STAT 5703 Homework 2 Exercise 5

Shijie He(sh3975), Yunjun Xia(yx2569), Shuyu Huang(sh3967)

3/1/2020

Part 1

Suppose at time $\tau^{(k)}$, there are in total $m_A = n_A^{(k)}$ patients at risk under treatment A and $m_B = n_B^{(k)}$ patients at risk under treatment B. $\{i_1, i_2, ..., i_{m_A}\}$ are the individuals at risk from group A and $\{j_1, j_2, ..., j_{m_B}\}$ are the individuals at risk from group B. Let P_A and P_B be the probability of a single death or relapse of treatment A and B. Under H_0, $P_A = P_B$. Thus, the probability that any m out of group A and any $m_d - m$ out of group B dying/relapsing is:

$$P(y^{(k)} = m, n_A^{(k)} = m_A, n_B^{(k)} = m_B, n_d^{(k)} = m_d)$$
$$= \binom{m_A}{m} P_A{}^m (1 - P_A)^{m_A - m} \binom{m_B}{m_d - m} P_B{}^{m_d - m} (1 - P_B)^{m_B - m_d + m}$$
$$= \binom{m_A}{m} \binom{m_B}{m_d - m} P_A{}^{m_d} (1 - P_A)^{m_A + m_B - m_d}$$

Since $P(n_A^{(k)} = m_A, n_B^{(k)} = m_B, n_d^{(k)} = m_d) = \binom{m_A + m_B}{m_d} P_A{}^{m_d} (1 - P_A)^{m_A + m_B - m_d}$, we would have

$$P(y^{(k)} = m \mid n_A^{(k)} = m_A, n_B^{(k)} = m_B, n_d^{(k)} = m_d)$$
$$= \frac{P(y^{(k)} = m, n_A^{(k)} = m_A, n_B^{(k)} = m_B, n_d^{(k)} = m_d)}{P(n_A^{(k)} = m_A, n_B^{(k)} = m_B, n_d^{(k)} = m_d)}$$
$$= \frac{\binom{m_A}{m} \binom{m_B}{m_d - m}}{\binom{m_A + m_B}{m_d}}$$

Thus, we can conclude that $y^{(k)} \sim Hypergeometric(n_A^{(k)} + n_B^{(k)}, n_A^{(k)}, n_d^{(k)})$.

Part 2

$p(y^{(k)} \mid n_A^{(k)}, n_B^{(k)}, n_d^{(k)}) \sim Hypergeometric(n_A^{(k)} + n_B^{(k)}, n_A^{(k)}, n_d^{(k)})$.

Thus,

$$\mathbb{E}[p(y^{(k)} \mid n_A^{(k)}, n_B^{(k)}, n_d^{(k)})] = \frac{n_A^{(k)} n_d^{(k)}}{n_A^{(k)} + n_B^{(k)}} = \frac{n_A^{(k)} n_d^{(k)}}{n^{(k)}}$$

and

$$Var(p(y^{(k)} \mid n_A^{(k)}, n_B^{(k)}, n_d^{(k)})) = \frac{n_d^{(k)} n_A^{(k)} n_B^{(k)} (n_A^{(k)} + n_B^{(k)} - n_d^{(k)})}{(n_A^{(k)} + n_B^{(k)})^2 ((n_A^{(k)} + n_B^{(k)}) - 1)} = \frac{n_d^{(k)} n_A^{(k)} n_B^{(k)} n_S^{(k)}}{(n^{(k)})^2 (n^{(k)} - 1)}$$

Part 3

$$Var[y^{(k)} - E^{(k)}] = Var[\mathbb{E}[y^{(k)} - E^{(k)} \mid n_A^{(k)}, n_B^{(k)}, n_d^{(k)}]] + \mathbb{E}[Var[y^{(k)} - E^{(k)} \mid n_A^{(k)}, n_B^{(k)}, n_d^{(k)}]]$$

We know:

$$\mathbb{E}[y^{(k)} - E^{(k)} \mid n_A^{(k)}, n_B^{(k)}, n_d^{(k)}] = E^{(k)} - E^{(k)} = 0$$

and

$$Var[y^{(k)} - E^{(k)} \mid n_A^{(k)}, n_B^{(k)}, n_d^{(k)}] = V^{(k)}$$

Thus,

$$Var[y^{(k)} - E^{(k)}] = Var[0] + \mathbb{E}[V^{(k)}] = \mathbb{E}[V^{(k)}]$$

Part 4

For different k, $y^{(k)} - E^{(k)}$ are independent. Thus,

$$Var[\sum_{k=1}^{K}(y^{(k)} - E^{(k)})] = \sum_{k=1}^{K} Var[y^{(k)} - E^{(k)})] = \sum_{k=1}^{K} \mathbb{E}[V^{(k)}]$$

Part 5

```r
library(readr)
library(dplyr)
library(magrittr)
transplant <- read_table2("transplant.txt", col_names = FALSE)[-(1:7),(1:3)]
colnames(transplant) = c("t", "type", "survive")
transplant = as.data.frame(transplant)
transplant %<>% mutate_if(is.character, as.numeric)
transplant <- transplant[order(transplant$t),]

ka_ind <- which(transplant$survive == 1)
y <- c()
e <- c()
v <- c()
for (i in 1:length(ka_ind)){
  if (transplant$type[ka_ind[i]] == 1){
    y <- c(y, 1)
  }else{
    y <- c(y, 0)
  }
  rest <- transplant[ka_ind[i]:nrow(transplant),]
  n_a <- sum(rest$type == 1)
  n_b <- sum(rest$type == 2)
  n <- n_a + n_b
  e <- c(e, n_a/n)
  v <- c(v, n_a*n_b*(n-1)/(n*n*(n-1)))
}
```

```r
z <- sum(y-e)/sqrt(sum(v))
cat("The p-value is :", pnorm(z)*2)
```

```
## The p-value is : 0.5444143
```

We fail to reject the null hypothesis. Thus, we can conclude that the survival function of the two groups are the same.