Survival Analysis Week 1

Petteri Mäntymaa

March 18, 2019

Estimation of survival function

Load the data set *Veterans administration lung cancer trial, cf. Kalbfleisch and Prentice, 2002* from the R `survival` package:

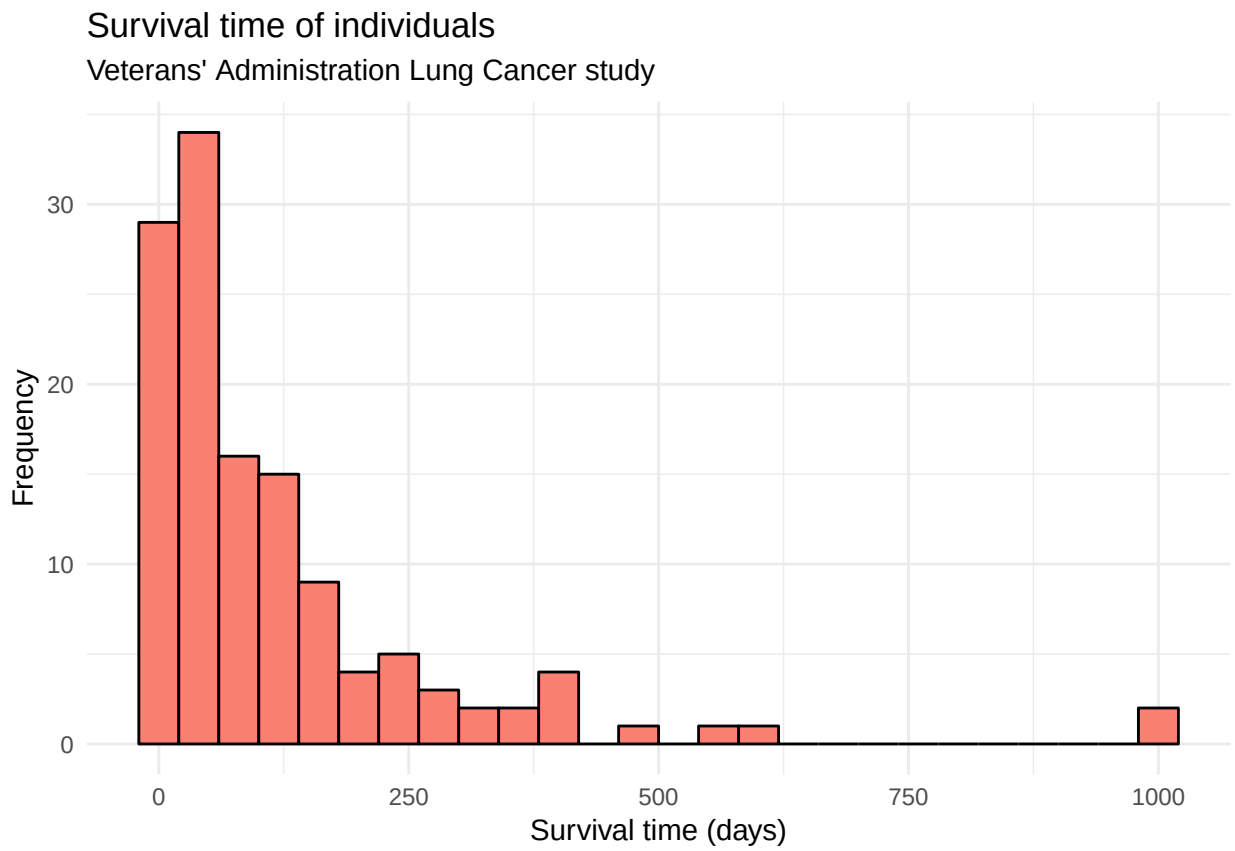
```
# Load and inspect the data
data(veteran)
str(veteran)
```

```
## 'data.frame':    137 obs. of  8 variables:
##  $ trt     : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ celltype: Factor w/ 4 levels "squamous","smallcell",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ time    : num  72 411 228 126 118 10 82 110 314 100 ...
##  $ status  : num  1 1 1 1 1 1 1 1 1 0 ...
##  $ karno   : num  60 70 60 60 70 20 40 80 50 70 ...
##  $ diagtime: num  7 5 3 9 11 5 10 29 18 6 ...
##  $ age     : num  69 64 38 63 65 49 69 68 43 70 ...
##  $ prior   : num  0 10 0 10 10 0 10 0 0 0 ...
```

1.

Plot a histogram of the survival times corresponding to uncensored observations (veteran$status == 1).

```
veteran %>%
  filter(status == 1) %>%
  ggplot() +
  aes(x = time) +
  geom_histogram(binwidth = 40, color = "black", fill = "salmon") +
  theme_minimal() +
  labs(title = "Survival time of individuals", subtitle = "Veterans' Administration Lung Cancer study") +
  xlab("Survival time (days)") +
  ylab("Frequency")
```



2.

Create an output file where the histogram is stored.

Did this, even though the plot is produced above

```
png("survivaltimes.png")
vet_surv <- veteran %>%
  filter(status == 1) %>%
  ggplot() +
  aes(x = time) +
  geom_histogram(binwidth = 40, color = "black", fill = "salmon") +
  theme_minimal() +
  labs(title = "Survival time of individuals", subtitle = "Veterans' Administration Lung Cancer study") +
  xlab("Survival time (days)") +
  ylab("Frequency")
print(vet_surv)
dev.off()
```

3.

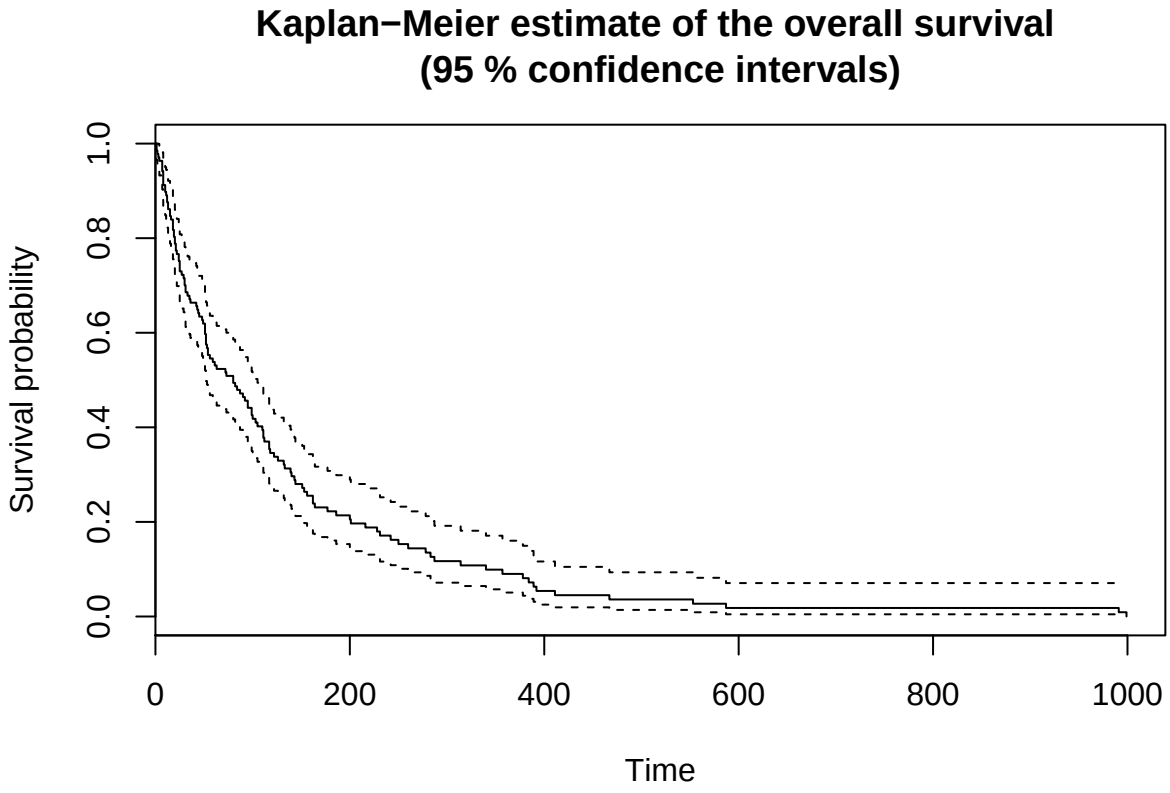
a.

Use the `survfit` routine in R to calculate the Kaplan-Meier estimate of the overall survival in the data.

In the survival routines of R, the response variable needs to be specified as a survival object. If the observed failure time variable is time and failure indicator variable is status, the response variable is created as `Surv(time, status)`

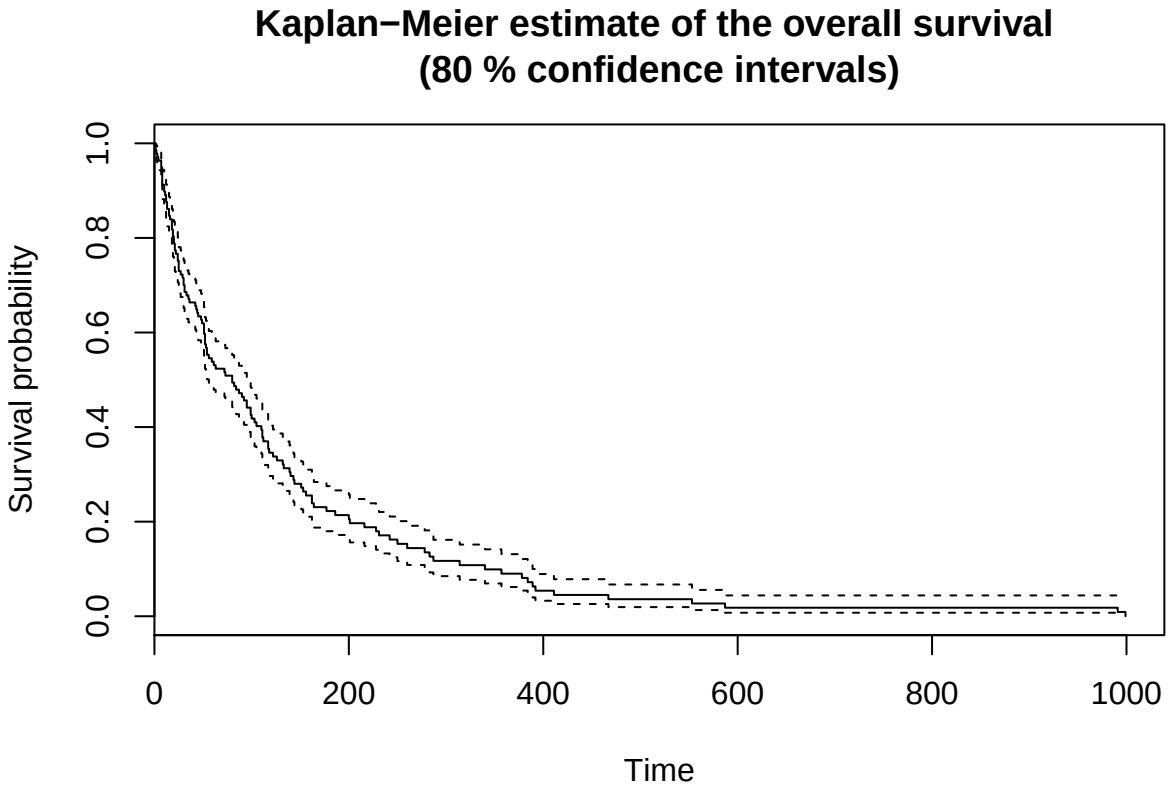
Applying the plot command to the output object from the `survfit` routine, you can draw the estimate and its confidence limits.

```
fit <- survfit(Surv(time, status) ~ 1, data = veteran)
plot(fit, xlab="Time", ylab="Survival probability", main = "Kaplan-Meier estimate of the overall surviva
```



Experiment with different confidence levels (e.g. 80% and 95%). You can also practice with the plot command options (e.g. `xlab`, `ylab`).

```
fit_twenty <- survfit(Surv(time, status) ~ 1, data = veteran, conf.int = 0.8)
plot(fit_twenty, xlab="Time", ylab="Survival probability", main = "Kaplan-Meier estimate of the overall
```

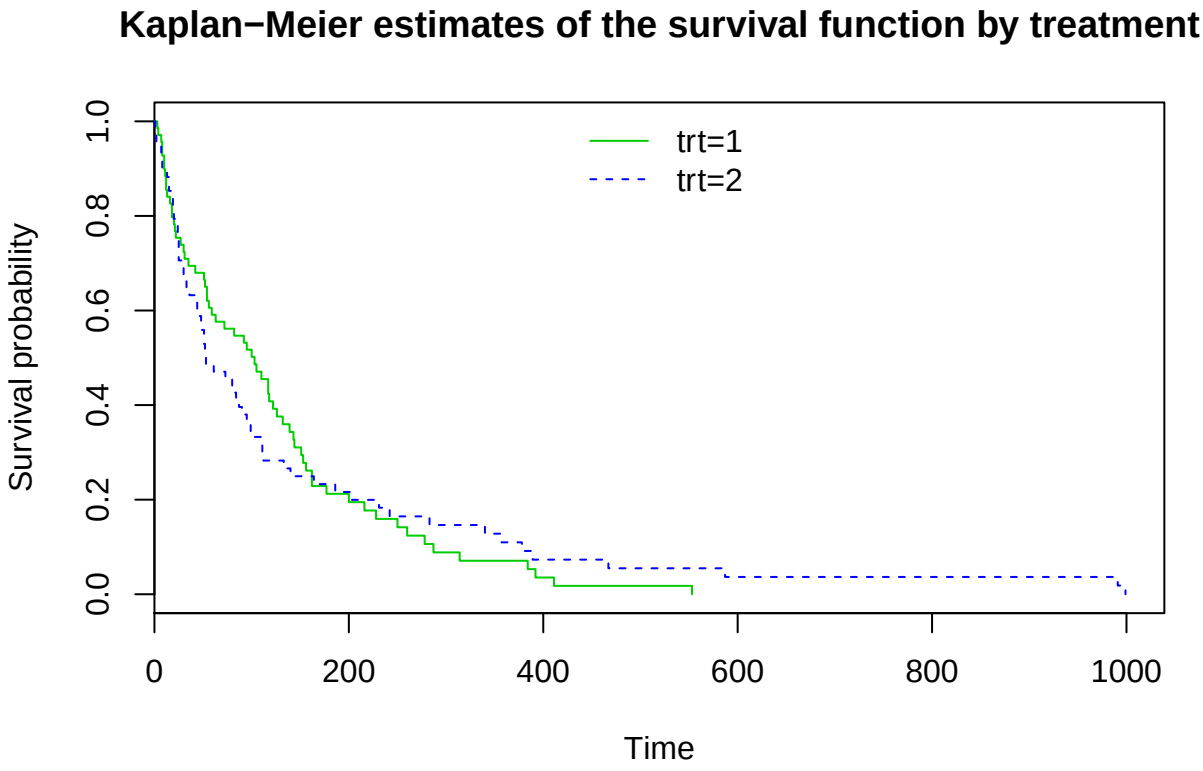


b.

Plot the Kaplan-Meier estimates of the survival functions separately for the two treatment groups (standard vs. test).

Does there appear to be a difference between the two groups in survival?

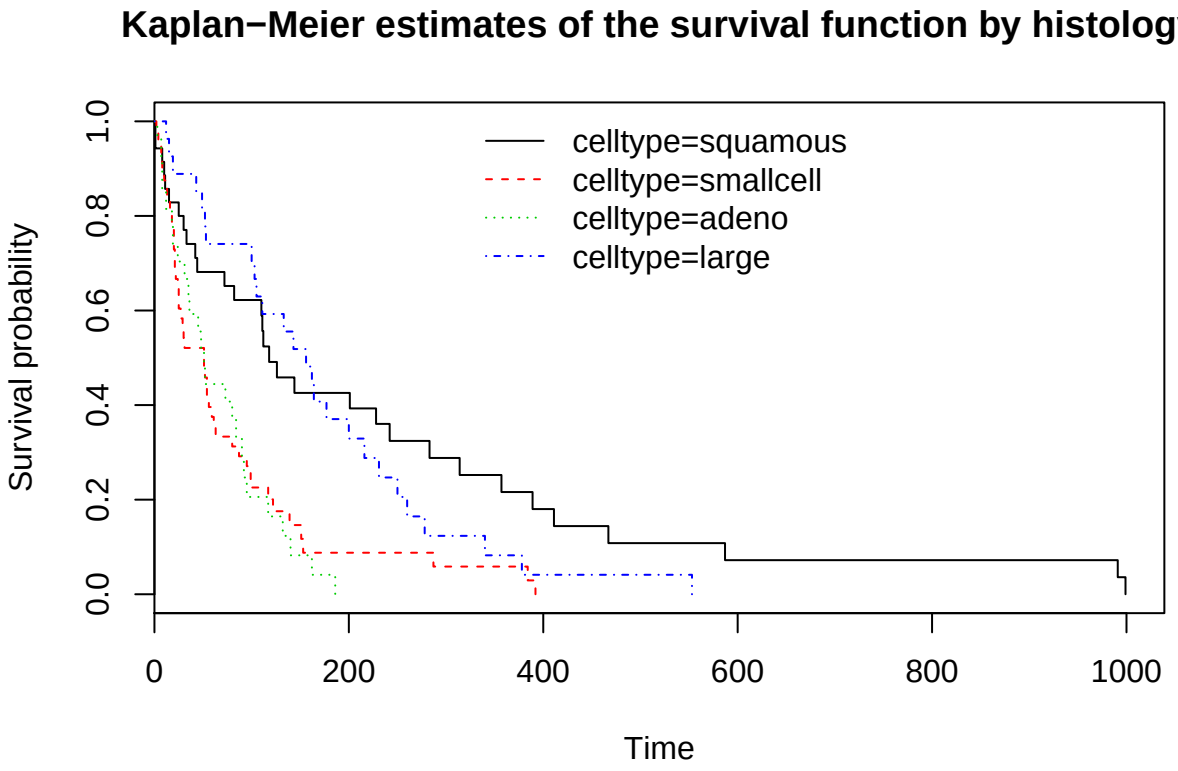
```
fit_treatment <- survfit(Surv(time, status) ~ trt, data = veteran)
{plot(fit_treatment, xlab="Time", ylab="Survival probability", col = 3:4, lty=1:2, main = "Kaplan-Meier
lL <- gsub("x=","",names(fit_treatment$strata))
legend(
  "top",
  legend=lL,
  col=3:4,
  lty=1:2,
  horiz=FALSE,
  bty='n')
}
```



It is very hard to judge if there is a difference in survival probabilities between the treatment groups. Survival probability decreases more sharply from the beginning of follow-up but evens out slightly on $Time > 200$.

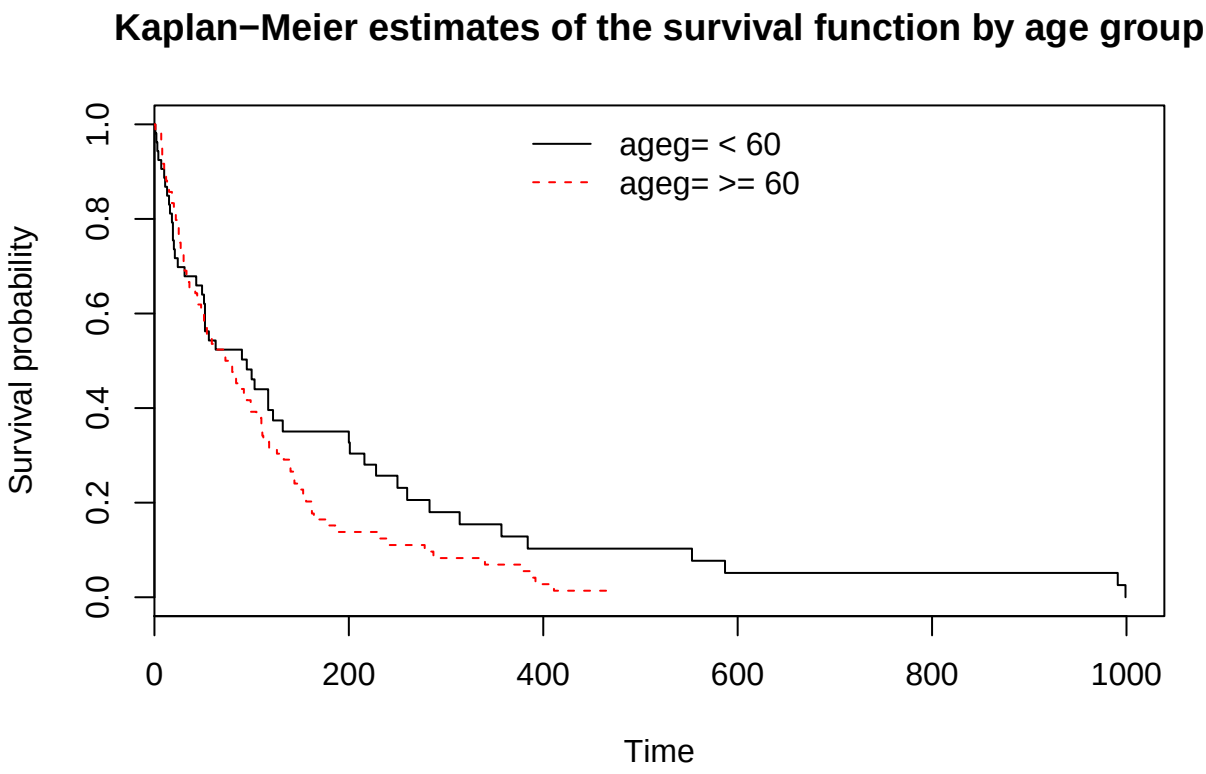
Irrespective of the treatment group, compare the survival in groups defined by the histological type of tumor (variable celltype). You may also like to explore the effect on survival of the other covariates in the data.

```
fit_hist <- survfit(Surv(time, status) ~ celltype, data = veteran)
{plot(fit_hist, xlab="Time", ylab="Survival probability",lty = 1:4, col = 1:4, main = "Kaplan-Meier esti
lLab <- gsub("x=","",names(fit_hist$strata))  ## legend labels
legend(
  "top",
  legend=lLab,
  col=1:4,
  lty=1:4,
  horiz=FALSE,
  bty='n')
}
```



There certainly seems to be a significant difference in survival probabilities especially between squamous and adeno.

```
veteran$ageg <- cut(veteran$age, breaks = c(0,60,100), labels = c(" < 60", " >= 60"), right = F)
fit_age <- survfit(Surv(time, status) ~ ageg, data = veteran)
{plot(fit_age, xlab="Time", ylab="Survival probability",lty = 1:4, col = 1:4, main = "Kaplan-Meier esti
lLab <- gsub("x=","",names(fit_age$strata))  ## legend labels
legend(
  "top",
  legend=lLab,
  col=1:4,
  lty=1:4,
  horiz=FALSE,
  bty='n')
}
```



c.

Compare the two treatments by the log-rank test. You can find this in the `survdiff` routine.

```
diff_treatment <- survdiff(Surv(time, status) ~ trt, data = veteran)
print(diff_treatment)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ trt, data = veteran)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=1 69       64     64.5   0.00388   0.00823
## trt=2 68       64     63.5   0.00394   0.00823
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.9
```

There does not seem to be any significant difference between treatment groups.

Compare then the effect of celltype on survival.

```
diff_cyto <- survdiff(Surv(time, status) ~ celltype, data = veteran)
print(diff_cyto)
```

```
## Call:
```

```
## survdiff(formula = Surv(time, status) ~ celltype, data = veteran)
##
##                      N Observed Expected (O-E)^2/E (O-E)^2/V
## celltype=squamous   35       31     47.7      5.82     10.53
## celltype=smallcell  48       45     30.1      7.37     10.20
## celltype=adeno      27       26     15.7      6.77      8.19
## celltype=large      27       26     34.5      2.12      3.02
##
##  Chisq= 25.4  on 3 degrees of freedom, p= 1e-05
```

As per our preliminary "hunch", there indeed seems to be very significant differerence between the histologies.

4.

Data matrix cervix contains grouped survival data for two cohorts of women, diagnosed with stage I or stage II cervix cancer.

Use the `lifetab` routine in R library `KMsurv` to create life tables for both groups.

Life table (stage 1)

```
cervix <- read.csv("data/cervix.dat", sep = ";")

tis_a <- c(cervix$year[cervix$stage == 1],NA)
ninit_a <- cervix$N[cervix$stage == 1][1]
nlost_a <- cervix$nlost[cervix$stage == 1]
nevent_a <- cervix$nfailure[cervix$stage == 1]

lt_a <- lifetab(tis_a, ninit_a, nlost_a, nevent_a)
lt_a
```

```
##       nsubs nlost nrisk nevent      surv        pdf     hazard    se.surv
## 1-2     110     5 107.5      5 1.0000000 0.04651163 0.04761905 0.00000000
## 2-3     100     7  96.5      7 0.9534884 0.06916496 0.07526882 0.02031114
## 3-4      86     7  82.5      7 0.8843234 0.07503350 0.08860759 0.03144341
## 4-5      72     8  68.0      3 0.8092899 0.03570397 0.04511278 0.03954839
## 5-6      61     7  57.5      0 0.7735859 0.00000000 0.00000000 0.04284029
## 6-7      54    10  49.0      2 0.7735859 0.03157494 0.04166667 0.04284029
## 7-8      42     6  39.0      3 0.7420110 0.05707777 0.08000000 0.04654751
## 8-9      33     5  30.5      0 0.6849332 0.00000000 0.00000000 0.05337208
## 9-10     28     4  26.0      0 0.6849332 0.00000000 0.00000000 0.05337208
## 10-NA    24     8  20.0      1 0.6849332         NA         NA 0.05337208
##           se.pdf  se.hazard
## 1-2   0.02031114 0.02128985
## 2-3   0.02521897 0.02842878
## 3-4   0.02726104 0.03345764
## 4-5   0.02022924 0.02603925
## 5-6          NaN        NaN
## 6-7   0.02193626 0.02945639
## 7-8   0.03186287 0.04615106
## 8-9          NaN        NaN
## 9-10         NaN        NaN
## 10-NA         NA         NA
```

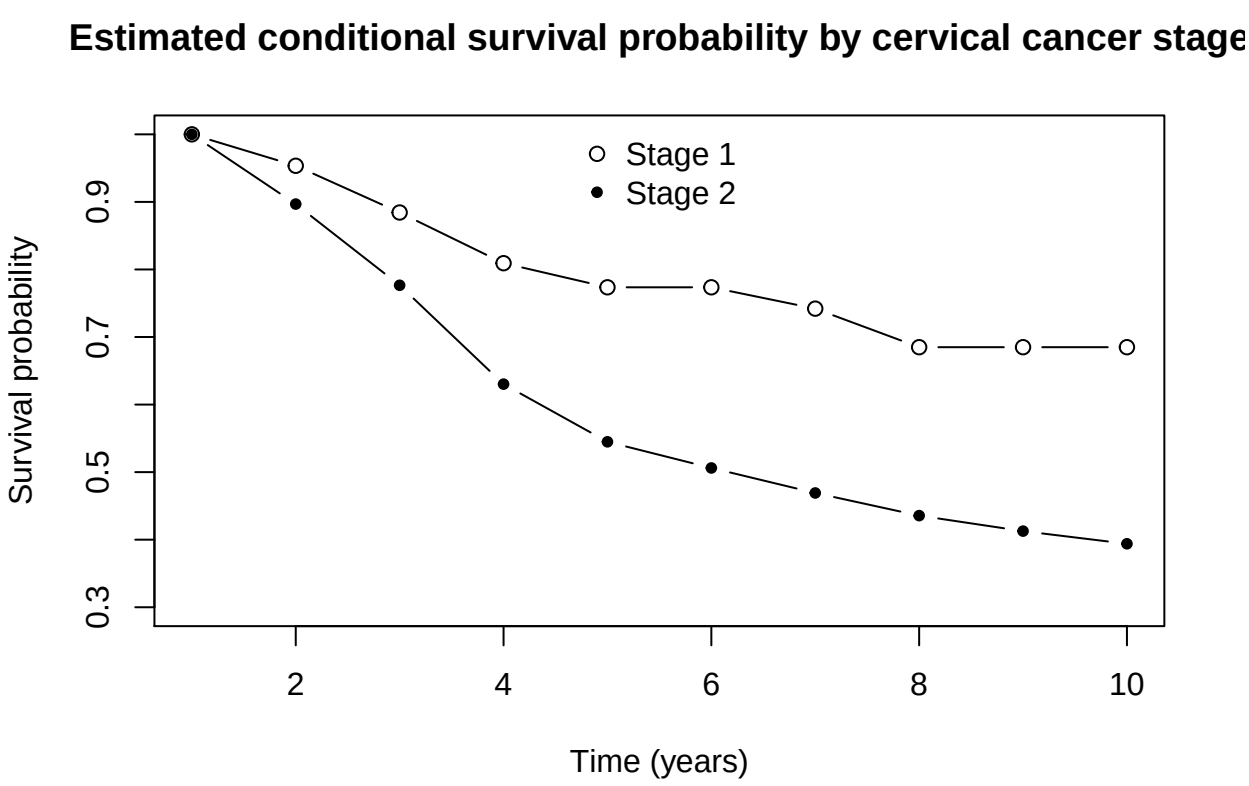
Life table (stage 2)

```
tis_b <- c(cervix$year[cervix$stage == 2],NA)
ninit_b <- cervix$N[cervix$stage == 2][1]
nlost_b <- cervix$nlost[cervix$stage == 2]
nevent_b <- cervix$nfailure[cervix$stage == 2]

lt_b <- lifetab(tis_b, ninit_b, nlost_b, nevent_b)
lt_b
```

```
##       nsubs nlost nrisk nevent      surv        pdf     hazard    se.surv
## 1-2     234     3 232.5     24 1.0000000 0.10322581 0.10884354 0.00000000
## 2-3     207    11 201.5     27 0.8967742 0.12016329 0.14361702 0.01995374
## 3-4     169     9 164.5     31 0.7766109 0.14635221 0.20805369 0.02759940
## 4-5     129     7 125.5     17 0.6302587 0.08537369 0.14529915 0.03259466
## 5-6     105    13  98.5      7 0.5448850 0.03872279 0.07368421 0.03412842
## 6-7      85     6  82.0      6 0.5061622 0.03703626 0.07594937 0.03469969
## 7-8      73     6  70.0      5 0.4691260 0.03350900 0.07407407 0.03530150
## 8-9      62    10  57.0      3 0.4356170 0.02292721 0.05405405 0.03581977
## 9-10     49    10  44.0      2 0.4126898 0.01875863 0.04651163 0.03629805
## 10-NA    37     6  34.0      4 0.3939311         NA         NA 0.03699242
##           se.pdf  se.hazard
## 1-2   0.01995374 0.02218467
## 2-3   0.02168584 0.02756776
## 3-4   0.02424416 0.03716481
## 4-5   0.01975252 0.03514710
## 5-6   0.01431319 0.02783111
## 6-7   0.01477609 0.03098383
## 7-8   0.01465906 0.03310420
## 8-9   0.01302118 0.03119672
## 9-10  0.01306399 0.03287979
## 10-NA         NA         NA
```

```
{plot(1:10, lt_a$surv, type = "b", pch = 21, ylim = c(0.3,1), xlab = "Time (years)", ylab = "Survival p
lines(1:10, lt_b$surv, type = "b", pch = 20)
leg <- c("Stage 1", "Stage 2")  ## legend labels
legend(
  "top",
  legend=leg,
  pch = c(21,20),
  horiz=FALSE,
  bty='n')
}
```



By comparing the conditional survival probabilities we can see the (anticipated) difference of survival probabilities between the cervical cancer stages.