

# Practical 4

09/03/2022

## Exercise 1

Consider the following small data set of survival times: 3, 4\*, 5, 6, 6\*, 8\*, 11, 14, 15, 16\* where \* indicates right-censored observations.

### Question (c)

Calculate and plot the Kaplan–Meier estimate  $\hat{S}(t)$  of the survival function  $S(t)$  by using R. Note that you have to type in the survival times and the event indicator see the lecture notes.

```
#specify the vector of survival and censored times
t<-c(3, 4, 5, 6, 6, 8, 11, 14, 15, 16)
#create an indicator variable to denote whether the subject was censored or not
#1=not censored
#0= censored
status<-rep(1,10)
#specify those who were censored
status[c(3,5,6,10)]<-1
#create your survival object in R
fit<-survfit(Surv(t,status)~1)
#check numerically your estimated survival function
summary(fit)
```

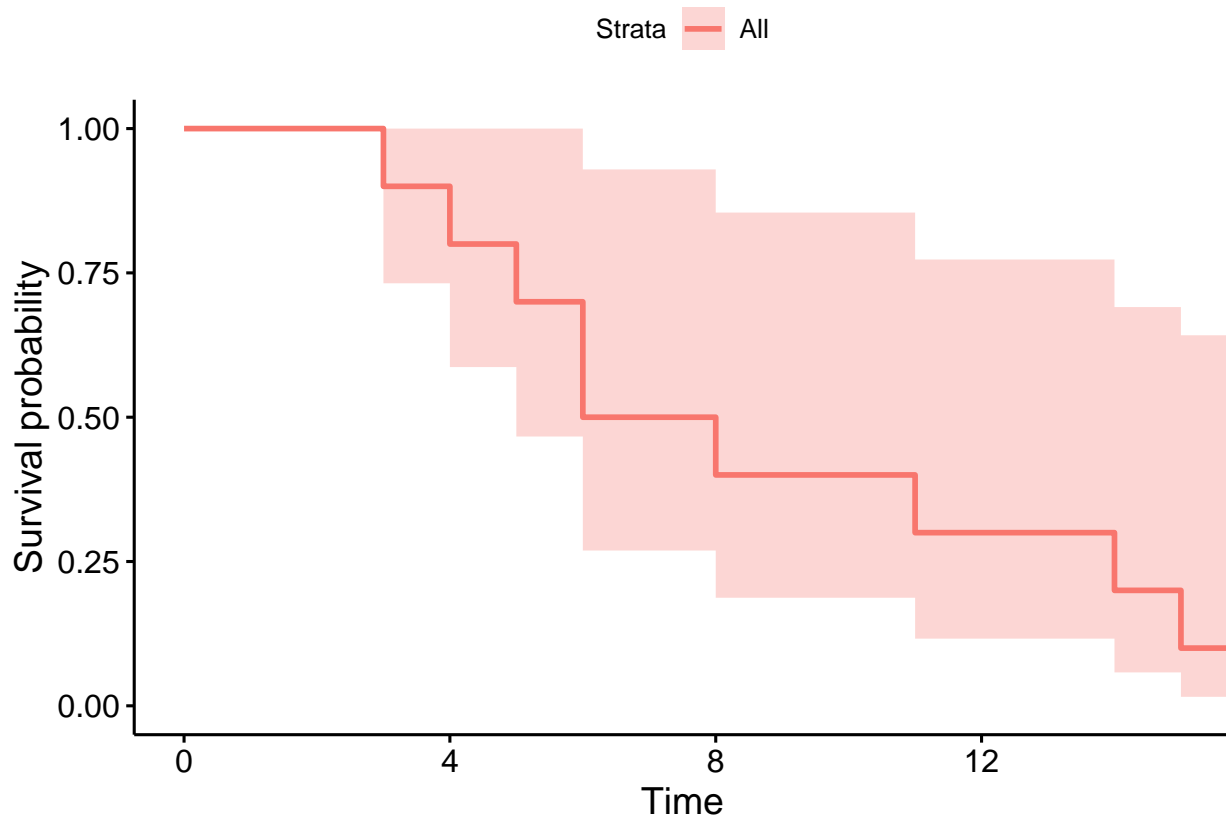
```
## Call: survfit(formula = Surv(t, status) ~ 1)
##
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    3      10        1     0.9  0.0949    0.7320      1.000
##    4       9        1     0.8  0.1265    0.5868      1.000
##    5       8        1     0.7  0.1449    0.4665      1.000
##    6       7        2     0.5  0.1581    0.2690      0.929
##    8       5        1     0.4  0.1549    0.1872      0.855
##   11       4        1     0.3  0.1449    0.1164      0.773
##   14       3        1     0.2  0.1265    0.0579      0.691
##   15       2        1     0.1  0.0949    0.0156      0.642
##   16       1        1     0.0   NaN           NA           NA
```

```
#check the median survival time
fit
```

```
## Call: survfit(formula = Surv(t, status) ~ 1)
##
##      n  events  median 0.95LCL 0.95UCL
##    10     10       7       5      NA
```

```
#plot the KM estimates for the survival function
ggsurvplot(fit,data=cbind(as.data.frame(t,status)))
```

```
## Warning in as.data.frame.numeric(t, status): 'row.names' non è un vettore di
## caratteri di lunghezza 10 -- omissso. Sarà un errore!
```



### Question (d)

Fit a survival curve assuming a Weibull distribution. Compute the MLEs for the two parameters using R

```
#fit the Weibull model
myfit1<-survreg(Surv(t,status)~1,dist="weibull")
summary(myfit1)

##
## Call:
## survreg(formula = Surv(t, status) ~ 1, dist = "weibull")
##               Value Std. Error      z      p
## (Intercept)  2.302      0.162 14.25 <2e-16
## Log(scale)  -0.728      0.252 -2.89 0.0039
##
## Scale= 0.483
##
## Weibull distribution
## Loglik(model)= -28.7  Loglik(intercept only)= -28.7
## Number of Newton-Raphson Iterations: 6
## n= 10

#shape parameter
gamma<-1/myfit1$scale
#scale parameter
lambda<-exp(-myfit1$coefficients[1]/myfit1$scale)
```

```
#alternatively use the ConvertWeibull function
ConvertWeibull(myfit1)
```

```
## $vars
##           Estimate      SE
## lambda 0.00852566 0.0114779
## gamma  2.07009189 0.5212427
```

## Question (e)

```
#check the p-value of log(scale)
#since p-value<0.05, the log of the shape parameter is different from 0
#hence an exponential model would not be adequate for such data
```

## Exercise 2

We will analyse the aml dataset which is available in R. The data is from a clinical trial to evaluate the efficacy of maintenance chemotherapy for acute myelogenous leukaemia. The trial was conducted by Embury et al. (1977) at Stanford University. After reaching a stage of remission through treatment by chemotherapy, patients were randomized into two groups. The first group received maintenance chemotherapy and the second group did not. The aim of the study was to see if maintenance chemotherapy increased the length of the remission. The variable time is length of the complete remission (in weeks). The variable cens is an indicator of right censoring. 1 indicates that the patient had a relapse and so time is the length of the remission. 0 indicates that the patient had left the study or was still in remission in October 1974, that is the length of remission is right-censored. The variable group corresponds to the group into which the patient was randomized. Group 1 received maintenance chemotherapy, group 2 did not. We will use R to analyse these data.

## Question (a)

Check the data for example use the function str.

```
data(aml)
str(aml)

## 'data.frame':   23 obs. of  3 variables:
## $ time : num  9 13 13 18 23 28 31 34 45 48 ...
## $ cens : num  1 1 0 1 1 0 1 1 0 1 ...
## $ group: num  1 1 1 1 1 1 1 1 1 1 ...

summary(aml)
```

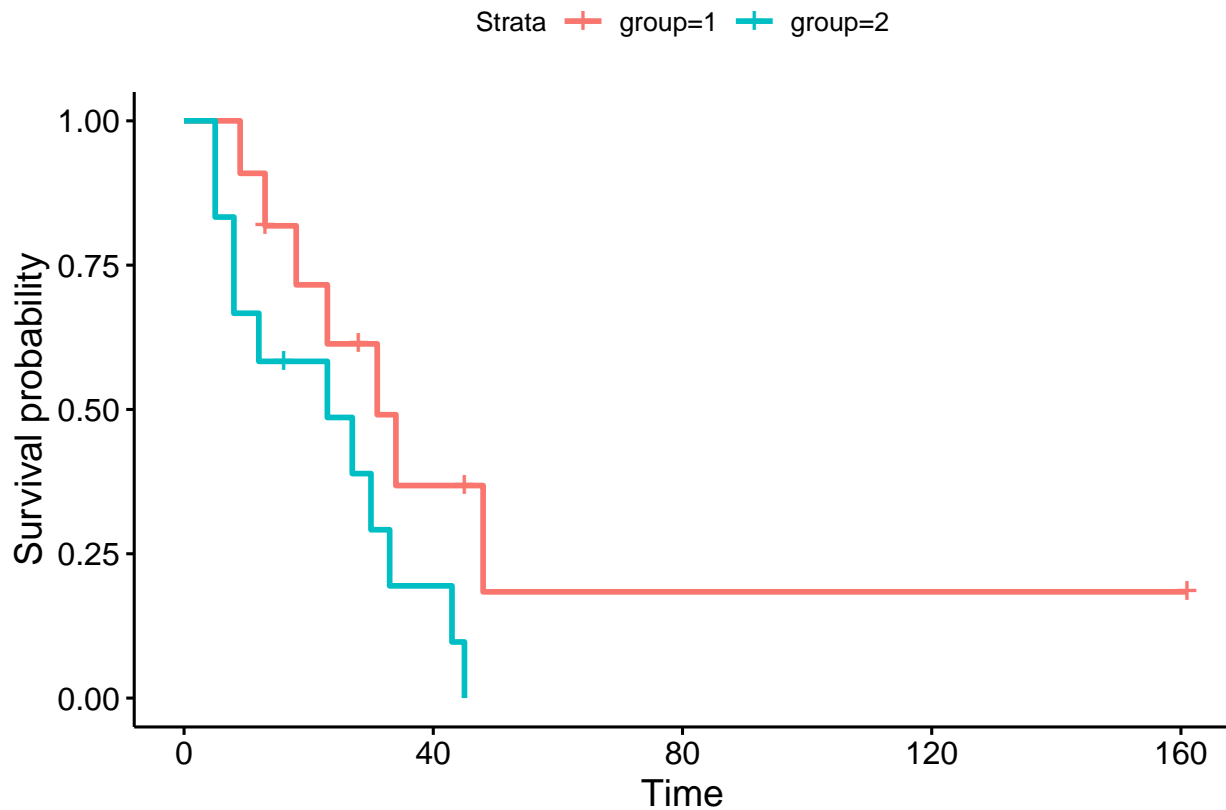
```
##           time           cens           group
## Min.      : 5.00   Min.      :0.0000   Min.      :1.000
## 1st Qu.: 12.50   1st Qu.:1.0000   1st Qu.:1.000
## Median : 23.00   Median :1.0000   Median :2.000
## Mean      : 29.48   Mean      :0.7826   Mean      :1.522
## 3rd Qu.: 33.50   3rd Qu.:1.0000   3rd Qu.:2.000
## Max.      :61.00   Max.      :1.0000   Max.      :2.000
```

## Question (b)

Plot the Kaplan Meier estimates of the survival distributions of the two groups in one plot.

```
#estimate the survival distributions in the two groups
fit<-survfit(Surv(time,cens)~group,data=aml)
```

```
#plot the survival distributions in the two groups
ggsurvplot(fit, data = aml)
```



### Question (c)

Use the logrank test to assess whether there is a significant difference in survival between the two groups. Formulate the null hypothesis and write down your conclusion.

```
#implement the logrank test
```

```
test<-survdif(Surv(time,cens)~group,data=aml)
log_rank<-(test$obs[1]-test$exp[1])^2/test$exp[1]+(test$obs[2]-test$exp[2])^2/test$exp[2]
print(log_rank)
```

```
## [1] 3.135172
```

```
critical_value<-qchisq(0.95,1)
print(critical_value)
```

```
## [1] 3.841459
```

```
p_value<-pchisq(log_rank,1,ncp = 0, lower.tail = FALSE)
print(p_value)
```

```
## [1] 0.07662031
```

```
#decision
```

```
#H0:S1(t)=S2(t) for every t>=0
log_rank>critical_value
```

```
## [1] FALSE
```

```
p_value<0.05
```

```
## [1] FALSE
```

```
#no significant difference in survival between the two groups
```

## Question (d)

Fit a **Cox Proportional Hazard model** and test whether the survival is the same in the two groups using a Wald test, a score test and a likelihood ratio test.

```
#fit a Cox PH model
```

```
coxph_model<-coxph(Surv(time,cens)~group,data=aml)
summary(coxph_model)
```

```
## Call:
```

```
## coxph(formula = Surv(time, cens) ~ group, data = aml)
```

```
##
```

```
## n= 23, number of events= 18
```

```
##
```

```
##      coef exp(coef) se(coef)      z Pr(>|z|)
```

```
## group 0.9155      2.4981      0.5119 1.788  0.0737 .
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
```

```
##      exp(coef) exp(-coef) lower .95 upper .95
```

```
## group      2.498      0.4003      0.9159      6.813
```

```
##
```

```
## Concordance= 0.619 (se = 0.063 )
```

```
## Likelihood ratio test= 3.38 on 1 df,  p=0.07
```

```
## Wald test              = 3.2 on 1 df,  p=0.07
```

```
## Score (logrank) test = 3.42 on 1 df,  p=0.06
```

```
#wald test
```

```
coxph_model$wald.test
```

```
## group
```

```
## 3.1983
```

```
#score test
```

```
coxph_model$score
```

```
## [1] 3.416734
```

```
#likelihood ratio test
```

```
lrt<-(-2)*(coxph_model$loglik[1]-coxph_model$loglik[2])
```

```
#no significant difference in survival as suggested by the
```

```
#p-values of the tests in the output of the model
```

```
#H0 is not rejected at 5% significance level
```

## Question (e)

Fit a **Proportional Hazard model with a Weibull baseline hazard** and test whether the survival is the same in the two groups using a Wald test and a likelihood ratio test.

```
#fit a PH model with Weibull baseline hazard
```

```
weibull_model<-survreg(Surv(time,cens)~group,data=aml,dist="weibull")
```

```
summary(weibull_model)
```

```
##
## Call:
## survreg(formula = Surv(time, cens) ~ group, data = aml, dist = "weibull")
##           Value Std. Error      z      p
## (Intercept)  5.038      0.644  7.82 5.1e-15
## group        -0.929      0.383 -2.43  0.015
## Log(scale)   -0.235      0.178 -1.32  0.188
##
## Scale= 0.791
##
## Weibull distribution
## Loglik(model)= -80.5   Loglik(intercept only)= -83.2
##   Chisq= 5.31 on 1 degrees of freedom, p= 0.021
## Number of Newton-Raphson Iterations: 5
## n= 23

#wald test
wald_test_weibull<-weibull_model$coefficients[2]/0.383
#likelihood ratio test
lrt<-(-2)*(weibull_model$loglik[1]-weibull_model$loglik[2])
#significant difference in survival between the two groups
#at 5% significance level
#H0 is rejected at 5% significance level
```

## Question (f)

Give the Kaplan-Meier estimate of the survival of non maintenance group at 10 weeks. Provide the corresponding standard error for this estimate.

```
summary(fit)
```

```
## Call: survfit(formula = Surv(time, cens) ~ group, data = aml)
##
##           group=1
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     9      11        1   0.909  0.0867   0.7541      1.000
##    13       10        1   0.818  0.1163   0.6192      1.000
##    18        8        1   0.716  0.1397   0.4884      1.000
##    23        7        1   0.614  0.1526   0.3769      0.999
##    31        5        1   0.491  0.1642   0.2549      0.946
##    34        4        1   0.368  0.1627   0.1549      0.875
##    48        2        1   0.184  0.1535   0.0359      0.944
##
##           group=2
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     5      12        2   0.8333  0.1076   0.6470      1.000
##     8      10        2   0.6667  0.1361   0.4468      0.995
##    12       8        1   0.5833  0.1423   0.3616      0.941
##    23       6        1   0.4861  0.1481   0.2675      0.883
##    27       5        1   0.3889  0.1470   0.1854      0.816
##    30       4        1   0.2917  0.1387   0.1148      0.741
##    33       3        1   0.1944  0.1219   0.0569      0.664
##    43       2        1   0.0972  0.0919   0.0153      0.620
##    45       1        1   0.0000    NaN      NA      NA
```

```

#Estimated survival at 10 weeks in non maintenance group is given 0.6667
#Estimated standard error: 0.1361
#The estimates are the same as at 8 weeks
#Notice that the last complete remission by 10 weeks was experienced at 8 weeks.

```

## Question (g)

Test the null hypothesis that the survival at 10 weeks is the same for the two groups.

```

#construct the test
#maintenance group
surv_10week_grp1<-fit$surv[1]
var_10week_grp1<-(fit$std.err[1])^2
summary(fit)

## Call: survfit(formula = Surv(time, cens) ~ group, data = aml)
##
##               group=1
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    9      11       1   0.909  0.0867   0.7541      1.000
##   13      10       1   0.818  0.1163   0.6192      1.000
##   18       8       1   0.716  0.1397   0.4884      1.000
##   23       7       1   0.614  0.1526   0.3769      0.999
##   31       5       1   0.491  0.1642   0.2549      0.946
##   34       4       1   0.368  0.1627   0.1549      0.875
##   48       2       1   0.184  0.1535   0.0359      0.944
##
##               group=2
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    5      12       2   0.8333  0.1076   0.6470      1.000
##    8      10       2   0.6667  0.1361   0.4468      0.995
##   12       8       1   0.5833  0.1423   0.3616      0.941
##   23       6       1   0.4861  0.1481   0.2675      0.883
##   27       5       1   0.3889  0.1470   0.1854      0.816
##   30       4       1   0.2917  0.1387   0.1148      0.741
##   33       3       1   0.1944  0.1219   0.0569      0.664
##   43       2       1   0.0972  0.0919   0.0153      0.620
##   45       1       1   0.0000   NaN      NA      NA

#non maintenance group
surv_10week_grp2<-fit$surv[12]
var_10week_grp2<-(fit$std.err[12])^2
#test
var_10_week<-var_10week_grp1+var_10week_grp2
test_10week<-(surv_10week_grp1-surv_10week_grp2)/sqrt(var_10_week)
print(test_10week)

## [1] 1.076033

#decision
#compare your test with the quantile at 0.975 percentile from N(0,1)
test_10week<qnorm(0.975)

## [1] TRUE

```

```
#calculate the p-value of your test
p_value_test<-2*pnorm(test_10week,lower.tail = F)
0.05<p_value_test
```

```
## [1] TRUE
```

```
#survival distributions at 10 weeks is not significantly different in the two groups
```

### Exercise 3

Survival data on 26 psychiatric inpatients admitted to the University of Iowa hospitals during the years 1935 to 1948. This sample is part of a larger study of psychiatric inpatients discussed by Tsuang and Woolson (1977). Data for each patient consists of age at first admission to the hospital, sex, number of years of follow-up (years from admission to death or censoring) and patient status at the follow-up time. Ultimately we want to know if male psychiatric patients and female psychiatric patients are more likely to die than the general public. The data from the hospitals can be found in the package KMsurv (use data(psych) to obtain the data).

```
#load the data
data("psych")
```

(c) Compute the correct Kaplan–Meier estimate

```
#specify the age at the end of the follow-up
psych$ageatfu<-psych$age+psych$time
#compute the correct Kaplan-Meier estimate
fit<-survfit(Surv(age,ageatfu,death)~1,data=psych)
summary(fit)
```

```
## Call: survfit(formula = Surv(age, ageatfu, death) ~ 1, data = psych)
##
##   time n.risk n.event censored survival std.err lower 95% CI upper 95% CI
##   47      21      1      0    0.952  0.0465    0.866    1.000
##   50      22      1      0    0.909  0.0613    0.797    1.000
##   52      21      1      1    0.866  0.0721    0.735    1.000
##   57      21      2      1    0.783  0.0856    0.632    0.970
##   59      18      2      1    0.696  0.0957    0.532    0.912
##   61      16      2      2    0.609  0.1016    0.439    0.845
##   63      11      1      2    0.554  0.1064    0.380    0.807
##   67       8      1      2    0.485  0.1134    0.306    0.767
##   69       5      2      1    0.291  0.1261    0.124    0.680
##   76       1      1      2    0.000    NaN      NA      NA
```

(d) Estimate the **hazard ratio** for females versus males using a Cox Proportional Hazard model. Interpret the results.

```
#fit a PH Cox Regression model taking into account delayed entries
cox_model_del<-coxph(Surv(age,ageatfu,death)~sex,data=psych)
summary(cox_model_del)
```

```
## Call:
## coxph(formula = Surv(age, ageatfu, death) ~ sex, data = psych)
##
##   n= 26, number of events= 14
##
##           coef exp(coef) se(coef)      z Pr(>|z|)
## sex 0.3900    1.4770    0.6102 0.639    0.523
```



```
##
##      exp(coef) exp(-coef) lower .95 upper .95
## sex      1.477      0.677      0.4466      4.884
##
## Concordance= 0.58 (se = 0.075 )
## Likelihood ratio test= 0.43 on 1 df,  p=0.5
## Wald test              = 0.41 on 1 df,  p=0.5
## Score (logrank) test = 0.41 on 1 df,  p=0.5
# Hazard Ratio for females vs males
exp(cox_model_del$coefficients)
```

```
##      sex
## 1.477014
```

(e) Check the difference with a Cox proportional hazard model fitted to the data ignoring delayed entry.

```
#fit a PH Cox Regression model ignoring delayed entries
cox_model_nodel<-coxph(Surv(ageatfu,death)~sex,data=psych)
summary(cox_model_nodel)
```

```
## Call:
## coxph(formula = Surv(ageatfu, death) ~ sex, data = psych)
##
##      n= 26, number of events= 14
##
##      coef exp(coef) se(coef)      z Pr(>|z|)
## sex 0.3091      1.3621      0.6082 0.508      0.611
##
##      exp(coef) exp(-coef) lower .95 upper .95
## sex      1.362      0.7341      0.4136      4.486
##
## Concordance= 0.562 (se = 0.075 )
## Likelihood ratio test= 0.27 on 1 df,  p=0.6
## Wald test              = 0.26 on 1 df,  p=0.6
## Score (logrank) test = 0.26 on 1 df,  p=0.6
#compare the two models using the AIC
AIC(cox_model_nodel,cox_model_del)
```

```
##      df      AIC
## cox_model_nodel  1 70.58342
## cox_model_del    1 69.36947
```