

표본조사의 올바른 모수추정법

- ▶ 일시 : 2014년 8월 14일(목) 오후 2시
- ▶ 장소 : 한국청소년정책연구원 10층 회의실
- ▶ 주최 : 한국청소년정책연구원

세부일정

표본조사의 올바른 모수추정법

시간	내용	비고
14:00 ~ 14:10	개회 및 발표자 소개	양계민 연구위원
14:10 ~ 14:20	환영사	노 혁 원장
14:20 ~ 15:40	콜로키움 발표	이계오(한국갤럽 자문교수)
15:40 ~ 16:00	질의응답	참석자 전원
16:00	폐회	양계민 연구위원

Contents

제 I 장 표본조사의 모수추정 방법	1
1. 단순임의추출법에서 추정 방법	2
2. 층화추출법에서 추정 방법	3
3. 계통추출법에서 추정 방법	5
4. 집락추출법에서 분산 추정	6
5. 2단 집락추출법에서 분산 추정	9
6. 비추정량에서 분산 추정	13
7. 회귀모형 추정 방법	14
 제 II 장 복잡표본조사에서 비선형추정량의 분산추정	17
1. 테일러 선형화 방법	18
2. 랜덤 그룹 방법	21
3. 잭나이프 방법	22
4. 일반화 분산함수 (Generalized Variance Functions)	23
 제 III 장 복잡표본조사에서 추정방법 예제	27
1. 층화집락추출법 예제	28
2. 비추정법 예제	30
3. 부차영역(domain) 추정 예제	31
4. 다양한 분산추정법 예제	32

제 1 장 표본조사의 모수추정 방법

1. 단순임의추출법에서 추정 방법

(1) SRSWOR(Simple Random Sampling Without Replacement)(N, n) ~ 표본={ $y_1, y_2, \dots, y_n$ }

① 표본평균과 표본분산

$$\bullet \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad ; \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

② $\bar{y}$ 와 s^2 의 기대값과 분산

$$\bullet E(\bar{y}) = \bar{Y}$$

$$\bullet V(\bar{y}) = \left(\frac{N-n}{N-1} \right) \frac{\sigma^2}{n} = (1-f) \frac{S^2}{n}, \quad \text{다만, } f = \frac{n}{N}$$

$$\bullet E(s^2) = S^2$$

③ $\bar{y}$ 의 분산과 표준오차의 추정량

$$\bullet \hat{V}(\bar{y}) = (1-f) \frac{s^2}{n}$$

$$\bullet \widehat{SE}(\bar{y}) = \sqrt{1-f} \frac{s}{\sqrt{n}}$$

(2) 모비율 추정과 분산 추정

① 모비율 P 의 불편추정량

$$\bullet \text{표본비율 } p = \frac{n_1}{n}$$

② 표본비율 p 의 분산

$$\bullet V(p) = \left(\frac{N-n}{N-1} \right) \frac{PQ}{n}, \quad \text{다만 } Q = 1 - P$$

③ $V(p)$ 의 불편추정량

- $\hat{V}(p) = (1-f) \frac{pq}{n-1}, \quad \text{다만 } q = 1-p$

④ p 의 표준오차 추정량

- $\widehat{SE}(p) = \sqrt{(1-f) \frac{pq}{n-1}}$

2. 층화추출법에서 추정 방법

(1) 모평균 $\overline{Y}$ 의 불편추정량 :

- $\overline{y_{st}} = \sum_{h=1}^L \frac{N_h}{N} \overline{y_h} = \sum_{h=1}^L W_h \overline{y_h}$

(2) $\overline{y_{st}}$ 의 분산과 불편추정량

- $V(\overline{y_{st}}) = \sum_{h=1}^L W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$

- $\widehat{V}(\overline{y_{st}}) = \sum_{h=1}^L W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}$

(3) $\overline{y_{st}}$ 의 표준오차 추정량

- $\widehat{SE}(\overline{y_{st}}) = \sqrt{\sum_{h=1}^L W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}}$

(4) 모비율 P 의 불편추정량

$$\bullet p_{st} = \frac{1}{N} \sum_{h=1}^L N_h p_h = \sum_{h=1}^L W_h p_h$$

(5) 분산 $V(p_{st})$ 의 불편추정량

$$\bullet \hat{V}(p_{st}) = \sum_{h=1}^L W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{p_h q_h}{n_h - 1}$$

■ 층화추출법에서 핵심 이슈 ■

(1) 층별 표본배분법

- 등배분 : $n_h = n/L$
- 비례배분 : $n_h = n \cdot N_h/N$
- 네이만배분 : $n_h = n \cdot N_h S_h / \sum_{h=1}^L N_h S_h$
- 최적배분 : $n_h = n \cdot [N_h S_h / \sqrt{c_h}] / \sum_{h=1}^L N_h S_h / \sqrt{c_h}$

(2) 층의 경계점 계산

- $W_h \sigma_h$ 의 균등화 배분 : Dalenius & Gurney
- $N_h \overline{Y_h}$ 의 균등화 배분 : Mahalanobis
- $W_h R_h$ 의 균등화 배분 : Ekman
- $\sqrt{f_y}$ 의 누적도수 균등화 배분 : Dalenius & Hodges
- 반복계산방법 : Lavalley & Hidioglou의 stratification 함수

3. 계통추출법에서 추정 방법

(1) 모평균 $\overline{Y}$ 의 추정량

- $\overline{y}_{sy} = \frac{1}{n} \sum_{i=1}^n y_i$

(2) $\overline{y}_{sy}$ 의 분산과 추정량

- $V(\overline{y}_{sy}) = \frac{1}{k} \sum_{i=1}^k (\overline{Y}_i - \overline{Y})^2 = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S_{wsy}^2$ (선형계통추출)

- $\widehat{V}_1(\overline{y}_{sy}) = \frac{N-n}{Nn^2} \sum_{i=1}^{n/2} (y_{2i} - y_{2i-1})^2$ (제1공식)

- $\widehat{V}_2(\overline{y}_{sy}) = \frac{N-n}{2Nn(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2$ (제2공식)

- $\widehat{V}(\overline{y}_{sy}) = \frac{N-n}{Nn(n-1)} \sum_{i=1}^n (y_i - \overline{y}_{sy})^2$ (임의모집단)

(3) 모비율 P 의 추정량

- 계통표본비율 : $p_{sy} = \frac{n_1}{n}$

(4) p_{sy} 의 분산 추정량

- $\widehat{V}_1(p_{sy}) = \frac{N-n}{Nn^2} \sum_{i=1}^{n/2} (y_{2i-1} - y_{2i})^2 = \frac{N-n}{Nn^2} R_1$ (제1공식)

- $\widehat{V}_2(p_{sy}) = \frac{N-n}{2Nn(n-1)} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2 = \frac{N-n}{2Nn(n-1)} R_2$ (제2공식)

- $\widehat{V}(p_{sy}) = (1-f) \frac{p_{sy}(1-p_{sy})}{n-1}$ (임의모집단)

여기서 R_1 = “관측값의 수열을 2개씩 짝지을 때 두 관측값이 1과 0, 또는 0과 1인 횟수” 이고,
 R_2 = “관측값의 수열에서 0 다음에 1이나 1 다음에 0이 나오는 횟수”이다.

4. 집락추출법에서 분산 추정

(1) 기호 정의

- N : 모집단에서 집락의 총개수
- n : 표본으로 뽑힌 집락의 개수
- M_i : i 번째 집락에서 조사단위의 개수(i 번째 집락 크기)
- $M_0 = \sum_{i=1}^N M_i$: 모집단에서 조사단위의 총개수(모집단 크기)
- $\bar{M} = M_0 / N$: 평균 집락 크기
- Y_{ij} : i 번째 집락의 j 번째 조사단위 특성값($j = 1, \dots, M_i, i = 1, \dots, N$)
- y_{ij} : i 번째 집락의 j 번째 관측값($j = 1, \dots, M_i, i = 1, \dots, n$)
- $Y_{i\cdot} = \sum_{j=1}^{M_i} Y_{ij}$: 모집단에서 i 번째 집락의 총계(모집단 집락총계)
- $y_{i\cdot} = \sum_{j=1}^{M_i} y_{ij}$: 표본집락에서 i 번째 집락의 총계(표본 집락총계)
- $\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} Y_{ij}$: i 번째 집락의 모평균
- $\bar{y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$: i 번째 표본집락의 평균
- $\bar{Y}_c = \frac{1}{N} \sum_{i=1}^N Y_{i\cdot}$: 모집단의 평균 집락총계
- $\bar{y}_c = \frac{1}{n} \sum_{i=1}^n y_{i\cdot}$: 표본의 평균 집락총계
- $\bar{Y} = \frac{1}{M_0} \sum_{i=1}^N \sum_{j=1}^{M_i} Y_{ij}$: 모평균

(2) 집락출법에서 모평균 추정

① 집락추출에서 M_0 를 알 때 모평균의 불편추정량

$$\bullet \quad \overline{y_{c1}} = \frac{1}{nM} \sum_{i=1}^n M_i \overline{y_i} = \frac{1}{nM} \sum_{i=1}^n y_i.$$

② $\overline{y_{c1}}$ 의 분산 추정량

$$\bullet \quad \hat{V}(\overline{y_{c1}}) = \left(\frac{N-n}{Nm^2} \right) \frac{1}{n-1} \sum_{i=1}^n (y_i - \overline{M y_{c1}})^2$$

③ 집락추출에서 M_0 를 모를 때 모평균의 편의추정량

$$\bullet \quad \overline{y_{c2}} = \frac{1}{n} \sum_{i=1}^n y_i$$

④ $\overline{y_{c2}}$ 의 분산 추정량

$$\bullet \quad \hat{V}(\overline{y_{c2}}) = \left(\frac{N-n}{Nn} \right) \frac{1}{n-1} \sum_{i=1}^n (y_i - \overline{y_{c2}})^2$$

⑤ 집락추출에서 M_0 와 무관한 모평균의 비형태 추정량

$$\bullet \quad \overline{y_{c3}} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n M_i}$$

⑥ $\overline{y_{c3}}$ 의 분산 추정량

$$\bullet \quad \hat{V}(\overline{y_{c3}}) = \left(\frac{N-n}{Nm^2} \right) \frac{1}{n-1} \sum_{i=1}^n (y_i - M_i \overline{y_{c3}})^2 \text{ (단, } \overline{M} = \hat{M} = \frac{1}{n} \sum_{i=1}^n M_i \text{로 대치 가능)}$$

⑦ 모비율 P 의 추정

- M_0 를 알 때(불편추정량)

$$- p_{c1} = \frac{1}{n\overline{M}} \sum_{i=1}^n a_i$$

$$- \hat{V}(p_{c1}) = \frac{N-n}{Nn(n-1)} \sum_{i=1}^n \left(\frac{a_i}{\overline{M}} - p_{c1} \right)^2$$

- M_0 를 모를 때(편향추정량)

$$- p_{c2} = \frac{1}{n} \sum_{i=1}^n \frac{a_i}{M_i}$$

$$- \hat{V}(p_{c2}) = \frac{N-n}{Nn(n-1)} \sum_{i=1}^n \left(\frac{a_i}{M_i} - p_{c2} \right)^2$$

- M_0 와 무관한 모비율의 비형태 추정량

$$- p_{c3} = \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n M_i}$$

$$- \hat{V}(p_{c3}) = \frac{N-n}{Nn\overline{M}^2(n-1)} \sum_{i=1}^n (a_i - p_{c3}M_i)^2$$

(다만, $\overline{M}$ 는 $\widehat{\overline{M}} = \frac{1}{n} \sum_{i=1}^n M_i$ 로 대체 가능)

5. 2단 집락추출법에서 분산 추정

(1) 기호 정의

- N : 모집단에서 1차 추출단위(psu)의 총개수
- n : 표본으로 뽑힌 psu(집락)의 개수
- M_i : i 번째 psu(집락)에서 있는 2차추출단위(ssu)의 총개수(i 번째 집락 크기)
- m_i : i 번째 표본 psu(집락)에서 표본크기
- $M_0 = \sum_{i=1}^N M_i$: 모집단에서 조사단위(ssu)의 총개수(모집단 크기)
- $\bar{M} = M_0/N$: psu당 평균 ssu 개수(평균 집락크기)
- Y_{ij} : i 번째 psu에서 j 번째 ssu의 특성값 ($j = 1, \dots, M_i, \quad i = 1, \dots, N$)
- y_{ij} : i 번째 표본 psu에서 j 번째 ssu의 관측값 ($j = 1, \dots, m_i, \quad i = 1, \dots, n$)
- $Y_{i.} = \sum_{j=1}^{M_i} Y_{ij}$: i 번째 psu의 모집단 총계
- $y_{i.} = \sum_{j=1}^{m_i} y_{ij}$: i 번째 표본 psu의 관측값 총계
- $Y_{..} = \sum_{i=1}^N Y_{i.}$: 모집단 특성값 총계
- $y_{..} = \sum_{i=1}^n y_{i.}$: 표본 관측값 총계
- $\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} Y_{ij}$: i 번째 psu의 모평균 (집락 모평균)
- $\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$: i 번째 표본 psu의 표본평균 (집락 표본평균)
- $\bar{Y} = \frac{1}{M_0} \sum_{i=1}^N \sum_{j=1}^{M_i} Y_{ij}$: (ssu당) 모평균

(2) 2단 집락출법에서 모평균 추정

① 이단추출에서 M_0 를 알 때 모평균의 불편추정량

$$\bullet \quad \overline{y_{m1}} = \frac{N}{nM_0} \sum_{i=1}^n M_i \overline{y_i}$$

② $\overline{y_{m1}}$ 의 분산 추정량

$$\bullet \quad \hat{V}(\overline{y_{m1}}) = \left(\frac{1}{n} - \frac{1}{N} \right) s_{1b}^2 + \frac{1}{nNM^2} \sum_{i=1}^n M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2$$

$$\bullet \quad s_{1b}^2 = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{M_i \overline{y_i}}{\overline{M}} - \overline{y_{m1}} \right)^2 = \frac{1}{M^2(n-1)} \left[\sum_{i=1}^n (M_i \overline{y_i})^2 - \frac{1}{n} \left(\sum_{i=1}^n M_i \overline{y_i} \right)^2 \right]$$

$$\bullet \quad s_i^2 = \frac{1}{m_i-1} \sum_{j=1}^{m_i} (y_{ij} - \overline{y_i})^2 = \frac{1}{m_i-1} \left(\sum_{j=1}^{m_i} y_{ij}^2 - m_i \overline{y_i}^2 \right)$$

③ 이단추출에서 M_0 를 모를 때 모평균의 편의추정량

$$\bullet \quad \overline{y_{m2}} = \frac{1}{n} \sum_{i=1}^n \overline{y_i}$$

• $\overline{y_{m2}}$ 의 분산 추정량 :

$$- \quad \hat{V}(\overline{y_{m2}}) = \left(\frac{1}{n} - \frac{1}{N} \right) s_{2b}^2 + \frac{1}{nN} \sum_{i=1}^n \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2$$

④ 이단추출에서 M_0 와 무관한 모평균의 비형태 추정량 :

$$\bullet \quad \overline{y_{m3}} = \frac{\sum_{i=1}^n M_i \overline{y_i}}{\sum_{i=1}^n M_i}$$

- $\overline{y_{m3}}$ 의 분산 추정량

$$- \hat{V}(\overline{y_{m3}}) = \left(\frac{1}{n} - \frac{1}{N} \right) s_{3b}^2 + \frac{1}{nNM^2} \sum_{i=1}^n M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2$$

(다만, $\overline{M}$ 가 미지일 때는 $\widehat{\overline{M}} = \frac{1}{n} \sum_{i=1}^n M_i$ 를 사용한다.)

- $s_{2b}^2 = \frac{1}{n-1} \sum_{i=1}^n \left(\overline{y_i} - \overline{y_{m2}} \right)^2 = \frac{1}{(n-1)} \left(\sum_{i=1}^n \overline{y_i^2} - \frac{\left(\sum_{i=1}^n \overline{y_i} \right)^2}{n} \right)$

- $s_{3b}^2 = \frac{1}{(n-1)M^2} \sum_{i=1}^n M_i^2 (\overline{y_i} - \overline{y_{m3}})^2$

⑤ 모총계의 추정

- M_0 와 무관한 추정량

$$- \widehat{Y_{m1}} = M_0 \overline{y_{m1}} = \frac{N}{n} \sum_{i=1}^n M_i \overline{y_i}$$

- M_0 를 사용하는 추정량

$$- \widehat{Y_{m2}} = M_0 \overline{y_{m2}} = \frac{M_0}{n} \sum_{i=1}^n \overline{y_i}$$

$$- \widehat{Y_{m3}} = M_0 \overline{y_{m3}} = \frac{M_0 \sum_{i=1}^n M_i \overline{y_i}}{\sum_{i=1}^n M_i}$$

⑥ 모비율 P 의 추정

- M_0 를 알 때(불편추정량)

$$- p_{m1} = \frac{N}{nM_0} \sum_{i=1}^n M_i p_i$$

$$- \hat{V}(p_{m1}) = \left(\frac{1}{n} - \frac{1}{N} \right) s_{1b}^2 + \frac{1}{M^2 n N} \sum_{i=1}^n \frac{M_i (M_i - m_i) p_i q_i}{m_i - 1}$$

$$- s_{1b}^2 = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{M_i p_i}{\overline{M}} - p_{m1} \right)^2 = \frac{1}{\overline{M}^2 (n-1)} \left[\sum_{i=1}^n (M_i p_i)^2 - \frac{1}{n} \left(\sum_{i=1}^n M_i p_i \right)^2 \right]$$

● M_0 를 모를 때(편향 추정량)

$$- p_{m2} = \frac{1}{n} \sum_{i=1}^n p_i$$

$$- \hat{V}(p_{m2}) = \left(\frac{1}{n} - \frac{1}{N} \right) s_{2b}^2 + \frac{1}{nN} \sum_{i=1}^n \left(\frac{M_i - m_i}{M_i} \right) \frac{p_i q_i}{m_i - 1}$$

$$- s_{2b}^2 = \frac{1}{n-1} \sum_{i=1}^n (p_i - p_{m2})^2 = \frac{1}{(n-1)} \left(\sum_{i=1}^n p_i^2 - n p_{m2}^2 \right)$$

● M_0 와 무관한 모비율의 비형태 추정량

$$- p_{m3} = \frac{\sum_{i=1}^n M_i p_i}{\sum_{i=1}^n M_i}$$

$$- \hat{V}(p_{m3}) = \left(\frac{1}{n} - \frac{1}{N} \right) s_{3b}^2 + \frac{1}{nN\overline{M}^2} \sum_{i=1}^n \left(\frac{M_i^2 (M_i - m_i)}{M_i} \right) \frac{p_i q_i}{m_i - 1}$$

(다만, $\overline{M}$ 가 미지일 때는 $\widehat{\overline{M}} = \frac{1}{n} \sum_{i=1}^n M_i$ 를 사용한다.)

$$- s_{3b}^2 = \frac{1}{(n-1)\overline{M}^2} \sum_{i=1}^n M_i^2 (p_i - p_{m3})^2$$

6. 비추정량에서 분산 추정

(1) 모비(ratio) R 의 추정량

- $\hat{R} = \bar{y} / \bar{x}$

(2) $\hat{R}$ 의 근사 편향

- $B(\hat{R}) = \left(\frac{N-n}{Nn} \right) \frac{1}{\bar{X}^2} (R S_x^2 - S_{xy})$

(3) $\hat{R}$ 의 근사 평균제곱오차

- $MSE(\hat{R}) = \left(\frac{N-n}{Nn} \right) \frac{1}{\bar{X}^2} (S_y^2 + R^2 S_x^2 - 2R S_{xy})$

(4) 평균제곱오차의 추정량

- $\widehat{MSE}(\hat{R}) = \left(\frac{N-n}{Nn} \right) \frac{1}{\bar{X}^2} (s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{xy})$

($\bar{X}$ 가 미지일 때는 $\bar{x}$ 를 사용한다.)

■ 예제 1 ■

$N=1620$ 가구에서 $n=30$ 가구를 추출해서 가구당 월수입(x)에서 우유값(y)의 비(ratio)를 추정하고

95% 신뢰구간을 구하자. 모평균 수입은 $\bar{X}=2250$ 과 기타 통계량들은 아래와 같다.

- $\bar{x} = 2606.5, \quad \bar{y} = 73.77$

- $s_x^2 = \frac{1}{n-1} (\sum x_i^2 - n \bar{x}^2) = 489759.15$

- $s_y^2 = \frac{1}{n-1} (\sum y_i^2 - n \bar{y}^2) = 417.88$

- $s_{xy} = \frac{1}{n-1} (\sum x_i y_i - n \bar{x} \bar{y}) = 12305.51$

- 비율의 추정값: $\hat{R} = \frac{\bar{y}}{x} = 0.02830$

평균제곱오차의 추정값 및 95% 신뢰구간은 다음과 같다.

- $$\begin{aligned}\widehat{\text{MSE}}(\hat{R}) &= \left(\frac{N-n}{Nn} \right) \frac{1}{X^2} (s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}) \\ &= \frac{1620-30}{(1620)(30)} \frac{1}{(2550)^2} [417.88 + (0.02830)^2(489759.15) \\ &\quad - 2(0.02830)(12305.51)] \\ &= 5.71713 \times 10^{-7}\end{aligned}$$

95% 신뢰구간은 다음과 같이 구할 수 있다.

- $$\begin{aligned}\hat{R} \pm 1.96 \sqrt{\widehat{\text{MSE}}(\hat{R})} &= 0.02830 \pm 1.96 \sqrt{5.71713 \times 10^{-7}} \\ &= 0.02830 \pm 0.00148 = (0.02682, 0.02978)\end{aligned}$$

7. 회귀모형 추정 방법

(1) 선형회귀추정량

- $\bar{y}_L = \bar{y} + \beta(\bar{X} - \bar{x})$

(2) β 의 값을 아는 경우 ($\beta = \beta_0$ 일 때)

- $\bar{y}_L = \bar{y} + \beta_0(\bar{X} - \bar{x})$

- $$\begin{aligned}V(\bar{y}_L) &= \left(\frac{N-n}{Nn} \right) [S_y^2 + \beta_0^2 S_x^2 - 2\beta_0 S_{xy}] \\ &= \left(\frac{N-n}{Nn} \right) S_y^2 (1 - \rho^2), \quad \beta_0 = \beta \text{ 일 때}\end{aligned}$$

- $$\hat{V}(\bar{y}_L) = \left(\frac{N-n}{Nn} \right) [s_y^2 + \beta_0^2 s_x^2 - 2\beta_0 s_{xy}]$$

(3) β 의 값을 모르는 경우

- $\overline{y_L} = \overline{y} + \hat{\beta}(\overline{X} - \overline{x})$, 여기서 $\hat{\beta} = \frac{s_{xy}}{s_x^2}$
- $B(\overline{y_L}) = -C(\overline{x}, \hat{\beta})$
- $\text{MSE}(\overline{y_L}) \approx \left(\frac{N-n}{Nn} \right) [S_y^2 + \beta^2 S_x^2 - 2\beta S_{xy}]$
 $= \left(\frac{N-n}{Nn} \right) S_y^2 (1 - \rho^2)$
- $\widehat{\text{MSE}}(\overline{y_L}) = \left(\frac{N-n}{Nn} \right) [s_y^2 + \hat{\beta}^2 s_x^2 - 2\hat{\beta} s_{xy}]$
 $= \left(\frac{N-n}{Nn} \right) s_y^2 (1 - r^2)$

■ 예제 2 ■

$N=961$ 명 교수 중 24명을 추출해서 랜덤하게 나이(x)와 혈압(y)을 측정하여 평균혈압을 추정하고, 95% 신뢰구간을 계산해보자. (단, 교수평균 나이 42.7세, $\beta = 0.8952$)

계산을 위해서 주어진 통계량은 아래와 같다.

- $\overline{x} = 44.13$, $\overline{y} = 7135.29$
- $s_x^2 = \frac{1}{n-1} (\sum x_i^2 - n\overline{x}^2) = 89.13$
- $s_y^2 = \frac{1}{n-1} (\sum y_i^2 - n\overline{y}^2) = 146.08$
- $s_{xy} = \frac{1}{n-1} (\sum x_i y_i - n\overline{x}\overline{y}) = 82.85$

평균 혈압의 회귀추정값은

$$\begin{aligned}\bullet \quad \overline{y}_L &= \bar{y} + \beta_0(\bar{X} - \bar{x}) \\ &= 135.29 + 0.8952(42.7 - 44.13) = 134.01\end{aligned}$$

$\overline{y}_L$ 의 분산 추정값은

$$\begin{aligned}\bullet \quad \hat{V}(\overline{y}_L) &= \left(\frac{N-n}{Nn} \right) [s_y^2 + \beta_0^2 s_x^2 - 2\beta_0 s_{xy}] \\ &= \frac{961-24}{(961)(24)} [146.08 + (0.8952)^2(89.13) - 2(0.8952)(82.85)] \\ &= 2.8102\end{aligned}$$

평균혈압의 95% 신뢰구간은

$$\begin{aligned}\bullet \quad \overline{y}_L \pm 1.96 \sqrt{\hat{V}(\overline{y}_L)} &= 134.01 \pm 1.96 \sqrt{2.8102} \\ &= (130.72, 137.30)\end{aligned}$$

보조변수를 사용하지 않으면 $\bar{y}$ 의 분산

$$\bullet \quad \hat{V}(\bar{y}) = \frac{N-n}{Nn} s_y^2 = \frac{961-24}{(961)(24)} (146.08) = 5.9347$$

$\overline{y}_L$ 의 $\bar{y}$ 에 대한 상대효율의 추정값

$$\bullet \quad \text{RE}(\overline{y}_L, \bar{y}) = \frac{5.9347}{2.8102} = 2.1118$$

Multicultural Adolescents Panel Survey

제 2 장 복잡표본조사에서 비선형추정량의 분산추정

1. 테일러 선형화 방법

- t_1 은 강도범죄에 의한 피해금액, t_q 는 강도범죄에 의한 손실금액을 나타낸다면 t_1/t_q 는 비선형통계량이므로 $a_1t_1 + a_2t_q$ 와 같은 선형결합으로 나타낼 수 없음 $\Rightarrow$ 분산추정을 선형함수로 표현할 수 없음
- 총계의 비선형함수 $h(t_1, t_2, \dots, t_k)$ 의 테일러정리를 이용하여 상수들 $a_0, a_1, \dots, a_k$ 에 대해서 비선형함수를 선형화할 수 있음

$$h(t_1, t_2, \dots, t_k) \cong a_0 + \sum_{i=1}^k a_i t_i \quad \text{-----}(1)$$

- 결론적으로 $V[h(t_1, t_2, \dots, t_k)]$ 는 근사적으로 $V(\sum_{i=1}^k a_i t_i)$ 로 계산 가능

■ 예제 1 ■

- 모비율 p 의 함수인 $\theta = p(1-p)$ 는 $\hat{\theta} = \hat{p}(1-\hat{p})$ 로 추정 가능하고, p 의 불편추정치는 $\hat{p}$, 분산 $V(\hat{p})$ 도 알고 있다면, 함수 $h(x) = x(1-x)$ 에 대해서 $\theta = h(p)$ 이고 $\hat{\theta} = h(\hat{p})$
- 함수 $h(\cdot)$ 는 x 의 비선형함수이므로 모수값인 p 에서 근사식을 아래와 같이 표현

$$\begin{aligned} h(\hat{p}) &\simeq h(p) + h'(p)(\hat{p}-p) \\ &= p(1-p) + (1-2p)(\hat{p}-p) \quad \text{-----}(2) \end{aligned}$$

- 분산의 근사값 식은 다음과 같음

$$V[h(\hat{p})] \simeq (1-2p)^2 V(\hat{p}-p) \quad \text{-----}(3)$$

- 여기서 $V(\hat{p})$ 는 알고 있으므로 $h(\hat{p})$ 분산의 근사값은 계산 가능

● 평균들의 비선형함수의 분산의 선형화추정량을 계산 과정은 다음과 같음

(1) 관심 함수를 표본평균들의 함수 형식으로 나타내면, 일반적으로 $\theta = h(\overline{y_{1U}}, \dots, \overline{y_{kU}})$ 이며, 예제

의 $\theta = h(\overline{y_U}) = h(p) = p(1-p)$ 와 같다.

(2) 각 인수별로 부분미분계수들을 계산하고 모수에서 계산된 부분미분계수들을 선형화 상수항 a_i 으로 이용한다.

(3) 테일러정리를 적용하여 선형화 추정치를 계산한다.

$$h(\overline{y_1}, \overline{y_2}, \dots, \overline{y_k}) \simeq h(\mu_1, \mu_2, \dots, \mu_k) + \sum_{i=1}^k a_i (\overline{y_i} - \mu_i)$$

여기서 a_i 는 i 번째 인수에 대한 부분미분계수의 값이다.

(4) 새로운 변수와 아래와 같이 정의하면 분산의 근사값은 쉽게 계산할 수 있다.

$$q_i = \sum_{j=1}^k a_j y_{ij} \text{ 와 같이 새로운 변수를 정의하면 추정치는 선형화형태인 } \hat{t}_q = \sum_{i \in S} w_i q_i \text{ 로 계산할 수 있고,}$$

분산의 근사값도 계산할 수 있다.

■ 예제 2 ■

비추정에서 분산의 근사값을 선형화방법으로 계산하면, 단순임의추출법에서 추정량은 $\hat{B} = \overline{y}/\overline{x} = \hat{t}_y/\hat{t}_x$ 이 되고 근사값은 아래와 같다.

$$\hat{B} - B = \frac{\overline{y} - B\overline{x}}{\overline{x}} \simeq \frac{\overline{y} - B\overline{x}}{\overline{x_U}} = \sum_{i \in s} \frac{y_i - Bx_i}{n x_U}$$

분산의 근사값은 아래와 같다.

$$V[\hat{B} - B] = \frac{1}{n^2 x_U^2 V[\sum_{i \in S} (y_i - Bx_i)]} \quad \text{-----(4)}$$

위의 근사값을 계산하기 위해서 테일러정리를 적용하는데 다음과 같은 절차로 계산한다.

(1) 비(B)를 모총계의 함수로 표현하면 $h(c,d) = d/c$ 와 같고 아래 식으로 나타낸다.

$$B = h(t_x, t_y) = \frac{t_y}{t_x}, \quad \hat{B} = h(\hat{t}_x, \hat{t}_y) = \frac{\hat{t}_y}{\hat{t}_x}$$

여기서 표본추정치 $\hat{t}_x$ 와 $\hat{t}_y$ 는 각각 비편향이라 가정한다.

(2) 부분 미분계수는 아래와 같다.

$$\frac{\delta h(c,d)}{\delta c} = \frac{-d}{c^2} \quad \text{과} \quad \frac{\delta h(c,d)}{\delta d} = \frac{1}{c}$$

위의 미분계수를 $c = t_x$ 와 $d = t_y$ 에서 계산하면 $-t_y/t_x^2$ 와 $1/t_x$ 이 된다.

(3) 테일러 정리를 이용하면 근사값은 아래와 같이 쓸 수 있다.

$$\begin{aligned} \hat{B} &= h(\hat{t}_x, \hat{t}_y) \\ &\simeq h(t_x, t_y) + \frac{\delta h(c,d)}{\delta c} \Big|_{t_x, t_y} (\hat{t}_x - t_x) + \frac{\delta h(c,d)}{\delta d} \Big|_{t_x, t_y} (\hat{t}_y - t_y) \end{aligned}$$

단계 (2)에서 주어진 부분미분계수를 적용하면 아래 식을 얻게 된다.

$$\hat{B} - B \simeq \frac{t_y}{t_x^2} (\hat{t}_x - t_x) + \frac{1}{t_x} (\hat{t}_y - t_y)$$

(4) $\hat{B}$ 의 평균제곱오차의 근사값은 아래 식으로 계산할 수 있다.

$$\begin{aligned} E[(\hat{B} - B)^2] &\simeq E \left[\left(\frac{-t_y}{t_x^2} (\hat{t}_x - t_x) + \frac{1}{t_x} (\hat{t}_y - t_y) \right)^2 \right] \\ &= \frac{t_y^2}{t_x^4} V(\hat{t}_x) + \frac{1}{t_x^2} V(\hat{t}_y) - 2 \frac{t_y}{t_x^3} COV(\hat{t}_x, \hat{t}_y) \\ &= \frac{1}{t_x^2} [B^2 V(\hat{t}_x) + V(\hat{t}_y) - 2BCOV(\hat{t}_x, \hat{t}_y)] \quad \text{-----(5)} \end{aligned}$$

SRSWOR(N, n)을 가정하면 $\hat{B}$ 의 평균제곱오차의 근사값은 $V(\hat{t}_x) = N^2(1-n/N)S_x^2/n$,

$V(\hat{t}_y) = N^2(1-n/N)S_y^2/n$ 와 $COV(\hat{t}_x, \hat{t}_y) = N^2(1-n/N)RS_xS_y/n$ 대입하면 아래와 같다.

$$E[(\hat{B}-B)^2] = \frac{N^2(N-n)}{Nnt_x^2} [S_y^2 + B^2 S_x^2 - 2BR S_x S_y]$$

평균제곱오차의 근사값의 추정치는 아래 식으로 계산 가능함.

$$\hat{E}[(\hat{B}-B)^2] = \frac{(N-n)}{Nnt_x^2} [s_y^2 + \hat{B}^2 s_x^2 - 2\hat{B}\hat{R} s_x s_y]$$

여기서 $\hat{B} = \bar{y}/\bar{x}$ 이고 $\hat{R} = s_{xy}/s_x s_y$ 이다.

2. 랜덤 그룹 방법

집락표본에서 1차 추출단위(PSU)를 R개 그룹으로 랜덤하게 구분하면 PSU에 속하는 관찰치들은 동일한 그룹에 속하게 되므로 각 그룹들은 역시 집락표본이 된다. 층화다단계표본에서는 랜덤그룹은 각 층으로부터 PSU들의 표본으로 구성되므로 만약에 가장 적은 층이 k개의 PSU들로 구성되었다면 기껏해야 k개의 랜덤그룹이 가능함에 유의해야 함.

만약 θ 가 비선형 통계량이라면 $\tilde{\theta}$ 는 본 표본으로부터 직접 계산된 추정량 $\hat{\theta}$ 와는 다를 수 있다. 비추정의 예제에서 $\tilde{\theta} = (1/R) \sum_{r=1}^R (\bar{y}_r / \bar{x}_r)$ 는 본표본에서 계산한 $\hat{\theta} = \hat{y}/\hat{x}$ 와는 다를 수 있다. $V(\hat{\theta})$ 를 추정에 사용 가능한 추정식은 아래와 같다.

$$\hat{V}_2(\hat{\theta}) = \frac{1}{R} \frac{\sum_{r=1}^R (\theta_r - \hat{\theta})^2}{R-1} \quad \text{-----(6)}$$

랜덤그룹방법을 이용한 분산 추정절차는 다음과 같다.

- (1) 주어진 표본설계에 의해서 추출된 본표본에서 첫 번째 그룹의 표본을 동일한 표본설계로 선정한다. 이 그룹을 $S_{(1)}$ 이라면 여기에서 모수 θ 에 대한 비편향 추정량 $\widehat{\theta}_{(1)}$ 을 계산한다.
- (2) 절차(1)과 동일한 추출방법으로 본 표본으로부터 두 번째 그룹의 표본 $S_{(2)}$ 을 독립적인 방법으로 추출한다. 이 부차표본($S_{(2)}$)에서부터 비편향추정량 $\widehat{\theta}_{(2)}$ 을 계산한다.
- (3) 위와 같은 절차를 반복적으로 수행하여 R개의 표본 그룹을 선정하고 각각의 표본그룹으로부터 모수의 추정량들 $\widehat{\theta}_{(1)}, \dots, \widehat{\theta}_{(R)}$ 들을 계산한다.
- (4) 관심모수(θ)에 대한 최종추정량을 아래 식으로 계산한다.

$$\tilde{\theta} = \frac{1}{R} \sum_{r=1}^R \widehat{\theta}_{(r)} \quad \text{-----}(7)$$

위 추정량의 분산은 아래와 같이 계산한다.

$$\widehat{V}_2(\hat{\theta}) = \frac{1}{R} \frac{1}{R-1} \sum_{r=1}^R (\theta_{(r)} - \tilde{\theta})^2 \quad \text{-----}(8)$$

3. 잭나이프 방법

크기 n 인 단순임의추출 표본에서 j 번째 단위를 제외하고 $\hat{\theta}$ 와 동일 형식 추정량을 $\widehat{\theta}_{(j)}$ 이라 하면 모평균 추정량인 표본평균($\hat{\theta} = \bar{y}$)에서 j 번째 단위를 제외한 잭나이프추정량은 $\widehat{\theta}_{(j)} = \sum_{i \neq j} y_i / (n-1)$ 이 된다. 단순임의추출에서 1개 단위를 제외한 잭나이프 분산추정량은 아래와 같이 정의한다.

$$\widehat{V}_{JK}(\hat{\theta}) = \frac{n-1}{n} \sum_{j=1}^n (\widehat{\theta}_{(j)} - \hat{\theta})^2 \quad \text{-----}(9)$$

여기서 상수 $(n-1)/n$ 가 왜 필요한지 $\hat{\theta} = \bar{y}$ 에서 알아보자.

$\hat{\theta} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ 에서 모든 n 개 복제에서 j 번째 단위 제외된 모평균 추정량.

$$\overline{y_{(j)}} = \frac{1}{n-1} \sum_{i \neq j} y_i = \frac{1}{n-1} (\sum_{i=1}^n y_i - y_j) = \bar{y} - \frac{1}{n-1} (y_j - \bar{y}) \quad \text{-----}(10)$$

분산추정식 (9)에서 평균추정치들의 평균에 대한 편차들 제곱 합은 아래와 같다.

$$\sum_{j=1}^n (\overline{y_{(j)}} - \bar{y})^2 = \frac{1}{(n-1)^2} \sum_{j=1}^n (y_j - \bar{y})^2 = \frac{1}{n-1} s_y^2$$

SRSWR에서 표본평균의 분산추정량은 s_y^2/n 이므로 잭나이프 분산추정량과 같다.

$$\widehat{V}_{JK}(\bar{y}) = s_y^2/n$$

만약, 원표본 추정량($\hat{\theta}$)과 잭나이프추정치들의 평균($\tilde{\theta} = \frac{1}{n} \sum_{j=1}^n \hat{\theta}_{(j)}$)이 같지 않으면 잭나이프 분산추정

량에서 $\hat{\theta}$ 대신에 $\tilde{\theta}$ 사용 가능.

$$\widehat{V}_{JK_1}(\hat{\theta}) = \frac{n-1}{n} \sum_{j=1}^n (\hat{\theta}_{(j)} - \tilde{\theta})^2 \quad \text{-----}(11)$$

4. 일반화 분산함수 (Generalized Variance Functions)

정부기관에서 발표하는 통계치들은 수천 페이지에 표로 작성되며 또는 세부영역에 대해서 추정치를 산출하여 표본오차의 크기가 필요한 경우가 많다. 이때 만일에 분산추정치를 표로 제시하지 않고 분산추정치를 산출할 수 있는 간단한 수식을 일반화 분산함수라고 한다.

미국 1990년 범죄피해조사에서 특정범죄유형별 피해자 또는 피해가구의 합계의 추정치를 $\hat{t}$ 라고 한다면 분산 추정식은

$$\hat{V}(\hat{t}) = a\hat{t}^2 + b\hat{t} \quad \text{-----}(12)$$

이 때, 모비율 추정치가 $\hat{p}$ 에 대한 분산 추정식은 아래와 같이 주어진다.

$$\hat{V}(\hat{p}) = \left(\frac{b}{x}\right)\hat{p}(1-\hat{p}) \quad \text{-----}(13)$$

여기서 $\hat{x}$ 는 비율추정에서 기준이 되는 추정모집단의 크기임

1990년 범죄피해조사 자료에서 계산한 a 와 b 가 각각 $a=0.00001833$ 과 $b=3725$ 이며, 20세~24세 연령대에 서 강도피해 비율($\hat{p}$)이 1.23%로 추정되었고, 해당 연령대 인구는 18,017,100명이라면 비율추정치($\hat{p}$)의

GVF에 의한 표준오차는 아래와 같다.

$$\sqrt{\frac{3725}{18017100}(0.0123)(1-0.0123)} = 0.0016$$

위의 근사 표준오차를 이용한 95% 근사 신뢰구간은 $0.0123 \pm (1.96)(0.0016)$ 이 된다.

만약에 t_i 를 특정계층에 속한 사람들의 수라면, 해당 계층에 속할 비율은 해당 특정계층에 속한 사람들의 비율로 $p_i = t_i/N$ 로 표현 가능하다.

비율(p_i)를 추정하는 조사에서 설계효과(def)를 d_i 라면 분산은

$$V(\hat{p}_i) \approx d_i \frac{p_i(1-p_i)}{n} = \frac{b_i}{N} p_i(1-p_i)$$

여기서 $b_i = d_i \cdot (N/n)$ 이므로 아래와 같이 쓸 수 있다.

$$V(\hat{t}_i) \approx d_i \cdot N^2 \frac{p_i(1-p_i)}{n} = a_i t_i^2 + b_i t_i$$

여기서 $a_i = -d_i/n$ 를 나타낸다.

다양한 추정치들에 대해서 설계효과들이 유사하다고 가정하여 $a_i \approx a$ 와 $b_i \approx b$ 이라면 상수인 a 와 b 는 식(10)와 식(11)을 이용하여 많은 통계량의 분산의 근사값 계산이 가능하다

일반화 분산함수를 계산하는 일반적인 과정을 설명하면 다음과 같다.

(1) 복제치와 다른 방법을 이용하여 k 개의 관심 모집단 합계들의 추정치($\hat{t}_1, \hat{t}_2, \dots, \hat{t}_k$)의 분산 추정치를 계산

하고 $\hat{t}_i$ 의 상대분산을 v_i 라 하면, $v_i = \widehat{V}(\hat{t}_i)/\hat{t}_i^2 = \widehat{CV}(\hat{t}_i)^2$, $i = 1, 2, \dots, k$ 와 같이 계산한다.

(2) v_i 와 $\hat{t}_i$ 에 관한 모형을 가정한다. 많은 종류의 조사에서 아래 모형을 사용한다.

$$v_i = \alpha + \frac{\beta}{\hat{t}_i} \quad \text{-----}(14)$$

위의 선형회귀모형에서 반응변수는 v_i 이고 설명변수는 $1/\hat{t}_i$ 이다.

- (3) 식(14)의 회귀계수 α 와 β 는 회귀분석방법으로 계산한다. 가중최소제곱 추정법으로 모수추정을 권장하고 있으며 큰 가중치를 갖는 조사항목은 작은 v_i 를 갖게 된다. 분산의 GVF추정치는 식(14)의 회귀식에서 예측값, $\hat{\alpha} + \hat{\beta}/\hat{t}$ 이다.

■ 예제 ■

2006년 미국 경상인구조사에서 실업자수를 추정하는데 수천종의 통계생산단위에 대한 추정치의 표준오차를 계산하는데 아래의 일반화 분산함수를 이용

$$V(\hat{X}) = -(\text{def})\left(\frac{N}{n}\right)\left(\frac{\hat{X}^2}{N}\right) + (\text{def})\frac{N}{n}\hat{X}$$

여기서 $a = -b/N$ 과 $b = (\text{def})N/n$ 이라 놓으면 위식은 아래와 같이 된다.

$$V(\hat{X}) = a\hat{X}^2 + b\hat{X} \quad \text{-----}(15)$$

〈 표 〉 경상인구조사에서 실업자수 추정치의 GVF 분산추정 회귀계수

실업자 구분	a	b
실업자 총계 또는 백인	-0.000016350	3095.55
흑인	0.000151396	3454.72
히스패닉인종	-0.000141225	3454.72

2012년 6월 실업자수 추정치가 9,100,000명인 경우에 2012년 6월 실업자수 추정치의 표준오차를 계산하고, 실업자수의 95%신뢰구간을 계산하면

$$se(\hat{X}) = \sqrt{(-0.000016350)(9100000)^2 + (3095.55)(9100000)} = 163754.6$$

표준오차가 163,754.6이므로 표본오차는 $(1.96)(163754.6)=320959$ 이고, 2012년 6월 전체실업자수에 대한 95% 신뢰구간은 $9100000 \pm (1.96)(163754.6)$ 이다.

제 3 장 복잡표본조사에서 추정방법 예제

1. 층화집락추출법 예제

(1) SAS data set 구축

- 7학년 ~ 9학년의 학생이 4,000명인 고등학교에서 주간 아이스크림 소비금액을 조사하기 위해서 617개 동아리를 추출틀로 사용하였고, 40명을 선정하여 주간 아이스크림 소비액이 10불 이상이면 과다 소비자로, 10불 이하이면 과소 소비자로 구분한 SAS data set이다.

```
data IceCreamStudy;
    input Grade StudyGroup Spending @@;
    if (Spending < 10) then Group='less';
    else Group='more';
    datalines ;

7 34 7 7 34 7 7 412 4 9 27 14
7 34 2 9 230 15 9 27 15 7 501 2
9 230 8 9 230 7 7 501 3 8 59 20
7 403 4 7 403 11 8 59 13 8 59 17
8 143 12 8 143 16 8 59 18 9 235 9
8 143 10 9 312 8 9 235 6 9 235 11
9 312 10 7 321 6 8 156 19 8 156 14
7 321 3 7 321 12 7 489 2 7 489 9
7 78 1 7 78 10 7 489 2 7 156 1
7 78 6 7 412 6 7 156 2 9 301 8

run;
```

- (2) 학년을 층으로 구분하고 동아리를 추출단위로 선정한 데이터분석에서 유한모집단수정계수 계산을 위한 SAS data set 입력 코드이다.

```
data StudyGroups;
    input Grade _total_;
    datalines;

7 608
8 252
9 403

run;
```

- (3) 층화추출에 대한 가중치를 계산하는 절차이다. 일반적으로 가중치는 별도로 산출하여 표본조사 데이터 세트에 포함되어 있으므로 별도의 산출 과정이 필요하지 않다.

```
data IceCreamStudy;
    set IceCreamStudy;

    if Grade=7 then Prob=8/608;
    if Grade=8 then Prob=3/252;
    if Grade=9 then Prob=5/403;

    Weight=1/Prob;

run;
```

- (4) 아이스크림 데이터를 층화집락추출법으로 분석하기 위한 SAS CODE이다.

```
title1 'Analysis of Ice Cream Spending'

proc surveymeans data=IceCreamStudy total=StudyGroups;
    strata Grade / list ;
    cluster StudyGroup;
    var Spending Group;
    weight Weight;

run;
```

- (5) 각 학년별로 동아리를 PSU로 추출하고, 학생을 선정하였는데, 여기서는 학년별로 동아리를 선정하지 않고, 직접 40명의 학생들을 선정하여 조사했다고 가정한 자료 분석 SAS CODE이다.
만약 급내 상관계수가 0에 가깝다면 cluster(동아리 집락)를 무시하고 분석해도 유사한 결과 나타난다.

```
data StudentTotal;
    input Grade _total_;
    datalines;
7 1824
8 1025
9 1151
;
run;

proc surveymeans data=IceCreamStudy total=StudentTotal;
    strata Grade / list
    var Spending Group;

run;
```

2. 비추정법 예제

(1) SAS data set 구축

- 경쟁력 강화를 위한 시장형태별로 경제적 영향 정도에 관한 정보를 조사하고자 한다. 전체적으로 800개 상위 기업들 중에서 66개 기업을 종사자수 기준 확률비례추출법으로 추출하였음

```
data company;
    length Type $14;
    input Type$ Asset sale value profit employee Weight;
    datalines;
Other    2764    1828    1850    144    18,7    9,6
energy   13246,2 4633,5 4387,7 462,9 24,3    42,6
Finance  3597,7  377,8   93     14     1,1    12,2
Hitech   6646,1  6414,2 2377,5 348,2 47,1    21,8
Transpo  1068,4  1689,8 1430,2 72,9   4,6     4,3
Manufa   1125    1719,4 1057,5 98,1   20,4    4,5
;
run;
```

- (2) 종사자수 1,000명 당 매출액(Sale)과 이익(Profit)을 추정하고자 한다. SAS data set에서 600개 기업의 자료를 입력한 후 분석하는 SAS CODE는 다음과 같다.

```
Proc surveymeans data=Company total=800 ratio ;
    Var Profit Sale Employee;
    Weight Weight;
    Ratio Profit Sale /Employee;
Run;
```

분산추정 방법을 테일러 선형화 이외의 방법을 적용하고자 할 경우에는 VARMETHOD = JACKKNIFE [또는 VARMETHOD = BRR]를 proc surveymeans 문장에 삽입하면 된다.

3. 부차영역(domain) 추정 예제

- (1) SAS 데이터는 비추정의 예제에서 이용한 상위 100개 기업에서 66개 기업을 PPS로 추출한 데이터 세트를 적용함
- (2) market type이 Hitech, Manufacturing, Medical, Other, Retail, Transportation 등 6종의 형태이지만 표본설계 또는 표본추출과정에서 반영하지 못하여 조사한 후에 마켓 유형별로 조사항목(asset, sale, value, profit와 employee)들 별로 통계치를 추정하고자 한다.
- (3) 조사항목의 측정단위는 아래와 같다.
 - Asset : 기업자본금을 100만달러 단위로 측정
 - Sale : 기업매출액을 100만달러 단위로 측정
 - Value : 기업 부가가치를 100만달러 단위로 측정
 - Employee : 기업종자수를 천 명단위로 조사함
 - Weight : 표본가중치를 나타냄
- (4) market type 별로 자본금, 매출액(Sale), 부가가치, 이익(Profit)과 종사자수 등의 평균 추정하고자 한다. 부차영역별로 분석하는 sas code는 다음과 같다

```
Proc surveymeans data=Company total=800 mean sum ;  
    Var asset sale value profit employee;  
    Weight Weight;  
    Domain type;  
Run;
```

분산추정 방법을 테일러 선형화 이외의 방법을 적용하고자 할 경우에는 VARMETHOD = JACKKNIFE 를 proc surveymeans 문장에 삽입하면 된다.

4. 다양한 분산추정법 예제

(1) 예제 1

- 7학년~9학년의 학생이 4000명인 고등학교에서 40명을 SRSWOR로 추출하여 가구소득과 어린이수가 주간 아이스크림소비금액에 얼마나 영향을 주는가를 분석하고자 한다.

```
Data icecream ;  
    input grade spending income kids @@;  
    datalines;  
7 7 39 2 7 7 38 1 8 12 47 1 9 10 47 4 7 1 34 4 7 10 43 2 7 3 44 4 8  
20 60 3 8 19 57 4 7 2  
35 2 7 2 36 1 9 15 51 1 8 16 53 1 7 6 37 4 7 6 41 2 7 6 39 2 9 15  
50 4 8 17 57 3 8 14 46  
2 9 8 41 2 9 8 41 1 9 7 47 3 7 3 39 3 7 12 50 2 7 4 43 4 9 14 46  
3 8 18 58 4 9 9 44 3 7  
2 37 1 7 1 37 2 7 4 44 2 7 11 42 2 9 8 41 2 8 10 42 2 8 13 46 1 7  
2 40 3 9 6 45 1 9 11 45 4 7 2 36 1 7 9 46 1  
;  
Run;
```

- 가구소득과 어린이 수가 아이스크림 소비금액에 어떤 영향을 주는가에 대한 분석은 회귀모형 분석을 적용하는데 여기에 사용할 SAS code이다.

```
title1 'Ice Cream Spending Analysis';  
title2 'Simple Random Sample Design';  
  
proc surveyreg data=IceCream total=4000 ;  
    class Kids;  
    model Spending = Income Kids / solution;  
run;
```

분산추정 방법을 테일러 선형화방법 이외의 방법을 적용하고자 할 경우에는 VARMETHOD = JACKKNIFE 를 proc surveyreg 문장에 삽입하면 된다.

(2) 예제 2

- 미국 샌프란시스코의 도시철도에서 승객들의 대기시간을 노선별, 기차별로 특성을 조사하였다. 조사할 노선은 지하철 'J-Church,' 'K-Ingleside,' 'L-Taraval,' 'M-Ocean View,' 'N-Judah,' 과 노상기차 'F-Market &Wharves.' 등이며, 각 노선별로 2개 기관차를 랜덤하게 선정한 후 모든 승객을 조사하였다.
- 층화집락추출법으로 분산추정(BRR)하기 위한 SAS CODE이다.

```
title 'MUNI Passenger Waiting Time Analysis Using BRR';

proc surveymeans data=MUNIsurvey mean varmethod=brr mean clm ;
    strata line;
    cluster vehicle;
    var waittime;
    weight weight;
run;
```

표본조사의 올바른 모수추정법

인 쇄 2014년 8월 11일

발 행 2014년 8월 13일

발행처 **한국청소년정책연구원**

서울특별시 서초구 태봉로 114

발행인 노 혁

등 록 1993. 10. 23 제 21-500호

인쇄처 (주)아르빛 전화 02)503-3223 대표 유채란

사전 승인없이 보고서 내용의 무단전재·복제를 금함.
구독문의 : (02) 2188-8844(정보자료·전산보안팀)