Nonparametric methods for Survival Analysis

One sample

```r
library(survival)
library(tidyverse)
```

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

## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## Warning: package 'dplyr' was built under R version 3.6.1

## -- Conflicts ------------------------------------------------------------------------------------
## 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))
```

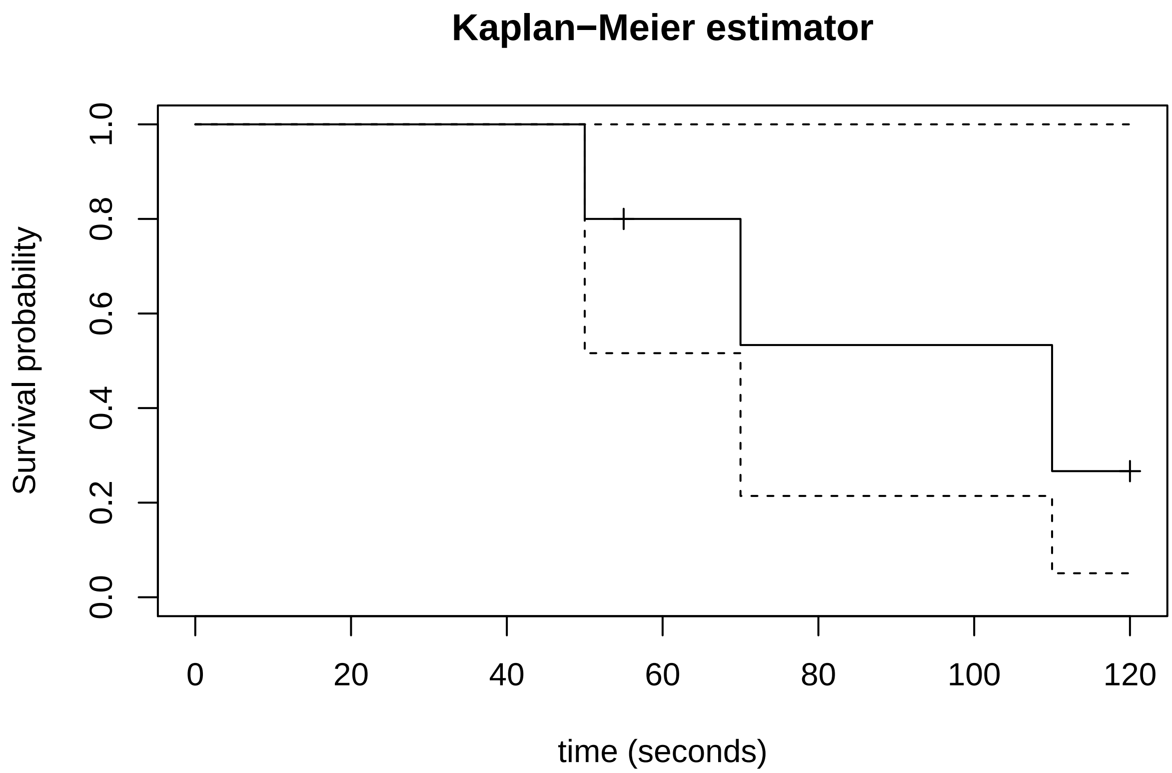
Kaplan-Meyer estimator

1: event, 0: no event, cencoring

```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
```

```r
#sensor is the mark in the line
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 #medium survival + interval
```

```
## 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   0.164        0.553            1
##    70      3       1    0.587   0.228        0.274            1
##   110      2       1    0.356   0.225        0.103            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

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

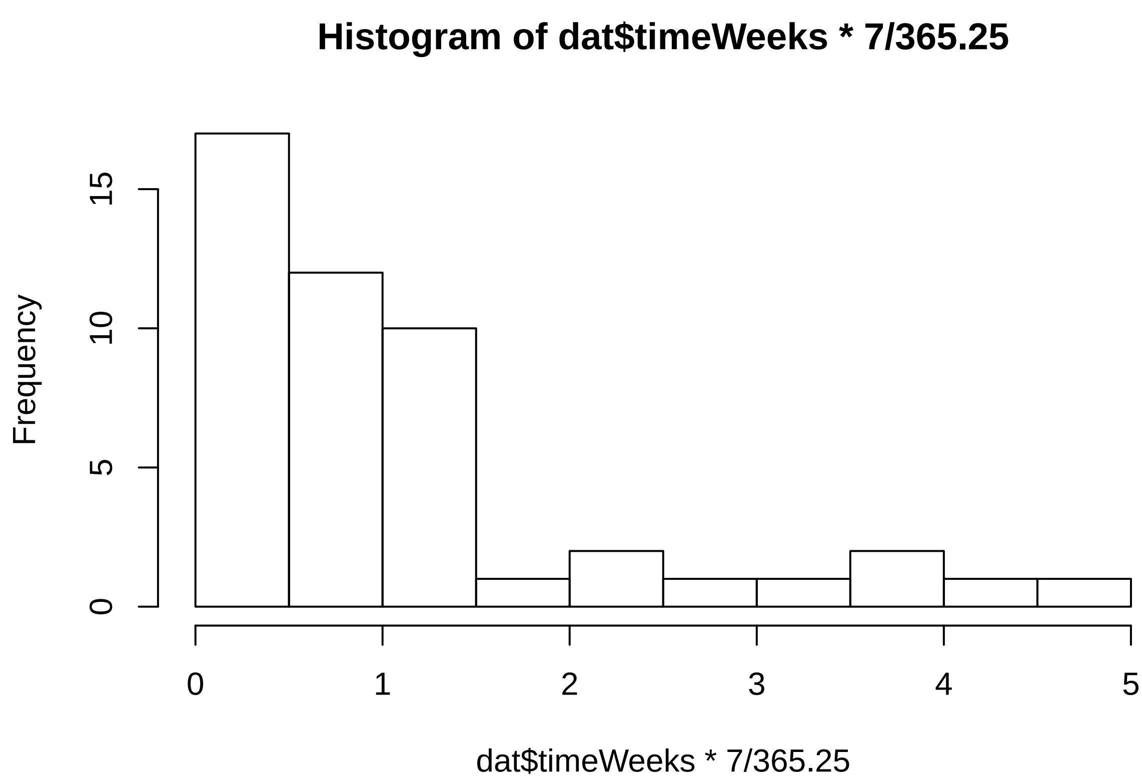
How many events, how many censored data points?

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

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

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

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



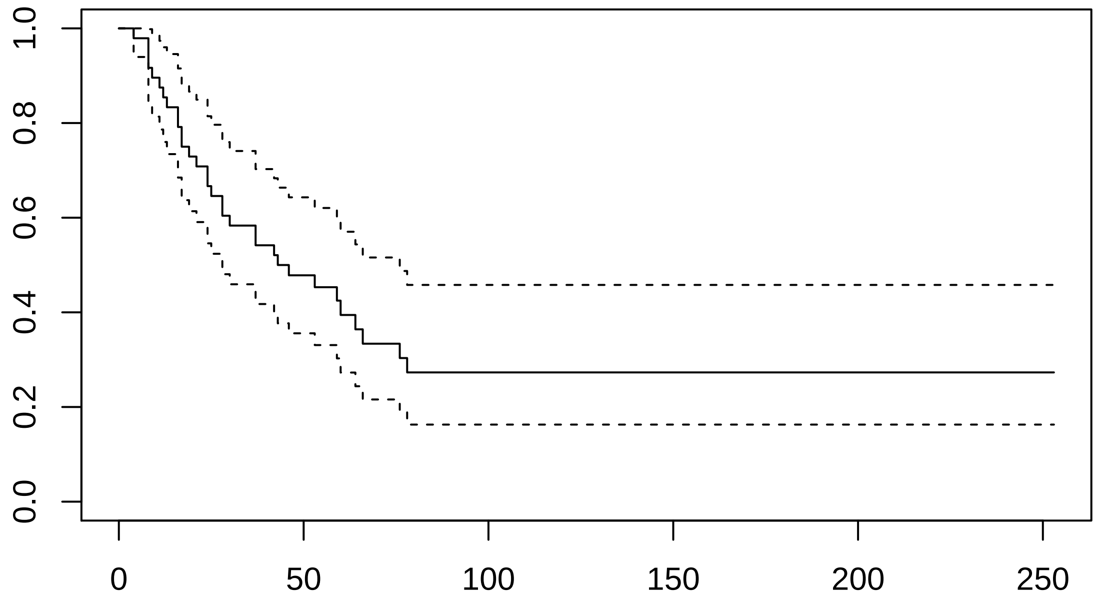
Kaplan-Meyer estimator

```
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
```

the median of time a patient see the progression is 44.5

```
plot(fit.KM)
```




```
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
```

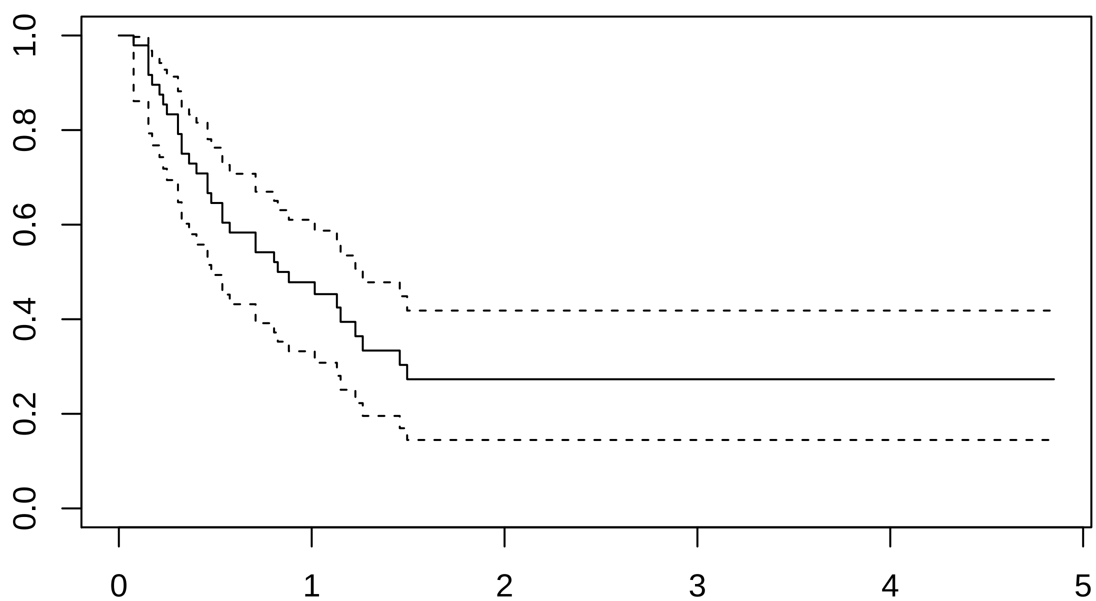
Time in weeks might be cumbersome to read: we can re-express it in years

```
#mutate create new data
dat <- mutate(dat, timeYears = timeWeeks * 7 / 365.25)
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
##  0.0767     48       1    0.979  0.0206        0.861        0.997
##  0.1533     47       3    0.917  0.0399        0.793        0.968
##  0.1725     44       1    0.896  0.0441        0.768        0.955
##  0.2108     43       1    0.875  0.0477        0.743        0.942
##  0.2300     42       1    0.854  0.0509        0.718        0.928
##  0.2491     41       1    0.833  0.0538        0.694        0.913
##  0.3066     40       2    0.792  0.0586        0.647        0.882
##  0.3258     38       2    0.750  0.0625        0.602        0.850
```

```
## 0.3641     36       1    0.729  0.0641        0.580        0.833
## 0.4025     35       1    0.708  0.0656        0.558        0.816
## 0.4600     34       2    0.667  0.0680        0.515        0.781
## 0.4791     32       1    0.646  0.0690        0.494        0.763
## 0.5366     31       2    0.604  0.0706        0.452        0.726
## 0.5749     29       1    0.583  0.0712        0.432        0.708
## 0.7091     28       2    0.542  0.0719        0.392        0.670
## 0.8049     26       1    0.521  0.0721        0.372        0.650
## 0.8241     25       1    0.500  0.0722        0.353        0.631
## 0.8816     23       1    0.478  0.0722        0.332        0.610
## 1.0157     19       1    0.453  0.0727        0.308        0.587
## 1.1307     16       1    0.425  0.0735        0.280        0.562
## 1.1499     14       1    0.394  0.0742        0.251        0.535
## 1.2266     13       1    0.364  0.0744        0.223        0.507
## 1.2649     12       1    0.334  0.0742        0.196        0.478
## 1.4565     11       1    0.303  0.0734        0.170        0.449
## 1.4949     10       1    0.273  0.0720        0.145        0.418
```

```r
plot(fit.KM)
```



Median survival

Question: what is the median survival time?

so median, 32 out of 48 see a progress, medium of 0.853 year with confidence interval (0.479,1.265)

```
fit.KM
```

```
## Call: survfit(formula = Surv(timeYears, delta) ~ 1, data = dat, conf.type = "log-log")
##
##       n  events  median 0.95LCL 0.95UCL
##  48.000  32.000   0.853   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': (how long we are able to follow up a subject)

```
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

H0: two group are the same, here we do not reject the null hypothesis.

```
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
```

Case study: the pancreatic dataset

```r
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'

```r
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

```r
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:

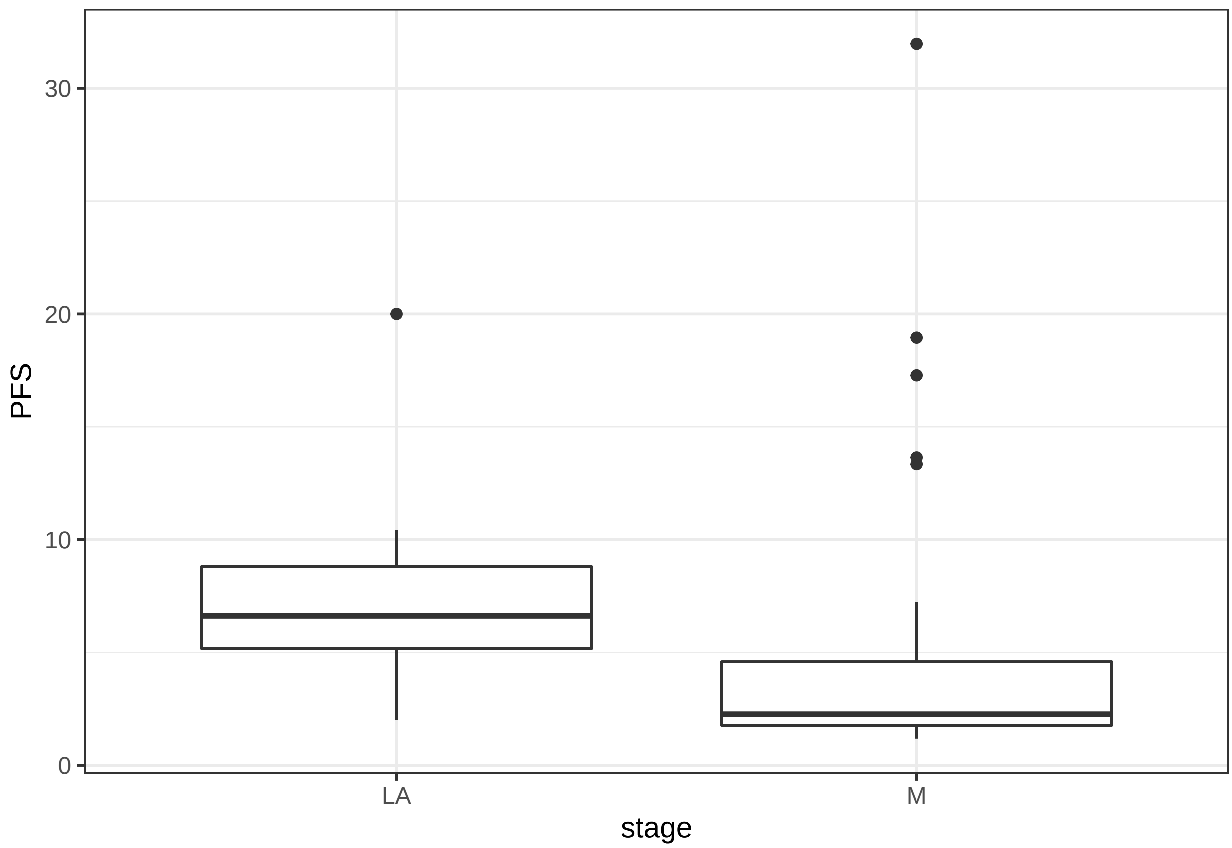
```r
dat <- mutate(dat,
  OS = as.numeric(OS) / 30.5,
  PFS = as.numeric(PFS) / 30.5
)
```

compare PFS in the 2 disease groups

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

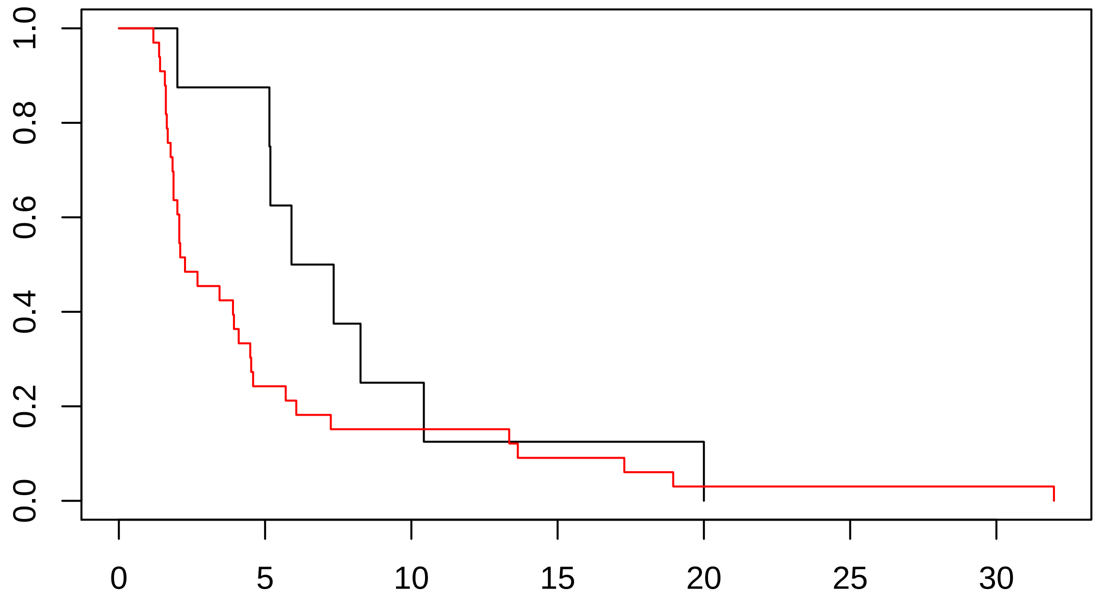
```r
library(ggplot2)
```

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



more generally, Kaplan-Meier estimates:

```r
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
```

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.

Can you repeat the analysis above, this time for OS?

Stratified logrank test: pharmacoSmoking dataset

The data

```r
dat <- pharmacoSmoking
head(dat)
```

```
##    id ttr relapse         grp age gender     race employment yearsSmoking
## 1  21 182       0   patchOnly  36   Male    white         ft           26
## 2 113  14       1   patchOnly  41   Male    white      other           27
## 3  39   5       1 combination  25 Female    white      other           12
## 4  80  16       1 combination  54   Male    white         ft           39
## 5  87   0       1 combination  45   Male    white      other           30
## 6  29 182       0 combination  43   Male hispanic         ft           30
##   levelSmoking ageGroup2 ageGroup4 priorAttempts longestNoSmoke
## 1        heavy     21-49     35-49             0              0
## 2        heavy     21-49     35-49             3             90
## 3        heavy     21-49     21-34             3             21
## 4        heavy       50+     50-64             0              0
## 5        heavy     21-49     35-49             0              0
## 6        heavy     21-49     35-49             2           1825
```

```r
summary(dat)
```

```
##        id              ttr            relapse                grp
##  Min.   :  1.00   Min.   :  0.00   Min.   :0.000   combination:61
##  1st Qu.: 33.00   1st Qu.:  8.00   1st Qu.:0.000   patchOnly  :64
##  Median : 67.00   Median : 49.00   Median :1.000
##  Mean   : 66.15   Mean   : 77.44   Mean   :0.712
```

```
##  3rd Qu.: 99.00   3rd Qu.:182.00   3rd Qu.:1.000
##  Max.   :130.00   Max.   :182.00   Max.   :1.000
##       age            gender          race     employment  yearsSmoking
##  Min.   :22.00   Female:81   black   :38   ft   :72   Min.   : 9.00
##  1st Qu.:41.00   Male  :44   hispanic: 8   other:39   1st Qu.:22.00
##  Median :49.00               other   : 2   pt   :14   Median :30.00
##  Mean   :48.84               white   :77              Mean   :30.88
##  3rd Qu.:56.00                                        3rd Qu.:39.00
##  Max.   :86.00                                        Max.   :56.00
##  levelSmoking ageGroup2  ageGroup4  priorAttempts     longestNoSmoke
##  heavy:89     21-49:66   21-34:16   Min.   :   0.00   Min.   :   0.0
##  light:36     50+  :59   35-49:50   1st Qu.:   1.00   1st Qu.:   7.0
##                          50-64:48   Median :   2.00   Median :  90.0
##                          65+  :11   Mean   :  12.68   Mean   : 539.7
##                                     3rd Qu.:   5.00   3rd Qu.: 365.0
##                                     Max.   :1000.00   Max.   :6205.0
```

Question: do the 2 treatment group differ significantly in terms of survival to relapse?

```r
survdiff(Surv(ttr, relapse) ~ grp, data = dat)
```

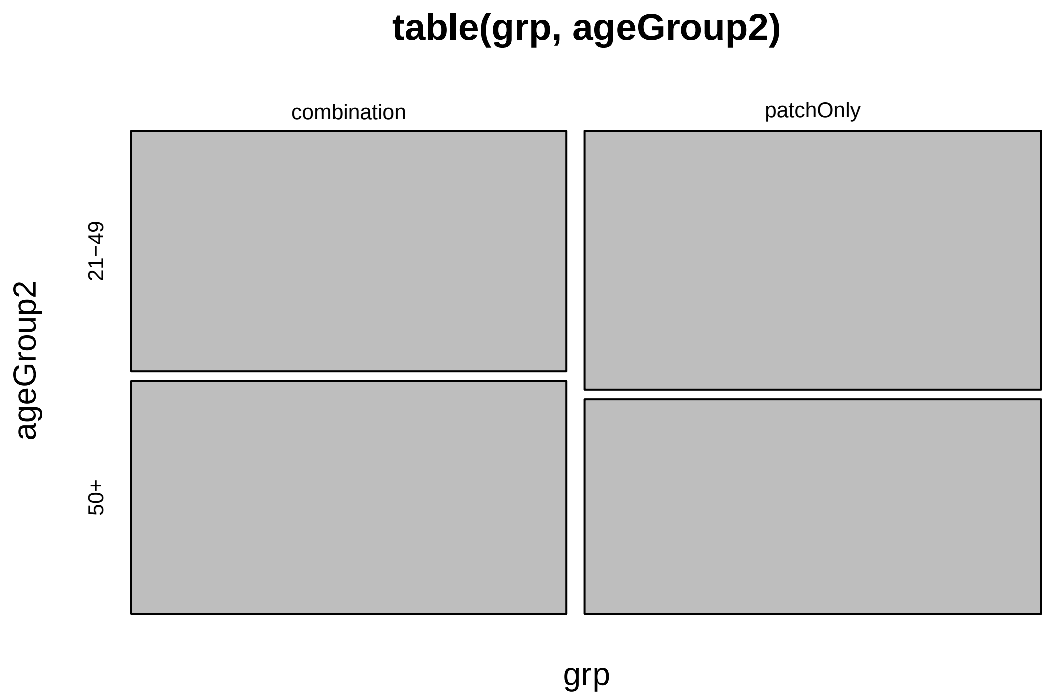
```
## Call:
## survdiff(formula = Surv(ttr, relapse) ~ grp, data = dat)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## grp=combination 61       37     49.9      3.36      8.03
## grp=patchOnly   64       52     39.1      4.29      8.03
##
##  Chisq= 8  on 1 degrees of freedom, p= 0.005
```

Critique: the 2 groups have different age distribution, which might confound our results. Lets investigate:

```r
with(dat, prop.table(table(grp, ageGroup2), 1))
```

```
##              ageGroup2
## grp               21-49       50+
##   combination 0.5081967 0.4918033
##   patchOnly   0.5468750 0.4531250
```

```r
with(dat, mosaicplot(table(grp, ageGroup2)))
```



stratified logrank test

```
survdiff(Surv(ttr, relapse) ~ grp + strata(ageGroup2), data = dat)
```

```
## Call:
## survdiff(formula = Surv(ttr, relapse) ~ grp + strata(ageGroup2),
##     data = dat)
##
##                    N Observed Expected (O-E)^2/E (O-E)^2/V
## grp=combination 61       37     49.1      2.99      7.03
## grp=patchOnly   64       52     39.9      3.68      7.03
##
##  Chisq= 7  on 1 degrees of freedom, p= 0.008
```

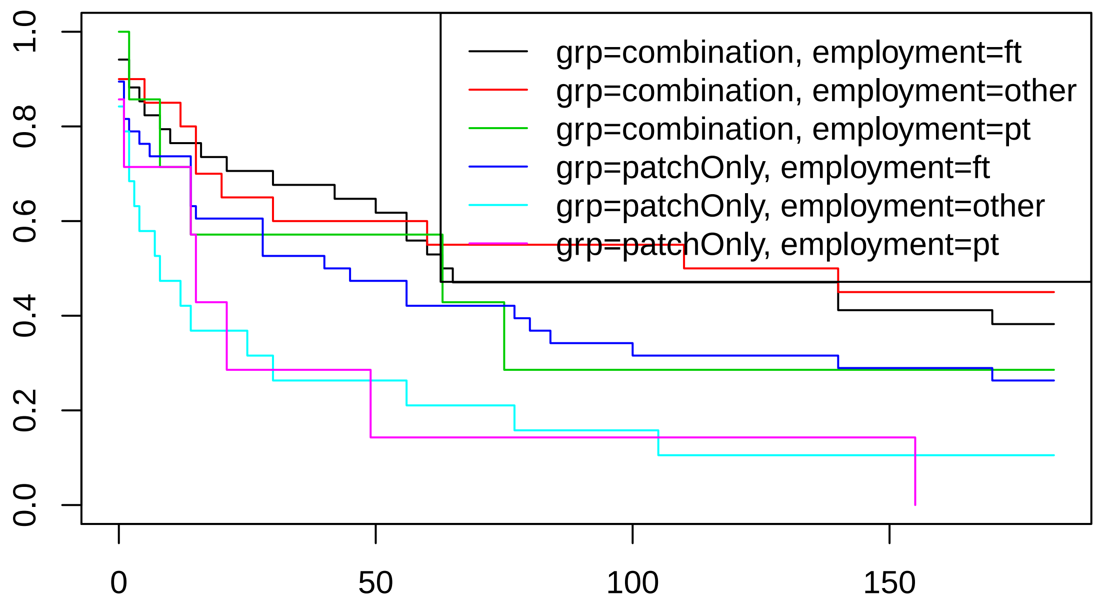
extra

```
fit.4 <- survfit(Surv(ttr, relapse) ~ grp + employment, data = dat)
fit.4
```

```
## Call: survfit(formula = Surv(ttr, relapse) ~ grp + employment, data = dat)
##
```

```
##                                    n events median 0.95LCL 0.95UCL
## grp=combination, employment=ft    34     21   64.0      50      NA
## grp=combination, employment=other 20     11  125.0      20      NA
## grp=combination, employment=pt     7      5   63.0       8      NA
## grp=patchOnly, employment=ft      38     28   42.5      14     140
## grp=patchOnly, employment=other   19     17    8.0       3      77
## grp=patchOnly, employment=pt       7      7   15.0       1      NA
```

```r
plot(fit.4, col = 1:6)
legend("topright", lty = 1, col = 1:6, legend = names(fit.4$strata))
```



The 3 'combination' curves seem all higher than the 3 'patchOnly' curves. Lets make a stratified test:

```r
survdiff(Surv(ttr, relapse) ~ grp + strata(employment), data = dat)
```

```
## Call:
## survdiff(formula = Surv(ttr, relapse) ~ grp + strata(employment),
##     data = dat)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## grp=combination 61       37     50.3      3.50      8.58
## grp=patchOnly   64       52     38.7      4.54      8.58
##
##  Chisq= 8.6  on 1 degrees of freedom, p= 0.003
```