표본조사론

2021-15115 권이태

September 12, 2024

Contents

1 표본조사와 추정 **3**
- 1.1 표본조사 . . . 3
- 1.2 오차의 분해 . . . 3
- 1.3 Horvitz-Thompson 추정량 . . . 3

2 비복원 단순확률추출 **6**
- 2.1 비복원 단순확률추출 . . . 6
- 2.2 표본수의 결정 . . . 6

3 층화추출 **8**
- 3.1 층화추출 . . . 8
- 3.2 층화추출의 표본배정 . . . 8
  - 3.2.1 비례배정 . . . 8
  - 3.2.2 최적배정 . . . 9
- 3.3 사후층화 . . . 10
- 3.4 층화를 위한 이중표집 . . . 10
- 3.5 부차모집단에서의 비교 . . . 11

4 일단계 집락추출 **12**
- 4.1 집락 크기가 같은 경우의 일단계 집락추출 . . . 12
- 4.2 표본설계효과 . . . 13
- 4.3 집락 크기가 다른 경우의 일단계 집락추출 . . . 14
- 4.4 평균 추정 . . . 15
- 4.5 총계 추정과 표본 크기의 선택 . . . 15

5 이단계 집락추출 **16**

6 층화집락추출 **18**

7 비추정과 회귀추정 **19**
- 7.1 비추정 . . . 19
  - 7.1.1 비추정과 SRS의 비교 . . . 20
  - 7.1.2 층화표집에서의 비추정량 . . . 21
- 7.2 회귀추정 . . . 22
  - 7.2.1 차이추정 . . . 22

8 계통추출 **23**
- 8.1 계통추출 . . . 23
- 8.2 반복계통추출 . . . 24

8.3 모형을 이용한 추정 24

9 모집단 크기의 추정 **25**
9.1 포획-재포획 추정 25
9.2 사각표집 26
9.3 임의화 반응 26

10 비균등 확률추출 **28**
10.1 PPS 추출 28
10.1.1 PPS 집락추출 28

11 분산추정법 **29**
11.1 선형화 분산 추정법 29
11.2 복제치 분산 추정법 29

12 무응답과 무응답대체법 **31**
12.1 단위무응답의 처리 31
12.2 항목무응답의 처리 32

Chapter 1

표본조사와 추정

1.1 표본조사

- 전수조사(census): 모든 조사 대상을 조사
  - 정확한 값을 계산할 수 있음
  - 조사비용이 많이 듦
- 표본조사: 일부를 표본으로 뽑아 조사
  - 조사비용이 적고 빠르게 진행할 수 있음
  - 훈련되지 않은 면접원 등으로 인한 비표본오차의 발생 위험이 적음
  - 표본오차의 발생

1.2 오차의 분해

- 비관측오차
  - 포함오차: 표본추출특(frame)이 목표모집단(population)과 같지 않아 생기는 오차
  - **표본오차**: 표본추출틀(frame)에서 일부 표본(sample)만을 관측하기 때문에 발생하는 오차
  - 무응답오차: 추출된 표본(sample)이 응답자(respondents) 집단과 다름에 따라 생성되는 오차
- 관측오차
  - 측정오차: 참값의 부정확한 측정에서 오는 오차; 면접원, 응답자, 도구, 면접 모드의 불완전성
  - 처리오차: 처리 과정에서 발생하는 오차

1.3 Horvitz-Thompson 추정량

Definition 1. *모집단 혹은 조사명부를 $U = \{1, 2, \cdots, N\}$라 할 때, 표본으로 얻을 수 있는 집합은 $\mathcal{A} = 2^U$ 이다.* ***표본분포****는 $\mathcal{A}$에서 정의된 확률밀도함수로,*

(a) $0 \le P(A) \le 1, \ \forall A \in \mathcal{A}$

(b) $\sum_{A \in \mathcal{A}} P(A) = 1$

을 만족한다.

Definition 2. ***일차 표본포함확률*** π_i*: i번째 대상이 표본에 포함될 확률*

$$\pi_i = P(i \in A) = \sum_{i \in A} P(A)$$

이차 표본포함확률*(결합표본포함확률)* π_{ij}*: i번째 대상과 j번째 대상이 모두 표본에 포함될 확률*

$$\pi_{ij} = P(i, j \in A) = \sum_{i,j \in A} P(A)$$

Definition 3.
확률표본추출*: 모든 $i \in U$에 대하여, $\pi_i > 0$인 표본추출 방법*
가측표본추출*: 모든 $i, j \in U$에 대하여, $\pi_{ij} > 0$인 표본추출 방법*

Definition 4. *확률표본추출로 추출된 표본에서 모집단 총계 $Y = \sum_{i=1}^{N} y_i$의* ***Horvitz-Thompson****(HT) 추정량은 아래와 같이 정의된다.*

$$\hat{Y}_{HT} = \sum_{i \in A} \frac{y_i}{\pi_i}$$

Theorem 1. *HT 추정량은 아래의 성질을 만족한다.*

$$\mathbb{E}[\hat{Y}_{HT}] = Y$$
$$\mathbb{V}(\hat{Y}_{HT}) = \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

Proof. 표본포함지시변수 $I_i = I(i \in A)$를 정의하면, HT 추정량은

$$\hat{Y}_{HT} = \sum_{i=1}^{N} \frac{y_i}{\pi_i} I_i$$

으로 주어진다. 한편 $\mathbb{E}[I_i] = \mathbb{E}[I(i \in A)] = \pi_i$이므로,

$$\mathbb{E}[\hat{Y}_{HT}] = \mathbb{E}\left[\sum_{i=1}^{N} \frac{y_i}{\pi_i} I_i\right] = \sum_{i=1}^{N} \frac{y_i}{\pi_i} \mathbb{E}[I_i] = \sum_{i=1}^{N} y_i = Y$$

를 얻는다. 마찬가지로, $\mathrm{Cov}(I_i, I_j) = \pi_{ij} - \pi_i \pi_j$임을 이용하면

$$\mathbb{V}(\hat{Y}_{HT}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \mathrm{Cov}(I_i, I_j) = \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

을 얻는다. □

Theorem 2. ***(SYG 분산 공식)*** *만약 표본수가 일정한 고정표본수 추출의 경우,*

$$\mathbb{V}(\hat{Y}_{HT}) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

와 같이 표현할 수 있다.

Proof.

$$\begin{aligned}\mathbb{V}(\hat{Y}_{HT}) &= \sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\left(\frac{y_i}{\pi_i}\right)^2 - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 \\ &= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\end{aligned}$$

이다. 이때 마지막 등식은 고정표본수 추출에서

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\left(\frac{y_i}{\pi_i}\right)^2 &= \sum_{i=1}^{N}\left(\frac{y_i}{\pi_i}\right)^2\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j) \\ &= \sum_{i=1}^{N}\left(\frac{y_i}{\pi_i}\right)^2\left(\sum_{j=1}^{N}\pi_{ij}-\pi_i\sum_{j=1}^{N}\pi_j\right) \\ &= \sum_{i=1}^{N}\left(\frac{y_i}{\pi_i}\right)^2(n\pi_i-n\pi_i)=0\end{aligned}$$

임에 따라 성립한다. □

Note 1. *한편 위로부터 $\pi_i \propto y_i$인 표본설계가 분산을 최소화함을 알 수 있다.*

Theorem 3. *가측표본추출 하에서 분산 $\mathbb{V}(\hat{Y}_{HT})$에 대한 비편향추정량은*

$$\hat{\mathbb{V}}(\hat{Y}_{HT}) = \sum_{i\in A}\sum_{j\in A}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

이며, 고정표본수 추출의 경우 SYG 공식을 이용하여

$$\hat{\mathbb{V}}(\hat{Y}_{HT}) = -\frac{1}{2}\sum_{i\in A}\sum_{j\in A}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$$

으로 추정할 수 있다.

Note 2. *일반적으로 HT 추정량은 적절한 조건에서 일치성과 근사적 정규성을 만족한다.*

Note 3. *모평균을 추정하고자 하는 경우, $\hat{Y}_{HT}$를 N으로 나눈 값을 추정량으로 사용할 수 있다.*

Note 4. *HT 추정량은 location-scale invariant하지 않다. 즉 $z_i = a + by_i$를 정의하는 경우,*

$$\hat{Z}_{HT} \neq aN + b\hat{Y}_{HT}$$

이 성립하지 않는다.

Note 5. *모집단 총계 $Y = \sum_{i=1}^{N} y_i$를*

$$\hat{Y} = \sum_{i\in A} w_i y_i + C$$

처럼 추정하는 경우, $C = 0$이라면 $\hat{Y}$가 불편추정량일 필요충분조건은 $w_i = 1/\pi_i$인 것이다.

Chapter 2

비복원 단순확률추출

2.1 비복원 단순확률추출

Definition 5. *크기가 N인 유한모집단으로부터 n개의 표본을 뽑을 때, 총 $\binom{N}{n}$개의 경우의 수가 존재한다. 이들에 동등한 확률을 주어*

$$P(A) = \begin{cases} \binom{N}{n}^{-1} & \text{if } |A| = n \\ 0 & o.w. \end{cases}$$

를 표본분포로 하는 추출법을 비복원 단순확률추출(SRS)라 한다.

Theorem 4. *비복원 단순확률추출의 일차 표본포함확률과 이차 표본포함확률은*

$$\pi_i = \frac{n}{N}$$

$$\pi_{ij} = \begin{cases} \dfrac{n}{N} & if\ i = j \\ \dfrac{n(n-1)}{N(N-1)} & if\ i \neq j \end{cases}$$

으로 주어진다.

Theorem 5. *비복원 단순확률추출의 HT 추정량은*

$$\hat{Y}_{SRS} = \frac{N}{n} \sum_{i \in A} y_i = N\bar{y}$$

이다.

Theorem 6. *비복원 단순확률추출에서 분산의 추정값은*

$$\hat{\mathbb{V}}(\hat{Y}_{SRS}) = N^2 \times \frac{1}{n}\left(1 - \frac{n}{N}\right) s^2$$

으로 주어진다.

2.2 표본수의 결정

Definition 6. *유의수준 α에서* ***오차한계*** *d는*

$$P(|\hat{\theta} - \theta| \leq d) = 1 - \alpha$$

를 만족하게 하는 값을 의미한다.

Theorem 7. *비복원 단순확률추출 하에서 모평균의 오차한계 d는 근사적으로*

$$d = z_{\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{\sigma^2}{n}}$$

으로 주어진다.

Theorem 8. *정해진 유의수준 α와 오차한계 d에 대하여, 필요한 표본의 크기는*

$$n = \frac{\sigma^2}{(d/z_{\alpha/2})^2 + \sigma^2/N} = \frac{\frac{1}{d^2}z_{\alpha/2}^2\sigma^2}{1+\frac{1}{N}\frac{1}{d^2}z_{\alpha/2}^2\sigma^2} \leq \frac{1}{d^2}z_{\alpha/2}^2\sigma^2$$

이다. 즉 복원추출을 할 때의 표본크기 $\frac{1}{d^2}z_{\alpha/2}^2\sigma^2$보다 더 적은 표본이 필요하다.

Note 6. *표본크기를 결정하기 위해서는 σ^2를 알아야 하는데, 일반적으로는 이를 모른다. 표본분산을 이용하거나, 다양한 비모수적 방법들을 사용할 수 있다. 모비율 추정의 경우 $\sigma^2 = p(1-p) \leq 0.25$이기에, 이를 이용할 수도 있다.*

Chapter 3

층화추출

3.1 층화추출

Definition 7. *모집단 U를 서로 겹치지 않고 모집단을 전부 포함하는 H개의 부차 모집단, 혹은* ***층*** $U_1, U_2, \cdots, U_H$*로 나눌 때, 각 층화된 모집단 U_h 내에서 층별 표본 A_h를 독립적으로 추출하는 방법을* ***층화추출****이라고 한다. 비복원 단순확률추출을 하는 경우 특정 부분에 표본이 몰릴 위험이 크지만, 층화추출을 하는 경우 이러한 위험에서 벗어날 수 있다.*

Definition 8. *층 h 내에서의 i번째 원소의 관심변수 y의 값을 y_{hi}라고 하면, 층 h 내에서의 y의 총계는 $Y_h = \sum_{i=1}^{N_h} y_{hi}$로 표현되고 $Y = \sum_{h=1}^{H} Y_h$이다. 이때*

$$\hat{Y}_{st} = \sum_{h=1}^{H} \hat{Y}_{HT,h}$$
$$\mathbb{V}(\hat{Y}_{st}) = \sum_{h=1}^{H} \mathbb{V}(\hat{Y}_{HT,h})$$

이다. 즉 각 층 내에서 추론을 수행한 뒤 이들을 단순히 합하여 추정량과 그들의 분산을 구할 수 있다.

Theorem 9. *만약 층 내에서 비복원 단순확률추출을 한다면,*

$$\hat{Y}_{st} = \sum_{h=1}^{N} \frac{N_h}{n_h} \sum_{i \in A_h} y_{hi} = \sum_{h=1}^{H} N_h \bar{y}_h$$

이다. 그 분산추정량은

$$\hat{\mathbb{V}}(\hat{Y}_{st}) = \sum_{h=1}^{H} \mathbb{V}(\hat{Y}_{HT,h}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}$$

이다. 모평균에 대해 논의하고자 하는 경우, 추정량을 $N = N_1 + N_2 + \cdots + N_H$으로 나누면 된다.

3.2 층화추출의 표본배정

전체 표본 크기 n이 주어진 경우, 각 층에 표본수 n_h를 배정할지에 대해 논의한다.

3.2.1 비례배정

Definition 9. *만약 $n_h = N_h \frac{n}{N}$이라면,* ***비례배정****이라 한다.*

Theorem 10. *만약 비례배정을 통해 추정량을 얻은 경우, 분산추정량은*

$$\hat{\mathbb{V}}(\hat{Y}_{st}) = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\sum_{h=1}^{H}\frac{N_h}{N}s_h^2$$

이다.

Theorem 11. *각 층내에서 N_h가 충분히 크다면,*

$$\begin{aligned}\mathbb{V}(\hat{Y}_{st}) &= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^2 \\ &= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{N_h-1}\sum_{i=1}^{N_h}(y_{hi}-\bar{Y}_h)^2 \\ &\approx \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-\bar{Y}_h)^2 \\ &= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{SSW}{N-1} \le \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{SST}{N-1} = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\sigma^2\end{aligned}$$

이 성립한다. 이때 SSW는 층내 분산을 의미한다. 따라서 비례배정의 경우 층화추출이 거의 대부분 단순임의추출보다 더 효율적이다.

Note 7.

$$\underbrace{\sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-\bar{Y})^2}_{SST} = \underbrace{\sum_{h=1}^{H}N_h(\bar{Y}_h-\bar{Y})^2}_{SSB} + \underbrace{\sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-\bar{Y}_h)^2}_{SSW}$$

Note 8. *이로부터 층화추출은 층내원소들은 동질적이고, 층간은 이질적이야 층화추출의 효율이 좋아짐을 알 수 있다. 단, 비례배정이 아닌 경우 이는 달라질 수 있다.*

3.2.2 최적배정

Definition 10. ***최적배정****은 주어진 비용 하에서 HT 추정량의 분산을 최소화하는 배정을 의미한다. 이는 아래 최적화 문제의 해가 되는 n_h를 결정하는 것과 같다.*

$$\begin{cases}\text{minimize} & \mathbb{V}(\hat{Y}_{st}) = \sum_{h=1}^{H}N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{\sigma_h^2}{n_h} \\ \text{subject to} & c_0+\sum_{h=1}^{H}c_hn_h \le C \\ & \sum_{h=1}^{H}n_h = n\end{cases}$$

Theorem 12. *최적배정의 결과, 최적표본수 n_h^*는 $n_h^* \propto N_h\sigma_h/\sqrt{c_h}$으로 표현된다.*

Proof. $(n_1,\cdots,n_H)$에 대한 라그랑주 승수법을 이용한다. 라그랑지안 $\mathcal{L}$을 라그랑지안 승수 λ에 대하여

$$\mathcal{L} = \sum_{h=1}^{H}N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{\sigma_h^2}{n_h} - (C-c_0-\sum_{h=1}^{H}c_hn_h)$$

로 정의하면,

$$\frac{\partial\mathcal{L}}{\partial n_h} = -\frac{N_h^2\sigma_h^2}{n_h^2}+c_h = 0$$

에서 최적배정을 찾을 수 있다. 따라서

$$n_h \propto \frac{N_h\sigma_h}{\sqrt{c_h}}$$

□

Note 9. *만약 모든 c_h가 상수 c라면, 이를 **네이만 배정**이라 한다.*

Note 10. *만약 모든 c_h가 상수이고 층내 분산 σ_h 역시 상수라면, 이는 비례배정과 동일하다.*

Note 11. *만약 두 층 간의 모평균 차이를 확인하고자 하는 경우, 최적배정은 $n_h^* \propto \sigma_h/\sqrt{c_h}$이다. 만약 σ_h와 c_h가 상수라면, 동등배정과 동일하다.*

Note 12. *전체 모평균의 추정에는 N_h에 비례하게 n_h를 설정해야 하고, 차이 추정에는 이것이 포함되지 않는다. 이에 대한 절충으로 $0 < \alpha < 1$에 대하여 N_h^α에 비례하는 배정을 수행할 수 있다. 이를 멱배정이라 한다.*

Note 13. *층화추출에서, 오차한계를 d로 하는 최소의 n은*

$$n = \frac{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \sigma_h^2/w_h}{(d/z_{\alpha/2})^2 + \sum_{h=1}^{H} \frac{N_h}{N}\sigma_h^2/N}$$

를 만족하게 결정한다. 이때 $w_h = n_h/n$이다.
만약 최적배정을 하는 경우,

$$n = \frac{(\sum_{h=1}^{H} N_h\sigma_h/\sqrt{c_h})(\sum_{h=1}^{H} N_h\sigma_h\sqrt{c_h})}{N^2(d/z_{\alpha/2})^2 + \sum_{h=1}^{H} N_h\sigma_h^2}$$

이다.

3.3 사후층화

Definition 11. *층의 모비율 N_h/N을 안다고 가정할 때, 먼저 비복원 단순확률추출을 통해 층화변수 Z_i와 관심변수 y_i를 얻었다고 가정하자. 이때 층화변수 Z_i를 기준으로 H개의 층으로 사후적으로 층화하여 추론하는 방법을 **사후층화**라고 한다.*

Theorem 13. *사후층화를 하는 경우, 총계의 추정량은*

$$\hat{Y}_{\text{post}} = \sum_{h=1}^{H} N_h\bar{y}_h$$

이다. 이때 $\bar{y}_h$를 계산할 때의 표본수 n_h가 확률변수라는 점이 일반적인 층화추출과 다르다.

Theorem 14. *사후층화로 얻은 추정량의 분산추정량은*

$$\hat{\mathbb{V}}(\hat{Y}_{\text{post}}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sum_{h=1}^{H}\frac{N_h}{N}s_h^2 + \frac{N^2}{n^2}\sum_{h=1}^{H}\left(1 - \frac{N_h}{N}\right)s_h^2$$

으로 주어진다. 만약 n이 크다면, 이는 비례배정에서와 거의 같아짐을 확인해볼 수 있다.

3.4 층화를 위한 이중표집

Definition 12. *만약 N_h/N을 모르는 경우, 두 phase로 나누어 층화추출을 할 수 있다.*
***Phase 1**: 큰 규모의 크기 n'인 sample을 표집하고, $a'_h = n'_h/n'$을 계산하여 N_h/N의 추정량으로 이용.*
***Phase 2**: 작은 규모의 크기 n인 sample을, n'_h명 중 n_h명을 표집함으로써 만듦.*

Theorem 15. *모평균에 대한 추정량은*

$$\bar{y}_{tp} = \sum_{h=1}^{H} a'_h \bar{y}_h$$

으로 주어진다. 만약 N을 안다면, 여기에 N을 곱하여 총계에 대한 추정량 역시 얻을 수 있다.

Theorem 16. *만약 n_h/N_h가 충분히 작고 N이 충분히 크며, n'이 커 a'_h/n'이 무시할 만하다면, 추정량의 분산은*

$$\hat{\mathbb{V}}(\bar{y}_{tp}) = \sum_{h=1}^{H} \left(\frac{a'^2_h s^2_h}{n_h} + \frac{a'_h (\bar{y}_h - \bar{y}_{tp})^2}{n'} \right)$$

으로 추정할 수 있다.

3.5 부차모집단에서의 비교

Definition 13. *특정한 부차모집단에 대한 지시변수가 δ_i일 때, 해당 부차모집단에 대한 모평균의 비추정량은*

$$\hat{\theta}_d = \frac{\sum_{i=1}^{n} \delta_i y_i / \pi_i}{\sum_{i=1}^{n} \delta_i / \pi_i}$$

로 주어진다.

Definition 14. *크기 N인 모집단에 크기가 N_1, N_2인 층(부차모집단)이 있다고 하자. N에서 n개의 SRS를 얻었고, 각각 n_1개와 n_2개의 개체를 표집하였다.*

$$\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_2$$

라고 할 때, Δ의 보수적인 분산추정량은

$$\hat{\mathbb{V}}(\hat{\Delta}) \approx \left(1 - \frac{n}{N}\right) \left(\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2} \right)$$

이다.

Chapter 4

일단계 집락추출

4.1 집락 크기가 같은 경우의 일단계 집락추출

Definition 15. *모집단이 N_I개의 집락들로 구분되어 $U_I = \{1, 2, \cdots, N_I\}$로 쓸 수 있다. U_i는 크기 M_i인 i번째 집락이다. y_{ij}를 i번째 집락에서 j번째 원소의 관심변수 값이라 하고*

$$Y_i = \sum_{j \in U_i} y_{ij},\ \bar{y}_i = \frac{Y_i}{M_i}$$

*라고 두자. **일단계 집락추출**에서는 U_I로부터 집락들이 표본추출단위인 표본 A_I를 추출하고 $n_I = |A_I|$개의 집락을 이용해 추론을 진행한다.*

Definition 16. *모든 집락의 크기가 일정해 $M_i = M$인 경우, 비복원 단순확률추출을 통해 N_I개의 집락 중 n_I개의 집락을 뽑아 선택된 집락의 모든 원소들을 관측하는 상황을 고려하자. 이때 모집단 총계 $Y = \sum_{i=1}^{N_I} Y_i$에 대한 추정량은*

$$\hat{Y}_{cl} = \frac{N_I}{n_I} \sum_{i \in A_I} Y_i$$

Definition 17. *모집단 평균을 추정하고자 하는 경우,*

$$\bar{y}_{cl} = \frac{\hat{Y}_{cl}}{N_I M} = \frac{1}{n_I} \sum_{i \in A_I} \bar{y}_i$$

이다.

Theorem 17. *모집단 평균의 분산은*

$$\mathbb{V}(\bar{y}_{cl}) = \frac{1}{n_I}\left(1 - \frac{n_I}{N_I}\right)\frac{1}{N_I - 1}\sum_{i=1}^{N_I}(\bar{y}_i - \bar{y}_{cl})^2 = \frac{1}{n_I}\left(1 - \frac{n_I}{N_I}\right)\frac{1}{N_I - 1}\frac{SSB}{M} = \frac{1}{n_I M}\left(1 - \frac{n_I}{N_I}\right)\sigma_b^2$$

으로 주어진다. 여기에서 SSB는 집락간 변동제곱합, σ_b^2은 $SSB/(N_I - 1)$이다.

Definition 18. ***급내상관계수***는 *각 집락 내에서 원소들의 상관관계를 의미한다.*

$$\rho = \frac{\mathrm{Cov}(y_{ij}, y_{ik} | j \neq k)}{\sqrt{\mathbb{V}(y_{ij})}\sqrt{\mathbb{V}(y_{ik})}} = \frac{1}{M-1}\frac{\sum_{i=1}^{N_I}\sum_{j \neq k}(y_{ij} - \bar{y}_{cl})(y_{ik} - \bar{y}_{cl})}{\sum_{i=1}^{N_I}\sum_{j=1}^{M}(y_{ij} - \bar{y}_{cl})^2} = 1 - \frac{M}{M-1}\frac{SSW}{SST}$$

이다. 이때 SST는 총변동제곱합, SSW는 집락내 변동제곱합이다.

Theorem 18. *급내상관계수는*

$$-\frac{1}{M-1} \le \rho \le 1$$

의 범위를 가지며, 최소값을 가지는 경우는 $SSB = 0$으로 집락내 평균 $\bar{y}_i$가 모두 동일한 경우이다. 반대로 최대값을 가지는 경우는 $SSW = 0$으로 집락 내의 모든 원소들이 동일한 경우이다.

Theorem 19. *N_I가 충분히 큰 경우에,*

$$\mathbb{V}(\hat{Y}_{cl}) \approx \mathbb{V}(\hat{Y}_{SRS}) \times (1 + (M-1)\rho)$$

이다.

Proof. $SST = SSB + SSW$이므로,

$$SSB = \frac{1}{M}(1 + (M-1)\rho)SST$$

가 성립한다. 이때

$$\begin{aligned}
\mathbb{V}(\hat{Y}_{cl}) &= \frac{1}{n_I}\left(1 - \frac{n_I}{N_I}\right)\frac{1}{N_I - 1}\frac{SSB}{M} \\
&= \frac{1}{n_I}\left(1 - \frac{n_I}{N_I}\right)\frac{1}{N_I - 1}\frac{(1 + (M-1)\rho)SST}{M^2} \\
&= \frac{1}{n_I M}\left(1 - \frac{n_I}{N_I}\right)\frac{SST}{N_I M - 1}(1 + (M-1)\rho) \times \frac{N_I M - 1}{(N_I - 1)M} \\
&\approx \frac{1}{n_I M}\left(1 - \frac{n_I}{N_I}\right)\sigma^2 \times (1 + (M-1)\rho) \\
&= \mathbb{V}(\hat{Y}_{SRS}) \times (1 + (M-1)\rho)
\end{aligned}$$

□

4.2 표본설계효과

Definition 19. ***표본설계효과****는 특정 표본설계 하에서의 분산을 같은 표본수의 비복원 단순확률추출에서의 분산으로 나눈 값, 즉*

$$\mathrm{deff}(\hat{Y}_p, \hat{Y}_{SRS}) = \frac{\mathbb{V}(\hat{Y}_p)}{\mathbb{V}(\hat{Y}_{SRS})}$$

으로 정의된다. 집락 크기가 같은 일단계 집락추출의 경우, 표본설계효과는 $1 + (M-1)\rho$이다.

Note 14. *만약 ρ가 음수이면, 즉 집락내의 원소의 성질이 이질적이면, 집락추출이 단순임의추출보다 효율적이다. 그렇지 않은 대부분의 경우에는 추정의 효율이 SRS에 비해 비효율적이다.*

Definition 20. ***유효표본수****는 특정 표본추출방법에서의 총 표본수가 n일 때, 동일한 효과를 내기 위해 SRS에서 몇 개의 표본이 필요한지를 의미한다.*

Note 15. *집락추출에서 유효표본수를 n^*이라고 하면,*

$$\frac{1}{n_I M}\left(1 - \frac{n_I}{N_I}\right)\sigma^2 \times (1 + (M-1)\rho) = \frac{1}{n^*}\left(1 - \frac{n^*}{N_I M}\right)\sigma^2$$

이 만족해야 하기에

$$n^* = \frac{n_I M}{1 + (M-1)\rho}$$

로 주어진다.

4.3 집락 크기가 다른 경우의 일단계 집락추출

Definition 21. *집락의 크기 M_i가 다른 경우, 모든 집락의 추출확률을 동일하게 하는 것은 최적이 아닐 수 있다. 각 집락의 일차표본포함확률을 $\pi_{Ii} = P(i \in A_I)$라 할 때, 모집단 총계의 추정량은*

$$\hat{Y}_{cl} = \sum_{i \in A_I} \frac{Y_i}{\pi_{Ii}}$$

이다.

Theorem 20. *고정표본수 집락추출의 경우, 위 추정량의 분산은*

$$\mathbb{V}(\hat{Y}_{cl}) = -\frac{1}{2} \sum_{i=1}^{N_I} \sum_{j=1}^{N_I} (\pi_{Iij} - \pi_{Ii}\pi_{Ij}) \left(\frac{Y_i}{\pi_{Ii}} - \frac{Y_j}{\pi_{Ij}} \right)^2$$

이다.

Definition 22. *M_i가 서로 다름에도 이를 무시하고 집락 모집단에서 비복원 단순확률추출을 사용하여 집락을 뽑는 방법을 **단순확률 집락추출**이라 한다. 이때 총계의 추정량은*

$$\hat{Y}_{SRC} = \frac{N_I}{n_I} \sum_{i \in A_I} Y_i$$

이며, 그 분산은

$$\mathbb{V}(\hat{Y}_{SRC}) = \frac{N_I^2}{n_I} \left(1 - \frac{n_I}{N_I} \right) \sigma_I^2$$

이다. 이때 σ_I는 Y_i들의 분산이다.

Definition 23. *집락크기 M_i가 다른 경우, **급락 동질성 계수**를 아래처럼 정의할 수 있다.*

$$\delta = 1 - \frac{\sum_{i \in U_I} \sum_{j \in U_i} (y_{ij} - \bar{y}_i)^2 / (N - N_I)}{\sum_{i \in U_I} \sum_{j \in U_i} (y_{ij} - \bar{y}_{SRC})^2 / (N-1)} = 1 - \frac{SSW/(N-N_I)}{SST/(N-1)}$$

이때 $\bar{y}_{SRC} = \bar{Y}$이다. 만약 집락크기가 모두 동일한 경우, δ는 급내상관계수와 동일하다.

Theorem 21.

$$K_I = \frac{N_I^2}{n_I} \left(1 - \frac{n_I}{N_I} \right),\ C* = \frac{1}{N_I - 1} \sum_{i=1}^{N_I} (M_i - \bar{M}) M_i \bar{y}_i^2$$

을 정의하면

$$\mathbb{V}(\hat{Y}_{SRC}) = \left(1 + \frac{N - N_I}{N_I - 1} \delta \right) \bar{M} \sigma^2 K_I + C * K_I$$

이며

$$\mathbb{V}(\hat{Y}_{SRS}) = \bar{M} \sigma_y^2 K_I$$

로 표현할 수 있으므로, 표본설계효과는

$$\text{deff}(SRC, SRS) = \left(1 + \frac{N - N_I}{N_I - 1} \delta \right) + \frac{C*}{\bar{M} \sigma_y^2}$$

이다. 따라서 $\delta > 0$이어 집락 내가 동질적이거나, $C > 0$이어 집락크기가 다르다면 추정의 효율이 감소하게 된다.*

4.4 평균 추정

Definition 24. *SRC에서 얻는 평균의 추정량은*

$$\bar{y}_{SRC} = \frac{\sum_{i\in A_I} M_i Y_i}{\sum_{i\in A_I} M_i}$$

이다.

Theorem 22. *SRC에서 얻는 평균추정량의 추정분산은*

$$\hat{\mathbb{V}}(\bar{y}_{SRC}) = \frac{1}{n_I \bar{M}^2}\left(1-\frac{n_I}{N_I}\right)\frac{1}{n_I-1}\sum_{i\in A_I} M_i^2(\bar{y}_i - \bar{y}_{SRC})^2$$

이다.

4.5 총계 추정과 표본 크기의 선택

Theorem 23. *만약 전체 개체수 M을 알면, τ_y는 $M\bar{y}_{SRC}$로 추정하며, 그 분산은*

$$\hat{\mathbb{V}}(M\bar{y}_{SRC}) = N_I^2\left(1-\frac{n_I}{N_I}\right)\frac{s_r^2}{n_I}$$

이다. 반면 개체수를 모르면, τ_y는

$$N_I\bar{y}_t = \frac{N}{n}\sum_{i=1}^{n} y_i$$

으로 추정하며, 그 분산은

$$\hat{\mathbb{V}}(N_I\bar{y}_t) = N_I^2\left(1-\frac{n_I}{N_I}\right)\frac{s_t^2}{n_I}$$

으로 추정한다. s_t^2는 y_i의 분산이다.

Theorem 24. *$\bar{y}_{SRC}$를 오차한계 B로 추정하려 한다면, 필요한 집락표본수는*

$$n^* = \frac{N_I^2 s_r^2}{N_I^2\left(\frac{B\bar{M}}{2}\right)^2 + N_I s_r^2}$$

이다.

Theorem 25. *총계를 추정하려 하는 경우, M을 안다면 오차한계 B로 할 때*

$$n^* = \frac{N_I^2 s_r^2}{N_I^2\left(\frac{B}{2N_I}\right)^2 + N_I s_r^2}$$

이며, M을 모른다면

$$n^* = \frac{N_I^2 s_t^2}{N_I^2\left(\frac{B}{2N_I}\right)^2 + N_I s_t^2}$$

이다.

Chapter 5

이단계 집락추출

Definition 25. ***이단계 집락추출****에서는 첫 번째 추출에서는 집락이 추출되고 두 번째 추출에서는 일단계에서 표본으로 추출된 집락으로부터 원소를 추출하는 방법을 의미한다. 이단계 집락추출에서는 집락 i의 모든 원소를 관측하지 못하므로 Y_i는 알지 못하며, 부차표본추출을 통하여 그 추정치 $\hat{Y}_i$를 얻고 이를 이용한다.*

Definition 26. *모집단 총계 Y의 비편향추정량은*

$$\hat{Y}_{2cl} = \sum_{i \in A_i} \frac{\hat{Y}_i}{\pi_{Ii}} = \sum_{i \in A_I} \sum_{k \in A_i} \frac{y_{ik}}{\pi_{k|i}\pi_{Ii}}$$

으로 표현되고 $\pi_{Ii} = P(i \in A_I), \pi_{k|i} = P(k \in U_i)$이다.

Theorem 26. *$\hat{Y}_{2cl}$의 분산은*

$$\mathbb{V}(\hat{Y}_{2cl}) = \underbrace{\mathbb{V}(\mathbb{E}[\hat{Y}_{2cl}|A_I])}_{V_{PSU}} + \underbrace{\mathbb{E}[\mathbb{V}(\hat{Y}_{2cl}|A_I)]}_{V_{SSU}}$$

로 각 표본추출단위를 추출함으로써 발생하는 분산의 합으로 표현할 수 있다. 이때

$$V_{PSU} = \sum_{i \in U_I} \sum_{j \in U_I} (\pi_{Iij} - \pi_{Ii}\pi_{Ij}) \frac{Y_i}{\pi_{Ii}} \frac{Y_j}{\pi_{Ij}}$$
$$V_{SSU} = \sum_{i \in U_I} \frac{V_i}{\pi_{Ii}} = \sum_{i \in U_I} \frac{1}{\pi_{Ii}} \sum_{k \in U_i} \sum_{l \in U_i} (\pi_{kl|i} - \pi_{k|i}\pi_{l|i}) \frac{y_{ik}}{\pi_{k|i}} \frac{y_{il}}{\pi_{l|i}}$$

이다.

Theorem 27. *아래와 같은 이단계 집락추출을 생각한다.*

1. *제 1단계 추출: N_I개의 집락으로부터 n_I개의 표본집락을 SRS로 추출*
2. *제 2단계 추출: 각 표본집락 i 내의 M_i개의 원소로부터 m_i개의 원소를 SRS로 추출*

이때

$$\hat{Y}_{2cl} = \frac{N_I}{n_I} \sum_{i \in A_I} \hat{Y}_i = \sum_{i \in A_I} \sum_{j \in A_i} \frac{N_I}{n_I} \frac{M_i}{m_i} y_{ij}$$

이며, 그 분산은

$$\mathbb{V}(\hat{Y}_{2cl}) = N_I^2 \left(1 - \frac{n_I}{N_I}\right) \frac{\sigma_I^2}{n_I} + \left(\frac{N_I}{n_I}\right) \sum_{i=1}^{N_I} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{\sigma_{2i}^2}{m_i}$$

이다. 이때 σ_{2i}^2은 i번째 집락에서의 분산이다.

Theorem 28. *이단계 집락추출에서 분산의 추정량은*

$$\hat{\mathbb{V}}(\hat{Y}_{2cl}) = \sum_{i\in A_I}\sum_{j\in A_I}\frac{\pi_{Iij}-\pi_{Ii}\pi_{Ij}}{\pi_{Iij}}\frac{\hat{Y}_i}{\pi_{Ii}}\frac{\hat{Y}_j}{\pi_{Ij}} + \sum_{i\in A_I}\frac{1}{\pi_{Ii}}\hat{V}_i$$

이다. 이단계 집락추출을 ***Theorem 24****와 동일한 방식으로 수행하는 경우,*

$$\hat{\mathbb{V}}(\hat{Y}_{2cl}) = N_I^2\left(1-\frac{n_I}{N_I}\right)\frac{s_I^2}{n_I} + \left(\frac{N_I}{n_I}\right)\sum_{i=1}^{N_I}M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{2i}^2}{m_i}$$

을 얻는다.

Theorem 29. *이단계 집락추출의 표본설계 효과는 n_I/N_I가 충분히 작을 때*

$$\mathrm{deff} = k(1+(\bar{m}-1)\delta)$$

이다. 이때 k는

$$k = \frac{(1+(\bar{M}-1)\delta)\sigma_y^2 + C^*/\bar{M}}{(1+(\bar{M}-1)\delta)\sigma_y^2}$$

이다.

Theorem 30. *만약 $M_1 = M_2 = \cdots = M_N = \bar{M}, m_1 = m_2 = \cdots = m_n = m$이라면,*

$$\hat{\mu} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}$$

이며,

$$\hat{\mathbb{V}}(\hat{\mu}) = \left(1-\frac{n}{N}\right)\frac{MSB}{nm} + \left(1-\frac{m}{\bar{M}}\right)\frac{MSW}{Nm}$$

이 된다. 또한

$$MSB = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{m}(\bar{y}_i-\hat{\mu})^2, \quad MSW = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i)^2$$

Theorem 31. *표집에 필요한 총비용이*

$$c = nc_1 + nmc_2$$

일 때,

$$m = \sqrt{\frac{\sigma_w^2 c_1}{\sigma_b^2 c_2}}$$

이면 분산이 최소화된다. 이후

$$\mathbb{V}(\hat{\mu}) = \frac{1}{n}\left(\sigma_b^2 + \frac{\sigma_w^2}{m}\right)$$

를 이용하여 n을 결정해주면 된다. 이때 σ_w^2은 MSW로, σ_b^2은 $(MSB-MSW)/m$으로 추정한다.

Chapter 6

층화집락추출

Definition 27. *모집단을 여러 개의 층으로 나눈 후 각각의 층에서 집락추출을 수행하는 방법을* ***층화집락추출****이라 한다. H개의 층이 있고 각 층에 N_{Ih}개의 집락이 있으며, 그 중 n_{Ih}개의 집락을 추출하여 조사를 진행하였다고 하자.*

이때 모평균의 추정량은

$$\bar{y}_c = \frac{\sum_{h=1}^{H} N_{Ih}\bar{Y}_{Ih}}{\sum_{h=1}^{H} N_{Ih}\bar{M}_{Ih}}$$

이며, 그 분산추정량은

$$\hat{\mathbb{V}}(\bar{y}_c) = \frac{1}{M^2}\sum_{h=1}^{H} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{s_{ch}^2}{n_h}$$

이다. 이때 M은 총 원소수로 추정해야 하는 값이며, s_{ch}^2은 h번째 층에서 얻은 $Y_i - \bar{y}_c m_i$의 분산이다.

Chapter 7

비추정과 회귀추정

7.1 비추정

Definition 28. *모집단의 보조정보가 알려져 있는 경우, 보조변수 X와 연구 대상인 Y 사이의 관계를 파악하여 이 관계를 Y의 모수 추정에 이용할 수 있다. 대표적으로,*

$$\sum_{i=1}^{N} y_i = \beta \sum_{i=1}^{N} x_i$$

가 만족할 때 이들의 비 β를 이용하는 ***비추정법****을 사용할 수 있다.*

Definition 29. ***모비율*** *R은 아래와 같이 정의된다.*

$$R = \frac{\tau_y}{\tau_x} = \frac{\mu_x}{\mu_y}$$

이때 τ_x, τ_y는 각각의 모합, μ_x, μ_y는 각각의 모평균이다. 이때 모비율 R의 추정량은

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}$$

이며, 그 분산추정량은

$$\hat{\mathbb{V}}(r) = \left(1 - \frac{n}{N}\right)\left(\frac{1}{\mu_x^2}\right)\frac{s_r^2}{n}$$

이다. 이때 s_r^2은 $y_i - rx_i$의 표본분산

$$s_r^2 = \frac{\sum_{i=1}^{n}(y_i - rx_i)^2}{n-1}$$

이다.

Note 16. *만약 μ_x를 모른다면 $\bar{x}$로 대체할 수 있다.*

Note 17. *모총계 τ_y는*

$$\hat{\tau}_y = r\hat{\tau}_x$$

으로 추정하며, 그 분산추정량은

$$\hat{\mathbb{V}}(\hat{\tau}_y) = N^2\left(1 - \frac{n}{N}\right)\frac{s_r^2}{n}$$

이다.

Note 18. *모평균 μ_y는*

$$\hat{\mu}_y = r\mu_x$$

으로 추정하며, 그 분산추정량은

$$\hat{\mathbb{V}}(\hat{\mu}_y) = \frac{1}{N^2}\left(1-\frac{n}{N}\right)\frac{s_r^2}{n}$$

이다.

Theorem 32. *r의 분산추정량은 $\hat{\rho} = \frac{s_{xy}}{s_x s_y}$를 이용하여*

$$\hat{\mathbb{V}}(r) = \left(1-\frac{n}{N}\right)\left(\frac{1}{\mu_x^2}\right)\frac{1}{n}(s_y^2 + r^2 s_x^2 - 2r\hat{\rho}s_x s_y)$$

처럼 쓸 수도 있다.

7.1.1 비추정과 SRS의 비교

Theorem 33. *비추정에서, 오차한계 d로 모비 β를 추정하기 위한 표본크기는*

$$n = \frac{N\sigma_r^2}{N(d\mu_x/z_{\alpha/2})^2 + \sigma_r^2}$$

으로 주어진다.

Theorem 34. *비추정에서, 오차한계 d로 모합 τ_y를 추정하기 위한 표본크기는*

$$n = \frac{N\sigma_r^2}{N(d/Nz_{\alpha/2})^2 + \sigma_r^2}$$

으로 주어진다.

Theorem 35. *비추정에서, 오차한계 d로 모평균 μ_y를 추정하기 위한 표본크기는*

$$n = \frac{N\sigma_r^2}{N(d/z_{\alpha/2})^2 + \sigma_r^2}$$

으로 주어진다.

Theorem 36. *비추정에서의 분산은*

$$\mathbb{V}(\hat{\mu}_y) = \left(1-\frac{n}{N}\right)\frac{\sigma_r^2}{n}$$

이며, SRS에서의 분산은

$$\mathbb{V}(\hat{\mu}_y) = \left(1-\frac{n}{N}\right)\frac{\sigma_y^2}{n}$$

*이다. 따라서 σ_r^2와 σ_y^2을 비교하여 둘의 효율을 비교할 수 있다. 이때 **Theorem 27**에서와 같이*

$$\sigma_r^2 = \sigma_y^2 + r^2\sigma_x^2 - 2r\rho\sigma_x\sigma_y$$

이므로, 상대효율성은

$$RE(SRC, SRS) = \frac{\mathbb{V}_{SRS}(\hat{\mu}_y)}{\mathbb{V}_{SRC}(\hat{\mu}_y)} = \left(1 + r^2\frac{\sigma_x^2}{\sigma_y^2} - 2r\rho\frac{\sigma_x}{\sigma_y}\right)^{-1}$$

으로 주어지고, 비추정량이 더 효율적일 조건은 $\rho > \frac{1}{2}r\frac{\sigma_x}{\sigma_y}$, 혹은

$$\frac{1}{2}\frac{CV(x)}{CV(y)} < \mathrm{corr}(x,y)$$

으로 쓸수도 있다. 따라서 일반적으로 x와 y의 correlation이 강할수록 상대효율성이 좋아진다.

7.1.2 층화표집에서의 비추정량

Definition 30. ***분리비추정량****은 각 층의 추정량을 얻는 데 비추정량을 사용하는 방법이다. 각 층에서,*

$$r_h = \frac{t_{yh}}{t_{xh}} = \frac{\bar{y}_h}{\bar{x}_h}$$

과 이를 통한

$$\hat{\tau}_{yh} = r_h \tau_{xh}$$

를 구하고, 이를 통해 모합 τ_y에 대한 분리비추정량

$$\hat{\tau}_y = \sum_{h=1}^{H} \hat{\tau}_{yh} = \sum_{h=1}^{H} b_h \tau_{xh}$$

을 얻을 수 있다. 그 분산추정량은

$$\hat{\mathbb{V}}(\hat{\tau}_y) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{rh}^2}{n_h}$$

이다.

Theorem 37. *모평균 μ_y에 대한 분리비추정량은*

$$\hat{\mu}_y = \frac{1}{N}\hat{\tau}_y = \sum_{h=1}^{H} \frac{N_h}{N} b_h \mu_{xh}$$

이며, 그 분산추정량은

$$\hat{\mathbb{V}}(\hat{\mu}_y) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{rh}^2}{n_h}$$

이다.

Definition 31. ***병합비추정량****은 각 층의 정보를 모두 합친 뒤 비추정을 수행하는 방법으로,*

$$r = \frac{\hat{\tau}_y}{\hat{\tau}_x} = \frac{\bar{y}_{st}}{\bar{x}_{st}}$$

이며 그 분산추정량은

$$\hat{\mathbb{V}}(r) = \frac{1}{\mu_x^2} \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{rh}^2}{n_h}$$

이다. 모합, 모평균에 대한 추정량은 각각 τ_x, μ_x를 곱하여

$$\hat{\tau}_y = r\tau_x, \quad \hat{\mathbb{V}}(\hat{\tau}_y) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{rh}^2}{n_h}$$

$$\hat{\mu}_y = r\mu_x, \quad \hat{\mathbb{V}}(\hat{\tau}_y) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{rh}^2}{n_h}$$

으로 각각이 주어진다.

Note 19. *일반적으로 분리비추정량이 더 효율적이지만, 요구하는 정보와 표본크기가 많아 비용이 많이 든다. 둘은 s_{rh}^2의 계산에서 사용하는 비 r이 층별비 r_h와 전체비 r이라는 점에서 다르다.*

7.2 회귀추정

Definition 32. *모집단 $\{(x_i, y_i), i = 1, 2, \cdots, N\}$에서 n개의 표본을 얻는다. 변수 x_i와 y_i가 선형의 관계 $y = \alpha + \beta x$를 가지고 있다고 가정할 때, α와 β에 대한 최소제곱추정량은*

$$\hat{\beta} = b = \frac{s_{XY}}{s_X^2} = \frac{rS_Y}{s_X}$$
$$\hat{\alpha} = a = \bar{Y} - b\bar{X}$$

이며, y의 평균 μ_y의 회귀 추정량은

$$\hat{\mu}_y = a + b\mu_x = \bar{Y} + b(\mu_x - \bar{X})$$

이다.

Theorem 38. *회귀추정량의 편의는*

$$bias(\hat{\mu}_y) = \mathbb{E}[\hat{\mu}_y - \mu_y] = -\mathrm{Cov}(b, \bar{X})$$

이며, 분산은

$$\hat{\mathbb{V}} = \left(1 - \frac{n}{N}\right) \frac{MSE}{n}$$

으로 추정된다.

7.2.1 차이추정

Definition 33. *μ_y에 대한 차이추정량은*

$$\hat{y} = \bar{y} + (\mu_x - \bar{x}) = \mu_x + (\bar{d})$$

로 주어진다. 그 분산추정량은

$$\hat{\mathbb{V}}(\hat{\mu}) = \left(1 - \frac{n}{N}\right) \frac{s_d^2}{n}$$

이다. 이때 $d_i = y_i - x_i$, s_d^2은 그 분산추정량이다.

Chapter 8

계통추출

8.1 계통추출

Definition 34. *단순확률추출을 할 때 표본을 뽑는 가장 간단한 방법은 표본추출틀 하에서 일정한 간격을 두고 표본을 선택하는 것이다. 이러한 추출방법을* ***계통추출****이라 한다.*

Note 20. *계통추출은 매우 간단하며, 전체 표집틀을 가지고 있지 않아도 되며, 비용이 적게 든다는 장점이 있다.*

Definition 35. *처음 k개 원소 중 하나를 임의로 고르고, 그 이후 k개마다 원소를 추출해 얻은 표본을 1 in k 계통표본이라 부른다.*

Note 21. *만약 전체 모집단의 원소 수를 안다면, 첫 원소를 $1/k$의 확률로 선택하는 대신 계통추출의 표본수에 비례하게 표본추출을 수행하는 방법 역시 사용할 수 있다. 둘째 방법은 N를 알 때만 사용할 수 있지만 불편추정량을 얻을 수 있는 반면, 첫째 방법은 N을 모를 때도 사용할 수 있지만 약간의 편의를 가진다.*

Theorem 39. *계통추출에서 모평균의 추정량은*

$$\mathbb{E}[\bar{y}_{sy}] = \frac{\sum_{i=1}^{n} y_i}{n}$$

이며, 그 분산추정량은 단순확률추출에서와 같이

$$\hat{\mathbb{V}}(\bar{y}_{sy}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

이다.

Note 22. *단, 추정 방법이 같다고 해서 실제 분산이 같다는 것은 아니다. 실제로는*

$$\mathbb{V}(\bar{y}_{sy}) = \frac{\sigma^2}{n}(1 + (n-1)\rho)$$

이며, 이때 ρ는 급내상관계수이다. 따라서 급내상관계수가 작은 경우에 분산이 작으며, 그렇지 않은 경우 분산이 크다. 계통추출된 표본 내의 원소들이 동질적일 때보다는 이질적일 때 그 분산이 작아진다.

Definition 36. *모집단이* ***랜덤모집단****이라는 것은 모집단의 원소들이 랜덤한 순서로 배치되어 있음을 의미한다.*
모집단이 ***순서모집단****이라는 것은 상승, 혹은 하락하는 순서로 배치된 모집단을 의미한다.*
주기모집단*은 모집단의 원소들이 일정한 주기성을 가지고 배치되어 있음을 의미한다.*

Note 23. *만약 순서모집단일 경우 급내상관계수가 낮아 계통추출에서 얻는 추정량의 분산이 작다. 따라서 분산추정량은 실제 분산을 과대평가한다. 반면 주기모집단일 경우 급내상관계수가 커 계통추출에서 얻는 추정량의 분산이 크다. 따라서 분산추정량은 실제 분산을 과소평가한다.*

Note 24. *모합, 모비율 등의 추정은 이를 통해 쉽게 수행할 수 있으며, 필요한 표본크기의 경우 SRS와 동일하다.*

8.2 반복계통추출

Definition 37. ***반복계통추출****은 N명의 모집단에서 n명을 계통표집하는 대신, n'명을 계통추출하는 것을 n_s번 반복하여*

$$(Y_{11}, Y_{12}, \cdots, Y_{1n'}), (Y_{21}, Y_{22}, \cdots, Y_{2n'}), \cdots, (Y_{n_s,1}, Y_{n_s,2}, \cdots, Y_{n_s,n'})$$

을 얻는 방법을 의미한다.

Definition 38. *n_s개의 표본평균 $\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_{n_s}$을 얻었다. 이때 평균 추정량은*

$$\hat{\mu} = \frac{1}{n_s} \sum_{i=1}^{n_s} \bar{y}_i$$

이며, 그 분산추정량은

$$\hat{\mathbb{V}}(\hat{\mu}) = \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n_s}$$

이다. 이때 s_y^2은 $\bar{y}_i$의 분산추정량이다.

8.3 모형을 이용한 추정

Theorem 40. *$N = nk$가 충분히 클 때,*

$$\rho = \frac{(k-1)nMSB - SST}{(n-1)SST} \approx \frac{MSB - MST}{(n-1)MST}$$

이다. 이때

$$MSB = \frac{n}{k-1} \sum_{i=1}^{k} (\bar{y}_i - \bar{\bar{y}})^2, \quad MST = \frac{1}{nk-1} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{\bar{y}})^2$$

이다.

Theorem 41. *선형추세가 있는 자료에서는 아래를 분산추정량으로 사용할 수 있다.*

$$\hat{\mathbb{V}}(\bar{y}_{sy}) = \left(1 - \frac{n}{N}\right) \frac{1}{2n(n-1)} \sum_{i=1}^{n-1} d_i^2$$

이때 $d_i = y_{i+1} - y_i$이다.

Chapter 9

모집단 크기의 추정

9.1 포획-재포획 추정

Definition 39. *집단의 크기가 N일 때, 해당 집단에 유입/유출이 없고 모든 개체가 포획될 확률이 동일하다고 하자. 그렇다면*

1. *t개의 개체를 포획하고 표지를 붙인 후 방류*
2. *일정 시간이 지난 후 n개의 개체를 포획하고 표지가 있는 개체의 개수 s을 확인*

했을 때, N의 추정량은

$$\hat{N} = \frac{nt}{s}$$

이며, 그 분산은

$$\hat{\mathbb{V}}(\hat{N}) = \frac{t^2 n(n-s)}{s^3}$$

으로 추정된다.

Proof. $\hat{N}$은 비추정량으로 표현할 수도 있다. 재포획 상황에서, i번째 물고기에 표지가 달린지 여부를 x_i로 표현하고, 모든 i에 대해 $y_i = 1$이라 하자. 그렇다면

$$\tau_x = t, \quad \tau_y = N$$

이며 재포획으로부터 얻은 $\beta = \frac{\tau_y}{\tau_x}$의 비추정량은

$$b = \frac{n}{s}$$

이다. 따라서 비추정량으로 얻은 Y의 추정량이 $\frac{nt}{s}$이다. 한편 그 분산은

$$\hat{\mathbb{V}}(\hat{N}) = \hat{N}^2 \left(1 - \frac{n}{\hat{N}}\right) \frac{s_e^2}{n} = \frac{nt^2(t-s)}{s^2} = \hat{N}^2 \left(1 - \frac{n}{\hat{N}}\right) \frac{n(n-s)}{s(n-1)} \approx \frac{t^2 n(n-s)}{s^3}$$

으로 추정된다. □

Definition 40. *앞서 고정된 n개의 개체를 포획하는 direct sampling과는 달리, inverse sampling에서는 재포획에서 s마리의 꼬리표 달린 개체가 포획될 때까지 임의표집한다. 즉*

$$n \sim NB(s, \frac{t}{N})$$

이다. 이때 N에 대한 비편향추정량은

$$\hat{N} = \frac{nt}{s}$$

이며, 그 추정분산은

$$\hat{\mathbb{V}}(\hat{N}) = \frac{t^2 n(n-s)}{s^2(s+1)}$$

으로 direct sampling과 유사하다.

9.2 사각표집

Definition 41. *분할된 구획인 사각에서 밀도를 측정하고, 밀도와 총면적을 곱하여 개체의 총수를 추정하는 방법을* ***사각표집****이라 한다. 총면적 A의 영역을 면적 a의 K개 사각으로 나눈 뒤, k개의 사각을 단순임의표집하여 개체수를 세면 i번째 사각에서의 개체수 Y_i이라 할 때 $Y_i \sim \text{Poisson}(\lambda a)$를 따른다. 이때 λ는 단위면적당 개체수를 의미한다.*

Theorem 42. *면적당 개체수 λ는*

$$\hat{\lambda} = \frac{1}{a}\bar{Y} = \frac{1}{ak}\sum_{i=1}^{k} Y_i$$

으로 추정한다. 또한 전체 개체수 N은

$$\hat{N} = \hat{\lambda} A$$

로 추정한다.

Theorem 43. *$\hat{N}$은 불편추정량이며, 그 분산추정량은*

$$\hat{\mathbb{V}}(\hat{N}) = A^2 \left(\frac{\hat{\lambda}}{ak}\right)$$

Definition 42. ***적재사각표집****은 개체수를 정확히 세는 대신 존재여부만을 기록하는 사각표집 방법을 의미한다. 푸아송모형을 가정하면*

$$Z_i = \begin{cases} 1 & I(Y_i \geq 1) \\ 0 & I(Y_i < 0) \end{cases}$$

이고, $P(Z_i = 1) = 1 - e^{-\lambda a}$이다. 그렇다면 $Z_i = 0$인 사각의 개수를 y라 하면

$$\hat{\lambda} = -\frac{1}{a}\log\left(\frac{y}{k}\right)$$

이며,

$$\hat{M} = \hat{\lambda} A, \quad \hat{\mathbb{V}}(\hat{M}) = \frac{A^2}{ka^2}(e^{\hat{\lambda} a} - 1)$$

9.3 임의화 반응

Definition 43. *민감한 질문을 해야 하는 경우, 전혀 무관한 무해한 질문을 동시에 하여 응답자로 하여금 임의선택된 질문에 대하여 답하도록 하는 절차이다. 예를 들어,*
(질문 1) 뇌물을 받은 적이 있습니까?
(질문 2) 뇌물을 받은 적이 없습니까?
그 다음 카드 7장에는 질문 1을, 카드 3장에는 질문 2를 적어 통에 넣고 임의로 1장의 카드를 선택하여 예, 아니오의 답만 기록한다. 그렇다면 뇌물을 받은 비율을 p_s, 질문 1을 뽑은 비율을 π라 하면, 예의 비율은

$$p = p_s\pi + (1 - p_s)(1 - \pi)$$

이며, 이로부터

$$p_s = \frac{p - (1 - \pi)}{2\pi - 1}$$

을 얻는다. 따라서

$$\hat{p}_s = \frac{\hat{p} - (1 - \pi)}{2\pi - 1}$$

이며,

$$\mathbb{E}[\hat{p}_s] = p_s,\ \hat{\mathbb{V}}(\hat{p}_s) = \frac{1}{(2\pi - 1)^2} \frac{\hat{p}(1 - \hat{p})}{n}$$

Chapter 10

비균등 확률추출

10.1 PPS 추출

Definition 44. *관심변수 y_i가 얼마나 큰지를 대략적으로 나타내는 크기척도에 비례하도록 표본을 복원추출하는 방법을* ***PPS*** *추출이라고 한다. 모집단이 $\{y_1, y_2, \cdots, y_N\}$이고 각 원소 i에서 크기척도가 M_i일 때, 원소 i의 일회 추출확률은*

$$p_i = \frac{M_i}{\sum_{i=1}^{N} M_i}$$

이다. 이때 모총계에 대한 비편향추정량은

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$$

이며, 이를 ***Hansen-Hurwitz****(HH) 추정량이라 한다.*

Theorem 44. *HH 추정량의 비편향 분산 추정량은*

$$\hat{\mathbb{V}}(\hat{\tau}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{\tau}\right)^2$$

이다.

10.1.1 PPS 집락추출

Definition 45. *집락추출에서 PPS 추출을 수행하는 경우, 집락의 크기를 크기척도로 이용할 수 있다. 그렇다면 i번째 집락 관측값들의 평균을 $\bar{y}_i$라 할 때*

$$\frac{y_i}{p_i} = \frac{y_i}{M_i} \times M = M\bar{y}_i$$

이므로,

$$\hat{\tau}_{pps} = \frac{M}{n}\sum_{i=1}^{n}\bar{y}_i, \quad \hat{\mu}_{pps} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$$

이다. 또한 그들의 분산추정량은

$$\hat{\mathbb{V}}(\hat{\tau}_{pps}) = \frac{M^2}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i - \hat{\mu}_{pps}), \quad \hat{\mathbb{V}}(\hat{\mu}_{pps}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i - \hat{\mu}_{pps})$$

로 주어진다. 이는 첫 단계를 PPS로 추출하는 모든 다단계 추출에서도 동일하다.

Chapter 11

분산추정법

Note 25. *좋은 분산 추정량의 성질로는 아래를 생각해볼 수 있다.*

- *비편향이거나, 편향이 매우 작을 것*
- *분산 추정량의 분산이 작아 안정적일 것*
- *음수 값을 취하지 않을 것*
- *계산이 복잡하지 않을 것*

Note 26. *비균등 확률추출의 경우 HT 추정량의 분산을 구하는 방법으로는 음의 분산추정량을 얻을 수 있다. 이러한 문제를 해결하기 위해,*

$$\hat{\mathbb{V}}(\hat{Y}_{HT}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\frac{ny_i}{\pi_i} - \hat{Y}_{HT})^2$$

을 ***단순분산추정량****으로 대용적으로 이용할 수 있다. 일반적으로 이는 보수적인 분산추정법이다.*

11.1 선형화 분산 추정법

Definition 46. *만약 $\bar{y}$가 참값 $\bar{Y}$에 대하여*

$$\bar{y} = \bar{Y} + O_p(n^{-1/2})$$

를 만족하고, $g(\bar{y})$를 추정량으로 사용한다면,

$$g(\bar{y}) = g(\bar{Y}) + \triangledown g(\bar{Y})^T (\bar{y} - \bar{Y}) + O_p(n^{-1})$$

이며, 분산 추정량은

$$\hat{\mathbb{V}}(g(\bar{y})) = \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial g(\bar{y})}{\partial y_i} \frac{\partial g(\bar{y})}{\partial y_j} \widehat{\mathrm{Cov}}(\bar{y}_i, \bar{y}_j)$$

으로 주어진다. 이때 $\bar{y}$는 p차원이다.

11.2 복제치 분산 추정법

Definition 47. ***랜덤 그룹 방법****은 G개의 표본그룹을 만든 후 각각에서 점추정치를 구하고, 그들의 변동으로부터 분산을 추정하는 방법이다. 독립적 랜덤 그룹법에서는*

1. *주어진 표본추출방법을 따라 첫째 표본 $A_{(1)}$과 그로부터 θ에 대한 $\hat{\theta}_{(1)}$을 구한다.*
2. *동일한 방법으로 $A_{(2)}$를 뽑고 $\hat{\theta}_{(2)}$를 구한다.*
3. *동일한 방식으로 G개의 표본 그룹을 뽑고 $\hat{\theta}_{(1)}, \cdots, \hat{\theta}_{(2)}$를 구한다.*
4. *모수에 대한 최종 점추정량으로*

$$\hat{\theta} = \frac{1}{G}\sum_{i=1}^{G}\hat{\theta}_{(i)}$$

를, 분산추정량으로

$$\hat{\mathbb{V}}(\hat{\theta}) = \frac{1}{G(G-1)}\sum_{i=1}^{G}(\hat{\theta}_{(i)} - \hat{\theta})^2$$

를 얻는다.

Definition 48. ***잭나이프 방법****을 이용해 분산을 추정할 수도 있다. i번째 관측값을 제외한 $n-1$개의 표본으로부터 얻은 추정량을 $\hat{\theta}^{(-i)}$라 할 때,*

$$\hat{\mathbb{V}}(\hat{\theta}) = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}^{(-i)} - \hat{\theta})^2$$

를 잭나이프 분산추정량으로 이용할 수 있다.

Chapter 12

무응답과 무응답대체법

Definition 49. *무응답은 조사 단위 자체가 응답을 하지 않는 단위무응답과, 일부 항목에 대해서만 응답을 하지 않는 항목무응답으로 나뉠 수 있다. 단위무응답의 경우 재조사나 무응답가중치조정을 통하여 처리하고, 항목무응답의 경우에는 결측값대체를 통하여 처리한다.*

12.1 단위무응답의 처리

Definition 50. ***가중치조정법****에서는 응답자들의 가중치를 조정하여 추정량의 편향을 줄이려 한다. 무응답이 없을 때 HT 추정량이*

$$\hat{Y}_{HT} = \sum_{i \in A} \frac{1}{\pi_i} y_i$$

이며, 응답확률을 ϕ_i라 하면

$$\hat{Y}_{NWA} = \sum_{i \in A_R} \frac{1}{\pi_i \hat{\phi}_i} y_i$$

이 된다. ϕ_i는 층정보 혹은 보조변수 등을 통해 추정할 수 있다. 이외에도 모집단에 대한 정보를 통해 사후층화와 유사한 방식으로 적절히 가중치를 조정할 수 있다.

Definition 51. ***재조사****를 하는 경우를 고려해 보자. 먼저 전체 N명 중 응답할 사람이 N_R, 무응답한 사람이 N_M명이라 하자. 그리고 일차 표본 크기가 n_R, n_M이며, 재조사를 통해 무응답자 중 ν의 비율만큼에게 조사를 다시 완료했다고 하자.*

모평균에 대한 비편향 점추정량은

$$\bar{y}* = \frac{n_R}{n} \bar{y}_R + \frac{n_M}{n} \bar{y}_M$$

이다. 이때 $\bar{y}_R$은 n_R명에게 얻은 관심변수 평균, $\bar{y}_M$은 νn_M명에게 얻은 관심변수 평균이다. 그리고 그 분산추정량은

$$\hat{\mathbb{V}}(\bar{y}*) = \left(1 - \frac{n}{N}\right) \frac{s^2}{n} + \frac{W_2 s_2^2}{n} \left(\frac{1}{\nu} - 1\right)$$

이다. s^2은 전체 표본에서 얻은 분산추정량, $W_2 = N_M/N$, s_2^2은 무응답 표본에서 얻은 분산추정량이다.

Theorem 45. *조사에 드는 비용을*

$$C = nc_0 + n_R c_1 + \nu n_M c_2$$

라고 하자. 이때의 최적재조사비율은

$$\nu = \sqrt{\frac{c_0 + c_1 N_R/N}{c_2} \times \frac{\sigma_2^2}{\sigma^2 - W_2 \sigma_2^2}}$$

이며, 분산 V_0를 위한 최소표본크기는

$$n = \frac{N(\sigma^2 + (\frac{1}{\nu} - 1)W_2\sigma_2^2)}{NV_0 + \sigma^2}$$

이다.

12.2 항목무응답의 처리

Definition 52. *항목무응답은 적당한 방식으로 결측된 값을 추정하여 대체함으로써 해결한다. 결측자료에 대한 최적예측치, 회귀모형 등을 사용할 수 있다.*