14강 비표본오차

정보통계학과 이기재교수

학/습/목/차

1. 표본조사 오차의 개요

2. 편향의 종류

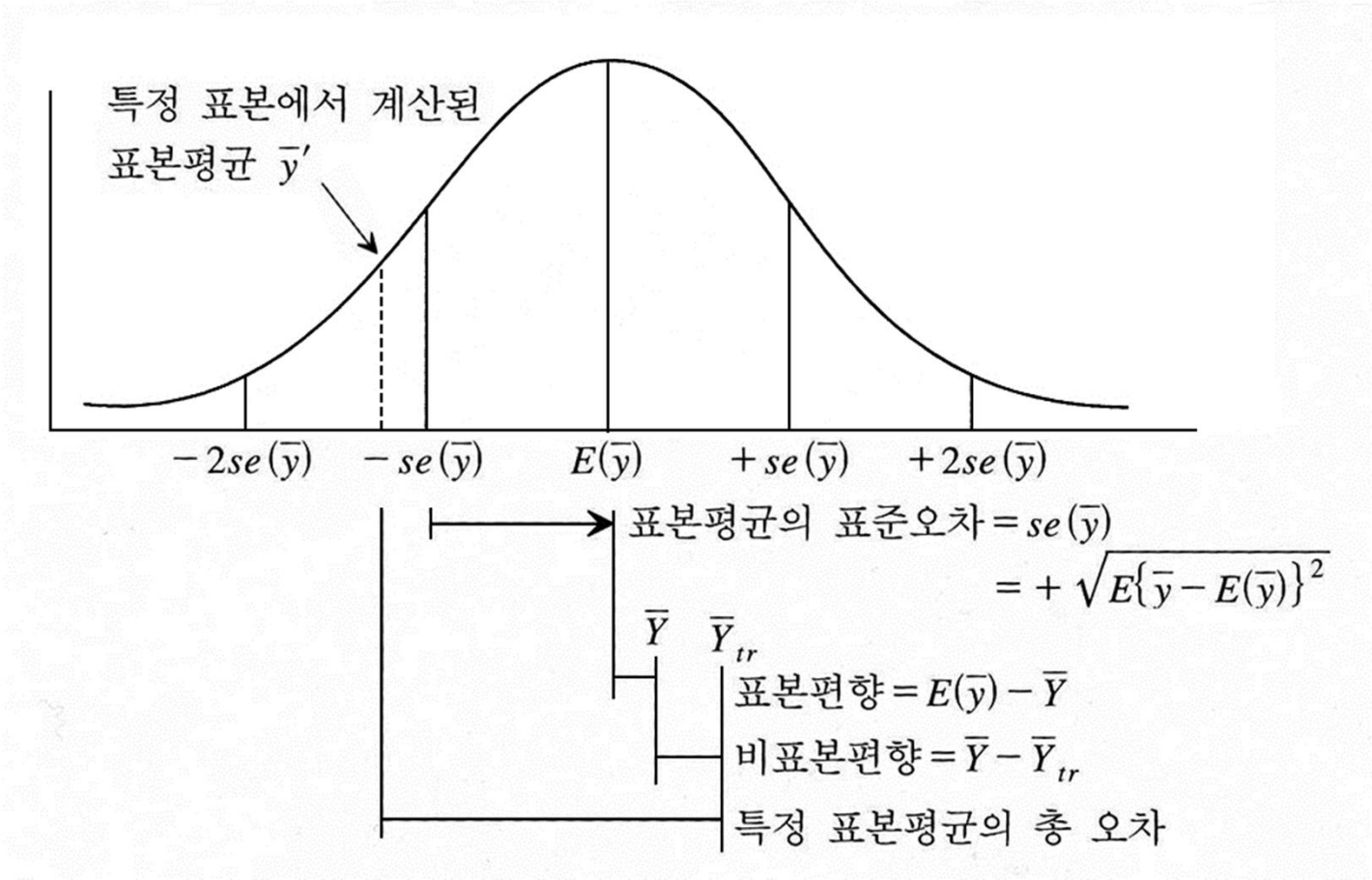
3. 비표본오차의 종류

4. 비표본오차의 측정

5. 항목 무응답 대체법

1. 표본조사 오차의 개요

- [그림 1] 표본평균($\bar{y}$) 의 확률 분포



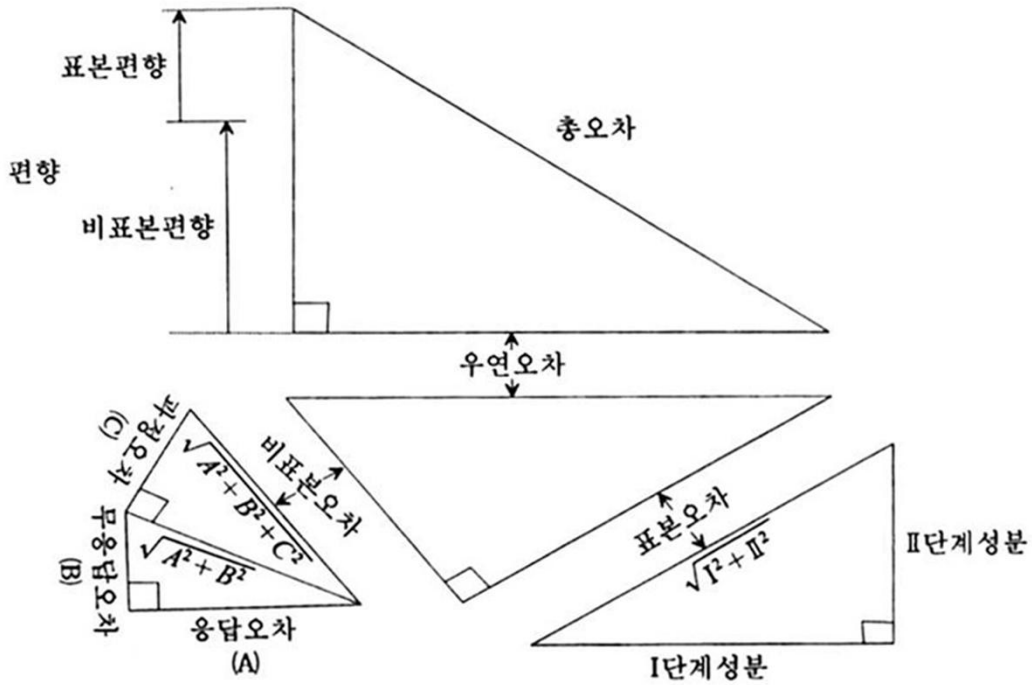
2. 표본조사에서 오차의 종류

- 우연오차(variable error)
  - 표본오차(sampling error) : 표본추출로 인해 발생한 오차
  - 비표본오차(non-sampling error)
    : 조사기획, 공표과정까지 발생한 오차

- 편향(bias)
  - 비확률변수 특성, 일종의 체계적인 오차
  - 표본편향과 비표본편향으로 구분

3. 표본조사 오차의 성분

[그림 2] 오차의 성분



- 총오차 = 편향(표본+비표본편향) + 우연오차(표본+비표본오차)

  표본오차 = 1단계 성분 + 2단계 성분

  비표본오차 = 응답오차 + 무응답오차 + 과정오차

학/습/목/차

1. 표본조사 오차의 개요

2. 편향의 종류

3. 비표본오차의 종류

4. 비표본오차의 측정

5. 항목 무응답 대체법

1. 편향이란?

- **편향(bias)**
  - 추정량의 기대값과 참값간의 차이
  - 표본편향(sampling bias)과 비표본편향(non-sampling bias) 으로 구분
  - 전수조사에서도 발생, 원인 복잡, 체계적 측정이 어려움
  - 편향을 줄이기 위해서는 표본조사 계획 단계, 표본설계, 본조사 과정, 자료집계 및 분석 과정 등의 표본조사 전 과정을 철저하게 관리해야 함

2. 표본편향의 특성과 종류

- 표본편향(sampling bias)의 특성
  - 표본조사를 통해서 모수를 추정하는 과정에서 발생하는 편향
  - 발생 원인에 따라 추출틀 편향, 일치성 편향, 상수성 통계적 편향 등으로 구분
  - 표본추출의 각 단계에서 발생 가능하고, 방향성을 갖고 있음

- 표본편향(sampling bias)의 종류
  - 프레임 표본편향 : 추출틀 부정확으로 발생
  - 일치성 편향
    : 일치추정량이지만 편향추정량을 사용할 때 발생
  - 상수성 통계적 편향 : 추정법에서 기인한 편향(중앙값)

3. 비표본 편향(nonsampling bias)

- (1) 조사가 불가능한 경우
  - 누락된 추출틀에서 표본추출하는 경우
  - 조사대상자의 일부가 응답거부한 경우

- (2) 조사는 가능하나 정확한 측정을 하지 못한 경우
  - 조사단위로부터 정확하게 측정하지 못한 경우에 발생
    ➡ 전수조사나 표본조사에서 모두 발생할 수 있음
  - 응답자의 의도적인 거짓응답(민감한 조사항목),
    면접조사 시 의도적인 조작 등
  - 코딩이나 계산 과정의 실수로 인한 오류

학/습/목/차

1. 표본조사 오차의 개요

2. 편향의 종류

3. 비표본오차의 종류

4. 비표본오차의 측정

5. 항목 무응답 대체법

1. 비표본오차의 특성 (1)

- 표본조사의 전 과정에서 부주의, 실수 등의 이유로 발생한 오차
- 표본조사와 전수조사 모두에서 발생할 있음
- 주요 발생 원인 : 계획단계, 조사과정, 집계과정 중 발생오차

조사 계획 단계 발생

조사 과정 중 발생

자료집계 및 분석 과정 발생

① 조사목적 불명확
② 조사범위 불분명
③ 과소범위 추출틀 사용
④ 잘못된 개념 정의

1. 비표본오차의 특성 (2)

- 표본조사의 전 과정에서 부주의, 실수 등의 이유로 발생한 오차
- 표본조사와 전수조사 모두에서 발생할 있음
- 주요 발생 원인 : 계획단계, 조사과정, 집계과정 중 발생오차

조사 계획 단계 발생

조사 과정 중 발생

자료집계 및 분석 과정 발생

① 조사원 교육 미흡
② 조사업무 감독 소홀
③ 거짓 응답 인한 오차
④ 무응답으로 인한 오차

1. 비표본오차의 특성 (3)

- 표본조사의 전 과정에서 부주의, 실수 등의 이유로 발생한 오차
- 표본조사와 전수조사 모두에서 발생할 있음
- 주요 발생 원인 : 계획단계, 조사과정, 집계과정 중 발생오차

조사 계획 단계 발생

조사 과정 중 발생

자료집계 및 분석 과정 발생

① 부호화 코딩 오차

② 분석 결과의 도표화 오차

③ 결과발표 인쇄과정 오차

2. 무응답오차 발생 원인과 대책

▪ 조사대상으로부터 자료수집 불가로 인한 오차

▶ 면접조사 중 발생 원인 및 대책

① 부재 : 방문일정 조정과 재방문 추진

② 응답거부 : 익명성 보장과 적극적인 협조 동기 부여

③ 항목 무응답 : 핫덱 대체, 콜덱 대체, 회귀 대체 등

▶ 우편조사에서 무응답 원인 및 대책

① 응답거부 : 부차표본추출 조사

② 응답 지연 : 독촉 우편 발송 또는 사전 공지

3. 응답오차 발생 원인과 대책

- 조사대상자로 얻은 관찰값과 참값 간의 차이를 응답오차라고 함
- 조사 성격 또는 조사 내용에 따라 상이한 특성이 있음

▶ 주요 발생 원인 및 대책

① 불충분한 조사원 감독 : 조사업무통제 체계 개선

② 조사원 경험과 전문성 미흡
: 실질적 교육훈련과 지침서 보정

③ 수집 과정상 문제점 : 수시점검 및 검증시스템 구축

4. 종합표 작성 중 발생 오차

- 수집된 자료 처리 분석과정 중에 발생하는 오차
- 편집, 코딩, 도표화 등에서 발생하는 오차

▶ 주요 발생 원인 및 대책

① 기초 데이터 검증 미흡 : 중간 감독자의 현장 점검 후 보완

② 데이터 처리 중 발생 오차
: 부호화 코딩 및 입력과정의 자동화

③ 공표 및 이용상의 오류 : 충실한 보고서 작성

학/습/목/차

1. 표본조사 오차의 개요

2. 편향의 종류

3. 비표본오차의 종류

4. 비표본오차의 측정

5. 항목 무응답 대체법

1. 일관성 점검

- 통계조사에서 추정치를 계산한 후 추정오차를 산출함
- 표본오차는 추정량의 분산이나 상대표준오차로 평가
- 비표본오차 : 6가지 측정방법으로 평가 가능

▶ 일관성 점검

- 조사내용의 정확성을 진단할 수 있는 항목을 질문 포함
  : 작물별 재배 면적과 총 재배 면적을 질문 문항에 포함

- 그래프를 이용한 이상값 검사 : 분포 형태 관찰
  : 가구당 가계비 조사에서 1인당 가계비의 그래프

2. 표본 점검과 사후표본조사 점검

▶ 표본 점검

- 조사과정의 일부를 독립적으로 이중실사(二重實査) 후 비교분석
- 대규모 조사에서 일부 조사단위를 랜덤추출하여 동일하게 조사
- 이중실사에 투입되는 면접원은 경험이 많은 우수 조사원임
- 설계와 점검작업이 성공적인 경우 비표본오차 규모와 특성 파악
- 조사원별 일부 선정 표본점검 후 이상 해당 조사원 자료 점검

2. 표본 점검과 사후표본조사 점검

▶ 사후표본조사 점검

- 대규모 본조사 후 비표본오차 측정을 위해 부차표본추출 재조사
- 잘 훈련된 전문 조사원이 본조사의 질과 오차 규모 평가
- 사후표본과 본조사의 조사값을 대응비교 분석하여 오차 원인과 특성 파악
- 사후표본조사는 본 조사를 마친 2주 후가 적당
  : 기억 오류 방지

3. 외부기록 점검과 품질관리기법

▶ 외부기록 점검

- 전수조사에서 비표본오차 평가를 위해 다른 리스트를 이용
  : 조사단위의 포함오차를 분석하거나 두 관찰값 차이 비교분석

- 인구주택 총조사에 대한 점검: 최근 출생기록부, 양로연금 수혜자, 학생명부 등을 이용할 수 있음
  – 총조사의 질(누락비율 등) 평가, 모집단의 규모 수정

3. 외부기록 점검과 품질관리기법

▶ 품질관리기법

- 조사결과도 제품과 같이 여러 절차를 거쳐 생산
  ➡ 통계적 품질관리기법 적용 가능

- 관리도, 합격샘플링기법
  : 대규모 조사 데이터의 질 평가 및 신뢰성 제고 활용

- 조사에서 비표본오차의 관리에 품질관리기법 적용 가능

4. 중복부차표본기법 – ① 연결부차표본

- 동일한 표본설계로 추출한 k개의 부차표본에서 추정치 계산
- k개 부차표본 추정치를 비교 분석하여 편향과 비표본오차 측정

- 조사원간의 편향 분석 목적, 부차표본을 연관성 갖도록 구성
- k개의 동질적인 부차표본을 각 조사원이 조사하도록 하여 부차표본별로 조사한 데이터에서 추정치를 산출하여 비교하는 방법
- 추정치 분석을 통해서 비표본편향의 차이는 파악할 수 있지만, 편향의 크기와 방향은 알 수 없음

4. 중복부차표본기법 – ② 독립부차표본

- 정확한 모수 추정이 주목적이고, 상이한 과정의 비교는 부차적 목적임
- 동일한 표본설계로 k개 부차표본을 독립적으로 추출하여 각 부차표본 별로 다른 조사원이 조사 후 추정치를 계산하여 비교 분석하는 방법
- 많은 특성을 조사하는 대규모 조사에서 총 변동을 분석할 때 효과적임

4. 중복부차표본기법 – ② 독립부차표본

- 모수(θ)를 추정하기 위해서 k개 독립부차표본을 조사하고, 추정치 계산한 후 추정량 ($\hat{\theta}$) 의 분산 추정치 아래 식 계산

$$\hat{V}(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{i=1}^{k} (\hat{\theta}_i - \hat{\theta})^2 \ , \ \hat{\theta} = \frac{1}{k} \sum_{i=1}^{k} \hat{\theta}_i$$

- 만일 모집단의 비(ratio, 比)를 추정할 경우 비 추정량($\bar{y}/\bar{x}$)의 분산은 아래 식으로 추정할 수 있음

$$\hat{V}(\bar{r}) = \frac{1}{k(k-1)} \sum_{i=1}^{k} (r_i - \bar{r})^2 \ , \ \bar{r} = \frac{1}{k} \sum_{i=1}^{k} r_i \ , \ r_i = y_i/x_i$$

학/습/목/차

1. 표본조사 오차의 개요

2. 편향의 종류

3. 비표본오차의 종류

4. 비표본오차의 측정

5. 항목 무응답 대체법

1. 무응답 대체법의 개요

단위 무응답 (unit nonresponse)	응답 거부 등의 이유로 발생한 무응답
항목 무응답 (item nonresponse)	일부 항목에서 자료수집 못한 경우

- 단위 무응답의 보정은 가중치 보정 : 응답자와 특성 유사성 가정

> 10명 조사 : 2명 무응답, 8명 유효응답이면 8명 데이터에
> 1.25(=10/8)의 가중치 보정하여 해석함

- 항목 무응답 문항별 응답자 수가 다를 경우 각 조사변수의 분석 결과에 통계적 정확도에 차이 발생

- 무응답 항목과 유사한 특성 값을 대체하여 응답한 것처럼 분석

2. 평균 대체(mean imputation)

- 조사대상자 유사 특성 별로 몇 개 층으로 분류 (무응답 대체 층)
- 무응답 대체 층 : 응답자가 무응답자보다 많도록 층 구분
- 무응답 항목 대체값을 동일 층 내의 응답자의 평균값을 이용
- 무응답자와 응답자간의 특성이 유사하다는 가정 사용
- 동일 층에 무응답자 많은 경우 표본분포가 왜곡 될 수 있음
- 대규모 조사에서 무응답 대체층의 방법에 따라 상이한 결과
- 추정량 분산을 과소 추정하는 경향이 있으므로 주의

남, 여 회사원 5명씩 월소득 조사에서 남여 각 1명씩 무응답

- 남자 4명의 평균 275만원 무응답 남자 회사원 월소득 대체
- 여자 4명의 평균 230만원 무응답 여자 회사원 월소득 대체

3. 최근방 대체(nearest neighbor imputation)

- 전체 표본을 유사한 특성별로 구분하여 대체 층을 만들고, 각 대체 층 내에서 지역별, 특성별로 기준을 정하여 정렬
- 항목 결측값 조사단위의 바로 앞 단위의 항목 응답값으로 대체
- 간편하지만 무응답이 많을 때 모평균이나 모분산 추정에 편향 발생

5명의 회사원을 근무년수 기준으로 정렬한 경우

- 3번째 남자가 무응답이면 두 번째 남자회사원 월소득 278만원을 대체값으로 적용함

- 일정한 기준으로 정렬한 경우 최근방 조사단위의 특성이 유사하다는 전제로 적용함
- 편리성이나 비용적인 측면에서 효과적인 대체법으로 평가됨

4. 회귀 대체(regression imputation)

- 항목 무응답 있는 문항을 종속변수로 하고, 그와 연관성이 큰 문항들을 독립변수로 하여 선형회귀모형을 적합하여 얻은 예측값을 이용하여 항목 무응답을 대체하는 방법
- 만일 무응답 항목과 독립변수 간의 회귀모형 관계가 과거 자료 분석 또는 타당한 이론적 근거로 주어졌다면 안정적인 무응답 대체 가능
- 남여 회사원의 월소득 조사에서 근무년수가 주어졌다면 월소득을 종속변수(y), 근무년수를 독립변수(x)로 하는 선형회귀모형을 가정할 수 있음
- 선형회귀추정식 y = 150 + 12.5 * x 이 주어졌으면, x가 11년인 남자 회사원이 월 소득을 응답하지 않은 경우는 287.5만원을 대체값으로 사용하게 됨

5. 핫덱 대체(hot-deck imputation)

- 무응답 대체 층 내에서 무응답 조사단위와 응답 조사단위가 무응답 항목에 대해서 유사할 것으로 가정 하에서 적용함
- 무응답 대체층 내에서 조사단위를 정렬할 적당한 변수가 없을 경우에 적용 가능
  ➡ 최근방 대체법을 적용할 수 없음
- 응답한 조사단위 중 랜덤하게 선정하여 결측치의 대체값 이용
  ➡ 생성되는 난수에 따라 다른 대체값을 얻게 됨
- 평균대체에서 발생하는 조사항목의 표본분포 왜곡 문제를 완화할 수 있음

이 강의는
강의용 휴대폰(U-KNOU 서비스 휴대폰)으로도
다시 볼 수 있습니다.