콜로키움자료집 19-S22

제21차 NYPI 역량강화 콜로키움

복합표본설계자료 분석의 개념 및 실제

일 시 2019년 8월 14일(수) 14:00~16:00

장 소 한국청소년정책연구원 7층 대회의실

주 최 한국청소년정책연구원

한국청소년정책연구원
National Youth Policy Institute

복합표본설계자료 분석의 개념 및 실제

박민규 (고려대학교 통계학과)

복합표본설계자료 분석의 개념 및 실제

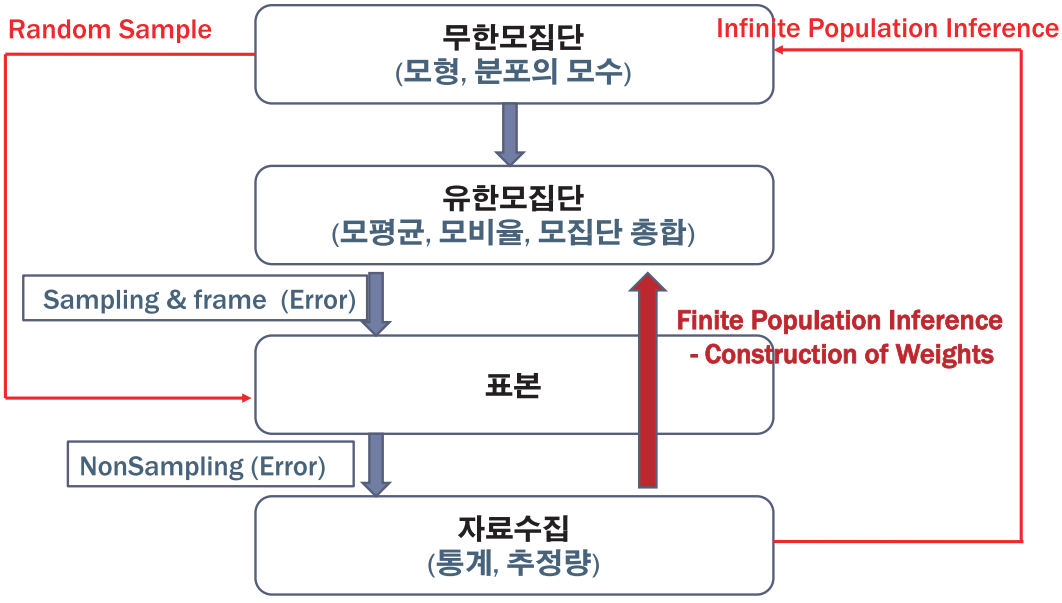
박민규 (고려대학교 통계학과)
2019-08-14

목차

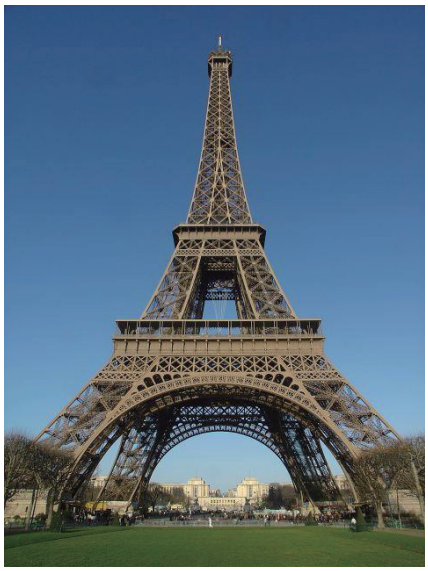
가중치(weight) ?

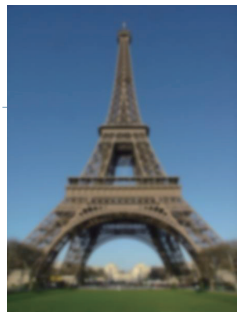
- 중앙선거여론조사공정심의위원회
  - 10: 한국리서치
  - 916: 리서치뷰
- 국민건강영양조사
  - 원시자료 -> 자료분석지침 -> FAQ
- 한국노동패널
  - 조사설계 및 결과 -> 가중치

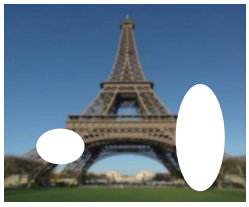
표본조사(Sample survey) 자료의 분석



가중치의 역할







표본조사 자료

	Variables			
	Weight	Y_1	⋯	Y_k
1	w_1	y_{11}	⋯	y_{1k}
2	w_2	Missing	⋯	y_{2k}
⋮	⋮	⋮	⋱	⋮
n_R	w_{nR}	y_{nR1}	⋯	Missing
⋮				
n	Missing			

개체 무응답 (Unit Nonresponse)

항목 무응답 (Item Nonresponse)

추정량 (Estimator)

- **모집단 총합 추정**

$$\hat{t} = \sum_{i=1}^{n_R} w_i y_i$$

- **모평균 추정**

$$\bar{y} = \frac{\sum_{i=1}^{n_R} w_i y_i}{\sum_{i=1}^{n_R} w_i}$$

- **모비율 추정**

$$\hat{p} = \frac{\sum_{i=1}^{n_R} w_i y_i}{\sum_{i=1}^{n_R} w_i}$$

$$y_i = \begin{cases} 1, & i \in U_g, \\ 0, & i \notin U_g. \end{cases}$$

가중치 어떻게 만들어지나?

- **표본 가중치**(sampling weight)
  - **확률표본설계**를 통해 생성되는 가중치
- **무응답 조정 가중치**(nonresponse adjusted weight)
  - Within cell(or group) adjustment
  - Post-stratification
  - Propensity score method

가중치 어떻게 만들어지나?

- **캘리브레이션 가중치**(calibration weight)
  - 알고 있는 모집단 정보와의 정합성을 위해
  - 추정량의 효율성을 높이기 위해
  - Raking Ratio (Rim) weight
  - Regression weight
  - Ratio weight

표본가중치

$$w_i = \frac{1}{\pi_i}$$

- 표본의 한 개체에 의하여 대표되는 모집단의 개체 수
- 일반적으로 $w_i \geq 1$
- *단순임의추출*: N/n
- *2단계 집락추출*: *N*개의 집락에서 *n* 개의 집락을 추출 한 후 M_i의 개체 중 n_i를 추출할 경우: $(N/n)(M_i/n_i)$
- π_i 개체 *i*가 표본에 포함 될 확률

무응답 보정 가중치 – Within Cell Adjustment

- Response Homogeneous Cell(RHC) 가정:

Cell	Initial weight	y	R	Adjusted weight
⋮	⋮	⋮	⋮	⋮
i	$w_{0,i1}$	y_{i1}	1	$w_{a,i1}$
i	$w_{0,i2}$	y_{i2}	1	$w_{a,i2}$
i	$w_{0,i3}$	y_{i3}	1	$w_{a,i3}$
i	$w_{0,i4}$	y_{i4}	1	$w_{a,i4}$
i	$w_{0,i5}$	y_{i5}	0	0
⋮	⋮	⋮	⋮	⋮

응답

무응답

$$w_{a,ij} = \frac{w_{0,ij}}{p_g}, \quad p_g = \frac{4}{5}$$

응답확률 추정치

무응답 보정 가중치

- 사후층화

$$\bar{y}_{post} = \sum_{g=1}^{G} \frac{N_g}{N} \tilde{y}_{Rg} = \sum_{g=1}^{G} \frac{N_g}{N} \bar{y}_{Rg}$$

Equal probability sample

- 표본추출이 단순임의추출법이나 계통추출법을 통해 이루어지고 그룹 (예: 성*연령) 모집단 분포가 주어진 경우, 추정량은 각 그룹의 상대크기를 이용한 그룹 표본 평균의 가중 합으로 정의 된다.

무응답 보정 가중치

- Propensity score

$$\bar{y}_{ps} = \sum_{i \in s_R} \frac{\alpha_i y_i}{\hat{p}_i}$$

$$\hat{p}_i = \hat{P}(R_i = 1)$$

$$\alpha_i = \left(\sum_{j \in s_R} \pi_j^{-1} \right)^{-1} \pi_i^{-1}$$

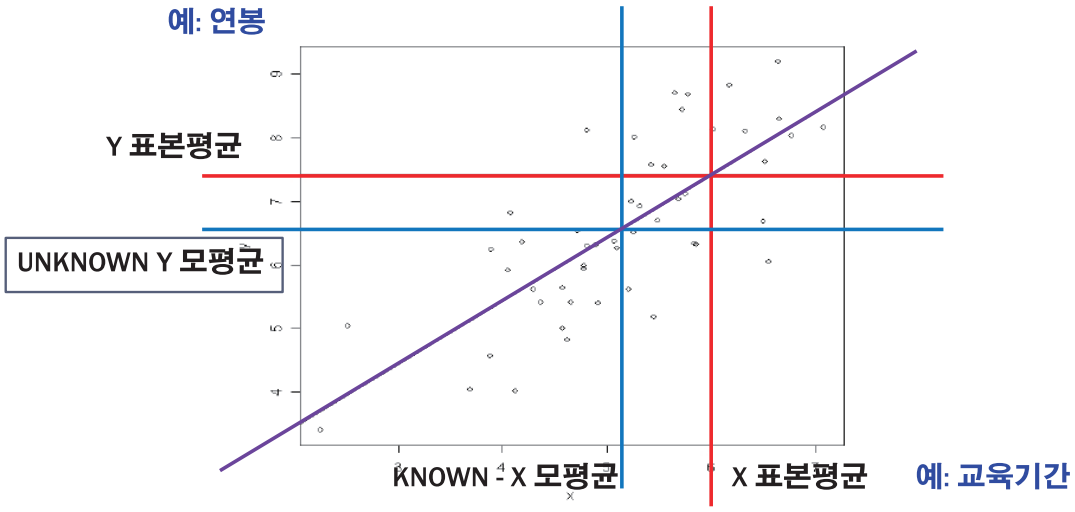
  - 로지스틱 회귀모형이 응답확률 예측을 위해 흔히 사용됨.
  - RHC 보정 방안은 propensity score 방안의 한 형태임.

현실에서의 무응답 보정

- Cross-sectional survey
  - 거의 이루어지지 않음.
  - 현장에서 표본 대체(substitution)가 이루어짐.
  - 조사의 비용(단가)이 표본 수에 의존.
- Panel survey
  - 일반적으로 2차 년도부터 무응답 보정이 propensity score 방안을 이용하여 이루어짐.
  - 가중치의 분산이 매우 커질 수 있음.

CALIBRATION

- 추정량의 효율성을 높이기 위하여



효율성을 높이기 위한 CALIBRATION

- Underlying Model
  - 일반 회귀추정량 $E(Y) = \beta_0 + \mathbf{x}\beta$
  - 사후층화추정량 $E(Y) = \mu_h, h = 1, \cdots, H$
  - 2 변수 레이킹 비 추정량 $E(Y) = \mu + \alpha_i + \beta_j$
- 사용되는 대부분의 추정량들은 회귀추정량이거나 혹은 근사적으로 이와 동일한 추정량들이다.

결과의 정합성을 위한 CALIBRATION

- 예를 들어 모집단의 성*연령 분포가 알려진 경우, 산출된 가중치를 표본에 적용하여 얻어지는 성*연령 분포가 주어진 모집단 분포와 일치하는 것이 일반적으로 요구됨.
- 이는 가중치 조정을 통해 표본의 모집단 대표성을 보조변수(auxiliary variable) 측면에서 확보하기 위함.
- 주로 사용되는 방안은 사후층화, 회귀, 레이킹 비 그리고 비 추정 방안임.

CALIBRATION 추정량

- 회귀추정량

$$\begin{aligned}\bar{y}_{reg} &= \bar{y}_{\pi} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\pi})\hat{\beta}_{\phi} \\ &= \sum_{i \in S} w_{i,reg} y_i\end{aligned}$$

$$\hat{\beta}_{\phi} = (\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{y}, \quad \boldsymbol{\Phi} = Diag(\phi_{ii},\ i = 1, \cdots, n)$$

$$w_{i,reg} = \alpha_i + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_{\pi}) \left(\sum_{i \in S} \mathbf{x}_i' \phi_{ii} \mathbf{x}_i \right)^{-1} \mathbf{x}_i' \phi_{ii}$$

- 사후층화 추정량
  - 보조변수가 아래와 같은 회귀추정량

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$$

$$x_{ig} = \begin{cases} 1, & \text{if } i \in U_g, \\ 0, & \text{elsewhere.} \end{cases}$$

CALIBRATION 추정량

- 비 추정량:
  - 보조변수가 하나이며 $\beta_0 = 0,\ \phi_{ii} = 1/x_i$ 인 회귀추정량

$$\bar{y}_{reg} = \bar{x}_N \frac{\bar{y}_\pi}{\bar{x}_\pi}$$

- 레이킹 비 추정량

고려된 모든 추정량이 Calibration Eqn. 을 만족함

$$w_{i,rak} = \alpha_i \exp\left(\lambda' \mathbf{x}\right),$$

where λ is a solution to

$$\sum_{i \in s} w_{i,rak} \mathbf{x}_i = \mathbf{t}_x.$$

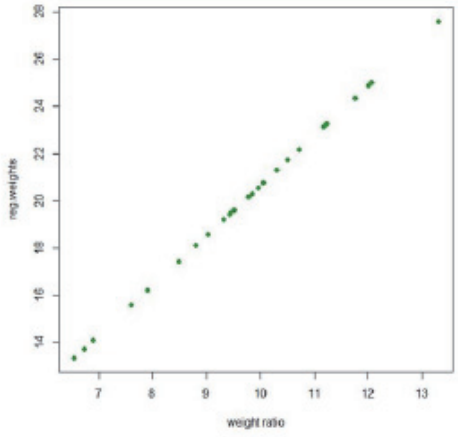
Calibration Eqn.

CALIBRATION 추정량

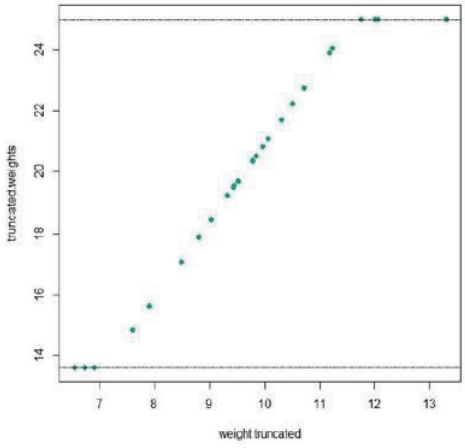
- Truncated Linear : 범위가 제한된 회귀 가중치
- Logit : 범위가 제한된 레이킹 비 가중치
  - 정합성이 요구되는 변수의 수가 매우 많은 경우, 회귀가중치와 레이킹 비 가중치는 극단적으로 크거나 혹은 음의 값을 갖는 문제를 야기할 수 있다. 이러한 문제를 해결할 수 있는 여러 방안 중 알고리즘을 통한 가중치 산출 방안으로 가중치의 상 · 하한을 조정하는 truncated linear 방법과 logit 방법을 고려할 수 있다.

CALIBRATION 추정량

The linear (regression) method

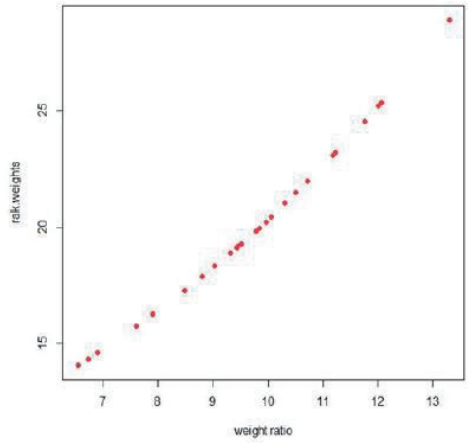


The truncated (L, U) method

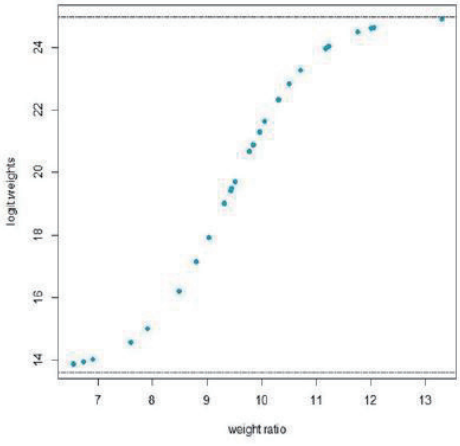


CALIBRATION 추정량

The raking ratio method



The logit (L, U) method



현실에서의 CALIBRATION

- Cross-sectional survey
  - 대부분의 일정 규모 이상의 조사에서 적용됨.
  - 무응답 조정이 필요한 경우, calibration을 통한 무응답 조정이 이루어짐.
- Panel survey
  - 무응답 조정 이 후 calibration이 일반적으로 이루어짐
- Calibration을 위한 변수로는 인구학적 변인과 가구 정보 그리고 산업분류 등이 흔히 사용됨.

가중치: 어떻게 사용할 것인가?

- Preliminary Questions

1. 추론의 대상과 표본추출 방안은?
   - 유한 모집단 / 무한모집단
   - 확률추출법 / 비확률추출법
2. 추론의 목적은?
   - 단순 모수 추정 / Statistical Modeling
3. 추론의 방법은?
   - Design Based / Model Based

가중치: 어떻게 사용할 것인가?

- **유한모집단/확률추출법/단순모수추정/Design-Based**
  - Survey Statistics에서의 가장 일반적인 형태.
  - 반드시 가중치의 적용이 필요함.
  - 모집단 총합(실업자 수 등), 모평균(평균 수입), 모비율(유병률), 모집단 회귀계수(Fertility on GDP) 등의 추정.
  - 비편향, 회귀, 사후층화, 비, 레이킹 추정량
  - SPSS, SAS, STATA 등을 이용한 추정 가능.
  - SAS : SURVEYMEANS, SURVEYREG Procedures
  - 교과서와 프로그램에서 제공하는 분산 추정량은 sampling variance만을 고려한 추정량임.

가중치: 어떻게 사용할 것인가?

- **무한모집단/Statistical Modeling/Model-Based**
  - Survey Statistics를 제외한 모든 통계 분야에서의 분석 방안
  - 가중치를 무시하고 무한모집단의 가정 하에서 분석이 이루어짐
  - 회귀분석, 분산분석, 로지스틱 회귀분석, 다변량 분석
  - SPSS, SAS, STATA 등을 이용한 분석 가능
  - 분석결과의 해석을 위해서는 무한 모집단의 가정에 대한 검증이 요구됨

가중치: 어떻게 사용할 것인가?

- 무한모집단/확률추출법/Statistical Modeling/Design-Based
  - 유한모집단을 무한모집단으로부터의 랜덤표본으로 간주하고 유한모집단에 대한 추론의 결과를 무한모집단으로 확대하는 분석방안
  - 가중치를 사용한 분석이 필요함.
  - SAS : SURVEYMEANS, SURVEYREG, SURVEYFREQ Procedures 등 기본적인 분석 procedure가 이용가능 함.

가중치: 어떻게 사용할 것인가?

- Example: Relation between Farm Area and Corn Yield

			Number of Farms	
Stratum	State	Region	Population	Sample
1	Iowa	1	100	3
2		2	50	5
3		3	15	3
4	Nebraska	1	30	6
5		2	40	2
Total			235	19

가중치: 어떻게 사용할 것인가?

- **Example: Relation between Farm Area and Corn Yield**
- **Result without using weight**

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.64378	8.98284	1.41	0.1773
FarmArea	1	0.18368	0.05556	3.31	0.0042

- **Result using weight**

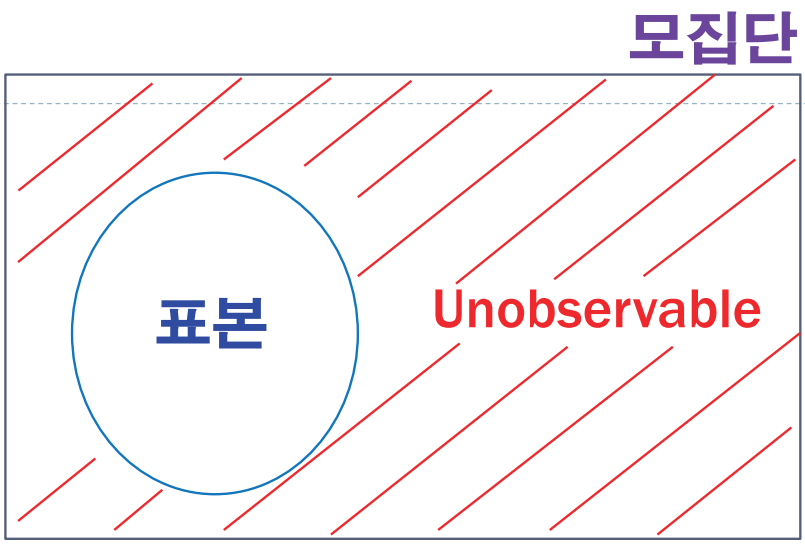
Estimated Regression Coefficients				
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	11.8162978	5.31981027	2.22	0.0433
FarmArea	0.2126576	0.04560949	4.66	0.0004

가중치: 어떻게 사용할 것인가?

- **무한모집단/확률추출법/Statistical Modeling/Design-Based**
  - **보다 복잡한 형태의 모형(Mixed model 등)을 위한 프로그램은 아직 제공되지 않음**
  - **Model parameter의 추정량에 대한 분산 추정 방안**
    - **Design-Based: Focus on sampling variance**
    - **Anticipated concept: Model expected design variance**
  - **Central Limit Theorem**

가중치: 어떻게 사용할 것인가?

- **유한모집단/비확률추출/단순모수/Model-Based**
  - 표본을 무한모집단으로부터의 랜덤표본으로 간주하고 유한모집단의 모수를 예측(Prediction)하는 관점에서 분석을 실시(Valliant, Dorfman, Royall)
  - 현재의 한국 조사 현실에서는 표면적으로 거의 사용되지 않는 접근방법
  - 그러나 실제 전화조사 등에서는 사용자가 인식하지 못하고 사용하는 경우가 빈번함
  - model failure에 추정량의 대한 강건성(robustness)을 항상 평가해야 하며 분산 추정에 많은 주의가 필요함



$$\sum_{k \in sample} y_k + \sum_{k \in not\ in\ sample} \hat{y}_k$$

가중치: 어떻게 사용할 것인가?

- **유한모집단/단순모수/Model-Based**
  - 예를 들어, RDD and(or) Quota Sampling을 통한 전화조사의 경우 흔히 지역*성*연령의 모집단 분포를 이용한 아래의 추정량을 사용한다.

$$\bar{y} = \frac{1}{N}\sum_i \sum_j \sum_k N_{ijk}\bar{y}_{ijk}$$

$$\hat{p} = \frac{1}{N}\sum_i \sum_j \sum_k N_{ijk}\hat{p}_{ijk}$$

  - ***i***(지역), ***j***(성), ***k***(연령대)

가중치: 어떻게 사용할 것인가?

- **유한모집단/단순모수/Model-Based**
  - 고려된 추정량은 아래의 모형 하에서 그 타당성을 갖는다.

$$y_{ijkl} \sim \left(\mu_{ijk}\ ,\ \sigma^2\right)$$

$$y_{ijkl} \sim Bernoulli\left(p_{ijk}\right)$$

  - 즉 $\left(\bar{y}_{ijk}\ ,\ \hat{p}_{ijk}\right)$ 는 $\left(\mu_{ijk}\ ,\ p_{ijk}\right)$ 의 추정량이며 이를 이용한 유한 모집단의 추정량이 유도된 것으로 생각할 수 있다.

가중치: 어떻게 사용할 것인가?

- 유한모집단/단순모수/Model-Based
- Design-based 추론에 근거한 분산 추정량의 사용은 이론적으로 타당하지 않음.
- 대안: 모평균 및 모비율에 대한 추정량의 분산 추정량은 주어진 모형 하에서 다음을 고려할 수 있음.

$$\hat{V}(\bar{y}) = \frac{s^2}{N^2} \sum_i \sum_j \sum_k \frac{N_{ijk}^2}{n_{ijk}},$$

$$\hat{V}(\hat{p}) = \frac{1}{N^2} \sum_i \sum_j \sum_k \frac{N_{ijk}^2}{n_{ijk}} \hat{p}_{ijk}(1 - \hat{p}_{ijk}),$$

$$s^2 = \frac{\sum_i \sum_j \sum_k (n_{ijk} - 1) s_{ijk}^2}{\sum_i \sum_j \sum_k n_{ijk} - 1}$$

예제 (Unrealistic but just for demo)

사후 층	모집단 크기	표본크기	표본비율 추정량
1110	60	13	0.5
2110	70	13	0.5
1120	80	13	0.5
2120	90	13	0.5
1210	100	12	0.5
2210	100	12	0.5
1220	200	12	0.5
2220	300	12	0.5

추정량	0.5	
SE 1	0.05	17% Underestimate
SE 2	0.06	

결론

- 타당한 가중치의 사용을 위해서는 가중치의 산출 과정에 대한 기본적인 이해가 필요하다.
- 가중치의 사용 여부는 추론의 대상, 목적 그리고 방법에 의하여 결정됨으로 명확히 이를 정의하는 것이 선행되어야 한다.
- Statistical Modeling이 목적인 경우 그리고 추론의 대상이나 목적이 명확하지 않을 때는 가중치를 적용한 방법을 사용하는 것이 안전한 접근방법일 수 있다.

결론

- 유한모집단의 모수 추정량에 대한 분산 추정은 매우 복잡한 문제이나 기본적으로 가중치를 고려한 표본분산을 사용해야 하며 이와 더불어 상당한 수준의 무응답이 발생한 경우 이에 대한 적절한 가정 하에서 무응답으로 인한 분산을 함께 고려해야 한다.
- 비확률표본추출법이 사용된 경우 전통적인 Design-based 추론보다는 Model-based 추론이 대안이 될 수 있으며 이 경우 모형의 타당성 및 분산 추정량의 적절성에 대한 검증이 요구된다.

결론

- 추론의 과정에 대한 확신이 없다면 표본 전문가에게 문의하는 것이 현명한 방안일 수 있다.

Q/A

MEMO