 공식임을 실험적으로 증명할 수 있었다. 실제조사에서 응답여부는 표본크기를 많이 좌우할 수 있다. 엄밀한 크기 공식에 의해 응답확률(응답률)이 고려된 표본크기를 확보할 수 있다면 설계의 정밀성을 높일 수 있는 효과를 가져오게 된다. 그러므로 본 연구에서 기존의 표본크기 공식에 응답률(응답확률)이 사용된 것을 연구함으로써 여러 외부 효과가 반영되어야 하는 표본크기 결정을 수리적 접근의 연구로 확장하는 발판을 마련한 것이라 할 수 있다.

한 가지 더 생각해야 할 것은 모집단의 변동계수의 추정을 위해서 두 시점의 자료가 필요하다. 그러나 표본크기를 결정하는 시점에서는 현 시점의 자료가 존재하지 않으므로 모변동계수의 추정을 위한 방안이 마련되어야 한다. 첫 번째 방법은 관심변수와 상관이 있는 보조변수의 모집단 정보를 사용하여 두 시점의 모변동계수를 대체하는 것이다. 두 번째 방법은 보조변수에 의한 모변동계수의 시점별 비를 구해서 과거 시점의 관심변수의 표본변동계수에 적용함으로써 현재 시점의 모변동계수를 추정하는 것이다. 한편 모집단의 변동계수는 시간의 변동에 변화하지 않는다는 성질을 (Park, 1989; Sung, 2012) 반영하여 두 시점의 모변동계수를 같다고 가정할 수도 있다.

본 연구에서의 표본크기 공식은 단순임의추출하에 유도되는 것이었다면 향후 연구로는 복합표본설계에서의 표본크기 공식들에 응답확률(응답률)이 반영된 경우로 확장할 수 있으며 사후추정 등과 같이 여러 가지 보정 등이 사용된 추정량에서 표본크기 공식을 고려해 볼 수도 있다. 또한 응답확률의 추정 문제를 포함한 표본크기로 확장될 수도 있다.

References

Han, G. S. and Lee, G. S. (2015). A note on the decision of sample size by relative standard error in successive occasions, *The Korean Journal of Applied Statistics*, **28**, 477–483.

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability, *The Canadian Journal of Statistics*, **35**, 501–514.

Kim, K. S. (2012). Sample size determination in repeated surveys with varying population sizes, *Survey Research*, **13**, 159–174.

Lee, I. and Park, M. (2015). A study on sample allocation for stratified sampling, *The Korean Journal of Applied Statistics*, **28**, 1047–1061.

Park, H. and Na, S. (2014). Decision of sample size on successive occasions, *The Korean Journal of Applied Statistics*, **27**, 513–521.

Park, H. N. (1989). *Statistical Survey* (2nd ed), Youngji Publishers, Seoul.

Sung, N. (2012). *Methodology of Sample Survey*, Freedom Academy, Gyeonggi.

Yoo, Y. and Shin, K. I. (2011). A study on the decision of sample size for panel survey design, *The Korean Journal of Applied Statistics*, **24**, 25–34.

계속조사에서 응답률을 반영한 표본크기

박현아[a] · 나성룡[a,1]

[a]연세대학교 정보통계학과

(2018년 5월 31일 접수, 2018년 7월 16일 수정, 2018년 7월 16일 채택)

요 약

조사목적에 부합하는 표본 자료를 얻기 위해서는 추출방법 및 조사방법 결정, 설문지 작성 등의 절차가 필요하며 중요한 결정 중 하나가 표본크기 공식의 적용이다. 표본크기 공식은 추출방법에 따른 목표오차와 총비용 등을 설정함으로써 결정되는데 본 논문에서는 단순임의추출에서 목표오차와 예상 응답률이 주어져 있을 때 과거 및 현재 시점의 모집단의 변동과 과거 자료의 추정오차 및 응답률을 사용한 표본크기 공식을 제안한다. 실제조사에서는 설계가중치 외에도 여러 가중치가 복합적으로 적용되는 추정량을 사용하고 있는데 본 논문에서는 설계가중치와 무응답 보정계수를 사용한 추정량에서의 표본크기 공식을 유도하며 이것은 시점별 조사방법이 달라질 경우 응답률에 차이가 발생하는 현상을 반영한 공식이 될 수 있다. 또한 모의 실험을 통하여 기존의 표본크기 공식과 비교함으로써 제안된 공식의 다양한 적용방안을 살펴본다.

주요용어: 계속조사, 표본크기, 응답률, 모집단크기, 목표오차

[1]교신저자: (26493) 강원도 원주시 연세대길 1, 연세대학교 정보통계학과. E-mail: nasr@yonsei.ac.kr