제 2장: 단순임의표집

유한모집단에서의 비복원 추출을 다룬다.

2.2 모수의 추정

유한모집단 $\{y_1, y_2, \ldots, y_N\}$ 에서 단순임의표본 $\{Y_1, Y_2, \ldots, Y_n\}$ 을 비복원 추출을 통하여 얻는다. 이 때 관심모수로는

- 모집단의 크기(population size), N
- 모합(population sum),
$$\tau = \sum_{i=1}^{N} y_i.$$
- 모평균(population mean),
$$\mu = \frac{1}{N}\tau = \frac{1}{N}\sum_{i=1}^{N} y_j.$$
- 모비율(population proportion), $p = \frac{1}{N}\tau$.
- 모분산(population variance),
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} \left(y_i - \mu\right)^2.$$
- 모표준편차(population standard deviation), $\sigma = \sqrt{\sigma^2}$
- 모 변동계수(population coefficient of variation)
$$\gamma = \frac{\sigma}{\mu}\left(= \sqrt{\frac{1-p}{p}}\right), \quad \mu, p > 0.$$
- 표집률: n/N.
- 표본 크기(sample size), n

- 표본 합(sample totla),

$$t = \sum_{i=1}^{n} Y_i.$$

- 표본평균(sample mean),

$$\widehat{\mu} = \frac{1}{n} t = \frac{1}{n} \sum_{i=1}^{n} Y_j = \bar{Y}.$$

- 표본비율(sample proportion), $\widehat{p} = \frac{1}{n} t$.

- 표본분산(sample variance),

$$\widehat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \widehat{\mu}\right)^2 \left(= \frac{n\widehat{p}(1-\widehat{p})}{n-1}\right).$$

- 표본표준편차(sample standard deviation), $\widehat{\sigma} = \sqrt{\widehat{\sigma}^2} = s$

- 표본 변동계수(sample coefficient of variation)

$$\widehat{\gamma} = \frac{\widehat{\sigma}}{\widehat{\mu}} \left(= \sqrt{\frac{n}{n-1} \frac{1-\widehat{p}}{\widehat{p}}}\right), \quad \widehat{\mu}, \widehat{p} > 0.$$

2.3 모평균 μ에 대한 추정

1. (Theorem 2.1)

   μ에 대한 추정량은 $\bar{Y}$ 이고 다음이 성립한다.

$$\mathrm{E}(\bar{Y}) = \mu, \quad \mathrm{var}(\bar{Y}) = \frac{\sigma^2}{n} \frac{N-n}{N-1}, \quad \mathrm{s.e.}(\bar{Y}) = \sqrt{\frac{\sigma^2}{n} \frac{N-n}{N-1}}.$$

2. 증명은 2.3.1절에 있다.

(증명) 교재 2.3.1 참조

$$Z_i = \begin{cases} 1 & \text{if } i\text{th 원소가 Sample에 선택} \\ 0 & \text{o.w.} \end{cases} \quad i=1,2,\ldots,N$$

$$E(Z_i) = \frac{n}{N}$$

$$var(Z_i) = E(Z_i^2) - E(Z_i)^2 = \frac{n}{N} - \left(\frac{n}{N}\right)^2 = \frac{n}{N}\left(\frac{N-n}{N}\right)$$

If $i \neq j$

$$E(Z_i Z_j) = P(Z_i=1, Z_j=1)$$

$$= P(Z_i=1)\, P(Z_j=1 \mid Z_i=1) = \frac{n}{N} \cdot \frac{n-1}{N-1}$$

$$Cov(Z_i, Z_j) = E(Z_i Z_j) - E(Z_i)E(Z_j)$$

$$= \frac{n-1}{N-1}\,\frac{n}{N} - \left(\frac{n}{N}\right)^2 = -\frac{1}{N-1}\left(\frac{N-n}{N}\right)\frac{n}{N}$$

이제 S = 크기가 n인 표본의 index set.

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{N} Z_i y_i$$

$$E(\bar{Y}) = E\left[\sum_{i=1}^{N} Z_i \frac{y_i}{n}\right] = \sum_{i=1}^{N} E(Z_i)\cdot\frac{y_i}{n}$$

$$= \sum_{i=1}^{N} \frac{n}{N}\cdot\frac{y_i}{n} = \frac{1}{N}\sum_{i=1}^{N} y_i = \mu$$

$$\mathrm{var}(\bar{Y}) = \mathrm{var}\left(\sum_{i=1}^{N} Z_i \frac{y_i}{n} \right)$$

$$= \frac{1}{n^2} \mathrm{var}\left(\sum_{i=1}^{N} Z_i \cdot y_i \right)$$

$$= \frac{1}{n^2} \left\{ \sum_{i=1}^{N} y_i^2 \,\mathrm{var}(Z_i) + \sum_{i=1}^{N} \sum_{j \neq i}^{N} y_i y_j \,\mathrm{Cov}(Z_i, Z_j) \right\}$$

$$= \frac{1}{n^2} \left\{ \frac{n}{N} \frac{N-n}{N} \sum_{i=1}^{N} y_i^2 - \frac{1}{N-1} \left(\frac{N-n}{N} \right) \left(\frac{n}{N} \right) \sum_{i=1}^{N} \sum_{j \neq i}^{N} y_i y_j \right\}$$

$$= \frac{1}{n^2} \frac{n}{N} \frac{N-n}{N} \left\{ \sum_{i=1}^{N} y_i^2 - \frac{1}{N-1} \left(\left(\sum_{i=1}^{N} y_i \right)^2 - \sum_{i=1}^{N} y_i^2 \right) \right\}$$

$$= \frac{1}{n^2} \frac{n}{N} \frac{N-n}{N} \cdot \frac{1}{N-1} \left\{ N \sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i \right)^2 \right\}$$

$$= \frac{1}{n} \frac{N-n}{N-1} \frac{1}{N^2} \left\{ N \sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i \right)^2 \right\}$$

$$= \frac{1}{n} \frac{N-n}{N-1} \cdot \sigma^2$$

Note that

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu)^2 = \frac{1}{N} \left\{ \sum_{i=1}^{N} y_i^2 - N \mu^2 \right\}$$

$$= \frac{1}{N^2} \left\{ N \sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i \right)^2 \right\}.$$

3. <u>(Theorem 2.2)</u>

$$\widehat{\mathrm{var}}(\overline{Y}) = \frac{s^2}{n}\frac{N-n}{N}$$

일 때

$$\mathrm{E}\big[\widehat{\mathrm{var}}(\overline{Y})\big] = \mathrm{var}(\overline{Y}) = \frac{1}{n}\frac{N-n}{N-1}\sigma^2.$$

(pf)

$$s^2 = \frac{1}{n-1} \sum_{i \in S} \frac{1}{n-1} (y_i - \bar{Y})^2$$

$$= \frac{1}{n-1} \sum_{i=1}^{N} z_i (y_i - \bar{Y})^2 = \frac{1}{n-1} \sum_{i=1}^{N} z_i \left((y_i - \mu) - (\bar{Y} - \mu) \right)^2$$

$$= \frac{1}{n-1} \left\{ \sum_{i=1}^{N} z_i (y_i - \mu)^2 - 2 \sum_{i=1}^{N} z_i (y_i - \mu)(\bar{Y} - \mu) + \sum_{i=1}^{N} z_i (\bar{Y} - \mu)^2 \right\}$$

$$= \frac{1}{n-1} \left\{ \sum_{i=1}^{N} z_i (y_i - \mu)^2 - n (\bar{Y} - \mu)^2 \right\}$$

$$E(s^2) = \frac{1}{n-1} \left\{ \sum_{i=1}^{N} E(z_i)(y_i - \mu)^2 - n E[\bar{Y} - \mu)^2 \right\}$$

$$= \frac{1}{n-1} \left\{ \frac{n}{N} \sum_{i=1}^{N} (y_i - \mu)^2 - n \, var(\bar{Y}) \right\}$$

$$= \frac{1}{n-1} \left\{ n \cdot \sigma^2 - n \cdot \left(\frac{N-n}{N-1} \right) \cdot \frac{\sigma^2}{n} \right\}$$

$$= \frac{1}{n-1} \left\{ n\sigma^2 - \frac{N-n}{N-1} \sigma^2 \right\} = \frac{1}{n-1} \frac{nN - n - N + n}{N-1} \sigma^2$$

$$= \frac{N}{N-1} \sigma^2.$$

따라서

$$E(\widehat{var}(\bar{Y})) = E\left[\frac{s^2}{n} \frac{N-n}{N} \right] = \frac{1}{n} \frac{N-n}{N} \cdot \frac{N}{N-1} \sigma^2$$

$$= \frac{1}{n} \frac{N-n}{N-1} \sigma^2 = var(\bar{Y})$$ 이다.

2.4 표본크기의 결정

(i) 오차한계(신뢰구간의 크기)를 이용한 방법과 (ii) 검정력을 이용한 방법이 있고 표본조사에 있어서는 전자를 사용한다.

1. $100(1-\alpha)\%$ 신뢰구간의 일반형태는

$$\widehat{\theta} \pm Z_{\alpha/2}\text{s.e.}(\widehat{\theta}) = \widehat{\theta} \pm \text{B}.$$

   여기서 B는 오차한계라 부른다.

2. 연구자는 유의수준 α가 정해진 상태에서 목표로하는 B의 수준에 맞게 표본의 크기를 정한다.

$$\begin{aligned}\text{B} &= Z_{\alpha/2} \cdot \text{s.e.}(\overline{Y}) = Z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}} \\ &\approx Z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}\frac{N-n}{N}}.\end{aligned}$$

$$n = \frac{Z^2_{\alpha/2} \cdot \sigma^2}{B^2 + Z^2_{\alpha/2}\frac{\sigma^2}{N}} = \frac{n_0}{1+\frac{n_0}{N}},$$

   이고 **여기서**

$$n_0 = \frac{1}{B^2}Z^2_{\alpha/2}\sigma^2.$$

   으로 복원추출시 표본 크기이다.

3. 표본크기를 결정하기 위하여는 σ^2를 알아야 하는데 보통 파일럿 스터디를 수행하여 (i) 표본분산을 이용하거나, (ii) 비모수적 방법들, 예를 들어

$$\sigma \approx \frac{\text{range}}{4} = \frac{1}{4}\big(Y_{(n)} - Y_{(1)}\big),$$

   또는 $1.35 \cdot \text{IQR}$나 $1.4826 \cdot \text{MAD}$를 사용한다. 여기서 IQR은 interquantile range, $\text{MAD} = \text{med}\big(Y_i - \overline{Y}\big)$이다.

4. 모합 τ를 추정할 때 표본크기의 공식은

$$n = \frac{Z_{\alpha/2}^2 \sigma^2}{\frac{B^2}{N^2} + \frac{Z_{\alpha/2}^2 \sigma^2}{N}}$$

이다 .

5. 모비율 p를 추정할 때는 $\overline{Y}$의 공식에서 σ^2 대신 $p(1-p)$를 사용

$$n = \frac{Z_{\alpha/2}^2 \cdot p(1-p)}{B^2 + \frac{Z_{\alpha/2}^2 \cdot p(1-p)}{N}}$$

이 경우 오차한계 B는 $p = 1/2$인 경우 최대값을 지니게 되고 최대오차한계를 원하는 수준으로 표본의 크기를 정한다.

$$\begin{aligned} n &= \left. \frac{Z_{\alpha/2}^2 \cdot p(1-p)}{B^2 + \frac{Z_{\alpha/2}^2 \cdot p(1-p)}{N}} \right|_{p=1/2} \\ &= \frac{Z_{\alpha/2}^2/4}{B^2 + \frac{Z_{\alpha/2}^2/4}{N}} = \frac{n_0}{1 + \frac{n_0}{N}}. \end{aligned}$$

여기서 $n_0 = \frac{Z_{\alpha/2}^2}{4B^2}$은 무한모집단(복원추출)을 위한 표본의 크기이다.

2.5 부-모집단, sub-population

1. 모집단의 부분집단에 관심이 있음. 부분집단이 관심집단인데 표집틀이 없는 경우 더 큰 집단을 모집단으로하여 표집.

   예, 으뜸대학 학생들 중 운전면허증 소지자를 대상으로 한 조사의 경우 으뜸대학 학생의 명단은 있으나 운전면허소지자의 명단은 없음. 부분집단이 관심집단인데 표집틀이 없는 경우 더 큰 집단을 모집단으로하여 표집.

2. 부-모집단의 크기를 N_1, 평균을 μ_1, 그리고 n개의 표본중 부-모집단에 해당하는 표본을

$$\{y_{11}, y_{12}, \ldots, y_{1,n_1}\}$$

이라 하자.

3. 평균 μ_1에 대한 추론을 생각하면

$$\widehat{\mu}_1 = \overline{Y}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} y_{1j}$$

이고

$$\mathrm{var}\big(\overline{Y}_1\big) : \widehat{\mathrm{var}}\big(\widehat{\mu}_1\big) = \frac{s_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1}\right).$$

그런데 앞의 예에서도 알 수 있듯이 <u>N_1을 알수가 없는 경우</u>, <u>$N : n \approx N_1 : n_1$의 관계</u>를 이용

$$\mathrm{var}\big(\overline{Y}_1\big) : \widehat{\mathrm{var}}\big(\widehat{\mu}_1\big) = \frac{s_1^2}{n_1}\left(\frac{N - n}{N}\right).$$

을 이용한다.

4. 총계와 모비율에 대한 추론도 비슷하게 계산한다.

2.6 두 모집단의 비교

1. 조사를 수행한 이후 특별한 두 부-모집단간 평균이나 비율의 차이가 있는지에 대한 검정을 수행한다.

2. 크기 N인 모집단($\mathcal{P}$)에 각각 크기가 N_1과 N_2인 두 개의 부-모집단($\mathcal{P}_1$과 $\mathcal{P}_2$)이 있음을 가정. N에서 n개의 SRS를 얻었더니 N_1에서 n_1개 N_2에서 n_2개의 개체가 표집되었다고 하자. 여기서 $n_1 + n_2 < n$.

3. 위의 경우를 두 부-모집단의로 부터의 서로 독립인 N_1에서 n_1을 N_2에서 n_2를 단순임의추출함과 동치라고 생각하고 계산한다.

   <u>두 모집단의 차이에 대한 검정을 위한 일반식</u>은 다음과 같다.

4. $\widehat{\theta}_1 - \widehat{\theta}_2$.

5. $\mathrm{var}\big(\widehat{\theta}_1 - \widehat{\theta}_2\big) = \mathrm{var}(\widehat{\theta}_1) + \mathrm{var}(\widehat{\theta}_2) - 2\mathrm{cov}\big(\widehat{\theta}_1, \widehat{\theta}_2\big)$.

6. 두 부-모집단의 모평균 차이에 대한 추론

$\widehat{\Delta} = \widehat{\mu}_1 - \widehat{\mu}_2 = \overline{Y}_1 - \overline{Y}_2.$

분산에 대한 추정량

$$\begin{aligned}
\widehat{\mathrm{var}}\big(\overline{Y}_1 - \overline{Y}_2\big) &= \widehat{\mathrm{var}}(\overline{Y}_1) + \widehat{\mathrm{var}}(\overline{Y}_2) \\
&= \frac{s_1^2}{n_1}\frac{N_1 - n_1}{N_1} + \frac{s_2^2}{n_2}\frac{N_2 - n_2}{N_2} \\
&\approx \frac{N-n}{N}\left\{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right\}.
\end{aligned}$$

7. 두 부-모집단의 (특정 사안에 대한) 모비율 차이에 대한 추론

$\widehat{\Delta} = \widehat{p}_1 - \widehat{p}_2.$

분산에 대한 추정량

$$\begin{aligned}
\widehat{\mathrm{var}}\big(\widehat{p}_1 - \widehat{p}_2\big) &= \widehat{\mathrm{var}}(\widehat{p}_1) + \widehat{\mathrm{var}}(\widehat{p}_2) \\
&= \frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1 - 1}\frac{N_1 - n_1}{N_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2 - 1}\frac{N_2 - n_2}{N_2} \\
&\approx \frac{N-n}{N}\left\{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1 - 1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2 - 1}\right\}.
\end{aligned}$$

<u>(부-모집단이 아닌) 모집단에서 모비율의 차이에 대한 추론</u>

8. <u>이항 모집단</u>에서 모비율 차이에 대한 추론

9. <u>다항 모집단</u>에서 모비율 차이에 대한 추론 ([a,b,c]를 가정한다)

$\widehat{\Delta} = \widehat{p}_a - \widehat{p}_b$

분산에 대한 추정량

$$
\begin{aligned}
\widehat{\mathrm{var}}\big(\widehat{p}_a - \widehat{p}_b\big) &= \widehat{\mathrm{var}}(\widehat{p}_a) + \widehat{\mathrm{var}}(\widehat{p}_b) - 2\mathrm{cov}(\widehat{p}_a, \widehat{p}_b) \\
&= \frac{\widehat{p}_a(1-\widehat{p}_a)}{n-1}\frac{N-n}{N} + \frac{\widehat{p}_b(1-\widehat{p}_b)}{n-1}\frac{N-n}{N} - 2[\mathrm{SOMETHING}] \\
&\approx \frac{N-n}{N}\left\{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1-1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2-1} - 2[\mathrm{SOMETHING}]\right\}.
\end{aligned}
$$

<u>부-모집단에서 모비율의 차이에 대한 추론</u>: 이 경우에 한하여 복잡하여 교재처럼 무한 모집단 처럼 취급하여 계산하겠다.

여성 유권자중 국민의힘당 과 더불어민주당의 지지율 차이

10. 교재의 예제 2.4