제 5장: 계통추출

5.1 정의및 개요

- The basic idea: Suppose that a sample of n names is to be selected from a long list. A simple way to make this selection is to choose an appropriate interval and to select names at equal interval along the list. For instance, every tenth name might be selected. If the starting point for this regular selection process is random, the result is a `systematic sampling`.

- Advantages of this method over SRS:
  - It is easier to draw a sample and often easier to execute without mistakes.
  - 전체 표집틀을 가지고 있지 않아도 된다. 공장의 컨베이어 벨트 형태의 공정관리, 선거 출구조사
  - systematic 한 편의가 발생 할 수 있다.
  - It can provide greater information per unit cost than SRS can provide.

- Definition: A sample obtained by randomly selecting one element from the first k elements in the frame and every kth element thereafter is called a *1 in k systematic sample*, with a random start.

  Example: Suppose that we have a population.

  $y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9, y_{10}, y_{11}, y_{12}$.

  We wish to select 1 in $k = 3$.

  시작점만 알면 전체 표본이 결정됨.

(1) Method A:

Randomly select a sampling unit from the first 3 sampling units, and then select every 3rd unit from this unit.

처음 $1, 2, \ldots, k$중 하나를 랜덤하게 뽑고 이후 k간격을 두고 표본을 추출하는 방법

For instance, if y_2 is the first sample, we have a systematic samples y_2, y_5, y_8, y_{11}.

- As is seen, the population has been divided into

$$n = \frac{N}{k} = \frac{12}{3} = 4$$

groups, and from each group we have selected 1 sampling unit.

The $k = 3$ possible systematic samples.

```
 #1    #2   #3
---------------
 y1    y2   y3
 y4    y5   y6
 y7    y8   y9
 y10   y11  y12
```

- A systematic sample is selected like this pattern:

$$j \quad j+k \quad j+2k \quad j+3k$$

where $j = 1, 2$, or 3 and $k = 3$.

- In general,

$$j \quad j+k \quad j+2k, \ldots, j+(n-1)k.$$

- **The probability that a systematic sample out of possible samples is selected is $1/k$.**

- In case of $N \neq nk$:

  Consider $k = 5$. Then

$$2 < \frac{N}{k} = \frac{12}{5} = 2\frac{2}{5} < 3.$$

  The $k = 5$ possible systematic samples.

```
 #1    #2   #3   #4   #5
---------------------------
 y1    y2   y3   y4   y5
 y6    y7   y8   y9   y10
 y11   y12
```

  As is seen, the sample size is therefore either 2 or 3.

  Note that the probability of selecting any one is 1/5 regardless of whether or not the sample size is 2 or 3.

(2) Method A:

Assume that $N = nk = 12$ and suppose that we wish to select a 1 in $k = 3$ sample. A sampling unit (say, the jth unit) is randomly selected from the population. Let j be the 8−th unit.

Then

$$\frac{j}{k} = \frac{8}{3} = 2 \quad \text{with remainder } r = 2.$$

- Note that $r = 2 < k = 3$, $r = 0, 1, 2$, where

$$\begin{aligned} r &= 1 \quad y_1 \\ r &= 2 \quad y_2 \\ r &= 0 \quad y_3 \end{aligned}$$

as the starting point.

The systematic samples we obtain will be as follows

```
 #1    #2   #3
---------------
 y1    y2   y3
 y4    y5   y6
 y7    y8   y9
 y10   y11  y12
```

- 전제 N개의 표본중 하나를 랜덤하게 선택하여 전체 표본을 원 위에 배치하고 이를 시작점으로 k개 단위로 뽑는다. 쉽게 시작점과 k로 나눈 나머지가 같은 점들을 모두 뽑는다고 이해하면 된다.

- 다음으로 $N \neq nk$인 경우를 살펴보자. Assume $N = 11$.

Then

$$\frac{j}{k} = \frac{8}{3} = 2 \quad \text{with remainder } r = 2$$

so that y_2 is the starting point and select every $k = 3$rd sampling unit.

All possible results will be

```
 #1    #2    #3
---------------
 y1    y2    y3
 y4    y5    y6
 y7    y8    y9
 y10   y11
```

- The characteristic of this method is that the probability of selecting the systematic sample will be n/N and not $1/k$.

(3) 방법 A 와 방법 B의 비교

- This can be easily shown as follows: The probability of selecting, say y_2, y_5, y_8, or y_{11} is 1/11 respectively. When any of these are selected we get the #2 systematic sample. Hence the probability of selecting this sample is

$$\frac{1}{11} + \frac{1}{11} + \frac{1}{11} + \frac{1}{11} = \frac{4}{11}.$$

- Similarly the probability of selecting (y_3, y_6, y_9) will be 3/11.

- The main difference between Method A and B is that although the same systematic samples are obtained, **there is a difference in the probability of selection.**

- 다음에서 살펴보겠지만 방법 A는 일반적으로 크지는 않지만 bias를 지니고 있고 방법 B는 불편 추정량을 제공한다. 추가로 in case that N is unknown, we can not use Method B.

5.2 추정과 표본의 크기 결정

item y_{ij} denotes the jth element of the ith systematic sample, where $j = 1, 2, \ldots, n$ and $i = 1, 2, \ldots, k$.

(1) Bias

Case 1: Method A and $N = nk$

- Let

$$\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$$

  be the sample mean for the ith systematic sample.

- Then

$$\begin{aligned} E(\bar{Y}_{\text{sy}}) &= \frac{1}{k}(\bar{Y}_1 + \bar{Y}_2 + \cdots + \bar{Y}_k) \\ &= \frac{1}{k}\frac{1}{n}(y_1 + y_2 + \cdots + y_N) \\ &= \mu \end{aligned}$$

- The sample mean is an unbiased estimator.

- Example: Suppose that we have a population (6,3,4,9,2,5,1,7,8).

```
#1    #2   #3
--------------
6     3    4        16/3+12/3+17/3=5=mu
9     2    5
1     7    8
```

Case 2: Method A and $N \neq nk$

- Given the population $(y_1, y_2, y_3, y_4, y_5, y_6, y_7)$, select a 1 in $k = 3$ systematic sample.
- $N/k = 7/3$

```
#1     #2    #3
---------------
y1     y2    y3
y4     y5    y6
y7
```

- The mean is $\mu = (y_1 + \cdots + y_7)/7$.
- The systematic mean is

$$\begin{aligned} E(\bar{Y}_{\text{sy}}) &= \frac{1}{3}(\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3) \\ &= \frac{1}{3}\left[\frac{1}{3}(y_1 + y_4 + y_7) + \frac{1}{2}(y_2 + y_5) + \frac{1}{2}(y_3 + y_6)\right] \\ &\neq \frac{1}{7}(y_1 + \cdots + y_7) = \mu. \end{aligned}$$

- $\bar{y}_{\text{sy}}$ is a biased estimator of μ.

Case 3: Method B and $N \neq nk$

- Given the population $(y_1, y_2, y_3, y_4, y_5, y_6, y_7)$, select a 1 in $k = 3$ systematic sample.
- $N/k = 7/3$

```
#1     #2    #3
---------------
y1     y2    y3
```

y4 y5 y6

y7

- The probabilities of samples #1, #2, and #3 are 3/7, 2/7, and 2/7, respectively.

- Hence,

$$\begin{aligned} E(\bar{y}_{\text{sy}}) &= \frac{3}{7}(\bar{Y}_1) + \frac{2}{7}(\bar{Y}_2) + \frac{2}{7}(\bar{Y}_3) \\ &= \frac{3}{7}\left[\frac{1}{3}(y_1 + y_4 + y_7)\right] + \frac{2}{7}\left[\frac{1}{2}(y_2 + y_5)\right] + \frac{2}{7}\left[\frac{1}{2}(y_3 + y_6)\right] \\ &= \frac{1}{7}(y_1 + \cdots + y_7) = \mu. \end{aligned}$$

- $\bar{Y}_{\text{SY}}$ is an unbiased estimator of μ.

(2) 추정량의 분산, The variance of $\bar{Y}_{\text{sy}}$

- 분산분석에서의 제곱합 분해를 생각하여

$$\sigma^2_{\text{within}} = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \mu)^2$$

라고 하면

$$\text{var}(\bar{Y}_{\text{SY}}) = \sigma^2 - \frac{N-k}{N}\sigma^2_{\text{within}}$$

이 된다.

- 이제 이 추정량의 분산을 비복원단순임의표집의 분산

$$\text{var}(\overline{Y}_{\text{srs}}) = \frac{1}{n}\frac{N-n}{N-1}\sigma^2$$

과 비교하면

$$\text{var}(\bar{Y}_{\text{SY}}) < \text{var}(\overline{Y}_{\text{srs}}) \Leftrightarrow \sigma^2_w > \frac{N}{N-1}\sigma^2 \approx \sigma^2.$$

- $\mathrm{var}(\bar{Y}_{\mathrm{sy}})$ can be rewritten as

$$\mathrm{var}(\bar{Y}_{\mathrm{sy}}) = \sigma^2 - \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2,$$

  where

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \mu)^2.$$

- The greater the variation within the systematic samples, the smaller $\mathrm{var}(\bar{Y}_{\mathrm{sy}})$ becomes.

- A large variation with in a systematic sample indicates that the sample is heterogeneous. Hence when the sampling units within a systematic sample are heterogeneous, the precision of systematic sampling will increase.

- Consider an extreme example with the population

  1 2 3 4 5 | 1 2 3 4 5 | 1 2 3 4 5

  The units in the population show periodicity. As an example of a systematic sample, there is 2 2 2 which is homogeneous. $\mathrm{var}(\bar{y}_{\mathrm{sy}})$ becomes large.

- *A question that naturally arises is: How do we measure this homogeneity or heterogeneity* ?

- A measure which expresses the degree of homogeneity in a systematic sample is the intraclass correlation coefficient (ICC, 급내상관계수) ρ_{w} defined as

$$\rho_{\mathrm{w}} = \frac{E[(y_{ij} - \mu)(y_{ij'} - \mu)]}{E[(y_{ij} - \mu)^2]}.$$

- Using ρ, we may express $\mathrm{var}(\bar{y}_{\mathrm{sy}})$ as

$$\mathrm{var}(\bar{Y}_{\mathrm{sy}}) = \frac{\sigma^2}{n}[1 + (n-1)\rho_{\mathrm{w}}].$$

  위 공식의 증명은 교재 158쪽에 있고 조교 실습 시간에 다룬다.

or calculation purpose, the intraclass correlation coefficient $\rho_{\rm w}$ is rewritten as

$$\rho_{\rm w} = \frac{2}{n-1} \sum_{i=1}^{k} \sum_{j<j'}^{n} [(y_{ij} - \mu)(y_{ij'} - \mu)] \frac{1}{N\sigma^2}.$$

- When $\rho_{\rm w}$ is large+positive, $\text{var}(\bar{Y}_{\rm Sy})$ is large.

  When $\rho_{\rm w}$ is small+positive or negative, $\text{var}(\bar{Y}_{\rm Sy})$ is small.

  When $\rho_{\rm w} = 0$, $\text{var}(\bar{Y}_{\rm Sy})$ is equal to $\text{var}(\bar{Y}_{\rm srs})$.

- **When the units are homogeneous in the systematic sample, $\rho_{\rm w}$ is large and positive. When the units are heterogeneous in the systematic sample, $\rho_{\rm w}$ is small and positive or negative.**

- Consider an example: the population 1 2 3 | 4 5 6 | 7 8 9. Note that $\mu = 5$.

#1	#2	#3
1	2	3
4	5	6
7	8	9

  We have found $\rho_{\rm w} = -21/60$ and $\text{var}(\bar{Y}_{\rm Sy}) = 2/3$.

- Consider another example: the population 1 4 7 | 2 5 8 | 3 6 9. Note that $\mu = 5$.

#1	#2	#3
1	4	7
2	5	8
3	6	9

We have found $\rho_{\mathrm{w}} = 51/60$ and $\mathrm{var}(\bar{Y}_{\mathrm{Sy}}) = 6$.

(3) 분산의 추정량, $\widehat{\mathrm{var}}(\bar{Y}_{\mathbf{sy}})$

- For calculation purpose, the intraclass correlation coefficient ρ_{w} is rewritten as

$$\rho_{\mathrm{w}} = \frac{2}{n-1}\sum_{i=1}^{k}\sum_{j<j'}^{n}[(y_{ij}-\mu)(y_{ij'}-\mu)]\frac{1}{N\sigma^2}.$$

- When ρ is large and positive, $\mathrm{var}(\bar{Y}_{\mathrm{Sy}})$ is large. When ρ_{w} is small and positive or negative, $\mathrm{var}(\bar{Y}_{\mathrm{Sy}})$ is small.

  When $\rho_{\mathrm{w}} = 0$, $\mathrm{var}(\bar{Y}_{\mathbf{sy}})$ is equal to $\mathrm{var}(\bar{Y}_{\mathrm{srs}})$.

  교재에서 계통추출에서는 N이 크고 이 경우 $\rho_{\mathrm{w}} \approx 0$ 임을 이야기 하고 있다.

(4) 모합의 추정량

- Estimator of the population mean μ:

$$\bar{Y}_{\mathrm{Sy}} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

- Variance of $\bar{Y}_{\mathrm{Sy}}$:

$$\widehat{\mathrm{var}}(\bar{Y}_{\mathrm{Sy}}) = \frac{N-n}{N}\frac{s^2}{n}.$$

  assuming a randomly ordered population.

- Estimator of the population total τ:

$$\widehat{\tau} = N\bar{Y}_{\mathrm{Sy}}.$$

- Estimated variance:

$$\widehat{\text{var}}(\widehat{\tau}) = \widehat{\text{var}}(N\bar{Y}_{\text{sy}}) = N^2\widehat{\text{var}}(\bar{Y}_{\text{sy}}).$$

assuming a randomly ordered population.

(5) 모비율의 추정량

- Estimator of the population proportion p:

$$\hat{p}_{\text{sy}} = \bar{Y}_{\text{sy}} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

where $Y_i = 0$ if the ith element sampled does not possess the specified characteristic and $Y_i = 1$ if it does.

- Estimated variance of $\widehat{p}_{\text{sy}}$:

$$\widehat{\text{var}}(\widehat{p}_{\text{sy}}) = \frac{\widehat{p}_{\text{sy}}\ \widehat{q}_{\text{sy}}}{n-1}\frac{N-n}{N}.$$

assuming a randomly ordered population.

(6) 표본크기의 결정

- 모평균, Sample size required to estimate μ with a bound B on the error of estimation:

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2},$$

where $D = B^2/4$.

- 모합,

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}$$

where $D = B^2/(4N^2)$.

- 모비율, Sample size required to estimate p with a bound B on the error of estimation:

$$n = \frac{Npq}{(N-1)D + pq},$$

where $q = 1 - p$ and $D = B^2/4$.

5.3 모집단의 구조

계통표집이 효율적이기 위하여는

We should try to make the sampling units in the sample heterogeneous and keep ρ_w small.

The practical questions is: what kind of populations tend to generate systematic samples with heterogeneous sampling units ?

1. 임의모집단, Sampling units in a population are in random order.
   - For instance, we wish to estimate the average weight of $N = 3000$ freshmen student at SNU and a list of their names is available.
   - The sampling units in the systematic samples are randomly ordered. Hence, the systematic sample may be treated as if it were a random sample.
   - ρ_w might be small, so $\text{var}(\bar{Y}_{sy})$ and $\text{var}(\overline{Y}_{srs})$ will be approximately equal.
2. 순서모집단, 자기상관모집단 Ordered population.
   - For instance, we wish to estimate the yield of corn and have a population of farms. We may order the farms according to area and select a systematic sample.
   - It is clear that the selection of systematic sample will provide a heterogeneous sample and $\text{var}(\bar{Y}_{sy})$ will be smaller than $\text{var}(\overline{Y}_{srs})$.

  - A systematic sample will cover the whole population and will avoid chances of selecting samples containing too many large or small farms. That is, a systematic sample will tend to be more representative of the population than a random sample.

3. 주기모집단, Populations with periodic variations.

  - For instance, the sales of supermarket are high on Friday and Saturday, and low on Monday and Tuesday, and will have a weekly periodicity.
  - The sampling units are clearly homogeneous and ρ_w will be large.
  - $\text{var}(\bar{Y}_{\text{Sy}})$ will be larger than $\text{var}(\overline{Y}_{\text{srs}})$.

5.4 분산추정

(1) 단일표본에서 추정량의 분산 추정

- For practical application, we need to find an estimator of $\text{var}(\bar{Y}_{\text{Sy}})$ based on a single sample.
- Under certain conditions, we may consider a systematic sample to be approximately equal to simple random sampling and we are thus able to use the sample variance to estimate $\text{var}(\overline{Y}_{\text{srs}})$.
- To estimate $\text{var}(\bar{Y}_{\text{Sy}})$, we will use

$$\text{var}(\overline{Y}_{\text{srs}}) = \frac{N-n}{N}\frac{s^2}{n}.$$

- 위 추정량은 순서모집단에서는 과대추정을 주기모집단에서는 과소추정을 하는 경향이 있다.

(2) 반복계통표집

- $N = 162$명의 모집단에서 $n = 18$명을 계통 표집하는 경우를 생각한다.
- 우리가 배운 계통 표집은 위 표집이 $1/9$ 계통표집이므로 $1, 2, \ldots, 9$의 숫자중 하나를 등확률로 뽑고 간격 9를 두어 나머지 17개의 표본을 뽑는다.
- 반복계통표집은 이를 대신하여 크기가 3인 $1/(9 \cdot 6) = 1/54$ 계통표본을 $n_s = 6$개 얻는 것이다.

  이를 위하여 $1 - 54$의 숫자 중 6개를 선택하고 각 선택된 숫자를 시작점으로 크기가 3인 $1/54$ 계통표본을 얻는다.
- 이렇게 얻어진 표본은

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

  과 이들의 표본평균들은

$$\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_{n_s}.$$

- 평균 추정량:

$$\widehat{\mu} = \frac{1}{n_s} \sum_{i=1}^{n_s} \overline{Y}_i.$$

- 추정량의 분산:

$$\mathrm{var}(\widehat{\mu}) = \frac{K - n_s}{K - 1} \frac{\sigma_m^2}{n_s}.$$

  이고

$$\sigma_m^2 = \frac{1}{K} \sum_{k=1}^{K} (\overline{y}_k - \mu)^2, \quad \overline{y}_k = \frac{1}{n'} \sum_{j=1}^{n'} y_{k+(j-1)K}.$$

- 추정량 분산의 추정:

$$\mathrm{var}(\widehat{\mu}) = \frac{K - n_s}{K} \frac{s_m^2}{n_s}$$

이고

$$s_m^2 = \frac{1}{n_s - 1} \sum_{i=1}^{n_s} \left(\overline{Y}_i - \widehat{\mu} \right)^2.$$

(3) 모형을 이용한 추정

- 임의모집단

- 층화효과

- 선형추세