

<http://sri.kostat.go.kr>

비확률표본을 위한 통계적 추론: 실증연구

2023. 8.



통계청
통계개발원



목 차

0. 수행개요	1
I. 서론	3
II. 추정방안	5
III. 모의실험	12
IV. 결론	22

□ 과제개요

- (과 제 명) 비확률표본을 위한 통계적 추론: 실증연구
- (연 구 자) 통계방법연구실 권순필 사무관, 권영미 주무관
- (과제구분) 자체과제
- (수행기간) 23.5월 ~ 12월 (8개월)

연구내용 \ 추진일정	2023년							
	5	6	7	8	9	10	11	12
• 연구계획 수립	■							
• 선행연구 검토 및 모의실험 설정		■	■	■				
• 중간자문				■				
• 시나리오별 모의실험 수행					■	■	■	
• 최종보고 및 보고서 완료							■	■

□ 특이사항

- 2022년 “비확률표본을 위한 통계적 추론” 후속연구

□ 연구목차(안)

제1장 서론

1. 연구배경
2. 연구내용

제2장 추정방안

1. 설정과 가정
2. 성향점수 가중치 방식(Propensity Score Weighting Method)
3. 보정 가중치 방식(Calibration Weighting Method)
4. 통대체 방식(Mass Imputation Method)
5. 이중강건 방식(Doubly Robust Method)

제3장 모의실험

1. 데이터 및 모의실험 설명
2. 모의실험 결과
 - 시나리오1 (확률표본 vs 비확률표본 설계 변경)
 - 시나리오2 (확률표본 vs 비확률표본 크기 변경)
 - 시나리오3 (관심변수가 이산형인 경우)

제4장 결론

1. 요약
2. 시사점 및 결론

□ 연구배경

○ 확률표본의 위기

- 확률표본이란 잘 알려진 확률표집이론에 따라 선정된 표본
- **확률표집은 고유한 이론적 체계 안에서 표본의 대표성과 결과의 객관성 확보 가능** → 공식통계 생산기관 선호
- 잘 정비된 표본추출틀, 정교한 표집설계, 표집설계에 의한 표본 추출, 완전한 응답이 전제되므로 고비용
- 최근 표본추출틀의 포함범위 감소, 무응답 증가, 조사비용의 급격한 증가로 확률표본의 선택과 유지 어려움

○ 비확률표본의 기회 및 도전

- 비확률표본이란 확률표본이 아닌 표본 혹은 데이터
- 한계에 봉착한 확률표본의 대안으로 저비용, 낮은 응답부담에 실시간으로 쏟아지는 대량의 비확률표본 활용 요구 증가
- 그러나 확률표집과 달리 **모든 비확률표집을 적절하게 포함하는 단일 프레임워크는 없으며**, 비확률표본의 선택편향, 과소포함, 미지의 추출확률 등 때문에 확률표집이론의 직접 적용 불가능
 - 비확률표본을 단순임의표본인 것처럼 다루는 경우 심각한 표본 선택편향 초래

○ 방법론에 대한 체계적인 검토 수행

- 2022년에 “비확률표본을 위한 통계적 추론” 연구를 통해 유한 모집단 추론에서 비확률표본이 재조명받게 된 이유와 모형기반 접근 등 방법론에 대한 종합적인 검토 수행

- 어떤 조건 하에서 비확률표본을 확률표본의 대안으로 사용할 수 있는지 확인

○ 실제 자료 적용을 통한 실증연구 수행 필요

- 비확률표집을 적절하게 포함하는 단일 프레임워크가 없기 때문에 경험적 연구에 의존할 수 밖에 없으며 이를 위해 실제 자료 적용을 통한 실증연구 필요

□ 연구내용

○ 비확률표본 추정 방안 검토

- 설정과 가정
- 성향점수 가중치 방식 (Propensity Score Weighting Method)
- 보정 가중치 방식 (Calibration Weighting Method)
- 통대체 방식 (Mass Imputation Method)
- 이중강건 방식 (Doubly Robust Method)

○ 가계금융복지조사 자료*를 이용한 모의실험을 통해 실증연구 수행

- 확률표본과 비확률표본의 상대적 크기 변화, 확률표본의 설계 변화 등 다양한 시나리오를 가정하여 비확률표본의 평균 및 중위수와 관련 신뢰구간 추정

* 통계청 조사 중 연속형 변수를 가지면서 표본규모가 가장 큼

□ 연구목적

- 비확률표본을 실무 수준에서 쉽게 접근할 수 있도록 이론 및 방법론, 경험적 연구 결과, 한계 등을 체계적으로 검토·정리하여, 국가통계 전문 연구기관으로서 통계생산 패러다임 변화 요구에 선제적으로 대응

□ 확률표본 vs 비확률표본

○ 확률표본

- 1) 표본추출틀의 조사단위는 확률적으로 선정 &
- 2) 모든 조사단위의 포함확률은 양수 &
- 3) 포함확률은 계산 가능해야 함.

○ 비확률표본

- 1) 표본추출틀의 조사단위가 확률적으로 선정되지 않음 or
 - 2) 일부 조사단위의 포함확률이 0 or
 - 3) 포함확률 계산이 불가능한 경우
- 1)선택편향, 2)과소포함, 3)미지의 추출확률로 표현 가능

□ 접근 방식

○ 비확률표본을 확률표본처럼 취급하여 확률표본 추론 프레임워크 사용

○ 무시가능성(ignorability)¹⁾ 혹은 무작위결측(Missing At Random; MAR)을 가정하면, 선택확률과 과소포함 문제가 해결되고 미지의 추출확률만 모형을 통한 계산 필요 => 평균제곱오차 (MSE)를 통해 품질평가 가능

$$- P(\delta_i = 1 | x_i, y_i) = P(\delta_i = 1 | x_i), \quad \pi_i^V(x_i) \equiv P(\delta_i = 1 | x_i) > 0, \quad \forall x_i$$

여기서, δ_i 는 비확률표본에 포함되면 1, 그렇지 않으면 0인 비확률표본 포함지시자, y_i , x_i 는 각각 관심변수와 보조변수, V 는 비확률표본

1) Rosenbaum & Rubin (1983)

□ 데이터 가정

○ 데이터 가정

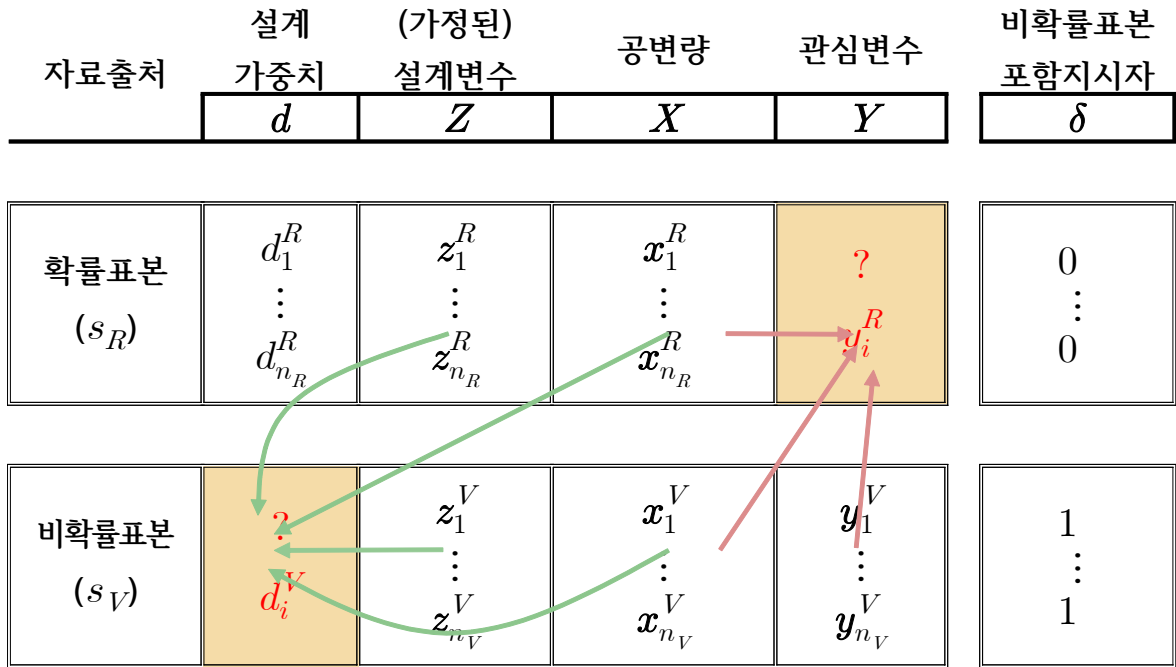
- 동일한 모집단을 대표하는 확률표본과 비확률표본이 존재하며, 두 표본은 서로 독립이고 자료 간 측정오차는 없음.
- 관심변수는 비확률표본에서만 관측되고, 비확률표본과 공통의 공변량을 갖는 유용한 확률표본이 존재

○ 표기

- $U = \{1, 2, \dots, N\}$: 크기 N 인 모집단
- y_i : 관심값(반응변수), x_i : 공변량(보조변수), $i = 1, 2, \dots, N$
- z_i : 설계변수(불균등 확률표집, 층화, 군집 등)
- $\mu_y = N^{-1} \sum_{i=1}^N y_i$: 반응변수의 유한모집단 평균
- Md_y : 유한모집단 반응변수의 중위수
- s_V : 크기 n_V 로 $\{(x_i, y_i), i \in s_V\}$ 데이터셋을 갖는 비확률표본
- s_R : 크기 n_R 로 $\{(x_i, d_i^R), i \in s_R\}$ 데이터셋을 갖는 확률표본
 - $d_i^R = 1/\pi_i^R$, π_i^R 는 S_r 의 포함확률
- δ_i : 비확률표본 포함지시자 $\begin{cases} \delta_i = 1, & \text{if } i \in s_V \\ \delta_i = 0, & \text{if } i \notin s_V \end{cases}, i = 1, 2, \dots, N$

□ 데이터 통합

- 비확률표본의 추출확률 계산을 위해서는 보조변수 x 를 갖는 신뢰할 만한 참조확률표본(reference probability sample)과 비확률표본을 연결하는 모형 필수
- 비확률표본 추론 접근 방식의 도식적 표현



주. ——— 가중치 방식, ——— 통대체 방식

[그림 1] 비확률표본 추론 접근 방식의 도식적 표현

- 가중치 방식은 비확률표본 s_V 의 미지의 추출확률 $\hat{\pi}_i^V$ 를 추정하여 비확률표본 s_V 를 확률표본인 것처럼 추정
- 통대체 방식은 확률표본 s_R 의 y_i^R 을 모두 결측으로 가정하고 추정치 \hat{y}_i^R 로 관심값을 통째로 대체하여 확률표본 s_R 의 관심값이 관측된 것처럼 추정
- 이중강건 방식은 가중치 방식과 통대체 방식의 결합을 통해 각 방식의 모형 오지정(model misspecification)에 로버스트하게 대응하는 추정 방식

□ 성향점수 가중치 방식 (Propensity Score Weighting Method)

○ 성향점수모형을 이용하여 비확률표본에 포함될 확률 π_i^V 를 추정

$$- \pi_i^V = P(\delta_i = 1 | \mathbf{x}_i, y_i) = P(\delta_i = 1 | \mathbf{x}_i), \text{ 여기서, } \begin{cases} \delta_i = 1, & \text{if } i \in s_V \\ \delta_i = 0, & \text{if } i \notin s_V \end{cases}$$

○ 대표적인 성향점수 모형은 로지스틱회귀모형

$$- \pi_i^V = \pi(\mathbf{x}_i, \boldsymbol{\theta}_0) = \exp(\mathbf{x}_i^T \boldsymbol{\theta}_0) / \{1 + \exp(\mathbf{x}_i^T \boldsymbol{\theta}_0)\}$$

○ 로지스틱회귀모형의 $\boldsymbol{\theta}$ 는 의사로그우도함수(pseudo-log-likelihood)

$l^*(\boldsymbol{\theta})$ 를 최대화 하는 $\hat{\boldsymbol{\theta}}$ 으로 추정 (Chen et al ((2020)

$$- l^*(\boldsymbol{\theta}) = \sum_{i \in S_V} \mathbf{x}_i^T \boldsymbol{\theta} - \sum_{i \in S_R} d_i^R \log\{1 + \exp(\mathbf{x}_i^T \boldsymbol{\theta})\}$$

$$- U(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l^*(\boldsymbol{\theta}) = \sum_{i \in S_V} \mathbf{x}_i - \sum_{i \in S_R} d_i^R \pi(\mathbf{x}_i, \boldsymbol{\theta}) \mathbf{x}_i$$

- 점수방정식 $U(\boldsymbol{\theta})$ 의 해는 Newton-Raphson 반복 절차로 찾음

초기값은 $\boldsymbol{\theta}^{(0)} = \mathbf{0}$

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + \{H(\boldsymbol{\theta}^{(m)})\}^{-1} U(\boldsymbol{\theta}^{(m)})$$

$$H(\boldsymbol{\theta}) = \sum_{i \in S_R} d_i^R \pi(\mathbf{x}_i, \boldsymbol{\theta}) \{1 - \pi(\mathbf{x}_i, \boldsymbol{\theta})\} \mathbf{x}_i \mathbf{x}_i^T$$

○ 특징

- 추정된 성향점수가 표본의 포함확률과 같은 역할을 하기 때문에
관심결과에 상관없이 모든 후속 분석에 적용할 수 있는 단일
조정 모형 생성

- 분위수와 같은 다른 유한모집단 모수로 확장이 간단하며, 관심
값이 바뀌어도 동일한 모형 이용

- 성향점수 모형이 잘못 지정된 경우, 특히 특정 단위가 $\hat{\pi}_i^V$ 에서
매우 작은 값을 가질 때 민감

□ 캘리브레이션 가중치 방식 (Calibration Weighting Method)

- 알려진 보조변수 총계에 맞도록 비확률표본의 가중치 d_i^V 를 직접 조정하는 방식

$$- \sum_{s_V} d_i^V x_i = \sum_{s_R} d_i^R x_i$$

- 캘리브레이션은 거리함수에 따라 다양한 형태를 지님.

유형	거리함수
GREG	$\frac{(x-1)^2}{2}, x \in (-\infty, \infty)$
raking ratio	$x \ln(x) - x + 1$, 만약 $x \in (0, \infty)$ $-x + 1$, 만약 $x = 0$
truncated linear	$\frac{(x-1)^2}{2}, x \in (-\infty, \infty), x \in (L, U), 0 \leq L < 1 < U$
logit	$[(x-L)\ln\frac{x-L}{1-L} + (U-x)\ln\frac{U-x}{U-1}]A^{-1}$, 만약 $x \in (L, U)$ $[(U-L)\ln\frac{U-L}{U-1}]A^{-1}$, 만약 $x \leq L$ $[(x-L)\ln\frac{x-L}{1-L} + (U-x)\ln\frac{U-x}{U-1}]A^{-1}$, 만약 $x \geq U, 0 \leq L < 1 < U$

출처 : Déville & Särndal (1992)

○ 특징

- 적용이 쉬워 비확률표본 추정 시 가장 자주 사용되는 방식으로 범주형 보조변수를 사용하는 Raking ratio 활용 빈도가 높음
- 캘리브레이션 방식은 설계가중치가 있는 표본을 사후적으로 보정하는데 주로 이용되는 방식으로, 가능한 원래의 설계가중치 π_i^{-1} 에 가깝게 설정되어야 하며 제약식도 만족해야 하기 때문에 Newton-Raphson 같은 반복 알고리즘이 필요할 수 있음

□ 통대체 방식 (Propensity Score Weighting Method)

- 표본과 비표본(non-sample)이 모두 동일한 모형을 따른다는 가정 하에, 비확률표본을 훈련 데이터로 사용하여 확률표본의 관심변수를 추정하는 방식으로 예측모형(Prediction model) 접근법이라고도 함
 - 가정 : $\{(\mathbf{x}_i, y_i), i \in U\}$ 는 모형 $y_i = m(\mathbf{x}_i) + \epsilon_i$, $i = 1, 2, \dots, N$ 의 무작위 표본
 - 예측모형 : $m(\mathbf{x}_i) = E_\xi(y_i | \mathbf{x}_i)$, ϵ_i 의 $E_\xi(\epsilon_i) = 0$, $V_\xi(\epsilon_i) = v(\mathbf{x}_i)\sigma^2$
- 예측모형의 대표적인 모형은 회귀모형
 - $m(\mathbf{x}_i) = \mathbf{x}_i^T \beta$ (즉, $\hat{y}_i = \mathbf{x}_i^T \hat{\beta}$), 여기서 $v(\mathbf{x}_i) = 1$
 - $\hat{\beta}$ 는 최소제곱법, 최대가능도추정 등으로 추정
- 특징
 - 특정 관심변수 y_i 에 대한 예측모형이기 때문에 다른 관심변수에 대해서는 다른 모형 설정 필요하여 후속 분석 어려움
 - 실제로 y 를 잘 설명하는 x 의 확보는 어려운 경우가 다수
 - 모형 지정이 잘 되는 경우에는 분산이 작아짐

□ 이중강건 방식 (Doubly Robust Method)

- 성향점수를 가중치의 역수로 이용하는 역확률가중(Inverse Probability Weighting; IPW) 추정량은 잘못 지정된 성향모형에 민감하기 때문에 예측모형을 통합하여 추정량의 효율성과 견고성 향상 가능 => 증강역확률가중(Augmented IPW; AIPW) 추정량

- AIPW 추정량은 두 모형 중 하나 이상의 모형이 올바르게 지정되면 일치추정량이기 때문에 이중강건 추정량이라고도 함

- 평균의 이중강건 추정량

$$\hat{\mu}_{DR} = \frac{1}{N} \sum_{i=1}^N \frac{R_i \{y_i - m(x_i, \hat{\beta})\}}{\pi(x_i, \hat{\theta})} + \frac{1}{N} \sum_{i=1}^N m(x_i, \hat{\beta}),$$

여기서 $\hat{\beta}$, $\hat{\theta}$ 는 각각 예측모형과 성향점수모형의 일치추정량

- 예측모형이 잘 지정되는 경우 일반적으로 분산이 작아지나
- 추론 과정에서 필요한 많은 가정과 모형 타당성에 대한 검증이 어려워 실제 적용은 어려움

III

모의실험

□ 모의실험 자료

○ 모집단 : 2021년 가계금융복지조사 가구마스터 자료²⁾

- $N = 18,187$ 가구
- 관심변수 Y : 연간가구경상소득
- 공변량(보조변수) : X

공변량	범주	내용
1.수도권여부 (geo)	2	수도권, 비수도권
2.가구주성별 (gender)	2	남성, 여성
3-1. 가구주연령 (age)		
3-2. 가구주연령 (age_g)	6	20대이하, 30대, 40대, 50대, 60대, 70대이상
4.가구주교육정도 (edu)	4	초졸이하, 중졸이하, 고졸이하, 대졸이상
5.혼인여부 (mrg)	4	미혼, 유배우자, 사별, 이혼
6.종사상지위 (stt)	4	상용임금, 임시일용, 자영업자, 기타
7. 가구원수		
8.연간가구소비지출 (exp1)		
9.연간가구비소비지출 (epx2)		

□ 시나리오

- 관심추정량 : 평균, 중위수
- 추정방안 : 성향점수, 캘리브레이션, 회귀모형, 이중강건 추정
 - 분산은 붓스트랩 추정

²⁾ 모의실험 설정이 잘 완료되면, 2022년 가계금융복지조사 가구 및 가구원 자료에 적용 예정

○ 평가 : 추정량 $\hat{\theta}$ 및 분산추정량 \hat{v} 에 대한 MCMC를 이용한 상대편향, 평균제곱오차, 포함확률

- 상대편향 $\%RB = \frac{1}{M} \sum_{m=1}^M \frac{\hat{\theta}^{(m)} - \theta}{\theta} \times 100$

- 평균제곱오차 $MSE = \frac{1}{M} \sum_{m=1}^M (\hat{\theta}^{(m)} - \theta)^2$

- 신뢰구간 포함확률 $\%CP = \frac{1}{M} \sum_{m=1}^M I(\mu_y \in CI^{(m)}) \times 100$

여기서, $CI^{(m)}$ 은 $[\hat{\theta}^{(m)} - 1.96 \times \sqrt{\hat{v}^{(m)}}, \hat{\theta}^{(m)} + 1.96 \times \sqrt{\hat{v}^{(m)}}]$

○ 검토사항

- 확률표본(S_R) : SRS, 층화, PPS 설정

• (가설) 동일한 규모의 표본이라면 보다 정밀한 설계에 기반한 확률표본을 참조하는 것이 추정의 효율을 개선시킬 것

- 비확률표본(S_V) : 선택편향, 과소포함 등 경우 설정

- 표본크기 : $n_R \gg n_V, n_R \approx n_V, n_R \ll n_V$

• (가설) 확률표본이든 비확률표본이든 규모가 클수록 추정의 효율이 개선될 것

- 변수형태 : 관심변수, 독립변수가 연속형 vs 이산형

□ 추정량

○ 평균

- 설계기반 $\hat{\mu}_R = \hat{N}_R^{-1} \sum_{i \in s_R} \frac{1}{\pi_i} y_i$,
- 단순(naive) $\hat{\mu}_{naive} = n_V^{-1} \sum_{i \in s_V} y_i$
- 성향점수 역확률가중(IPW) $\hat{\mu}_{ipw} = \hat{N}_V^{-1} \sum_{i \in s_V} \frac{1}{\pi_i} y_i$,
- 통대체(MI) $\hat{\mu}_{reg} = \hat{N}_R^{-1} \sum_{i \in s_R} \frac{1}{\pi_i} \hat{y}_i$
- 이중강건(DR)

$$\begin{aligned} \hat{\mu}_{dr} &= \hat{N}_V^{-1} \sum_{i \in s_V} \frac{1}{\pi_i} \{y_i - m(x_i, \hat{\beta})\} + \hat{N}_R^{-1} \sum_{i \in s_R} \frac{1}{\pi_i} m(x_i, \hat{\beta}) \\ &= \hat{N}_V^{-1} \sum_{i \in s_V} \frac{1}{\pi_i} \{y_i - m(x_i, \hat{\beta})\} + \hat{\mu}_{reg} \end{aligned}$$

$$\text{여기서 } \hat{N}_R = \sum_{i \in s_R} \frac{1}{\pi_i}, \quad \hat{N}_V = \sum_{i \in s_V} \frac{1}{\pi_i}$$

○ 중위수 ($q = 0.5$)

$$\text{- 표본분포함수 } \hat{F}_n(\theta) = \frac{\sum_{k=1}^n w_k I(y_k \leq \theta)}{\sum_{k=1}^n w_k}, \quad -\infty < \theta < \infty,$$

- 관심변수를 순위데이터로 표기 $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$

- $y_{(k)}$ 지점에서 표본분포함수의 점프 폭 $\lambda_{(k)} = w_{(k)} \left/ \sum_{l=1}^n w_l \right.$ 이면,

- 중위수는
$$\begin{cases} \hat{\theta}_q = \frac{1}{2}(y_{(i)} + y_{(i+1)}), & \text{if } \sum_{k=1}^i \lambda_{(k)} = q \\ \hat{\theta}_q = y_{(i+1)}, & \text{if } \sum_{k=1}^i \lambda_{(k)} < q < \sum_{k=1}^{i+1} \lambda_{(k)} \end{cases}$$

- 추정량별 y_k 및 점프 폭 $\lambda_{(k)} = w_{(k)} \left/ \sum_{l=1}^n w_l \right.$

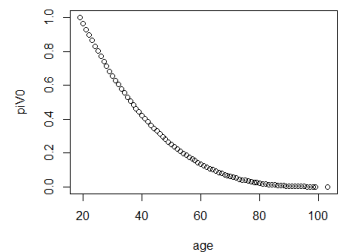
- 설계기반 $\hat{\theta}_{0.5, R}$: $y_l, w_l = (\pi_l)^{-1}, l \in s_R$
- 단순 $\hat{\theta}_{0.5, naive}$: $y_l, w_l = 1$
- 성향점수 $\hat{\theta}_{0.5, ipw}$: $y_l, w_l = (\hat{\pi}_i)^{-1}, l \in s_V$
- 통대체 $\hat{\theta}_{0.5, reg}$: $\hat{y}_l, w_l = (\pi_l)^{-1}, l \in s_R$
- 이중강건 $\hat{\theta}_{0.5, dr}$: 성향점수가 추정된 비확률표본 데이터셋과 관심변수가 추정된 확률표본 데이터셋의 가중치를 동일하게 조정 후 통합하여 분포함수 생성

□ 모의실험 설정 (중간)

- 모집단 : 연간가구소득 상위 5%인 가구는 제외, $N=17,277$
- 확률표본 (S_R) : SRS 표집, $n_R=300, 500, 1,000, 2,000$
- 비확률표본 (S_V) : $n_V=300, 500, 1,000, 2,000$

- $\pi_{B,i} = \frac{(\text{최고령} - \text{연령})^3}{(\text{최고령} - \text{최저령})^3}$ 로 포아송 표집

- Castro-Martin et al. (2020)의 설정으로 연령이 높아질수록 참여확률 감소



○ 관심변수 (y) : 연간가구경상소득

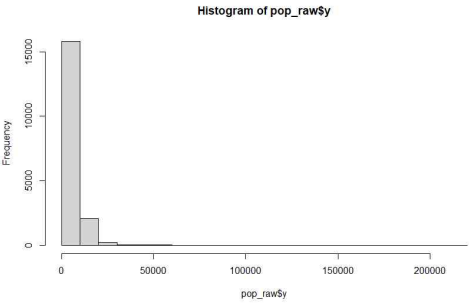
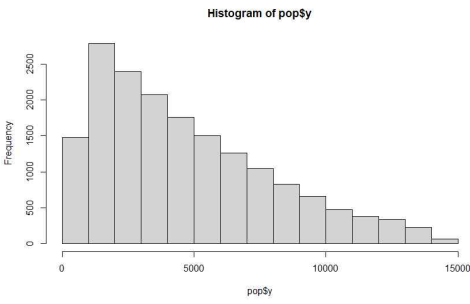
○ 보조변수 (x_i)

- 성향점수모형 ($x_{PS,i}$) : 성별(남,여), 교육수준(중졸이하, 고졸이하, 대졸이상), 연령수준(20대이하, 30대~50대, 60대이상)
- 예측모형 ($x_{SP,i}$) : 연령, 비소비지출(exp2)

○ 모집단 특성

		비중(%)	income (만원)			
			최소	평균	중위값	최대
계		100.0	40	5,538	4,177	219,905
성별	남	72.8	40	6,488	5,208	219,905
	여	27.2	41	2,996	2,066	61,896
교육	대졸+	36.7	40	8,089	6,599	219,905
	고졸	32.5	41	5,224	4,360	61,424
	중졸	10.7	41	3,752	2,986	32,037
	초졸-	20.1	50	2,326	1,624	62,752
연령	20대-	2.3	70	3,482	2,984	34,492
	30대	10.0	52	6,272	5,345	112,389
	40대	18.1	40	7,264	6,283	219,905
	50대	22.3	41	7,404	5,990	103,525
	60대	22.9	50	5,294	4,064	88,712
	70대+	24.3	50	2,672	1,793	61,424
종사상 지위	상용	37.2	52	7,789	6,515	112,389
	임시	12.7	50	3,497	2,727	22,592
	자영	23.5	41	6,086	4,783	219,905
	기타	26.5	40	2,872	1,796	52,081
지역	수도권	32.4	50	6,300	4,679	219,905
	비수도권	67.6	40	5,172	3,966	112,389

주. $r(income, exp1) = 0.68$, $r(income, exp2) = 0.82$, $r(income, age) = -0.25$

	가금복 100%	가금복 95%
평균	5,538	4,700
분산	31,073,507	10,688,401
왜도	6.98	0.84
첨도	160.62	2.95
분포		

○ 표본분포 ($n_R = 1,000$, $n_V = 1,000$)

(단위 : %)

	구분	모집단	SRS	STR	PPS	비확률표본
성별	남	72.80	72.80	72.80	72.80	76.80
	여	27.20	27.20	27.20	27.20	23.20
교육	대졸+	36.70	36.70	36.80	36.70	54.80
	고졸	32.50	32.50	32.50	32.50	34.30
	중졸	10.70	10.70	10.70	10.70	5.37
	초졸-	20.10	20.10	20.10	20.10	5.47
연령	20대-	2.30	2.35	2.35	2.35	8.65
	30대	10.00	9.98	9.97	9.98	25.50
	40대	18.10	18.10	18.10	18.10	29.70
	50대	22.30	22.30	22.30	22.30	21.20
	60대	22.90	22.90	23.00	22.90	11.30
	70대+	24.30	24.30	24.30	24.30	3.69
종사상 지위	상용	37.20	37.20	37.20	37.20	57.00
	임시	12.70	12.70	12.80	12.70	10.30
	자영	23.50	23.50	23.50	23.50	21.10
	기타	26.50	26.50	26.50	26.50	11.60
혼인 상태	미혼	10.50	10.50	10.50	10.50	23.00
	유배우자	62.40	62.40	62.40	62.40	61.90
	사별	15.80	15.80	15.80	15.80	4.68
	이혼	11.40	11.40	11.30	11.40	10.40
지역	수도권	32.40	32.40	32.40	32.40	36.50
	비수도권	67.60	67.60	67.60	67.60	63.50

□ 실험결과

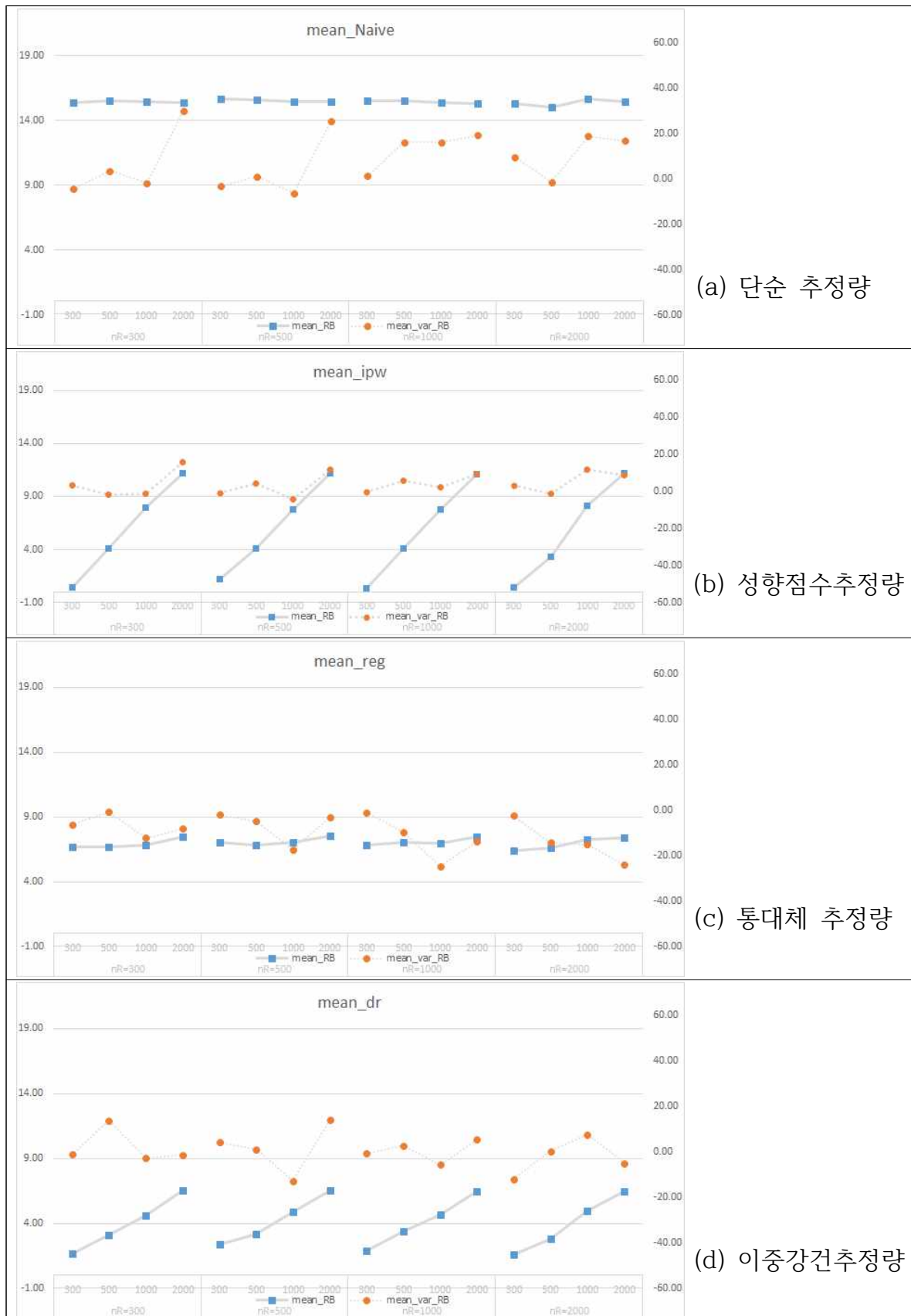
○ 참조확률표본이 *SRS*이고 크기가 500인 경우에 비확률표본의 크기만 변경 (캘리브레이션 추정량은 미수행)

○ 평균

$n_R = 500$	추정량	평균 %RB	MSE	분산 %RB	%CP
$n_V = 300$	$\hat{\mu}_R$	0.21	21,317	0.79	0.95
	$\hat{\mu}_{naive}$	15.62	574,024	-3.50	0.01
	$\hat{\mu}_{ipw}$	1.18	72,431	-1.07	0.95
	$\hat{\mu}_{reg}$	7.07	161,919	-1.81	0.70
	$\hat{\mu}_{dr}$	2.36	69,143	4.17	0.94
$n_V = 500$	$\hat{\mu}_R$	-0.15	20,000	6.55	0.96
	$\hat{\mu}_{naive}$	15.53	552,743	0.82	0.00
	$\hat{\mu}_{ipw}$	4.07	72,709	3.91	0.85
	$\hat{\mu}_{reg}$	6.80	139,468	-4.78	0.62
	$\hat{\mu}_{dr}$	3.16	61,091	1.27	0.90
$n_V = 1000$	$\hat{\mu}_R$	-0.19	19,685	8.75	0.95
	$\hat{\mu}_{naive}$	15.38	533,393	-6.67	0.00
	$\hat{\mu}_{ipw}$	7.81	152,874	-4.23	0.20
	$\hat{\mu}_{reg}$	7.06	141,225	-17.17	0.42
	$\hat{\mu}_{dr}$	4.91	84,001	-12.77	0.72
$n_V = 2000$	$\hat{\mu}_R$	-0.14	21,681	-1.89	0.94
	$\hat{\mu}_{naive}$	15.43	529,904	25.07	0.00
	$\hat{\mu}_{ipw}$	11.22	286,101	11.38	0.00
	$\hat{\mu}_{reg}$	7.51	147,357	-3.20	0.30
	$\hat{\mu}_{dr}$	6.52	112,404	14.11	0.43

- 비확률표본을 가중치 보정 후 추정한 결과가 보정없이 추정한 결과에 비해 평균의 상대편향, 평균제곱오차, 추정 신뢰구간에 모 평균 포함 정도가 모두 뚜렷한 개선을 보였음.
- 다만, 비확률표본도 크기가 클수록 추정량을 개선시킬 거라는 예측은 잘 맞지 않았는데, 성향점수 추정량에서는 비확률표본의 크기가 클수록 오히려 추정이 비효율적으로 나타남
 - 성향점수 추정에 대해 재검토 필요해 보임
- 상대편향이 클수록 신뢰구간 포함확률이 떨어짐, 비확률표본 추론의 성공은 편향의 감소 여부와 직결

○ 시나리오별 평균의 상대편향과 MSE

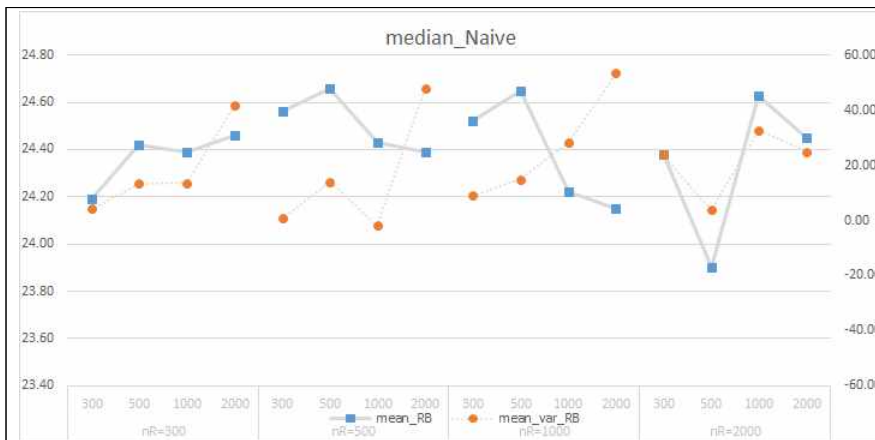


○ 분위

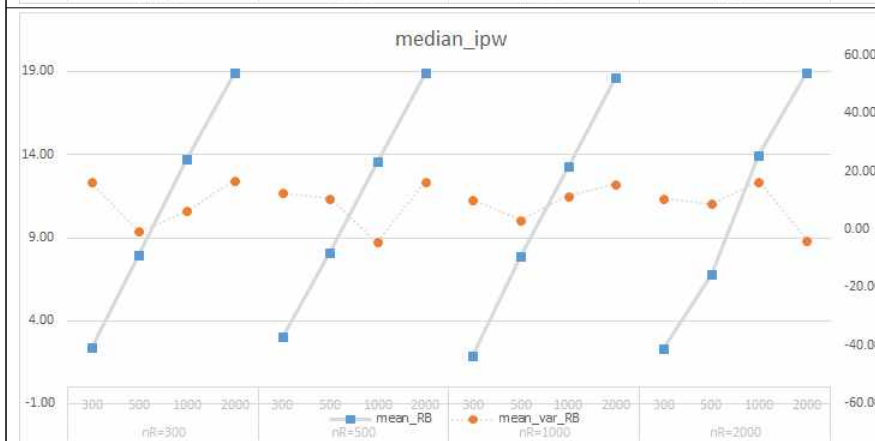
$n_R = 500$	추정량	평균 %RB	MSE	분산 %RB	%CP
$n_V = 300$	$\hat{\mu}_R$	0.55	42,377	-1.77	0.93
	$\hat{\mu}_{naive}$	24.56	1,001,417	0.78	0.03
	$\hat{\mu}_{ipw}$	3.07	141,064	12.54	0.93
	$\hat{\mu}_{reg}$	8.55	207,902	-38.93	0.69
	$\hat{\mu}_{dr}$	7.07	174,213	-31.44	0.83
$n_V = 500$	$\hat{\mu}_R$	-0.23	35,958	14.67	0.95
	$\hat{\mu}_{naive}$	24.66	976,859	14.12	0.00
	$\hat{\mu}_{ipw}$	8.13	170,037	10.71	0.80
	$\hat{\mu}_{reg}$	8.65	190,739	-39.48	0.63
	$\hat{\mu}_{dr}$	8.77	184,610	-33.44	0.61
$n_V = 1000$	$\hat{\mu}_R$	-0.30	37,875	11.48	0.95
	$\hat{\mu}_{naive}$	24.43	944,113	-1.96	0.00
	$\hat{\mu}_{ipw}$	13.63	323,848	-4.36	0.17
	$\hat{\mu}_{reg}$	9.69	226,580	-49.14	0.44
	$\hat{\mu}_{dr}$	11.42	261,243	-46.19	0.16
$n_V = 2000$	$\hat{\mu}_R$	0.04	39,234	6.84	0.94
	$\hat{\mu}_{naive}$	24.39	927,788	48.08	0.00
	$\hat{\mu}_{ipw}$	18.98	572,224	16.59	0.00
	$\hat{\mu}_{reg}$	11.39	295,841	-48.77	0.32
	$\hat{\mu}_{dr}$	14.22	366,380	-44.04	0.01

- 비확률표본을 가중치 보정 후 추정한 결과가 보정없이 추정한 결과에 비해 중위수의 상대편향, 평균제곱오차, 추정 신뢰구간에 모평균 포함 정도가 모두 뚜렷한 개선을 보였음.
- 평균과 마찬가지로 중위수의 IPW 추정량도 불안정하며 IPW 추정량과 REG 추정량을 결합한 DR 추정량 역시 불안정하며 그 정도도 평균에 비해 큼
- 이같은 경향은 확률표본보다는 비확률표본의 크기에 더 영향을 받는 것으로 보임

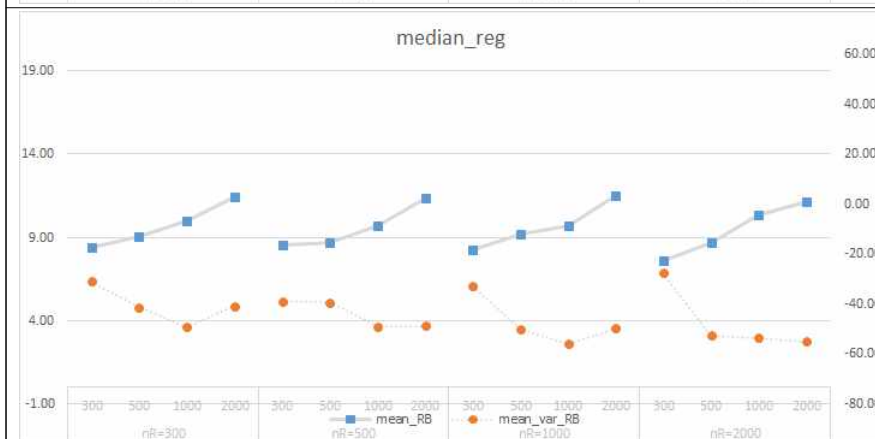
○ 시나리오별 중위수의 상대편향과 MSE



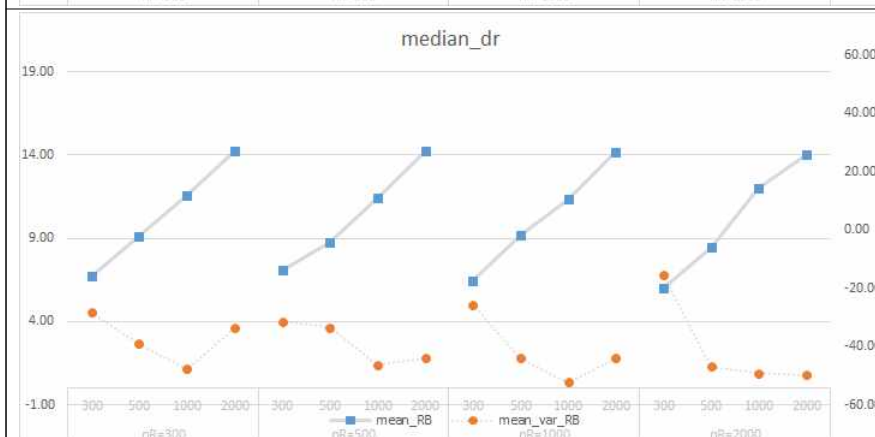
(a) 단순 추정량



(b) 성향점수 추정량



(c) 통대체 추정량



(d) 이중강건 추정량

□ 비확률표본의 편향 감소 경향 뚜렷

- 고품질 참조확률표본과의 모형 설정을 통해 비확률표본의 참여 확률이나 확률표본의 미관측 관심변수를 추정하는 추론은 비확률 표본을 단순히 추론하는 것에 비해 편향 및 분산 경향 뚜렷
- 그러나 비확률표본의 크기가 커질수록 추정 결과가 불안정하게 나오는 것에 대한 검토 필요. 특히, ipw 추정에서 그 경향이 크게 나타남

□ 성향점수 모형 프로그램의 재검토 필요

- Chen et al. (2020)이 제안한 방식을 2022년 연세대 임종호 교수와 공동연구를 통해 작성한 프로그램으로 성향점수를 추정하였는데, 이에 대한 재검토 필요해 보임

□ 실제로 어떤 분야에 적용이 가능할지 고민

- 통계청의 조사 자료는 대부분 고품질 확률표본이거나 포괄범위가 넓은 행정자료로 본 연구 결과를 적용하거나 지속할 수 있는지

□ 확률표본의 크기가 미치는 영향은 작음

- 고품질 확률표본이라면 크기는 추론의 성능에 크게 영향을 미치지 않는 것으로 보임
- 확률표본의 설계가 미치는 영향은 실험 예정

□ 기타사항

- PC 컴퓨팅 능력 한계 : $n_R = 2000$, $n_V = 2000$ 인 경우, MC 500, 붓스트랩 500회 수행에 40시간 소요, 보다 정밀한 실험 어려움
- 중간까지의 실험은 공공용 가구마스터 자료를 이용하였는데, 본 실험은 RAS (원격접근시스템)에서 수행 예정으로 연구자의 제어가 쉽지 않은 상태에서 모의실험이 원활히 수행될지 우려됨