Population-based cancer survival: measures and non-parametric estimation

Facilitators: B. Rachet and A. Belot

Acknowledgement: We adapted the contents of a practical which was originally written by Maja Pohar Perme.

The goal of this exercise is to get acquainted with the use of relative survival in R. Load the libraries 'survival' and 'relsurv'.

```
> library(survival)
> library(relsurv)
> data(colrec)
> source("frpop.r")
```

1 The data

We have two sets of data:

- The data on 5971 patients diagnosed with colon or rectum cancer between January 1st, 1994 and December 31st, 2000. Data set `colrec`
- The French population mortality tables. Data set *frpop*

1.1 Inspect the colrec dataset

In this session we will use the colrec dataset that contains anonymised data on 5971 patients diagnosed with colon or rectum cancer between January 1st, 1994 and December 31st, 2000. The original dataset is integrated in the relsurv package (object `colrec`).

Task 1: Check the data:

(i) Check the distribution of age and sex.
Hint: age is reported in days (use 365.241 for changes between days and years), sex: 1=male, 2=female

(ii) Find the first and last date of diagnosis.
Hint: use `as.Date` function to get dates

(iii) Check the distribution of follow-up times.
Hint: reported in days

(iv) How many patients have died during the follow-up time?
Hint: variable `stat`

1.2 Understand the mortality tables structure

In our analysis, we will assume that all patients come from France, thus we will use the French population tables.

The population tables can be organized in various ways (e.g. number of demographical variables, cutpoints), the survival package object `ratetable` has been provided to help in working with the different formats.

The tables for this practical have already been put into the `ratetable` format, so you can source them into R.

For future reference, the solutions describe how the tables can be obtained.

The obtained ratetable is a three-dimensional array that contains the daily hazards for every combination of a person's age, year and sex.

Task 2: Examine the population table object:

- According to which variables are the tables split?
  Hint: `attributes(frpop), dimid`
- Find the dimensions of the object.
  Hint: `attributes(frpop), dim`
- What is the span of calendar years covered?
  Hint: `attributes(frpop), cutpoints`
- Find the value of the daily hazard for individuals aged 80 in 2003
- Estimate the probability of surviving from 80 to 81 for a man aged 80 in 2003.

2 Overall and expected survival

Task 3:

(i) Plot overall survival.

(ii) Estimate 5 and 10-year overall survival.

(iii) Plot expected survival on the same graph.

(iv) Estimate 10-year expected survival in the population. Interpret the results.

Note that the demographic varaibles have to be organized to match the ratetable object format and names. In the frpop, there are three dimensions:

- age is named 'age', reported in days
- calendar year is named 'year', reported in the date format
- sex is named 'sex', the first category is 'male', the second 'female'.

Hint: there are two options for the consolidation of the variables:

- *change the observed data set to match the ratetable*
- *use the `rmap` argument*

Hint: use the xscale-argument in plot.survfit to convert the time unit from days to years.

We see that the first couple of years are critical for patients (many deaths occur), but from the fifth year onwards the hazard decreases. The 10-year overall survival is 0.27.

3 Estimating crude and net survival

We are interested in 5-year survival, we will estimate 5-year net and crude survival.

Task 4:

(i) Limit the follow-up time to 5 years (censor all individuals after 5 years)

(ii) Estimate 5-year crude mortality, interpret your results

(iii) Estimate 5-year net survival, interpret your results

Interpretation for crude mortality:
We already know that 0.64 patients died. 0.57 died due to cancer and 0.07 died due to other causes.

Two interpretations for net survival:

- On average, the ratio between the patient survival and that of their counterparts is 0.41
- In a hypothetical world, where patients could die of cancer only, 41 % of patients would still be alive after 5 years.

4 Net survival with respect to covariates

Task 5: Compare the 5-year net survival with respect to sex and age. To this end, split individuals to below and above 65 years.
Hint: use the function `rs.diff`

No difference can be observed between men and women. On the other hand, age seems to be a very important factor, with younger patients having a lower excess hazard and hence a better net survival.

5 Longer follow-up times

Task 6:

(i) Check the amount of information in your data for individuals aged over 80 after 10 or 15 years of follow-up.
Hint: use the function `nessie`

(ii) Plot 15-year net survival for a sensible subgroup of patients.

We can observe that:

- For individuals above 90 (or the total sample containing them) even the 5-year net survival is a stretch.
- If 10-year net survival is of interest, one should limit to individuals under 90
- If 15-year net survival is of interest, one should limit to individuals under 85.

Note that: if a very small proportion of individuals from a certain subgroup remains, they get a very large weight. If no individuals from a certain subgroup remain at risk, their values do not get represented in the average (they are taken to be equal to average)