Nonparametric methods for Survival Analysis

One sample

```r
library(survival)
```

```
##
## Attaching package: 'survival'

## The following object is masked from 'package:rpart':
##
##     solder
```

```r
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------- tidyv

## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0

## -- Conflicts ---------------------------------------------------------------------------- tidyverse_c
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Entering right-censored data in R

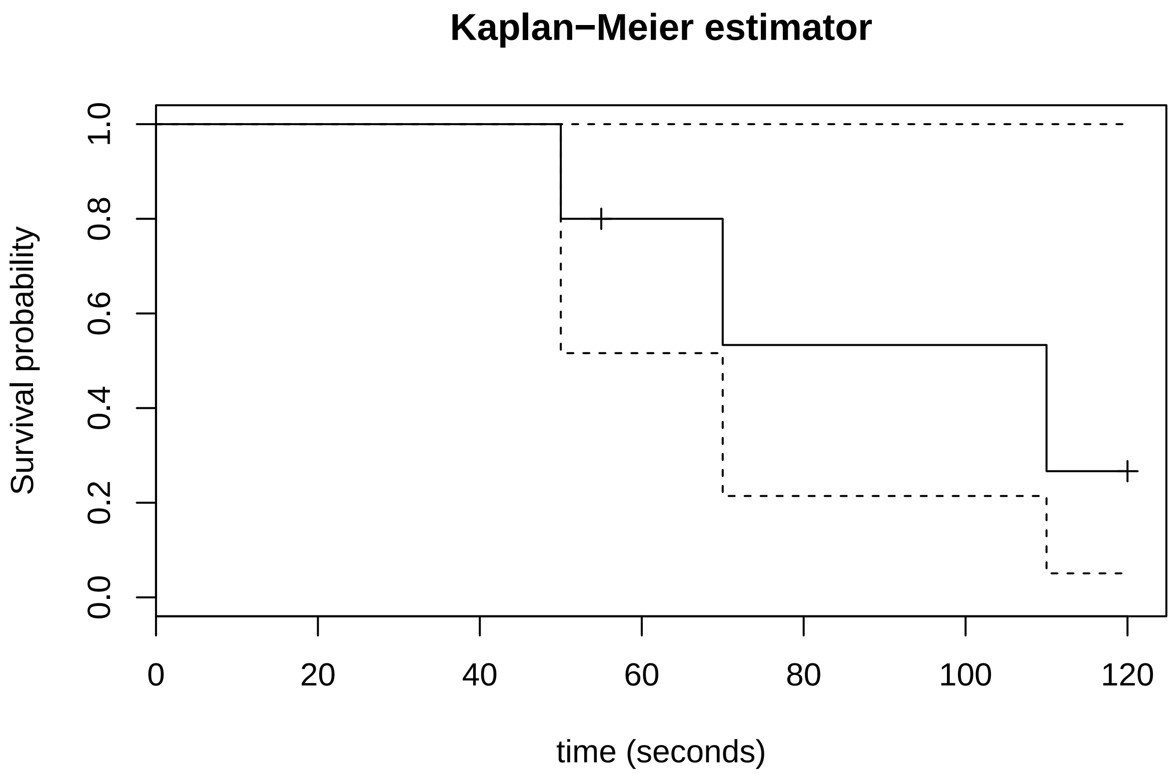
```r
dat <- data.frame(ratID = paste0("rat", 1:5),
                  time = c(55, 50, 70, 120, 110),
                  status = c(0, 1, 1, 0, 1))
```

Kaplan-Meyer estimator

```r
fit.KM <- survfit(Surv(time, status) ~ 1, data = dat)
summary(fit.KM)
```

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = dat)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    50      5       1    0.800   0.179       0.5161            1
##    70      3       1    0.533   0.248       0.2142            1
##   110      2       1    0.267   0.226       0.0507            1
```

```
plot(fit.KM, mark.time = TRUE,
     main = "Kaplan-Meier estimator",
     ylab = "Survival probability",
     xlab = "time (seconds)")
```



Question: what is the median survival time?

```
fit.KM
```

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = dat)
##
##       n  events  median 0.95LCL 0.95UCL
##       5       3     110      70      NA
```

Nelson-AAlen estimator

```
fit.NA <- survfit(Surv(time, status) ~ 1, data = dat, type = "fh")
summary(fit.NA)
```

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = dat, type = "fh")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    50      5       1    0.819   0.183       0.5282            1
##    70      3       1    0.587   0.273       0.2356             1
##   110      2       1    0.356   0.301       0.0677             1
```

```
fit.NA
```

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = dat, type = "fh")
##
##       n  events  median 0.95LCL 0.95UCL
##       5       3     110      70      NA
```

Case study: the Xelox trial

```
library(asaur)
dat <- gastricXelox
```

How many events, how many censored data points?

```
table(dat$delta)
```

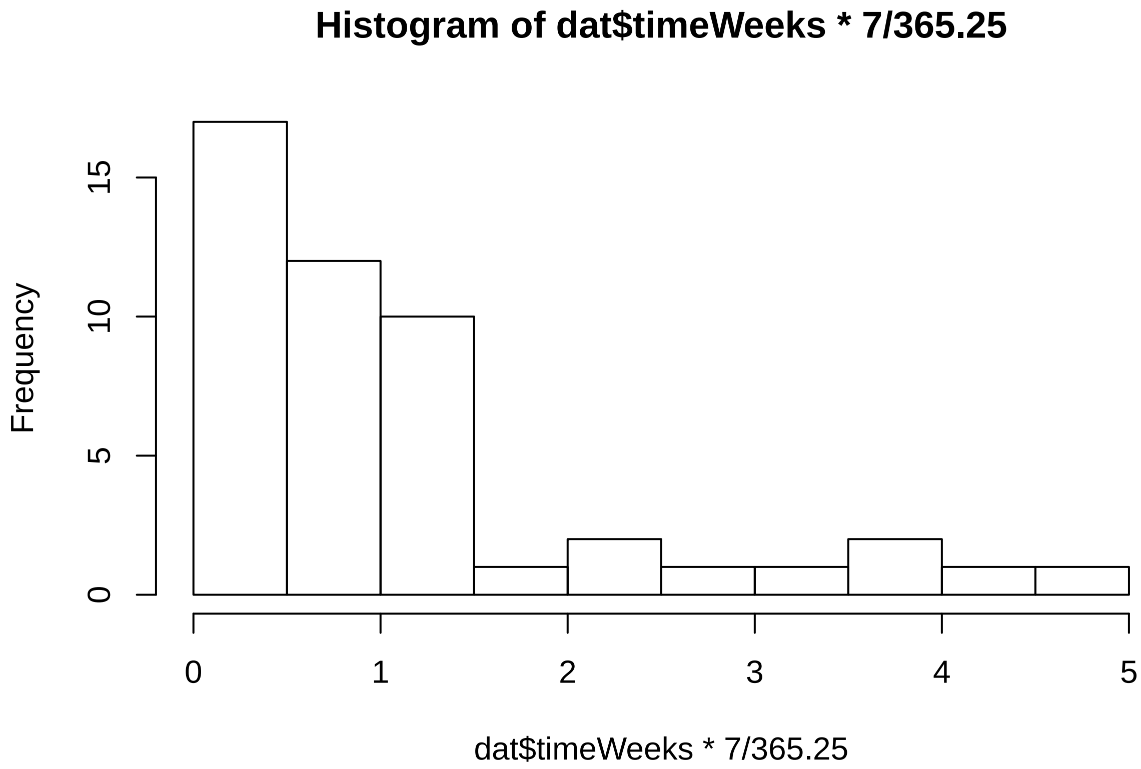
```
##
##  0  1
## 16 32
```

```
summary(dat)
```

```
##    timeWeeks          delta
##  Min.   :  4.00   Min.   :0.0000
##  1st Qu.: 18.50   1st Qu.:0.0000
##  Median : 43.00   Median :1.0000
##  Mean   : 59.71   Mean   :0.6667
##  3rd Qu.: 64.50   3rd Qu.:1.0000
##  Max.   :253.00   Max.   :1.0000
```

How the Progress Free Survival times data looks like (ignoring censoring info)?

```
hist(dat$timeWeeks * 7 / 365.25)
```



Kaplan-Meyer estimator

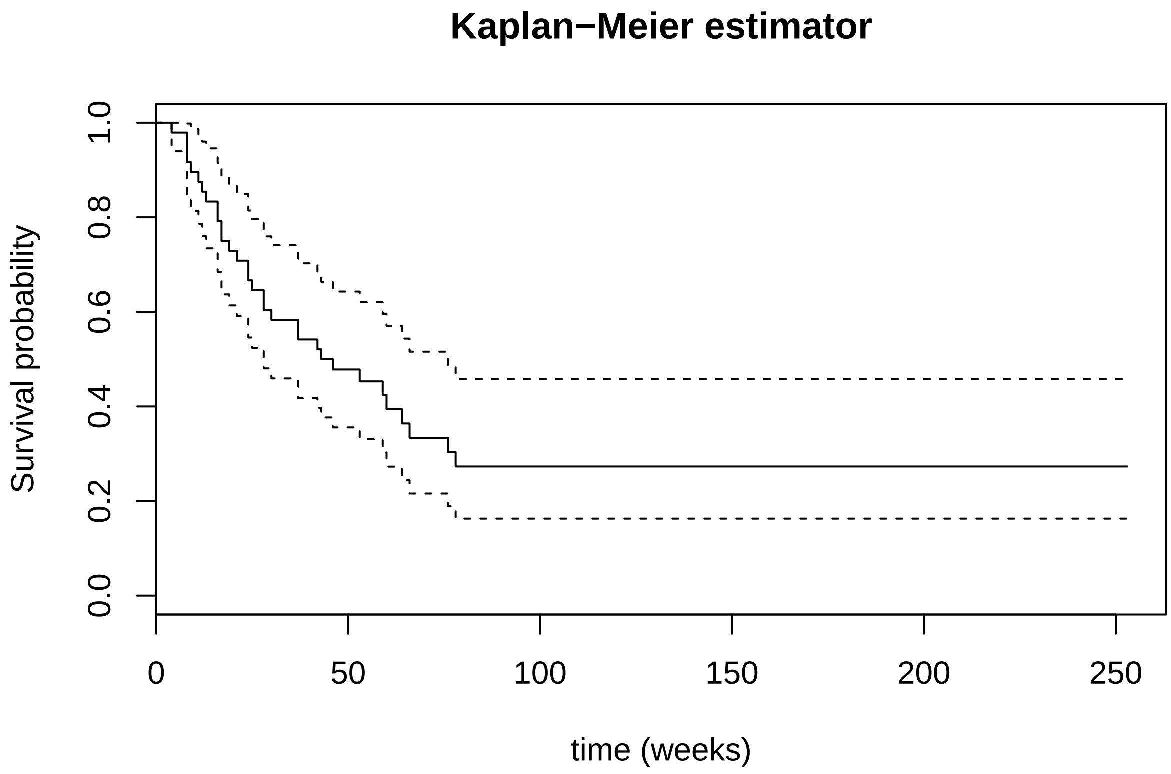
```
fit.KM <- survfit(Surv(timeWeeks, delta) ~ 1, data = dat)
summary(fit.KM)
```

```
## Call: survfit(formula = Surv(timeWeeks, delta) ~ 1, data = dat)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     4     48       1    0.979  0.0206        0.940        1.000
##     8     47       3    0.917  0.0399        0.842        0.998
##     9     44       1    0.896  0.0441        0.813        0.987
##    11     43       1    0.875  0.0477        0.786        0.974
##    12     42       1    0.854  0.0509        0.760        0.960
##    13     41       1    0.833  0.0538        0.734        0.946
##    16     40       2    0.792  0.0586        0.685        0.915
##    17     38       2    0.750  0.0625        0.637        0.883
##    19     36       1    0.729  0.0641        0.614        0.866
##    21     35       1    0.708  0.0656        0.591        0.849
##    24     34       2    0.667  0.0680        0.546        0.814
##    25     32       1    0.646  0.0690        0.524        0.796
##    28     31       2    0.604  0.0706        0.481        0.760
##    30     29       1    0.583  0.0712        0.459        0.741
##    37     28       2    0.542  0.0719        0.418        0.703
##    42     26       1    0.521  0.0721        0.397        0.683
##    43      25        1    0.500  0.0722        0.377         0.663
##    46      23        1    0.478  0.0722        0.356         0.643
##    53      19        1    0.453  0.0727        0.331         0.620
##    59      16        1    0.425  0.0735        0.303         0.596
##    60      14        1    0.394  0.0742        0.273         0.570
##    64      13        1    0.364  0.0744        0.244         0.544
##    66      12        1    0.334  0.0742        0.216         0.516
##    76      11        1    0.303  0.0734        0.189         0.487
##    78      10        1    0.273  0.0720        0.163         0.458
```

```
fit.KM
```

```
## Call: survfit(formula = Surv(timeWeeks, delta) ~ 1, data = dat)
##
##       n  events  median 0.95LCL 0.95UCL
##    48.0    32.0    44.5    28.0    76.0
```

```
plot(fit.KM,
     main = "Kaplan-Meier estimator",
     ylab = "Survival probability",
     xlab = "time (weeks)")
```

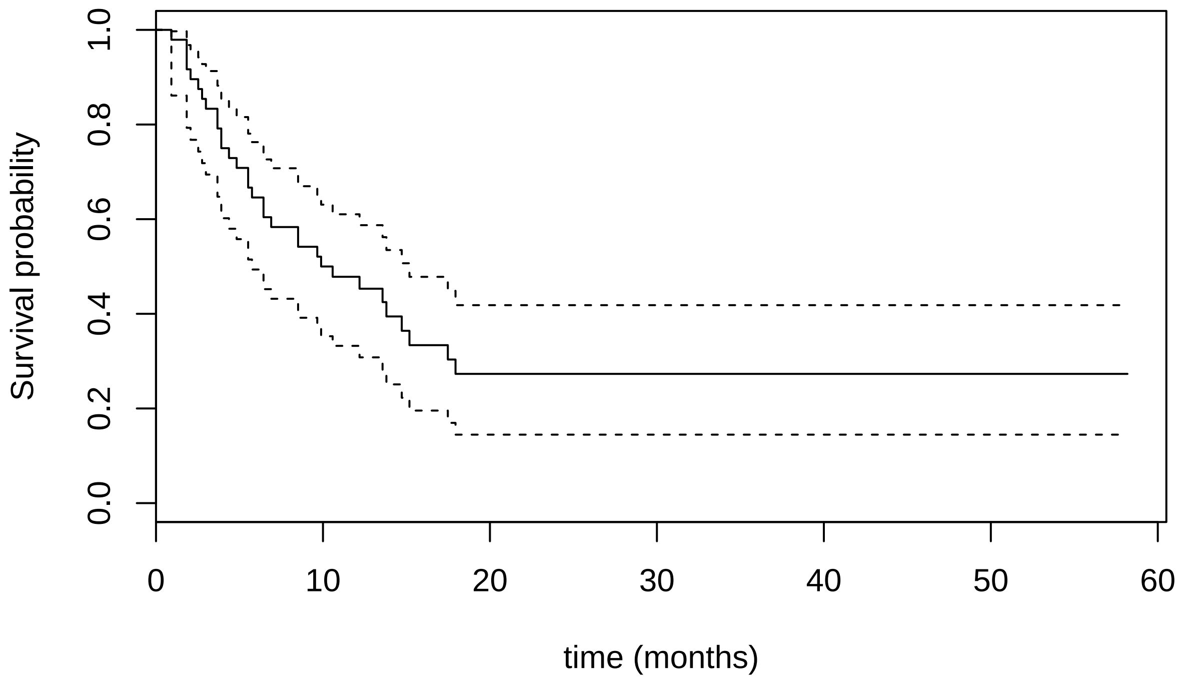


Express the Progress_Free Survival(PFS) times in months

```r
dat <- mutate(dat, timeYears = timeWeeks * 7 / 365.25 * 12)
fit.KM <- survfit(Surv(timeYears, delta) ~ 1, data = dat, conf.type = "log-log")
summary(fit.KM)
```

```
## Call: survfit(formula = Surv(timeYears, delta) ~ 1, data = dat, conf.type = "log-log")
##
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   0.92     48       1    0.979  0.0206        0.861        0.997
##   1.84     47       3    0.917  0.0399        0.793        0.968
##   2.07     44       1    0.896  0.0441        0.768        0.955
##   2.53     43       1    0.875  0.0477        0.743        0.942
##   2.76     42       1    0.854  0.0509        0.718        0.928
##   2.99     41       1    0.833  0.0538        0.694        0.913
##   3.68     40       2    0.792  0.0586        0.647        0.882
##   3.91     38       2    0.750  0.0625        0.602        0.850
##   4.37     36       1    0.729  0.0641        0.580        0.833
##   4.83     35       1    0.708  0.0656        0.558        0.816
##   5.52     34       2    0.667  0.0680        0.515        0.781
##   5.75     32       1    0.646  0.0690        0.494        0.763
##   6.44     31       2    0.604  0.0706        0.452        0.726
##   6.90     29       1    0.583  0.0712        0.432        0.708
##   8.51     28       2    0.542  0.0719        0.392        0.670
##   9.66     26       1    0.521  0.0721        0.372        0.650
##   9.89     25       1    0.500  0.0722        0.353        0.631
##  10.58     23       1    0.478  0.0722        0.332        0.610
##  12.19     19       1    0.453  0.0727        0.308        0.587
##  13.57     16       1    0.425  0.0735        0.280        0.562
##  13.80     14       1    0.394  0.0742        0.251        0.535
##  14.72     13       1    0.364  0.0744        0.223        0.507
##  15.18     12       1    0.334  0.0742        0.196        0.478
##  17.48     11       1    0.303  0.0734        0.170        0.449
##  17.94     10       1    0.273  0.0720        0.145        0.418
```

```r
plot(fit.KM,
     ylab = "Survival probability",
     xlab = "time (months)")
```

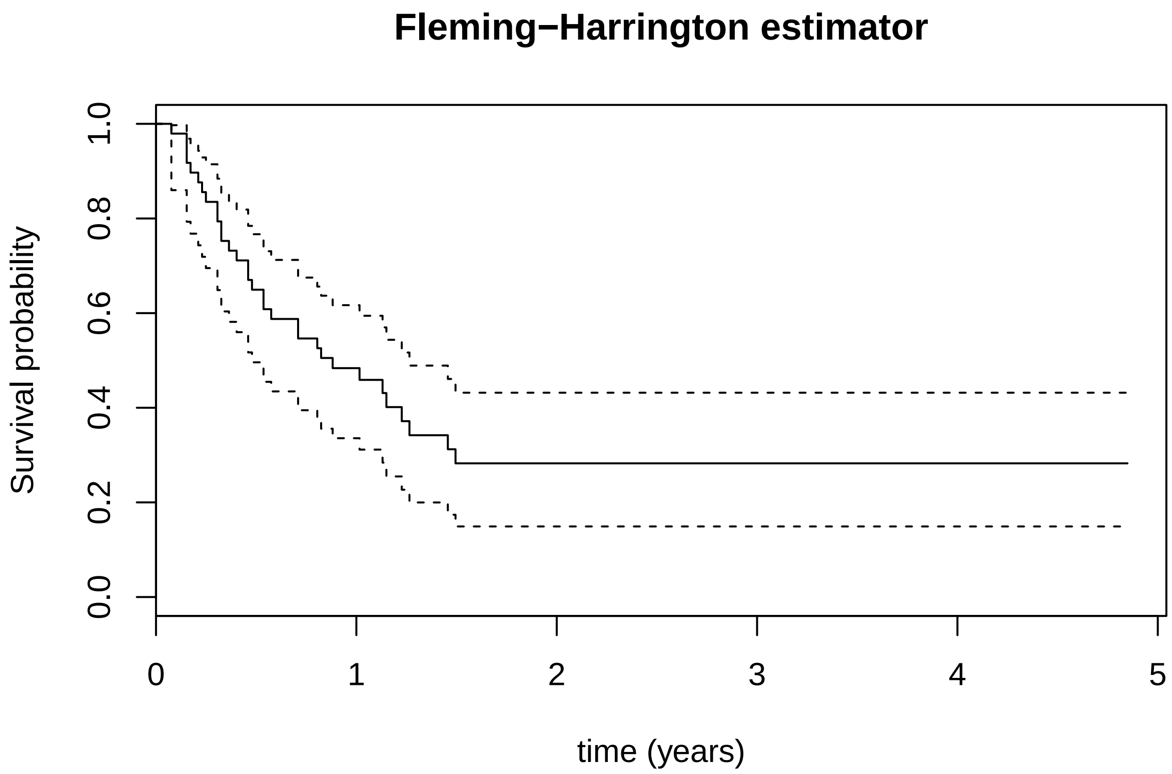


Time in weeks might be cumbersome to read: we can re-express it in years

```
dat <- mutate(dat, timeYears = timeWeeks * 7 / 365.25)
fit.KM <- survfit(Surv(timeYears, delta) ~ 1, data = dat, type = "fh", conf.type = "log-log")
summary(fit.KM)
```

```
## Call: survfit(formula = Surv(timeYears, delta) ~ 1, data = dat, type = "fh",
##     conf.type = "log-log")
##
##    time n.risk n.event survival std.err lower 95% CI upper 95% CI
##  0.0767     48       1    0.979  0.0206        0.860        0.997
##  0.1533     47       3    0.918  0.0399        0.793        0.969
##  0.1725     44       1    0.897  0.0441        0.768        0.956
##  0.2108     43       1    0.876  0.0478        0.743        0.943
##  0.2300     42       1    0.856  0.0510        0.719        0.929
##  0.2491     41       1    0.835  0.0539        0.695        0.915
##  0.3066     40       2    0.794  0.0588        0.649        0.884
##  0.3258     38       2    0.753  0.0627        0.604        0.852
##  0.3641     36       1    0.732  0.0644        0.581        0.836
##  0.4025     35       1    0.711  0.0659        0.560        0.819
##  0.4600     34       2    0.670  0.0684        0.517        0.784
##  0.4791     32       1    0.650  0.0694        0.496        0.767
##  0.5366     31       2    0.608  0.0711        0.455        0.731
##  0.5749     29       1    0.588  0.0717        0.435        0.712
##  0.7091     28       2    0.546  0.0725        0.395        0.675
##  0.8049     26       1    0.526  0.0728        0.375        0.656
##   0.8241      25         1     0.505  0.0729         0.356         0.637
##   0.8816      23         1     0.484  0.0731         0.336         0.617
##   1.0157      19         1     0.459  0.0736         0.312         0.594
##   1.1307      16         1     0.431  0.0745         0.284         0.570
##   1.1499      14         1     0.401  0.0755         0.255         0.544
##   1.2266      13         1     0.372  0.0760         0.227         0.517
##   1.2649      12         1     0.342  0.0760         0.200         0.489
##   1.4565      11         1     0.312  0.0755         0.174         0.461
##   1.4949      10         1     0.283  0.0745         0.149         0.432
```

```
plot(fit.KM,
     main = "Fleming-Harrington estimator",
     ylab = "Survival probability",
     xlab = "time (years)")
```



Median survival

Question: what is the median survival time?

```
fit.KM
```

```
## Call: survfit(formula = Surv(timeYears, delta) ~ 1, data = dat, type = "fh",
##     conf.type = "log-log")
##
```

```
##       n  events  median 0.95LCL 0.95UCL 
##  48.000  32.000   0.882   0.479   1.265 
```

Note that the definition of censoring depends on what's the quantity of interest. If we're interested in measuring the follow-up time, delta is to be 'inverted' (1- delta):

```
dat <- mutate(dat, delta_followUp = 1 - delta)
fit.followUp <- survfit(Surv(timeYears, delta_followUp) ~ 1, data = dat, conf.type = "log-log")
fit.followUp
```

```
## Call: survfit(formula = Surv(timeYears, delta_followUp) ~ 1, data = dat, 
##     conf.type = "log-log")
## 
##       n  events  median 0.95LCL 0.95UCL 
##   48.00   16.00    2.30    1.13    3.58 
```

Nonparametric comparison of two samples

Entering right-censored data in R

```
dat <- data.frame(ratID = paste0("rat", 1:5),
                  time = c(55, 50, 70, 120, 110),
                  status = c(0, 1, 1, 0, 1),
                  group = c(0, 1, 0, 1, 1))
```

The logrank test

```
fit.logrank <- survdiff(Surv(time, status) ~ group, data = dat)
fit.logrank
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ group, data = dat)
## 
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## group=0 2        1    0.733    0.0970     0.154
## group=1 3        2    2.267    0.0314     0.154
## 
##  Chisq= 0.2  on 1 degrees of freedom, p= 0.7 
```

For the rats depr. example: A: non sd, B: sd, Logrank Test: p-value = 0.7 Do not reject H0: SA(t) = SB(t) Conclusion: This is a nonparametric test

Case study: the pancreatic dataset

1. What's the medain Overall Survival of a patient with Locally Advanced(LA) pancreatic cancer?
2. Provide confidence interval

3. Do the two stages experiece significantly different survival?
4. What's the probability of surviving more than a year within more than a yar within each group?

```
library(asaur)

dat <- pancreatic
head(dat)
```

```
##   stage    onstudy progression      death
## 1     M 12/16/2005    2/2/2006 10/19/2006
## 2     M   1/6/2006   2/26/2006  4/19/2006
## 3    LA   2/3/2006    8/2/2006  1/19/2007
## 4     M  3/30/2006           .  5/11/2006
## 5    LA  4/27/2006   3/11/2007  5/29/2007
## 6     M   5/7/2006   6/25/2006 10/11/2006
```

- M: metastatic
- LA: locally advanced

This dataset requires some preprocessing before proper survival analysis.

1. parse 'onstudy', 'progression' and 'death' dates correctly
2. compute progression free survival times and overall survival times (this dataset has no censored data)

step 1: parse dates

Check the manual page of 'as.Date'

```
fmt <- "%m/%d/%Y"
dat <- mutate(dat,
  onstudy = as.Date(as.character(onstudy), format = fmt),
  progression = as.Date(as.character(progression), format = fmt),
  death = as.Date(as.character(death), format = fmt)
)
head(dat)
```

```
##   stage    onstudy progression      death
## 1     M 2005-12-16  2006-02-02 2006-10-19
## 2     M 2006-01-06  2006-02-26 2006-04-19
## 3    LA 2006-02-03  2006-08-02 2007-01-19
## 4     M 2006-03-30        <NA> 2006-05-11
## 5    LA 2006-04-27  2007-03-11 2007-05-29
## 6     M 2006-05-07  2006-06-25 2006-10-11
```

step 2: compute survival times

```
dat <- mutate(dat,
  OS = difftime(death, onstudy, units = "days"),
  PFS = ifelse(!is.na(progression), difftime(progression, onstudy, units = "days"), OS)
)
```

Note: OS and PFS are expressed in days. We want them in months:

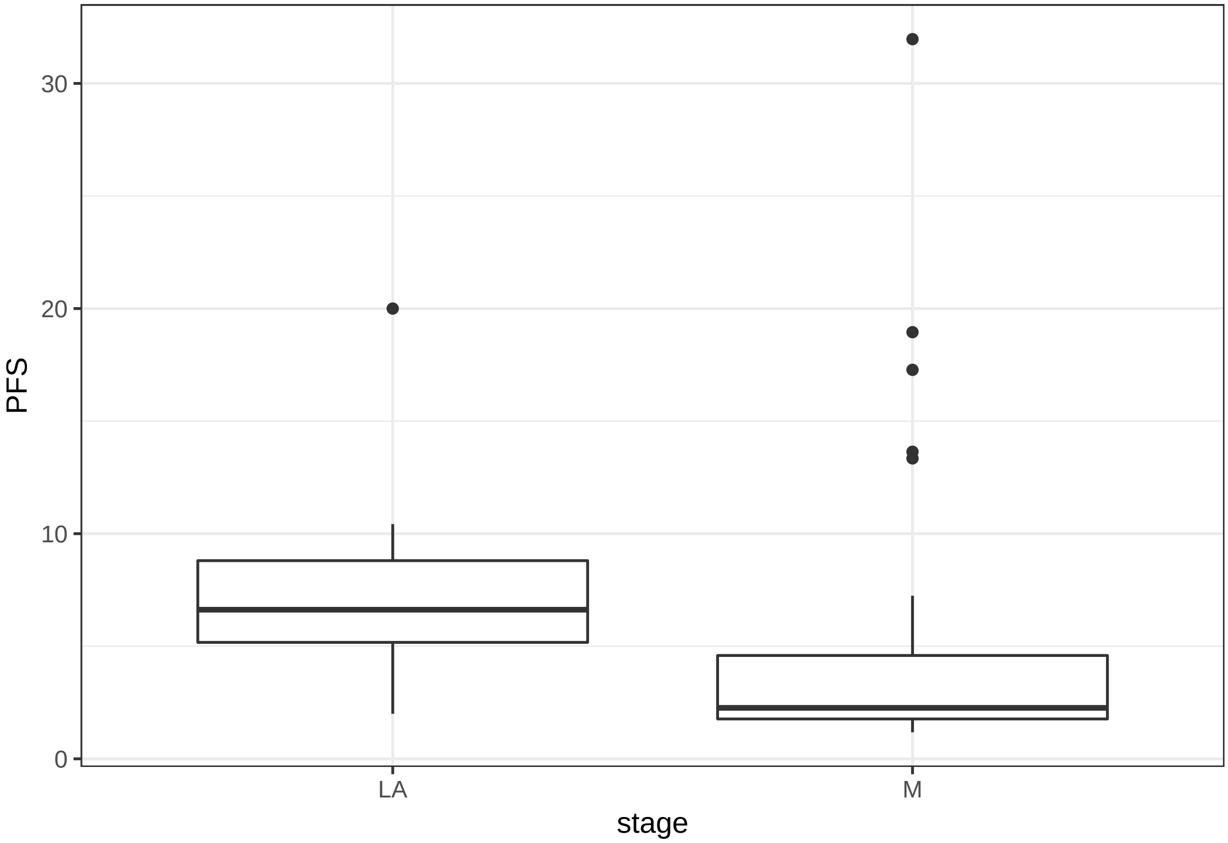
```
dat <- mutate(dat,
  OS = as.numeric(OS) / 30.5,
  PFS = as.numeric(PFS) / 30.5
)
```

compare PFS in the 2 disease groups

As we have no censoring, we can produce use simple boxplots:

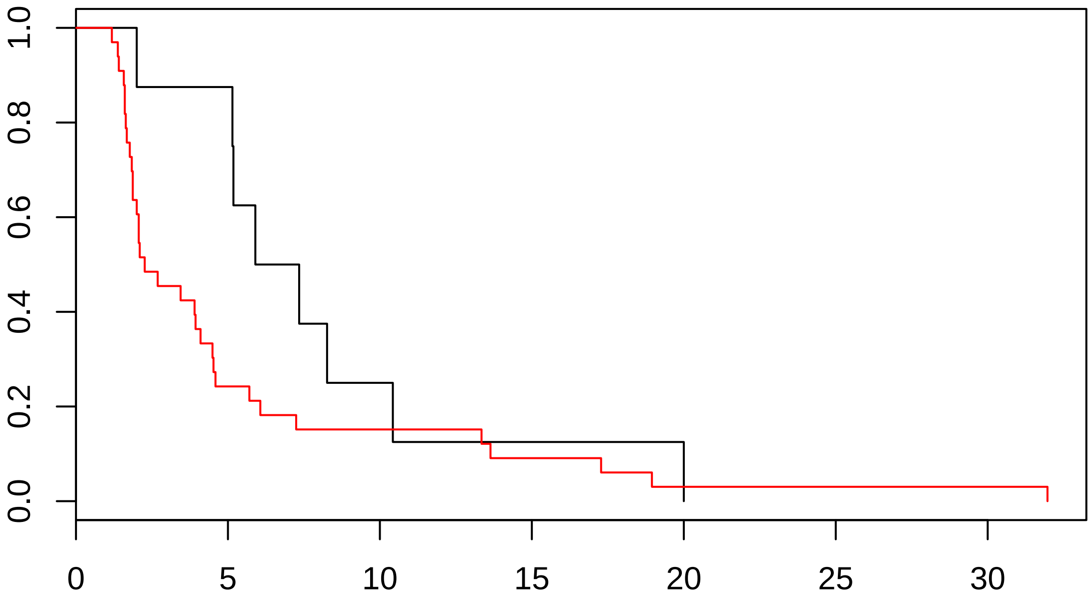
```
library(ggplot2)
```

```
ggplot(dat, aes(stage, PFS)) +
  geom_boxplot() +
  theme_bw()
```



more generally, Kaplan-Meier estimates:

```
fit.KM <- survfit(Surv(PFS) ~ stage, data = dat, conf.type = "log-log")
plot(fit.KM, col = 1:2)
```



```
fit.KM
```

```
## Call: survfit(formula = Surv(PFS) ~ stage, data = dat, conf.type = "log-log")
##
##           n events median 0.95LCL 0.95UCL
## stage=LA  8      8   6.62    2.00    10.4
## stage=M  33     33   2.26    1.87     4.1
```

The logrank test

```
survdiff(Surv(PFS) ~ stage, data = dat)
```

```
## Call:
## survdiff(formula = Surv(PFS) ~ stage, data = dat)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## stage=LA  8        8     12.3      1.49      2.25
## stage=M  33       33     28.7      0.64      2.25
##
##  Chisq= 2.2  on 1 degrees of freedom, p= 0.1
```

Cannot reject null hypothesis

What’s the estimated probability of not experiencing a cancer progression for (at least) 1 year?

```r
summary(fit.KM, time = 12)
```

```
## Call: survfit(formula = Surv(PFS) ~ stage, data = dat, conf.type = "log-log")
##
##                 stage=LA
##         time       n.risk      n.event     survival      std.err
##     12.00000      1.00000      7.00000      0.12500      0.11693
## lower 95% CI upper 95% CI
##      0.00659      0.42271
##
##                 stage=M
##         time       n.risk      n.event     survival      std.err
##      12.0000       5.0000      28.0000       0.1515       0.0624
## lower 95% CI upper 95% CI
##       0.0553       0.2922
```

It is similar in the 2 groups, namely between 13% and 15%. Said otherwise, chances are high that the cancer is going to make a comeback within one year.

How about OS?

```r
survdiff(Surv(OS) ~ stage, data = dat)
```

```
## Call:
## survdiff(formula = Surv(OS) ~ stage, data = dat)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## stage=LA  8        8     9.74    0.3093     0.425
## stage=M  33       33    31.26    0.0963     0.425
##
##  Chisq= 0.4  on 1 degrees of freedom, p= 0.5
```

```r
summary(survfit(Surv(OS) ~ stage, data = dat, conf.type = "log-log"), time = 12)
```

```
## Call: survfit(formula = Surv(OS) ~ stage, data = dat, conf.type = "log-log")
##
##                 stage=LA
##         time       n.risk      n.event     survival      std.err
##       12.000        3.000        5.000        0.375        0.171
## lower 95% CI upper 95% CI
##        0.087        0.674
##
##                 stage=M
##         time       n.risk      n.event     survival      std.err
##      12.0000       7.0000      26.0000       0.2121       0.0712
## lower 95% CI upper 95% CI
##       0.0935       0.3625
```