HW 1

2016314895, 정희철, 통계학과

1) $Y_{11}, \dots, Y_{1n} \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$, $Y_{21}, \dots, Y_{2n} \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$ 이고, 2-sample case with known variances 일 때, $n = \dfrac{(Z_{\frac{\alpha}{2}}+Z_\beta)^2(\sigma_1^2+\sigma_2^2)}{(\Delta-\Delta_0)^2}$ 임을 보여라.

pf) i) $\beta = P(\text{Type II Error}) = P(|Z_0| < Z_{\frac{\alpha}{2}} \mid H_1)$

$= P(-Z_{\frac{\alpha}{2}} < Z_0 < Z_{\frac{\alpha}{2}} \mid H_1)$, 이 때 $Z_0 \sim N\left(\dfrac{\sqrt{n}(\Delta-\Delta_0)}{\sqrt{Var(Y_1+Y_2)}}\right)$을 따르는데, 여기서 $\Delta - \Delta_0 = \mu_1 - \mu_2 - (\bar{y}_1 - \bar{y}_2)$ 이고,

$var(Y_1+Y_2) = \sigma_1^2 + \sigma_2^2 + cov(Y_1, Y_2)$ 이지만, $Y_1 \perp\!\!\!\perp Y_2$ 이므로 $cov(Y_1, Y_2) = 0$ 이 되어 $\sigma_1^2 + \sigma_2^2$ 로 정리된다.

$\Rightarrow Z_0 \sim N\left(\dfrac{\sqrt{n}(\Delta-\Delta_0)}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)$

$= P\left(-Z_{\frac{\alpha}{2}} - \dfrac{\sqrt{n}(\Delta-\Delta_0)}{\sqrt{\sigma_1^2+\sigma_2^2}} < Z_0 - \dfrac{\sqrt{n}(\Delta-\Delta_0)}{\sqrt{\sigma_1^2+\sigma_2^2}} < Z_{\frac{\alpha}{2}} - \dfrac{\sqrt{n}(\Delta-\Delta_0)}{\sqrt{\sigma_1^2+\sigma_2^2}} \,\middle|\, H_1\right)$, 각 변수들을 표준화시켜 $N(0,1)$ 분포를 따르게 하면,

$= \Phi\left(Z_{\frac{\alpha}{2}} - \dfrac{\sqrt{n}(\Delta-\Delta_0)}{\sqrt{\sigma_1^2+\sigma_2^2}}\right) - \Phi\left(-Z_{\frac{\alpha}{2}} - \dfrac{\sqrt{n}(\Delta-\Delta_0)}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)$, 이 때 후자는 0에 근접하므로

$= \Phi\left(Z_{\frac{\alpha}{2}} - \dfrac{\sqrt{n}(\Delta-\Delta_0)}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)$

$= 1 - \Phi\left(-Z_{\frac{\alpha}{2}} + \dfrac{\sqrt{n}(\Delta-\Delta_0)}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)$, 이 때 cdf 가 $1-\beta$ 와 동일하므로 다음 식이 유도된다.

ii) $Z_\beta = -Z_{\frac{\alpha}{2}} + \dfrac{\sqrt{n}(\Delta-\Delta_0)}{\sqrt{\sigma_1^2+\sigma_2^2}}$

$\sim\ Z_\beta + Z_{\frac{\alpha}{2}} = \dfrac{\sqrt{n}(\Delta-\Delta_0)}{\sqrt{\sigma_1^2+\sigma_2^2}}$

$\sim\ \dfrac{(Z_\beta + Z_{\frac{\alpha}{2}})\sqrt{\sigma_1^2+\sigma_2^2}}{\Delta-\Delta_0} = \sqrt{n}$

$\therefore\ n = \dfrac{(Z_\beta + Z_{\frac{\alpha}{2}})^2(\sigma_1^2+\sigma_2^2)}{(\Delta-\Delta_0)^2}$ 임을 확인할 수 있다.

2) a) cell means model : $y_{ij} = \mu_i + \varepsilon_{ij}$, where $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

b) y_{ij} : an individually observed values of different treatments

"range of y_{ij}" $= [max(y_{ij}), min(y_{ij})] = [103, 7]$

μ_i : theoretical mean the observations of the i^{th} treatment

"range of μ_i" $= [max(\mu_i), min(\mu_i)] = [84.8, 40.6]$

ε_{ij} : random errors

"range of ε_{ij}" $= [max(\varepsilon_{ij}), min(\varepsilon_{ij})] = [max(y_{ij}-\mu_i), min(y_{ij}-\mu_i)] = [35.4, -33.6]$

c) Unbiased Estimate of $\sigma = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$

We have $n = 20$, $\bar{x} = 58.5$

$\therefore\ \hat{\sigma} = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i-58.5)^2}{19}} = 550.6842$

d)

	Degree of Freedom	Sum Sq	Mean Sq	F-value	p-value
Treatments	3	5242	1747.3	5.354	0.00957
Residuals	16	5221	326.3		
Total	19	10463	2073.6		

e) $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

Ha : at least one of the treatment means is significantly different from the others.

We have F-value of 5.354, which has the p-value of $0.00957 > \alpha = 0.05$.

$\therefore$ We can reject the null hypothesis ; at least one of the treatment means is significantly different from the others.

3) a) $E(MS_{Treatment}) \stackrel{?}{=} \sigma^2 + \frac{n}{a-1} \sum_{i=1}^{a} \tau_i^2$

$\Rightarrow MS_{Treatment} = \frac{SS_{treatment}}{a-1}$, where $SS_{treatment} = \sum_{i=1}^{a} n(\bar{y}_{i.} - \bar{y}_{..})^2$

$$E(MS_{treatment}) = \frac{1}{a-1} E(SS_{treatment})$$

$$= \frac{1}{a-1} E\left[n \sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2\right]$$

$$= \frac{1}{a-1} E\left[\sum_{i=1}^{a} \frac{y_{i.}^2}{n} - \frac{y_{..}^2}{N}\right]$$, $y_{i.} = \sum_{j=1}^{n} y_{ij} = \sum_{j=1}^{n} \mu + \tau_i + \varepsilon_{ij}$ and $y_{..} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{n} \mu + \tau_i + \varepsilon_{ij}$

$$= \frac{1}{a-1} E\left[\frac{1}{n}\sum_{i=1}^{a}\left(\sum_{j=1}^{n} \mu + \tau_i + \varepsilon_{ij}\right)^2 - \frac{1}{N}\left(\sum_{i=1}^{a}\sum_{j=1}^{n} \mu + \tau_i + \varepsilon_{ij}\right)^2\right]$$

$$= \frac{1}{a-1} E\left[\frac{1}{n}\sum_{i=1}^{a}\left(n^2\mu^2 + n^2\tau_i^2 + \varepsilon_{i.}^2\right) - \frac{1}{N}\left(N^2\mu^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}\tau_i^2 + \varepsilon_{..}^2\right)\right]$$

$$= \frac{1}{a-1} E\left[\cancel{N\mu^2} + n\sum_{i=1}^{a}\tau_i^2 + \frac{1}{n}\sum_{i=1}^{a}\varepsilon_{i.}^2 - \cancel{N\mu^2} - \frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{n}\tau_i^2 - \frac{1}{N}\varepsilon_{..}^2\right]$$, where $E(\tau_i^2) = \sigma_\tau^2$ and $E\left(\sum_{i=1}^{a}\sum_{j=1}^{n}\tau_i^2\right) = an^2$

$$= \frac{1}{a-1}\left[N\sigma_\tau^2 + a\sigma^2 - n\sigma_\tau^2 - \sigma^2\right]$$

$$= \sigma^2 + \frac{n}{a-1}\left[(a-1)\sigma_\tau^2\right]$$, $\sigma_\tau^2 = E(\tau_i^2) = \frac{\sum_{i=1}^{a}\tau_i^2}{a-1}$

$$= \sigma^2 + \frac{n}{a-1}\left[\sum_{i=1}^{a}\tau_i^2\right]$$

$\therefore$ $E(MS_{Treatment}) = \sigma^2 + \frac{n}{a-1}\sum_{i=1}^{a}\tau_i^2$

b) $E(MS_E) \stackrel{?}{=} \sigma^2$

$MS_E = \frac{SS_E}{N-a}$, where $SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - 2\bar{y}_{i.}y_{ij} - \bar{y}_{i.}^2 = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \sum_{i=1}^{a} n\bar{y}_{i.}^2 = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{1}{n}\sum_{i=1}^{a} y_{i.}^2$

$$= \frac{1}{N-a}\left\{\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{1}{n}\sum_{i=1}^{a} y_{i.}^2\right\}$$

$$E(MS_E) = \frac{1}{N-a} E\left\{\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{1}{n}\sum_{i=1}^{a} y_{i.}^2\right\}$$, where it is given that $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ and $y_{i.} = \sum_{j=1}^{n} \mu + \tau_i + \varepsilon_{ij}$

$$= \frac{1}{N-a} E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(\mu + \tau_i + \varepsilon_{ij})^2 - \frac{1}{n}\sum_{i=1}^{a}\left(\sum_{j=1}^{n}\mu + \tau_i + \varepsilon_{ij}\right)^2\right]$$

$$= \frac{1}{N-a} E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(\mu + \tau_i + \varepsilon_{ij})^2 - \frac{1}{n}\sum_{i=1}^{a}\left(\sum_{j=1}^{n}\mu + \tau_i + \varepsilon_{ij}\right)^2\right]$$

$$= \frac{1}{N-a} E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}\mu^2 + \tau_i^2 + \varepsilon_{ij}^2 - \frac{1}{n}\sum_{i=1}^{a} n^2\mu^2 + n^2\tau_i^2 + \varepsilon_{i.}^2\right]$$

$$= \frac{1}{N-a} E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}\mu^2 + \tau_i^2 + \varepsilon_{ij}^2 - N\mu^2 - n\sum_{i=1}^{a}\tau_i^2 - \frac{1}{n}\sum_{i=1}^{a}\varepsilon_{i.}^2\right]$$

$$= \frac{1}{N-a} E\left[N\mu^2 + n\sum_{i=1}^{a}\tau_i^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}\varepsilon_{ij}^2 - N\mu^2 - n\sum_{i=1}^{a}\tau_i^2 - \frac{1}{n}\sum_{i=1}^{a}\varepsilon_{i.}^2\right]$$

$$= \frac{1}{N-a} E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}\varepsilon_{ij}^2 - \frac{1}{n}\sum_{i=1}^{a}\varepsilon_{i.}^2\right]$$

$$= \frac{1}{N-a}\left\{\sum_{i=1}^{a}\sum_{j=1}^{n}E(\varepsilon_{ij}^2) - \frac{1}{n}\sum_{i=1}^{a}E(\varepsilon_{i.}^2)\right\}$$

$$= \frac{1}{N-a}\left\{N\sigma^2 - a\sigma^2\right\}$$

$$= \sigma^2$$

$\therefore$ $E(MS_E) = \sigma^2$

c) $SS_T = SS_{treatment} + SS_E = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$

It is defined that $MS_{treatment} = \frac{SS_{treatment}}{a-1}$ and $MS_E = \frac{SS_E}{a(n-1)}$

$F = \frac{W_1/df_1}{W_2/df_2} \sim F_{df_1, df_2}$, with $W_1 \sim \chi^2_{df_1}$, $W_2 \sim \chi^2_{df_2}$

If we define $W_1 = \frac{SS_{treatment}}{\sigma^2}$ and $W_2 = \frac{SS_E}{\sigma^2}$. Then it is known that under $\tau_i = 0$ for all i's, $\frac{SS_{treatment}}{\sigma^2} \sim \chi^2_{a-1}$, and $\frac{SS_E}{\sigma^2} \sim \chi^2_{N-a}$

$\therefore$ $F = \frac{\left(\frac{SS_{treatment}}{\sigma^2}\right)/(a-1)}{\left(\frac{SS_E}{\sigma^2}\right)/(N-a)} = \frac{SS_{treatment}/a-1}{SS_E/N-a} = \frac{MS_{treatment}}{MS_E} \overset{H_0}{\sim} F_{a-1, N-a}$