Homework 3 Cong Zhang

Cong Zhang

1. **Textbook problem 2.1** Listed below are values of survival time in years for 6 males and 6 females from the *WHAS100* study. Right-censored times are denoted by a "+" as a superscript.

   Males: 1.2, 3.4, 5.0^+, 5.1, 6.1, 7.1

   Females: 0.4, 1.2, 4.3, 4.9, 5.0, 5.1^+

   Using these data, compute the following **without a software package**:

   a. (2 pts) The Kaplan-Meier estimate of the survival function for each gender.
   b. (2 pts) Pointwise 95 % confidence intervals for the survival functions estimated in problem (1a).
   c. (2 pts) Pointwise 95 % confidence interval estimates of the 50th percentile of the survival time distribution for each gender.
   d. (2 pts) The estimated mean survival time for each gender using all available times, upto 7.1.
   e. (2 pts) A graph of the estimated survival functions for each gender computed in problem (1a) along with the point wise and overall 95 % limits computed in problem (1b).

   Answer

   The Kaplan-Meier estimate and 95 % confidence intervals of the survival function for each gender were shown in Table 1&2 and Figure 1. 95% confidence intervals were computed using log-log approaches.

   95 % confidence interval estimates of the 50th percentile of the survival time and estimated mean survival time were shown in Table 3.

```
> library(dplyr)
> library(tibble)
> library(tidyverse)
> library(knitr)
> library(kableExtra)
> fit.surv <- function(dat){
+     dat %>% mutate(n.risk = c(N, N - cumsum(event+censored))[-7] ) %>%
+         filter(censored == 0) %>%
+         mutate(km_s = (n.risk - event)/n.risk, S.km= cumprod(km_s)) %>%
+         mutate(kk = event/(n.risk*(n.risk - event)),
+               std_err = sqrt((S.km)^2*cumsum(kk)),
+               logsd = sqrt(1/(-log(S.km))^2*cumsum(kk)),
+               lowerCI = log(-log(S.km)) + qt(0.975, Inf)*logsd,
+               upperCI = log(-log(S.km)) - qt(0.975, Inf)*logsd) %>%
+         mutate(lowerCI = ifelse(exp(-exp(lowerCI))<0, 0, exp(-exp(lowerCI))),
+               upperCI = ifelse(exp(-exp(upperCI))>1, 1, exp(-exp(upperCI)) )) %>%
+         select( time, n.risk, event, S.km, std_err, lowerCI, upperCI )
+ }
>
> N = 6
> m <- tibble(time = c(1.2, 3.4, 5.0, 5.1, 6.1, 7.1),
+                 event = c(1, 1, 0, 1, 1, 1),
+                 censored = c(0, 0, 1, 0, 0, 0))
> male <- fit.surv(m)
>
> kable(round(male, 3), caption = "KM estimator for male")
```

Table 1: KM estimator for male

time	n.risk	event	S.km	std_err	lowerCI	upperCI
1.2	6	1	0.833	0.152	0.273	0.975
3.4	5	1	0.667	0.192	0.195	0.904
5.1	3	1	0.444	0.222	0.066	0.785
6.1	2	1	0.222	0.192	0.010	0.615
7.1	1	1	0.000	NaN	NA	NA

```
> f <- tibble(time = c(0.4, 1.2, 4.3, 4.9, 5.0, 5.1),
+             event = c(1, 1, 1, 1, 1, 0),
+             censored = c(0, 0, 0, 0, 0, 1))
> female <- fit.surv(f)
> kable(round(female, 3), caption = "KM estimator for female")
```

Table 2: KM estimator for female

time	n.risk	event	S.km	std_err	lowerCI	upperCI
0.4	6	1	0.833	0.152	0.273	0.975
1.2	5	1	0.667	0.192	0.195	0.904
4.3	4	1	0.500	0.204	0.111	0.804
4.9	3	1	0.333	0.192	0.046	0.676
5.0	2	1	0.167	0.152	0.008	0.517

```
> get.median <- function(dat)
+ {
+     dat %>% mutate(t50 = min(time[S.km <= 0.5]),
+                    u50 = max(time[S.km >= (0.5 + 0.1)]),
+                    l50 = min(time[S.km <= (0.5 - 0.1)]),
+                    f50 = (S.km[time == u50] - S.km[time == l50])/(l50 - u50),
+                    se50 = std_err[time == t50]/f50,
+                    lower50 = t50 - qt(0.975, Inf)*se50,
+                    upper50 = t50 + qt(0.975, Inf)*se50,
+                    prod = max(cumsum(c(1, S.km)[-6]*diff(c(0, time))))
+     ) %>%
+         filter(time == 1.2) %>%
+         select(t50, lower50, upper50, prod)
+ }
> t <- round(rbind(get.median(male), get.median(female)), 3)
> colnames(t) <- c("Median", "LowerCL", "UpperCL", "E(MeanSurvTime)")
> rownames(t) <- c("Male", "Female")
> kable(t,caption = "Median survival time")
```

Table 3: Median survival time

	Median	LowerCL	UpperCL	E(MeanSurvTime)
Male	5.1	2.454	7.746	4.833
Female	4.3	-0.141	8.741	3.467

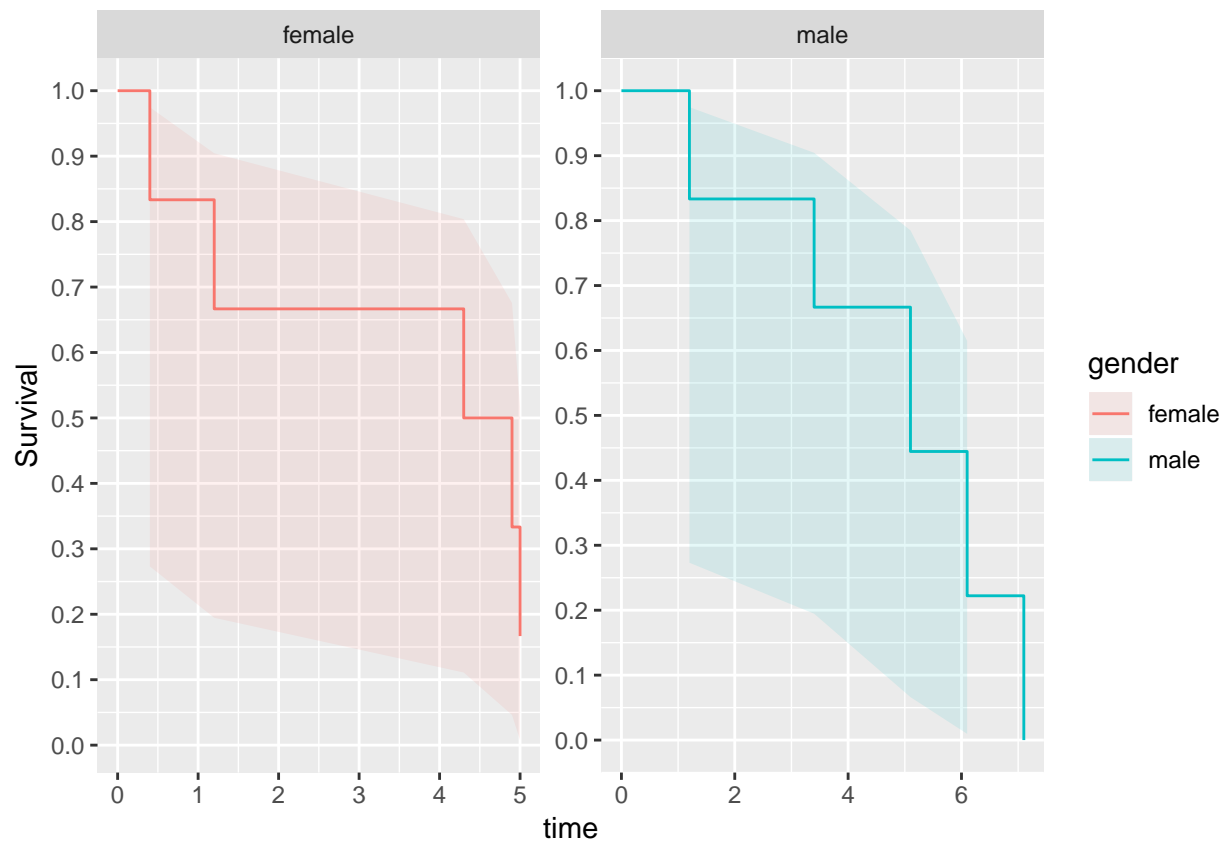Figure 1: Survival curves and 95% CI

```
> library(ggplot2)
> data <- bind_rows(list(male = male, female = female), .id = "gender") %>%
+ add_row(gender = "male",time = 0, n.risk = 0, S.km = 1, .before = 1) %>%
+ add_row(gender = "female",time = 0, n.risk = 0, S.km = 1, .before = 1)
> ggplot(data = data, aes(colour = gender, shape=gender)) +
+ geom_step(aes(x = time, y = S.km))+
+     facet_wrap(gender ~ ., scale = "free") +
+ scale_y_continuous(breaks=seq(0,1,by=0.1)) +
+ geom_ribbon(aes(x = time, ymax = upperCI, ymin = lowerCI, fill = gender),
+             linetype = 0,alpha = 0.10) +
+ ylab("Survival")
```