2강 표본조사의 기본개념

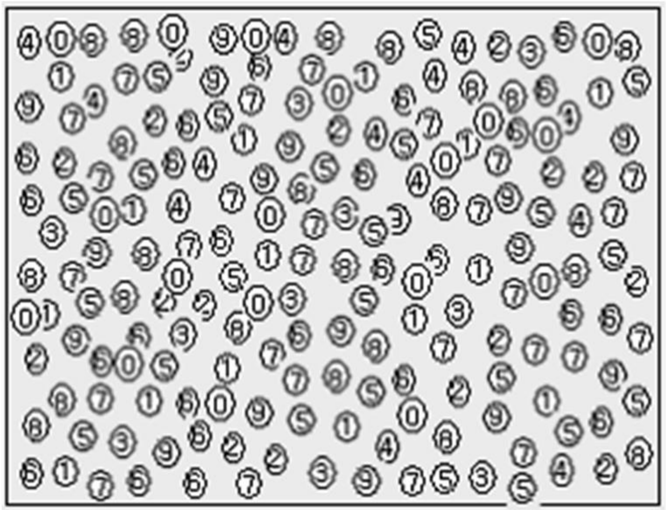
정보통계학과 이기재교수

학/습/목/차

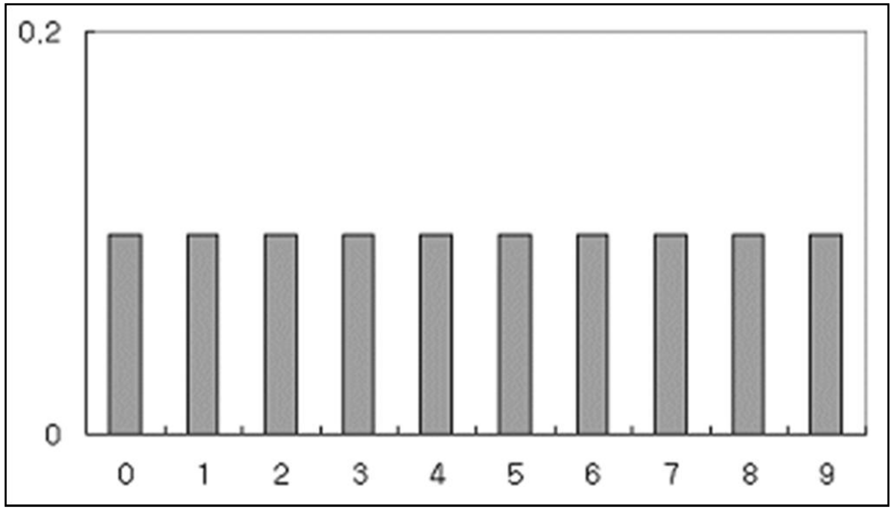
1. 모집단 분포
2. 표본분포
3. 추정
4. 엑셀을 활용한 실습

가상의 모집단 예

모집단의 분포



모집단 분포의 히스토그램



가상의 모집단 예

1. 모수

- 모집단의 특성값

  (예) 모평균, 모분산, 모비율, 모총계 등

2. 모평균

- 모집단의 중심위치의 척도
- $\mu = E(y)$

3. 모분산

- 모집단에서 각 단위들이 모평균으로부터 흩어진 정도
- $$\begin{aligned} V(y) &= E(y-\mu)^2 \\ &= \sum_y (y-\mu)^2 \cdot p(y) \\ &= \sum_y y^2 p(y) - \mu^2 \\ &= \sigma^2 \end{aligned}$$

가상의 모집단 예

예제

- $$E(y) = \sum_y y \cdot p(y) \\ = 0 \cdot p(0) + 1 \cdot p(1) + \cdots + 9 \cdot p(9)$$

- $$\sigma^2 = \sum_y y^2 p(y) - \mu^2 \\ = \frac{1}{10}(0^2 + 1^2 + \cdots + 9^2) - (4.5)^2 \\ = 8.25$$

모수추정

표본조사의 목적은
표본의 데이터로 모수(모집단 특성치)를 추론하는 것

모수	추정량
모평균 (μ)	표본평균 ($\bar{y} = \sum_{i=1}^{n} y_i / n$)
모분산 (σ^2)	표본분산 ($s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1)$)

모수추정

예제 1–1

▶ 6개의 표본 데이터 : 3 0 9 8 5 2

▶ 표본평균 :

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{6}(3+0+9+8+5+2) = 4.5$$

▶ 표본분산 :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{6-1}\left[(3-4.5)^2 + (0-4.5)^2 + \cdots + (2-4.5)^2\right] = 12.3$$

학/습/목/차

1. 모집단 분포
2. 표본분포
3. 추정
4. 엑셀을 활용한 실습

표본분포

1. 표본추출변동

▶ 동일한 모집단에서 같은 표본추출방법으로 같은 크기의 표본을 추출할지라도 각 표본에서 계산된 추정량의 값은 표본마다 달라지는 것

2. 표본분포

▶ 같은 크기의 확률표본을 무한 반복해서 추출할 때 각 표본으로부터 계산되는 추정량이 이루는 분포

[표본평균의 표본분포]

가능한 모든 표본으로부터 표본평균( )$\bar{y}$ 계산하여

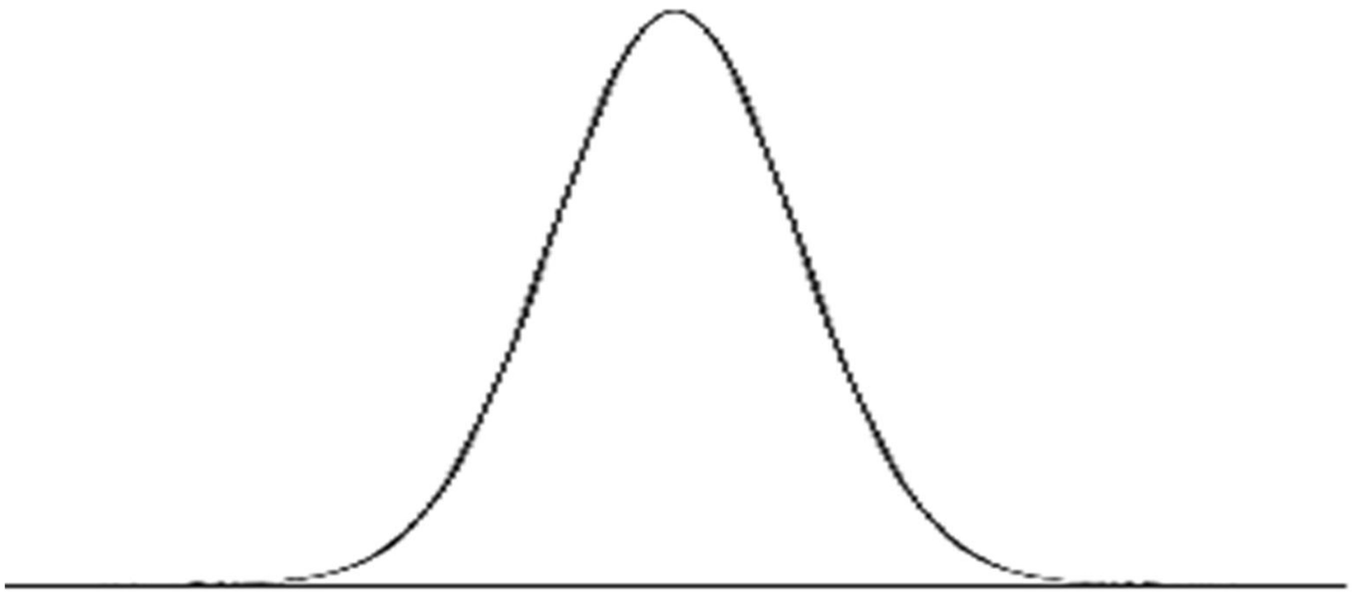
구한 표본평균($\bar{y}$의 분포

〈참고〉중심극한정리(central limit theorem)

표본분포

〈참고〉 중심극한정리(central limit theorem)

표본크기가 커지면 표본평균은 근사적으로 정규분포를 따름



무한집단에서 표본평균의 분포

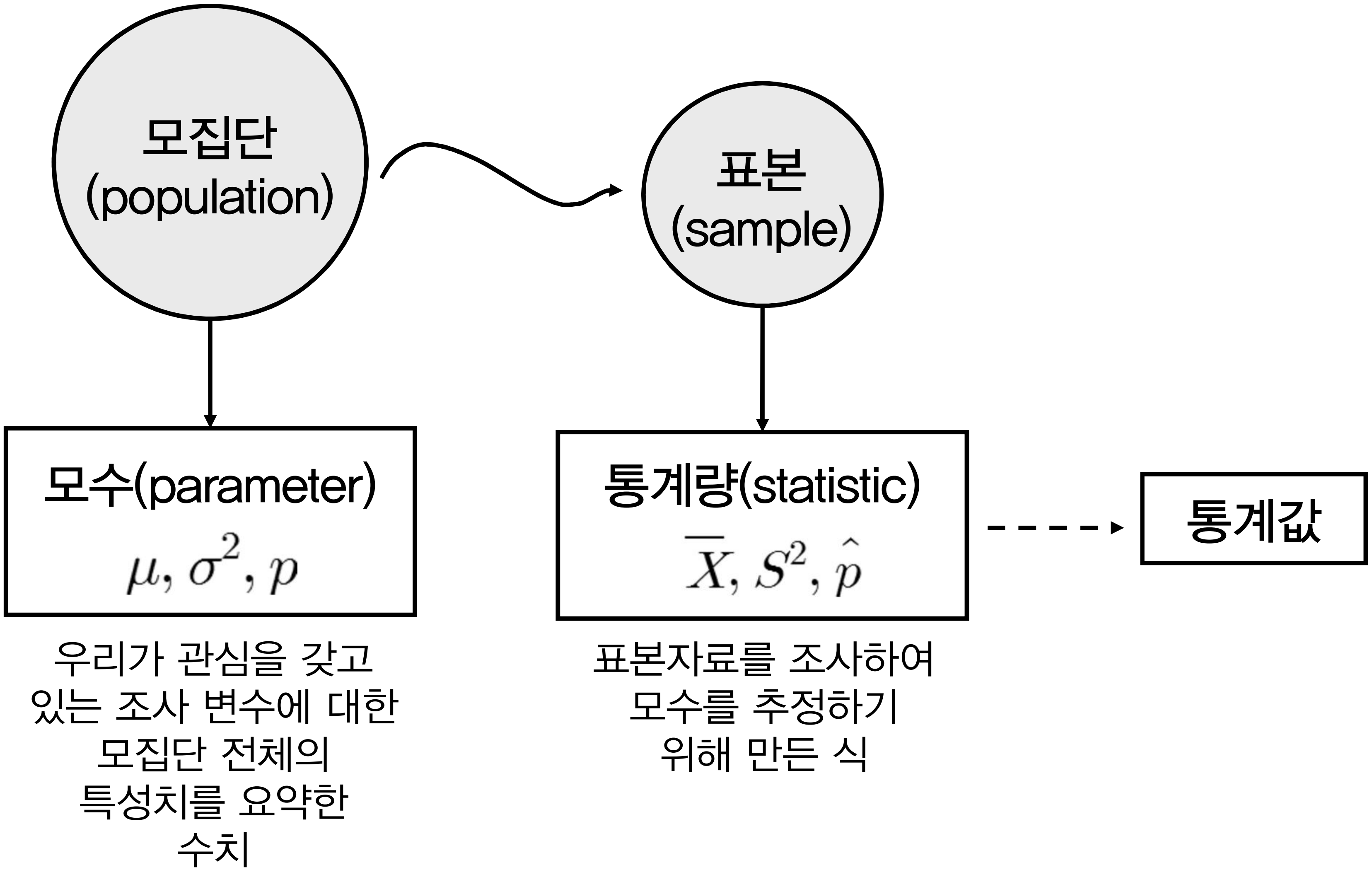
학/습/목/차

1. 모집단 분포
2. 표본 분포
3. 추정
4. 엑셀을 활용한 실습

추정의 핵심내용

- 추정량 유도
- 추정량의 정도(precision) 파악을 위한 추정량 분산 계산

표본추출과 추정량



모집단
(population)
표본
(sample)
모수(parameter)
μ, σ^2, p
통계량(statistic)
$\overline{X}, S^2, \hat{p}$
통계값
우리가 관심을 갖고
있는 조사 변수에 대한
모집단 전체의
특성치를 요약한
수치
표본자료를 조사하여
모수를 추정하기
위해 만든 식

바람직한 추정량의 성질

1. 비편향성(unbiasedness)

- 반복해서 표본을 추출할 때 표본으로부터 계산된 통계치가 모수를 과대 또는 과소 추정하는 경향이 없는 것

편향

비편향

2. 효율성(efficiency)

- 추정량 $\hat{\theta}_1, \hat{\theta}_2$을 비교할 때 만약 $\hat{\theta}_1$의 분산이 $\hat{\theta}_2$의 분산보다 작다면 $\hat{\theta}_1$ 이 $\hat{\theta}_2$보다 효율적이라고 함

(추정량의 정확도 평가 척도)

효율적

비효율적

바람직한 추정량의 성질

1	일반적으로 두 추정량이 모두 비편향 추정량이거나 편향을 무시할 수 있는 경우에는 두 추정량 중에서 분산이 작은 추정량을 사용해야 함
2	어떤 추정량의 분산이 작다는 의미는 같은 크기의 표본을 다시 반복해서 추출하여 통계치를 구한다고 할 때 구해진 통계치는 현재 구한 통계치와 유사한 값을 나타낼 것이라는 확률적 보증이라고 할 수 있음

표본오차 (sampling error)

표본에서 구한 결과와 센서스의 결과(모수)의 차이

▶ 표본오차

= ｜모집단의 참값(모수) – 모수에 대한 추정치｜ = $\left|\hat{\theta} - \theta\right|$

모집단의 일부를 표본추출하여 조사하여 추정함으로써 발생하는 우연적 오차

표본오차의 통계적 표현

- 추정량의 표준오차(standard error)
  - ▶ $\sqrt{V(\hat{\theta})}$
- 추정량의 100(1− α)% 신뢰구간(confidence interval)
  - ▶ $\hat{\theta} \pm z_{\alpha/2} \sqrt{V(\hat{\theta})}$
- 오차의 한계(bound of error)
  - ▶ $B = z_{\alpha/2} \sqrt{V(\hat{\theta})}$
- 추정량의 상대표준오차

  : 추정량의 정도(精度)를 나타내는 상대적 기준

  - ▶ $RSE(\hat{\theta}) = \dfrac{\sqrt{V(\hat{\theta})}}{\theta} \times 100$

〈참고〉 추정량의 변동계수(coefficient of variation : CV)라고도 함

표본오차의 통계적 표현

- 예제 1-1

⊙ 2명 조사하여 1명 지지한 경우와 2,000명 조사하여 1,000명 지지한 경우 모두 지지율은 50%인가?

▶ 2명 조사 : 오차의 한계 50%,
오차가 너무 커서 현재 추정값은 정보로서 가치 없음

▶ 2,000명 조사 : 오차의 한계 2%,
현재 추정값은 가치 있는 정보라고 할 수 있음

▶ 오차의 한계 계산: 여론조사 정확도에 대한 통계학적 근거

표본오차의 통계적 표현

예제 1–2

⊙ 도시가구들의 월평균 소득액 추정을 위한 표본조사

▶ 상대표준오차 $\widehat{RSE}(\hat{\theta}) = \dfrac{\sqrt{\hat{V}(\hat{\theta})}}{\theta} \times 100\ (\%)$

$$= \frac{10}{150} \times 100\ (\%) = 6.67\ (\%)$$

목표정도 (target precision)	표본조사를 기획할 때 설정하는 오차의 수준
달성정도 (attained precision)	표본조사 결과 얻어진 데이터로부터 계산한 오차의 수준

학/습/목/차

1. 모집단 분포
2. 표본분포
3. 추정
4. 엑셀을 활용한 실습

다음 페이지 〈실습하기〉에서 자세히 다룸