9강 계통추출법

정보통계학과 이기재교수

학/습/목/차

계통추출법 적용 사례 1

자동차 회사에서 지난 해 자동차 구매자를 대상으로 사용상 불만 사항을 조사한 사례

- 자동차 구매자 6만 명 중 1,000명을 표본으로 추출하여 조사할 예정
- 단순임의추출법에 의한 표본추출법
  - 고객들에게 1~ 60000번 사이의 고유번호 지정
  - 난수표에서 1~ 60000번 사이의 1000개 난수를 뽑아 이에 대응하는 표본추출

    ➔ 실제 표본추출 작업에 무척 번거로움

계통추출법 적용 사례 2

계통추출법에 의한 표본추출법

- 주어진 표본크기 1,000명은 60명 중 1명을 뽑는 셈
- 1에서 60사이의 숫자 중에서 임의의 난수를 난수표에서 선택
- 선택된 난수로부터 매 60번째 되는 단위들을 선택하여 1,000개의 표본 추출
- 예를 들어 1에서 60사이의 난수로 38이 뽑힌 경우

  38 98 158 218 278 338 59,978

계통추출법 적용 사례 3

- 사례 :
  - 어느 백화점의 입구에서 쇼핑을 마친 100명을 대상으로 조사하는 경우

    - 전체 쇼핑객 수 N을 알 수 없기 때문에 단순임의추출법을 적용하기 곤란함
    - 일정한 간격(예: 50명 중의 1명)으로 쇼핑객을 조사할 수 있음

- 사례 :
  - 생산현장에서 품질관리를 위한 샘플링검사

    - 어떤 조립라인에서 생산된 제품 중에서 매 50개째 되는 제품만을 검사하는 경우
      ➔ 1/50 계통표본

1/k 계통추출법(1-in-k systematic sampling) 정의

모집단의 추출틀에서 처음의 k개 단위들 중에서 랜덤하게 하나의 단위를 추출하고, 그 이후 매 k번째 간격마다 하나씩의 단위를 표본으로 추출하는 표본추출방법

계통표본 추출방법 1

- 추출틀에서 처음 k개 중에서 하나의 난수 $r(1 \le r \le k)$ 선택

  ➔ 랜덤출발점(starting point) : r

- 이후 $r, r+k, r+2k, \ \ldots \ , r+(n-1)k$ 들을 표본으로 추출

  ➔ 추출간격(sampling interval) : $k = \dfrac{N}{n}$

계통표본 추출방법 2

- $N = kn$인 경우
  - 랜덤출발점 : r , 추출간격 : $k = \frac{N}{n}$

	임의출발점					
	1	2	…	r	…	k
모집단 단위	U_1	U_2	…	U_r	…	U_k
	U_{1+k}	U_{2+k}	…	U_{r+k}	…	U_{2k}
	⋮	⋮	⋮	⋮	⋮	⋮
	$U_{1+(n-1)k}$	$U_{2+(n-1)k}$	…	$U_{r+(n-1)k}$	…	U_{nk}

  - 계통추출은 k개의 집단(표의 열에 해당)에서 하나의 집단을 뽑는 것과 동일함

계통표본 추출 사례 1

- 52명의 학생 중에서 6명의 학생을 계통추출법으로 추출하는 경우
  - 52명의 학생에게 1부터 52까지의 일련번호 부여
  - $N = 52$ 이고, $n = 6$ 이므로 추출간격 $k = \frac{N}{n} = \frac{52}{6} = 8.67$
  - 추출간격을 $k = 9$ 로 정하고 1에서 9사이의 난수를 하나 선택
  - 난수로 3이 뽑혔다면 다음 학생들이 표본으로 선택

    3, 12, 21, 30, 39, 48

계통표본 추출 사례 2 : 분수간격법

- 52명의 학생 중에서 6명의 학생을 계통추출법으로 추출하는 경우
  - 난수표에서 $0 < r \leq 8.67$ 의 난수를 선택
  - 난수표의 세 자리 숫자에서 $r = 1.04$ 가 선택되었다면

$$
\begin{aligned}
&1 < r = 1.04 < 2 \Rightarrow U_2 \\
&9 < 1.04 + 8.67 < 10 \Rightarrow U_{10} \\
&18 < 1.04 + 2 \times 8.67 < 19 \Rightarrow U_{19} \\
&27 < 1.04 + 3 \times 8.67 < 28 \Rightarrow U_{28} \\
&35 < 1.04 + 4 \times 8.67 < 36 \Rightarrow U_{36} \\
&44 < 1.04 + 5 \times 8.67 < 45 \Rightarrow U_{45}
\end{aligned}
$$

계통추출법의 장점

1	표본추출이 간편함

- 표본추출과정의 선택오차(selection error)를 줄일 수 있음

2	단순임의추출법의 대용으로 사용할 수 있음

3	일반적으로 모집단의 전체를 잘 반영함

계통추출법의 단점

1	조사된 표본자료로부터 추정량의 표준오차를 계산할 수 없음

- 단순임의표본으로 가정하여 표본오차 추정
- 모집단의 유형에 따라 실제보다 과대 또는 과소 평가될 수 있음

2	계통표본은 추출틀의 형태에 따라 추정의 정도(精度)에 차이가 생김

- 추출틀이 주기성을 갖고 있을 때는 계통추출법을 사용하면 곤란함

학/습/목/차

모평균의 추정 1

- 모평균의 추정량 : $\hat{\mu} = \bar{y}_{sy} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- 모평균 추정량의 분산 : $V(\bar{y}_{sy}) = \frac{N-1}{N}\frac{S^2}{n}[1+(n-1)\rho]$

  여기서 ρ : 급내상관계수

- 급내상관계수 ρ : 계통표본 내 표본단위들의 동질성의 정도를 나타냄

➔ 이론적인 값으로 실제 문제에서 추정할 수 없음

〈참고〉 단순임의추출법의 경우 : $V(\bar{y}) = \frac{N-n}{N}\frac{S^2}{n}$

모평균의 추정 2

- 모집단의 조사단위들이 랜덤하게 배열되어 있다고 가정할 수 있다면

  ➔ 계통추출법은 단순임의추출법의 대용으로 사용될 수 있음

  ➔ 얻어진 계통표본을 단순임의표본으로 간주함

  - $\hat{V}(\bar{y}_{sy}) = \frac{N-n}{N}\frac{s^2}{n}$, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_{sy})^2$
  - 모평균에 대한 $100(1-\alpha)\%$신뢰구간

    $\left(\bar{y}_{sy} - z_{\alpha/2}\sqrt{\hat{V}(\bar{y}_{sy})}, \bar{y}_{sy} + z_{\alpha/2}\sqrt{\hat{V}(\bar{y}_{sy})}\right)$

모평균의 추정 사례 1

- 어느 제약회사에서 생산되는 드링크류의 정량(120ml) 확인 목적의 조사

- 매 100번째 생산되는 병을 표본으로 선택하는 1/100계통추출법 적용

120.0	119.7	120.1	120.3
119.1	119.8	120.3	119.8
118.7	120.1	119.8	118.7
120.5	118.7	119.1	119.3
117.5	119.3	119.5	119.7

모평균의 추정 사례 2

- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 119.5$
- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 0.522$

  여기서, $N = kn = 100 \times 20 = 2{,}000$

- 모평균 μ 의 95% 신뢰구간 :

$$\bar{y}_{sy} \pm z_{\alpha/2}\sqrt{\hat{V}(\bar{y}_{sy})} \leftrightarrow 119.5 \pm 2\sqrt{0.0258}$$
$$\leftrightarrow 119.5 \pm 0.32$$

모총계의 추정

- 모총계의 수학적 표현 : $\tau = \sum_{i=1}^{N} y_i = N\mu$
- 모총계의 추정량 : $\hat{\tau}_{sy} = N\bar{y}_{sy}$
- $\hat{V}(\hat{\tau}_{sy}) = N^2 V(\bar{y}_{sy}) = N^2 \frac{N-n}{N} \frac{s^2}{n}$
- 모총계에 대한 $100(1-\alpha)\%$ 신뢰구간

$$\left(\hat{\tau}_{sy} - z_{\alpha/2}\sqrt{\hat{V}(\hat{\tau}_{sy})}, \hat{\tau}_{sy} + z_{\alpha/2}\sqrt{\hat{V}(\hat{\tau}_{sy})}\right)$$

모비율 추정 : 랜덤모집단 가정

- $\hat{p}_{sy} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- $\hat{V}(\hat{p}_{sy}) = \frac{N-n}{N}\frac{\hat{p}_{sy}(1-\hat{p}_{sy})}{n-1}$
- 모비율에 대한 $100(1-\alpha)\%$ 신뢰구간

$$\left(\hat{p}_{sy} - z_{\alpha/2}\sqrt{\hat{V}(\hat{p}_{sy})}\,,\ \hat{p}_{sy} + z_{\alpha/2}\sqrt{\hat{V}(\hat{p}_{sy})}\right)$$

모비율 추정 예 (1)

어느 후보자에 대한 유권자의 지지율 파악 목적

- 어떤 지역의 총 유권자수 $N = 5,775$ 명
- 유권자 명부에서 1/6 계통표본 추출($n = 962$ 명)
- 962명의 표본 중 652명이 지지한다고 응답

  ➔ 지지율의 95% 신뢰구간은?

모비율 추정 예 (2)

- 지지율의 추정값

$$\hat{p}_{sy} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{652}{962} = 0.678$$

- 유권자 명부에 나열 순서와 지지 여부와는 서로 독립적

$$\begin{aligned}\hat{V}(\hat{p}_{sy}) &= \frac{N-n}{N}\frac{\hat{p}_{sy}\hat{q}_{sy}}{n-1} \\ &= \frac{5,775-962}{5,775} \cdot \frac{(0.678)(0.322)}{961} \\ &= 0.000189\end{aligned}$$

- 지지율에 대한 95% 신뢰구간

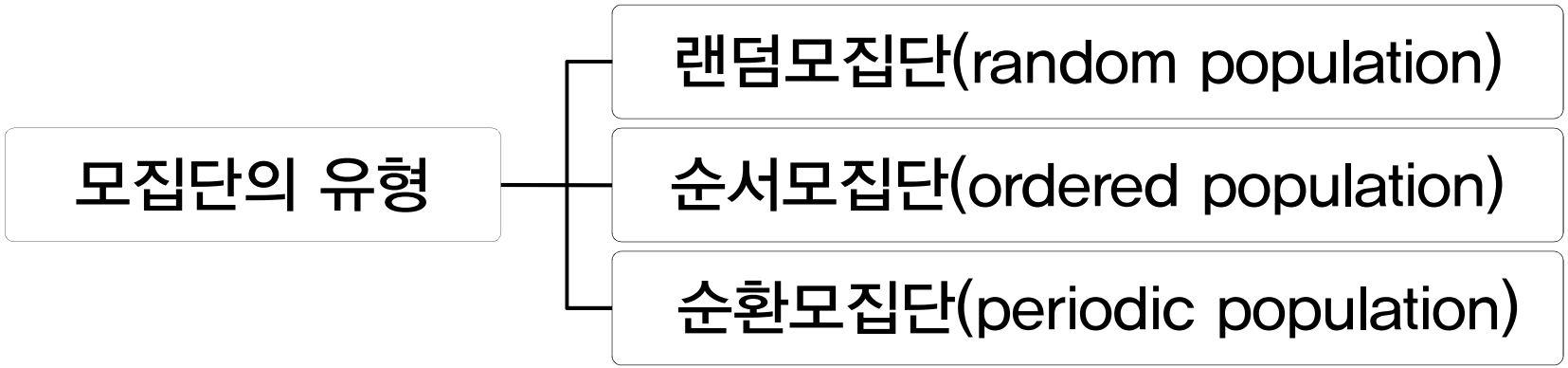
$$\begin{aligned}\hat{p}_{sy} \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{p}_{sy})} &\leftrightarrow 0.678 \pm 2\sqrt{0.000189} \\ &\leftrightarrow 0.678 \pm 0.0275\end{aligned}$$

표본의 크기 결정 : 랜덤모집단 가정

- 모평균 추정의 경우
  - $100(1-\alpha)\%$ 신뢰도에서 오차의 한계가 B 이내
  - $$n = \frac{N(z_{\alpha/2}S)^2}{NB^2 + (z_{\alpha/2}S)^2} = \frac{n_0}{1+\frac{n_0}{N}} \quad , \quad n_0 = \frac{(z_{\alpha/2}S)^2}{B^2}$$
- 모비율 추정의 경우
  - $$n = \frac{n_0}{1+\frac{n_0}{N}} \quad , \quad n_0 = \frac{z_{\alpha/2}^2\, p\,(1-p)}{B^2}$$

학/습/목/차

모집단의 유형



- 모집단 유형에 따라서 추정의 정확도가 달라짐

랜덤모집단

- 추출단위가 관심변수 값과 아무 관련 없이 랜덤하게 배열된 경우
  - 대부분의 모집단은 랜덤모집단으로 간주할 수 있음
  - 급내상관계수 ρ 값을 0으로 간주할 수 있음
    - ➔ 단순임의추출법에서와 같은 추정방법 사용
    - ➔ 단순임의추출법과 계통추출법은 동등(equivalent)하다고 말함

랜덤모집단의 예

일련번호	추출단위	특성치	일련번호	추출단위	특성치
1	U_1	27	9	U_9	18
2	U_2	3	10	U_{10}	24
3	U_3	9	11	U_{11}	2
4	U_4	10	12	U_{12}	29
5	U_5	7	13	U_{13}	12
6	U_6	21	14	U_{14}	20
7	U_7	2	15	U_{15}	28
8	U_8	22	16	U_{16}	4

순서모집단

- 추출단위가 관심변수 값의 크기 순으로 나열되어 있는 경우

예

- 소득을 추정할 때 과거 소득 자료 크기 순으로 모집단을 나열한 경우
  - 계통표본 내의 단위들이 이질적이어서 급내상관계수 ρ 값이 음수로 나타남
    - ➔ 단순임의추출법에 비해 효율적임

순서모집단의 예

일련번호	추출단위	특성치	일련번호	추출단위	특성치
1	U_7	2	9	U_9	18
2	U_{11}	2	10	U_6	21
3	U_2	3	11	U_8	22
4	U_{16}	4	12	U_{10}	24
5	U_5	7	13	U_{14}	20
6	U_3	9	14	U_1	27
7	U_4	10	15	U_{15}	28
8	U_{13}	12	16	U_{12}	29

순환모집단

- 조사단위의 배열이 관심변수 값을 기준으로 주기적으로 변동하는 경우
  - 계통표본 내 단위들이 동질적이어서 급내상관계수가 0보다 큼
    - ➔ 단순임의추출법에 비해 효율이 떨어짐
    - ➔ 계통추출법을 사용하는 것은 피해야 함
  - 계통추출법을 적용하기에 앞서 모집단에 대한 충분한 검토가 필요함

순환모집단의 예

일련번호	추출단위	특성치	일련번호	추출단위	특성치
1	U_7	2	9	U_2	3
2	U_5	7	10	U_4	10
3	U_9	18	11	U_8	22
4	U_{14}	20	12	U_{15}	28
5	U_{11}	2	13	U_{16}	4
6	U_3	9	14	U_{13}	12
7	U_6	21	15	U_{10}	24
8	U_1	27	16	U_{12}	29

반복계통추출법 (참고)

- 여러 개의 임의출발점을 택하여 여러 개의 계통표본을 뽑는 방법
  - 모집단에 대한 가정 없이 추정량의 분산 계산 가능
  - $\hat{\mu} = \bar{y}_{sy} = \frac{1}{n}\sum_{i=1}^{n} y_i$ 의 분산 추정
    - $\bar{y}_1,\ \bar{y}_2,\ \cdots,\ \bar{y}_m$: m 개의 독립적인 계통표본
    - 모평균 추정량 : $\hat{\mu} = \sum_{i=1}^{m} \frac{\bar{y}_i}{m}$
    - 추정량의 분산 : $\hat{V}(\hat{\mu}) = \left(\frac{N-n}{N}\right) \frac{\sum_{i=1}^{m} (\bar{y}_i - \hat{\mu})^2}{m(m-1)}$

계통추출법에서 추정량의 분산 추정

- 추출된 계통표본을 단순임의표본으로 간주하여 계산
  - 단순임의추출법의 분산 추정식 이용
  - 랜덤모집단인 경우에 사용됨
- 반복계통추출법 이용
  - 모집단에 대한 가정 없이 추정량의 분산 계산 가능

학/습/목/차

1. 계통추출법의 개념과 장단점
2. 모평균과 모총계에 대한 추정
3. 모집단의 유형
4. 엑셀을 활용한 실습

〈실습하기〉에서 자세히 다룸

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