

1 제 1장: 표본조사의 개요

1.1 조사와 실험

1. 조사: survey

관찰연구, 인과성, sampling design

2. 실험: experiment

요인 (factor), 조절 (control), experimental design

3. 예제: 흡연과 폐암의 관계

4. 시청률조사, 물가지수 선정, 선거에서 후보의 득표율 예측, 기업의 시장조사 등등

1.2 모집단과 임의표본

1. 모집단

2. 표본

3. 대표성 (representativeness)

4. 임의표본 (random sample)과 임의화(randomization)

5. 센서스 (census), 인구및 주택 총조사

6. 전수조사

1.3 용어

1. 원소 (element), 관찰단위 (observation unit)

2. 표적모집단 (target population)

3. 표집단위(추출단위, sampling unit), 기본단위

가구조사에서 반을 뽑고 그 안의 모든 가구를 조사하는 경우

4. 일차 표집단위, 이차 표집단위

5. 표집틀 (sampling frame), 틀 (frame): list frame, map frame, card frame 등등

6. 표집틀 모집단 (sampling frame population), 표집 모집단 (sampled population)

[그림 1.1]

1.4 추정의 정확성

1. 확률변수(random variable), 기대값(expectation), 표본분포(sampling distribution)

2. 교재의 복원 추출과 비복원 추출 예제, 예제 1.1, 1.2.

- 크기 N 인 모집단에서 n 개를 비복원 추출하는 확률은 $P(S) = 1/\binom{N}{n}$
- 모집단의 어느 특정 단위 U_i 가 표본에 포함되는 확률은 $P(U_i \in S) = n/N$
- 어느 특정 단위 U_i 가 m 번째 추출에서 뽑히는 확률은 $1/N$ 이다.

3. 관심있는 모수 θ 에 대한 추정량 $\hat{\theta}$

4. 편향 (bias) = $\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$

5. 분산 (variance):

$$\text{var}(\hat{\theta}) = \sigma^2(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2 = E(\hat{\theta}^2) - (E(\hat{\theta}))^2.$$

6. 상대표준오차 (relative standard error) 또는 변동계수 (coefficient of variation, CV):

$$\text{CV}(\hat{\theta}) = \frac{\sigma(\hat{\theta})}{E(\hat{\theta})}.$$

7. 평균제곱오차(MSE, mean squared error)

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 \\ &= E(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2 \\ &= \text{var}(\hat{\theta}) + \text{bias}^2(\hat{\theta}). \end{aligned}$$

8. 상대효율 (relative efficiency): 두 추정량 $\hat{\theta}_1$ 과 $\hat{\theta}_2$ 가 존재할 때, relative efficiency of $\hat{\theta}_1$ to $\hat{\theta}_2$ is defined as

$$\text{MSE}(\hat{\theta}_2) / \text{MSE}(\hat{\theta}_1).$$

1.5 신뢰구간

1. Interval estimation, Confidence interval, Confidence level

2. Central limit theorem:

Y_1, Y_2, \dots, Y_n are IID from a distribution with (μ, σ^2) and $0 < \sigma^2 < \infty$ 이면

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \Rightarrow N(0, 1).$$

이를 다른 식으로 표현하면

$$\sup_z \left| P \left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq z \right) - \Phi(z) \right| \rightarrow 0.$$

3. Confidence interval for population μ or CI for general parameter θ :

$$\hat{\theta} \pm z_{\alpha/2} \cdot \text{s.e.}(\hat{\theta}).$$

4. 표준오차(standard error, s.e.): 추정량의 표준편차
5. 오차한계 (error bound): $z_{\alpha/2} \cdot \text{s.e.}(\hat{\theta})$

1.6 여론조사방법

여론조사 (opinion poll)은 어떤 특정한 사안에 대하여 대중의 의견을 묻는 표본조사이다. 자세한 내용은 각자 책을 정독하기 바라고 일부 중요한 개념만을 살핀다.

1. 자료수집 방법 으로는 개별 면접, 전화조사, 설문조사, 직접관찰(교통량 조사, 농축산물 통계) 등이 있다.
2. 1.7.2절을 보면 표집 방법은 확률표집과 비확률표집으로 분류되고 확률표집이란 계획된 임의화(randomization)에 의하여 추출된 표집을 일컫고 이 경우 모집단의 각 원소들이 표본으로 선택될 확률이 0보다 크면 그 확률들을 정확히 알고 있어서 추정값의 신뢰성 등을 계산할 수 있다. 반면 비확률표집은 이러한 성질을 지니지 못하는 표집방법이다.

비확률표집이 수리적 분석이 불가능 하나 종종 전문가들의 주관적인 지식들을 이용할 수 있어 확률표집보다 정확할 수도 있다.

예시: 20대 대선에서의 전화 여론조사

3. 층화추출(stratified sampling)과 집락추출(cluster sampling)

층화추출: 한국은행 물가지수 조사: 백화점/ 대형 슈퍼체인/ 슈퍼/ 편의점

집락추출:

만일 전국 단위의 유치원 생들을 상대로한 조사의 경우 유치원의 명단은 구하기 쉽지만 유치원 재원생의 명단은 구하기 어렵다. 따라서 유치원을 표집단위로 하고 sample된 유치원에 대하여 원생들을 전수조사하거나 simple random sampling을 하게된다.

집락추출에서는 각 집락의 크기를 사전에 알기가 힘든 경우가 많고 이에 따른 어려움이 있을 수 있다.

4. 계통추출법:

공장의 컨베이어 벨트에서 일정 간격을 두고 표본을 뽑아 불량 검사를 한다.

출구조사에서도 사전에 계획된 투표소에 대하여 일정 간격으로 투표자를 뽑아 조사를 한다.

5. 전국단위 조사 부서

대부분의 조사 + 인구및 주택 총조사: 통계청

농업통계: 농림수산부

산업통계: 산업자원부

경제통계: 한국은행, 통계청

6. 인터넷조사

최근 컴퓨터 이용의 확대와 더불어 internet/smart survey에 대한 관심이 증대되고 있다. 교재에서 언급한 바와 같이 인터넷조사는 지리적 공간에 대한 제약이 없고 조사 비용이 저렴하다는 장점이 있으나 단점으로는 모집단을 대표하는 확률 표집방법이 알려져 있지 않아 모집단에 대한 대표성이 떨어진다는 점이다.

현재 한국에서 시행되는 대표적인 인터넷조사는 방송사의 인기순위조사(음악프로, 교양프로의 순위 조사)를 생각 할 수있다.

7. 조사오류

표본조사의 오류로는 크게 표집오차와 비표집오차가 있다.

표집오차란 모집단 전체를 조사하지 않고 일부를 조사함으로써 생기는 오차를 의미한다.

비표집오차란 표본간의 변동에 기인한다고 볼 수 없는 오차로 다양한 원인이 있을 수 있다. 예를 들어 편향된 질문, 우리나라의 경우 국민의힘, 더민당 효과, 언더독 효과등을 생각할 수 있다.

비표집오차의 한 종류인 포함오차(error of coverage)에 대하여도 교재의 내용을 파악하기 바란다.

8. 무응답

교재에서 설명한 판별분석을 통한 imputation 또는 K-근접법등에 대하여 알아본다.

9. 표본조사의 단계

- 명확한 조사목적에 대한 진술
- 표집모집단에 대한 규정
- 수집될 데이터의 선정 - 설문설계
- 정밀도의 결정
- 측정방법의 결정
- 표집틀의 설정
- 표본선택
- 사전조사 또는 예비조사
- 실사계획및 구성
- 데이터 요약과 분석
- 차후조사를 위한 정보획득

10. 공직선거법 108조(여론조사의 결과공표금지 등)

<http://law.go.kr/법령/공직선거법/제108조>