응용통계연구 제 9권 2호
1996년 9월 pp. 161-169

왜도(Skewness)가 심한 모집단에서의 절사법효과에 관한 연구

한 근 식[1] , 김 용 철[2]

요 약

표본크기 결정은 표본설계시 중요한 부분이며 정도(Precision)를 높이면서 비용, 시간등을 고려하여 최적화(Optimal)된 표본의 크기를 구하려 할 때 모집단분포가 심한 왜도 (highly skewed)를 보이거나 소수의 모집단요소들이 모집단총계의 대부분을 차지하는 경우가 있다. 이에 대해 Neyman의 최적할당법과 절사법(cut-off method)응용방법의 효율성을 사례를 이용하여 비교하였다.

1. 서 론

여러 분야의 표본설계를 계획할 때 문제는 표본의 크기를 결정하는 것이라 할 수 있다. 기대되는 답은 항상 간단한 것 만은 아니다. 모집단으로부터 단순 임의 표본을 추출할 경우 흔히 표본평균 $\bar{y}$가 정규분포를 따른다는 가정 하에 상대오차 r을 조정하여 표본의 크기를 다음과 같이 결정한다.

$$n=\frac{\left(\frac{tS}{r\bar{Y}}\right)^2}{1+\frac{1}{N}\left(\frac{tS}{r\bar{Y}}\right)^2} \qquad (1)$$

그러나 실제 표본조사시 지역통계산출, 조사비용 등을 고려하여 단순 임의 추출보다는 층화추출을 하게 되는데 이때 널리 활용되는 표본크기의 결정법이 Neyman의 최적할당법이다. 그러나 모집단의 분포가 심한 왜도를 보이거나 소수의 모집단요소들이 모집단총계의 대부분을 차지하는 경우 Neyman의 최적할당법은 지나치게 큰 표본크기를 할당하여 많은 조사비용이 요구될 뿐아니라 비표본오차(non-sampling error)가 크게 되어 정도를 감소시킬 수 있다. 이와 같은 현상은 사업체조사(business survey)에서 특히 심한데 이러한 모집단에 대한 표본크기의 할당은 모집단을 전수층(take-all stratum)과 표본층(take-some stratum)으로 구분하여 주어진 정도와 신뢰계수를 만족하는 각 층의 크기를 결정하는 방법인 절사법을 응용함으로써 Neyman의 최적할당보다 적은 표본크기를 산출할 수 있게 되는데 이는 Deming(1960)에 의해 처음으로 제안되었다.

---

1) (447-791) 경기도 오산시 양산동 411, 한신대학교 전산통계학과 조교수.
* 이 연구는 한신대학교 학술연구비에 의해서 지원되었음.
2) (449-714) 경기도 용인시 용인대학교 전산통계학과 전임강사.

본 연구에서는 층의 수가 2개가 적절하다고 생각되는 Cochran(94쪽)의 1920년도 미국 주요 도시 인구자료와 통계청에서 시행한 사업체 조사자료를 이용하여 층화추출을 할때 각 층의 표본배분에 관해 Neyman 의 최적할당법과 절사법의 응용을 통한 표본크기의 결정과 그 효율성을 알아보고 수집된 자료의 순위와 크기와의 그림을 통해 어떤방법을 선택할 것인가를 사례를 통해 제시하였다. 제 2장에서는 Neyman의 최적할당법에 의한 표본크기 할당법의 문제점을 제시하고 절사법의 응용을 통한 표본크기를 결정하며 제 3 장에서는 두방법의 효율성을 사례를 통해 비교하였다.

2. 최적할당법과 절사법

본문에 들어가기 전에 본 연구에서 이용된 기호(Notation)를 설명하면 다음과 같다.

W_1 : 전수층(take-all-stratum)의 비율

W_2 : 표본층(take-some-stratum)의 비율

$\overline{Y}$: 모평균

$\overline{Y_1}$: 전수층의 모평균

$\overline{Y_2}$: 표본층의 모평균

$\overline{y}$: 표본평균

$\overline{y_1}$: 전수층의 표본평균

$\overline{y_2}$: 표본층의 표본평균

N_h : h번째 부모집단의 크기

n_h : h번째층의 표본크기

n : 표본크기

2.1 Neyman의 최적할당법

흔히 사용하는 Neyman할당의 경우 각층의 표본크기의 식(Cochran, Sukhatme)은 다음과 같다.

$$n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} * n \qquad (2)$$

실제 표본조사시 층의 경계를 구하기 위해 Dalenius & Hodges(1959)의 Cum $\sqrt{freq}$ 방법이 많이 사용되고 있으며 식(2)에 의해 표본크기를 구하면 $n_1 \geq N_1$인 경우가 흔히 발생하게 된다.

실례로 통계청에서 실시하고 있는 사업체 조사자료를 보면 모집단내의 극히 일부 요소들이 모집단총계의 70% 내지 90%를 차지하는 경우가 흔하다. 이러한 자료는 왜도가 심하며 크기와 순위와의 관계를 나타내는 그림을 살펴보면 X축과 Y축으로 긴 꼬리를 형성함을 다음절에서 설명되는 각 사례를 통해 볼 수 있다. 이러한 모집단에 대해 식(2)의 Neyman할당법을 적용하면 첫번째 층의 표본의 크기 n_1이 그 층의 부모집단크기 N_1 보다 크게 할당되는 것($n_1 > N_1$)을 볼 수 있는데 이와같은 경우 $n_1 = N_1$으로 첫번째 층에서 표본을 추출한 후 나머지 층에서 $n_2 = n - N_1$개의 요소를 추출하도록 권하고 있다(Cochran 1977). 이는 2개의 층만을 고려할 경우 어느 한 층에 대해서는 전수조사를 실시하고 나머지 층에 대해서는 표본조사를 실시하는 것을 의미하는데 이러한 경우 다음절에서 설명하는 절사법과 그 방법이 유사하다고 하겠다.

2.2 절사법

모집단을 전수층(take-all stratum)과 표본층(take-some-stratum)으로 구분할 때 주어진 조건들(d, t)을 만족하는 전수층과 표본층의 크기를 결정하는 방법으로 다음과 같은 절차를 따른다.

주어진 신뢰계수, t, 정도, d, 그리고 $N_1 = k$ 의 값을 고정시킨 후 주어진 자료를 이용하여 최적의 n_2를 결정한다. 그러면 최적의 표본크기 n은 $N_1 + n_2$이다. 이와같은 방법으로 k의 값이 1부터 N 까지 반복한 후에 n의 값이 최소가 되는 $N_1 = k^0$을 구할 수 있다. 이때 k^0이 전수층의 크기가 되고 n_2는 표본층의 크기가 된다.

2개의 층만을 갖는 층화추출에서의 분산은 다음과 같은 식에 의해서 구해진다.

$$V(\overline{y_{st}}) = \frac{N_1 - n_1}{N_1} \frac{W_1^2 \, S_1^2}{n_1} + \frac{N_2 - n_2}{N_2} \frac{W_2^2 \, S_2^2}{n_2} \tag{3}$$

절사법에서는 모집단의 각 단위들을 큰 단위에서 작은 단위로 나열하였을 때 첫번째 층을 전수층으로 나머지 층을 표본층으로 정하기 때문에 전수층의 표본의 크기는 $n_1 = N_1$ 이므로 식(3)의 우측 첫항은 항상 0 이된다. 한편, 사전에 주어진 t값과 정도, $d = t\sqrt{V(\overline{y})}$를 식(3)에 대입후 n_2에 대해 정리하면 크기가 N_2인 표본층에서의 표본크기 결정식은 다음과 같다.

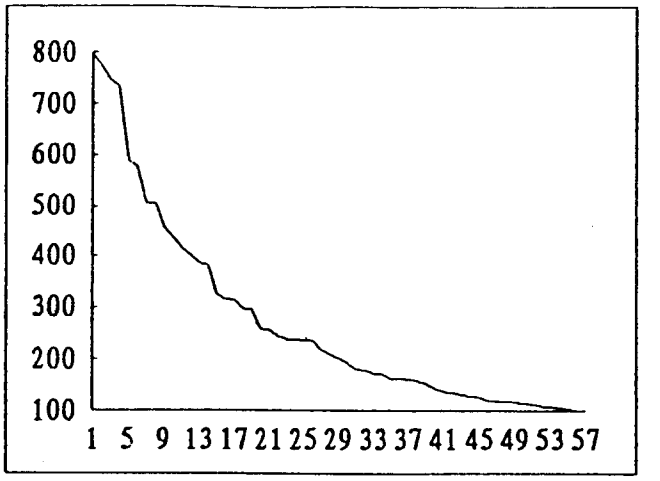
$$n_2 = \frac{\dfrac{W_2^2 t^2 CV^2(\overline{y_2})}{k^2 R^2}}{1 + \dfrac{1}{N_2} \dfrac{W_2^2 t^2 CV^2(\overline{y_2})}{k^2 R^2}}$$

여기서 $k=\frac{d}{y}$ 이고 $R=\frac{\bar{y}}{y_2}$ 이다.

3. 사례비교

위 두방법을 비교하기 위해서 Cochran의 미국 주요도시의 인구자료를 이용하였으며 표본크기 n은 절사법에 의해 산출된 값을 이용하였다.

797 314 172 121
773 298 172 120
748 296 163 119
734 258 162 118
588 256 161 118
577 243 159 116
507 238 153 116
507 237 144 113
457 235 138 113
438 235 138 110
415 216 138 110
401 208 138 108
387 201 136 106
381 192 132 104
324 180 130 101
315 179 126 100



<그림 1> 크기순으로 나열한 주요 도시 인구 자료와 크기와 순위와의 관계를 나타낸 그림

Neyman할당법을 적용하기 위해 층의 경계는 Dalenius & Hodges(1959)의 Cum $\sqrt{freq}$ 방법을 이용하고 표본평균과 그 분산은 다음과 같은 층화추출의 식을 이용하였다.

$$\overline{y_{st}}=\sum_{h=1}^{2} W_h \overline{y_h}$$

$$V(\overline{y_{st}})=\sum_{h=1}^{2} W_h^2 \frac{S_h^2}{n_h}(1-f_h)$$

여기서 $f_h=\frac{n_h}{N_h}$ 이다.

Cochran의 자료에서 t=1.96과 1.645 그리고 d=0.05의 조건으로 표본크기를 구하면 $n_1 > N_1$ 이 되어 첫번째 층에서 $1-f_h$항이 0이 되고 t=1.00일때 $N_1 > n_1$임을 알 수 있다. t와 d=0.05의 값에 따른 각 층의 $\overline{Y_1}$, $\overline{y_2}$, $\bar{y}$의 평균과 $\overline{y_1}$, $\overline{y_2}$, $\bar{y}$의 표준오차가 표본의 크기와 함께 <표 1>에 정리되었다. 절사법 사용시 표본층에서의 표본은 계통추출에 의해 추출되었으며 모평균의 추정치는 $\bar{y} = \overline{Y_1} + \widehat{Y_2} = \overline{Y_1} + \overline{y_2}$에 의해 그리고 분산은 $\frac{N_2 - n_2}{N_2} \frac{W_2^2 \, S_2^2}{n_2}$에 의해 구해졌다. <표 2>에서는 절사법에 의한 t와 d=0.05의 각 조건에 따른 $\overline{Y_1}$, $\overline{y_2}$, $\bar{y}$의 평균과 $\overline{y_1}$, $\overline{y_2}$, $\bar{y}$의 표준오차가 표본크기와 함께 정리되었다.

<표 1> 미국 주요도시 인구의 Neyman 할당법에 의한 표본 할당량과 그 추정량

상대오차=0.05

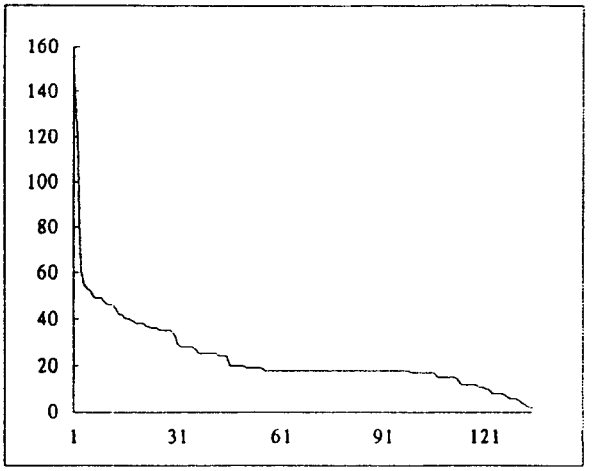
t 값	모집단의 크기		표본의 크기			층별 평균 추정량			층별 표준오차		
	N_1	N_2	n_1	n_2	n	$\bar{y}_1 = \overline{Y_1}$	$\bar{y}_2$	$\bar{y}$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}$
1.960	19	45	21	15	34	487.21	156.42	254.62	0.0	47.585	7.0536
1.645	19	45	20	13	32	487.21	156.39	254.60	0.0	47.415	7.7973
1.000	19	45	16	10	26	489.28	156.22	254.48	38.784	47.292	9.2735

<표 2> 미국 주요 도시 인구의 절사법에 의한 표본할당과 그 추정량

상대오차=0.05

t 값	모집단의 크기		표본의 크기			층별 평균 추정량			층별 표준오차		
	N_1	N_2	n_1	n_2	n	$\bar{y}_1 = \overline{Y_1}$	$\bar{y}_2$	$\bar{y}$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}$
1.960	25	39	25	9	34	428.96	142.60	254.45	0.0	35.821	6.382
1.645	24	40	24	8	32	437.04	144.87	254.44	0.0	38.099	7.529
1.000	17	47	17	9	26	509.58	163.55	255.45	0.0	53.875	11.875

주요 도시 인구 자료에서 전수층의 비율은 65%이었으며 Neyman의 최적할당을 이용하면 <표 1>에서 보듯이 t=1.96일때 첫번째 층에 할당된 표본크기 n_1=21은 부모집단의 크기 N_1=19보다 크다는 것을 볼 수 있으며 이는 t=1.645 일때도 마찬가지이다. t=1.00일 때는 두개의 층에서 모두 표본을 추출하게 되며 이때 추정치의 표준오차가 상당히 크다는 것을 알 수 있다. 한편 <표 2>에서 보는 바와 같이 절사법을 응용하면 t = 1.96일 때 전수층에서 $N_1 = n_1$=25를 표본층에서 n_2=9개의 요소를 할당하며 이때의 추정치의 표준오차는 $std(\bar{y}_{Cutoff})$= 6.382로서 Neyman할당법에 의한 표준오차 $std(\bar{y}_{Neyman})$ = 7.0536보다 적음을 알 수 있다.



<그림 2> 경기 및 오락용품 임대업의 크기와 순위와의 관계를 나타낸 그림

경기 및 오락용품 임대업의 경우 <표3>에서 보듯이 t=1.96일 때 Neyman의 최적할당에 의하면 $n_1 > N_1$이나 추정치의 표준오차를 살펴 보면 Neyman의 방법을 따를 경우 $std(\bar{y}_{Neyman})$ = 0.6039, <표 4>를 통해 알 수 있듯이 절사법을 응용하면 $std(\bar{y}_{Cutoff})$=0.6541로서 Neyman의 방법이 더 적은 표준오차를 제공한다. 이 자료는 모집단 총계중 전수층의 비율이 47%로 비교적 낮은 편이며 크기 순으로 배열된 자료의 그림이 1사분면에서 $Y=X$에 비대칭인 것을 <그림 2>를 통해서 볼 수 있다.

한편 <그림 2>의 X축과 같은 형태로 Y=X에 대해 대칭형태인 자료, 이러한 형태의 자료는 <그림 1>과 유사하나 양축의 꼬리부분이 <그림 1>보다는 두터운 모양이며 전수층이 차지하는 비율이 높을 경우 모의 실험 분석결과 Neyman의 할당법보다 절사법이 적은 표준오차를 제공하였다. 그러나 전수층의 비율을 50%, 40%, 30%로 낮출 경우 크기와 순위와의 그림에서 Y=X에 대한 대칭성은 사라지게되며 Neyman의 방법이 적절함을 볼 수 있었다.

<표 3> 경기 및 오락용품 임대업자료의 Neyman 할당법에 의한 표본할당과 그추정량

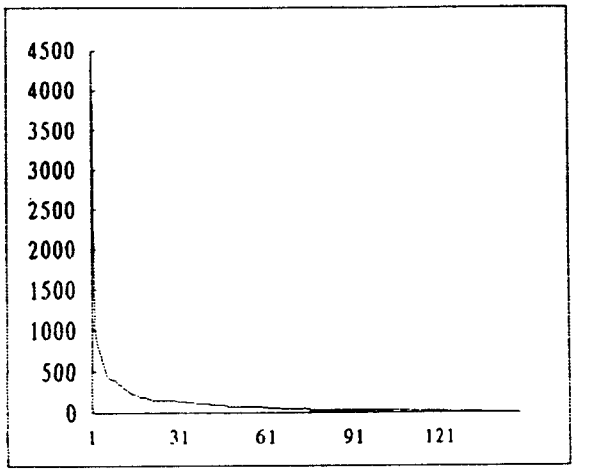
상대오차=0.05

t 값	모집단의 크기		표본의 크기			층별 평균 추정량			층별 표준오차		
	N_1	N_2	n_1	n_2	n	$\bar{y}_1 = \bar{Y}_1$	$\bar{y}_2$	$\bar{y}$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}$
1.960	38	97	42	26	64	44.10	16.09	23.977	0.0	5.0094	0.603
1.645	38	97	37	20	57	44.12	16.09	23.982	23.713	4.9899	0.736
1.000	38	97	28	14	42	44.22	16.08	24.005	23.366	4.8970	1.078

<표 4> 경기 및 오락용품 임대업자료의 절사법에 따른 표본할당량과 그 추정량

상대오차=0.05

t 값	모집단의 크기		표본의 크기			층별 평균 추정량			층별 표준오차		
	N_1	N_2	n_1	n_2	n	$\bar{y}_1 = \bar{Y}_1$	$\bar{y}_2$	$\bar{y}$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}$
1.960	33	102	33	31	64	46.75	16.419	23.833	0.0	5.777	0.654
1.645	30	105	30	27	57	48.60	17.185	24.166	0.0	5.540	0.711
1.000	25	110	25	17	42	51.40	17.647	23.897	0.0	7.296	1.316



<그림 3> 승용차 임대업자료의 크기와 순위와의 관계를 나타낸 그림

승용차 임대업 조사자료의 전수층의 비율은 약 80%이며 크기와 순위와의 관계를 나타낸 그림은 <그림 3>에서 보는 바와 같이 1사분면에서 $Y=X$에 대해 대칭임을 볼 수 있다. <표 5>에서 보는 바와 같이 t의 각 값에 대해 $n_1 > N_1$이다. 이 자료에 대해서도 도시인구 자료와 마찬가지로 절사법의 응용이 적은 표준오차를 제공함을 <표 6>을 통해 볼 수 있다.

<표 5> 승용차 임대업자료의 Neyman 할당법에 의한 표본 할당량과 그 추정량

상대오차=0.05

t 값	모집단의 크기		표본의 크기			층별 평균 추정량			층별 표준오차		
	N_1	N_2	n_1	n_2	n	$\bar{y}_1 = \bar{Y}_1$	$\bar{y}_2$	$\bar{y}$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}$
1.960	33	116	54	28	61	415.54	47.189	128.77	0.0	30.663	3.923
1.645	33	116	50	24	57	415.54	47.253	128.82	0.0	30.719	4.347
1.000	33	116	40	12	45	415.54	47.200	128.78	0.0	30.148	6.419

<표 6> 승용차 임대업자료의 절사법에 의한 표본할당과 그 추정량

상대오차=0.05

t 값	모집단의 크기		표본의 크기			층별 평균 추정량			층별 표준오차		
	N_1	N_2	n_1	n_2	n	$\bar{y}_1 = \bar{Y}_1$	$\bar{y}_2$	$\bar{y}$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}$
1.960	46	103	46	15	61	330.09	37.93	128.09	0.0	21.402	3.531
1.645	40	109	40	17	57	364.45	41.00	127.83	0.0	24.584	4.007
1.000	22	127	22	23	45	549.86	55.43	129.08	0.0	39.468	6.348

4. 결 론

표본의 크기를 정함에 있어서 모집단의 극히 일부요소가 모집단총계의 절대다수를 차지하는 경우 모집단의 요소들을 크기순으로 나열한 후 2차원 상에 그래프를 그려 보면 원점을 중심으로 1사분면에서 X축과 Y축의 꼬리가 무한대로 퍼져나가는 형태(1사분면에서 hyperbola 형태)일 때 절사법을 사용하면 Neyman의 방법보다 적은 표본크기로 정도높은 추정치를 구할 수 있다는 것을 보였다. 특히 전국의 숙박업체 조사와 같이 극히 몇몇의 호텔이 전체 숙박업소 소득의 절대다수를 차지하는 경우와 같은 사업체 조사의 경우는 절사법을 사용하는 것이 유용하다고 하겠다. 그러나 자료를 크기순으로 재배열한 그림이 1사분면에서 원점을 중심으로 비대칭일 때는 지나치게 적은 표본크기 n으로 인하여 표본층의 크기가 매우적게 추정되어 절사법을 적용한다 하더라도 추정치의 변량(variation)이 크게 되어 적절하지 못하다.

참고문헌

[1] 사업체 기초통계 조사보고서 (1994). 통계청.

[2] 절사법 표본설계응용 (1991). 통계청.

[3] Cochran, W. G. (1977). Sampling Technique, 3rd edition, John Wiley & sons, New york.

[4] Dalenius, T. & Hodges, J. L. (1959). Minimum Variance Stratification, *Journal of American Statistical Association* 54, 88-101.

[5] Deming, W. E. (1960). Sample Design in Business Research, John Wiley & Sons, New York.

[6] MATLAB The Math Works, Inc. Natick. Massachusetts.

[7] Sukhatme, P.V., Sukhatme, B.V, Sukhatme, S. and Asok, C (1984). Sampling Theory of Surveys, With Applications. Iowa State University Press, Ames, Iowa.

A Study for the Efficiency of the Cut-off Method in Highly Skewed Populations

Geun-Shik Han[3], Yong-Chul Kim[4]

Abstract

In the design of the sampling, it is important to make a decision about the size of the sample to be selected from the population. We often have a problem to get the optimal size of the sample to be considered for cost and time expended for selecting sample unit from highly skewed population. In this case, we give a graphical criterion with Take-all Stratum rate to choose a method and also illustrate the efficiency between the Neyman allocation and the cut-off method with real data.

3) Assistant Professor, Department of Computer Science and Statistics, Hansin University, Osan, 447-270, Korea.

4) Full-time Lecturer, Department of Computer Science and Statistics, Yongin University, Yongin, 449-714, Korea.