표본크기 결정

* 수업시간 : 화07,08 목07,08

* 교 수 명 : 변 종 석

* 이메일 : jsbyun@hs.ac.kr

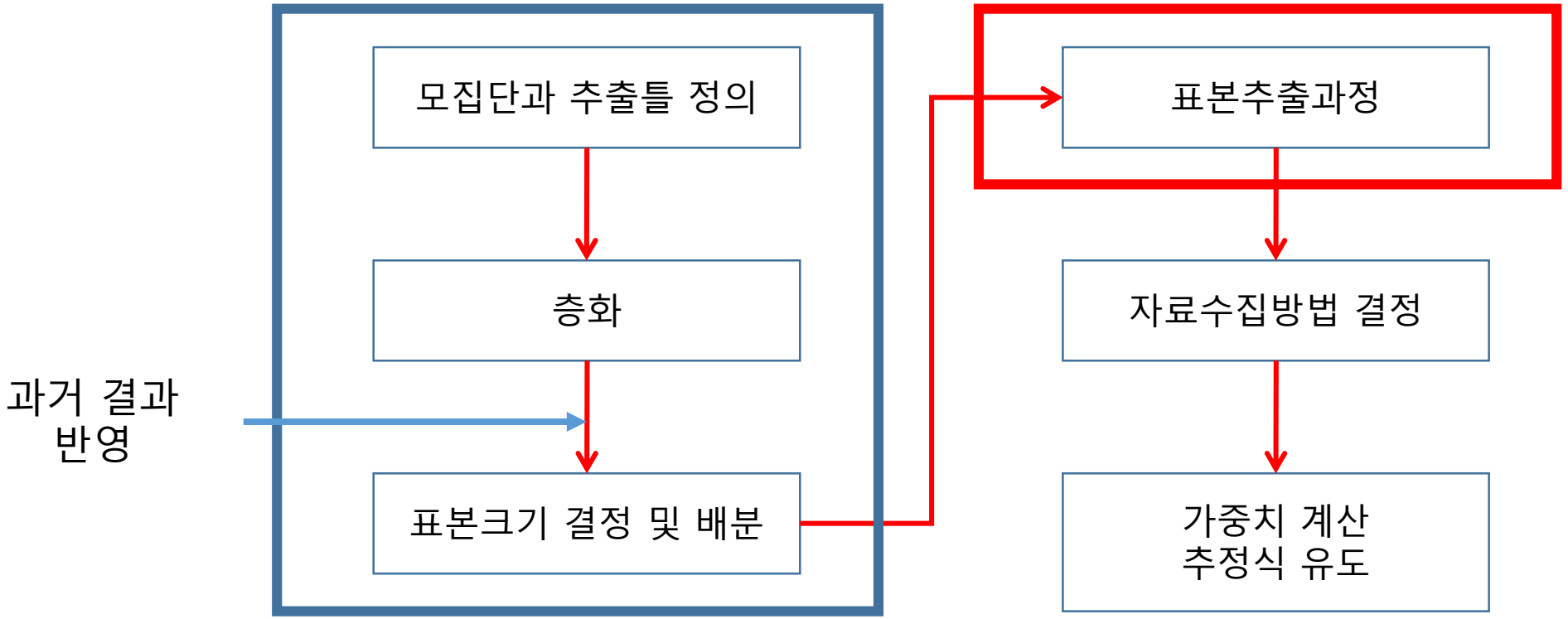
[주제 2] 통계조사의 개요

2-4. 표본크기 결정 과정

[1] 표본크기 결정 과정
[2] 표본크기 계산

2-4. 표본크기 결정 과정

- **[Review] 표본설계의 주요 과정**



1. 표본크기 결정 과정

1) 모집단 분석

- 조사모집단 설정
- 표본추출틀 분석
  - 모집단 특성 파악
  - 층화 및 집락화를 위한 분석 : 군집분석, 분산분석 등
  - 조사의 주요 문항(변수) 특성 분석 : 평균, 표준편차, cv 등

2) 층화/집락화

- 대표성 : 모집단 포함율 및 이질성 반영
- 정확성 : 표본오차 감소
- 효율성 : 표본추출 편의성 및 경제성

3) 표본자료 분석

- 설계 시점 이전의 표본조사 자료 분석
- 목적
  - 현행 표본설계의 표본오차 달성 수준 평가
  - 새로운 표본설계의 목표 오차 설정에 활용
  - 표본크기 결정에 반영

4) 목표 오차 설정

- 조사 목표 및 분석 단위(별)의 목표 오차 수준 설정
- 참고1 : 목표 오차 설정 유형
  - 절대 오차 $|\theta - \hat{\theta}|$
  - 상대 오차 $\frac{|\theta - \hat{\theta}|}{\theta}$
  - 상대표준오차 $CV_{\hat{\theta}} = \frac{\hat{\sigma}_{\hat{\theta}}}{\hat{\theta}} \approx \frac{s_{\hat{\theta}}}{\hat{\theta}}$
- 참고2 : 정확성 평가 기준(상대표준오차의 허용 범위)
  - 캐나다 통계청
  - Kish 기준
  - 호주 통계청
  - 통계청

1. 캐나다 통계청의 표본조사 기준

- 0.00% ~ 4.99 : 매우 우수(Excellent)
- 5.00% ~ 9.99% : 우수(Very Good)
- 10.00% ~ 14.99% : 좋음(Good)
- 15.00% ~ 24.99% : 허용 가능(Acceptable)
- 25.00% ~ 34.99% : 주의사항과 함께 사용가능(Use with caution)
- 35.00% : 공표 시 신뢰불가(Too unreliable to publish)

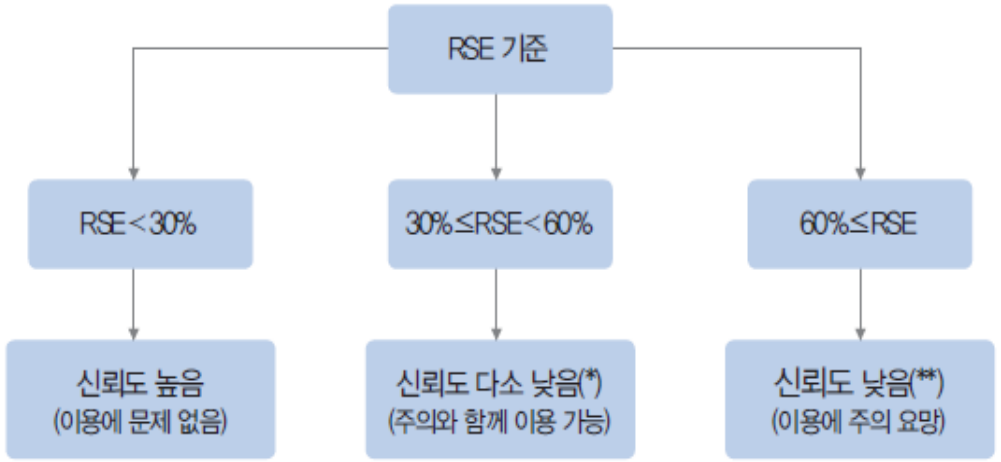
2. Kish 기준

- 10% 이하 : 우수(Sufficiently good)
- 20% 이하 : 허용 가능(tolerable)

* 출처 : Kish(1965), Survey Sampling, p. 218

3. 호주 통계청의 표본조사 기준

- 상대표준오차가 25% 이하는 대부분 목적에 그대로 사용
- 상대표준오차 25~50%는 * 표시를 하여 주의하여 사용
- 50% 이상은 **을 표시하여 신뢰가 부족하니 이용 시 주의 바람



5) 목표 표본크기 결정

- 조사 목표 및 분석 단위의 목표 오차 달성에 필요한 크기 결정
- 고려 사항
  - 조사 목표 : 모수
  - 모집단 크기
  - 목표 오차 : 모분산
  - 신뢰수준
  - 응답률
  - 조사 비용
  - 기타 : 조사 기간, 조사원의 인력 등

2. 표본크기 계산

- 표본크기 결정을 위한 기본 고려 요소
  - 조사목적과 여건을 고려하여 결정
  - 일반적으로 표본추출방법, 목표 오차, 신뢰수준, 분포, 비용, 응답률 등을 고려해 결정

1. 단순확률추출법의 표본크기 계산 과정

가) 목표 정도(=추정오차한계) 설정 ; 표본조사에서 예상하는 표본오차(=추정오차)에 대한 목표 수준(절대 오차 기준, B)를 미리 설정

추정오차(Error of Estimation) $= \left|\theta - \hat{\theta}\right| < B$

여기서, θ 는 모수, $\hat{\theta}$ 은 표본 추정량을 의미

나) 신뢰수준($1-\alpha$) 설정 $\Pr\left(\left|\theta - \hat{\theta}\right| < B\right) = 1-\alpha$

여기서, $B = 1.96\,\sigma_{\hat{\theta}}$ 의미

다) 최소 비용조건에서 목표 오차수준을 달성할 수 있는 적절한 표본 크기를 결정
- 일반적으로 비용보다는 최소 분산을 갖는 조건에서 오차한계를 달성할 수 있는 표본크기를 계산

1) 모평균 추정을 위한 표본크기 결정(단순확률추출법 가정)

- 산출 식 : 95% 신뢰수준에서 모평균 추정을 위한 표본크기의 결정

  - 표본추출 : 복원 추출

$$n_0 = 1.96^2 \frac{\sigma^2}{B^2} \qquad \left(\text{참고} \ : \ B = 1.96\,\sigma_{\bar{X}} = 1.96\,\frac{\sigma}{\sqrt{n}} \right)$$

  - 표본추출 : 비복원 추출 혹은 유한모집단

$$n = \frac{1.96^2\sigma^2}{B^2 + \frac{1.96^2\sigma^2}{N}} = \frac{n_0}{1 + n_0/N}$$

$$\left(\text{참고} \ : \ B = 1.96\,\sigma_{\bar{X}} \approx 1.96\,\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N}} \right)$$

- 예: N = 5000인 A기업에서 직원의 평균 문화비를 조사하는 경우, 추정오차 한계가 1,000원이내인 결과를 95% 신뢰수준에서 얻고자 할 때 필요한 표본크기는?(과거 조사 결과가 없어서 모분산 파악을 위한 50명을 사전조사한 결과, 표준편차는 6000원으로 파악된 것으로 가정)

$$n_0 = 1.96^2 \frac{6000^2}{1000^2} = 138.3$$

$$n = \frac{n_0}{1 + n_0/N} = \frac{138.3}{1 + 138.3/5000}$$

$$= \frac{1.96^2 \times 6000^2}{1000^2 + \frac{1.96^2 \times 6000^2}{5000}} \approx 135$$

2) 모비율 추정을 위한 표본크기 결정(단순확률추출법 가정)

- 산출 식 : 95% 신뢰수준에서 모비율 추정을 위한 표본크기의 결정

  - 표본추출 : 복원 추출

$$n_0 = 1.96^2 \frac{PQ}{B^2} \qquad \left(\because \quad B = 1.96\,\sigma_{\hat{p}} = 1.96 \frac{\sqrt{PQ}}{\sqrt{n}} \right)$$

  - 표본추출 : 비복원 추출 혹은 유한모집단

$$n = \frac{1.96^2 PQ}{B^2 + \frac{1.96^2 PQ}{N}} = \frac{n_0}{1 + n_0 / N}$$

$$\left(\because \quad B = 1.96\,\sigma_{\hat{p}} \approx 1.96 \frac{\sqrt{PQ}}{\sqrt{n}} \sqrt{\frac{N-n}{N}} \right)$$

- 예: N = 5000인 A기업에서 새로운 복지제도에 찬성하는 직원의 비율을 조사하고자 하는 경우, 추정오차 한계가 0.05이내인 결과를 95% 신뢰수준에서 얻고자 할 때 필요한 표본크기는?

$$n_0 = 1.96^2 \frac{0.5 \times 0.5}{0.05^2} = 384.2$$

$$n = \frac{n_0}{1 + n_0 / N} = \frac{384.2}{1 + 384.2 / 5000}$$

$$= \frac{1.96^2 \times 0.5 \times 0.5}{1000^2 + \frac{1.96^2 \times 0.5 \times 0.5}{5000}} \approx 357$$

참고 : 일반적으로 모비율 추정을 위한 표본크기 계산에서 모비율의 분산에 대한 사전 정보가 없다면 모분산이 최대가 되는 PQ(=0.25=05*0.5)를 사용하여 최대 표본오차를 기준으로 표본크기를 결정

- **사례1 : 모비율 추정을 목적으로 하는 표본조사의 표본크기 계산**

1) 표본크기 및 목표오차

○ 「가공식품 소비자태도 조사」의 표본크기는 가용 예산과 조사 소요시간 등의 조사에 필요한 제반 여건을 고려하고, 작성되는 통계의 표본오차 수준을 검토하여 2,000가구 내외를 목표로 하였다. 본 조사에서는 전국에서 200개 표본 조사구를 추출하고, 각 표본 조사구로부터 10가구의 표본가구를 추출하여 조사하는 것을 원칙으로 한다.

○ 본 조사의 표본오차는 95% 신뢰수준에서 ± 2.18%p이내가 된다.

2. 일반적인 표본조사에서의 표본크기 관련 식

- 상대표준오차를 이용한 표본크기 결정
  - 표본조사의 목표 상대표준오차(=Relative Standard Error)를 설정하여 표본크기를 결정

$$n_o = \left(\frac{CV_{\hat{\theta}}}{d} \right)^2 \quad \Rightarrow \quad n = \frac{n_0}{1 + n_0 / N}$$

여기서 CV(=표준편차/평균)는 모집단의 변동계수,
d는 목표 상대표준오차(추정량의 변이계수), N은 모집단 크기

$$\hat{d} = CV_{\hat{\theta}} / \sqrt{n}$$

- **사례2 : 상대표준오차 기준 표본크기 계산**

사회서비스 수요 실태조사의 표본은 행정자치부 주민등록인구통계(2015년 6월 기준)의 세대수에 따라 규모를 산정하였으며, 95% 신뢰수준 하에서 모집단 크기가 작은 제주특별자치도와 세종특별자치시를 제외한 지역별 상대표준오차가 7.0% 내외가 되도록 설계하였다.

- 계속조사 : 이전 조사 결과를 반영한 표본크기 결정
  - 이전 조사의 목표 상대표준오차(=Relative Standard Error)를 이용하여 표본크기를 결정

$$n_0 = n'\left(\frac{CV'}{CV}\right)^2 \quad \Rightarrow \quad n = \frac{n_0}{1 + n_0/N}$$

여기서 n' 은 이전조사의 표본크기, CV' 는 이전 조사의 상대표준오차,
CV 는 목표 상대표준오차, N은 모집단 크기

● 사례3 : 계속조사의 표본크기 계산

현행 조사에 포함되지 않은 교육서비스업(P), 보건업 및 사회복지서비스업(Q)을 제외한 산업들은 2013년, 2014년, 2015년의 상대표준오차를 이용하여 각 년도 기준으로 새로운 표본설계의 산업중분류별 표본크기(m_{gh})를 다음과 같은 식으로 계산하고, 3개년 평균으로 절충하여 표본크기를 산출한다.

$$m_{gh}^{*} = n_{gh} \cdot \left(\frac{CV}{CV^{*}}\right)^2$$

$$m_{gh}^{1} = \frac{m_{gh}^{*}}{1 + \frac{m_{gh}^{*}}{N_{gh}}}$$

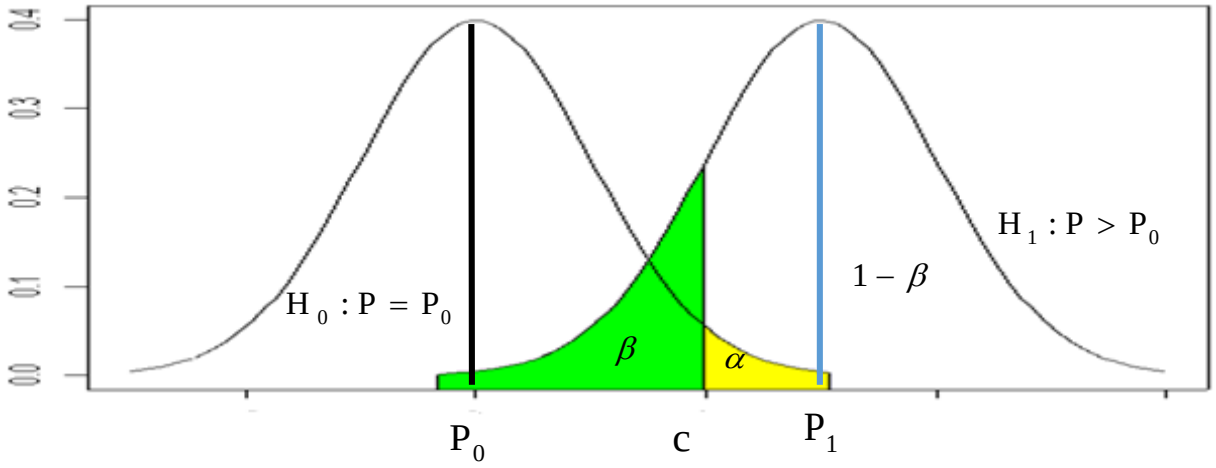
- 여기서 N_{gh}는 g산업대분류, h산업중분류의 모집단 크기, m_{gh}^{*}는 무한모집단을 가정한 표본크기, m_{gh}^{1}는 유한모집단 수정계수를 반영한 표본크기이다. 즉, m_{gh}^{1}는 g산업대분류, h산업중분류의 기본적으로 필요한 표본크기이다.

- 참고 : 표본크기 결정
  - 통계학적 연구 방법에서는 추론 및 분석의 오차 수준을 고려하여 표본 크기를 결정함
    - 표본조사연구
    - 실험연구
    - 관찰연구
  - 비교연구 및 인과성분석도 분석의 오차 수준을 고려하여 표본크기를 결정함
    - 임상실험 연구
    - 동등성 연구

3. 가설검정 및 비교연구를 위한 표본크기 결정

- 검정력(=power)을 고려한 표본크기를 계산

- 예1 : 한 모집단에서 모비율의 가설검정을 위한 표본크기(단측검정)



$$c = P_0 + z_{1-\alpha}\sqrt{P_0(1-P_0)/n} \quad \text{under} \quad H_0$$

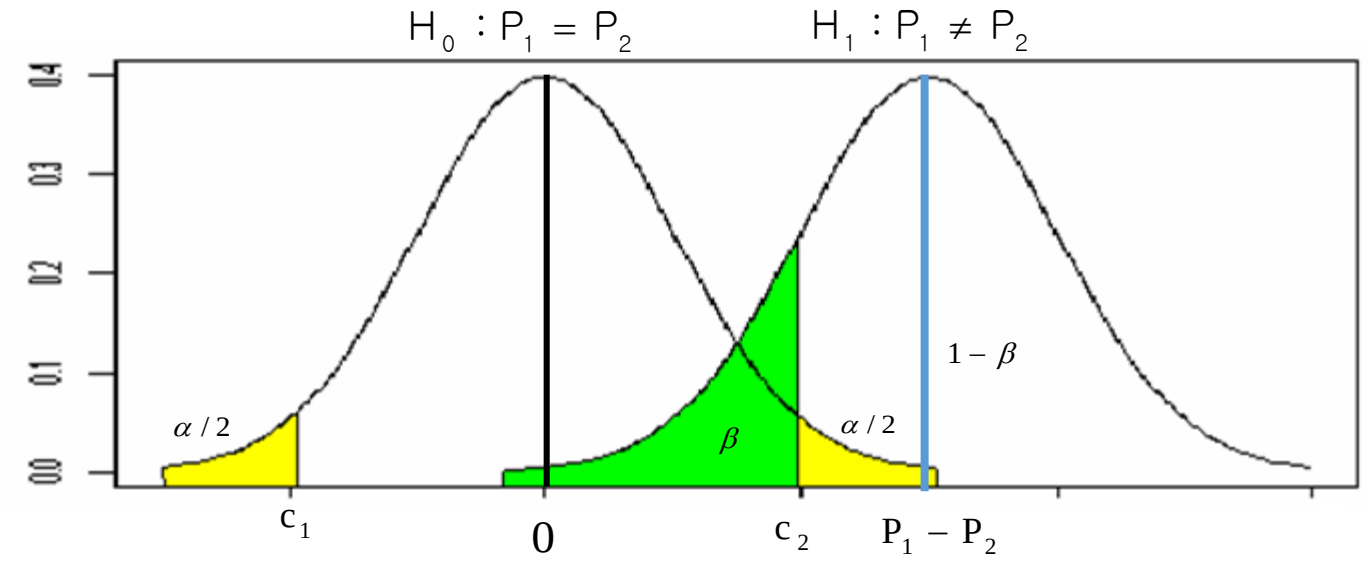
$$c = P_1 - z_{1-\beta}\sqrt{P_1(1-P_1)/n} \quad \text{under} \quad H_1$$

$$\Rightarrow \quad c = P_0 + z_{1-\alpha}\sqrt{P_0(1-P_0)/n} = P_1 - z_{1-\beta}\sqrt{P_1(1-P_1)/n}$$

$$\Downarrow$$

$$n = \frac{\left\{z_{1-\alpha}\sqrt{P_0(1-P_0)} + z_{1-\beta}\sqrt{P_1(1-P_1)}\right\}^2}{(P_1-P_0)^2}$$

- 예2 : 두 모비율 차이에 대한 가설 검정(양측 검정)



$$c_2 = 0 + z_{1-\alpha/2}\sqrt{2\bar{P}(1-\bar{P})/n} \quad \text{under} \quad H_0, \quad \text{where} \quad \bar{P} = (P_1 + P_2)/2$$

$$c_2 = (P_1 - P_2) - z_{1-\beta}\sqrt{(1/n)\left\{P_1(1-P_1) + P_2(1-P_2)\right\}} \quad \text{under} \quad H_1, n_1 = n_2 = n$$

$$\Rightarrow \quad n = \frac{\left\{ z_{1-\alpha/2}\sqrt{2\,\bar{P}(1-\bar{P})} + z_{1-\beta}\sqrt{P_1(1-P_1) + P_2(1-P_2)} \right\}^2}{(P_1 - P_2)^2}$$

감사합니다

한신대학교