Exam 1

Cong Zhang

Instructions

- Due date: **Tuesday, November 13**.
- This is a open resource exam, but you are not allowed to ask post exam questions online.
- You are not allowed to collaborate with classmates and/or people outside of class.
- Please circle or highlight your final answer.
- The toal possible point is 80.

Violation of this agreement will result in an **F** on this exam and it will be averaged in as a 0%.

---

1. Use the complete *WHAS100* dataset and gender as the group indicator, compute the log-rank statistic, Q, presented as Equation (9) on page 60 of note 2 with $\omega_i = 1$. Use Q to compute a p-value to test the null hypothesis of $H_o : S_0(t) = S_1(t)$. Do this **without a software package** (6 pts), and verify the calculation with `survdiff` (4 pts).

```r
library(survival)
library(dplyr)
library(ggplot2)
data(whas100, package = "survMisc")
Get.Surv <- function(dat)
{
    N_male = sum(dat$gender == 0)
    Male <- as.tibble(dat) %>% filter(gender == 0) %>%
        select(lenfol, fstat) %>%
        mutate(cernsored = 1*(fstat == 0)) %>%
        arrange(lenfol) %>%
        group_by(lenfol) %>%
        summarise_all(funs(sum(., na.rm = TRUE))) %>%
        mutate(n.risk = c(N_male, N_male - cumsum(fstat+cernsored))[1:length(lenfol)],
               MaleFailer = fstat,
               MaleNonFailer = n.risk - fstat) %>%
        select(lenfol, MaleFailer, MaleNonFailer, n.risk)

    N_female = sum(dat$gender == 1)
    Female <- as.tibble(dat) %>% filter(gender == 1) %>%
        select(lenfol, fstat) %>%
        mutate(cernsored = 1*(fstat == 0)) %>%
        arrange(lenfol) %>%
        group_by(lenfol) %>%
        summarise_all(funs(sum(., na.rm = TRUE))) %>%
        mutate(n.risk = c(N_female, N_female - cumsum(fstat+cernsored))[1:length(lenfol)],
               FeFailer = fstat,
               FeNonFailer = n.risk - fstat) %>%
        select(lenfol, FeFailer, FeNonFailer, n.risk)
    tab <- Male %>% full_join(Female, by = "lenfol")%>%
        arrange(lenfol) %>%
        fill(MaleNonFailer, n.risk.x, FeNonFailer, n.risk.y, .direction = "up") %>%
        fill(MaleNonFailer, n.risk.x, FeNonFailer, n.risk.y) %>%
        replace_na(list(MaleFailer = 0, FeFailer = 0)) %>%
        mutate(time = lenfol,
               n0 = n.risk.x, n1 = n.risk.y, n = n1+n0,
               d0 = MaleFailer, d1 = FeFailer, d = d1+d0,
               E_d1 = n1*d/n,
               Var_d1 = n1*n0*d*(n-d)/(n^2*(n-1))) %>%
        filter(d1 != 0 | d0 != 0) %>%
        select(time, n1, d1, n0, d0, n, d, E_d1, Var_d1) %>%
        mutate(S_km_1= cumprod((n1-d1)/n1),
               S_km_0= cumprod((n0-d0)/n0),
               distSCurve = abs(S_km_1 - S_km_0))
    return(tab)
}

tl <- Get.Surv(whas100)
Q = sum(tl$d1 - tl$E_d1)^2/sum(tl$Var_d1)
Q
```

```
## [1] 3.971377
```

```r
pvalue = pchisq(Q, df=1, lower.tail=FALSE)
pvalue
```

```
## [1] 0.04627992
```

```r
survdiff(Surv(lenfol, fstat) ~ gender, data = whas100)
```

```
## Call:
## survdiff(formula = Surv(lenfol, fstat) ~ gender, data = whas100)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## gender=0 65       28     34.6      1.27      3.97
## gender=1 35       23     16.4      2.68      3.97
##
##  Chisq= 4  on 1 degrees of freedom, p= 0.0463
```

Answer

We got log-rank statistic $Q = 3.97$ and p-value $= 0.046 < 0.05$. So we reject H_0 at 5% level of signifiacne and concluded there is a difference in survival time between two genders.

That was in consistant with the ourpur from `survdiff` as shown below.

2. There are many ways to form a basis for survival curve comparison. Here are some:
   a. The numerator term in Q without the square:

$$D_1 = \sum_{i=1}^{D} \{d_{1i} - E(d_{1i})\}.$$

   b. The largest distance between the two curves:

$$D_2 = \max |S_1(t) - S_0(t)|.$$

   c. The difference between the median survival times:

$$D_3 = S_1^{-1}(0.5) - S_0^{-1}(0.5)$$

   d. The difference between the mean survival times:

$$D_4 = \int_0^{t_{(n)}} \{S_1(u) - S_0(u)\}\, du,$$

   where $t_{(n)}$ is the maximum observed survival time.

   Compute each of the above statistic for the $WHAS100$ dataset (5 pts ×4).

```r
Get.D <- function(dat)
{
    ##A
    D1 <- sum(dat$d1 - dat$E_d1)
    ##B
    D2 <- max(dat$distSCurve)
    ###C
    Tm_0 <- min(dat$time[dat$S_km_0 <= 0.5])
    Tm_1 <- min(dat$time[dat$S_km_1 <= 0.5])
    D3 <- Tm_1 - Tm_0
    ###D
    D4 <- sum((c(1, dat$S_km_1) - c(1, dat$S_km_0))[1:length(dat$S_km_1)]*diff(c(0, dat$time)))
    return(tibble(D1, D2, D3, D4))
}
D <- Get.D(tl)
kable(round(D, 3), caption = "Statistics for survival curve comparison")
```

Table 1: Statistics for survival curve comparison

D1	D2	D3	D4
6.62	0.385	-818	-428.741

Answer

The four statistics were computed and shown in Table 1.

3. The statistics computed in (2) do not provide meaningful interpretations when standing along. We will use a permutation approach to test for the null hypothesis of $H_o : S_0(t) = S_1(t)$ based on these statistics. The idea of a permutation test is simple. The general procedure can be summarized into the following steps:
   i. Compute the desired statistic based on the observed data; we will call this the observed statistic.
   ii. Permute the data under the null.
   iii. Compute the statistics for each possible permutation in Step ii.; we will call these permutation statistics.
   iv. Draw conclusion based on where the observed statistic stands among the permutation statistics.

   The statistics we computed in (2) are the observed statistics in Step i. If the null hypothesis of $H_o : S_0(t) = S_1(t)$ is true, then one can randomly shuffle the group indicator to generate different permutations (Step ii) and the statistics for these permutations should be similar (Step iii).

   a. (5 pts×4) Generate 5000 permutation and, for each of the permutation, compute the four statistics presented in (2). We will call the permutated statistics D^*_{1i}, D^*_{2i}, D^*_{3i} and D^*_{4i} for $i = 1, \ldots, 5000$. Create a histgram for these permutated statistics and print the `summary`.
   b. (5 pts×4) Compute the p-value based on these statistics by

$$p = 2 \cdot \frac{\min(N_1, N_2)}{5000},$$

   where $N_1 = \#\{D \geq D^*\}$, $N_2 = \#\{D \leq D^*\}$, and # means the "number of", e.g., N_1 is the number of these permutated statistics less than or equal to the observed statistic.

```r
permute.once <- function(dat)
{
    dat$gender <- sample(dat$gender)
    sf <- Get.Surv(dat)
    return(Get.D(sf))
}
N <- 5000
i <- 1
DD <- tibble(D1= numeric(N),
             D2= numeric(N),
             D3= numeric(N),
             D4= numeric(N))

while(i <= N){
    a <- as.numeric(permute.once(whas100))
    if(!(Inf %in% abs(a))){
        DD[i,] = a
        i = i+1}}
kable(summary(DD), caption = "Summary of permutated statistcs")
```

Table 2: Summary of permutated statistcs

D1	D2	D3	D4
Min. :-12.015	Min. :0.1041	Min. :-1576.00	Min. :-736.09
1st Qu.: -2.039	1st Qu.:0.2369	1st Qu.: -390.00	1st Qu.:-163.25
Median : 0.156	Median :0.3038	Median : -136.00	Median : -13.39
Mean : 0.135	Mean :0.3091	Mean : -37.57	Mean : -12.42
3rd Qu.: 2.312	3rd Qu.:0.3779	3rd Qu.: 356.00	3rd Qu.: 134.13
Max. : 10.286	Max. :0.6491	Max. : 1153.00	Max. : 726.53

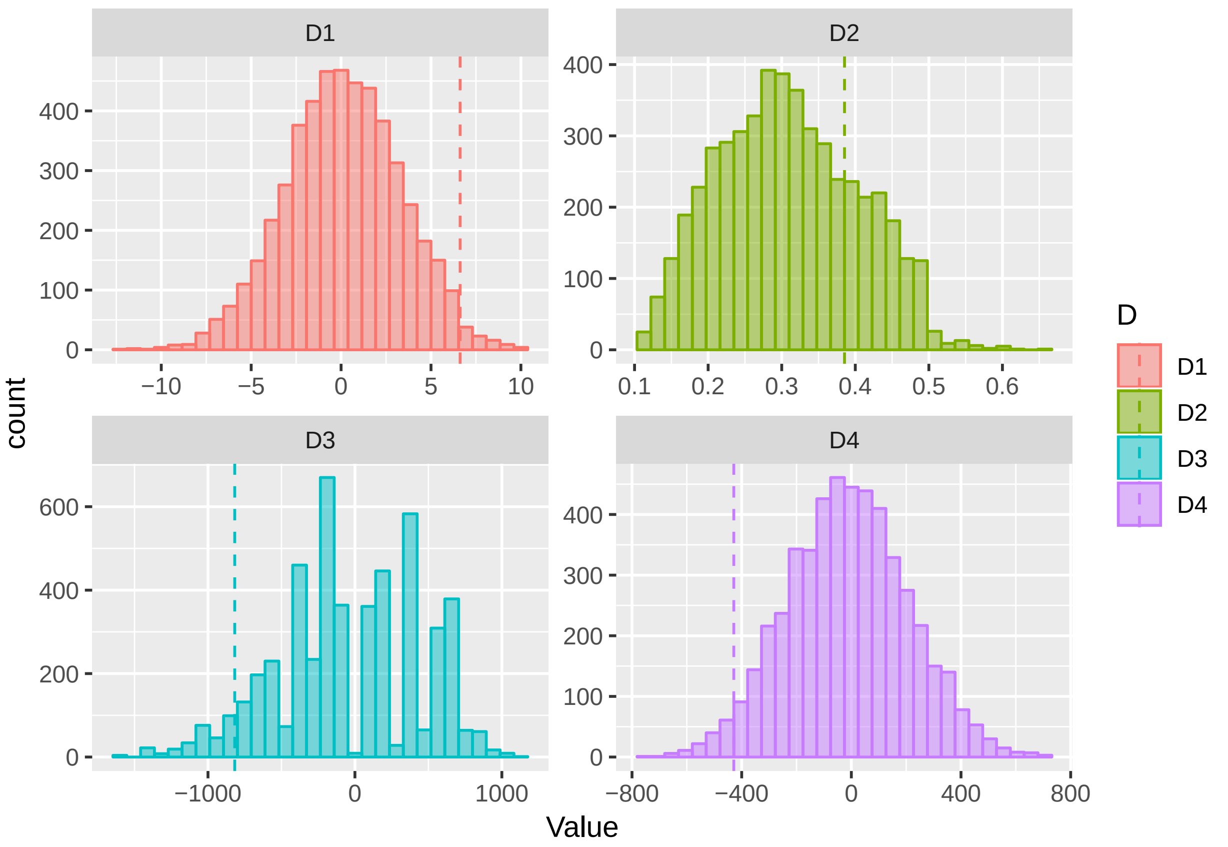

Figure 1: Histogram of permuted statistics

```
p1 <- 2*min(sum(D$D1 < DD$D1)/N, 1- sum(D$D1 < DD$D1)/N)
p2 <- 2*min(sum(D$D2 < DD$D2)/N, 1- sum(D$D2 < DD$D2)/N)
p3 <- 2*min(sum(D$D3 < DD$D3)/N, 1- sum(D$D3 < DD$D3)/N)
p4 <- 2*min(sum(D$D4 < DD$D4)/N, 1- sum(D$D4 < DD$D4)/N)
kable(data.frame(p1, p2, p3, p4), caption = "p-values by difference statistcs")
```

Table 3: p-values by difference statistcs

p1	p2	p3	p4
0.0332	0.4664	0.118	0.0568

```
DD <- gather(DD, key = D, Value, D1:D4)
D <- gather(D, key = D, Value, D1:D4)
ggplot(DD, aes(x=Value, fill=D, color=D))+
    geom_histogram(alpha=0.5)+
    facet_wrap(D ~ ., scale = "free") +
    geom_vline(data=D, aes(xintercept=Value, color=D),
               linetype="dashed")
```

Answer

Permutation of four statistcs and p-values were computed and and shown in Table 2 and Table 3. Histgram of those permuted statistics were ploted with observed D indicated by dashline (Figure 1).

4. Another method to compare two survival curves is to consider a sign test. Suppose we have two groups of uncensored survival times:

   Males: $x_1, x_2, \ldots, x_{n_0}$.

   Females: $y_1, y_2, \ldots, y_{n_1}$.

   The sign test looks at the statistic

$$U = \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \text{sgn}(x_i - y_j),$$

   where sgn$(\cdot)$ is the sign function. In the pretense of right censoring, survival times can not be compared directly and a modified version of $U = \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} U_{ij}$ is considered, where

$$U_{ij} = \begin{cases} 1 & \text{if } x_i > y_j, y_j \text{ is uncersored.} \\ -1 & \text{if } x_i < y_j, x_i \text{ is uncersored.} \\ 0 & \qquad\quad \text{otherwise.} \end{cases}$$

   a. (5 pts) Compute U for the WHAS100 dataset.
   b. (5 pts) Create a histgram for these permutated statistics and print the `summary`. Obtain a permutation p-value based on 5000 permutations.

```
compute.U <- function(dat){
    idx <- (dat$gender == 0)
    n0 <- sum(idx)
    n1 <- sum(1-idx)
    dat0 <- dat[idx,c("lenfol", "fstat", "gender")]
    dat0 <- dat0[rep(1:n0, each = n1),]
    dat1 <- dat[!idx,c("lenfol", "fstat", "gender")]
    dat1 <- dat1[rep(1:n1, n0),]
    U <- (dat0$lenfol - dat1$lenfol)
    sum(U*dat1$fstat > 0) - sum(U*dat0$fstat < 0)
}
permute.u <- function(dat){
    dat$gender = sample(dat$gender)
    compute.U(dat)
}
N = 5000
U <- compute.U(whas100)
U
```

```
## [1] 459
```

```
u <- replicate(N, permute.u(whas100))
cat("Summary of permutated U")
```

```
## Summary of permutated U
```

```
summary(u)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -864.000 -175.250   -6.000   -5.584  158.000  877.000
```

```
p <- 2*min(sum(U < u)/N, 1- sum(U < u)/N)
p
```

```
## [1] 0.0696
```

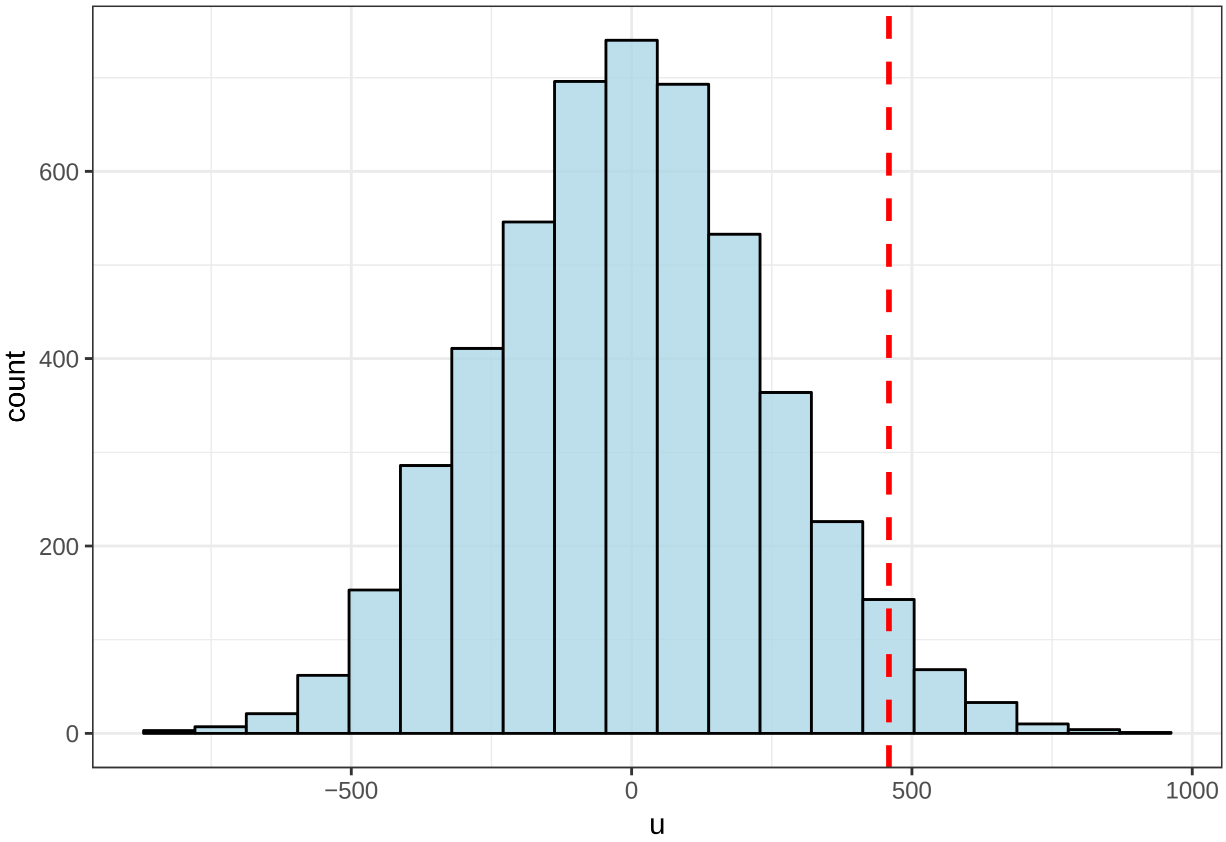

Figure 2: Histogram of permuted statistics in sign test

```r
ggplot(data.frame(u), aes(x=u)) +
    geom_histogram(alpha=0.8, color="black", fill="lightblue", bins = 20)  +
    geom_vline(aes(xintercept=U),
               color="red", linetype="dashed", size=1) +
    theme_bw()
```

Answer

U for the WHAS100 dataset was 459 and the summary of permutated U was shown above. We got p-value of 0.0696 so we cannot reject H_0 and we concluded there is no difference between the two survival curves. Histgram of permuted statistics was shown in Figure 2 with observed U indicated by dashline.