HW2 Exercise 5

Wen Fan(wf2255), Banruo Xie(bx2168), Hanjun Li(hl3339)

3/7/2020

```
library(readr)
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
```

```
transplant <- read_table2("~/Documents/Columbia/STAT W5703/HW/HW2/transplant.txt",
    col_names = c("time", "type", "Indicator"), skip = 8)
```

```
## Parsed with column specification:
## cols(
##   time = col_double(),
##   type = col_double(),
##   Indicator = col_double()
## )
```

```
# preprocessing
transplant$type <- factor(transplant$type)
type1_index <- which(transplant$type==1)
type2_index <- which(transplant$type==2)
death_index <- which(transplant$Indicator==1)
```

Problem 1

Suppose the relapse rates for both treatment group are the same. i.e. $P_A = P_B = p$. We also see that the total number of death $n_d^{(k)}$ is independent of the number of at risk groups $n_A^{(k)}$ and $n_B^{(k)}$, so

$$\begin{aligned}
P(y^{(k)} = m | n_A^{(k)} = m_A, n_B^{(k)} = m_B, n_d^{(k)} = m_d) &= \frac{P(y^{(k)} = m, n_A^{(k)} = m_A, n_B^{(k)} = m_B, n_d^{(k)} = m_d)}{P(n_A^{(k)} = m_A, n_B^{(k)} = m_B, n_d^{(k)} = m_d)} \\
&= \frac{P(y^{(k)} = m, n_d^{(k)} = m_d | n_A^{(k)} = m_A, n_B^{(k)} = m_B) P(n_A^{(k)} = m_A, n_B^{(k)} = m_B)}{P(n_d^{(k)} = m_d | P(n_A^{(k)} = m_A, n_B^{(k)} = m_B)} \\
&= \frac{\binom{m_A}{m} p^m (1-p)^{m_A - m} \binom{m_B}{m_d - m} p^{m_d - m} (1-p)^{m_B - m_d + m}}{\binom{m_A + m_B}{m_d} p^{m_d} (1-p)^{m_A + m_B - m_d}} \\
&= \frac{\binom{m_A}{m}\binom{m_B}{m_d - m}}{\binom{m_A + m_B}{m_d}}
\end{aligned}$$

Hence, it is a $HyperGeometric(n_A^{(k)} + n_B^{(k)}, n_A^{(k)}, n_d^{(k)})$

Problem 2

From previous problem, we have the conditonal probability

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

Using the formulas for the expected value and variance of a hypergeometric distribution: given $h(x; N, n, k)$,

$$E[X] = \frac{n * k}{N}$$

$$Var(X) = \frac{n * k * (N - k) * (N - n)}{N^2 * (N - 1)}$$

so

$$E^{(k)} = \frac{n_A^{(k)} n_d^{(k)}}{n^{(k)}}$$

$$V^{(k)} = \frac{n_A^{(k)} n_d^{(k)} (n^{(k)} - n_A^{(k)})(n^{(k)} - n_d^{(k)})}{(n^{(k)})^2 (n^{(k)} - 1)} = \frac{n_A^{(k)} n_d^{(k)} n_B^{(k)} n_s^{(k)}}{(n^{(k)})^2 (n^{(k)} - 1)}$$

Hence, shown.

Problem 3

We have

$$\begin{aligned} var(y^{(k)} - E^{(k)}) &= var(E[y^{(k)} - E^{(k)} | n_A^{(k)}, n_B^{(k)}, n_d^{(k)}]) + E[var(y^{(k)} - E^{(k)} | n_A^{(k)}, n_B^{(k)}, n_d^{(k)})] \\ &= var(E[y^{(k)} | n_A^{(k)}, n_B^{(k)}, n_d^{(k)}] - E^{(k)}) + E[V^{(k)}] \\ &= 0 + E[V^{(k)}] \\ &= E[V^{(k)}] \end{aligned}$$

Problem 4

We have

$$var[\sum_{k=1}^{K} (y^k - E^{(k)})] = \sum_{k=1}^{K} var[y^{(k)} - E^{(k)}] + 2 \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} cov[y^{(i)} - E^{(i)}, y^{(j)} - E^{(j)}]$$

Let the condition $(n_A^{(i)}, n_B^{(i)}, n_d^{(i)}, n_A^{(j)}, n_B^{(j)}, n_d^{(j)})$ be C. By the Law of Total Variance,

$$cov[y^{(i)} - E^{(i)}, y^{(j)} - E^{(j)}] = cov[E[y^{(i)} - E^{(i)} | C], E[y^{(j)} - E^{(j)} | C]] + E[cov[y^{(i)} - E^{(i)}, y^{(j)} - E^{(j)} | C]]$$

We have $E[y^{(i)} - E^{(i)} | C] = E[y^{(j)} - E^{(j)} | C] = 0$ and $cov[y^{(i)} - E^{(i)}, y^{(j)} - E^{(j)} | C] = cov[y^{(i)}, y^{(j)} | C]$ from problem 3, and $y^{(i)}$, $y^{(j)}$ are two independent hypergeometric variables; hence,

$$cov[y^{(i)} - E^{(i)}, y^{(j)} - E^{(j)} | C] = cov[y^{(i)}, y^{(j)} | C] = 0$$

. As a result,

$$\begin{aligned} cov[y^{(i)} - E^{(i)}, y^{(j)} - E^{(j)}] &= cov[E[y^{(i)} - E^{(i)}|C], E[y^{(j)} - E^{(j)}|C]] + E[cov[y^{(i)} - E^{(i)}, y^{(j)} - E^{(j)}|C]] \\ &= 0 + 0 = 0 \end{aligned}$$

and

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

Problem 5

In our dataset, for each k, we have 1 death if the data is not censored, so $n_d^k = 1, \forall k$

```
n <- nrow(transplant)
k <- length(death_index)
na <- length(type1_index)
nb <- length(type2_index)
nd <- 1
df <- transplant[order(transplant$time),] %>% filter(Indicator==1)
y = numeric(k)
E = numeric(k)
Var = numeric(k)
for (i in seq_len(k)){
  E[i] <- na*nd/(na+nb)
  Var[i] <- na*nb*nd*(na+nb-nd)/((na+nb)^2*(na+nb-1))
  if(df[i,]$type==1){
    y[i] <- 1
    na <- na - 1
  } else {
    y[i] <- 0
    nb <- nb - 1
  }
}

Z <- sum(y-E)/sqrt(sum(Var))
pnorm(Z)
```

```
## [1] 0.2828261
```

Using the logrank_test from "coin" library, we got the same conclusion:

```
library(coin)
logrank_test(Surv(time, Indicator)~type, data=transplant, distribution = "asymptotic")
```

```
##
##  Asymptotic Two-Sample Logrank Test
##
## data:  Surv(time, Indicator) by type (1, 2)
## Z = 0.61754, p-value = 0.5369
## alternative hypothesis: true theta is not equal to 1
```

Since the resulting p-value is insignificant, we do not have enough evidence to reject H_0.