Group: Zining Fan, Mutian Wang, Siyuan Wang
UNI: zf2234, mw3386, sw3418

Graded Homework 2- Exercise 5

1. We want to calculate the conditional distribution, which is $P(y^{(k)}|(n_A^{(k)}, n_B^{(k)}, n_d^{(k)})$. So the total number of relapsed $n_d^{(k)}$ is given.
   We have $n_A^{(k)} + n_B^{(k)}$ patients, among which $n_A^{(k)}$ are from group A, $n_B^{(k)}$ are from group B. We know that there are $n_d^{(k)}$ patients of relapses from $n_A^{(k)} + n_B^{(k)}$.
   Under H_0, the survival function of the two groups are the same.$S_A = S_B$. So the probability of individual patient being relapse is equal among the two groups.
   So the occurance of relapses is random in the larger group of $n_A^{(k)} + n_B^{(k)} = m_A + m_B$, which is randomly selecting m out of $m_A + m_B$.
   In the $n_d^{(k)} = m_d$ relapse cases, the probability of $y^{(k)} = m$ coming from group A is:

   $$P(m|(m_A, m_B, m_d)) = \frac{\binom{m_A}{m})\binom{m_B}{m_d - m})}{\binom{m_A+m_B}{m_d}}$$

   which is:

   $$P(y^{(k)}|(n_A^{(k)}, n_B^{(k)}, n_d^{(k)}) = \frac{\binom{n_A^{(k)}}{y^{(k)}})\binom{n_B^{(k)}}{n_d^{(k)} - y^{(k)}})}{\binom{n_A^{(k)}+n_B^{(k)}}{n_d^{(k)}}}$$

   So it's a HyperGeometric distribution with parameter$(n_A^{(k)} + n_B^{(k)}, n_A^{(k)}, n_d^{(k)})$

2. From point 1, we know that $P(y^{(k)}|(n_A^{(k)}, n_B^{(k)}, n_d^{(k)})$ has a hypergeometric distribution.

   $$P(y^{(k)}|(n_A^{(k)}, n_B^{(k)}, n_d^{(k)}) = \frac{\binom{n_A^{(k)}}{y^{(k)}})\binom{n_B^{(k)}}{n_d^{(k)} - y^{(k)}})}{\binom{n_A^{(k)}+n_B^{(k)}}{n_d^{(k)}}}$$

   So the conditional expectation is given by:

   $$E^{(k)} = \frac{m \times m_A}{m_A + m_B} = \frac{y^{(k)} \times n_A^{(k)}}{n_A^{(k)} + n_B^{(k)}} = \frac{y^{(k)} n_A^{(k)}}{n^{(k)}}$$

   as $n^{(k)} = n_A^{(k)} + n_B^{(k)}$.
   The conditional variance is given by:

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

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

   as $n_s^{(k)} = n_A^{(k)} + n_B^{(k)} - n_d^{(k)}$.

3. According to conditional variance formula, we can fraction $Var[y^{(k)} - E^{(k)}]$ under H_0:

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

as $E[y^{(k)}|(n_A^{(k)}, n_B^{(k)}, n_d^{(k)})] = E^{(k)}$.

4. Under H_0 we have:

$$
\begin{aligned}
&Var[\sum_{k=1}^{K}(y^{(k)} - E^{(k)})] \\
&= Var[E[\sum_{k=1}^{K}(y^{(k)} - E^{(k)})|(n_A^{(K)}, n_B^{(K)}, n_d^{(K)})]] + E[Var[\sum_{k=1}^{K}(y^{(k)} - E^{(k)})|(n_A^{(K)}, n_B^{(K)}, n_d^{(K)})]] \\
&= Var[E[\sum_{k=1}^{K-1}(y^{(k)} - E^{(k)})|(n_A^{(K)}, n_B^{(K)}, n_d^{(K)}) + (y^{(K)} - E^{(K)})|(n_A^{(K)}, n_B^{(K)}, n_d^{(K)})]]+ \\
&E[Var[\sum_{k=1}^{K-1}(y^{(k)} - E^{(k)})|(n_A^{(K)}, n_B^{(K)}, n_d^{(K)}) + (y^{(K)} - E^{(K)})|(n_A^{(K)}, n_B^{(K)}, n_d^{(K)})]] \\
&= Var[E[\sum_{k=1}^{K-1}(y^{(k)} - E^{(k)})|(n_A^{(K)}, n_B^{(K)}, n_d^{(K)})] + E[(y^{(K)} - E^{(K)})|(n_A^{(K)}, n_B^{(K)}, n_d^{(K)})]]+ \\
&E[Var[\sum_{k=1}^{K-1}(y^{(k)} - E^{(k)})|(n_A^{(K)}, n_B^{(K)}, n_d^{(K)})] + Var[(y^{(K)} - E^{(K)})|(n_A^{(K)}, n_B^{(K)}, n_d^{(K)})]] \\
&(According\ to\ iid\ assumption\ of\ k) \\
&= Var[E[\sum_{k=1}^{K-1}(y^{(k)} - E^{(k)})|(n_A^{(K)}, n_B^{(K)}, n_d^{(K)})] + 0] + E[Var[\sum_{k=1}^{K-1}(y^{(k)} - E^{(k)})|(n_A^{(K)}, n_B^{(K)}, n_d^{(K)})]+ \\
&Var[(y^{(K)} - E^{(K)})|(n_A^{(K)}, n_B^{(K)}, n_d^{(K)})]] \\
&= Var[\sum_{k=1}^{K-1}(y^{(k)} - E^{(k)})] + E[Var[(y^{(K)} - E^{(K)})|(n_A^{(K)}, n_B^{(K)}, n_d^{(K)})]] \\
&= Var[\sum_{k=1}^{K-1}(y^{(k)} - E^{(k)})] + E[V^{(K)}]
\end{aligned}
$$

Similarly, we can apply this process to every other value of k, and get that:

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

5. H_0 : the survival functions of group A and group B are identical.
We implement the test to the sample in 1.2 using r. We first bin the time based on the interger point of the time values and then calculate the Z value with the binned times as k. We encouter some cases where $V^{(K)}$ are null and $E^{(k)}$ are 0. In these case, as it would not affect the result significantly, we recode the null value as 0 and encorporate them into our

calculation.

```
library(dplyr)
df <- read.table("hw2/df.txt", sep=" ")
names(df) <- c("time","types","survive")
df <- df[order(df$time),]
df$k <- floor(df$time)
for (i in c(1:nrow(df))){
  df$n_a[i] <- nrow(df %>% filter((types == 1 & k == df$k[i])))
  df$n_b[i] <- nrow(df %>% filter((types == 2 & k == df$k[i])))
  df$n_d[i] <- nrow(df %>% filter(survive == 1 & k == df$k[i]))
  df$y_k[i] <- nrow(df %>% filter((types == 1 & k == df$k[i] & survive == 1)))
}
df_5 <- distinct(df[c("k","n_a","n_b","n_d","y_k")])
for (i in c(1:nrow(df_5))){
  df_5$E_k <- (df_5$n_a*df_5$n_d)/(df_5$n_a + df_5$n_b)
  df_5$V_k <- (df_5$n_a*df_5$n_b*df_5$n_d*(df_5$n_a+df_5$n_b-df_5$n_d))/(((df_5
}
df_5$V_k[is.na(df_5$V_k)] <- 0
Z = (sum(df_5$y_k-df_5$E_k))/sqrt(sum(df_5$V_k))
Z
```

Here we compute the Z value and get -1.32141.

By CLT we know that Z has N(0,1) distribution. As -1.32141 > -1.96, we failed to reject H_0.

So we cannot reject the hypothesis that the two survival functions are identical.