Epidemiology for PhD students

Bendix Carstensen Steno Diabetes Center Copenhagen
Gentofte, Denmark
http://BendixCarstensen.com/EpiPhD/F2019

Department of Biostatistics, University of Copenhagen, Spring 2019

From /home/bendix/teach/Epi/KU-epi/slides/slides.tex Monday 8th April, 2019, 16:46

Splitting the follow-up

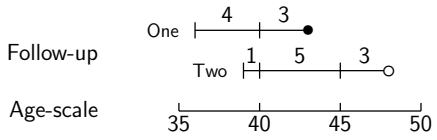
Tuesday 9 April 2019

Epidemiology for PhD students
Department of Biostatistics, University of Copenhagen,Spring 2019

http://BendixCarstensen.com/EpiPhD/F2019 rec-split

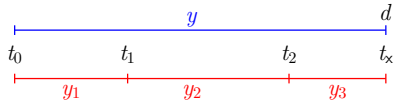
Stratification by age

- If follow-up is rather short, age at entry is OK for age-stratification.
- If follow-up is long, use stratification by categories of **current age**, both for no. of events, D, and risk time, Y.
- — (D, Y) is the fundamental observation in follow-up studies.



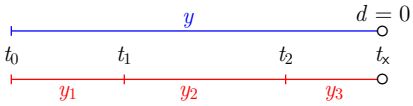
Representation of follow-up data

- In a cohort study we have records of (**Events**,**Risk time**).
- Follow-up data for each individual must have (at least) three variables:
  - Date of entry — `entry` — date variable.
  - Date of exit — `exit` — date variable
  - Status at exit — `fail` — indicator-variable (0/1)
- Specific for each **type** of outcome.



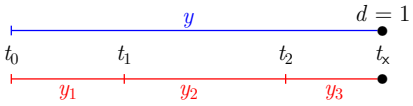
Probability	log-Likelihood
$\mathrm{P}(d \text{ at } t_x \mid \text{entry } t_0)$	$d\log(\lambda) - \lambda y$
$= \mathrm{P}(\text{surv } t_0 \to t_1 \mid \text{entry } t_0)$	$= 0\log(\lambda) - \lambda y_1$
$\times \mathrm{P}(\text{surv } t_1 \to t_2 \mid \text{entry } t_1)$	$+ 0\log(\lambda) - \lambda y_2$
$\times \mathrm{P}(d \text{ at } t_x \mid \text{entry } t_2)$	$+ d\log(\lambda) - \lambda y_3$

. . . assuming that the rate λ is constant



Probability	log-Likelihood
$\mathrm{P}(\text{surv } t_0 \to t_x \mid \text{entry } t_0)$	$0\log(\lambda) - \lambda y$
$= \mathrm{P}(\text{surv } t_0 \to t_1 \mid \text{entry } t_0)$	$= 0\log(\lambda) - \lambda y_1$
$\times \mathrm{P}(\text{surv } t_1 \to t_2 \mid \text{entry } t_1)$	$+ 0\log(\lambda) - \lambda y_2$
$\times \mathrm{P}(\text{surv } t_2 \to t_x \mid \text{entry } t_2)$	$+ 0\log(\lambda) - \lambda y_3$

. . . assuming that the rate λ is constant



Probability	log-Likelihood
$\mathrm{P}(\text{event at } t_x \mid \text{entry } t_0)$	$1\log(\lambda) - \lambda y$
$= \mathrm{P}(\text{surv } t_0 \to t_1 \mid \text{entry } t_0)$	$= 0\log(\lambda) - \lambda y_1$
$\times \mathrm{P}(\text{surv } t_1 \to t_2 \mid \text{entry } t_1)$	$+ 0\log(\lambda) - \lambda y_2$
$\times \mathrm{P}(\text{event at } t_x \mid \text{entry } t_2)$	$+ 1\log(\lambda) - \lambda y_3$

. . . assuming that the rate λ is constant

Aim of dividing time into bands:

- Compute rates in different bands of:
  - age
  - calendar time
  - disease duration
  - . . .
- Allow rates to vary along the timescale:

$$\begin{array}{r} 0\log(\lambda) - \lambda y_1 \\ +\,0\log(\lambda) - \lambda y_2 \\ +\,d\log(\lambda) - \lambda y_3 \end{array} \quad \longrightarrow \quad \begin{array}{r} 0\log(\lambda_1) - \lambda_1 y_1 \\ +\,0\log(\lambda_2) - \lambda_2 y_2 \\ +\,d\log(\lambda_3) - \lambda_3 y_3 \end{array}$$

Prerequisites of splitting time

Origin: The date where the time scale is 0:

- Age — 0 at date of birth
- Disease duration — 0 at date of diagnosis
- Occupation exposure — 0 at date of hire
- Time scale is always **time since** some origin.

Intervals: How should the scale be subdivided:

- 1-year classes? 5-year classes?
- Equal length — not necessarily.

Cohort with 3 persons:

```
Id    Bdate    Entry     Exit St
 1 14/07/52 04/08/65 27/06/97  1
 2 01/04/54 08/09/72 23/05/95  0
 3 10/06/87 23/12/91 24/07/98  1o
```

- Define strata: 10-years intervals of current **age**.
- Split Y for every subject accordingly
- Treat each segment as a separate unit of observation.
- Keep track of exit status, D, in each interval.

Splitting the follow up

	subj. 1	subj. 2	subj. 3
Age at **E**ntry:	13.06	18.44	4.54
Age at e**X**it:	44.95	41.14	11.12
Status at exit:	Dead	Alive	Dead
Y	31.89	22.70	6.58
D	1	0	1

Where did the pieces go?

	subj. 1		subj. 2		subj. 3		$\sum$	
Age	Y	D	Y	D	Y	D	Y	D
0–	0.00	0	0.00	0	5.46	0	5.46	0
10–	6.94	0	1.56	0	1.12	1	8.62	1
20–	10.00	0	10.00	0	0.00	0	20.00	0
30–	10.00	0	10.00	0	0.00	0	20.00	0
40–	4.95	1	1.14	0	0.00	0	6.09	1
$\sum$	31.89	1	22.70	0	6.58	1	60.17	2

Time-splitting with SAS: %Lexis

```
%Lexis( data=a, entry=Entry, exit=Exit, fail=St,
        origin=bdate, scale=365.25, breaks=0 to 80 by 10 ) ;
```

```
id       Bdate       Entry        Exit  St     risk left

 1  14/07/1952  03/08/1965  14/07/1972   0   6.9432   10
 1  14/07/1952  14/07/1972  14/07/1982   0  10.0000   20
 1  14/07/1952  14/07/1982  14/07/1992   0  10.0000   30
 1  14/07/1952  14/07/1992  27/06/1997   1   4.9528   40
 2  01/04/1954  08/09/1972  01/04/1974   0   1.5606   10
 2  01/04/1954  01/04/1974  31/03/1984   0  10.0000   20
 2  01/04/1954  31/03/1984  01/04/1994   0  10.0000   30
 2  01/04/1954  01/04/1994  23/05/1995   0   1.1417   40
 3  10/06/1987  23/12/1991  09/06/1997   0   5.4634    0
 3  10/06/1987  09/06/1997  24/07/1998   1   1.1211   10
```


Time-splitting with Stata stset, stsplit

```
stset  Exit,  failure(St==1)  entry(Entry)  origin(Bdate) /*
          */  scale(365.25) id(Id)

stsplit cAge, at(40(10)70) after(Bdate)

gen py = _t - _t0

table cAge, c(sum _d sum py) format(%9.2f)
```

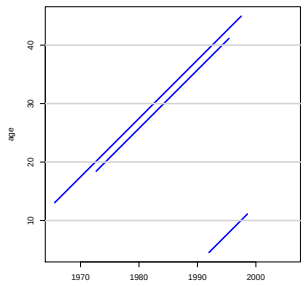

Time-splitting with R Lexis, splitLexis

```
library( Epi )

Lx <- Lexis( entry = list( per = Entry,
                           age = Entry-Bdate ),
             exit = list( per = Exit ),
       exit.status = factor( St, labels=c("Alive","Dead") ),
             data = coh )

Ls <- splitLexis( Lx, breaks=seq(0,100,10), time.scale="age" )
```

```
lex.id      per    age lex.dur lex.Cst lex.Xst Id    Bdate    Entry     Exit St
     1 1965.589 13.056   6.943   Alive   Alive  1 1952.533 1965.589 1997.485  1
     1 1972.533 20.000  10.000   Alive   Alive  1 1952.533 1965.589 1997.485  1
     1 1982.533 30.000  10.000   Alive   Alive  1 1952.533 1965.589 1997.485  1
     1 1992.533 40.000   4.952   Alive    Dead  1 1952.533 1965.589 1997.485  1
     2 1972.686 18.439   1.560   Alive   Alive  2 1954.246 1972.686 1995.388  0
     2 1974.246 20.000  10.000   Alive   Alive  2 1954.246 1972.686 1995.388  0
     2 1984.246 30.000  10.000   Alive   Alive  2 1954.246 1972.686 1995.388  0
     2 1994.246 40.000   1.141   Alive   Alive  2 1954.246 1972.686 1995.388  0
     3 1991.974  4.536   5.463   Alive   Alive  3 1987.437 1991.974 1998.559  1
     3 1997.437 10.000   1.121   Alive    Dead  3 1987.437 1991.974 1998.559  1
```

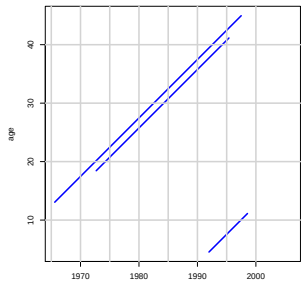

Time-splitting with R Lexis, splitLexis

```
plot( Ls, col="blue", lwd=3 )
```



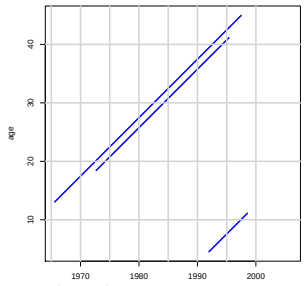
Time-splitting with R Lexis, splitLexis

```
Ls <- splitLexis( Ls, breaks=seq(1900,2000,5), time.scale="per" )
plot( Ls, col="blue", lwd=3 )
```



Time-splitting with R Lexis, splitMiltu

```
library( popEpi )
Ls <- splitMulti( Ls, age=seq(0,100,10), per=seq(1900,2000,5),
plot( Ls, col="blue", lwd=3 )
```



What happens when splitting time?

- **From:** one record per person
- **To:** many records per person,
- — each representing a short piece of follow-up time.
- **Same** total no. events
- **Same** total follow-up time (PYs)
- Likelihood contributions from intervals from one person are **conditionally** independent
- Likelihood contributions from different persons are independent
- $\Rightarrow$ D variates can be treated as independent Poisson variates with mean λY

What happens when splitting time?

- **From:** one record per person
- **To:** many records per person,
- $\Rightarrow$ allows different rates in different intervals.
- start point of an interval represented on all time scales:
  - what is the age here
  - what date is it here
  - what is the disease duration here
  - ...
- $\Rightarrow$ allows modeling of rates as continuous function of the **timescales** as represented in each interval

Your turn now: IHD data

The following exercise is designed to illustrate how follow-up time is subdivided in order to produce the table of events and person-years. Furthermore the aim is to show you that tabulated data and time-split data gives the same results if only age and exposure are used as variables.

We will first analyze frequency records as above (these are almost identical to Table 22.6 in C & H). Next, we shall read the individual records and construct the corresponding table of cases and person-years.

1. Import the program `ihd-lexis-sol.sas` (from the folder `http://bendixcarstensen.com/EpiPhD/F2019/sas`) to the program editor.
   Run the first part of the program — the part reading the tabulated data and `proc genmod`. Compare with the results from the results table in Clayton & Hills.
2. Next, use the second part of the program to read the individual records from the file `diet.txt`, including the `proc print` and check on the output that it looks reasonable and that you understand what each line in the data represents.

Time-splitting with SAS I

3. Now you should import the macro %Lexis and use it to split into the age intervals 40–50, 50–60 and 60–70 years:
   In order to use this you must first load it from the appropriate folder folder on the net:

```
* This will list the included code in your log-window ;
options source2 ;

filename lexispr url "http://BendixCarstensen.com/Lexis/Lexis.sas";
%inc lexispr ;
```

   Once you have specified `%inc lexispr ;` and run that line in SAS, SAS will know the macro %lexis and you can use it in the rest of the session.

Time-splitting with SAS II

4. The time-splitting is now done by running the SAS-macro %Lexis
   A SAS-macro is a piece of SAS-program (normally quite long) where certain small parts of the program can be changed when the program is run. The SAS-convention is that names of such programs start with a "%".
   To use the %Lexis macro we must specify the follow-up information from the input file:
   - Date of entry into the study — `doe`
   - Date of exit from the study — `dox`
   - Status at exit from the study — `chd` (1 if CHD occurred at `dox`, 0 otherwise).

   Moreover, we must decide which timescale to split the data on. In this case we want to split along the scale "current age", i.e. time since date of birth.

Time-splitting with SAS III

5. To this end we must specify:
   - The **origin** of the time-scale, i.e. where the time-scale is 0, in this case date of birth — `dob`.
   - The **intervals** where we want the follow-up grouped, here ages 40–50, 50–60 and 60–70.
   - As a purely technical thing we need to specify the conversion between the **scale** in which time is measured in the input dataset (in this case days) and in the specification of the grouping (in this case years) — 365.25.

   In the case of %Lexis we must supply these 6 parameters in order to specify how to split time.
   Finally we must tell the program where the original data is, where the time-split data has to go, and what the name of the age-variable should be.

Time-splitting with SAS IV

This looks like this (you do not have to write the stuff between the `/*...*/`):

```
%Lexis( data   = ihdindiv,        /* Dataset with original data       */
        out    = ihdsplit,        /* Dataset where time-split data go */
        entry  = doe,             /* Date of entry                    */
        exit   = dox,             /* Date of exit                     */
        fail   = chd,             /* Event (failure) indicator        */
        origin = dob,             /* Origin of the time-scale         */
        breaks = 40 to 70 by 10,  /* Where to split the time scale    */
        scale  = 365.25,          /* Conversion from days to years:   */
                                  /* from: scale of entry/exit        */
                                  /*   to: scale of breaks            */
        left   = agr );           /* The name of the new age-variable */
```

Run this piece of SAS code.

(In the top of the file `http://BendixCarstensen.com/Lexis/Lexis.sas` are some more detailed explanations of how to use %Lexis).

Tabulation of time-split data with SAS I

6. How many records are in the resulting dataset (`ihdsplit`)
7. Take at look at the resulting data file, for example the first 20 records:

```
proc print  data = ihdsplit (obs=20) ;
run ;
```

   How does this compare with the the original dataset?

8. Use %PYtab to tabulate IHD-cases and person-years by exposure and age-group. You must first get this from the net as you did with the %Lexis macro:

```
filename pytabpr url "http://BendixCarstensen.com/Lexis/PYtab.sas";
%inc pytabpr ;
```


Tabulation of time-split data with SAS II

Once you have imported the macro you can use it:

```
%PYtab( data  = ihdsplit,
        class = exposure agr,
        fail  = chd,
        risk  = risk,
        scale = 1000 ) ;
```

Compare with the sums from the table given in the first data step in `ihd-lexis.sas`

What about the Cox-model?

Data for Cox-regression has only one record per person:

- Assumes (the baseline) rate to vary arbitrarily over time
- — internally in the program, the data is split
- Time-dependent covariates require multiple records per person
- Additional time-scales require multiple records per person
- Main time scale and other time scales modeled differently

What happens when splitting time?

We are actually mimicking a **continuous** surveillance of the study population — the smaller the intervals, the closer we get.

For each little piece of follow up we attach the relevant covariates:

- Fixed covariates. (sex, genotype, . . .)
- Deterministically time-varying covariates:
  age, time since entry, calendar time
- Non-deterministically varying covariates:
  (current smoking habits, occupational exposure, . . .)

Analysis of results from %Lexis

- D — events in the variable `fail`.
- Y — risk time = difference: `exit - entry`.
  Enters in the model via $\log(Y)$ as offset.
- Covariates are:
  - timescales (age, calendar time, time since entry)
  - other variables for this person
    (constant or *assumed* constant in each interval).
- Model rates using the covariates in `proc genmod`
- Note: there is no difference in how time-scales and other covariates are treated in the model — they are all covariates.

From split to aggregate data

- Each interval contribute $d\log(\lambda) - \lambda y$ to the log-likelihood.
- All intervals with the same set of covariate values (age,exposure,...) have the same λ.
- The log-likelihood contribution from these is

$$\sum d\log(\lambda) - \lambda \sum y = D\log(\lambda) - \lambda Y$$

  — the same as from **aggregated** data.
- The log-likelihood is the same for split data and aggregated data — no need to tabulate first.
- (...except possibly for computing time)

Your turn again:

9. Use `proc genmod` to estimate the effect of age and exposure from the split dataset. How does the estimates compare with those based on the initially tabulated dataset?
10. Add an interaction between age and exposure and check that you get the same test for interaction as with the grouped data.
11. Compare the type 3 likelihood ratio statistic (`Chi-square`) for the interaction with the deviance of the model without interaction for the grouped data.