유한모집단에서 분포함수추정량 비교

박혜균 [1] 김규성[2]

요 약

이 논문에서는 유한모집단 분포함수에 대한 추정량들을 소개하고, 이론적인 측면과 경험적인 측면으로 비교하였다. 분포함수 추정량은 설계기반 특성을 갖는 추정량과 모형기반 특성을 갖는 추정량으로 구분되며, 각각 설계기반 특성과 모형기반 특성을 갖는다. 수치적인 비교를 위하여 분포함수 추정량들을 2000년 인구주택총조사의 서울 가구수와 가구원수 데이터에 적합하여 비교하였다.

주요용어 : 분포함수추정량, 유한모집단.

1. 서론

크기가 N인 유한모집단을 가정하고, 모집단 조사단위에서는 (x_i, y_i), $i=1,\ldots,N$를 관측한다고 하자. 이때 보조변수 x_i는 표집 전에 알려진 값으로 가정하고 관심변수 y_i는 표집 후에 관측하는 값이라고 하자. 유한모집단에서 임의의 값 t에서 분포함수는 다음과 같이 정의된다.

$$F(t)=\frac{1}{N}\sum_{i=1}^{N}\Delta(t-y_i) \tag{1}$$

여기서 $\Delta(\cdot)$는 다음과 같이 정의된 지시함수이다.

$$\Delta(a)=\begin{cases}1, & \text{만일 } a\geq 0\\ 0, & \text{기타}\end{cases} \tag{2}$$

우리의 관심은 주어진 t 값에서 분포함수 $F(t)$를 추정하는 것이다. 일반적인 표집이론에서와 마찬가지로 분포함수 $F(t)$의 추정량은 설계기반추정량과 모형기반추정량으로 구분할 수 있다. 그리고 각각은 설계기반 특성과 모형기반 특성을 그대로 가지고 있다. 다음절에서는 설계기반 분포함수 추정량과 모형기반 분포함수 추정량을 소개하고 3절에서는 소개된 분포함수 추정량을 2000년 인구주택총조사 데이터에 적합하여 수치적으로 비교·분석하기로 한다.

2. 분포함수 추정량

무한모집단에서 IID 표본에 기초한 통상적인 추정량은 다음과 같은 표본분포함수이다.

$$\hat{F}_0(t)=\frac{1}{n}\sum_{j\in s}\Delta(t-y_j) \tag{3}$$

임의의 고정된 값 t에서 위의 추정량은 일치추정량이며 표본의 크기가 크면 근사적으로 평균이 $F(t)$이고 분산이 $F(t)(1-F(t))/n$인 근사정규분포로 수렴함이 알려져 있다(Serfling,

1) 서울시립대 통계학과 석사과정, 서울시 동대문구 전농3동 90.
2) 서울시립대 통계학과 교수, 서울시 동대문구 전농3동 90.

1980, p. 57). 그런데 유한모집단에서 서로 다른 추출확률을 갖는 표본을 고려하고, 또한 사용가능한 보조변수가 주어진 경우를 고려하면 (3)에서 주어진 표본분포함수는 최선의 추정량이 아닐 수 있다. 왜냐하면 추출확률과 보조변수를 추정에 이용하여 추정의 효율을 높일 수 있기 때문이다. 표본추출확률에 근거하여 제안된 추정량들이 설계기반추정량이며 관심변수와 보조변수의 관련성에 근거하여 만들어진 추정량들이 모형기반 추정량이다.

2.1 설계기반 분포함수추정량

설계기반 분포함수추정량들은 표본의 추출확률에 기초한 추정량들이다. 따라서 포함확률 π_j가 추정량의 주요 성분이 된다. 모평균의 추정에서와 마찬가지로 포함확률만을 이용하여 만든 Hajek 추정량을 고려할 수 있다.

$$\hat{F}_0(t) = \sum_{j\in s} \frac{\Delta(t-y_j)}{\pi_j} \Big/ \sum_{j\in s} \frac{1}{\pi_j} \tag{4}$$

모평균 추정의 Hajek 추정량과의 차이는 지시함수 $\Delta(t-y_j)$가 관심변수 y_j에 대한 선형함수가 아니기 때문에 관심변수와 포함확률이 근사적으로 비례한다 하더라도 추정량의 분산이 0으로 수렴하지는 않는다는 점이다.

설계기반추론에서 널리 이용되는 비추정량은 분포함수추정에서도 적용할 수 있다. 분포함수에 대한 비추정량은 다음과 같다.

$$\hat{F}_r(t) = \left(\frac{1}{N} \sum_{j\in s} \frac{\Delta(t-y_j)}{\pi_j} \Big/ \sum_{j\in s} \frac{\Delta(t-\hat{R}x_j)}{\pi_j} \right) \sum_{i=1}^{N} \Delta(t-\hat{R}x_i) \tag{5}$$

여기서 $\hat{R} = (\sum_{k\in s} y_k/\pi_k)/(\sum_{k\in s}(x_k/\pi_k)$이다. 관심변수 y가 보조변수 x에 비례할 때 비추정량 $\hat{F}_r(t)$는 분포함수값 $F(t)$가 된다. 즉, 관심변수가 보조변수와 비례할 때 비추정량 $\hat{F}_r$의 효율은 증가한다.

보조변수를 이용한 또 다른 설계기반 추정량은 다음과 같은 편차추정량이다.

$$\hat{F}_d(t) = \frac{1}{N} \left(\sum_{j\in s} \frac{\Delta(t-y_j)}{\pi_j} + \sum_{i=1}^{N} \Delta(t-\hat{R}x_i) - \sum_{j\in s} \frac{\Delta(t-\hat{R}x_j)}{\pi_j} \right) \tag{6}$$

식 (5)에서 소개된 비추정량과 달리 편차추정량은 설계비편향이다.

Kuk & Mak(1989)은 분할표를 이용하여 분포함수를 추정하는 방법을 고안하였다. 만일 M_x를 보조변수 x에 대한 모집단 중위수라고 하고

$$\hat{F}_1(t) = \sum_{j\in s} \Delta(M_x - x_j)/n, \quad \hat{F}_2(t) = \sum_{j\in s} \Delta(x_j - M_x)/n,$$

라고 하며 $N_x = \sum_{j=1}^{N} \Delta(M_x - x_j)$라 하자. Kuk& Mak이 제안한 추정량은 다음과 같다.

$$\hat{F}_{KM} = \frac{1}{N}[N_x \hat{F}_1(t) + (N-N_x)\hat{F}_2(t)] \tag{7}$$

Kuk & Mak의 추정량은 관심변수와 보조변수의 결합분포를 고려하여 구한 추정량인 점이 특징이다.

2.2 모형기반 분포함수추정량

분포함수추정에 모형을 처음으로 도입한 사람은 Chambers & Dunstan(1986)이다. 그들은 다음과 같은 비모형을 고려하였다.

$$Y_i = \beta x_i + v(x_i)\varepsilon_i, \quad i=1,\ldots,N \tag{8}$$

여기서 $\varepsilon_i \sim i.i.d.\ (0,\sigma^2)$이며, $v(x_i)$는 알려진 보조변수 x_i의 함수인데, Chambers & Dunstan이 고려한 함수는 $v(x_i)=\sqrt{x_i}$였다. 비모형을 고려하여 Chambers & Dunstan이 제안한 모형비편향 분포함수추정량은 아래와 같다.

$$\hat{F}_{CD}(t) = \frac{1}{N}\left(\sum_{j\in s}\Delta(t-y_j) + \frac{1}{n}\sum_{i\notin s}\sum_{j\in s}\Delta\left(\frac{t-b_n x_i}{v(x_i)} - \frac{y_j - b_n x_j}{v(x_j)} \right) \right) \tag{9}$$

여기서, $b_n = \sum_{j\in s}\frac{x_j y_j}{v(x_j)} \Big/ \sum_{j\in s}\frac{x_j^2}{v(x_j)}$ 이며 $v(x_j)=\sqrt{x_j}$이다. $\hat{F}_{CD}$의 특징은 비모형을 기초로 만들었기 때문에 적용하고자 하는 데이터가 비모형에 적합이 잘되면 $\hat{F}_{CD}$의 효율은 높으며 만일 모형이 적합되지 않으면 추정의 효율은 감소한다.

설계비편향성과 모형비편향성을 동시에 지니는 추정량에 대한 개발은 Rao, Kover and Matel(1990)에 의해 이루어졌다. 그들은 모형보조추론의 방법에 기초하여 근사적으로 설계비편향이면서 모형비편향인 추정량을 제안하였다.

$$\hat{F}_{RKM}(t) = \frac{1}{N}\left\{ \sum_{j\in s}\frac{\Delta(t-y_j)}{\pi_j} + \sum_{i=1}^{N}\hat{G}_i - \sum_{j\in s}\frac{\hat{G}_{jc}}{\pi_j} \right\} \tag{10}$$

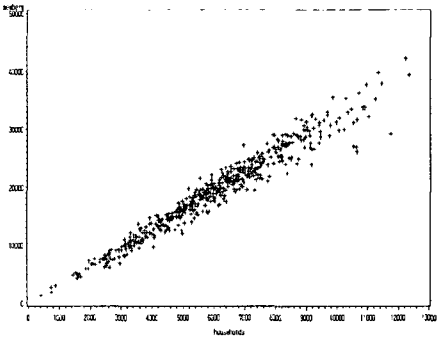
여기서

$$\hat{G}_i = \frac{\sum_{j\in s}\Delta\left(\frac{t-\hat{R}x_i}{\sqrt{x_i}} - \frac{y_j-\hat{R}x_j}{\sqrt{x_j}} \Big/ \pi_j \right)}{\sum_{j\in s}1/\pi_j}, \quad \hat{G}_{jc} = \frac{\sum_{k\in s}\frac{\pi_j}{\pi_{jk}}\Delta\left(\frac{t-\hat{R}x_j}{\sqrt{x_j}} - \frac{y_k-\hat{R}x_k}{\sqrt{x_k}} \right)}{\sum_{k\in s}\pi_j/\pi_{jk}}$$

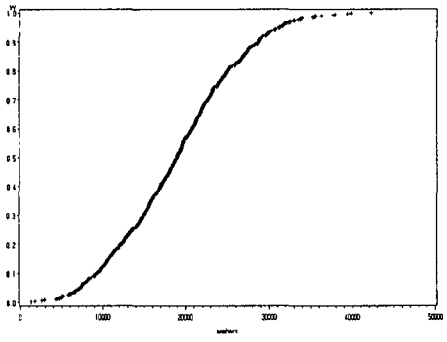
이다.

3. 모의실험

앞에서 소개한 분포함수추정량들을 수치적으로 비교하기 위하여 모이실험을 실시하였다. 모의실험을 위하여 2000년 인구주택총조사보고서(서울)에서 서울의 가구수(x_i)와 가구원수(y_i)를 변수로 하였으며 모집단 크기는 522이다: $\{(x_i,\ y_i) : i=1,\ldots,522\}$. (<그림 1>, <그림 2>).



<그림 1> 가구수와 가구원수의 산점도



<그림 2> 가구원수의 분포함수

모집단에서 크기 20, 40, 60, 80인 표본을 단순임의추출방법으로 1,000개 반복 추출하고, 각각의 추출된 표본으로부터 앞에서 소개한 분포함수 추정량을 계산하였다. 그리고 분포함수의 효율을 알아보기 위하여 상대표준오차(relative mean error, RME)와 상대제곱근평균제곱오차(relative root mean square error, RRMSE)을 계산하였다.

$$\mathrm{RME}(t)=\frac{1}{A}\sum_{s=1}^{A}(\hat{F}^{s}(t)-F(t))/F(t),\quad A=1,000 \tag{11}$$

그리고

$$\mathrm{RRMSE}(t)=\sqrt{\frac{1}{A}\sum_{s=1}^{A}(\hat{F}^{s}(t)-F(t))^{2}/F(t)^{2}} \tag{12}$$

이다. 그리고 t 값은 누적확률 p에 대응하는 관심변수의 분위수이다. 모의실험에 이용된 분위수는 $p=0.01,\ 0.05,\ 0.10,\ 0.25,\ 0.50,\ 0.75,\ 0.90,\ 0.95,\ 0.99$이다.
아래의 <표 1>에 RME가 주어져 있으며 <표 2>에 RRMSE가 나타나 있다.

<표 1> RME

표본 크기	추정 방법	분위수								
		p=0.01	p=0.05	p=0.10	p=0.25	p=0.50	p=0.75	p=0.90	p=0.95	p=0.99
20	F_0	0.3927	0.0536	0.0297	0.0146	0.0013	-0.0034	-0.0021	-0.0004	0.0008
	F_r	-1.3193	-0.6160	-0.1347	0.0647	0.0150	0.0022	0.0067	0.0032	-0.0014
	F_d	0.3837	0.0287	0.0398	0.0109	0.0035	-0.0005	0.0026	0.0017	-0.0016
	F_{KM}	0.3730	0.0815	0.0359	0.0141	-0.0018	-0.0047	-0.0022	-0.0005	0.0006
	F_{CD}	0.2272	-0.0409	0.0094	-0.0038	0.0302	0.0065	0.0135	-0.0029	-0.0039
	F_{RKM}	0.4602	0.0540	0.0352	0.0068	-0.0007	-0.0035	-0.0019	-0.0018	-0.0054
40	F_0	0.2730	0.0517	0.0139	0.0163	0.0041	-0.0004	-0.0004	-0.0004	0.0003
	F_r	-1.2303	-0.1017	0.0770	0.0408	0.0112	0.0037	0.0037	0.0026	-0.0002
	F_d	0.2526	0.0597	0.0172	0.0210	0.0057	0.0018	0.0028	0.0012	-0.0007
	F_{KM}	0.2816	0.0640	0.0194	0.0225	0.0063	0.0005	-0.0003	-0.0002	0.0004
	F_{CD}	0.1709	-0.0276	0.0225	0.0044	0.0212	0.0076	0.0103	-0.0010	-0.0030
	F_{RKM}	0.2639	0.0603	0.0205	0.0129	0.0005	-0.0016	-0.0015	-0.0030	-0.0040
60	F_0	0.3161	0.0773	0.0351	0.0013	-0.0003	-0.0014	0.0021	0.0010	0.0014
	F_r	-1.1028	0.1444	0.0873	0.0121	0.0074	0.0013	0.0027	0.0015	-0.0002
	F_d	0.2742	0.0706	0.0218	-0.0012	0.0035	0.0008	0.0019	0.0007	-0.0002
	F_{KM}	0.3184	0.0845	0.0377	0.0061	0.0020	-0.0003	0.0025	0.0013	0.0014
	F_{CD}	0.1543	-0.0045	0.0407	-0.0086	0.0146	0.0054	0.0106	0.0001	-0.0024
	F_{RKM}	0.2738	0.0690	0.0309	-0.0030	-0.0018	-0.0027	-0.0012	-0.0030	-0.0030
80	F_0	0.3664	0.0785	0.0292	0.0176	0.0011	0.0035	0.0006	-0.0008	0.0000
	F_r	-0.8994	0.2309	0.0617	0.0242	0.0022	0.0027	0.0011	0.0007	-0.0007
	F_d	0.3138	0.0678	0.0183	0.0133	-0.0003	0.0018	0.0007	0.0001	-0.0009
	F_{KM}	0.3535	0.0767	0.0278	0.0160	-0.0010	0.0026	0.0004	-0.0009	-0.0001
	F_{CD}	0.1159	0.0091	0.0304	-0.0036	0.0139	0.0076	0.0086	-0.0003	-0.0032
	F_{RKM}	0.3120	0.0682	0.0230	0.0127	-0.0040	-0.0004	-0.0030	-0.0041	-0.0039

분위수가 클수록 상대편향은 감소한다. 분위수가 작을때 비추정량은 다른추정량 보다 상대적으로 큰 편향을 보인다. 그리고 분위수가 작을 때 편향이 작은 추정량은 Chambers & Dunstan 추정량이다. 그러나 분위수가 커지면 편향을 전반적으로 줄어든다.

<표4.2>RRMSE

표본 크기	추정 방법	분위수								
		p=0.01	p=0.05	p=0.10	p=0.25	p=0.50	p=0.75	p=0.90	p=0.95	p=0.99
20	F_0	3.3662	1.2984	0.8887	0.5238	0.3033	0.1725	0.1013	0.0667	0.0298
	F_r	1.1397	0.8236	0.6414	0.3216	0.1432	0.0812	0.0666	0.0534	0.0155
	F_d	1.9848	0.8281	0.4119	0.2480	0.1504	0.0882	0.0722	0.0572	0.0168
	F_{KM}	3.3706	1.3218	0.8576	0.4442	0.1517	0.1467	0.0992	0.0674	0.0307
	F_{CD}	0.7725	0.4026	0.2937	0.1613	0.1145	0.0726	0.0490	0.0326	0.0171
	F_{RKM}	1.7305	0.7284	0.3631	0.2096	0.1294	0.0815	0.0638	0.0498	0.0157
40	F_0	2.2569	0.9124	0.6183	0.3454	0.2101	0.1212	0.0691	0.0482	0.0209
	F_r	1.1069	0.7396	0.4795	0.1890	0.1064	0.0569	0.0473	0.0390	0.0116
	F_d	1.2639	0.5642	0.2710	0.1722	0.1109	0.0598	0.0498	0.0410	0.0127
	F_{KM}	2.2793	0.9208	0.5992	0.3008	0.1029	0.0989	0.0664	0.0473	0.0209
	F_{CD}	0.6945	0.3536	0.2250	0.1226	0.0852	0.0529	0.0379	0.0275	0.0140
	F_{RKM}	1.1387	0.5025	0.2388	0.1463	0.0921	0.0528	0.0429	0.0352	0.0118
60	F_0	1.8371	0.7501	0.5054	0.2809	0.1659	0.0971	0.0558	0.0402	0.0167
	F_r	1.0704	0.7207	0.3400	0.1491	0.0844	0.0459	0.0400	0.0324	0.0100
	F_d	1.0381	0.4575	0.2124	0.1429	0.0850	0.0471	0.0408	0.0330	0.0102
	F_{KM}	1.8416	0.7468	0.4822	0.2380	0.0800	0.0795	0.0528	0.0395	0.0166
	F_{CD}	0.5807	0.2971	0.1949	0.1077	0.0669	0.0397	0.0324	0.0233	0.0116
	F_{RKM}	0.9503	0.4050	0.1937	0.1195	0.0688	0.0417	0.0345	0.0290	0.0102
80	F_0	1.5707	0.6337	0.4316	0.2431	0.1433	0.0841	0.0478	0.0323	0.0141
	F_r	1.0137	0.7192	0.2958	0.1277	0.0729	0.0402	0.0342	0.0265	0.0096
	F_d	0.9090	0.3933	0.1799	0.1200	0.0739	0.0412	0.0347	0.0270	0.0097
	F_{KM}	1.5479	0.6183	0.4083	0.1956	0.0696	0.0692	0.0453	0.0319	0.0142
	F_{CD}	0.4864	0.2677	0.1587	0.0845	0.0557	0.0367	0.0269	0.0192	0.0100
	F_{RKM}	0.8392	0.3477	0.1654	0.0972	0.0589	0.0372	0.0295	0.0233	0.0092

평균제곱오차의 관점에서는 Chambers & Dunstan의 추정량이 가장 작게 나타난다. 그 이유는 모집단으로 사용된 데이터가 비모형에 잘 적합되기 때문인 것으로 보인다. 다음으로 비추정량과 Rao et al.의 추정량이 높은 효율을 보이는 것으로 나타나는데 그 이유는 Chambers & Dunstan과 마찬가지로 보조변수를 추정에 이용하고 있기 때문이다. 보조변수를 이용하지 않는 표본분포함수는 상대적으로 큰 평균제곱오차를 보인다.

4. 결론

본 논문에서는 유한모집단에서 분포함수추정량에 관하여 고찰하였다. 분포함수추정량은 설계기반추정량과 모형기반추정량으로 구분할 수 있으며 각각 추출확률과 모형을 중요시하는 특징이 있다. 모의실험결과 서울의 가구수 및 가구원수 모집단에는 Chambers & Dunstan의 추정량

이 우수한 것으로 나타났으며, 비추정량과 Rao, Kovar & Mantel 추정량도 우수한 것으로 나타났다.

본 논문에서는 표본추출을 단순임의추출만 했기 때문에 추출확률 차이에 의한 추정량의 비교는 하지 못하였다. 불균등추출확률에 의한 효과를 보기 위한 모의실험은 향후에 수행할 예정이다.

참고문헌

[1] Chambers, R.L. and Dunstan, R. (1986). Estimating distribution functions from survey data, *Biometrika*, vol. 73,597-604.

[2] Kuk, A.Y.C. and Mak, T.K. (1989). Median Estimation in the Presence of Auxiliary Information, *Journal of Royal Statistical Society, Ser B*, 51, 261-299.

[3] Mukhopadhyay, P. (2000). *Topics in Survey Sampling*, New York: Springer, chap 6, 165-201.

[4] Rao, J.N.K. and Kover, J.G. and Mantel, H.J. (1990). On Estimating Distribution Functions and Quantiles from Survey Data Using Auxiliary Infromation. *Biometrika*, 77, 365-375.

[5] Serfling, R.J. (1980). Approximation theorems of mathematical statistics. John Wiley and sons.