AISUM ℚ2: 가구방문조사를 위한 표본설계 시 '1차추출단위들'을 추출하기 위한 최적의 방법은?

AISUM ℚ2: What is the best way to select primary sampling units in sample design for face-to-face household surveys?

▷질문 설명

가구방문조사(가구방문면접조사, face-to-face household surveys)는 면접원(조사원)이 1인 가구 또는 2인 이상 가구(households)에 살고 있는 '개인'을 직접 방문하여 진행하는 면접조사를 말한다. 이러한 조사를 진행하려면 '방문할 가구들'의 주소 목록, 즉 '가구표본(household sample)' 주소 목록이 있어야한다. 그런데 '가구표본'은 어떻게 얻어지는(추출되는) 것일까? 표본을 추출하는 절차를 '표본설계(sample design, survey sampling)'라고 하며, '가구표본'을 추출하기 위한 표본설계를 '가구표본설계(sample design for household surveys)'라고 한다. '가구표본설계' 시에는 일반적으로 층화다단계집락추출법(stratified multi- stage cluster sampling)이라는 표본추출법(sampling methods)이 사용되며, 이를 지역표본추출법(area sampling)이라고도 한다.

예를 들어, 전국적으로 '가구표본'을 추출할 때는 다음과 같은 층화다단계집락추출법을 사용할 수 있다. 우선 전국을 '17개 특별·광역시/도' 지역들(서울, 부산, 경기도, 강원도, 세종 등)로 구분한다(층화한다). 다음으로 각 '특별·광역시/도'별로 있는 '구/군/시'(종로구, 가평군, 수원시 등) 중 일부를 표본으로 추출하고, 추출된 각 '구/군/시'별로 '동/읍/면' 중 일부를 표본으로 추출하며, 추출된 각 '동/읍/면'별로 '통/반' 중 일부를 표본으로 추출한 뒤, 추출된 각 '통/반'별로 일부 '가구들'을 표본으로 추출한다. 이 방법을 개략도로 표현하면

아래와 같다.

① ② ③ ④
특별·광역시/도 → (표본)구/군/시 → (표본)동/읍/면 → (표본)통/반 → (표본)가구

이러한 다단계의 방법은 전국의 '모든 가구들'의 주소 목록(표본추출틀, sampling frame)이 없어도 '가구표본'을 추출할 수 있는 장점이 있으며 특히, 전국의 모든 '구/군/시'가 아니고, 표본으로 추출된 '구/군/시'에 있는 '가구표본'만을 방문하면 되므로 조사비용도 크게 절감할 수 있다. 이 방법은 층화를 한 후 4개 단계(①, ②, ③, ④)를 거쳐 최종적으로 '가구표본'을 추출하게 되므로 '층화4단계집락추출법'이라고 한다. 이 추출법의 명칭에서 '집락(clusters)'은 '구/군/시', '동/읍/면', '통/반', '가구'와 같은 단위를 일컫는 것으로서 '추출단위(sampling units)'라고도 한다.

이와 같은 층화다단계집락추출법을 사용할 때, 순서 상 가장 먼저 추출될 수 있는 '추출단위(집락)'를 '1차추출단위(primary sampling units; PSUs)'라고 한다. 앞의 예에서는 '구/군/시'(종로구, 가평군, 수원시 등)가 해당된다. '2차추출단위(secondary sampling units; SSUs)'는 '동/읍/면'이 해당되며, '3차추출단위(tertiary sampling units; TSUs)'는 '통/반'이 해당된다. '가구'는 '최종추출단위(ultimate sampling units; USUs)'에 해당된다.

이들 중 '1차추출단위'는 영문명이 "more important than anything else"의 뜻을 갖는 'primary'라는 단어로 시작되므로 표본설계 관점에서 '가장 중요한 추출단위'라는 의미가 원래 담겨있다. 그도 그럴 수밖에 없는 것이 어떤 '1차추출단위'가 추출되느냐에 따라서 뒤이어 줄줄이 추출되는 '2차' '3차' '최종' 추출단위들이 달라지기 때문에 이것은 당연한 의미이기도 하지만, 어떤 '1차추출단위들(PSUs)'이 표본으로 추출되느냐에 따라서 '추정(조사결과)의 정확성'이 높아질 수도 있고 낮아질 수도 있기 때문에 그 중요성을 부각시키기 위해 붙여진 명칭이다.

1) 그렇다면 이렇게 중요한 '1차추출단위들(PSUs)'을 표본으로 추출하기 위해서는 어떤 방법을 사용하는 것이 좋을까? '추정의 정확성'이 높은 방법이 존재할까?

2) 그리고 'model-based πPS sampling'이라는 새로운 차원(개념)의 표본추출방법론은 무엇이며, 왜 필요한 것일까?

▶나침반 보기

1)그렇다면 이렇게 중요한 '1차추출단위들(PSUs)'을 표본으로 추출하기 위해서는 어떤 방법을 사용하는 것이 좋을까? '추정의 정확성'이 높은 방법이 존재하는가?

1차추출단위들을 추출할 때는 단순임의추출법(simple random sampling)과 같이 모든 단위들이 동일한 확률로 추출되도록 하는 '동일확률추출법(equal probability of selection method)'보다는 크기비례확률추출법(probability proportional to size sampling; PPS sampling)과 같은 비동일확률추출법(unequal probability sampling method)을 사용하는 것이 바람직하다. 여기서 '크기(size)'는 각 1차추출단위(집락)의 '크기'를 말하며, 가구조사(household surveys)의 경우는 그 크기로서 '가구수(number of households)'를 흔히 사용한다.

크기비례확률추출법은 복원추출방법과 비복원추출방법으로 다시 구분된다. 복원추출방법은 영문으로 'PPS sampling with replacement,' 비복원추출방법은 'PPS sampling without replacement'이며, 각각 '복원크기비례확률추출법'과 '비복원크기비례확률추출법'으로 번역될 수 있다. 일반적으로 비복원크기비례확률추출법이 복원크기비례확률추출법보다 정확성이 높다.

국내의 통계청을 포함한 중앙정부기관들에서 진행하는 가구방문조사에서는 흔히 '조사구(평균 60가구로 구성됨)'라는 1차추출단위들을 '크기비례확률계통추출법(PPS systematic sampling)'이라는 '비복원크기비례확률추출법'을 이용하여 추출한다. 그런데 이 방법은 Madow(1949)가 개발한 것인데, 표본추출절차가 비교적 간단하여 사용하기 편리한 장점이 있지만, 일반적인 계통추출법들이 갖는 기본적인 문제(추출단위들이 나열된 '순서의 특성'에 따라 분산추정을 해야 하지만 그 '순서의 특성'을 파악하는 것이 쉽지 않음)를 마찬가지로 가지고 있어, 다른 비복원크기비례확률추출법을 사용할 필요가 있다(Cochran, 1977, Sampling Techniques, 265쪽).

이에 사용할 수 있는 비복원크기비례확률추출법은 매우 다양하며, 이미 개발된 방법들 중 적절한 방법을 선택해서 사용할 수 있다. 예를 들어, (아이디어가) 기발한 비복원크기비례확률추출법이기는 하지만, 표본추출절차가 복잡할 뿐만 아니라 '표본의 크기(추출되는 추출단위들의 개수)'가 커질수록 더욱 복잡해져 사용을 꺼려왔던 Sampford(1967)의 방법(Brewer(1963) 방법의 확장형), Murthy(1957)의 방법 등을 2000년대 들어서 SAS, SPSS, R 등과 같은 통계소프트웨어를 통해 연구자들이 쉽고 편리하게 구현할 수 있게 되었으므로 이

방법들을 사용하는 것도 바람직할 것이다. 동국대 서베이앤헬스폴리시리서치센터에서는 'Sampford (1967)의 방법'을 실제 가구표본설계에서 자주 사용해왔다. 이 방법이 처음으로 제시된 논문은 다음과 같다.

Sampford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

본 센터에서 'Sampford 방법'을 사용하게 된 주된 이유들 중 하나는, 추정의 정확성을 결정하는 절대적 기준인 '추정치들의 분산(variance of estimates)'이 복원크기비례확률추출법(PPS sampling with replacement)보다 항상 작기 때문이다. 이에 대한 이론과 증명은 다음 논문에서 확인할 수 있다.

Gabler, S. (1981). A comparison of Sampford's sampling procedure versus unequal probability sampling with replacement. *Biometrika*, 68, 725-727.

그리고 'Sampford 방법'은 매우 정교한 방법으로서, 표본설계와 관련한 많은 논문들과 저서들에서 좋은 평가와 함께 "약방의 감초"처럼 자주 언급된다. 'Sampford 방법'과 같은 비복원크기비례확률추출방법을 "πPS sampling" 또는 "inclusion probability proportional to size sampling (IPPS sampling, 크기비례포함확률추출법)"이라고도 한다. 'Sampford 방법'을 포함하여 지금까지 개발된 많은 "πPS sampling" 방법들에 대한 자세한 설명은 다음 참고문헌들에서 확인할 수 있다.

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

Brewer, K. R. W. and Hanif, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer-Verlag.

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer.

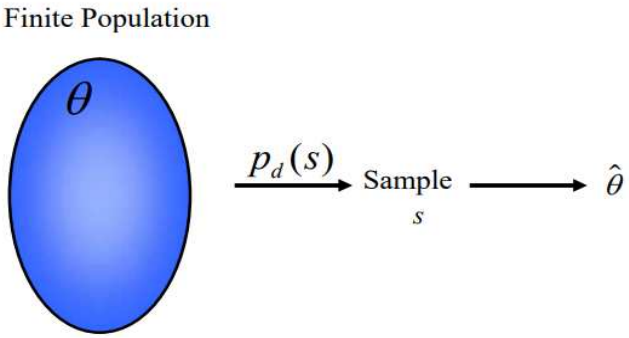
2) 그리고 'model-based πPS sampling'이라는 새로운 차원(개념)의 표본추출방법론은 무엇이며, 왜 필요한 것일까?

동국대 서베이앤헬스폴리시리서치센터는 10년간의 심혈을 기울인 이론과 실제 연구를 통해, 'model-based πPS sampling'이라는 '새로운 차원(개념)의 표본추출방법론'을 개발하였다. 이 방법론에 대한 전체적인 설명은 아래 논문(총 32쪽)에 담겨 있다(URL이 맞지 않는 경우, 센터 홈페이지 메뉴의 'Publications' → 'Technical Reports' 에서 직접 확인할 수 있다)

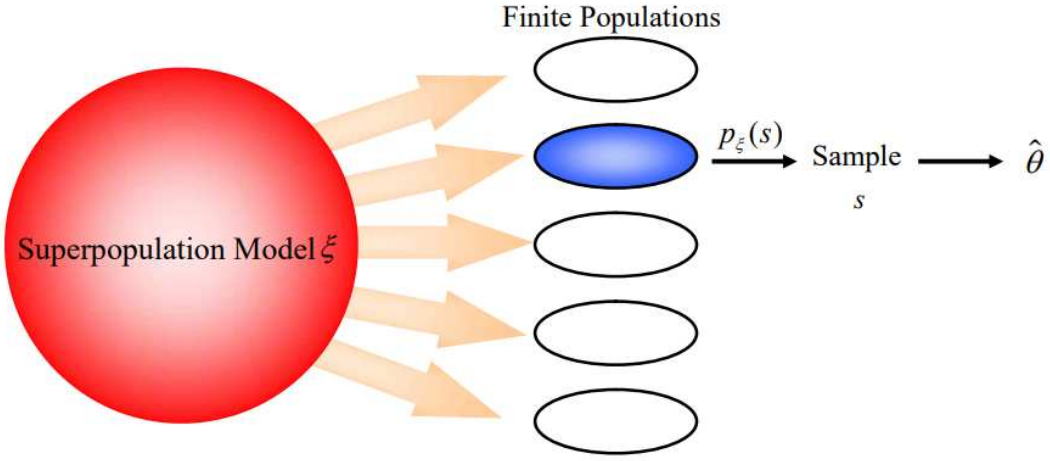
Kim, S. W., Heeringa, S. G., Hong, S. J. and Park, S. H. (2023). Some Methods of Model-Based Sampling. *Dongguk University, Survey & Health Policy Research Center Technical Report*. Available from: https://shprc.dongguk.edu/article/reports/detail/177568?pageIndex=1&

이 논문에서는 Sampford(1967)의 방법과 같은 기존의 "πPS sampling" 방법들을 'design-based πPS sampling'으로 규정하면서, 이 방법들이 근본적으로 표본추출절차(특정 알고리즘)에만 의존하기 때문에, 모집단 특성에 민감하여 '추정치들의 분산'이 일정하지 않으므로, 이런 문제점을 보완할 수 있는 'model-based πPS sampling' 방법론을 제시하였다.
'model-based πPS sampling'은 연구하고자 하는 모집단의 특성을 수리적으로 반영한 '초모집단모형(superpopulation models)'을 도입하여 '최적화 이론(optimization theory)'을 기반으로 한 메커니즘을 사용함으로써 '추정치들의 분산'을 효과적으로 최소화시킬 수 있는 표본추출방법론이다. 아울러 SAS/OR과 같은 최적화 소프트웨어를 사용하여 실제로 손쉽게 구현이 가능하다.
다음 그림들은 'design-based πPS sampling'과 'model-based πPS sampling' 표본추출방법론이 개념적으로 서로 어떻게 다른지를 보여준다. 'design-based πPS sampling'은 연구하고자 하는 "유한모집단(finite populations)"에서 표본을 추출하는 통상적인 개념인 반면에 'model-based πPS sampling'은 "유한모집단"도 '모체'인 "초모집단(superpopulation)"으로부터 추출된 하나의 표본으로 간주하여 2개 단계(초모집단 → 유한모집단 → 표본)로 표본을 추출하는 개념이다. $p_d(s)$ 또는 $p_\xi(s)$는 각 방법론에서 표본(sample)의 추출확률을 나타낸다.

Mechanism of Design-Based πPS Sampling
(One-Step Sampling)



Mechanism of Model-Based πPS Sampling
(Two-Step Sampling)



이 그림들에 있는 2가지 표본추출방법론이 실제로 구현되면 어떤 결과를 낳게 되는지는 논문의 24쪽에 있는 'Figure 2'(다음 쪽 참조)를 보면 알 수 있다. 이것은 연구 결과의 백미 중 하나로서 "model-based πPS sampling"과 기존의 "design-based πPS sampling"이 표본설계 시 실제로 '추정치들의 분산'의 최소화를 포함하여 어떤 차이가 발생하는지, 그리고 "model-based π PS sampling"이 어떤 특별한 장점이 있는지를 잘 보여준다. 'Figure 2'의 상세한 설명은 논문을 확인하기 바란다.

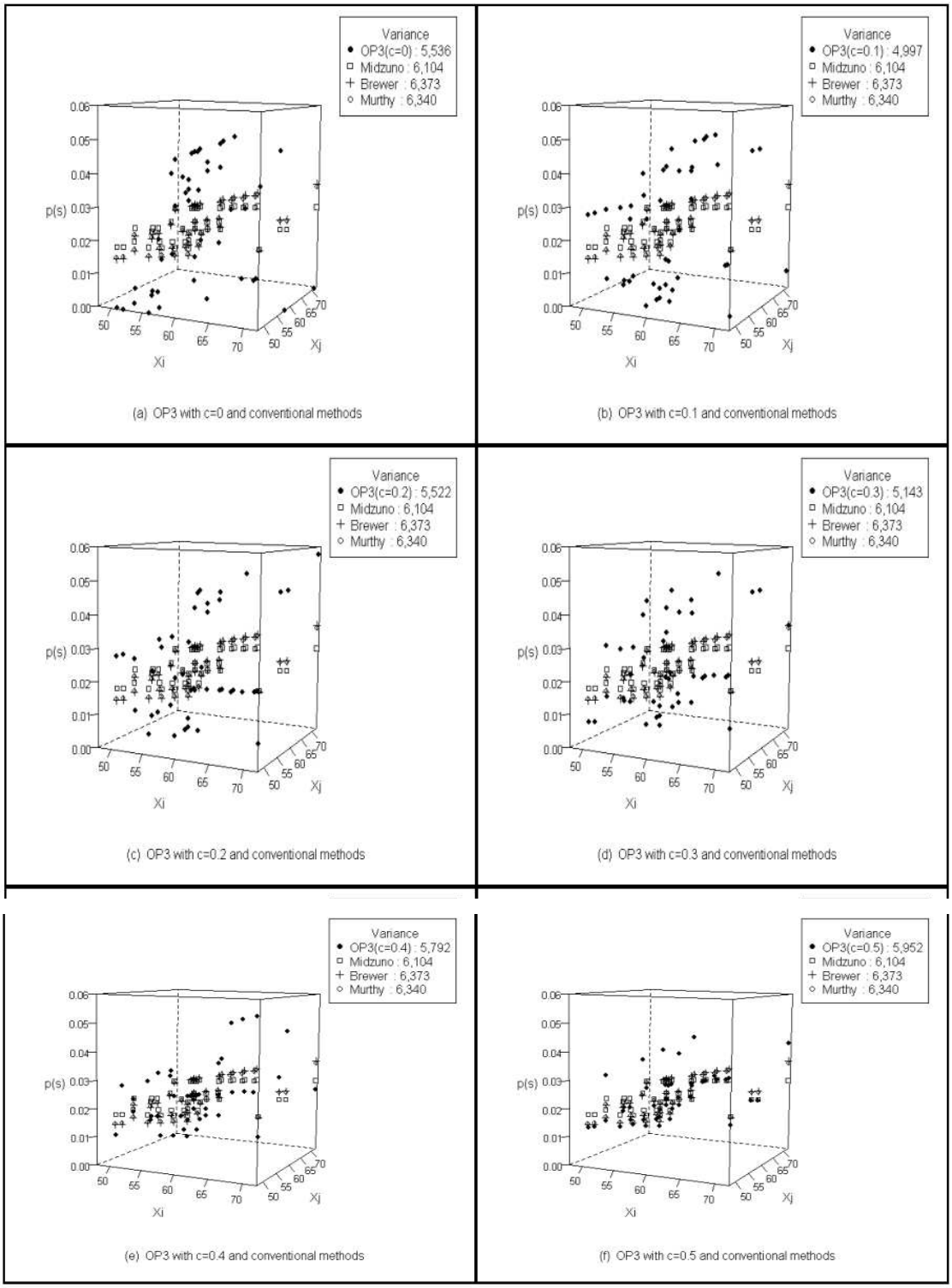


Figure 2. *Comparison of sampling designs by the values of the auxiliary variable and the corresponding variances between model-based sampling method using OP3 with a different value of c and three conventional design-based sampling methods; those obtained from conventional methods are repeatedly shown in each panel for the convenience of comparison.*

논문의 ‘Conclusion Remarks’에는 향후 연구 방향 및 과제들이 언급이 되어 있다. 후학들과 다른 연구자들이 “model-based πPS sampling”에 대한 가치와 잠재성을 잘 살펴서 연구를 더욱 진척시켜 나가기를 진심으로 바란다.

♣일화(에피소드)

대학연구기관인 본 센터에서는 자체적으로 순수 연구 목적의 많은 서베이들을 직접 수행하면서도 선임연구진들이 이론적인 연구에도 많은 시간을 할애했다. 이유는 표본설계에 사용할 수 있는 보다 실용적인 표본추출법을 찾기(선택하기) 위해서였다. 과거 초창기 1940대 출간된 것부터 최근에 이르기까지 많은 논문들과 책들을 읽고 검토를 하였다.

그런데 이들 참고문헌에서 다룬 모의실험(simulations)이나 경험적 연구(empirical studies)의 결과들을 종합해서 놓고 보면, 어떤 특정 표본추출법이 다른 방법들에 비해 명백히 탁월한 ‘추정의 정확성’을 갖는다고 결론을 내리기가 쉽지 않았다. “왜 그런 것일까?” “우리가 참고한 문헌들의 숫자가 적어서일까?” 의문이 들었다. 국내에서 구하기 어려운 논문들을 외국에서 직접 복사를 해오고 좀 더 방법론을 보강해보았지만 별다른 진전이 없었다.

그러던 중, 문득 처음부터 다시 시작해보자는 생각이 들었다. 그렇게 초창기 표본추출법 논문들부터 다시 읽기 시작하였다. 저자들이 당시에 표본추출법을 고안해낸 의도를 혹시라도 잘못 이해한 것이 있는지 확인하며, 각 논문의 서론 부분을 더욱 집중해서 읽었고, 천신만고 끝에 그 의문에 대한 실마리를 오래된 몇 편의 논문들에 나와 있는 몇 줄 되지 않는 문장들에서 찾을 수 있었다. 이 문장들에서는 “잡으려 하지만 잘 잡히지 않는 것이 있다”는 것을 나지막하게 피력하고 있었다. 그것은 한마디로 “모집단(의 특성)을 닮도록 표본을 추출하는 것”이었다. 이 실마리를 따라 계속 문헌연구를 진행하였고, 잠정적으로 내린 결론은 기존의 “design-based πPS sampling”의 메커니즘으로는 모집단(의 특성)을 닮도록 표본을 추출하는데 한계가 있다”는 것이었다. 따라서 우리는 전혀 다른 차원(개념)의 표본추출방법론이 필요하다는 것을 깨닫게 되었다. “모집단(의 특성)을 닮도록 표본을 얻기 위해” 우리는 ‘모집단’ 자체를 연구하게 되었고 “superpopulation models(초모집단모형)”에 심취하게 되었다. 이 모형은, 무한계(無限界), 즉 경계가 없는 모집단을 나타내는 모형이었고, 그 이론에 깊이 들어갈수록 새로운 체계들이 펼쳐졌다. 이 모형을 연구한 적지 않은 통계학자나 서베이방법론 학자들이 있다는 것을 알게 되었고, 이 분들의 연구

결과(대부분 ‘표본추출법’이 아닌 ‘추정’과 관련됨)를 참고하여 무한계(無限界)를 건너는데 필요한 ‘이론적 디딤돌’인 정리들(theorems)을 정립할 수 있었다. 이를 기반으로 센터에서 오랜 동안 연구하고 활용해왔던 “최적화 이론”을 적용하여 새로운 개념의 표본추출방법론을 개발하게 되었다. 이 표본추출방법론에 대한 내용은 Kim 등(2023)이 쓴 ‘Some Methods of Model-Based Sampling’ 논문에 상세히 나와 있다. 이 논문의 ‘Figure 2’는 우리가 개발한 새로운 차원의 표본추출방법론의 ‘정수(精髓)’를 보여주는 것인데 그래프들을 완성한 뒤 그 기쁨은 이루 말할 수 없었다.

‘참고문헌’으로 사용하시는 경우 다음과 같이 넣으시기 바랍니다.

동국대학교 서베이앤헬스폴리시리서치센터(2023). AISUM Q2: 가구방문조사를 위한 표본설계 시 ‘1차추출단위들’을 추출하기 위한 최적의 방법은? pp1-9. Retrieved from https://shprc.dongguk.edu/article/aisum/list