THE ANALYSIS FACTOR

Survival Analysis Workshop

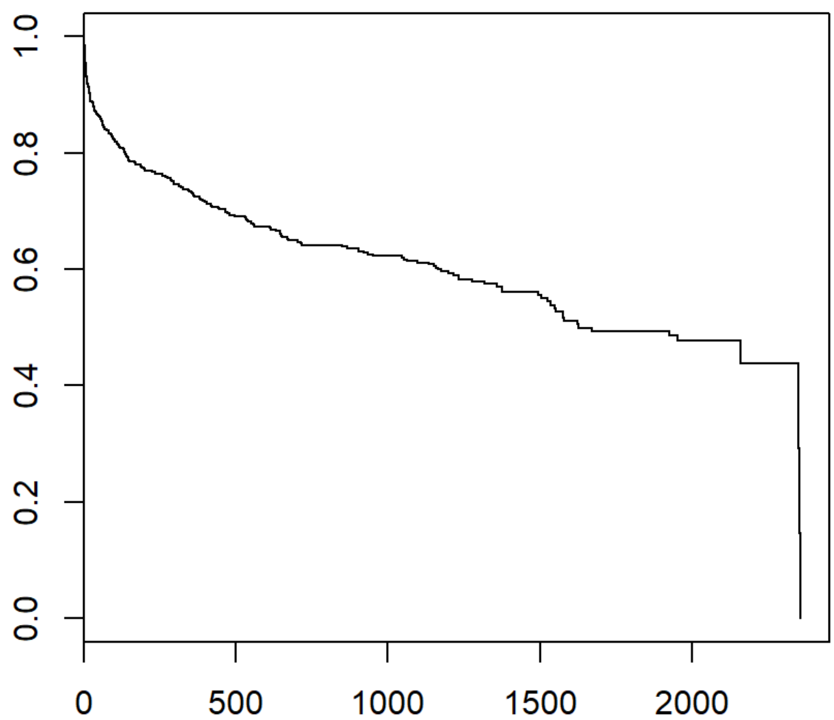
Exercises: Module 1

1. **Open the WHAS500 data set in the software program of your choice.**

a. Produce a table of counts for fstat, to indicate which patients have died and which have been censored.

```
##   id age gender hr sysbp diasbp      bmi cvd afb sho chf av3 miord mitype
## 1  1  83      0 89   152     78 25.54051   1   1   0   0   0     1      0
## 2  2  49      0 84   120     60 24.02398   1   0   0   0   0     0      1
## 3  3  70      1 83   147     88 22.14290   0   0   0   0   0     0      1
## 4  4  70      0 65   123     76 26.63187   1   0   0   1   0     0      1
## 5  5  70      0 63   135     85 24.41255   1   0   0   0   0     0      1
## 6  6  70      0 76    83     54 23.24236   1   0   0   0   1     0      0
##   year  admitdate    disdate      fdate los dstat lenfol fstat
## 1    1 01/13/1997 01/18/1997 12/31/2002   5     0   2178     0
## 2    1 01/19/1997 01/24/1997 12/31/2002   5     0   2172     0
## 3    1 01/01/1997 01/06/1997 12/31/2002   5     0   2190     0
## 4    1 02/17/1997 02/27/1997 12/11/1997  10     0    297     1
## 5    1 03/01/1997 03/07/1997 12/31/2002   6     0   2131     0
## 6    1 03/11/1997 03/12/1997 03/12/1997   1     1      1     1
```

b. Draw a Kaplan-Meier plot for overall survival.



c. Estimate the 25th, 50th, and 75th quantiles for overall survival.

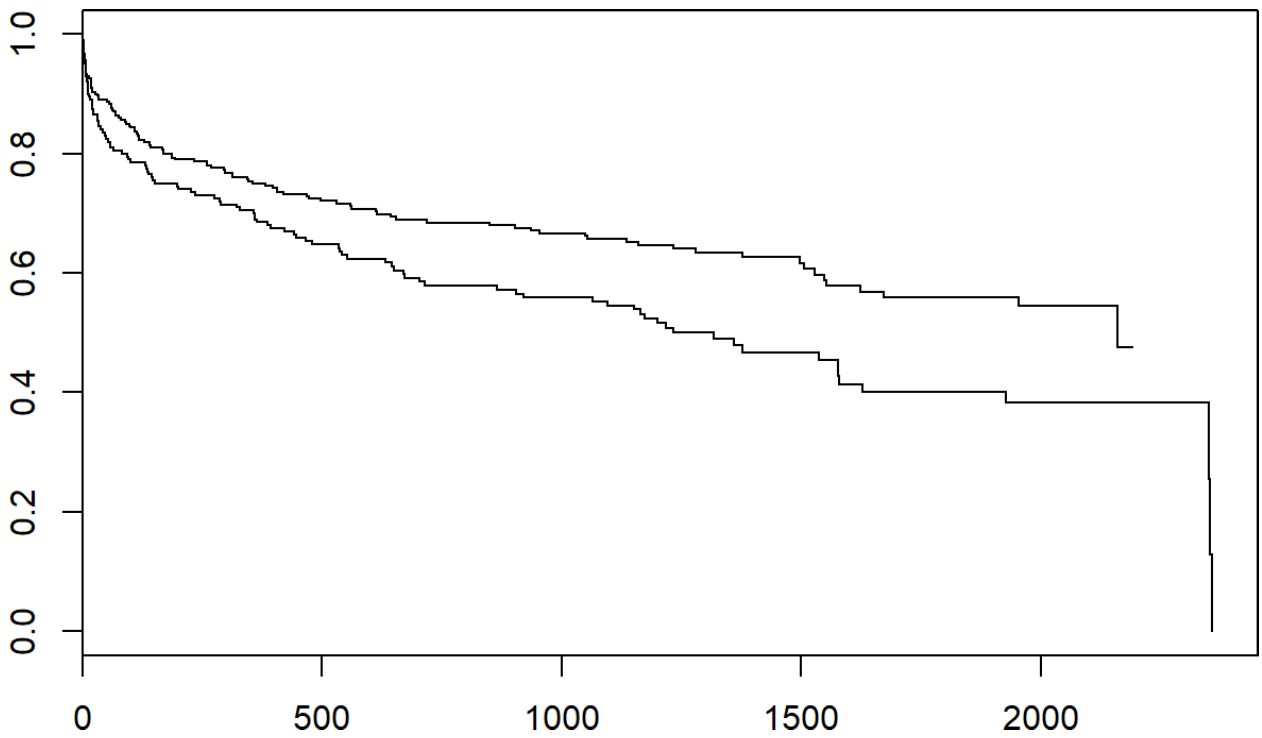
```
## $quantile
##   25   50   75
##  296 1627 2353
##
## $lower
##   25   50   75
##  166 1527 2350
##
## $upper
##  25  50  75
## 422  NA  NA
```

2. **Use the WHAS500 data set for this problem.**

a. Produce a crosstabulation of fstat and gender. Are you comfortable with the number of deaths in each group?

```
##
##       0   1
##   0 189 111
##   1  96 104
```

b. Draw Kaplan-Meier curves for males and females.



c. Calculate median survival with confidence intervals for males and females.

```
## $quantile
##                   25   50   75
## whas500$gender=0 368 2160   NA
## whas500$gender=1 174 1317 2353
```

```
##
## $lower
##                    25   50   75
## whas500$gender=0 187 1671   NA
## whas500$gender=1  83  905 2350
##
## $upper
##                    25   50 75
## whas500$gender=0 644   NA NA
## whas500$gender=1 385 1627 NA
```

d. Calculate the log rank test for males versus females. Interpret your result.

```
## Call:
## survdiff(formula = whas500_surv ~ whas500$gender)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## whas500$gender=0 300      111    130.7      2.98      7.79
## whas500$gender=1 200      104     84.3      4.62      7.79
##
##  Chisq= 7.8  on 1 degrees of freedom, p= 0.00525
```

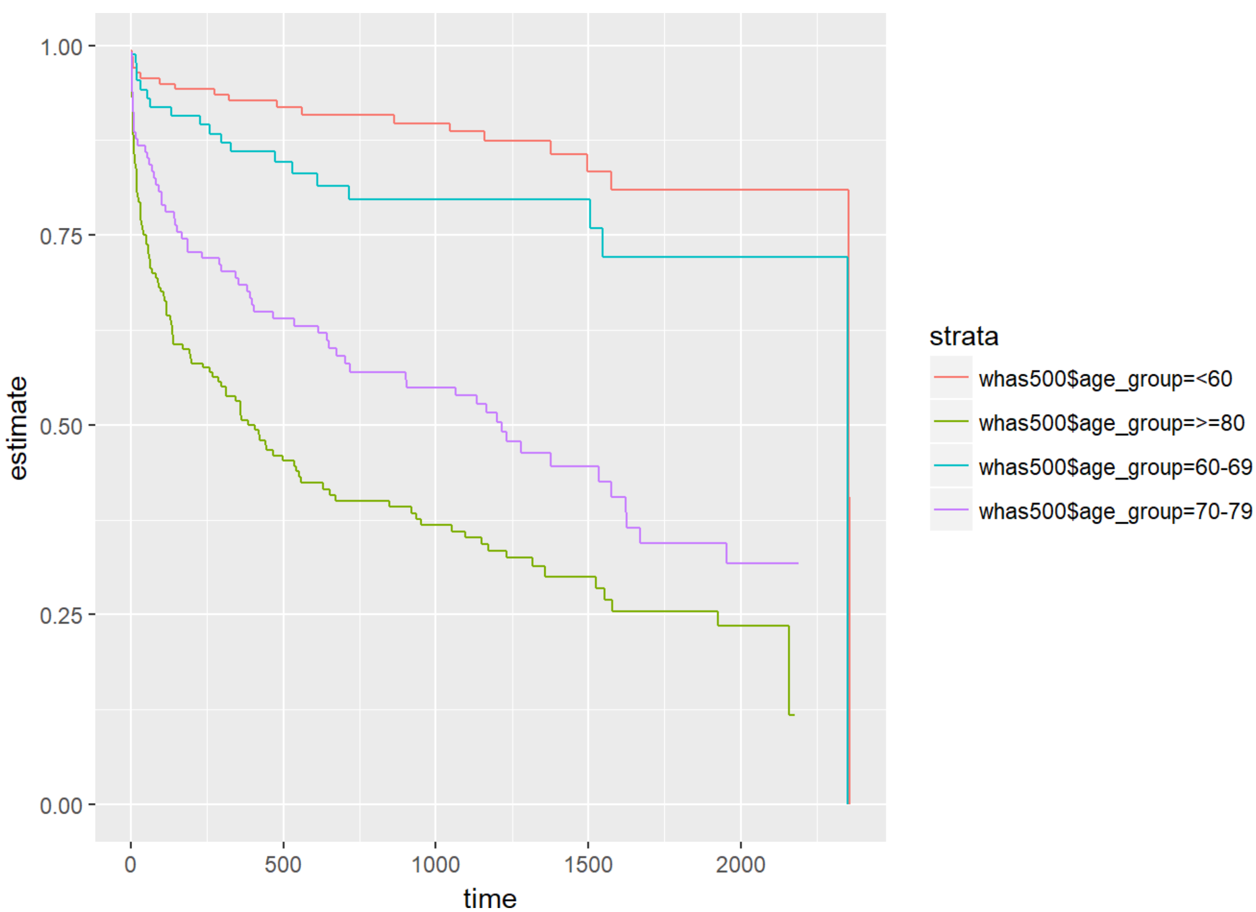
Interpretation: the p-value is less than 0.05, so you would reject the null hypothesis and conclude that the survival probabilities are different for women versus men.

3. Use the WHAS500 data set for this problem.

a. Produce age groups <60, 60-69, 70-79, and >=80. Compute a crosstabulation of this variable with fstat. Are you comfortable with the number of deaths in each group?

```
##
##              0    1
##    <60     118   20
##    60-69    67   19
##    70-79    50   64
##    >=80     50  110
```

b. Draw Kaplan Meier curves for each age group.



c. Calculate the median survival time with confidence intervals for each age group.

```
## $quantile
##                              25   50   75
## whas500$age_group=<60   2353.0 2353 2358
## whas500$age_group=60-69 1548.0 2350 2350
## whas500$age_group=70-79  166.0 1217   NA
## whas500$age_group=>=80    45.5  385 1926
##
## $lower
##                            25   50   75
## whas500$age_group=<60   1577 2353 2353
## whas500$age_group=60-69  612   NA   NA
## whas500$age_group=70-79   81  704 1671
## whas500$age_group=>=80    20  259 1317
##
## $upper
##                           25   50 75
## whas500$age_group=<60     NA   NA NA
## whas500$age_group=60-69   NA   NA NA
## whas500$age_group=70-79 405 1627 NA
## whas500$age_group=>=80  108  654 NA
```

d. Calculate the log rank test for age groups. Interpret your results.

```
## Call:
## survdiff(formula = whas500_surv ~ whas500$age_group)
##
## n=498, 2 observations deleted due to missingness.
##
##                           N Observed Expected (O-E)^2/E (O-E)^2/V
## whas500$age_group=<60   138       20     71.5     37.08     57.45
## whas500$age_group=60-69  86       19     41.1     11.86     14.78
## whas500$age_group=70-79 114       64     47.2      5.97      7.73
## whas500$age_group=>=80  160      110     53.2     60.52     82.07
```

```
##
##  Chisq= 118  on 3 degrees of freedom, p= 0
```

4. **(Only for those who are brave) The following are times for catheters in infants. A "+" means that the catheter was removed because it was no longer needed. Times without a + mean that the catheter was removed because it failed. Occlusion and infection were the two major reasons for failure. Treating failures as an event and removal because it was no longer needed as a censored observation, estimate the Kaplan-Meier survival curve by hand, showing all your intermediate calculations.**

1+, 1+, 1+, 1+, 1+, 1+, 1+, 1+, 1, 1, 2+, 2+, 2, 2, 3+, 3, 4+, 4, 5+, 5+, 5+, 5+, 5+, 5, 5, 5, 6, 6, 7, 10, 10, 12, 12, 13

First, set up a data frame for the ten time points (1-7, 10, 12, 13)

```
km <- data.frame(
  t=c(1:7, 10, 12, 13),
  n=rep(-1, 10),
  d=rep(-1, 10),
  c=rep(-1, 10),
  p=rep(-1, 10),
  s=rep(-1, 10))
```

We will fill in these numbers soon enough. Here's the key: t = time n = number at risk d = number of failures c = number censored p = conditional probability s = survival probability

1+, 1+, 1+, 1+, 1+, 1+, 1+, 1+, 1, 1, 2+, 2+, 2, 2, 3+, 3, 4+, 4, 5+, 5+, 5+, 5+, 5+, 5, 5, 5, 6, 6, 7, 10, 10, 12, 12, 13

There are 34 observations total. At t1=1, there are d1=2 deaths and c1=8 ceonsored values. The survival probability is equal to the conditional probability.

```
km$n[1] <- 34
km$d[1] <-  2
km$c[1] <-  8
km$p[1] <- 1-km$d[1]/km$n[1]
km$s[1] <- km$p[1]
km
```

```
##     t  n  d  c          p          s
## 1   1 34  2  8  0.9411765  0.9411765
## 2   2 -1 -1 -1 -1.0000000 -1.0000000
```

```
## 3    3 -1 -1 -1 -1.0000000 -1.0000000
## 4    4 -1 -1 -1 -1.0000000 -1.0000000
## 5    5 -1 -1 -1 -1.0000000 -1.0000000
## 6    6 -1 -1 -1 -1.0000000 -1.0000000
## 7    7 -1 -1 -1 -1.0000000 -1.0000000
## 8   10 -1 -1 -1 -1.0000000 -1.0000000
## 9   12 -1 -1 -1 -1.0000000 -1.0000000
## 10 13 -1 -1 -1 -1.0000000 -1.0000000
```

1+, 1+, 1+, 1+, 1+, 1+, 1+, 1+, 1, 1, **2+, 2+, 2, 2,** 3+, 3, 4+, 4, 5+, 5+, 5+, 5+, 5+, 5, 5, 5, 6, 6, 7, 10, 10, 12, 12, 13

At t2=2, there are n2=24 at risk, and you have d2=2 failures and c2=2 censored observations. The survival probability is equal to the product of the first two conditional probabilities.

```
km$n[2] <- 24
km$d[2] <-  2
km$c[2] <-  2
km$p[2] <- 1-km$d[2]/km$n[2]
km$s[2] <- km$p[1]*km$p[2]
km
```

```
##      t  n  d  c          p          s
## 1    1 34  2  8  0.9411765  0.9411765
## 2    2 24  2  2  0.9166667  0.8627451
## 3    3 -1 -1 -1 -1.0000000 -1.0000000
## 4    4 -1 -1 -1 -1.0000000 -1.0000000
## 5    5 -1 -1 -1 -1.0000000 -1.0000000
## 6    6 -1 -1 -1 -1.0000000 -1.0000000
## 7    7 -1 -1 -1 -1.0000000 -1.0000000
## 8   10 -1 -1 -1 -1.0000000 -1.0000000
## 9   12 -1 -1 -1 -1.0000000 -1.0000000
## 10 13 -1 -1 -1 -1.0000000 -1.0000000
```

1+, 1+, 1+, 1+, 1+, 1+, 1+, 1+, 1, 1, 2+, 2+, 2, 2, **3+, 3,** 4+, 4, 5+, 5+, 5+, 5+, 5+, 5, 5, 5, 6, 6, 7, 10, 10, 12, 12, 13

At t3=3, there are n3=20 at risk, and you have d3=1 failure and c3=1 censored observation. The survival probability is equal to the product of the first three conditional probabilities.

```
km$n[3] <- 20
km$d[3] <-  1
```

```
km$c[3] <-  1
km$p[3] <- 1-km$d[3]/km$n[3]
km$s[3] <- km$p[1]*km$p[2]*km$p[3]
km
```

```
##     t  n  d  c          p          s
## 1   1 34  2  8  0.9411765  0.9411765
## 2   2 24  2  2  0.9166667  0.8627451
## 3   3 20  1  1  0.9500000  0.8196078
## 4   4 -1 -1 -1 -1.0000000 -1.0000000
## 5   5 -1 -1 -1 -1.0000000 -1.0000000
## 6   6 -1 -1 -1 -1.0000000 -1.0000000
## 7   7 -1 -1 -1 -1.0000000 -1.0000000
## 8  10 -1 -1 -1 -1.0000000 -1.0000000
## 9  12 -1 -1 -1 -1.0000000 -1.0000000
## 10 13 -1 -1 -1 -1.0000000 -1.0000000
```

1+, 1+, 1+, 1+, 1+, 1+, 1+, 1+, 1, 1, 2+, 2+, 2, 2, 3+, 3, **4+, 4,** 5+, 5+, 5+, 5+, 5+, 5, 5, 5, 6, 6, 7, 10, 10, 12, 12, 13

At t4=4, there are n4=18 at risk, and you have d4=1 failure and c4=1 censored observations. The survival probability is equal to the product of the first four conditional probabilities.

```
km$n[4] <- 18
km$d[4] <-  1
km$c[4] <-  1
km$p[4] <- 1-km$d[4]/km$n[4]
km$s[4] <- km$p[1]*km$p[2]*km$p[3]*km$p[4]
km
```

```
##     t  n  d  c          p          s
## 1   1 34  2  8  0.9411765  0.9411765
## 2   2 24  2  2  0.9166667  0.8627451
## 3   3 20  1  1  0.9500000  0.8196078
## 4   4 18  1  1  0.9444444  0.7740741
## 5   5 -1 -1 -1 -1.0000000 -1.0000000
## 6   6 -1 -1 -1 -1.0000000 -1.0000000
## 7   7 -1 -1 -1 -1.0000000 -1.0000000
## 8  10 -1 -1 -1 -1.0000000 -1.0000000
## 9  12 -1 -1 -1 -1.0000000 -1.0000000
```

```
## 10 13 -1 -1 -1 -1.0000000 -1.0000000
```

1+, 1+, 1+, 1+, 1+, 1+, 1+, 1+, 1, 1, 2+, 2+, 2, 2, 3+, 3, 4+, 4, **5+, 5+, 5+, 5+, 5+, 5, 5, 5,** 6, 6, 7, 10, 10, 12, 12, 13

At t5=5, there are n5=16 at risk, and you have d5=3 failures and c5=5 censored observations. The survival probability is equal to the product of the first five conditional probabilities.

```
km$n[5] <- 16
km$d[5] <-  3
km$c[5] <-  5
km$p[5] <- 1-km$d[5]/km$n[5]
km$s[5] <- km$p[1]*km$p[2]*km$p[3]*km$p[4]*km$p[5]
km
```

```
##      t  n  d  c          p          s
## 1    1 34  2  8  0.9411765  0.9411765
## 2    2 24  2  2  0.9166667  0.8627451
## 3    3 20  1  1  0.9500000  0.8196078
## 4    4 18  1  1  0.9444444  0.7740741
## 5    5 16  3  5  0.8125000  0.6289352
## 6    6 -1 -1 -1 -1.0000000 -1.0000000
## 7    7 -1 -1 -1 -1.0000000 -1.0000000
## 8   10 -1 -1 -1 -1.0000000 -1.0000000
## 9   12 -1 -1 -1 -1.0000000 -1.0000000
## 10  13 -1 -1 -1 -1.0000000 -1.0000000
```

1+, 1+, 1+, 1+, 1+, 1+, 1+, 1+, 1, 1, 2+, 2+, 2, 2, 3+, 3, 4+, 4, 5+, 5+, 5+, 5+, 5+, 5, 5, 5, **6, 6,** 7, 10, 10, 12, 12, 13

At t6=6, there are n6=8 at risk, and you have d6=2 failures and c6=0 censored observations. The survival probability is equal to the product of the first six conditional probabilities.

```
km$n[6] <-  8
km$d[6] <-  2
km$c[6] <-  0
km$p[6] <- 1-km$d[6]/km$n[6]
km$s[6] <- km$p[1]*km$p[2]*km$p[3]*km$p[4]*km$p[5]*km$p[6]
km
```

```
##      t  n  d  c          p          s
## 1    1 34  2  8  0.9411765  0.9411765
## 2    2 24  2  2  0.9166667  0.8627451
```

```
## 3    3 20  1  1  0.9500000  0.8196078
## 4    4 18  1  1  0.9444444  0.7740741
## 5    5 16  3  5  0.8125000  0.6289352
## 6    6  8  2  0  0.7500000  0.4717014
## 7    7 -1 -1 -1 -1.0000000 -1.0000000
## 8   10 -1 -1 -1 -1.0000000 -1.0000000
## 9   12 -1 -1 -1 -1.0000000 -1.0000000
## 10  13 -1 -1 -1 -1.0000000 -1.0000000
```

1+, 1+, 1+, 1+, 1+, 1+, 1+, 1+, 1, 1, 2+, 2+, 2, 2, 3+, 3, 4+, 4, 5+, 5+, 5+, 5+, 5+, 5, 5, 5, 6, 6, **7,** 10, 10, 12, 12, 13

At t7=7, there are n7=6 at risk, and you have d7=1 failure and c7=0 censored observations. The survival probability is equal to the product of the first seven conditional probabilities.

```
km$n[7] <-  6
km$d[7] <-  1
km$c[7] <-  0
km$p[7] <- 1-km$d[7]/km$n[7]
km$s[7] <- km$p[1]*km$p[2]*km$p[3]*km$p[4]*km$p[5]*km$p[6]*km$p[7]
km
```

```
##       t  n  d  c          p          s
## 1    1 34  2  8  0.9411765  0.9411765
## 2    2 24  2  2  0.9166667  0.8627451
## 3    3 20  1  1  0.9500000  0.8196078
## 4    4 18  1  1  0.9444444  0.7740741
## 5    5 16  3  5  0.8125000  0.6289352
## 6    6  8  2  0  0.7500000  0.4717014
## 7    7  6  1  0  0.8333333  0.3930845
## 8   10 -1 -1 -1 -1.0000000 -1.0000000
## 9   12 -1 -1 -1 -1.0000000 -1.0000000
## 10  13 -1 -1 -1 -1.0000000 -1.0000000
```

1+, 1+, 1+, 1+, 1+, 1+, 1+, 1+, 1, 1, 2+, 2+, 2, 2, 3+, 3, 4+, 4, 5+, 5+, 5+, 5+, 5+, 5, 5, 5, 6, 6, 7, **10, 10,** 12, 12, 13

At t8=10, there are n8=5 at risk, and you have d8=2 failures and c8=0 censored observations. The survival probability is equal to the product of the first eight conditional probabilities.

```
km$n[8] <-  5
km$d[8] <-  2
```

```
km$c[8] <-  0
km$p[8] <- 1-km$d[8]/km$n[8]
km$s[8] <- km$p[1]*km$p[2]*km$p[3]*km$p[4]*km$p[5]*km$p[6]*km$p[7]*km$p[8]
km
```

```
##     t  n  d  c          p          s
## 1   1 34  2  8  0.9411765  0.9411765
## 2   2 24  2  2  0.9166667  0.8627451
## 3   3 20  1  1  0.9500000  0.8196078
## 4   4 18  1  1  0.9444444  0.7740741
## 5   5 16  3  5  0.8125000  0.6289352
## 6   6  8  2  0  0.7500000  0.4717014
## 7   7  6  1  0  0.8333333  0.3930845
## 8  10  5  2  0  0.6000000  0.2358507
## 9  12 -1 -1 -1 -1.0000000 -1.0000000
## 10 13 -1 -1 -1 -1.0000000 -1.0000000
```

1+, 1+, 1+, 1+, 1+, 1+, 1+, 1+, 1, 1, 2+, 2+, 2, 2, 3+, 3, 4+, 4, 5+, 5+, 5+, 5+, 5+, 5, 5, 5, 6, 6, 7, 10, 10, **12, 12,** 13

At t9=12, there are n9=3 at risk, and you have d9=2 failures and c9=0 censored observations. The survival probability is equal to the product of the first nine conditional probabilities.

```
km$n[9] <-  3
km$d[9] <-  2
km$c[9] <-  0
km$p[9] <- 1-km$d[9]/km$n[9]
km$s[9] <- km$p[1]*km$p[2]*km$p[3]*km$p[4]*km$p[5]*km$p[6]*km$p[7]*km$p[8]*km$p[9]
km
```

```
##     t  n  d  c          p          s
## 1   1 34  2  8  0.9411765  0.9411765
## 2   2 24  2  2  0.9166667  0.8627451
## 3   3 20  1  1  0.9500000  0.8196078
## 4   4 18  1  1  0.9444444  0.7740741
## 5   5 16  3  5  0.8125000  0.6289352
## 6   6  8  2  0  0.7500000  0.4717014
## 7   7  6  1  0  0.8333333  0.3930845
## 8  10  5  2  0  0.6000000  0.2358507
## 9  12  3  2  0  0.3333333  0.0786169
```

```
## 10 13 -1 -1 -1 -1.0000000 -1.0000000
```

1+, 1+, 1+, 1+, 1+, 1+, 1+, 1+, 1, 1, 2+, 2+, 2, 2, 3+, 3, 4+, 4, 5+, 5+, 5+, 5+, 5+, 5, 5, 5, 6, 6, 7, 10, 10, 12, 12, **13**

At t10=12, there is n10=1 at risk, and you have d10=1 failure and c10=0 censored observations. The survival probability is equal to the product of all ten conditional probabilities.

```
km$n[10] <-  1

km$d[10] <-  1

km$c[10] <-  0

km$p[10] <- 1-km$d[10]/km$n[10]

km$s[10] <- km$p[1]*km$p[2]*km$p[3]*km$p[4]*km$p[5]*km$p[6]*km$p[7]*km$p[8]*km$p[9]*
km$p[10]

km
```

```
##     t  n d c         p         s

## 1   1 34 2 8 0.9411765 0.9411765

## 2   2 24 2 2 0.9166667 0.8627451

## 3   3 20 1 1 0.9500000 0.8196078

## 4   4 18 1 1 0.9444444 0.7740741

## 5   5 16 3 5 0.8125000 0.6289352

## 6   6  8 2 0 0.7500000 0.4717014

## 7   7  6 1 0 0.8333333 0.3930845

## 8  10  5 2 0 0.6000000 0.2358507

## 9  12  3 2 0 0.3333333 0.0786169

## 10 13  1 1 0 0.0000000 0.0000000
```

Well, that was a lot of work, but it was worth it.

Let’s input the data to check our work.

```
t <- c(1:7, 10, 12, 13)


t <- c(

  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2,

  3, 3, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6,

  7, 10, 10, 12, 12, 13)


i <- c(

  0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1,
```

```
  0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1,
  1, 1, 1, 1, 1, 1)

catheter <- data.frame(t=t, i=i)
catheter_surv <- Surv(catheter$t, catheter$i)
catheter_km <- summary(survfit(catheter_surv~1))
catheter_km
## Call: survfit(formula = catheter_surv ~ 1)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     34       2   0.9412  0.0404       0.8653        1.000
##     2     24       2   0.8627  0.0647       0.7448        0.999
##     3     20       1   0.8196  0.0745       0.6859        0.979
##     4     18       1   0.7741  0.0831       0.6272        0.955
##     5     16       3   0.6289  0.1013       0.4587        0.862
##     6      8       2   0.4717  0.1227       0.2834        0.785
##     7      6       1   0.3931  0.1249       0.2109        0.733
##    10      5       2   0.2359  0.1142       0.0913        0.609
##    12      3       2   0.0786  0.0746       0.0122        0.505
##    13      1       1   0.0000     NaN           NA           NA
data.frame(our.calc=km$s, r.calc=catheter_km$surv)
##     our.calc    r.calc
## 1  0.9411765 0.9411765
## 2  0.8627451 0.8627451
## 3  0.8196078 0.8196078
## 4  0.7740741 0.7740741
## 5  0.6289352 0.6289352
## 6  0.4717014 0.4717014
## 7  0.3930845 0.3930845
## 8  0.2358507 0.2358507
## 9  0.0786169 0.0786169
## 10 0.0000000 0.0000000
```

They match. Hooray!