

제 3장: 총화임의표집

3.1 개요와 용어

1. 예제:

- 전자제품을 생산하는 기업 대상 조사: 대기업, 중견기업, 중소기업

- 유품대학의 성형수술 경험에 대한 조사:

유품대학 총학생의 수는 10000명이고 남학생이 6000명 여학생이 4000명이다.

- 우리나라 대통령 선거 여론조사: 지역, 연령별 총화

2. 특정 보조변수에 따른 부-모집단별 관심있는 특성치의 평균이 매우 다른 경우 총화를 한 이후 단순임의추출을 한다. 이를 총화추출(표집) 또는 총화임의표집이라고 한다.

3. 총화표집에서는 총내원소들은 동질적이고 층간은 이질적이다.

4. 종종 층을 나누는 것이 어려울 수 있다. 예를 들어 소규모 지역조사에서 도시지역과 농촌지역으로 나눈다고 할 때 도시와 농촌지역의 구분이 어렵고 또 다른 예로 유품기업 직원들의 한 달 외식비에 대하여 조사를 하는 경우 연봉을 기준으로 층을 나눈다면 층을 몇 개로 나누는게 좋은 지도 명확하지 않다.

3.2 총화표본의 추출과 표기

총화추출은 N 개의 표집단위로 구성된 모집단을 H 개의 층으로 나누고

$$N = \sum_{h=1}^H N_h$$

크기가 N_h 인 각 층으로부터 크기가 n_h 인 표본을 단순임의표집한다,

$$n = \sum_{h=1}^H n_h.$$

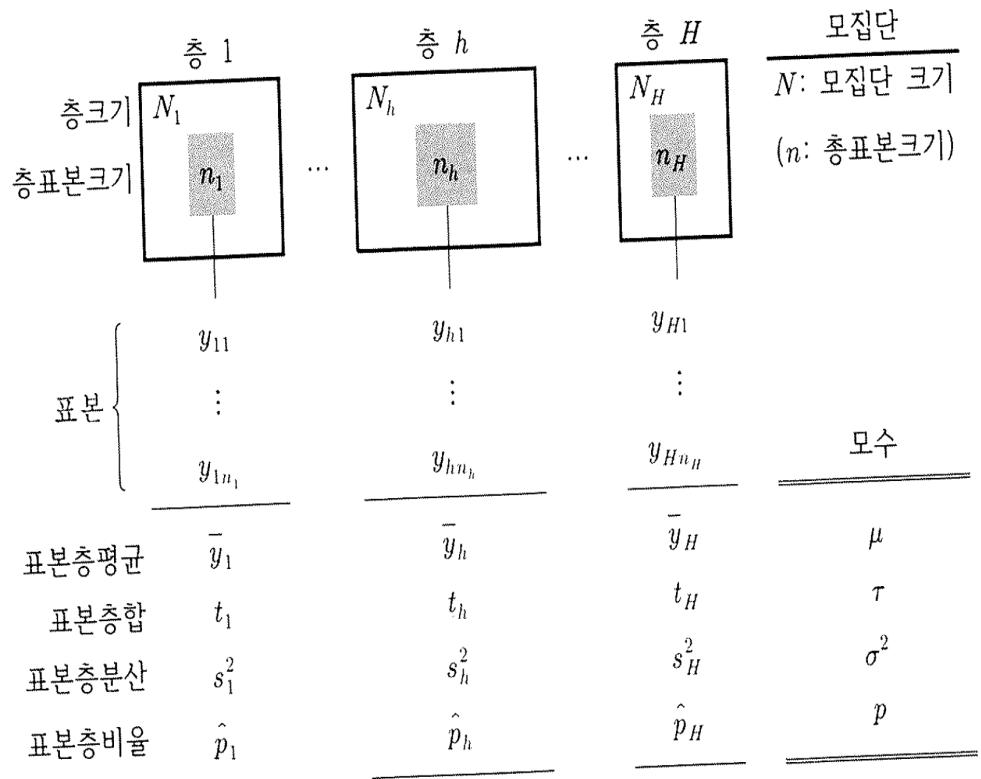


그림 3.1 층화표집

모집단 특성에 대한 표기

모집단 특성에 대한 표기	
H	총의 개수
N_h	h 번째 층의 크기
y_{hj}	$\sum_h h$ 의 j 번째 단위 ($h = 1, \dots, H; j = 1, \dots, N_h$)
$W_h = N_h/N$	$\sum_h h$ 의 층비중(stratum weight)
$\tau_h = \sum_{j=1}^{N_h} y_{hj}$	$\sum_h h$ 의 모합
$\mu_h = \tau_h/N_h$	$\sum_h h$ 의 모평균
$p_h = \tau_h/N_h$	$\sum_h h$ 의 모비율
$\sigma_h^2 = \sum_{j=1}^{N_h} (y_{hj} - \mu_h)^2 / N_h$	$\sum_h h$ 의 모분산
$\tau = \sum_{h=1}^H \tau_h$	모합
$\mu = \tau/N = \sum_{h=1}^H \tau_h/N = \sum_{h=1}^H \frac{N_h}{N} \mu_h = \sum_{h=1}^H W_h \mu_h$	모평균
$p = \sum_{h=1}^H W_h p_h$	모비율
$\sigma^2 = \sum_{h=1}^H \sum_{j=1}^{N_h} (y_{hj} - \mu)^2 / N$	모분산

표본 특성에 대한 표기

아래의 표에서 y_{hj} 는 Y_{hj} 로, \bar{y}_h 는 \bar{Y}_h 로 바꾼다.

표본 특성에 대한 표기	
y_{hj}	총 h 의 j 번째 단위 ($h = 1, \dots, H; j = 1, \dots, n_h$)
n_h	h 번째 층의 표본크기
$n = \sum_{h=1}^H n_h$	총표본의 수
$f_h(w_h) = n_h/n$	총 h 의 표집률(총 배정률)
$\bar{y}_h = \frac{1}{n_h} \sum_{j=1}^{n_h} y_{hj}$	총 h 의 표본평균
$t_h = n_h \bar{y}_h$	총 h 의 표본합
$s_h^2 = \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2 / (n_h - 1)$	총 h 의 표본분산

3.3 추정

(1) 모총계와 모평균의 추정

모총계

- 추정량: 층 h 에 대하여 $\hat{\tau}_h = N_h \bar{Y}_h$ 이다. 따라서

$$\hat{\tau}_{\text{st}} = \sum_{h=1}^H \hat{\tau}_h = \sum_{h=1}^H N_h \bar{Y}_h.$$

- 추정량의 분산:

$$\text{var}(\hat{\tau}_{\text{st}}) = \sum_{h=1}^H N_h^2 \text{var}(\bar{Y}_h) = \sum_{h=1}^H N_h^2 \frac{\sigma_h^2}{n_h} \left(\frac{N_h - n_h}{N_h - 1} \right).$$

- 분산 추정량:

$$\text{var}(\widehat{\tau}_{\text{st}}) = \sum_{h=1}^H N_h^2 \widehat{\text{var}}(\bar{Y}_h) = \sum_{h=1}^H N_h^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{s_h^2}{n_h}.$$

모평균

$$\mu = \sum_{h=1}^H W_h \mu_h.$$

- 추정량: 총 h 에 대하여 $\widehat{\mu}_h = \bar{Y}_h$ 이다. 따라서

$$\widehat{\mu}_{\text{st}} = \sum_{h=1}^H W_h \widehat{\mu}_h = \sum_{h=1}^H W_h \bar{Y}_h.$$

- 추정량의 분산:

$$\text{var}(\widehat{\mu}_{\text{st}}) = \sum_{h=1}^H W_h^2 \text{var}(\bar{Y}_h) = \sum_{h=1}^H W_h^2 \frac{\sigma_h^2}{n_h} \left(\frac{N_h - n_h}{N_h - 1} \right).$$

- 분산 추정량:

$$\widehat{\text{var}}(\widehat{\mu}_{\text{st}}) = \sum_{h=1}^H W_h^2 \widehat{\text{var}}(\bar{Y}_h) = \sum_{h=1}^H W_h^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{s_h^2}{n_h}.$$

예제 3.1

으뜸대학교 통계학과 학생 $N = 300$ 명 중 남학생은 $N_1 = 212$ 명 $N_2 = 88$ 명이다. $n = 60$ 명에 대하여 한 학기 동안 읽은 책의 평균 권수(μ)를 조사하고자 한다. 총화 표집을 이용 $n_1 = 42$ 명 $n_2 = 18$ 명을 추출한 조사 결과가 다음과 같을 때 평균 권수의 95% 신뢰구간을 계산하여라. 남학생과 여학생의 평균 권수에 차이가 있다고 할 수 있는가?

총	N_h	n_h	\bar{Y}_h	s_h^2
남학생	212	42	20.2	230.7
여학생	88	18	30.5	40.3
합계	300	60		

(2) 모비율의 추정

자료의 값이 0 또는 1인 경우의 모평균. 따라서 모평균의 추정량과 동일한 추정량을 사용. 유일하게 조심하여야 할 부분은

$$s_h^2 = \frac{n_h}{n_h - 1} \hat{p}_h(1 - \hat{p}_h)$$

이다.

모비율

$$\mu = \sum_{h=1}^H W_h p_h.$$

- 추정량: 총 h 에 대하여 모비율의 추정량 \hat{p}_h 이다. 따라서

$$\hat{p}_{\text{st}} = \sum_{h=1}^H W_h \hat{p}_h$$

- 추정량의 분산:

$$\text{var}(\hat{p}_{\text{st}}) = \sum_{h=1}^H W_h^2 \text{var}(\hat{p}_h) = \sum_{h=1}^H W_h^2 \frac{\sigma_h^2}{n_h} \left(\frac{N_h - n_h}{N_h - 1} \right).$$

- 분산 추정량:

$$\begin{aligned} \widehat{\text{var}}(\hat{p}_{\text{st}}) &= \sum_{h=1}^H W_h^2 \widehat{\text{var}}(\hat{p}_h) \\ &= \sum_{h=1}^H W_h^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{s_h^2}{n_h} \\ &= \sum_{h=1}^H W_h^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1}. \end{aligned}$$

(3) 총내분산과 총간분산

본 절에서는 추가적으로 단순임의추출(simple random sample)과 층화추출(stratified sample)의 평균추정에서의 효율성을 비교한다. 단 층화추출에서 비례배정 (proportional allocation)을 가정한다.

$$n_h = n \cdot \frac{N_h}{N}.$$

1. \bar{Y}_n and \bar{Y}_{st}
2. 두 추정량 모두 불편추정량이어서 추정량의 분산의 크기를 통하여 효율성을 비교한다.
3. 추정량의 분산:

$$\begin{aligned} \text{var}(\bar{Y}_n) &= \frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right) \\ &= \frac{N-n}{N} \cdot \frac{\sigma^2}{N} \cdot \frac{N}{n}. \end{aligned} \quad (1)$$

$$\begin{aligned} \text{var}(\bar{Y}_{\text{st}}) &= \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \cdot \frac{N_h - n_h}{N_h - 1} \cdot \frac{\sigma_h^2}{n_h} \\ &= \frac{N-n}{N} \cdot \left\{ \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{\sigma_h^2}{N_h} \right\} \cdot \frac{N_h}{n_h} \\ &= \frac{N-n}{N} \cdot \left\{ \sum_{h=1}^H \frac{N_h}{N^2} \sigma_h^2 \right\} \cdot \frac{N}{n}. \end{aligned}$$

따라서

$$\underline{\frac{\sigma^2}{N}} \quad \text{과} \quad \underline{\sum_{h=1}^H \frac{N_h}{N^2} \sigma_h^2}$$

을 비교한다.

$$\begin{aligned}
\frac{1}{N}\sigma^2 &= \frac{1}{N^2} \sum_{h=1}^H \sum_{j=1}^{N_h} (y_{hj} - \mu)^2 \\
&= \frac{1}{N} \left\{ \frac{1}{N} \sum_{h=1}^H \sum_{j=1}^{N_h} (y_{hj} - \mu_h)^2 + \frac{1}{N} \sum_{h=1}^H \sum_{j=1}^{N_h} (\mu_h - \mu)^2 \right\} \\
&\geq \frac{1}{N} \left\{ \frac{1}{N} \sum_{h=1}^H N_h \sigma_h^2 \right\} \\
&= \sum_{h=1}^H \frac{N_h}{N^2} \sigma_h^2
\end{aligned}$$

이다.

따라서

$$\text{RE} = \frac{\text{var}(\bar{Y}_n)}{\text{var}(\bar{Y}_{\text{st}})} \geq 1.$$

4. 위의 계산에서 알 수 있듯 총화추출에서 모분산은 총내분산과 총간분산의 합으로 표현된다.

$$\sigma^2 = \frac{1}{N} \sum_{h=1}^H \sum_{j=1}^{N_h} (y_{hj} - \mu)^2 = \sum_{h=1}^H \frac{N_h}{N} \sigma_h^2 + \sum_{h=1}^H \frac{N_h}{N} (\mu_h - \mu)^2$$

$$\equiv \sigma_w^2 + \sigma_b^2$$

분산분석표(비례배정) $n_h/N_h = n/N$

변동요인	자유도	제곱합
총간	$H-1$	$SSB = \sum_{h=1}^H \sum_{j=1}^{N_h} (\mu_h - \mu)^2$
총내	$N-H$	$SSW = \sum_{h=1}^H \sum_{j=1}^{N_h} (y_{hj} - \mu_h)^2$
계	$N-1$	$SST = \sum_{h=1}^H \sum_{j=1}^{N_h} (y_{hj} - \mu)^2$

3.4 표본배정

(1) 모평균

3.4.1 최적배정

표본배정은 표본의 크기 n 이 주어졌을 때 총 표본 n 을 각 층에 몇 개를 배정할지, 즉 n_h 의 결정에 대한 문제이다. 본 수업에서는 각 층의 조사비용이 다른 경우 총 비용의 제약조건 하에서 층화 추정량의 분산을 최소로 하는 배정방법에 대하여 일차적으로 공부하고 몇몇 다른 배정 방법들에 대하여도 공부한다.

- 구체적으로 비용함수가

$$C = c_0 + \sum_{h=1}^H c_h n_h = c_0 + n \sum_{h=1}^H c_h w_h$$

로 주어져 있을 때 비용조건을 만족시키고 $\text{var}(\bar{Y}_{\text{st}})$ 이 최소가 되도록 표본배정을 한다.

- 이 때, 풀어야 하는 최적화 문제는 (set $c_0 = 0$)

$$\begin{aligned} \text{minimize} \quad & \text{var}(\bar{Y}_{\text{st}}) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \cdot \frac{N_h - n_h}{N_h} \cdot \frac{\sigma_h^2}{n_h} \\ \text{subject to} \quad & C - \sum_{h=1}^H c_h n_h = 0. \end{aligned}$$

$$\begin{aligned} \text{minimize} \quad & \text{var}(\bar{Y}_{\text{st}}) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \cdot \frac{N_h - nw_h}{N_h} \cdot \frac{\sigma_h^2}{nw_h} \\ \text{subject to} \quad & C - n \sum_{h=1}^H c_h w_h = 0. \end{aligned}$$

- 라그랑즈 승수법(Lagrangian multiplier method)을 이용한다.

■ 라그랑즈 승수법

목적: minimize (or, maximize) $f(x_1, \dots, x_n)$

x_1, \dots, x_n

등식 조건: $g(x_1, \dots, x_n) = 0$

해법: $L = f - \lambda g$ 로 놓고, $\frac{\partial L}{\partial x_1} = 0, \dots, \frac{\partial L}{\partial x_n} = 0, g = 0$ 의 $(n+1)$ 개 방정식을 $(x_1, \dots, x_n, \lambda)$ 의 $(n+1)$ 개 미지수에 대하여 푼다.

•

$$\mathcal{L} = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \cdot \frac{N_h - nw_h}{N_h} \cdot \frac{\sigma_h^2}{nw_h} - \lambda \left(C - \sum_{h=1}^H c_h nw_h \right)$$

따라서

$$\frac{\partial \mathcal{L}}{\partial w_h} = -\frac{N_h^2}{N^2} \frac{\sigma_h^2}{nw_h^2} + \lambda n c_h = 0, \quad h = 1, 2, \dots, H, \quad (2)$$

이고

$$w_h = \frac{1}{\sqrt{\lambda}} \frac{N_h}{N} \frac{\sigma_h}{n \sqrt{c_h}} \quad \text{또는} \quad n_h = \frac{1}{\sqrt{\lambda}} \frac{N_h}{N} \frac{\sigma_h}{\sqrt{c_h}}$$

를 얻게 된다.

(P1) 이제 위 식에서 λ 를 구하면 되는데 이는

$$\frac{\partial \mathcal{L}}{\partial \lambda} = C - \sum_{h=1}^H c_h n_h = 0$$

의 방정식으로 부터, 즉 $C = \sum_{h=1}^H c_h n w_h$ 의 비용제약조건으로,

$$\frac{1}{\sqrt{\lambda}} = \frac{C}{\sum_{h=1}^H \frac{N_h}{N} \sqrt{c_h} \sigma_h}$$

이고 따라서

$$w_h = \frac{1}{n} C \cdot \frac{N_h \sigma_h / \sqrt{c_h}}{\sum_{h=1}^H \sqrt{c_h} N_h \sigma_h} \quad (3)$$

또는

$$n_h = nw_h = C \cdot \frac{N_h \sigma_h / \sqrt{c_h}}{\sum_{h=1}^H \sqrt{c_h} N_h \sigma_h} \quad (4)$$

를 얻는다.

- 최적배정에서 추정량의 분산은

$$\text{Var}(\bar{Y}_{\text{st}}) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{N_h - nw_h}{N_h - 1} \frac{\sigma_h^2}{nw_h} \quad (5)$$

의 공식에 (3)의 값을 대입하여 계산한다.

이제 (P1)지점으로 돌아가 교재의 계산을 따라가면 교재에서는:

- 이제 위 식에서 λ 를 구하면 되는데 이는 본디

$$\frac{\partial \mathcal{L}}{\partial \lambda} = C - \sum_{h=1}^H c_h nw_h = 0$$

즉 $C = \sum_{h=1}^H c_h nw_h$ 의 비용제약조건으로 부터 계산하여야 한다.

(P2) 하지만 우리가 먼저 원하는 오차한계를 위한 총 표본크기를 계산하여 놓았으므로

$$\sum_{h=1}^H n_h = n \quad \text{or} \quad \sum_{h=1}^H w_h = 1$$

의 식과 (3)을 이용하여

$$w_h^* = \left(\frac{N_h \sigma_h / \sqrt{c_h}}{\sum_k N_k \sigma_k / \sqrt{c_k}} \right) \quad (6)$$

를 계산한다.

- 마지막으로 위에서 계산된 w_h 를 이용하여 다음 절에서 사용하게 되는 총 표본의 크기 공식을 이용하여

$$n^* = \frac{\left(\sum_{h=1}^H N_h \sigma_h / \sqrt{c_h} \right) \left(\sum_{h=1}^H N_h \sigma_h \sqrt{c_h} \right)}{N^2 D + \sum_{h=1}^H N_h \sigma_h^2}.$$

를 다시 계산한다.

- 추정량의 최소분산은 꼭 계산하여 보자.

- 이렇게 계산한 w_h^* 와 n^* 는 총비용 조건

$$C = c_0 + \sum_{h=1}^H n^* w_h^* c_h$$

를 만족시키지 못한다.

다시 (P2)로 돌아가 (P2)는 위의 결론(표본설계에서 잘알려 공식임)을 얻기 위하여는 어떻게 수정이 되어야 하는가?

(P3) 다음의 문제, 즉 “(분산) \times (비용)” 을 최소로 하는 표본 배정 $\{n_h, h = 1, 2, \dots, H\}$ 을 찾는 문제를 생각한다.

$$\text{minimize } \left\{ \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{\sigma_h^2}{n_h} - \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{\sigma_h^2}{N_h} \right\} \times \left\{ \sum_{h=1}^H c_h n_h \right\}$$

위의 문제는

$$\text{minimize } \left\{ \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{\sigma_h^2}{n_h} \right\} \times \left\{ \sum_{h=1}^H c_h n_h \right\} \quad (7)$$

와 동치이고 또한

$$\text{minimize } \left\{ \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{\sigma_h^2}{w_h} \right\} \times \left\{ \sum_{h=1}^H c_h w_h \right\} \quad (8)$$

와도 동치이다. Cauchy-Schwartz 부등식에 의하여 최소값이

$$w_h \propto N_h \sigma_h / \sqrt{c_h}$$

일 때 성립하는 것을 알 수 있다. 따라서

$$w_h^* = \left(\frac{N_h \sigma_h / \sqrt{c_h}}{\sum_k N_k \sigma_k / \sqrt{c_k}} \right) \quad (9)$$

를 얻게 된다.

- 최적배정하에서 충화 추정량의 분산을 계산하여 본다.

3.4.2 다른 특별한 최적배정

- 네이만배정(Neyman allocation) ($c_h = c$, 층별단위 관측 비용을 전부 같게 놓는 경우의 최적 배정)

$$n_h = \left(\frac{N_h \sigma_h}{\sum_{h=1}^H N_h \sigma_h} \right) n, \quad n = \frac{\left(\sum_{h=1}^H N_h \sigma_h \right)^2}{N^2 D + \sum_{h=1}^H N_h \sigma_h^2}.$$

- 비례배정(proportional allocation) ($n_h \propto N_h$)

$$n_h = \left(\frac{N_h}{N} \right) n, \quad n = \frac{\sum_{h=1}^H N_h \sigma_h^2}{ND + \sum_{h=1}^H N_h \sigma_h^2 / N}.$$

- 동등배정 (equal allocation)

$$n_h = n/H.$$

- 각각의 배정에 대하여 총화 추정량의 분산을 계산하여 보자.

3.4.3 총 표본크기의 결정

앞 절에서 표본의 최적배정을 수행함에 있어 총표본의 크기 n 이 주어졌다고 가정을 하고 최적 배정을 계산하였다. 본 강의록에서 순서를 바꾸어 최적배정을 먼저 설명하고 총표본크기 n 을 설명하고자 하는데 이는 총표본 크기 n 의 결정이 표본배정의 방법에 의존하기 때문인다.

표본배정의 방식이 정하여져 있다고 가정하자, 즉 w_h 들이 주어져 있다고 가정한다. 우리가 이야기한 앞의 최적배정들에 해당하는 총화추정량의 분산은

$$\text{var}(\bar{Y}_{\text{st}}) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \cdot \frac{N_h - n_h}{N_h - 1} \cdot \frac{\sigma_h^2}{n_h} \approx \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \cdot \frac{N_h - n_h}{N_h} \cdot \frac{\sigma_h^2}{n_h}$$

이제 $w_h = n_h/n$ 을 배정률이라고 하고 이를 위의 식에 대입하면

$$\text{var}(\bar{Y}_{\text{st}}) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \cdot \frac{N_h - nw_h}{N_h} \cdot \frac{\sigma_h^2}{nw_h}$$

와 같이 적을 수 있다.

이제 오차한계를 B 로 만드는 n 을 계산하면

$$Z_{\alpha/2} \sqrt{\text{var}(\bar{Y}_{\text{st}})} = B \quad (10)$$

를 채 정돈

$$\text{var}(\bar{Y}_{\text{st}}) = \left(\frac{B}{Z_{\alpha/2}} \right)^2 = D$$

라 하고 n 을 계산하면

$$n = \frac{\sum_{h=1}^H N_h^2 \sigma_h^2 / w_h}{N^2 D + \sum_{h=1}^H N_h \sigma_h^2}$$

이다.

우리가 앞에서 계산한 모든(총비용 제약조건의 최적배정 제외) 최적배정에서 w_h 는 총표본 크기 n 에 의존하지 않는 값이었다. 만일 w_h 가 n 에 의존한다면 (10)에서 꽤 복잡한 방정식을 수치적으로 풀어야 한다.

예제 3.5

최대오차허용한계 = ±2.

총	N_h	σ_h	c_h
남학생	212	10	1
여학생	88	5	2
$N = 300$			

공식 (3.23)의 계산 과정은 다음과 같다.

항목	남학생	여학생	합계
$N_h \sigma_h / \sqrt{c_h}$	2120	311.13	2431.13
$N_h \sigma_h \sqrt{c_h}$	2120	622.25	2742.25
$N_h \sigma_h^2$	21200	2200	23400
w_h	0.872	0.128	1

$$D = 2^2 / 4 = 1 \text{이므로}$$

$$n = \frac{2431.13 (2742.25)}{300^2 (1) + 23400} = 58.8 \approx 59$$

(2) 모합의 추정

모합 τ 를 추정할 때 표본 크기의 결정 공식은 모평균 추정시와 동일하다. 단, 이 경우 모든 공식에서 D 를 다음과 같이 놓기만 하면 된다.

$$D = \left(\frac{B}{Z_{\alpha/2}N} \right)^2.$$

(3) 모비율의 추정

모평균 추정시 사용하는 모든 공식들에서 σ_h^2 대신 $p_h(1 - p_h)$ 를 사용하면 된다.

- 총 표본크기:

$$n = \frac{\sum_{h=1}^H N_h^2 \sigma_h^2 / w_h}{N^2 D + \sum_{h=1}^H N_h \sigma_h^2}, \quad (\sigma_h^2 = p_h(1 - p_h))$$

를 생각하면

$$n = \frac{\sum_{h=1}^H N_h^2 p_h(1 - p_h) / w_h}{N^2 D + \sum_{h=1}^H N_h p_h(1 - p_h)} \quad (D = B^2 / 4).$$

을 얻는다.

- 최적배정의 경우 n_h 와 n 을 구하는 공식은

$$n_h = \left(\frac{N_h \sqrt{p_h(1 - p_h) / c_h}}{\sum_{h=1}^H N_h \sqrt{p_h(1 - p_h) / c_h}} \right) n$$

$$n = \frac{\left(\sum_{h=1}^H N_h \sqrt{p_h(1 - p_h) / c_h} \right) \left(\sum_{h=1}^H N_h \sqrt{p_h(1 - p_h) c_h} \right)}{N^2 D + \sum_{h=1}^H N_h p_h(1 - p_h)}.$$

- 네이만 배정 ($c_h = c$)

$$n_h = \left(\frac{N_h \sqrt{p_h(1 - p_h)}}{\sum_{h=1}^H N_h \sqrt{p_h(1 - p_h)}} \right) n$$

$$n = \frac{\left(\sum_{h=1}^H N_h \sqrt{p_h(1 - p_h)} \right) \left(\sum_{h=1}^H N_h \sqrt{p_h(1 - p_h)} \right)}{N^2 D + \sum_{h=1}^H N_h p_h(1 - p_h)}.$$

- 비례배정 ($n_h \propto N_h$)

$$n_h = \left(\frac{N_h}{N} \right) n, \quad n = \frac{\sum_{h=1}^H N_h p_h (1 - p_h)}{ND + \sum_{h=1}^H N_h p_h (1 - p_h)/N}.$$

3.5 사후총화(post-stratification)

- 총의 모비율 N_h/N 을 있다고 가정한다.
- SRS $(Z_i, Y_i), i = 1, 2, \dots, n$, 을 얻고 층화변수 Z_i 를 기준으로 H 개의 층으로 층화한다.
- 모평균의 추정량은

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

이다.

여기서 \bar{Y}_h 를 계산할 때 표본의 수 n_h 는 다행분포를 따르는 확률 변수이다. 따라서 $\bar{Y}_{\text{post}} \neq \bar{Y}_{\text{st}}$.

- 추정량 \bar{Y}_{post} 의 평균은 μ , 분산은

$$\begin{aligned} \text{var}(\bar{Y}_{\text{post}}) &= \text{var}\left[\underbrace{\text{E}(\bar{Y}_{\text{post}} | \mathbf{Z})}_{=\mu}\right] + \text{E}[\text{var}(\bar{Y}_{\text{post}} | \mathbf{Z})] \\ &= \text{E}_{\mathbf{Z}} \left[\sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{N_h - n_h}{N_h - 1} \frac{\sigma_h^2}{n_h} \right]. \end{aligned}$$

- 따라서 분산의 추정량으로는

$$\widehat{\text{var}}(\bar{Y}_{\text{post}}) \approx \left(1 - \frac{n}{N}\right) \sum_{h=1}^H \frac{N_h}{N} \frac{s_h^2}{n_h}$$

을 제안한다.

- 사후총화 방법은 n_h 들이 충분히 크다면 비례배정에 기반한 층화추정량과 매우 유사한다.

3.6 총화를 위한 이중표집(double sampling, two phase sampling)

- 교재의 notation을 따르지 않고

$$w_h = \frac{N_h}{N} = \text{총 } n \text{의 모집단 비율}$$

$$W_h = \frac{n'_h}{n'} = \text{첫 번째 phase에서 총 } n \text{의 표본 비율.}$$

이라 적는다.

- 모평균 μ 에 대한 추정량은

$$\bar{Y}_{\text{tp}} = \sum_{h=1}^H W_h \bar{Y}_h.$$

추정량의 평균은

$$\begin{aligned} E(\bar{Y}_{\text{tp}}) &= E[E(\bar{Y}_{\text{tp}} | \mathbf{n}')] \\ &= E\left[\sum_{h=1}^H W_h \mu_h\right] = \mu \end{aligned}$$

이고 분산은

$$\begin{aligned} \text{var}(\bar{Y}_{\text{tp}}) &= \text{var}[E(\bar{Y}_{\text{tp}} | \mathbf{n}')] + E[\text{var}(\bar{Y}_{\text{tp}} | \mathbf{n}')] \\ \text{var}_{\mathbf{n}'}[E(\bar{Y}_{\text{tp}} | \mathbf{n}')] &= \dots \\ &= \frac{1}{n'} \sum_{h=1}^H w_h (\mu_h - \mu)^2. \\ E_{\mathbf{n}'}[\text{var}(\bar{Y}_{\text{tp}} | \mathbf{n}')] &= \dots \\ &= E_{\mathbf{n}'} \left[\frac{W_h^2 s_h^2}{n_h} \frac{N_h - n_h}{N_h - 1} \right]. \end{aligned}$$

따라서 분산의 추정량으로는

$$\widehat{\text{var}}(\bar{Y}_{\text{tp}}) \approx \sum_{h=1}^H \left[\frac{W_h^2 s_h^2}{n_h} + \frac{W_h (\bar{Y}_h - \bar{Y}_{\text{tp}})^2}{n'} \right]$$

을 사용한다.

- 모비율의 추정에 있어서는 위의 분산 추정량의 공식이

$$\widehat{\text{var}}(\widehat{p}_{\text{tp}}) \approx \sum_{h=1}^H \left[\frac{W_h^2 \widehat{p}_h (1 - \widehat{p}_h)}{n_h - 1} + \frac{W_h (\widehat{p}_h - \widehat{p}_{\text{tp}})^2}{n'} \right]$$

로만 바뀐다.

3장 부록: Horvitz-Thompson 추정량

1 Inclusion probability

정의: S 는 표본, A 는 추출 단위

(1) 일차표본 포함 확률 (first order inclusion probability)

$$\begin{aligned}\pi_i &= i \text{ 번째 개체가 표본에 포함될 확률} \\ &= P(i \in S) \\ &= \sum_{A: i \in A} P(A)\end{aligned}$$

(2) 이차표본 포함 확률 (second order inclusion probability)

$$\begin{aligned}\pi_{ij} &= 개체 i, j 가 같이 표본에 포함될 확률 \\ &= P(i, j \in S) \\ &= \sum_{A: i, j \in A} \pi_{ij}\end{aligned}$$

(3) 표본수를 n 이라 하면

$$\begin{aligned}\sum_{i=1}^N \pi_i &= n \\ \sum_{i=1}^N \pi_{ij} &= n\pi_j\end{aligned}$$

이다.

Proof. $i = 1, 2, \dots, n$, 에 대해서,

$$Z_i = \begin{cases} 1 & i \in A \\ 0 & i \notin A \end{cases}$$

라 하자. 이때 $\sum_{i=1}^N Z_i = n$, $\mathbb{E}(Z_i) = P(Z_i = 1) = \pi_i$ 므로,
 $\sum_{i=1}^N \pi_i = n$ 을 얻는다.

비슷하게, $\sum_{i=1}^N Z_i Z_j = nZ_j$, $\mathbb{E}(Z_i Z_j) = P(Z_i = 1, Z_j = 1) = \pi_{ij}$ 므로,
 $\sum_{i=1}^N \pi_{ij} = n\pi_j$ 을 얻는다. \square

(예제 A.1) 모집단 $U = \{1, 2, 3\}$ 에서 다음과 같은 표본 추출법을 생각하여 보자.

$$P(A) = \begin{cases} 0.5 & \text{if } A = \{1, 2\} \\ 0.25 & \text{if } A = \{1, 3\} \\ 0.25 & \text{if } A = \{2, 3\} \end{cases}$$

여기서 $\pi_1 = 0.5 + 0.25 = 0.75$, $\pi_2 = 0.5 + 0.25 = 0.75$, $\pi_3 = 0.25 + 0.25 = 0.5$ 이다.

2 Horvitz-Thompson (HT) 추정량

총계

$$\tau = \sum_{i=1}^N y_i$$

의 추정량으로

$$\hat{\tau}_{HT} = \sum_{i \in S} \frac{1}{\pi_i} y_i$$

을 사용한다. $\hat{\tau}_{HT}$ 는 τ 의 불편추정량 이고, IPW(inverse probability weight) 추정량이라고도 불리운다.

Theorem 1.

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

Proof.

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

임을 기억한다.

$$\begin{aligned}\mathbb{E}(\hat{\tau}_{HT}) &= \sum_{i=1}^N \frac{y_i}{\pi_i} \mathbb{E}(Z_i) \\ &= \sum_{i=1}^N \frac{y_i}{\pi_i} \pi_i = \sum_{i=1}^N y_i = \tau\end{aligned}$$

이다.

비슷하게, 아래의 식을 얻는다.

$$\begin{aligned}
 \text{Var}(\hat{\tau}_{HT}) &= \text{Var}\left(\sum_{i=1}^N \frac{y_i}{\pi_i} Z_i\right) \\
 &= \sum_{i=1}^N \sum_{j=1}^N \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \text{Cov}(Z_i, Z_j) \\
 &= \sum_{i=1}^N \sum_{j=1}^N \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} (\pi_{ij} - \pi_i \pi_j).
 \end{aligned}$$

이제 분산공식의 두 번째 식을 보이자. 표기의 편의성을 위해 $\Delta_{ij} = \pi_i \pi_j - \pi_{ij}$ 라고 하자. 이 때 정의에 의해서 $\Delta_{ij} = \Delta_{ji}$ 임을 염두해 두자. 그러면

$$-\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (\Delta_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 = -\sum_{i=1}^N \sum_{j=1}^N \Delta_{ij} \left(\frac{y_i}{\pi_i} \right)^2 + \sum_{i=1}^N \sum_{j=1}^N \Delta_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

이고

$$\sum_{i=1}^N \sum_{j=1}^N \Delta_{ij} \left(\frac{y_i}{\pi_i} \right)^2 = \sum_{i=1}^N \left(\frac{y_i}{\pi_i} \right)^2 \sum_{j=1}^N \Delta_{ij} = \sum_{i=1}^N \left(\frac{y_i}{\pi_i} \right)^2 (n\pi_i - n\pi_i) = 0 \quad (1)$$

이다. 따라서 두 번째 식을 얻는다. \square

마지막으로 분산 $\text{Var}(\hat{\tau}_{HT})$ 의 추정량으로서,

$$\widehat{\text{Var}}(\hat{\tau}_{HT}) = -\frac{1}{2} \sum_{i \in S} \sum_{j \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \quad (\text{SYG}) \quad (2)$$

를 사용한다. 이때,

$$\mathbb{E}(\widehat{\text{Var}}(\hat{\tau}_{HT})) = \text{Var}(\hat{\tau}_{HT})$$

임을 아래와 같이 확인할 수 있다.

$$\begin{aligned}
\mathbb{E}(\widehat{\text{Var}}(\hat{\tau}_{HT})) &= \mathbb{E}\left(-\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 Z_i Z_j\right) \\
&= -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \mathbb{E}(Z_i Z_j) \\
&= -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \pi_{ij} \\
&= -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2
\end{aligned}$$

(예제 A.1) 의 연속: 모집단 $U = \{1, 2, 3\}$ 에서 다음과 같은 표본 추출법을 생각하여 보자.

$$P(A) = \begin{cases} 0.5 & \text{if } A = \{1, 2\} \\ 0.25 & \text{if } A = \{1, 3\} \\ 0.25 & \text{if } A = \{2, 3\} \end{cases}$$

앞에서 살펴본 바와 같이 여기서

$$\pi_1 = 0.5 + 0.25 = 0.75, \pi_2 = 0.5 + 0.25 = 0.75, \pi_3 = 0.25 + 0.25 = 0.5$$

이고,

$$\hat{\tau}_{HT} = \begin{cases} \frac{y_1}{0.75} + \frac{y_2}{0.75} & \text{if } S = \{1, 2\} \\ \frac{y_1}{0.75} + \frac{y_3}{0.5} & \text{if } S = \{1, 3\} \\ \frac{y_2}{0.75} + \frac{y_3}{0.5} & \text{if } S = \{2, 3\} \end{cases}$$

이다.

따라서

$$\begin{aligned}\mathbb{E}(\hat{\tau}_{HT}) &= 0.5 \left(\frac{y_1}{0.75} + \frac{y_2}{0.75} \right) + 0.25 \left(\frac{y_1}{0.75} + \frac{y_3}{0.5} \right) + 0.25 \left(\frac{y_2}{0.75} + \frac{y_3}{0.5} \right) \\ &= y_1 + y_2 + y_3.\end{aligned}$$

3 모평균 μ 의 추정

모평균 $\mu = \frac{1}{N}(y_1 + y_2 + \dots + y_N)$ 의 추정량으로

$$\hat{\mu}_{HT} = \frac{1}{N} \hat{\tau}_{HT}$$

를 사용한다.

- Markov 부등식에 의해서,

$$P(|\bar{Y}_{HT} - \mu| > \varepsilon) \leq \frac{1}{\varepsilon^2} \text{Var}(\bar{Y}_{HT})$$

가 성립한다.

- 다음의 중심극한정리도 성립한다.

$$\frac{\bar{Y}_{HT} - \mu}{\sqrt{\text{Var}(\bar{Y}_{HT})}} \xrightarrow{d} N(0, 1) \text{ as } n \rightarrow \infty$$

4 비복원 단순 임의 추출 (SRS)

- Inclusion probability:

$$P(A) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{if } |A| = n \text{ for all } A \\ 0 & \text{otherwise} \end{cases}$$

따라서

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

- 모집단 총계에 관한 HT 추정량 :

HT 추정량은

$$\hat{\tau}_{HT} = \frac{N}{n} \sum_{i \in S} y_i = N\bar{Y}$$

이고, 여기서 \bar{Y} 는 표본평균이다.

- HT 추정량의 평균과 분산

$$\begin{aligned} \mathbb{E}(\hat{\tau}_{HT}) &= \tau \\ \text{Var}(\hat{\tau}_{HT}) &= -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \\ &= \frac{1}{2} \frac{N}{n} \frac{N-n}{N(N-1)} \sum_{i=1}^N \sum_{j=1}^N (y_i - y_j)^2 \end{aligned} \tag{3}$$

마지막 등식은 $\pi_i = \frac{n}{N}$, $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$ ($i \neq j$ 일 때)를 대입하면 쉽게 확인할 수 있다.

- 식 (3)에 $\mu = \frac{1}{N}(y_1 + \dots + y_N)$ 에 대해서

$$\begin{aligned}\sum_{i=1}^N \sum_{j=1}^N (y_i - y_j)^2 &= 2 \sum_{i=1}^N \sum_{j=1}^N y_i^2 - 2 \sum_{i=1}^N \sum_{j=1}^N y_i y_j \\ &= 2N \left(\sum_{i=1}^N y_i^2 - N\mu^2 \right) = 2N \sum_{i=1}^N (y_i - \mu)^2\end{aligned}$$

이므로, $\sum_{i=1}^N \sum_{j=1}^N (y_i - y_j)^2 = 2N \sum_{i=1}^N (y_i - \mu)^2$ 이다. 따라서

$$\text{Var}(\hat{\tau}_{HT}) = \frac{N^2}{n} \frac{N-n}{N-1} \sigma^2$$

이 성립한다. 이때 $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2$ 이다.

- HT 추정량의 분산의 추정량 : (SYG) 공식으로부터

$$\widehat{\text{Var}}(\hat{\tau}_{HT}) = -\frac{1}{2} \sum_{i \in S} \sum_{j \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

이고, $\pi_i = \frac{n}{N}$, $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$ ($i \neq j$ 일 때)를 대입하면

$$\hat{\text{Var}}(\hat{\tau}_{HT}) = \frac{1}{2} \frac{N}{n} \frac{N(N-n)}{n(n-1)} \sum_{i \in S} \sum_{j \in S} (y_i - y_j)^2$$

을 얻는다.

이때 sample mean \bar{Y} 에 대해서

$$\sum_{i \in S} \sum_{j \in S} (y_i - y_j)^2 = 2n \sum_{i \in S} (y_i - \bar{Y})^2$$

을 대입하면

$$\widehat{\text{Var}}(\hat{\tau}_{HT}) = N^2 \frac{1}{n} \frac{N-n}{N} s^2$$

을 얻는다. 이 때, $s^2 = \frac{1}{n-1} \sum_{i \in S} (y_i - \bar{Y})^2$ 이다.

5 층화 추출(층화 단순 임의 추출)

- 모집단 $U = U_1 \cup U_2 \cup \dots \cup U_H$,
이 때 $U_i \cap U_j = \emptyset$ for $i \neq j$.
각 층 U_h 에서 층별표본 S_h 를 비복원 단순임의 추출한다.
각 $i = 1, 2, \dots, H$ 에 대해서 $U_i = \{y_{i1}, \dots, y_{iN_i}\}$ 이다.
- 총계 τ 의 HT 추정량은 각 층별 총계 τ_h 의 HT 추정량의 합으로, 다음과
같이 표현된다.

$$\begin{aligned}\hat{\tau}_{HT} &= \sum_{h=1}^H \hat{\tau}_h^{HT} \\ &= \sum_{h=1}^H \frac{N_h}{n_h} \sum_{i \in S_h} y_{hi} = \sum_{h=1}^H N_h \bar{Y}_h\end{aligned}$$

이때 \bar{Y}_h 는 층별표본 S_h 의 산술평균이다.

- τ 의 HT 추정량의 평균과 분산, 그리고 τ 의 HT 추정량의 분산의 추정

량은 다음과 같다.

$$\begin{aligned}\mathbb{E}(\hat{\tau}^{HT}) &= \sum_{h=1}^H N_h \mu_h \\ \text{Var}(\hat{\tau}^{HT}) &= \sum_{h=1}^H N_h^2 \text{Var}(\bar{Y}_h) = \sum_{h=1}^H N_h^2 \frac{N_h - n_h}{N_h - 1} \frac{\sigma_h^2}{n_h} \\ \hat{\text{Var}}(\hat{\tau}^{HT}) &= \sum_{h=1}^H N_h^2 \frac{1}{n_h} \frac{N_h - n_h}{N_h} s_h^2\end{aligned}$$

여기서 $\mu_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi}$, $\sigma_h^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} (y_{hi} - \mu_h)^2$, $s_h^2 = \frac{1}{n_h-1} \sum_{i \in A_h} (y_{hi} - \bar{Y}_h)^2$ 이다.

3장 강의 3 P 16 ... 214

$$\begin{aligned}
 ① &= \text{var}_{\underline{n}'} \left\{ E \left[\sum_{h=1}^H W_h \bar{Y}_h \mid \underline{n}' \right] \right\} \\
 &= \text{var}_{\underline{n}'} \left\{ \sum_{h=1}^H W_h \cdot \mu_h \right\} \\
 &= \sum_{h=1}^H \mu_h^2 \cdot \text{var}(W_h) + \sum_{h_1 \neq h_2} \mu_{h_1} \mu_{h_2} \text{cov}(W_{h_1}, W_{h_2}) \\
 &= \sum_{h=1}^H \mu_h^2 \frac{w_h(1-w_h)}{n'} - \sum_{h_1 \neq h_2} \mu_{h_1} \mu_{h_2} \frac{w_{h_1} w_{h_2}}{n'} \\
 &= \sum_{h=1}^H \mu_h^2 \frac{w_h(1-w_h)}{n'} - \frac{1}{n'} \left\{ \sum_{h=1}^H w_h \mu_h \right\}^2 + \frac{1}{n'} \sum_{h=1}^H \mu_h^2 w_h^2 \\
 &= \sum_{h=1}^H \mu_h^2 \frac{w_h}{n'} - \frac{1}{n'} \left(\sum_{h=1}^H \mu_h w_h \right)^2 \\
 &= \frac{1}{n'} \sum_{h=1}^H w_h \left\{ \mu_h - \frac{1}{n'} \sum_{h=1}^H \mu_h w_h \right\}^2
 \end{aligned}$$

$$② = E_{\underline{n}'} \left\{ \text{var} \left[\sum_{h=1}^H W_h \bar{Y}_h \mid \underline{n}' \right] \right\}$$

$$\begin{aligned}
 &= E_{\underline{n}'} \left\{ \sum_{h=1}^H W_h^2 \cdot \text{var}(\bar{Y}_h \mid \underline{n}') \right\} \\
 &= E_{\underline{n}'} \left\{ \sum_{h=1}^H W_h^2 \frac{\sigma_{e_h}^2}{n'_h} \frac{N_h - n'_h}{N_h - 1} \right\}
 \end{aligned}$$

$$W_h = \frac{n'_h}{n}$$

$$w_h = \frac{N_h}{N}$$