8강 층화임의추출법(3)

정보통계학과 이기재교수

학/습/목/차

1. 표본배분 – 네이만배분법, 최적배분법
2. 사후층화
3. 층화임의추출법과 단순임의추출법의 비교
4. 엑셀을 활용한 실습

(1) 네이만배분법

- 각 층의 크기와 층별 변동의 정도를 동시에 고려한 표본배정 방법
- 변동이 큰 층에 대해서는 상대적으로 많은 표본을 배정
- 층별 조사비용은 별 차이가 없고, 변동의 정도가 많이 나는 경우에 적당
- 배분공식 : $n_h = n \times \dfrac{N_h S_h}{\sum_{k=1}^{H} N_k S_k}$, $h = 1, 2, \ldots, H$
- 네이만배분일 때의 표본크기 결정 공식 : $n = \dfrac{\left(\sum_{h=1}^{H} N_h S_h\right)^2}{N^2 D + \sum_{h=1}^{H} N_h S_h^2}$

(1) 네이만배분법

예제 4-8

▶ 네이만배분법 적용 예제

층	규모(종업원수)	제조업체의 수(N_h)	표본분산(s_h^2)
1	49인 이하	18,000	80^2
2	50 – 99인	4,000	200^2
3	100 – 249인	2,000	600^2
4	250인 이상	500	$1,900^2$

$$\sum_{k=1}^{4} N_k S_k = 18,000 \times 80 + 4,000 \times 200 + 2,000 \times 600 + 500 \times 1,900 = 4,390,000$$

- 층1 : $n_1 = 500 \times \dfrac{18,000 \times 80}{4,390,000} \fallingdotseq 164$ ▪ 층2 : $n_2 = 500 \times \dfrac{4,000 \times 200}{4,390,000} \fallingdotseq 91$
- 층3 : $n_3 = 500 \times \dfrac{2,000 \times 600}{4,390,000} \fallingdotseq 137$ ▪ 층4 : $n_4 = 500 \times \dfrac{500 \times 1,900}{4,390,000} \fallingdotseq 108$

(2) 최적배분법

- 주어진 비용 하에서 추정량의 분산을 최소화시키거나
  주어진 분산의 범위 하에서 비용을 최소화시키는 방법

- 층별로 단위당 조사비용에 차이가 있는 경우에 쓰이는 방법

- 배분공식 : $n_h = n \times \dfrac{N_h S_h / \sqrt{c_h}}{\sum_{k=1}^{H} N_k S_k / \sqrt{c_k}}$, $h = 1, 2, \ldots, H$

- 최적배분일 때의 표본크기 결정 공식 :

$$n = \frac{\left(\sum_{h=1}^{H} N_h S_h / \sqrt{c_h}\right)\left(\sum_{k=1}^{H} N_k S_k / \sqrt{c_k}\right)}{N^2 D + \sum_{h=1}^{H} N_h S_h^2}$$

(2) 최적배분법

예제 4-9

▶ 최적배분법 적용 예제

층	규모(종업원수)	제조업체의 수(N_h)	표본분산 (s_h^2)	조사비용(c_h)
1	49인 이하	18,000	80^2	1
2	50 – 99인	4,000	200^2	1
3	100 – 249인	2,000	600^2	2
4	250인 이상	500	$1{,}900^2$	3

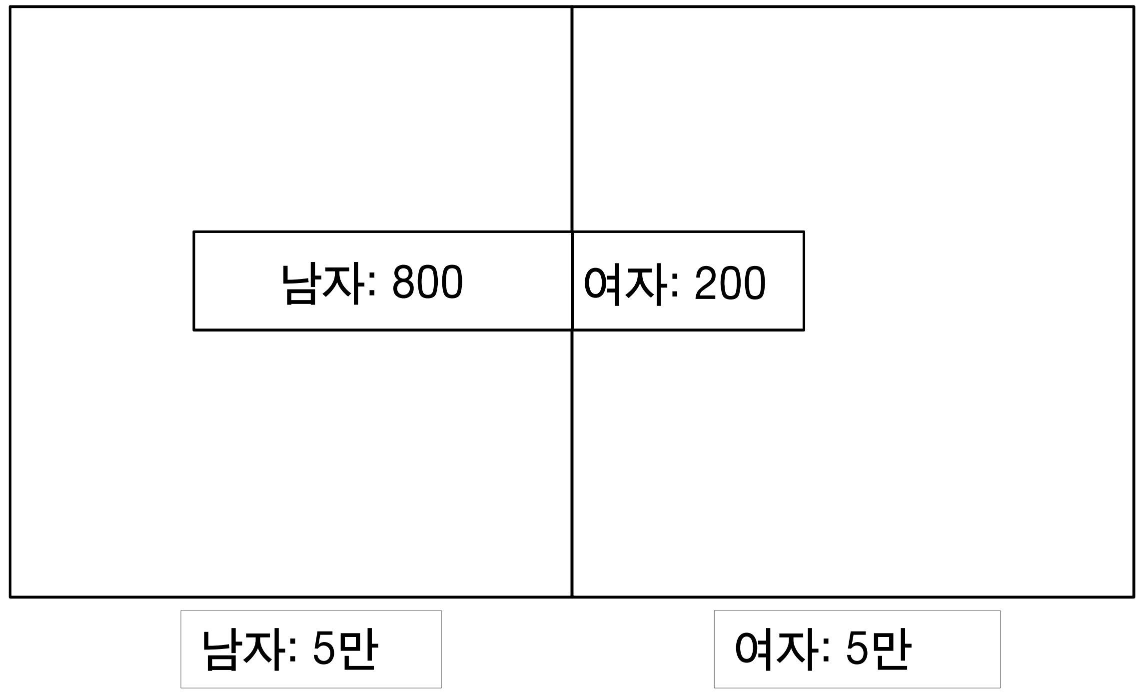
$$\sum_{k=1}^{H} N_k S_k / \sqrt{c_k} = \frac{18,000 \times 80}{\sqrt{1}} + \frac{4,000 \times 200}{\sqrt{1}} + \frac{2,000 \times 600}{\sqrt{2}} + \frac{500 \times 1,900}{\sqrt{3}} = 3,637,010$$

- 층1 : $n_1 = 500 \times \frac{1,440,000}{3,637,010} \fallingdotseq 198$
- 층2 : $n_2 = 500 \times \frac{800,000}{3,637,010} \fallingdotseq 110$
- 층3 : $n_3 = 500 \times \frac{848,528}{3,637,010} \fallingdotseq 117$
- 층4 : $n_4 = 500 \times \frac{548,482}{3,637,010} \fallingdotseq 75$

학/습/목/차

1. 표본배분 – 네이만배분법, 최적배분법
2. 사후층화
3. 층화임의추출법과 단순임의추출법의 비교
4. 엑셀을 활용한 실습

사후층화의 필요성을 나타내는 예 : 평균 몸무게 추정



- 남자 표본평균 = 55kg, 여자 표본평균 = 45kg
- 남자가 표본에 과다하게 반영 ➔ 과다 추정의 가능성

사후층화(post-stratification)의 개념

- 단순임의표본을 이용할 경우는 이미 알고 있는 모집단 특성 비율을 반영 못함
- 단순임의추출을 이용했지만 추정단계에서 모집단의 사전정보를 이용
- 층화임의추출 : 표본설계 단계에서 층화변수를 기준으로 층화
  사후층화 : 표본추출이 이루어지고 난 이후 표본의 데이터를 층화

예

▶ 도시 시민들의 평균 몸무게 추정

▪ 평균의 사후 층화 추정량 = (남자의 평균 × 0.5) +
(여자의 평균 × 0.5)
= (55 × 0.5) +(45 × 0.5) = 50kg

〈참고〉 단순 표본평균 = 53kg

모평균에 대한 사후층화 추정량

- 표본데이터
  - $y_1, y_2, \ldots, y_n$ ➔ 사후층화 : $(y_{11}, y_{12}, \ldots, y_{1n_1}), \ldots (y_{H1}, y_{H2}, \ldots, y_{Hn_H})$
- 사후층화추정량
  - $\bar{y}_{post} = \sum_{h=1}^{H} W_h \bullet \bar{y}_{ph}$, 여기서 $W_h = \frac{N_h}{N}$, $\bar{y}_{ph} = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$
- 분산추정량
  - $\hat{V}(\bar{y}_{post}) = \frac{N-n}{Nn} \bullet \sum_{h=1}^{H} W_h s_{ph}^2 + \frac{1}{n^2}\sum_{h=1}^{H}(1-W_h)s_{ph}^2$

모평균에 대한 사후층화 추정량

예제 4-10

▶ 사후층화 적용 예제

❖ 모집단 - 도매계정:40%, 소매계정 : 60%

❖ 표 본 - 도매계정:70%, 소매계정 : 30%

도매계정	소매계정
$n_1 = 70$ $\bar{y}_{p1} = 520$ $s_{p1} = 210$	$n_2 = 30$ $\bar{y}_{p2} = 280$ $s_{p2} = 90$

- 평균에 대한 사후층화 추정값 : $\bar{y}_{post} = 0.4 \times 520 + 0.6 \times 280 = 376$
- 분산추정 : $\hat{V}(\bar{y}_{post}) = \frac{1}{n} \cdot \sum_{h=1}^{2} W_h s_{ph}^2 + \frac{1}{n^2} \sum_{h=1}^{2} (1 - W_h) s_{ph}^2 = 227.97$
- 신뢰구간 : $376 \pm 2\sqrt{227.97} \leftrightarrow 376 \pm 30$

학/습/목/차

1. 표본배분 – 네이만배분법, 최적배분법
2. 사후층화
3. 층화임의추출법과 단순임의추출법의 비교
4. 엑셀을 활용한 실습

층화추출은 반드시 단순임의추출보다 더 효과적인가?

- 적절한 층화변수 선택 (층화변수와 조사변수의 상관관계가 높을 때)

예제 4-11

▶ 부적절한 층화의 예 – 체인망으로 층화

층	N_h	n_h	표본데이터							$\bar{y}_h$	s_h^2
1	30	5	92	88	100	108	96			96.8	59.2
2	42	7	89	97	91	103	109	99	98	98.0	46.3
3	36	6	106	94	98	91	91	94		95.7	32.3
4	36	6	90	108	92	89	111	92		97.0	96.0
합계	144	24									

1. 층화추출일 때의 평균 판매량 추정

- $\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{4} N_h \bar{y}_h = 96.9$
- $\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h}\frac{s_h^2}{n_h} = 2.01$

층화추출은 반드시 단순임의추출보다 더 효과적인가?

- 적절한 층화변수 선택 (층화변수와 조사변수의 상관관계가 높을 때)

예제 4-11

▶ 부적절한 층화의 예 – 체인망으로 층화

층	N_h	n_h	표본데이터							$\bar{y}_h$	s_h^2
1	30	5	92	88	100	108	96			96.8	59.2
2	42	7	89	97	91	103	109	99	98	98.0	46.3
3	36	6	106	94	98	91	91	94		95.7	32.3
4	36	6	90	108	92	89	111	92		97.0	96.0
합계	144	24									

2. 단순임의추출법일 때의 추정

- $\bar{y} = \frac{1}{24}(92 + 88 + \cdots + 89 + 111 + 92) = 96.9$
- $\hat{V}(\bar{y}) = \frac{N-n}{n} \cdot \frac{s^2}{n} = 1.77$

층화추출은 반드시 단순임의추출보다 더 효과적인가?

- 적절한 층화변수 선택 (층화변수와 조사변수의 상관관계가 높을 때)

예제 4-11

▶ 부적절한 층화의 예 – 체인망으로 층화

층	N_h	n_h	표본데이터							$\bar{y}_h$	s_h^2
1	30	5	92	88	100	108	96			96.8	59.2
2	42	7	89	97	91	103	109	99	98	98.0	46.3
3	36	6	106	94	98	91	91	94		95.7	32.3
4	36	6	90	108	92	89	111	92		97.0	96.0
합계	144	24									

3. 두 추정값의 비교

- 단순임의추출에서의 추정분산 = 1.77

〈 층화추출에서의 추정분산 = 2.01

- 체인망을 층화기준으로 삼은 것은 부적절

〈참고〉 좋은 층화 : 층내 동질적, 층간 이질적

층화추출은 반드시 단순임의추출보다 더 효과적인가?

- 적절한 층화변수 선택 (층화변수와 조사변수의 상관관계가 높을 때)

예제 4-12

▶ 적절한 층화의 예 – 각 점포의 전년 판매량 기준 층화

층	N_h	n_h	표본데이터	$\bar{y}_h$	s_h^2
1(대)	41	7	100 109 108 111 106 103 96	104.7	28.6
2(중)	43	7	94 98 91 97 91 108 98	96.7	33.9
3(소)	60	10	92 88 89 99 90 92 91 89 94 92	91.6	10.0
합계	144	24			

층화추출은 반드시 단순임의추출보다 더 효과적인가?

- 적절한 층화변수 선택 (층화변수와 조사변수의 상관관계가 높을 때)

예제 4-12

▶ 적절한 층화의 예 – 각 점포의 전년 판매량 기준 층화

1. 층화추출일 때의 평균 판매량 추정

- $$\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{3} N_h \bar{y}_h = \frac{1}{144}(41 \times 104.7 + 43 \times 96.7 + 60 \times 91.6) = 96.9$$

- $$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h}\frac{s_h^2}{n_h} = \frac{1}{144^2}[41(41-7) \times \frac{28.6}{7} + 43(43-7) \times \frac{33.9}{7} + 60(60-10) \times \frac{10}{10}] = 0.78$$

층화추출은 반드시 단순임의추출보다 더 효과적인가?

▪ 적절한 층화변수 선택 (층화변수와 조사변수의 상관관계가 높을 때)

예제 4-12

▶ 적절한 층화의 예 – 각 점포의 전년 판매량 기준 층화

2. 단순임의추출법일 때의 추정

▪ $\bar{y} = 96.9$, $\hat{V}(\bar{y}) = 1.77$

3. 두 추정값의 비교

▪ 단순임의추출에서의 추정분산 = 1.77 〉 층화추출에서의 추정분산 = 0.78

▪ 적절한 층화로 인해 효율을 2.3배(= 1.77/0.78) 높인 사례

학/습/목/차

1. 표본배분 – 네이만배분법, 최적배분법
2. 사후층화
3. 층화임의추출법과 단순임의추출법의 비교
4. 엑셀을 활용한 실습

〈실습하기〉에서 자세히 다룸

이 강의는

강의용 휴대폰(U-KNOU 서비스 휴대폰)으로도

다시 볼 수 있습니다.