

6강 총화임의추출법(1)

정보통계학과 이기재교수

학/습/목/차

1. 총화임의추출법의 개요

2. 총화

3. 모평균 추정

4. 엑셀을 활용한 실습

총화임의추출법의 개요

예

- ▶ 전국 서점에서의 월간 도서 판매량 조사
전국 서점 35,050개 중 300개의 표본을 추출

소규모 서점 (매장 30평 미만)	중규모 서점 (30평 이상 ~ 100평 미만)	대규모 서점 (100평 이상)
서점수 30,000	서점수 5,000	서점수 50

단순임의추출법의 문제점

- ▶ 대규모 서점이 표본에 전혀 뽑히지 않을 가능성 높음



해결책

- ▶ 서점을 규모별로 구분한 후 각 규모별로 표본을 추출

층화임의추출법의 정의

층화(stratification)

- 모집단을 서로 겹쳐지지 않게 몇 개의 부분군으로 나누는 일
- 층화는 층화임의추출법의 효율을 결정하는 가장 중요한 작업

층(stratum)

- 모집단을 겹치지 않게 몇 개의 부분군으로 나누었을 때 각각의 부분군

층화임의추출법(stratified random sampling)

- 모집단을 먼저 서로 겹치지 않도록 여러 개의 층으로 분할한 후, 각 층별로 단순임의추출법을 적용시켜 표본을 얻는 방법

층화임의추출방법이 널리 사용되는 이유

1	단순임의추출법에 비해 추정의 정도를 높일 수 있음 ✓ 같은 층에 속한 단위들이 동질적(homogeneous) 일수록 효과적
2	전체 추정뿐 아니라 각 층별로도 추정도 가능 ✓ 지역별, 특성별 통계작성이 가능
3	표본의 대표성 제고 및 조사관리 편리, 조사비용 절감

학/습/목/차

1. 총화임의추출법의 개요

2. 총화

3. 모평균 추정

4. 엑셀을 활용한 실습

층화변수 (1)

층화변수(stratification variable)

모집단을 몇 개의 층으로 나누려고 할 때 각 추출단위가 어느 층에 속하는지를 구분하기 위해 기준으로 사용되는 변수

층화변수의 선택 시 고려사항

- ▶ 주변수와의 관련성
- ▶ 활용 가능한 보조정보의 종류
- ▶ 통계의 작성단위 : 전국 or 시도별 or 시군별

총화변수 (2)

좋은 총화변수

사전에 모집단 단위들의 정보를 쉽게 알 수 있으면서도
조사하고자 하는 주변수와 밀접한 관련이 있는 보조변수

총화변수 선택의 예

도서판매량 조사	서점의 매장면적
여론조사	유권자의 지역, 성별, 연령, 학력
사업체조사	전년도 매출액, 종업원 수
농작물생산량 조사	해당 경지에서 재배하는 작물의 종류

총화변수 (3)

❖ 예제 4-1

▶ 총화변수 결정 사례

어느 대학교 학생들의 주당 평균 학습시간 조사

■ 총화변수

: 학습시간과 밀접한 관련을 지니면서도 쉽게 구할 수 있는 변수

✓ 성별 : 남, 여

✓ 전공 : 인문, 이공, 예체능

✓ 학년 : 저학년(1-2학년), 고학년(3-4학년)

■ 총의 수: $2 \times 3 \times 2 = 12$ 개

층의 최적경계점 결정

질적 층화변수와 양적 층화변수

- 질적 층화변수 : 변수값에 따라 층 구분
- 양적 층화변수 : 층의 경계점을 나누는 방법 필요

층의 최적경계점(optimum point of stratification)

- ▶ 층화변수가 양적 변수
- ▶ 모집단을 개의 층으로 나누려면 개의 경계점을 결정해야 함
- ▶ 주어진 여건 하에서 추정값의 분산을 최소화시킬 수 있도록
경계점 결정
- ▶ Dalenius & Hodges의 누적도수제공근법

누적도수제곱근법의 사용 예

예제 4-2

▶ 13,435개 미국 은행들의 기업대출금 비율(%) 자료

비율	도수(f)	\sqrt{f}	$\Sigma\sqrt{f}$
0-5%	3,464	58.9	58.9
5-10%	2,516	50.2	109.1
10-15%	2,157	46.4	155.5
15-20%	1,581	39.8	195.3
20-25%	1,142	33.8	229.1
25-30%	746	27.3	256.4
30-35%	512	22.6	279
35-40%	376	19.4	298.4
40-45%	265	16.3	314.7
45-50%	207	14.4	329.1
50-55%	126	11.2	340.3
55-60%	107	10.3	350.6
60-65%	82	9.1	359.7
65-70%	50	7.1	366.8
70-75%	39	6.2	373
75-80%	25	5.0	378
80-85%	16	4.0	382
85-90%	19	4.4	386.4
90-95%	2	1.4	387.8
95-100%	3	1.7	389.5

▪ 총의 개수 : $H = 6$

▪ $F = \sum \sqrt{f(y)} = 389.5$

▪ 경계점 간격

$$B = F/H = 389.5/6 = 64.9$$

학/습/목/차

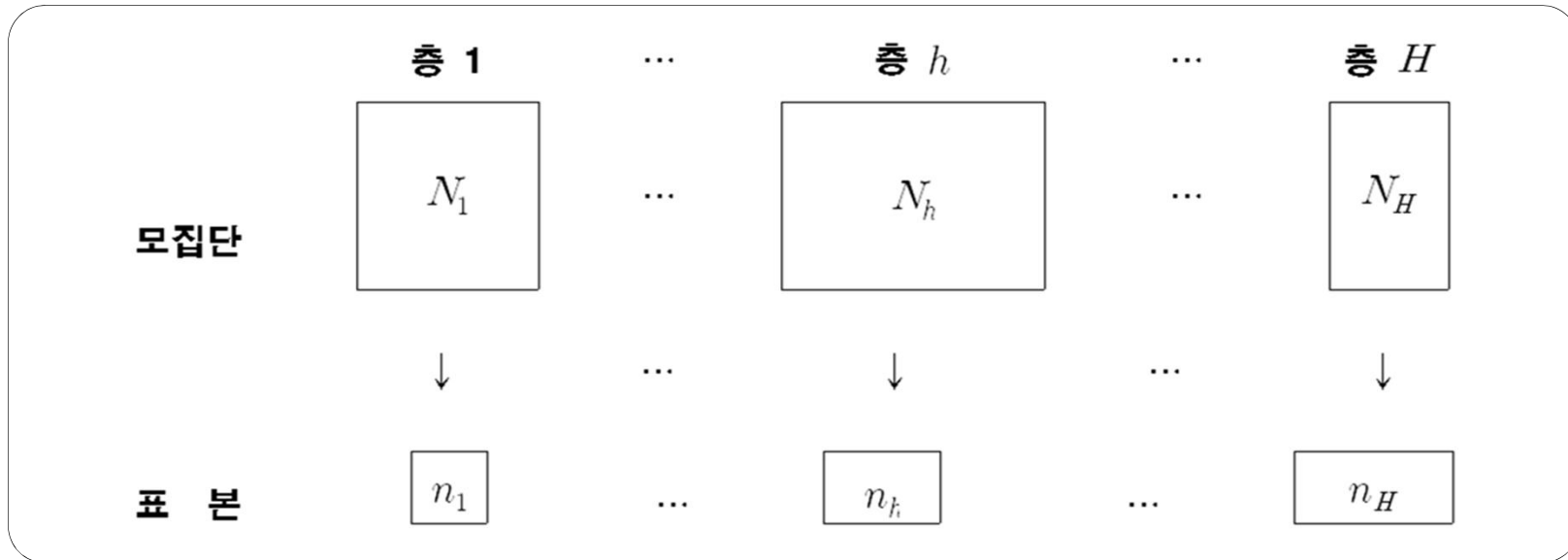
1. 총화임의추출법의 개요

2. 총화

3. 모평균 추정

4. 엑셀을 활용한 실습

기호



- ▶ H = 모집단의 전체 층의 수
- ▶ $N_h = h$ 번째 층 내 추출단위의 수, $h = 1, 2, \dots, H$
- ▶ $N = N_1 + N_2 + \dots + N_H =$ 모집단의 전체 추출단위의 수
- ▶ $n_h = h$ 번째 층에서 뽑히는 표본의 크기, $h = 1, 2, \dots, H$
- ▶ $n = n_1 + n_2 + \dots + n_H =$ 전체 표본의 크기
- ▶ $y_{hi} = h$ 번째 층에 속한 i 번째 추출단위에서의 관찰값,
 $h = 1, 2, \dots, H ; i = 1, 2, \dots, n_h$

모집단의 특성값

$$\tau_h = \sum_{i=1}^{N_h} y_{hi}$$

h 번째 층의 모총계

$$\mu_h = \tau_h / N_h$$

h 번째 층의 모평균

$$\tau = \sum_{h=1}^H \tau_h = \sum_{h=1}^H N_h \mu_h = \sum_{h=1}^H \sum_{i=1}^{N_h} y_{hi}$$

모집단의 모총계

$$\mu = \frac{\tau}{N} = \frac{1}{N} \sum_{h=1}^H N_h \mu_h = \frac{1}{N} \sum_{h=1}^H \sum_{i=1}^{N_h} y_{hi}$$

모집단의 모평균

모평균의 추정 (1)

1. h 번째 층의 추정량 ($h = 1, 2, \dots, H$)

- ▶ h 번째 층의 데이터

$$y_{h1}, y_{h2}, \dots, y_{hn_h}$$

- ▶ h 번째 층의 평균추정량

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$$

- ▶ h 번째 층의 분산추정량

$$V(\bar{y}_h) = \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h}, \quad S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \mu_h)^2$$

$$\hat{V}(\bar{y}_h) = \frac{N_h - n_h}{N_h} \frac{s_h^2}{n_h}, \quad s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$$

모평균의 추정 (2)

2. 모집단 전체에 대한 모평균(μ)에 대한 추정

▶ 모평균 추정량

$$\overline{y}_{st} = \frac{1}{N} \sum_{h=1}^H N_h \overline{y}_h$$

▶ 분산추정

$$V(\overline{y}_{st}) = \frac{1}{N^2} \{ N_1^2 V(\overline{y}_1) + N_2^2 V(\overline{y}_2) + \cdots + N_H^2 V(\overline{y}_H) \}$$

$$= \frac{1}{N^2} \sum_{h=1}^H N_h^2 V(\overline{y}_h) = \frac{1}{N^2} \sum_{h=1}^H N_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h}$$

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

▶ 신뢰구간

$$\overline{y}_{st} \pm z_{\alpha/2} \sqrt{\hat{V}(\overline{y}_{st})}$$

모평균의 추정 (3)

❖ 예제 4-3

▶ 모평균 추정 사례

도시	2,000	1,200	1,500	1,000	900	1,300	3,000	1,500	1,000	1,800
농촌	600	450	200	300	1,200	900	600	750	1,000	300

- $\bar{y}_{\text{도시}} = 1,520(\text{명})$, $\bar{y}_{\text{농촌}} = 630(\text{명})$
- $\hat{V}(\bar{y}_{\text{도시}}) = \frac{100-10}{100} \cdot \frac{630^2}{10} = 35,721$
- $\hat{V}(\bar{y}_{\text{농촌}}) = \frac{50-10}{50} \cdot \frac{331^2}{10} = 8,765$
- $\bar{y}_{st} = \frac{1}{N} \sum_{h=1}^H N_h \bar{y}_h = \frac{1}{150} (100 \cdot 1,520 + 50 \cdot 630) = 1,223.3$
- $\hat{V}(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^H N_h^2 \hat{V}(\bar{y}_h) = \frac{1}{150^2} [100^2 \cdot (35,721) + 50^2 \cdot (8,765)] = 16,850$
- $\bar{y}_{st} \pm z_{\alpha/2} \sqrt{\hat{V}(\bar{y}_{st})} \Leftrightarrow 1,223 \pm 260$

학/습/목/차

1. 총화임의추출법의 개요

2. 총화

3. 모평균 추정

4. 엑셀을 활용한 실습

↳ <실습하기>에서 자세히 다룸



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