Journal of the Korean Data & Information Science Society
2012, **23**(4), 677–682

http://dx.doi.org/10.7465/jkdi.2012.23.4.677
한국데이터정보과학회지

시뮬레이션을 이용한 임상자료의 샘플링 방법 연구[†]

손기철[1] · 김달호[2]

[1]대구가톨릭대학교 의과대학· [2]경북대학교 통계학과·

접수 2012년 5월 28일, 수정 2012년 6월 17일, 게재확정 2012년 6월 22일

요약

여러 분야에서 자료를 수집하기 위하여 모집단에서 표본을 추출하는 표본설계를 하고 있다. 특히 실험군과 대조군이 존재하는 임상자료에서는 모집단에서 집단별 일반적 변수들의 특성이 표본에 잘 반영되어야 하므로 더욱더 표본설계가 중요한 문제이다. 즉 모집단에서 집단별로 일반적 변수들이 가지는 빈도, 중심척도 그리고 산포척도 등이 표본에서도 동일하게 나타나야 한다. 그러나 주로 이루어지는 표본설계는 매우 복잡하고 어려워 일반 연구자가 사용하는데 있어 어려움을 겪는다. 따라서 본 논문에서는 시뮬레이션을 이용하여 집단별로 일반적 특성을 유지할 수 있는 임상자료의 샘플링 방법을 연구하였다. 또한 중환자실에 있는 환자자료에 적용하여 모집단과 표본의 일반적 변수의 특성값을 계산하여 보았고 통계적 가설검정을 이용하여 모집단과 표본집단에서 일반적 변수들의 값의 차이 여부를 비교하여 보았다.

주요용어: 샘플링, 시뮬레이션, 일반적 특성, 임상자료, 표본설계.

1. 서론

많은 임상자료를 수집하기 위하여 모집단의 모든 대상을 조사하기 보다는 표본을 정하고 이를 조사하여 연구에 사용한다. 즉 표본설계를 통한 샘플링을 한다. 이러한 표본설계는 모집단의 기본적 특성들이 가지는 정보를 표본에서도 동일하게 구성하도록 하는 것이 주요 원칙이다. 예를 들어, 기본적 특성에 해당하는 변수인 성별, 나이, BMI (Body Mass Index) 등의 변수 값은 모집단에서의 비율이나 평균값이 표본에서도 동일하여야 한다. 샘플링을 하는 방법 중 하나인 단순임의추출법으로는 모집단과 표본이 동질의 기본적 특성을 가지도록 하는 것이 거의 불가능하다. 또한 표본설계의 방법에는 여러 가지가 존재하고 이런 방법들이 복잡하여 표본설계의 전문가가 아닌 이상 사용하기가 어렵다. 물론 모집단에 해당하는 대상을 전수조사하면 더할 나위 없이 이상적이지만 임상자료에서는 특히 비용과 시간의 문제를 고려한다면 샘플링이 필요할 수밖에 없다. 따라서 본 논문에서는 전형적인 표본설계 방법이 아닌 시뮬레이션을 이용한 샘플링 방법을 연구하였다.

Cochran (1977), Kish (1965), Bellhouse (1988) 등은 표본설계이론을 전문적이고 자세하게 다루고 있는 문헌들이다. 표본추출을 하는 방법은 대표적으로 단순임의추출, 집락추출, 군집추출, 계통추출 등이 있다. 그 외에도 편향을 적게 하고 모집단을 잘 반영하는 표본을 얻기 위한 연구가 현재도 계속해서 진행되고 있다. Kim 등 (2010)은 중소기업실태조사를 위한 표본설계를 연구하였다. 이때 변동계수

[†] 본 연구는 보건복지부 권역별 전문질환설치·지원-류마티스·퇴행성관절염센터 (과제고유번호 : 090-091-2700-2744-300)의 지원에 의하여 이루어진 것임.

[1] (705-718) 대구광역시 남구 대명4동 3056-6, 대구가톨릭대학교 의과대학, 전임강사.
[2] 교신저자: (702-701) 대구광역시 북구 산격동 1370, 경북대학교 자연과학대학 통계학과, 교수.
E-mail: dalkim@knu.ac.kr

(CV)를 각 층 (stratum)별로 설정함으로써 표본을 구성하였다. 또한 모집단의 기본적 특성을 잘 반영하지만 기존의 표본 보다 표본수를 줄여 표본조사의 업무량을 줄이기 위한 연구도 이루어지고 있다. Kim 등 (1994)은 농업 기본통계 및 가축통계 조사를 위한 표본설계를 연구하였다. 이때 전수조사가구 및 유의표본의 기준을 시도별로 조정하여 효율은 높이고 표본수를 줄여 설계의 효과를 높였다. Heo와 Chang (2010)은 경남지역교육청 수요자 만족도조사를 위한 표본설계에 관한 연구를 하였다. 이를 위해 지역별 평가에 필요한 최소표본을 우선배분한 후, 나머지는 지역별 학급수에 비례하여 배분하고, 표본학교는 지역과 학교설립유형별로 층화하여 비례추출하였다. 또한 표본학교 내에서 조사대상 학생은 2단 집락추출하였다. 이와 같이 여러 분야에서 표본설계에 관한 연구가 이루어지고 있다. 한편 Choi와 Kim (2010)은 엑셀의 매크로 기능을 이용하여 표본추출 과정과 방법, 모수와 통계량을 비교하는 프로그램을 연구하였다. 이를 통하여 빠른 시간 내에 모집단에서 일부분의 표본을 추출하고 모수와 통계량의 차이를 비교하고 분석할 수 있게 하였다. 본 논문은 다음과 같이 구성하였다. 2절에서 시뮬레이션을 이용한 샘플링 방법을 소개한다. 그리고 3절에서는 임상자료에 2절에서 소개한 방법을 적용하여 보고 4절에서 논의를 기술하고자 한다.

2. 샘플링 방법

모집단 자료에서 일반적 변수들의 특성을 유지하면서 샘플링을 하여 표본자료를 구성하는 방법은 다음과 같다. 우선 모집단 자료가 하나의 그룹변수 (G_p)와 T개의 변수들로 구성되고 데이터의 수는 N개로 가정하자. 또한 변수들 중 K개는 일반적 변수들 $(X_{p1}, \cdots, X_{pK})$ 그리고 그 외의 $T-K$개의 나머지 변수들 $(X_{pK+1}, \cdots, X_{pT})$로 구성되고 그룹은 편의를 위해 두 집단으로 가정한다면 각 그룹에 해당하는 데이터의 수는 각 N_1과 N_2이다. 여기서 $N_1 + N_2 = N$이다. 모집단의 특성을 파악하기 위해 모집단 자료에서 그룹에 대해서 일반적 변수들의 특성값을 계산한다. 예를 들어, X_{pk}, $k = 1, \cdots, K$ 변수가 양적변수라면 (2.1) 식과 같이 그룹1에 대한 평균 $M_1(X_{pk})$과 그룹2에 대한 평균 $M_2(X_{pk})$을 계산할 수 있다.

$$M_1(X_{pk}) = (1/N_1)\sum_{i=1}^{N_1} X_{pki}, M_2(X_{pk}) = (1/N_2)\sum_{i=1}^{N_2} X_{pki}, k = 1, \cdots, K. \tag{2.1}$$

또한, X_{pk}변수가 이분형 질적 변수이고 0과 1로 입력되어 있다면 (2.2)식과 같이 그룹1에 대한 퍼센트 값 $P_1(X_{pk})$과 그룹2에 대한 퍼센트 값 $P_2(X_{pk})$을 계산할 수 있다.

$$P_1(X_{pk}) = (1/N_1)\sum_{i=1}^{N_1} X_{pki}, P_2(X_{pk}) = (1/N_2)\sum_{i=1}^{N_2} X_{pki}, k = 1, \cdots, K. \tag{2.2}$$

만약 질적 변수의 범주가 3개 이상인 경우는 각 그룹에 대한 퍼센트 값을 각각의 범주에 대하여 계산하여야 한다. 예를 들어, X_{pk}변수가 질적 변수이고 0, 1, 2로 입력되어 있다면 X_{pk}의 값이 0이면 1, 그 외는 0의 값을 가지는 X_{1pk}변수와 X_{pk}의 값이 1이면 1, 그 외는 0의 값을 가지는 X_{2pk}변수 그리고 X_{pk}의 값이 2이면 1, 그 외는 0의 값을 가지는 X_{3pk}변수를 생성하여 그룹1에 대한 퍼센트 값 $P_1(X_{1pk})$, $P_1(X_{2pk})$, $P_1(X_{3pk})$와 그룹2에 대한 퍼센트 값 $P_2(X_{1pk})$, $P_2(X_{2pk})$, $P_2(X_{3pk})$를 (2.2)식과 유사한 방법으로 계산할 수 있다. 따라서 질적 변수의 범주가 3개 이상인 경우는 그룹1에 대한 퍼센트 값으로 $P_1(X_{pk})$대신에 $P_1(X_{1pk})$, $P_1(X_{2pk})$, $P_1(X_{3pk})$를 고려하여야 한다. 그러나 본 논문에서는 이분형 질적 변수에 대한 언급만 하기로 한다.

다음으로 모집단 자료에서 샘플링을 하여 표본을 구성한다. 여기서 표본 데이터의 수는 n개로 가정하자. 샘플링은 각 데이터에 0과 1사이의 균일분포에서 난수를 발생하여 난수의 값이 n/N보다 작은 값에 해당하는 데이터로만 표본을 구성한다. 이로 구성된 표본자료는 모집단 자료와 동일하게 하나의 그룹변수 (G_s), K개의 일반적 변수들 ($X_{1s}, \cdots, X_{sK}$) 그리고 나머지 변수들 ($X_{sK+1}, \cdots, X_{sT}$)이다. 표본 데이터의 수는 각 그룹별로 n_1과 n_2개를 추출하여 전체 표본 데이터의 수는 $n(= n_1 + n_2)$개 이다. 모집단과 동일한 관점에서 표본의 특성을 파악하기 위해 표본 자료에서 그룹에 대해서 일반적 변수들의 특성 값을 계산한다. 예를 들어, X_{sk}, $k = 1, \cdots, K$변수가 양적변수라면 (2.3)식과 같이그룹1에 대한 평균 $M_1(X_{sk})$과 그룹2에 대한 평균 $M_2(X_{sk})$을 계산할 수 있다.

$$M_1(X_{sk}) = (1/n_1)\sum_{i=1}^{n_1} X_{ski}, M_2(X_{sk}) = (1/n_2)\sum_{i=1}^{n_2} X_{ski}, k = 1, \cdots, K. \tag{2.3}$$

X_{sk}변수가 질적 변수이고 0과 1로 입력되어 있다면 (2.4)식과 같이 그룹1에 대한 퍼센트 값 $P_1(X_{sk})$과 그룹2에 대한 퍼센트 값 $P_2(X_{sk})$을 계산할 수 있다.

$$P_1(X_{sk}) = (1/n_1)\sum_{i=1}^{n_1} X_{ski}, P_2(X_{sk}) = (1/n_2)\sum_{i=1}^{n_2} X_{ski}, k = 1, \cdots, K. \tag{2.4}$$

다음으로 그룹별로 모집단과 표본에 해당하는 일반적 변수의 특성 값의 차이를 계산한다. 다시 말해서, k번째 일반적 변수가 양적변수라면 $M_1(X_{pk}) - M_1(X_{sk})$과 $M_2(X_{pk}) - M_2(X_{sk})$를 계산하고, 질적 변수라면 $P_1(X_{pk}) - P_1(X_{sk})$과 $P_2(X_{pk}) - P_2(X_{sk})$를 계산한다. 그런 후, (2.5)식과 같이 차이 값의 절댓값을 모두 더한 D를 계산한다.

$$\begin{aligned} D = \sum_{k=1}^{K} (&|M_1(X_{pk}) - M_1(X_{sk})| + |M_2(X_{pk}) - M_2(X_{sk})| \\ &+ |P_1(X_{pk}) - P_1(X_{sk})| + |P_2(X_{pk}) - P_2(X_{sk})|) \end{aligned} \tag{2.5}$$

마지막으로 위에서 설명한 모집단에서 표본을 구성하여 D값을 계산하는 일련의 과정을 H번 반복하여 H개의 $D^{(h)}$, $h = 1, \cdots, H$를 계산한다. 그리고 H개의 $D^{(h)}$값 중 가장 작은 값에 해당하는 표본을 최종 표본으로 결정한다.

3. 임상자료에의 적용

사용한 자료는 카네기 멜론 대학교 (Carnegie Mellon University)에서 제공하는 자료 및 관련 이야기 도서관 (http://lib.stat.cmu.edu/DASL/DataArchive.html)에 공개된 자료이며, 중환자실 (Intensive care unit; ICU)에 입원한 성인 환자에 관한 임상자료로써 200명으로 구성되어 있다. 자료의 변수는 환자고유번호 (ID), 환자상태 (vital status; 0=살아있음, 1=죽음), 나이 (age), 성별 (sex; 1=남성, 2=여성), 인종 (race; 1=백인, 2=흑인, 3=그외), 진료과구분 (service; 0=내과, 1=외과), 암구분 (cancer; 0=없음, 1=있음), 만성병구분 (chronic disease; 0=없음, 1=있음) 등으로 구성되어 있으며 Table 3.1에 일부를 제시하였다. 이 자료에서 환자상태는 그룹변수로 이고 나이, 성별 그리고 인종은 일반적 특성 변수이다. 이 자료를 모집단 자료로 간주하고 여기에서 총 50명의 표본을 추출하되, 2절에서 제시한 방법을 적용하여 집단별로 일반적 특성을 모집단과 유사하게 유지하면서 표본이 추출되는 과정을 예시하고자 한다.

Table 3.1 Patients data in adult intensive care unit (ICU)

No.	ID	Vital status	Age	Sex	Race	Service	Cancer	Chronic disease	···
1	4	1	87	2	1	2	1	1	
2	8	0	27	2	1	1	1	1	
3	12	0	59	1	1	1	1	1	
4	14	0	77	1	1	2	1	1	
5	27	1	76	2	1	2	1	1	
					⋮				
196	921	1	50	2	2	1	1	1	
197	923	0	20	1	1	2	1	1	
198	924	0	73	2	2	1	1	2	
199	925	0	59	1	1	1	1	1	
200	929	0	42	1	1	2	1	1	

중환자실의 성인 환자 자료 전체에서 집단 (환자상태)별 일반적 변수 (나이, 성별, 인종)에 대해 정리한 모집단 특성값은 Table 3.2와 같이 나타났다. 환자상태 변수에서 '살아있음 (alive)' 집단에 해당하는 환자 수는 160명이고 '죽음 (dead)' 집단에 해당하는 환자 수는 40명으로 4:1의 비율을 나타내고 있다. 나이의 값은 '살아있음' 집단과 '죽음' 집단에서 각각 55.650과 65.125의 평균값을 나타내고 있다. 또한 남 (male), 여 (female) 성별의 비율은 '살아있음' 집단에서 각 62.5%와 37.5%를 '죽음' 집단에서 각 60.0%와 40.0%를 나타내고 있다. 마지막으로 백인 (white), 흑인 (black) 그리고 그 외 (others) 인종의 비율은 '살아있음' 집단에서 각 86.3%, 8.8%, 5.0%를 '죽음' 집단에서 각 92.5%, 2.5%, 5.0%를 나타내고 있다. 표본자료의 크기는 50명으로 정하였고 모집단에서 집단별 비율을 유지하기 위하여 '살아있음' 집단에서 40명의 환자와 '죽음' 집단에서 10명의 환자로 표본자료를 구성하였다.

Table 3.2 Population characteristics on age, sex and race for patients

Variable			Vital status	
			Alive ($N_1 = 160$)	Dead ($N_2 = 40$)
Quantitative mean (SD)	Age		55.650 (20.428)	65.125 (16.649)
Qualitativecount (%)	Sex	Male	100 (62.5)	24 (60.0)
		Female	60 (37.5)	16 (40.0)
	Race	White	138 (86.3)	37 (92.5)
		Black	14 (8.8)	1 (2.5)
		Others	8 (5.0)	2 (5.0)

2절에서 소개한 방법으로 10,000번 (D) 의 반복을 시행하였고 그 중 D값이 가장 작은 자료로 표본자료를 구성하였다. 표본자료를 이용하여 집단 (환자상태)별 일반적 특성변수 (나이, 성별, 인종)에 대해 정리한 통계량은 Table 3.3과 같이 나타났다. 그리고 각각의 집단에서 전체자료와 표본자료를 비교하기 위하여 양적변수 (나이)에 대해서는 독립 2표본 t-test와 질적 변수 (성별, 인종)에 대해서는 χ^2-test를 실시하여 나온 유의확률 값 (p-value)도 함께 Table 3.3에 제시하였다. 결과는 각 집단별로 모든 일반적 특성변수에 대하여 전체자료와 표본자료의 차이가 나지 않는 것으로 나타났다. 즉 모집단의 일반적 특성을 시뮬레이션을 통하여 구성한 표본 집단에서도 잘 반영하고 있는 것으로 나타났다.

Table 3.3 Parameter values, statistics, and test results of difference on age, sex and race for patients

Variable			Vital status					
			Alive			Dead		
			Population ($N_1 = 160$)	Sample ($n_1 = 40$)	P-value P-value	Population ($N_2 = 40$)	Sample ($n_2 = 10$)	P-value P-value
Quantitativemean (SD)	Age		55.650 (20.428)	55.201 (22.120)	0.450	65.125 (16.649)	60.100 (17.278)	0.401
Qualitative count (%)	Sex	Male	100 (62.5)	22 (55.0)	0.384	24 (60.0)	5 (50.0)	0.567
		Female	60 (37.5)	18 (45.0)		16 (40.0)	5 (50.0)	
	Race	White	138 (86.3)	34 (85.0)	0.630	37 (92.5)	8 (80.0)	0.450
		Black	14 (8.7)	5 (12.5)		1 (2.5)	1 (10.0)	
		Others	8 (5.0)	1 (2.5)		2 (5.0)	1 (10.0)	

4. 논의

본 논문에서는 시뮬레이션을 이용하여 집단별로 일반적 특성을 유지하면서 샘플링을 하는 방법을 연구하였다. 이를 위하여, 우선 모집단에서 난수를 발생하여 두 집단의 비율과 동일하게 표본을 구성하였다. 다음으로 모집단과 표본에서 집단별 일반적 변수들의 특성값을 계산하였다. 이 과정을 여러 번 반복하여 각 경우에서 모집단과 표본의 일반적 변수의 차이값에 대한 합을 계산하였고 이 값이 가장 작은 표본을 10,000개의 후보표본 중 최종표본으로 설정하였다. 이 방법을 중환자실의 환자 자료에 적용하여 200명의 모집단자료에서 50명의 표본자료를 추출하였고 각 집단에서 일반적 변수의 특성값을 계산하여 모집단자료의 특성을 표본이 잘 반영하는지를 비교하여 보았다. 결과는 모집단의 특성을 표본이 잘 반영하는 것으로 나타났다. 그렇지만 이 논문에서 제시한 방법은 반복수를 작게 한다면 표본이 모집단을 잘 반영하지 못할 수 있으므로 여러 번 반복하여 계산하는 것을 추천한다.

참고문헌

Bellhouse, D. R. (1988). Systematic sampling. In *Handbook of Statistics 6: Sampling*, edited by Krishnaiah, P. R. and Rao, C. R., North-Holland, Amsterdam, 125-145.

Choi H. S. and Kim T. Y. (2010). A study on sampling using the function of excel. *Journal of the Korean Data & Information Science Society*, **21**, 481-491.

Cochran, W. G. (1977). *Sampling techniques,* 3rd ed., Wiley, New York.

Heo S. Y. and Chang D. J. (2010). A sample survey design for service satisfaction evaluation of regional education offices. *Journal of the Korean Data & Information Science Society*, **21**, 669-679.

Kim K. S., Jeon J. W. and Park H. N. (1994). Sample designs of the farm population survey and the livestock survey. *The Korean Journal of Applied Statistics*, **1**, 1012-1020.

Kim D. H., Hwang J. S. and Kwak S. G. (2010). A sample design for the survey on actual state of SMEs. *Journal of the Korean Data & Information Science Society*, **21**, 1021-1029.

Kish, L. (1965). *Survey sampling*, Wiley, New York.

Journal of the Korean Data & Information Science Society
2012, **23**(4), 677–682

http://dx.doi.org/10.7465/jkdi.2012.23.4.677
한국데이터정보과학회지

Study for the sampling method using simulation in clinical data[†]

Ki Cheul Sohn[1] · Dal Ho Kim[2]

[1]School of Medicine, Catholic University of Daegu
[2]Department of Statistics, Kyungpook National University

Received 28 May 2012, revised 17 June 2012, accepted 22 June 2012

Abstract

There are lots of sampling design which is determined for sample survey in various fields. Especially, it is important problem for clinical data because basic characteristic variables by group which consist of experiment group and control group in population should be reflect to sample. Therefore, frequencies, center scales and dispersion scales of variables by group in population should be similar in sample. But usual sampling design is very complicate so it is difficult to use in practice for researcher. In this paper, we consider the sampling method using simulation. We applied the proposed method to colon cancer data from a hospital. We compare basic characteristic variables between population and sample with mean, frequency and statistic hypothesis test.

Keywords: Basic characteristic, clinical data, sampling, sampling-design, simulation.

[†] This study was supported by the grant of Korea Ministry of Health & Welfare, Republic of Korea (Project No : 090-091-2700-2744-300).
[1] Full time instructor, School of Medicine, Catholic University of Daegu, Daegu 705-718, Korea.
[2] Corresponding author: Professor, Department of Statistics, College of Natural Sciences, Kyungpook National University, Daegu 702-701, Korea. E-mail: dalkim@knu.ac.kr