제 12회 국가통계방법론 심포지엄

비확률표본 활용방법론

2022. 9. 28 (水) 통계센터

김 규 성

(서울시립대학교 통계학과, kskim@uos.ac.kr)

서울시립대학교
UNIVERSITY OF SEOUL

차례

1. 배경
2. 비확률표집의 개념
3. 비확률표집의 유형
4. 비확률표집의 역사적 사례
5. 비확률표본을 이용한 통계적 추론
6. 확률데이터와 비확률데이터의 정확성 비교
7. 토의

1. 배경

- **조사연구의 발전 단계** (Groves, 2011)
  - 1930년~1960년 : 발명의 시대
    - 데이터 수집 설계의 구성, 조사에서 통계정보 생성 기법 발명
  - 1960년~1990년 : 확장의 시대
    - 조사방법 사용의 증가
  - 1990년~현재 : 유기적 데이터로 보완된 설계된 데이터
    - 조사참여율 감소, 데이터 수집의 대안 모드 증가
    - 표집 프레임의 약화, 디지털 생태계에서 **유기적 데이터** 생산
- **도전받는 조사** : 조사 참여 감소, 조사 비용 증가, 대체 데이터 부상

조사환경 변화에 대한 대응 (Miller, 2017)

- 응답자를 **모집**하는 방법에 대한 새로운 생각
- **커뮤니케이션 기술**을 적용하는 새로운 방법
- **조사**와 **비조사** 데이터를 혼합하는 새로운 접근 방식

- 조사데이터는 행정기록, 비정형 데이터를 포함하는 다양한 출처 데이터의 구성물이 될 수 있음

2. 비확률표집의 개념

- **표집** (sampling, 혹은 표본추출) : 모집단 전체에 대한 추론을 위하여 단위의 정보를 수집할 목적으로 모집단의 일부 단위를 선택하는 것
- 표집의 유형 **:**
  - 확률표집, 비확률표집
- **확률표집** **:**
  - 확률화 원리에 기초하여 모집단에서 단위를 선택하는 것을 말함
  - 모든 모집단 구성단위는 양수의 포함확률을 가짐
  - 랜덤 프로세스에 의하여 표본 선정
- **비확률표집** : 확률 메커니즘이 없는 표집 방법

확률표집

- 전제조건 :
  - **고정모집단(표집틀)** : 유한모집단, 모집단 구성단위 식별 가능
  - 양수의 **포함확률** : 모든 모집단 단위의 포함확률은 양수(+)임
  - **표집 메커니즘** : 랜덤 프로세스에 의한 표집
- 장점 :
  - 표본에서 **추정치**와 **표집오차**를 산출할 수 있어 모집단 추론 가능
- 단점 :
  - 비확률표집보다 더 복잡하고, 더 많은 시간과 비용이 필요

비확률표집

- **비확률표집** :
  - 모집단에서 주관적인 방법으로 표본 단위를 선택
- 장점
  - 완전한 조사프레임을 요구하지 하지 않기 때문에 데이터를 얻는 빠르고, 쉽고, 저렴한 방법임
- 단점
  - 표본에서 모집단으로 결과를 일반화할 수 있는지가 불분명
  - 표본의 대표성에 대한 강한 가정이 필요 : 종종 위험한 가정임
  - 신뢰할 수 있는 추정치와 표집오차 추정치 계산 어려움

3. 비확률표집의 유형

- 유의 선정 (Purposive Selection)
- 판단 표집(Judgement Sampling)
- 편의 표집(Convenience Sampling)
- 우발적 표집(Haphazard Sampling)
- 자원자 표집(Volunteer Sampling)
- 할당 표집(Quota Sampling)
- 네트웍 표집 (Network Sampling)
- 수정된 확률표집(Modified Probability Sampling)

유의 선정 (Purposive Selection)

- ‘대표적 표본’을 찾으려는 연구자의 판단이나 아이디어를 따르는 표집
  - 연구자가 명시적으로 기준을 제시할 수도 있음
  - 연구자가 어떤 기준을 만족할 때까지 순차적으로 추가 가능

- 전문가 선택 : 주제 분야 전문가가 단위를 선택
  - 예) 특정 특성이 가장 높은 단위를 선택 (모드 선택).
- 사례 연구 : 정성적 기법으로 단위선택 후 심층 연구
  - 전문가 선택의 극단적인 예

판단 표집(Judgement Sampling)

- 모집단 구성과 행동에 대한 기존 아이디어를 기반으로 표집
  - 모집단에 대한 지식을 가진 전문가가 단위선택을 결정
  - 즉, 전문가는 대표적 표본으로 간주되는 단위를 의도적으로 선택
- 판단 표집은 전문가의 견해에 영향을 받으며, 우발적 표집보다 훨씬 더 편향될 수 있음

편의 표집

- 잠재적 참가자를 쉽게 찾거나 모집할 수 있는 것이 주요 고려사항
- 현재 접근할 수 있는 단위를 선택하는 방식
  - 자원자 표본, 관찰연구용 표본, 눈덩이 표본
  - 이벤트 표집 : 스포츠, 음악 행사 등에서 모집
  - 쇼핑몰 앞 표집 : 쇼핑몰에서 고객에게 접근하여 조사에 모집
  - 거리 표집 : 거리에서 사람들을 조사에 참여시키는 모집
  - 가상공간 표집 : 웹 방문자를 팝업 창에서 모집
  - 사회적 유대 표집 : 친구, 동료, 지인 등을 모집
- 추론을 뒷받침하는 이론이 뚜렷하지 않음

우발적 표집(Haphazard Sampling)

- 목적 없이 임의적인 방식으로 단위 선택
  - 표집계획이 거의 또는 전혀 필요하지 않음
- 우발적 표집은 모집단이 균질하다고 가정함
  - 만일 모집단 단위들이 균질하면 랜덤 표본이 됨
- 예) 거리의 행인 인터뷰
  - 거리에서 우연히 지나가는 사람 인터뷰

자원자 표집(Volunteer Sampling)

- 자원자는 조사 목적에 적합한 특성 집합이 되도록 선별되어야 함
  - 예) 특정 질병이 있는 사람 중 자원자
- 개별단위 접촉이 분명하지 않아, 참여 결정이 응답자에 크게 의존함
- 미디어, 포스터, 전단지, 웹 등에서 참여 요청
  - 참여자 의사가 자체적으로 결정되어 회신
  - 웹 배너 모집에 자체 선정으로 참여
- 선택편향의 영향을 크게 받지만 때때로 필요한 방법임
  - 예) 윤리적인 이유로 특정 건강 상태를 가진 사람에게 일부 의료 실험이 요청되는 경우, 자원자 표본이 필요함

할당 표집(Quota Sampling)

- 비확률표집의 가장 일반적인 형태
- 모집단의 사회-인구통계적 특성을 반영한 편의 표집의 일종
  - 다양한 하위 모집단에 대하여 목표 표본 크기 정함
  - 각 하위 모집단에서 목표에 만족하도록 단위를 선택
- 모집단에 대한 추론
  - 각 하위 모집단에서 표본 단위와 나머지가 유사하다는 가정 필요
  - 그러나 이러한 강한 가정은 거의 타당하지 않음

네트웍 표집 (Network Sampling)

- 목표모집단이 매우 희소하여 원하는 표본수 표집을 위한 확률표집은 너무 많은 비용이 들어가는 경우에 대한 대안적 방법
- 조사 대상이 사회적으로 연결된 하위 모집단일 때 유용
  - 초기에 소수의 적격한 표본단위 식별, 이후 충분한 신뢰 구축
  - 첫 번째 응답자 네트웍을 통해 다른 적격한 단위에 연결
  - 원하는 표본 크기 달성될 때까지 순차적 연결

수정된 확률표집(Modified Probability Sampling)

- 확률표집과 비확률표집의 조합. 일반적으로,
  - 초기 단계는 확률표집 기반
  - 마지막 단계는 비확률표집, 보통 할당표집 사용
- 예)
  - 지리적 영역 : 확률 표집설계 사용하여 선정
  - 선정된 표본 영역에서는 개인의 할당표본 선택

확률표집과 비확률표집의 혼합

- 확률표집 과정에서 개입되는 비확률성
  - 조사모집단의 일부가 생략되는 경우

    ⇒ 해당 단위의 포함확률은 0이 됨
  - 낮은 응답률

    ⇒ 무응답 단위의 조사 참여율은 0이 됨
- 비확률표집 과정에서 확률성 부여
  - 포함범위가 넓고, 응답률이 높으며, 특히 목표모집단 규모가 작을 때 (예: 이벤트 참석자, 기관 방문자 등)

    ⇒ 비확률표집 유형 중 일부는 확률표집으로 전환 가능

4. 비확률표집의 역사적 사례

- 그랜트의 런던인구 추정(1662년)
- 라플라스의 프랑스 인구 추정(1812년)
- 키에르의 노르웨이 공식통계(1897년, 키에르)
  - 표본을 사용한 모집단 전체 추론
  - ‘대표적인 표본’의 개념
  - 대다수 통계전문가는 표본 개념을 거부
  - 표본조사와 전수조사의 병행
- 대표적인 표본 선정 방법 채택 (1925년, ISI, Bowley)
  - 랜덤 표집, 유의 선정 병행

비확률표집의 역사적 사례

- 지니와 갤바니 (1929년) : 유의선정
- 확률표집 방법의 옹호 (1934년 네이만)
- 미국 대통령 선거
  - 1936년 : 리터러리 다이제스트 조사
  - 1948년 : 갤럽 조사
  - 확률표집 이론으로 비확률표집 결과를 설명
- 모형기반 추론 (1970년, Royall) : 베스트 추정량
- 지니와 갤바니의 후예들
  - 큐브 방법 (2004년, Deville & Tille) : 균형 표집

5. 비확률표본을 이용한 통계적 추론

- **통계적 추론** (Baker 등 2013)
  - 목표모집단의 특성에 대한 추정치를 산출하고, 추정치의 신뢰성에 대한 측정치를 제공하는 일련의 절차
- 통계적 추론의 특징
  - 추정을 하고, 그 추정의 정확성을 판단하기 위해 이론적 근거와 명확한 가정이 필요
  - 이론적 근거 없는 데이터 수집과 추정은 통계적 추론에 부적절
- 확률표집과 비확률표집은 통계적 추론 방식에서 차이를 보임

확률표집의 정당성

- 확률표집의 정당성:
  - 수학적 원리(Fisher 1925; Neyman 1934; Sarndal 외 1992)에 기초한 확률표집 이론
  - 추정의 정확성 계산(신뢰구간 또는 오차한계 등) 가능, 추정 방법의 보편적 타당성을 제공
- 확률표본 조사는 데이터 생성 과정의 세부 사항을 제공하므로, 모집단 포함범위, 표본추출 및 무응답 편향의 사후 조정 가능

비확률표집의 정당성

- 비확률표집의 정당성의 근거 (Cornesse 외 2020):
  - 표본과 나머지 부분의 특성을 설명하는 모형 가정에 기초함
  - 이러한 모형 가정은 일반적으로 검증이 불가능
  - 모형과 연구 주제와의 관련성은 연구원의 믿음에 기초함
- 모형 가정의 형태(Deville 1991; Elliott & Valliant 2017) :
  - 유사 랜덤화(quasi-randomization)
  - 초모집단 모형화(superpopulation modeling)
  - 이중 강건 추정(doubly robust estimation, Yang & Kim 2020)
- 비확률표집을 정당화하는 일반적인 통계 이론은 없음

비확률표집 정당성의 개념화

- 비확률표집의 정당화는 네 가지 유형의 주장에서 비롯될 수 있음 (Cornesse 외 2020):
  - (i) 특정 질문을 조사하는 모든 표본은 동일한 추론을 산출
  - (ii) 질문과 연관된 특정 표본설계는 관심모집단을 반영하는 결론을 생성
  - (iii) 일련의 분석 단계는 표본과 모집단 간의 차이를 설명
  - (iv) 표본과 분석방법의 특정 조합은 정확한 모집단 추정치를 생성

- 따라서 특정 연구질문에 특정 방법이 적절한지 여부는 (i) 관심 질문, (ii) 표본, (iii) 사용된 조정 절차에 대한 근거 주장에 따라 다름

유사 랜덤화 (Valliant 2020)

- 표본을 확률 메커니즘을 통해 얻은 것처럼 간주하지만, 실제 해당 메커니즘은 알 수 없음
- 표본에 포함될 의사(pseudo)-추출 확률은 표본 데이터와 외부 데이터를 함께 사용하여 추정됨

- 보정 가중치 (calibration weights)
- 성향점수 가중치 (Propensity Score Weights)

보정 가중치 (Calibration Weights)

- **보정 가중치** :
  - 단위 수준의 참조 데이터 사용
  - 어떤 특성의 표본 가중합이 동일 특성의 모집단 총계에 일치하도록 가중치 산출
- 해당 특성의 모집단 총계는 센서스 데이터, 공식통계 또는 고품질로 간주되는 다른 데이터 소스에서 나옴

- **기초 가정** : 알려진 편차의 원인이 조정되면 나머지 오차는 모집단 특정 단위의 조사 참여 여부와 무관하다는 것

성향점수 가중치 (Propensity Score Weights)

- **성향점수 가중치** :
  - 비확률표본의 개별 단위에서 조사데이터 수집 후에 수행
- 기본 절차
  - 비확률표본 데이터를 참조 데이터세트, 일반적으로 대규모 확률표본조사 데이터와 병합
  - 두 데이터 세트의 변수를 사용하여 모형(로짓 또는 프로빗)을 적합한 후 개별단위가 비확률표본에 속할 확률을 예측(Rosenbaum and Rubin 1983, 1984; Lee 2006 등).
- 추정된 포함확률의 역수로 **성향점수 가중치** 구성

의사-설계기반 추정 (Pseudo design-based Est.)

- **의사 설계기반 추정** (Elliott 2009; Elliott and Valliant 2017)
  - 비확률표본과 확률표본을 결합
  - 공통변수들을 기반으로 비확률표본 단위의 유사 포함확률 추정
- 포함확률 추정 방법
  - (예측) 비확률 및 확률 데이터 연결하여 비확률조사 참여 확률 예측
  - (매칭 및 대체) 두 조사 데이터를 매칭하여 확률표본 단위의 포함확률을 비확률표본의 가장 가까운 단위에 기부
- 의사 포함확률이 모든 비확률표본 단위에 할당되면 확률표집된 것으로 간주하여 분석 처리
- 설계기반 재표집 분산추정법

초모집단 모형화 및 이중 강건 추정

- 분석변수의 관찰값을 어떤 확률모형에 의해서 생성된 것으로 간주
- 모형은 표본에서 추정되며, 외부 모집단 통제 데이터와 함께 표본을 모집단에 투영하는데 사용됨
- 유사 랜덤화와 초모집단 모형화가 결합된 경우 이중 강건 추정이 됨

6. 확률데이터와 비확률데이터의 정확성 비교

- 확률 및 비확률 표본조사의 경험적 정확성 평가 (Cornesse 외 2020)
- 많은 연구에서 확률조사가 비확률조사보다 정확도가 더 높다고 결론
  - 투표 행동(Chang and Krosnick 2009; Sturgis et al. 2018)
  - 건강 행동(Yeager, Krosnick, Chang, Javitz, Levendusky 2011)
  - 소비 행동(Szolnoki 및 Hoffmann 2013
  - 성적 행동 및 태도(Legleye 외 2018)
  - 사회 인구 통계(Malickhotra 외 2009; Yeager 등 2011)
- 서로 다른 주제의 서로 다른 시대와 국가의 모든 연구는 확률표본 조사가 비확률표본조사보다 더 정확한 추정을 이끌어낸다고 결론

Table 1. Studies on the Accuracy of Probability (PS) and Nonprobability (NPS) Sample Surveys

Study	Country	Benchmark	PS modes studied	PS more accurate than NPS?[a]	Did (re)weighting sufficiently reduce NPS bias?[b]
Blom et al. (2018)	Germany	Census data Election data	F2F, web	Yes (univariate)	No (raking)
MacInnis et al. (2018)	USA	High-quality PS	Phone, web	Yes (univariate)	No (poststratification)
Dassonneville et al. (2018)	Belgium	Census data	F2F	Yes (univariate)	No (unspecified approach) – univariate
		Election outcome		No (multivariate)	N/A – multivariate
Legleye et al. (2018)	France	Census data	Phone	Yes (univariate)	N/A
Pennay et al. (2018)	Australia	High-quality PS	Phone	Yes (univariate)	No (poststratification)
Sturgis et al. (2018)	UK	Election outcome	F2F	Yes (univariate)	No (raking, propensity weighting, matching)
Dutwin and Buskirk (2017)	USA	High-quality PS	F2F, phone	Yes (univariate)	No (propensity weighting, matching, raking)
Sohlberg et al. (2017)	Sweden	Election outcome	Phone	Yes (univariate)	N/A
Brüggen et al. (2016)	Netherlands	Population register	F2F, web	Yes (univariate) Yes (bivariate)	No (Generalized Regression Estimation) – univariate, bivariate
Kennedy et al. (2016)	USA	High-quality PS	Web, mail	No (univariate) No (multivariate)	N/A – univariate N/A – multivariate
Pasek (2016)	USA	High-quality PS	Phone	Yes (univariate) No (bivariate) Yes (longitudinal)	No (raking, propensity weighting) – univariate, bivariate, longitudinal

Table 1. *Continued*

Study	Country	Benchmark	PS modes studied	PS more accurate than NPS?[a]	Did (re)weighting sufficiently reduce NPS bias?[b]
Gittelman et al. (2015)	USA	High-quality PS	Phone	No (univariate)	No (poststratification)
Ansolabehere and Schaffner (2014)	USA	High-quality PS Election outcome	Phone, mail	No (univariate) No (multivariate)	N/A
Erens et al. (2014)	UK	High-quality PS	CASI	Yes (univariate)	N/A
Steinmetz et al. (2014)	Netherlands	High-quality PS	Web	No (univariate) No (bivariate)	Yes (propensity weighting) – univariate, bivariate
Ansolabehere and Rivers (2013)	USA	High-quality PS Election outcome	F2F	No (univariate)	N/A
Szolnoki and Hoffmann (2013)	Germany	High-quality PS	F2F, phone	Yes (univariate)	N/A
Chan and Ambrose (2011)	Canada	Unspecified	Web	No (univariate)	N/A
Scherpenzeel and Bethlehem (2011)	Netherlands	Election outcome Unspecified	F2F, web	Yes (univariate)	N/A
Yeager et al. (2011)	USA	High-quality PS	Phone, web	Yes (univariate)	No (poststratification)
Chang and Krosnick (2009)	USA	High-quality PS	Phone, web	Yes (univariate)	No (raking)
Walker, Pettit, and Rubinson (2009)[c]	USA	Unspecified	Phone, mail	Yes (univariate)	N/A
Loosveldt and Sonck (2008)	Belgium	Census data	F2F	No (univariate)	No (poststratification, propensity weighting)

정확성 평가 : 표집효과와 모드효과

- 표집효과(선택 편향)에서 모드 효과(측정 편향)의 혼재 가능성
  - 확률 표본조사 : 오프라인(대면, 전화 조사) 수행
  - 비확률 표본조사 : 비확률 온라인 패널을 통해 수행
- 표집효과와 모드 효과 분리 가능
  - 오프라인 확률조사 v.s. 온라인 확률조사 (모드효과) 비교
  - 온라인 확률조사 v.s. 온라인 비확률조사 (표집효과) 비교
- 대부분은 확률조사가 비확률조사보다 더 정확하다는 결론

  (Chang and Krosnick 2009; Scherpenzeel and Bethlehem 2011; Blom et al. 2018, MacInnis et al. 2018).

7. 토의

- 확률표집 이론은 잘 정립된 수학적 원리를 기반으로 하지만, 비확률표집은 단일 프레임워크가 없음
- 비확률표본에서 추론 도출은 잠재적 정당성이 있지만, 근거가 명확하지 않은 경우가 많으며, 강력한 모형 가정이 필요함
- 비확률표본을 사용하는 경우 기본 가정을 기반으로 다양한 모형 접근 방식 중에서 신중하게 선택할 필요 있음
- 실증적 증거는 응답률이 감소하는 시대에도 확률조사의 정확도가 비확률조사보다 일반적으로 높음을 보여줌

공식통계 관련 토의 (Valliant 2022)

- (공식통계) 변화하는 데이터 수요에 신속한 대응이 요구됨
- (공식통계) 새로운 데이터 수집 방법에 적응할 필요 있음
- (공식통계) 데이터 통합 관련한 품질 문제
  - 통합 데이터는 조사데이터와 동일한 품질인가?
  - 통합 데이터의 모집단은 조사데이터 모집단과 동일한가?
  - 행정데이터의 단위는 조사데이터의 단위와 정의가 동일한가?
- (공식통계) 사회가 용인하는 수준의 기밀 및 노출 보호