응용통계연구 (2011)
24(1), 25–34

DOI: 10.5351/KJAS.2011.24.1.025

패널조사 표본설계 시 표본크기 결정에 관한 연구

유양상[1] · 신기일[2]

[1]한국외국어대학교 통계학과, [2]한국외국어대학교 통계학과

(2010년 8월 접수, 2010년 11월 채택)

요 약

패널자료를 이용하여 부모집단(subpopulation) 총계를 추정할 경우 추이확률(transition probability)을 사용할 수 있다. 패널조사는 같은 표본을 계속 조사하기 때문에 한번 정해진 표본크기는 자료분석에 중대한 영향을 미친다. 본 논문에서는 장애인고용패널조사에서 추이확률을 이용한 부모집단 총계 추정사례를 살펴보고, 부모집단 총계 추정에 표본크기가 얼마나 영향을 미치는지 살펴보았다.

주요용어: 패널자료, 추이확률, 가중치 보정, 벤치마킹 보정.

1. 서론

패널자료는 종단면과 횡단면 정보를 모두 사용할 수 있기 때문에 정확한 통계적 추론이 가능하다. 이러한 이유로 국내외적으로 많은 패널조사가 이루어지고 있으며 국내에서는 약 16개의 패널조사 자료가 얻어지고 있다. 패널조사의 가장 큰 특징은 종단면 정보를 얻기 위해 한번 정해진 표본을 계속해서 조사하는 것이다. 이는 표본설계를 다시하지 않는다는 장점이 있는 반면 추출된 표본이 최신의 모집단 특성을 잘 반영하지 못한다는 단점이 있다. 이를 극복하기 위한 방법의 하나가 벤치마킹 보정이다. 벤치마킹 보정방법은 사후층화보정방법의 하나로 이미 국내의 많은 보고서에서 사용되고 있는 방법이다. 사후층화보정방법에 관한 많은 연구가 진행되고 있으며 이에 관한 내용은 김규성 (2005)과 김석과 신기일 (2008)을 참고하기 바란다.

장애인 경제활동상태를 파악하기 위한 장애인고용패널조사가 2006년 기초연구를 시작으로 2008년부터 실시되었다. 이 패널조사의 주된 목적은 장애인들의 경제활동이 동적인 관점에서 어떻게 변하는지를 파악하는 것이다. 이러한 목적을 달성하기 위한 기초 자료가 각 경제활동상태의 비율과 총계이다. 즉 취업률, 비경제활동비율, 그리고 실업률과 이를 기반으로 얻어진 총계이다. 장애인고용패널조사에 관한 자세한 내용은 장애인고용촉진공단 고용개발원 (2006, 2009a)을 참조하기 바란다.

장애인 경제활동상태는 분기별로 얻어지는 등록장애인명부에 나타나지 않는 변수로 설문에 의해 조사해야 하는 항목이며 경제활동상태의 비율은 3년마다 한국보건사회연구원에서 실시하는 장애인고용실태조사에서 얻어진다. 특히 패널조사를 위한 표본설계 당시에는, 실업률이 작기 때문에 실업자에 관한 정확한 정보를 얻기 위해 과대 배정(over sampling)을 하게 된다. 따라서 경제활동상태별 비율은 장애인고용패널조사에서 단순히 얻어지지 않는다.

이 논문은 2010년도 한국외국어대학교 학술연구비 지원에 의해 이루어진 것임.
[2]교신저자: (449-791) 경기도 용인시 모현면 왕산리 산 89, 한국외국어대학교 통계학과, 교수.
E-mail: keyshin@hufs.ac.kr

또한 한번 표본으로 설정되면 패널조사의 특성상 모집단이 변화하여도 계속 같은 표본이 조사되기 때문에 시점별로 표본 비율의 변화가 부모집단 비율의 변화를 따르지 않을 수 있다. 따라서 다른 행정자료 또는 벤치마킹 자료에서 얻어진 정보와 표본으로 선택된 자료의 경제활동상태 변화를 이용하여 실업률, 취업률 등을 추정하여야 한다.

먼저 $t-1$ 시점의 장애인별 경제활동상태가 알려져 있고 t 시점에서도 장애인별 경제활동상태가 조사되었다고 하자. 이런 경우에 t 시점의 경제활동상태 비율을 파악하는 것이 장애인고용패널조사의 중요한 내용 중의 하나이다. 이때 적용 가능한 방법 중의 하나가 추이확률행렬(transition probability matrix)을 사용하는 것이다. 즉 $t-1$ 시점에서 취업 상태인 사람은 t 시점에서도 취업 상태일 수 있고, 실업 상태일 수도 있으며 비경제활동상태일 수도 있다. 또한 $t-1$ 시점에서 실업 상태인 사람은 t 시점에서 취업, 실업 또는 비경제활동 상태일 수도 있다. 이렇게 여러 경우가 발생할 때 추이확률행렬을 사용하게 되면 변화된 취업, 실업 그리고 비경제활동 상태의 비율을 어렵지 않게 계산할 수 있다. 본 논문에서는 추이확률행렬을 이용하여 t 시점의 경제활동상태 비율과 총계를 추정하는 방법을 살펴보았다. 특히 부모집단 비율과 총계 추정에 표본크기가 얼마나 영향을 주는지 살펴보았다.

본 논문의 구성은 다음과 같다. 먼저 2절에서 추이확률에 관하여 간단히 살펴보았다. 3절에서는 장애인고용패널조사 자료를 이용하여 실제 추이확률을 이용한 경제활동상태 비율과 총계 추정 방법을 설명하였으며 또한 표본크기가 경제활동상태 비율과 총계 추정에 얼마나 영향을 주는지에 관하여 살펴보았다. 끝으로 토의 및 결론은 4절에 있다.

2. 추이확률

유한개의 상태를 갖고 있는 마코프 연쇄(Markov chain)를 고려하자. 이제 X_k는 k 시점의 변수이고 i, j와 s_{n-2}는 상태를 나타낸다고 하자. 그리고 모든 상태는 상호교류가 가능하다고 가정하자. 그러면 잘 알려진 마코프 성질은 다음과 같다.

$$p_{ij}^{(n)} = \Pr\{X_n = j | X_{n-1} = i, X_{n-2} = s_{n-2}, \cdots\} = \Pr\{X_n = j | X_{n-1} = i\} = p_{ij}. \tag{2.1}$$

마코프 성질의 가장 큰 특징은 과거의 여러 상태 정보가 주어졌어도 최근의 상태에 있을 확률은 바로 전 시점의 정보에만 영향을 받는 다는 것이다. 또한 (2.1) 식에서 $p_{ij}^{(n)}$이 시점 n에 무관할 경우, $p_{ij}^{(n)} = p_{ij}$를 정상추이확률(stationary transition probability)이라 부르고 이러한 경우의 연쇄를 정상마코프연쇄라 부른다. 이러한 추이확률을 이용하여 만든 행렬을 추이확률행렬이라 부른다. 예를 들어 상태집합 $S = \{0, 1, 2\}$라 하자. 즉 상태는 세 가지가 있다. 그리고 각 상태는 상호교류한다고 하자. 즉 $p_{ij} > 0$이고 $\sum_j p_{ij} = 1$이다. 그러면 추이확률행렬은 다음과 같이 표시 될 수 있다.

$$P = \begin{pmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{pmatrix}. \tag{2.2}$$

이 추이확률행렬을 이용하게 되면 n 시점의 변수 X_n이 상태 i에 있을 확률을 구할 수 있다. 예를 들어 상태집합 $S = \{0, 1, 2\}$라 하고 변수 X_0가 시점 '0'에서 상태 i에 있을 확률을 $\pi_i^{(0)}, i = 0, 1, 2$이라 하자. 즉 $\pi_i^{(0)} = \Pr\{X_0 = i\}$이고 $\sum_{i=0}^{2} \pi_i^{(0)} = 1$이다. 이제 $\pi^{(0)} = (\pi_0^{(0)}, \pi_1^{(0)}, \pi_2^{(0)})$라 표시하자. 그러면 n 시점 후의 변수 X_n이 상태 i에 있을 확률 $\pi^{(n)} = (\pi_0^{(n)}, \pi_1^{(n)}, \pi_2^{(n)})$은 다음과 같이 구해진다.

$$\pi^{(n)} = \left(\pi_0^{(n)}, \pi_1^{(n)}, \pi_2^{(n)}\right) = \left(\pi_0^{(0)}, \pi_1^{(0)}, \pi_2^{(0)}\right) \times \begin{pmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{pmatrix}^n. \tag{2.3}$$

표 **3.1.** 경제활동상태별 자료

		2차 패널조사			합계
		취업자	실업자	비경활	
1차 패널조사	취업자	1629	52	110	1,791
	실업자	57	52	93	202
	비경활	197	77	2410	2,684
합계		1,883	181	2,613	4,677

결국 n 시점 변수 X_n이 상태 i에 있을 확률은 $n-1$ 시점의 확률에 추이확률행렬 P를 곱하여 얻어지게 된다. 자세한 내용은 최기헌 (1998) 또는 Taylor와 Karlin (1984)을 살펴보기 바란다.

3. 장애인고용실태조사 자료분석 및 모의실험

이 절에서는 장애인고용실태조사에서 얻어진 자료를 분석함으로써 패널자료 분석에서 어떻게 추이확률을 사용하여 부모집단 총계를 추정하는지 살펴보았다. 부모집단 총계를 추정하기 위해서는 부모집단 비율이 추정되어야 하며 추정된 비율에 모집단 총계를 곱합으로써 부모집단 총계가 추정된다. 따라서 부모집단 비율 추정이 중요하므로 본 모의실험에서는 부모집단 비율과 이를 이용하여 추정된 부모집단 총계 추정에 표본크기가 얼마나 영향을 미치는지 살펴보았다.

3.1. 자료분석

장애인고용패널조사에서 추이확률행렬을 이용하여 경제활동상태 비율을 구하는 과정을 살펴보자. 먼저 상태는 취업상태, 실업상태 그리고 비경제활동상태 세 가지로 나누어진다. 2008년에 1차 조사가 이루어졌으며 그 당시 취업률이 0.409264, 실업률이 0.048483 그리고 비경제활동비율이 0.542253으로 구해져 있다. 즉 $\pi^{(0)} = (\pi_0^{(0)}, \pi_1^{(0)}, \pi_2^{(0)}) = (0.409264, 0.048483, 0.542253)$이다. 다음으로

T_{00} : 1차 조사에서 취업자였고 2차 조사에서도 취업자인 사람 수

T_{01} : 1차 조사에서 취업자였고 2차 조사에서는 실업자인 사람 수

T_{02} : 1차 조사에서 취업자였고 2차 조사에서는 비경제활동인 사람 수

T_{10} : 1차 조사에서 실업자였고 2차 조사에서도 취업자인 사람 수

T_{11} : 1차 조사에서 실업자였고 2차 조사에서는 실업자인 사람 수

T_{12} : 1차 조사에서 실업자였고 2차 조사에서는 비경제활동인 사람 수

T_{20} : 1차 조사에서 비경제활동인 사람이었고 2차 조사에서도 취업자인 사람 수

T_{21} : 1차 조사에서 비경제활동인 사람이었고 2차 조사에서는 실업자인 사람 수

T_{22} : 1차 조사에서 비경제활동인 사람이었고 2차 조사에서는 비경제활동인 사람 수

라 하자. 이에 해당되는 자료를 표로 구성하면 표 3.1과 같다.

표 3.1의 결과와 $p_{ij} = T_{ij}/T_{i\cdot}$, $T_{i\cdot} = \sum_{j=0}^{2} T_{ij}$를 이용하여 추이확률을 구하면 다음과 같다.

이제 $\pi^{(0)} = (\pi_0^{(0)}, \pi_1^{(0)}, \pi_2^{(0)}) = (0.409264, 0.048483, 0.542253)$가 구해졌고, 표 3.2를 추이확률행렬로

표 **3.2.** 경제활동상태별 추이확률

		2차 패널조사		
		취업자	실업자	비경활 인구
1차 패널조사	취업자	0.909548	0.029034	0.061418
	실업자	0.282178	0.257426	0.460396
	비경활	0.073398	0.028689	0.897914

표시하면

$$P = \begin{pmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 0.909548 & 0.029034 & 0.061418 \\ 0.282178 & 0.257426 & 0.460396 \\ 0.073398 & 0.028689 & 0.897914 \end{pmatrix} \tag{3.1}$$

가 된다. 따라서 2차 년도 경제활동상태 비율은 식 (2.2)를 이용하여 구할 수 있다. 즉

$$\begin{aligned} \pi^{(1)} &= (0.409264, 0.048483, 0.542253) \times \begin{pmatrix} 0.909548 & 0.029034 & 0.061418 \\ 0.282178 & 0.257426 & 0.460396 \\ 0.073398 & 0.028689 & 0.897914 \end{pmatrix} \\ &= (0.4257247, 0.0399194, 0.5343335) \end{aligned}$$

이다. 따라서 2009년의 취업률은 약 0.426으로 2008년의 0.409에 비해 증가한 것으로 추정되며 실업률은 감소한 것으로 그리고 비경제활동 비율도 감소한 것으로 추정된다. 추정된 비율과 등록장애인명부의 총 장애인수를 이용하면 각 경제상태별 부모집단 총계가 추정된다. 이에 관한 자세한 내용은 장애인고용촉진공단 고용개발원 (2009b)을 살펴보기 바란다.

3.2. 모의실험

각 경제활동상태별 비율 추정의 핵심은 추이확률행렬의 원소를 구성하는 확률을 정확히 추정하는 것이다. 이 확률은 자료로부터 추정되기 때문에 자료의 크기에 따라 그 정확도(precision)가 달라진다. 따라서 적은 수의 표본을 사용할 경우 추이확률행렬의 원소 추정의 정확도가 나빠질 수 있으며 그 결과로 경제활동상태 비율 추정값과 총계 추정값에 영향을 줄 수 있다. 따라서 이 절에서는 표본규모에 따른 경제활동상태 비율과 총계 추정의 민감도를 살펴보았다. 먼저 모의실험을 간단히 하기 위해 초기 확률을 다음의 두 경우로 결정하였다. 즉 $\pi^{(0)} = (\pi_0^{(0)}, \pi_1^{(0)}, \pi_2^{(0)}) = (0.40,\ 0.05,\ 0.55)$와 $(0.3,\ 0.3,\ 0.4)$를 사용하였다. $(0.40,\ 0.05,\ 0.55)$는 장애인고용패널조사 결과와 유사하게 $\pi_1^{(0)} = 0.05$로 작은 수를 사용하였고, 두 번째는 $\pi_i^{(0)}$, $i = 1, 2, 3$가 거의 유사한 값이 되도록 선택하였다. 확률의 순서는 상태의 순서로 분석에 영향을 주지 않으므로 장애인고용패널조사와 같이 두 번째 확률에 작은 수를 사용하였다.

다음으로 추이확률행렬 P의 영향을 살펴보기 위해 다음 두 경우의 행렬을 사용하였다. P_1의 경우 장애인고용실태조사 결과와 유사한 값으로 원소를 정하였다. 특징을 보면 상태 '0'과 상태 '2'의 대각원소 값이 크고, 상태 '0'에서 상태 '1'로, 그리고 상태 '0'에서 상태 '2'로 갈 확률이 매우 작다는 것이다. 또한 상태 '2'에서 상태 '0'과 상태 '1'로 갈 확률도 매우 작다. P_2의 경우는 모든 원소가 작지 않은 경우의 추이확률행렬을 고려하여 결정하였다. 물론 상태 '0'과 상태 '2'의 대각원소는 큰 값으로 결정하였다. 이는 상태 '0'과 상태 '2'의 경우 같은 상태에 있을 확률이 일반적으로 크기 때문이다.

초기행렬 두 경우와 추이확률행렬 두 경우의 조합이 각각 Case 1에서 Case 4가 되며 각 Case의 π_0, π_1 그리고 P는 다음과 같다.

표 **3.3.** Case 1의 백분위수 비교

π	n_1	Percentiles				
		5%	10%	50%	90%	95%
π_0^1	30	0.409569	0.411424	0.418758	0.426422	0.428504
	50	0.409742	0.412111	0.419025	0.425444	0.427227
	100	0.411939	0.413342	0.419168	0.424730	0.426087
	200	0.412073	0.413479	0.419063	0.424396	0.425865
	500	0.412306	0.413894	0.419256	0.424233	0.425594
π_1^1	30	0.030625	0.032099	0.037915	0.044051	0.045997
	50	0.031479	0.032707	0.037742	0.043042	0.044096
	100	0.032790	0.033913	0.037974	0.042071	0.043168
	200	0.033635	0.033635	0.038083	0.041333	0.042250
	500	0.034127	0.034127	0.037977	0.040844	0.041761
π_2^1	30	0.533670	0.535399	0.543139	0.550600	0.552371
	50	0.534313	0.536854	0.543273	0.550207	0.551919
	100	0.535418	0.537122	0.543077	0.548908	0.550648
	200	0.536021	0.537323	0.542948	0.548292	0.550042
	500	0.536161	0.537378	0.542894	0.547883	0.549850

Case 1: $\pi^{(0)} = \left(\pi_0^{(0)}, \pi_1^{(0)}, \pi_2^{(0)}\right) = (0.40, 0.05, 0.55)$, $P_1 = \begin{pmatrix} 0.90 & 0.03 & 0.07 \\ 0.30 & 0.30 & 0.40 \\ 0.08 & 0.02 & 0.90 \end{pmatrix}$

$\pi^{(1)} = (0.419, 0.038, 0.543)$

Case 2: $\pi^{(0)} = \left(\pi_0^{(0)}, \pi_1^{(0)}, \pi_2^{(0)}\right) = (0.3, 0.3, 0.4)$, $P_1 = \begin{pmatrix} 0.90 & 0.03 & 0.07 \\ 0.30 & 0.30 & 0.40 \\ 0.08 & 0.02 & 0.90 \end{pmatrix}$

$\pi^{(1)} = (0.392, 0.107, 0.501)$

Case 3: $\pi^{(0)} = \left(\pi_0^{(0)}, \pi_1^{(0)}, \pi_2^{(0)}\right) = (0.40, 0.05, 0.55)$, $P_2 = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.3 & 0.4 \\ 0.2 & 0.1 & 0.7 \end{pmatrix}$

$\pi^{(1)} = (0.405, 0.15, 0.445)$

Case 4: $\pi^{(0)} = \left(\pi_0^{(0)}, \pi_1^{(0)}, \pi_2^{(0)}\right) = (0.3, 0.3, 0.4)$, $P_2 = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.3 & 0.4 \\ 0.2 & 0.1 & 0.7 \end{pmatrix}$

$\pi^{(1)} = (0.38, 0.19, 0.43)$

본 논문은 표본규모가 경제활동상태의 비율과 총계 추정에 미치는 민감도를 살펴보는 것이 주 목적이므로 이를 고려하기 위해 다음과 같은 표본규모를 모의실험에서 사용하였다. 흔히 패널자료의 규모는 5,000에서 10,000개 정도이고, 장애인고용패널조사를 기본으로 하기 위해 전체 자료 수를 5,000개로 정하였다. 또한 소표본의 특성을 알아내기 위해 상태 '0'과 '2'는 많은 수를 그리고 상태 '1'은 상대적으로 작은 수를 배분하였다. 즉 $n = (n_0, n_1, n_2)$에서 (n_0, n_1, n_2)= (2485, 30, 2485), (2475, 50, 2475), (2450, 100, 2450), (2400, 200, 2400), (2250, 500, 2250)이 사용되었다. 모의실험에서 사용된 표본 수를 살펴보면 단지 n_1의 영향을 고려한 것으로 볼 수 있으나 n_1 대신 n_0 또는 n_2를 변화시켜도 같은 결

표 **3.4.** Case 2의 백분위수 비교

π	n_1	Percentiles				
		5%	10%	50%	90%	95%
π_0^1	30	0.351932	0.361610	0.391469	0.424266	0.434044
	50	0.361242	0.368000	0.391778	0.416646	0.425283
	100	0.367959	0.375184	0.392735	0.410867	0.417020
	200	0.375958	0.379458	0.392604	0.405229	0.409042
	500	0.380211	0.383000	0.391422	0.400822	0.403300
π_1^1	30	0.066780	0.076016	0.106760	0.137444	0.146780
	50	0.076161	0.083373	0.107474	0.132707	0.141313
	100	0.083561	0.088612	0.106326	0.125224	0.130846
	200	0.091062	0.094104	0.106229	0.118916	0.122354
	500	0.096200	0.098466	0.106755	0.115377	0.117955
π_2^1	30	0.459457	0.468773	0.501247	0.537445	0.547304
	50	0.464889	0.472848	0.499495	0.526768	0.532949
	100	0.477173	0.482337	0.500204	0.519316	0.523888
	200	0.483896	0.487438	0.501813	0.514521	0.518354
	500	0.488967	0.491578	0.501544	0.510578	0.513244

표 **3.5.** Case 3의 백분위수 비교

π	n_1	Percentiles				
		5%	10%	50%	90%	95%
π_0^1	30	0.393181	0.395651	0.404722	0.414492	0.421605
	50	0.394222	0.396929	0.404818	0.413434	0.415389
	100	0.395036	0.397474	0.405265	0.413066	0.414796
	200	0.395719	0.397490	0.404531	0.412688	0.414875
	500	0.395211	0.397456	0.405133	0.412906	0.414817
π_1^1	30	0.139764	0.141994	0.149945	0.157792	0.159690
	50	0.140727	0.143242	0.150303	0.157768	0.159717
	100	0.141240	0.143378	0.150000	0.156796	0.158612
	200	0.142229	0.143958	0.150146	0.156542	0.158250
	500	0.142000	0.143811	0.150189	0.156844	0.158428
π_2^1	30	0.433511	0.435565	0.445478	0.454435	0.456569
	50	0.433258	0.435394	0.444869	0.453626	0.456045
	100	0.434633	0.437026	0.444811	0.452770	0.454878
	200	0.435125	0.437771	0.445083	0.452177	0.454021
	500	0.434833	0.437039	0.444772	0.452667	0.454900

과를 주기 때문에 n_1만을 변화시켰으며 이를 이용하여 소표본의 영향력을 확인하였다. 본 실험에서 사용한 반복수는 1,000번이다.

모의실험 결과로 얻어진 $\pi^{(1)} = (\pi_0^{(1)}, \pi_1^{(1)}, \pi_2^{(1)})$의 추정된 비율 분포가 표 3.3에서 3.6에 작성되었다. 또한 총계 추정에 미치는 영향력을 보기위해 MSE와 Bias를 구했으며 그 결과를 표 3.7에서 3.10에 작성하였다.

표 3.3을 살펴보면 표본 수 $n = (n_0, n_1, n_2)$의 크기에 상관없이 $\pi^{(1)} = (\pi_0^{(1)}, \pi_1^{(1)}, \pi_2^{(1)})$은 매우 안정적인 분포를 보이고 있다. 이러한 결과는 표 3.5에서도 확인할 수 있다. 그러나 표 3.4와 3.6을 살펴보면

표 **3.6.** Case 4의 백분위수 비교

π	n_1	Percentiles				
		5%	10%	50%	90%	95%
π_0^1	30	0.337807	0.346197	0.379819	0.412475	0.423521
	50	0.347566	0.355010	0.380232	0.407687	0.414657
	100	0.355959	0.362327	0.380112	0.398214	0.402939
	200	0.361000	0.365125	0.379521	0.392646	0.396938
	500	0.367511	0.369956	0.379667	0.389811	0.392711
π_1^1	30	0.149135	0.157636	0.189416	0.222183	0.233018
	50	0.160833	0.166798	0.189636	0.216030	0.224131
	100	0.166378	0.171714	0.189918	0.207786	0.213980
	200	0.173792	0.177667	0.190521	0.204771	0.209583
	500	0.178189	0.181000	0.190789	0.199556	0.201522
π_2^1	30	0.384316	0.395292	0.429095	0.466579	0.478229
	50	0.394202	0.400869	0.428586	0.457505	0.465788
	100	0.406092	0.411643	0.430010	0.448663	0.453969
	200	0.409333	0.414813	0.430021	0.444417	0.448729
	500	0.416933	0.419644	0.429800	0.439900	0.443211

표 **3.7.** Case 1의 MSE와 Bias

n_i	MSE			Bias		
	π_0	π_1	π_2	π_0	π_1	π_2
30	133611141	82075195	142553588	−44.38	126.14	−81.75
50	99787661	56933110	108552724	96.06	−83.57	−12.48
100	78812958	38899318	88784910	−411.83	195.83	216.00
200	75052725	27998906	78912413	233.83	99.25	−333.08
500	73814649	22206282	73992232	69.28	−89.13	19.84

표 **3.8.** Case 2의 MSE와 Bias

n_i	MSE			Bias		
	π_0	π_1	π_2	π_0	π_1	π_2
30	2591645140	2572530990	2793746392	−605.27	−1027.28	1632.55
50	1443103706	1480481162	1776204152	615.35	145.17	−760.52
100	759348538	731307728	924700411	−447.59	−594.69	1042.28
200	411049465	363494861	462853618	−187.25	577.66	−390.41
500	179848308	166016302	219149825	260.26	−582.62	322.35

표본 수에 따라 비율 분포가 달라지는 것을 확인할 수 있다. 이는 표본 수가 비율 분포에 영향을 미치고 있음을 보여주는 결과이다. 큰 차이를 보이고 있는 부분은 표본 수가 100개 이하인 경우이다. 물론 전국 추정을 할 경우 표본 규모를 100개 이하로 하는 경우는 별로 없겠지만 표본 수가 100개 이상인 경우에도 차이를 보이고 있다.

먼저 Case 1과 Case 3의 경우를 살펴보면 추이확률행렬이 P_1에서 P_2로 변하였다. 이러한 변화에도 불구하고 분포에는 큰 영향을 주고 있지 않다. 그러나 case 2와 case 4의 경우, 각각에 대응되는 case와 비교한다면 추이확률행렬은 동일하나 초기 확률이 변하였으며 분포에 큰 영향을 주고 있다. 이는 초기 확률의 분포에 따른 표본크기가 분포에 영향을 주고 있음을 보여주는 결과이다.

표 **3.9.** Case 3의 MSE와 Bias

n_i	MSE			Bias		
	π_0	π_1	π_2	π_0	π_1	π_2
30	226294808	145623582	217940340	494.25	−499.67	5.41
50	182128899	120597387	194221952	770.06	−164.12	−605.93
100	153766252	102327406	162201956	−239.02	300.02	−61.00
200	157330826	96296713	150456196	−436.08	−346.70	782.79
500	143642406	96806364	145777949	426.08	−6.33	−419.75

표 **3.10.** Case 4의 MSE와 Bias

n_i	MSE			Bias		
	π_0	π_1	π_2	π_0	π_1	π_2
30	2540827596	2613269902	2945035774	670.62	559.23	−1229.85
50	1598979941	1614824215	1713750228	−2964.28	2025.13	939.15
100	864066542	822460799	1023085103	410.61	8.04	−418.65
200	456777902	408777506	513433423	232.00	−12.08	−219.91
500	251781609	190799358	267075202	−342.66	302.26	40.40

다음으로 표 3.7에서 3.10을 살펴보자. 먼저 표 3.7과 3.9에서 표본 크기는 MSE에 영향을 주는 것으로 나타났다. 그러나 전술한 것처럼 전국 추정의 경우 표본 수를 100개 이하로 하는 경우가 흔하지 않기 때문에 그 이상을 살펴보면 MSE 기준으로 거의 일치하는 것을 확인할 수 있다. 반면 표 3.8과 3.10을 비교하면 표본크기가 총계 추정에 큰 영향을 주고 있는 것을 확인할 수 있다. 즉 표 3.7와 표 3.9 결과와 달리 표본수가 증가할수록 MSE는 계속 감소하고 있다. 이는 부모집단 비율 분포에서 확인한 결과와 일치한다. 반면에 MSE 결과와는 달리 모든 표에서 전체적으로 Bias는 매우 작은 것으로 나타났다.

이상의 결과는 소규모 모의실험 결과이므로 일반적인 결과로 확장시키는 것에 다소 무리가 따를 수 있다. 그러나 특별한 경우를 제외하고 초기 비율과 추이확률행렬 그리고 표본크기가 부모집단 비율 및 총계 추정에 영향을 준 다는 결론을 내릴 수 있다.

4. 토의 및 결론

본 연구에서는 패널자료를 이용한 총계추정 시 추이확률을 이용하는 방법을 살펴보았다. 패널조사에서는 한번 추출된 표본이 계속적으로 조사되기 때문에 부모집단 비율 및 총계 추정에 배분된 표본크기는 매우 중요하다. 특히 추이확률의 정확한 추정값은 부모집단 비율추정에 직접적인 영향을 주기 때문에 정확한 비율 추정은 중요하다. 본 논문에서는 부모집단에 배분된 표본크기가 부모집단 비율 및 총계 추정에 미치는 영향을 살펴보았다. 특히 case 1은 장애인고용패널조사 결과의 타당성을 보기 위한 것으로 장애인고용패널의 경우 약 $(n_0, n_1, n_2) = (1800, 200, 2600)$의 표본수를 사용하였으며 3절의 모의실험 결과를 적용한다면 다른 표본크기, 즉 n_1에 200 이상을 배분했을 경우와 거의 같은 결과를 얻는다고 할 수 있다. 그러나 이 결과는 매우 특수한 경우라 판단된다. 결론적으로 배분된 표본크기, 초기 비율 그리고 추이확률의 추정값이 종합적으로 부모집단 비율 추정에 영향을 미친다고 판단된다. 많은 경우 표본설계 시에는 추이확률행렬의 형태가 알려져 있지 않고 또한 초기확률도 알려져 있지 않기 때문에 표본설계 시 표본크기를 결정할 때에는 적정 크기의 표본수가 필요하며 특히 작은 부모집단이 존재할 때 그 부모집단의 비율과 총계 추정의 정도를 높이기 위해서는 과대 배정을 고려할 필요가 있다.

참고문헌

김규성 (2005). 표본의 대표성과 추정의 효율성, <조사연구학회>, **6**, 39–62.

김석, 신기일 (2008). 상관관계와 표본크기에 따른 BLS 무응답 보정의 효율성 비교, <응용통계연구>, **22**, 1301–1313.

최기헌 (1998). 확률과정론 입문, 자유아카데미.

한국보건사회연구원 (2009). 2008년 장애인 실태조사.

한국장애인고용촉진공단 고용개발원 (2006). 장애인고용패널조사 기초연구.

한국장애인고용촉진공단 고용개발원 (2009a). 제 1차 장애인고용패널조사.

한국장애인고용촉진공단 고용개발원 (2009b). 장애인고용패널조사 무응답 대체 및 가중치 부여방안 연구.

Taylor, H. M. and Karlin, S. (1984). *An Introduction to Stochastic Modeling*, Academic Press, Inc.

A Study on the Decision of Sample Size for Panel Survey Design

Yangsang Yoo[1] · Key-Il Shin[2]

[1]Department of Statistics, Hankuk University of Foreign Studies
[2]Department of Statistics, Hankuk University of Foreign Studies

Abstract

The transition probability can be used for the estimation of subpopulation total in panel data analysis. In this paper a real data analysis is performed and the sensitivity of the sample size allocated in the subpopulation is examined by small simulation studies.

Keywords: Panel data, transition probability, weight adjustment, benchmarking adjustment.

This research was supported by the research fund of Hankuk University of Foreign Studies(2010).
[2]Corresponding author: Professor, Department of Statistics, Hankuk University of Foreign Studies, Yongin, Kyonggi 449-791, Korea. E-mail: keyshin@hufs.ac.kr