Assignment 3

Luis Nicolas Luarte Rodriguez

```
library(tidyverse)
library(knitr)
library(ggplot2)
library(knitr)
library(kableExtra)
library(survival)
```

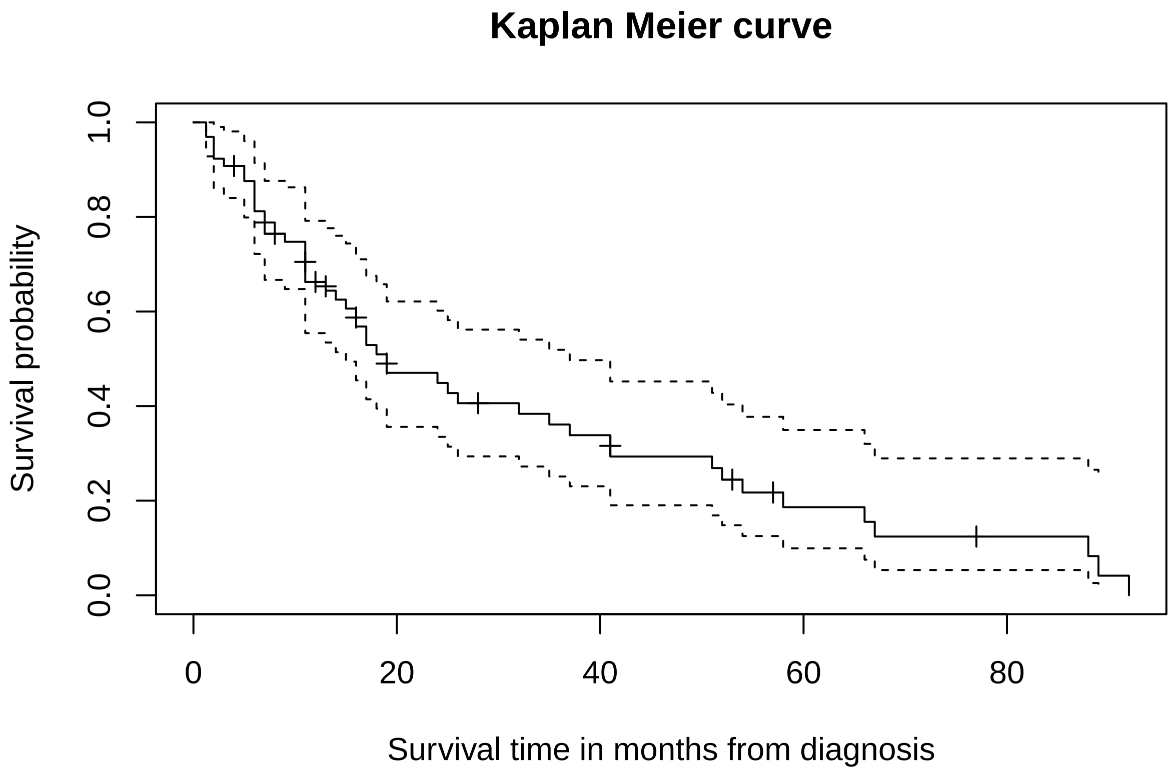
```
# load the dataset
setwd('/home/nicoluarte/uni/PHD/stat_course')
dataSet <- as_tibble(read.csv('survival65.csv', sep="\t"))
head(dataSet)
```

```
## # A tibble: 6 x 11
##    TIME STATUS LOGBUN   HGB PLATELET   AGE LOGWBC FRACTURE LOGPBM PROTEIN
##   <dbl>  <int>  <dbl> <dbl>    <int> <int>  <dbl>    <int>  <dbl>   <int>
## 1  1.25      1   2.22   9.4        1    67   3.66        1   1.95      12
## 2  1.25      1   1.94  12          1    38   3.99        1   1.95      20
## 3  2         1   1.52   9.8        1    81   3.88        1   2          2
## 4  2         1   1.75  11.3        0    75   3.81        1   1.26       0
## 5  2         1   1.30   5.1        0    57   3.72        1   2          3
## 6  3         1   1.54   6.7        1    46   4.48        0   1.93      12
## # ... with 1 more variable: CALCIUM <int>
```

Use the Kaplan Meier method to estimate the distribution of survival time for the total sample.

a) Plot the Kaplan Meier curve for the total sample.

```
s <- Surv(time=dataSet$TIME, event=dataSet$STATUS)
plot(s,
     main = 'Kaplan Meier curve',
     xlab = 'Survival time in months from diagnosis',
     ylab = 'Survival probability',
     mark.time=TRUE)
```



b) What is the mean survival time and standard error?

```
estimate <- survfit(s ~ 1)
print(estimate, print.rmean=TRUE)
```

```
## Call: survfit(formula = s ~ 1)
##
##           n      events      *rmean *se(rmean)      median     0.95LCL     0.95UCL
##       65.00       48.00       32.15        3.99       19.00       15.00       37.00
##     * restricted mean with upper limit =   92
```

Mean survival is 32.15 and its standard error 3.99

c) What is the median survival time and 95% CI?

```
print(estimate)
```

```
## Call: survfit(formula = s ~ 1)
##
##        n  events  median 0.95LCL 0.95UCL
##       65      48      19      15      37
```

The median survival time is 19 and its confidence interval at 95% (15 - 37)

d) How many censored observations are there?

```
# as per the data set description a status of 0 means censored
# then
sum(dataSet['TIME'] == 0)
```

```
## [1] 0
```

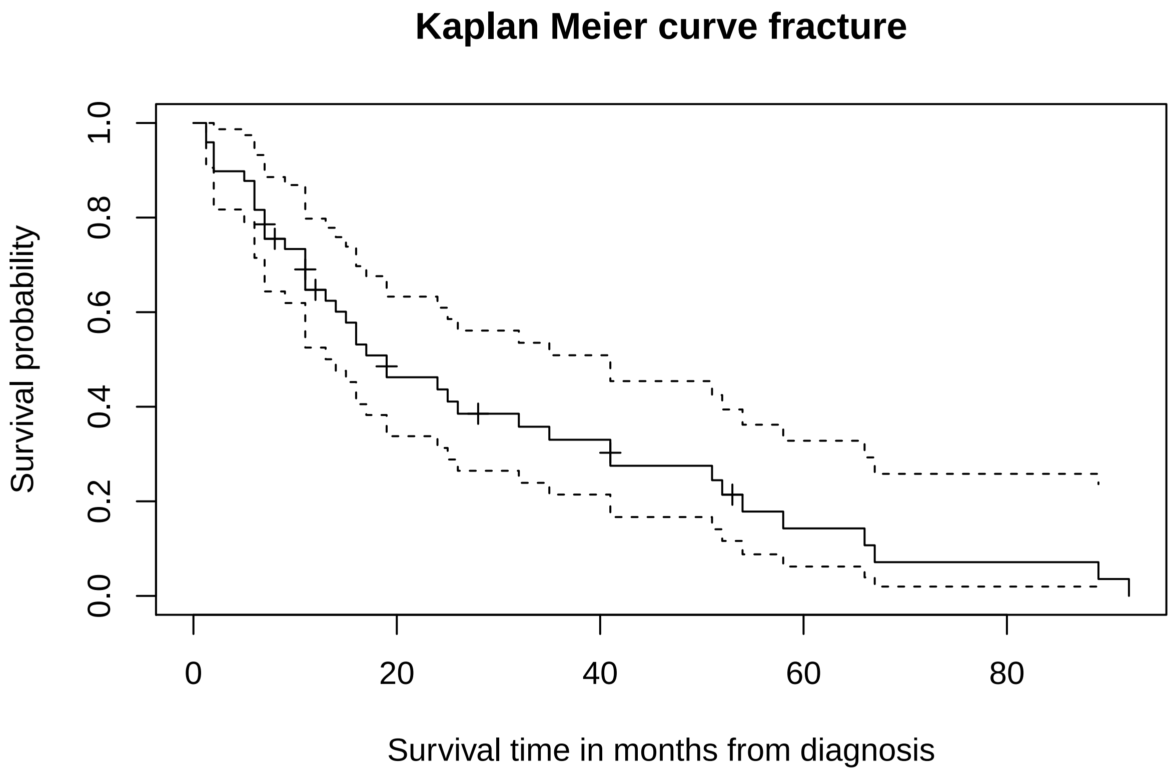
```
# defines the number of censored observations
```

17

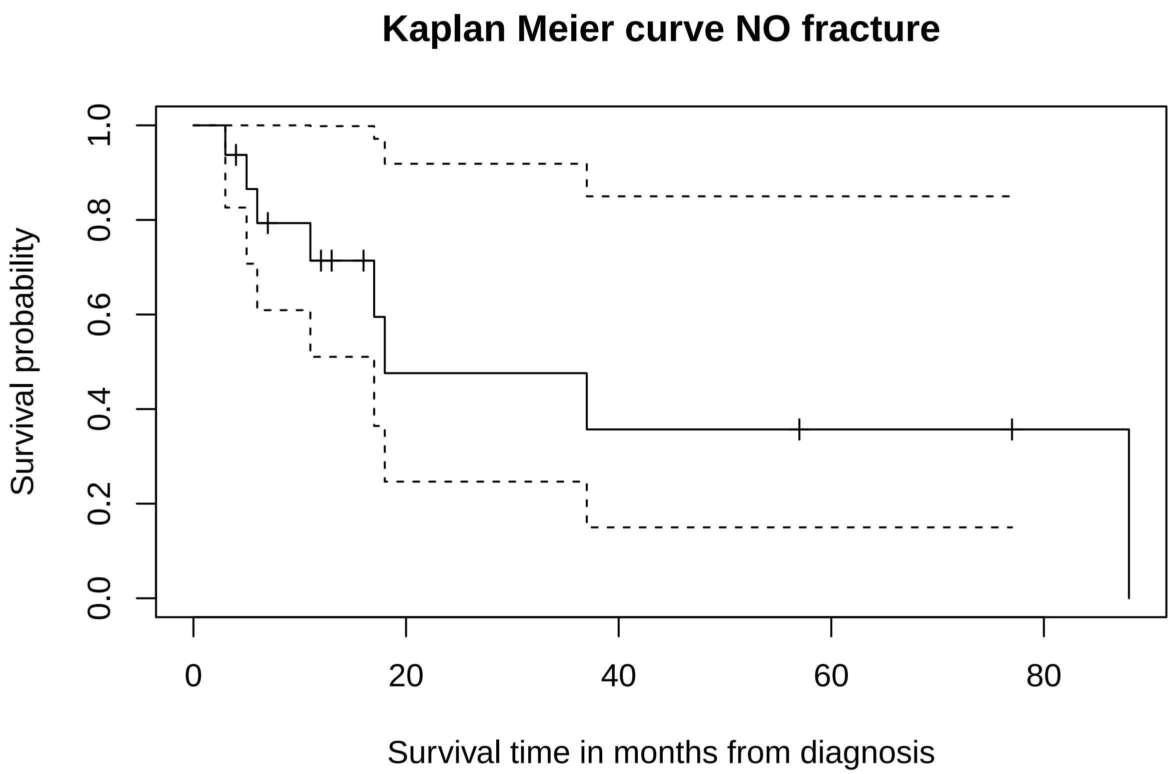
Using the Kaplan Meier method to estimate whether fracture influences survival time.

a) Plot on Kaplan Meier curve for the patients with fracture and another one for the patients without fracture

```
# create to data sets
fractureYes <- dataSet %>% filter(FRACTURE == 1)
fractureYesS <- Surv(time=fractureYes$TIME, event=fractureYes$STATUS)
fractureNo <- dataSet %>% filter(FRACTURE == 0)
fractureNoS <- Surv(time=fractureNo$TIME, event=fractureNo$STATUS)
plot(fractureYesS,
     main = 'Kaplan Meier curve fracture',
     xlab = 'Survival time in months from diagnosis',
     ylab = 'Survival probability',
     mark.time=TRUE)
```



```
plot(fractureNoS,
     main = 'Kaplan Meier curve NO fracture',
     xlab = 'Survival time in months from diagnosis',
     ylab = 'Survival probability',
     mark.time=TRUE)
```



b) What is the median survival time and 95% CI in each group?

```
yesEst <- survfit(fractureYesS ~ 1)
noEst <- survfit(fractureNoS ~ 1)
print(yesEst)
```

```
## Call: survfit(formula = fractureYesS ~ 1)
##
##       n  events  median 0.95LCL 0.95UCL
##      49      40      19      14      41
```

```
print(noEst)
```

```
## Call: survfit(formula = fractureNoS ~ 1)
##
##       n  events  median 0.95LCL 0.95UCL
##      16       8      18      17      NA
```

For the fracture group: median = 19, CI = (14, 41). For the no fracture group: median = 18, CI = (17, infinity)

c) Interpret the Log Rank test and the survival curves.

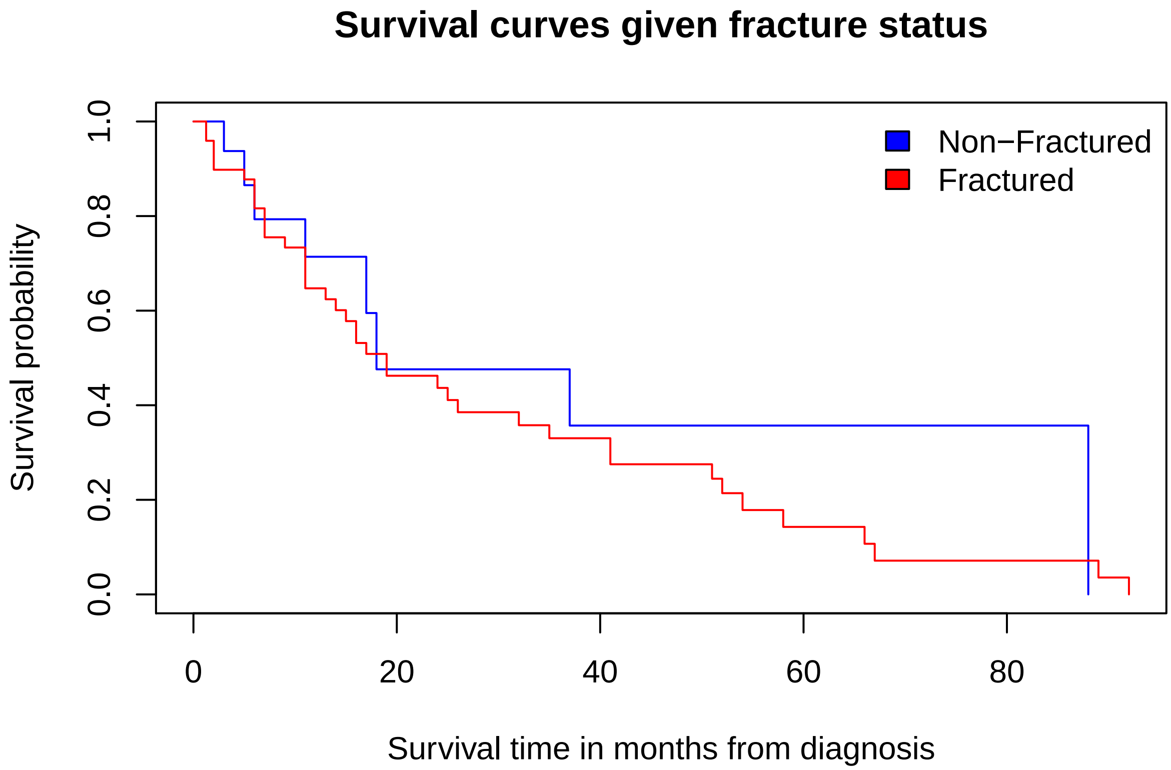
```
# Log Rank test
survdiff(Surv(time=dataSet$TIME, event=dataSet$STATUS) ~ dataSet$FRACTURE)
```

```
## Call:
## survdiff(formula = Surv(time = dataSet$TIME, event = dataSet$STATUS) ~
##     dataSet$FRACTURE)
##
```

```
##                        N Observed Expected (O-E)^2/E (O-E)^2/V
## dataSet$FRACTURE=0 16        8     10.7     0.694     0.941
## dataSet$FRACTURE=1 49       40     37.3     0.200     0.941
##
##  Chisq= 0.9  on 1 degrees of freedom, p= 0.3
```

The log-rank test considers the survival function from each subdivision (fracture or not), and test the hypothesis that there's no difference in such function. Tested under the Chi squared distribution, the difference between the expected values are non-significant p = 0. Given that, the interpretation is the estimation of the survival function does not differ given the fracture/no fracture condition, and as such, the expected survival is not affected by having a fracture.

```
plot(survfit(Surv(time=dataSet$TIME, event=dataSet$STATUS) ~ dataSet$FRACTURE), main = "Survival curves
legend("topright", legend=c("Non-Fractured", "Fractured"), fill=c("blue", "red"), bty="n")
```



If curves are observed, while they follow the same "shape" is clear that no-fracture group lacks resolution because of having less observations. However, up to the last observation, there's is a big gap with no observations, and thus the estimated survival function can be misleading, this can be supported by the big confidence intervals for the no-fracture group.