2. 단순임의표집(SRS, simple random sampling)

유한모집단에서 비복원추출

선택될 확률 : $1/{}_N C_n$

2.2 모수의 추정

- 모집단 크기 : N ➡ n
- 모합 : $\tau = \sum_{i=1}^{N} Y_i \Rightarrow t = \sum_{i=1}^{n} y_i$
- 모평균 : $\mu = \sum_{i=1}^{N} Y_i/N \Rightarrow \hat{\mu} = \bar{y} = t/n = \sum_{i=1}^{n} y_i/n$
- 모비율 : $p = \tau/N \Rightarrow \hat{p} = t/n$
- 모분산 : $\sigma^2 = \sum (Y_i - \mu)^2/N \ (= p(1-p)) \Rightarrow \widehat{\sigma^2} = s^2 = \sum_{n} (y_i - \bar{y})^2/(n-1) \ (= n\hat{p}(1-\hat{p})/(n-1))$
- 모변동계수 : $\gamma = \sigma/\mu \ (= \sqrt{(1-p)/p}\ ; \mu, p \neq 0) \Rightarrow \hat{\gamma} = s/\bar{y} \ (= \sqrt{n(1-\hat{p})/(n-1)\hat{p}}\ ; \bar{y}, \hat{p} \neq 0)$

예: 콜레스테롤 수준과 혈압

항목	평균	표준편차	단위	변동계수
수축혈압 수준	130	15	mmHg	0.115
콜레스테롤 수준	200	40	mg/100mL	0.200
이완혈압 수준	60	8	mmHg	0.133

(1) 모평균 μ에 대한 추정

- $E(\overline{Y}) = \mu$
- $Var(\overline{Y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$ ➡ $\widehat{Var}(\overline{Y}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right)$ $\left(E(S^2) = \frac{N}{N-1}\sigma^2\right)$
- $se(\overline{Y}) = \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} \Rightarrow \widehat{se}(\overline{Y}) = \sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)}$
- $\frac{N-n}{N-1}, \frac{N-n}{N}$: 유한모집단수정(fpc : finite population correction)
- 95% 오차한계 : $2\widehat{se}(\overline{Y}) = 2\sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)}$
- μ에 대한 95% 신뢰구간 : $\bar{y} \pm 1.96\sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)}$

(2) 모합 $\tau = N\mu$에 대한 추정

- $Var(\hat{\tau}) = Var(N\overline{Y}) = N^2 Var(\overline{Y}) = N^2 \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$ ➡ $\widehat{Var}(\tau) = N^2 \frac{s^2}{n}\left(\frac{N-n}{N}\right)$
- $se(\hat{\tau}) = \sqrt{N^2 \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} \Rightarrow \widehat{se}(\hat{\tau}) = \sqrt{N^2 \frac{s^2}{n}\left(\frac{N-n}{N}\right)}$
- 95% 오차한계 : $2\widehat{se}(\hat{\tau}) = 2\sqrt{N^2 \frac{s^2}{n}\left(\frac{N-n}{N}\right)}$
- τ에 대한 95% 신뢰구간 : $\hat{\tau} \pm 1.96\sqrt{N^2 \frac{s^2}{n}\left(\frac{N-n}{N}\right)}$

(3) 모비율 p에 대한 추정

- $E(\hat{p}) = p$
- $Var(\hat{p}) = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$ ➡ $\widehat{Var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N}\right)$
- $se(\hat{p}) = \sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}$ ➡ $\widehat{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N}\right)}$
- 95% 오차한계 : $2\widehat{se}(\hat{p}) = 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N}\right)}$
- p에 대한 95% 신뢰구간 : $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N}\right)}$

(4) 추정량의 변동계수

$cv(\hat{\theta}) = \frac{se(\hat{\theta})}{\theta}$ ➡ $\widehat{cv}(\hat{\theta}) = \frac{\widehat{se}(\hat{\theta})}{\hat{\theta}}$

$cv^2(\hat{\theta})$: 상대분산(relative variance, rel-variance)

① 모평균 μ

$$cv(\overline{Y}) = \frac{se(\overline{Y})}{\mu} = \frac{\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}}{\mu} = \frac{\gamma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \Rightarrow$$

$$\widehat{cv}(\overline{Y}) = \frac{\widehat{se}(\overline{Y})}{\overline{y}} = \frac{\sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)}}{\overline{y}} = \frac{\hat{\gamma}}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$$

$\gamma = \sigma/\mu$, $\hat{\gamma} = s/\overline{y}$

② 모비율 p

$$cv(\hat{p}) = \frac{se(\hat{p})}{p} = \frac{\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}}{p} = \frac{\gamma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \Rightarrow$$

$$\widehat{cv}(\hat{p}) = \frac{\widehat{se}(\hat{p})}{\hat{p}} = \frac{\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N}\right)}}{\hat{p}} = \frac{\hat{\gamma}}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$$

$\gamma = \sqrt{\frac{1-p}{p}}$, $\hat{\gamma} = \sqrt{\frac{n}{n-1}\frac{1-\hat{p}}{\hat{p}}}$

2.4 표본크기의 결정

$p(|\hat{\theta}-\theta| \leqq B) = 1-\alpha,\ \hat{\theta} \pm z_{\alpha/2} \quad se(\widehat{\theta}) \qquad B$: 오차한계

$p(|\frac{\hat{\theta}-\theta}{\theta}| \leqq B) = 1-\alpha$(상대 정밀도)

(1) 평균추정

$$B = z_{\alpha/2} se(\overline{y}) = z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}} \approx z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}\frac{N-n}{N}}$$

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

$$n = \frac{z_{\alpha/2}^2\sigma^2}{(B\mu)^2 + \frac{z_{\alpha/2}^2\sigma^2}{N}} = \frac{z_{\alpha/2}^2\gamma^2}{B^2 + \frac{z_{\alpha/2}^2\gamma^2}{N}},\ (\text{ 상대정밀도, } \sigma \approx range/4)$$

(2) 모합 τ추정

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

(3) 모비율 p추정

$$n = \frac{z_{\alpha/2}^2 \quad p(1-p)}{B^2 + \frac{z_{\alpha/2}^2 \quad p(1-p)}{N}} \leqq \frac{z_{\alpha/2}^2/4}{B^2 + \frac{z_{\alpha/2}^2/4}{N}} = \frac{n_0}{1+\frac{n_0}{N}},\ n_0 = z_{\alpha/2}^2/4B^2$$

예 2.2

시험 중 부정행위를 저지른 학생들의 비율을 알고자 설문 조사를 계획한다.

95% 신뢰도에서 최대 3%의 오차한계를 유지

통계학부 N=1251, n=?

1) 모비율 추정

$$n_0 = z_{\alpha/2}^2/4B^2 = \frac{1.96^2}{4\times 0.03^2} \approx 1067$$

$$n = \frac{n_0}{1+\frac{n_0}{N}} = \frac{1067}{1+\frac{1067}{1251}} = 576$$

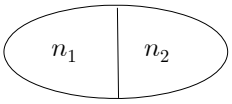
2) 모평균 추정(평균수면 시간), 95% 신뢰도에서 오차한계 0.25, 수면시간은 최소 4시간-12시간

$$\sigma \approx range/4 = 8/4 = 2$$

$$n_0 = z_{\alpha/2}^2\sigma^2/B^2 = \frac{1.96^2\times 2^2}{0.25^2} = 256$$

$$n = \frac{n_0}{1+\frac{n_0}{N}} = \frac{256}{1+\frac{256}{1251}} = 216$$

2.5 부 모집단(subpopulation)



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

(1) 부 모집단 평균 μ_1추정

- $\overline{y_1} = \sum_{j=1}^{n_1} y_{1j}/n_1$
- $\widehat{Var}(\overline{y_1}) = \frac{s_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1}\right), \quad s_1^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \overline{y_1})^2}{n_1 - 1}$
- N_1를 모르는 경우 : $\widehat{Var}(\overline{y_1}) = \frac{s_1^2}{n_1}\left(\frac{N-n}{N}\right)$(가정 $N : N_1 = n : n_1$)

(2) 부 모집단 합 τ_1추정

- $\hat{\tau_1} = N_1\overline{y_1} = \frac{N_1}{n_1}\sum_{j=1}^{n_1} y_{1j}$
- $\widehat{Var}(\hat{\tau_1}) = N_1^2\frac{s_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1}\right)$
- N_1를 모르는 경우 : $\hat{\tau_1} = \frac{N}{n}\sum_{j=1}^{n_1} y_{1j}$, $\widehat{Var}(\hat{\tau_1}) = N^2\left(\frac{N-n}{N}\right)\frac{s_u^2}{n}$,

$$s_u^2 = \frac{\sum_{j=1}^{n_1} y_{1j}^2 - \left[\left(\sum_{j=1}^{n_1} y_{1j}\right)^2/n\right]}{n-1} = \frac{(n_1-1)s_1^2 + n_1(1-n_1/n)\overline{y_1^2}}{n-1}$$

(3) 부 모집단 비율 p_1추정

	핸드폰 있음	핸드폰 없음	합계
흡연	n_{11}	n_{12}	
비흡연	n_{21}	n_{22}	
	n_1	n_2	n

- $\hat{p_1} = n_{11}/n_1$
- $\widehat{Var}(\hat{p_1}) = \frac{\hat{p_1}(1-\hat{p_1})}{n_1 - 1}\left(\frac{N_1 - n_1}{N_1}\right)$ • N_1를 모르는 경우 : $\widehat{Var}(\hat{p_1}) = \frac{\hat{p_1}(1-\hat{p_1})}{n_1 - 1}\left(\frac{N-n}{N}\right)$

예 2.3 전체 학생 N=3000, n=300명을 임의표집하였다.
핸드폰을 갖고 있는 전체학생들이 한 달에 사용하는 용돈의 총액은 얼마인가 ?

$n_1 = 40$(핸드폰을 갖고 있는 학생), $\overline{y_1} = 60$, $s_1 = 20$

$$\hat{\tau}_1 = \frac{N}{n}\sum_{j=1}^{n_1} y_{1j} = \frac{3000}{300}(40 \times 60) = 24{,}000$$

$$s_u^2 = \frac{\sum_{j=1}^{n_1} y_{1j}^2 - \left[\left(\sum_{j=1}^{n_1} y_{1j}\right)^2 / n\right]}{n-1} = \frac{(n_1-1)s_1^2 + n_1(1-n_1/n)\overline{y_1^2}}{n-1} = 469.565$$

$$\widehat{Var}(\hat{\tau}_1) = N^2\left(\frac{N-n}{N}\right)\frac{s_u^2}{n} = 12{,}678{,}255$$

용돈의 총액에 대한 95%신뢰구간 $(24{,}000 \pm 2\sqrt{12{,}678{,}255} = (16{,}878.7,\ 31{,}121.3)$

2.6 평균차이에 대한 검정

$Var(\widehat{\theta_1}-\widehat{\theta_2}) = Var(\widehat{\theta_1}) + Var(\widehat{\theta_2}) - 2Cov(\widehat{\theta_1},\widehat{\theta_2})$

(1) 모 평균 차이에 대한 95%신뢰구간 ; $\overline{y_1}-\overline{y_2} \pm 2\sqrt{\widehat{Var}(\overline{y_1}) + \widehat{Var}(\overline{y_2})}$

(2) 모 비율 차이에 대한 95%신뢰구간 ; $\widehat{p_1}-\widehat{p_2} \pm 2\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$

(3) 다항표집에서 모 비율 차이에 대한 95%신뢰구간 ; $\widehat{p_1}-\widehat{p_2} \pm 2\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n} + \frac{2\widehat{p_1}\widehat{p_2}}{n}}$

예 2.4 N=300, n=90

1) 남녀 간 주 당 평균 컴퓨터 사용 시간에 차이가 있는가?

남자 : $N_1 = 200, n_1 = 50, \quad \overline{y_1} = 14, s_1 = 5$

여자 : $N_2 = 100, n_2 = 40, \quad \overline{y_2} = 10, s_2 = 4$

$$\widehat{Var}(\overline{y_1}) = \frac{s_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1}\right) = 0.375, \quad \widehat{Var}(\overline{y_2}) = \frac{s_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2}\right) = 0.240$$

$$\overline{y_1} - \overline{y_2} \pm 2\sqrt{\widehat{Var}(\overline{y_1}) + \widehat{Var}(\overline{y_2})} = (2.43, \quad 5.57)$$

2) 핸드폰 사용금지 비율에 남녀간 차이가 있을까?

핸드폰 사용	남	여	합계
금지	30(60%)	20(50%)	
제한적 허용	15(30%)	8(20%)	
무조건 허용	5(10%)	12(30%)	
합계	n_1=50	n_2=40	n

$$\hat{p_1} - \hat{p_2} \pm 2\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} = (-0.11, \ 0.31)$$

3) 여자로서 제한적 허용과 무조건 허용을 선택한 학생들 간 비율 차이가 있는가?

$$\hat{p_1} - \hat{p_2} \pm 2\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n} + \frac{\hat{p_2}(1-\hat{p_2})}{n} + \frac{2\hat{p_1}\hat{p_2}}{n}} = (-0.32, \ 0.12)$$