

* Derive some equation

這兩個推導用手寫的

Design of experiment 106354003
Homework 02 (3/13) 林健宏

Question 01.
Derive $\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^g n_i (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$

<pf> $\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..})^2$
 $= \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + 2 \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.}) (\bar{y}_{i.} - \bar{y}_{..}) + \sum_{i=1}^g \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2$
 (其中 $\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.}) (\bar{y}_{i.} - \bar{y}_{..}) = \sum_{i=1}^g (\bar{y}_{i.} - \bar{y}_{..}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})$)
 $\sum_{j=1}^{n_i} y_{ij} - \bar{y}_{i.} = \sum_{j=1}^{n_i} y_{ij} - n_i \bar{y}_{i.} = 0$
 $\therefore \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.}) (\bar{y}_{i.} - \bar{y}_{..}) = 0$

$\Rightarrow \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^g \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2$
 $= \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^g n_i (\bar{y}_{i.} - \bar{y}_{..})^2$

又 $\bar{y}_{i.} - \bar{y}_{..} = \alpha_i$, $\therefore \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \underbrace{\sum_{i=1}^g n_i \alpha_i^2}_{SS_T} + \underbrace{\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}_{SS_{T-t}} + \underbrace{\sum_{i=1}^g \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2}_{SSE}$

Question 02. (1) 假設非常態下
Derive $E(\hat{\sigma}^2) = \sigma^2$, $\hat{\sigma}^2 = MSE = \frac{SSE}{N-g} = \frac{\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}{N-g}$
 其中 $\frac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}{\sigma^2} \sim \chi^2(n_i-1)$, ($\because y_{ij} = \bar{y}_{i.} + \varepsilon_{ij}$, $\varepsilon_{ij} \sim N(0, \sigma^2)$)
 \downarrow
 $\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 \sim \sum_{i=1}^g \chi^2(n_i-1) = \chi^2(\sum_{i=1}^g n_i - g) = \chi^2(N-g)$
 $\therefore E\left(\frac{\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}{\sigma^2}\right) = N-g$
 $\therefore E\left(\frac{\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}{N-g}\right) = \sigma^2$, 又 $\frac{\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}{N-g} = MSE$
 $\therefore E(MSE) = \sigma^2 \Rightarrow MSE \text{ is unbiased to } \sigma^2$

$$\frac{\sum_{j=1}^g \sum_{i=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}{n-1}$$

(2) 假設不為常態下.

$$\begin{aligned} E(MSE) &= E\left(\frac{SSE}{N-g}\right) = \frac{1}{N-g} E\left(\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2\right) \\ &= \frac{1}{N-g} E\left(\sum_{i=1}^g \sum_{j=1}^{n_i} y_{ij}^2 - 2 \sum_{i=1}^g \sum_{j=1}^{n_i} \bar{y}_{i.} y_{ij} + \sum_{i=1}^g \sum_{j=1}^{n_i} \bar{y}_{i.}^2\right) \\ &= \frac{1}{N-g} E\left(\sum_{i=1}^g \sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^g n_i \bar{y}_{i.}^2\right) \\ &= \frac{1}{N-g} \left(\sum_{i=1}^g \sum_{j=1}^{n_i} E(y_{ij}^2) - \sum_{i=1}^g n_i E(\bar{y}_{i.}^2)\right) \\ &= \frac{1}{N-g} \sum_{i=1}^g \sum_{j=1}^{n_i} (\sigma^2 + \mu_i^2) - \sum_{i=1}^g n_i \left(\frac{\sigma^2}{n_i} + \mu_i^2\right) \\ &= \frac{1}{N-g} (N\sigma^2 + \sum_{i=1}^g n_i \mu_i^2 - g\sigma^2 - \sum_{i=1}^g n_i \mu_i^2) \\ &= \frac{1}{N-g} (N-g)\sigma^2 = \sigma^2 \end{aligned}$$

* Exercise 2.4

As part of a larger experiment, Dale (1992) looked at six samples of a wetland soil undergoing a simulated snowmelt. Three were randomly selected for treatment with a neutral pH snowmelt; the other three got a reduced pH snowmelt. The observed response was the number of Copepoda removed from each microcosm during the first 14 days of snowmelt.

Reduced pH			Neutral pH		
256	159	149	54	123	248

Using randomization methods, test the null hypothesis that the two treatments have equal average numbers of Copepoda versus a two-sided alternative.

H_0 : The two treatments have equal average numbers of Copepoda.

H_1 : The two treatments don't have equal average numbers of Copepoda.

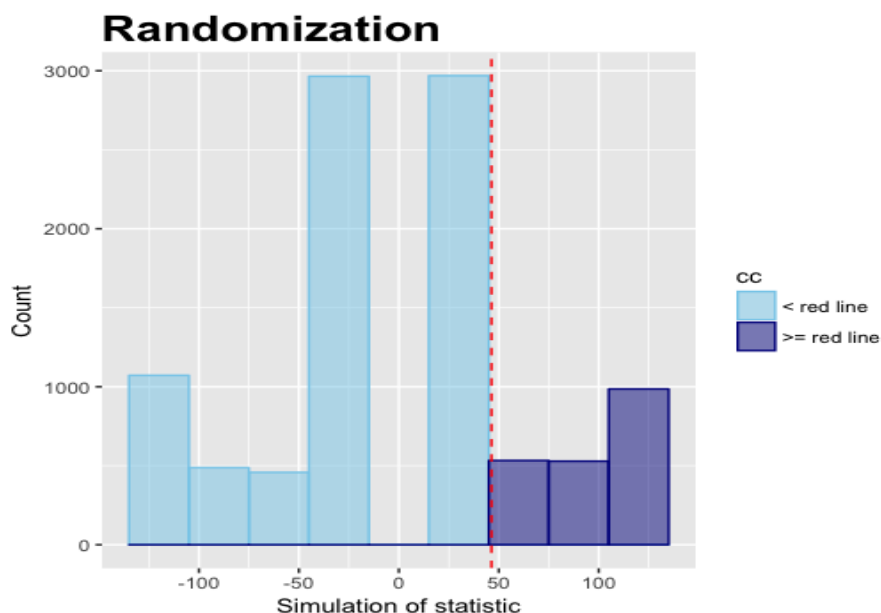
這邊只列出結果及圖(Code附在Appendix)

在 H_0 為真之下，將這六個數字隨機抽取三個為一組加總後減去剩下三個數的加總得到一個差值。將這樣的模式進行10000次的模擬後繪製成直方圖，並計算超過檢定統計量

$k = ((256+159+149)-(54+123+248)) / 3$ 的比例乘2(雙尾檢定)，即為p value。

模擬結果p value為0.4 (將模擬實驗進行100次，收集每次實驗的p value，再取平均值，標準差0.00403)，因為p value = 0.4 > 0.05，不拒絕 H_0 的假設。(紅線為臨界值 46.333，P值為大於紅線值的機率)

Conclusion: The two treatments have equal average numbers of Copepoda.



* Appendix

<http://rpubs.com/YaPi/370646>

```
library(ggplot2)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##   filter, lag

## The following objects are masked from 'package:base':
##   intersect, setdiff, setequal, union

# 實驗設計 Exercise 2.4#
x1<- c(256,159,149)
x2 <- c(54,123,248)
x <- c(x1,x2)
#observation#
ob <- (sum(x1) - sum(x2))/3
# simulation
n = 1; pv = NULL
while(n <= 1000){
  sumd <- NULL
  for(i in 1:10000){
    y <- sample(x,3,F)
    tmp <- sum(x) - sum(y)
    sumd[i] <- (sum(y) - tmp)/3}
  p <- sum(sumd >= ob)/10000
  pv <- c(pv,p)
  n = n + 1}
# 雙尾檢定 P value #
P_value <-mean(pv) * 2; sd(pv)

## [1] 0.003520883

P_value

## [1] 0.3000982

# 繪圖
sumd <- as.data.frame(sumd)
sumd <- mutate(sumd,cc = (sumd >= ob) *1 )
sumd$cc[sumd$cc == 1] <- '>= red line'
sumd$cc[sumd$cc == 0] <- '< red line'
sumd$cc <- sumd$cc %>% as.factor()
tmp <- ggplot(sumd,aes(x = sumd, fill=cc, color = cc))+ geom_histogram(position="identity", alpha = 0.5, binwidth = 30)
tmp+theme(plot.title = element_text(colour = "black", face = "bold", size = 20, vjust = 0.3)) +
  geom_vline(aes(xintercept=ob),linetype="dashed",color="red")+
  scale_color_manual(values=c("skyblue", "darkblue","black"))+
  scale_fill_manual(values=c("skyblue", "darkblue","black"))+
  labs(title="Randomization ",x="Simulation of statistic", y = "Count")
```