Splitting the follow-up
C&H 6

Bendix Carstensen

Steno Diabetes Center
& Department of Biostatistics, University of Copenhagen

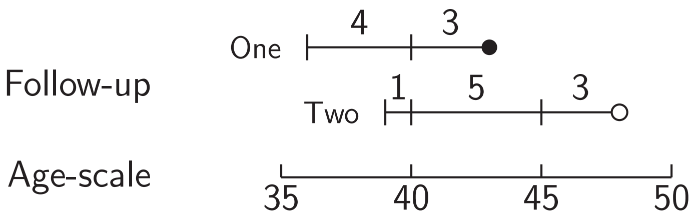
bxc@steno.dk
http://BendixCarstensen.com

PhD-course in Epidemiology,
Department of Biostatistics,
Tuesday 23 March 2015

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.



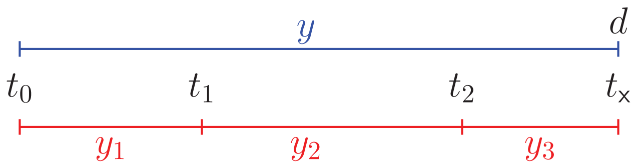
Representation of follow-up data

In a cohort study we have records of:
Events and **Risk time**.

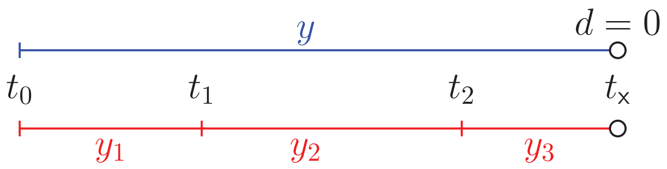
Follow-up data for each individual must have (at least) three variables:

- Date of entry — `entry` — date variable.
- Date of exit — `exit` — date variable
- Status at exit — `fail` — indicator-variable (0/1)

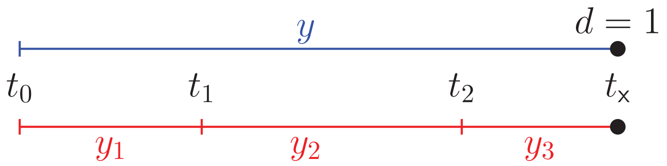
Specific for each *type* of outcome.



Probability	log-Likelihood
$\mathrm{P}(d \text{ at } t_\text{x}\|\text{entry } t_0)$	$d\log(\lambda) - \lambda y$
$= \mathrm{P}(\text{surv } t_0 \rightarrow t_1\|\text{entry } t_0)$	$= 0\log(\lambda) - \lambda y_1$
$\times\, \mathrm{P}(\text{surv } t_1 \rightarrow t_2\|\text{entry } t_1)$	$+\, 0\log(\lambda) - \lambda y_2$
$\times\, \mathrm{P}(d \text{ at } t_\text{x}\|\text{entry } t_2)$	$+\, d\log(\lambda) - \lambda y_3$



Probability	log-Likelihood
$\mathrm{P}(\text{surv } t_0 \rightarrow t_\text{x}\|\text{entry } t_0)$	$0\log(\lambda) - \lambda y$
$= \mathrm{P}(\text{surv } t_0 \rightarrow t_1\|\text{entry } t_0)$	$= 0\log(\lambda) - \lambda y_1$
$\times\, \mathrm{P}(\text{surv } t_1 \rightarrow t_2\|\text{entry } t_1)$	$+\, 0\log(\lambda) - \lambda y_2$
$\times\, \mathrm{P}(\text{surv } t_2 \rightarrow t_\text{x}\|\text{entry } t_2)$	$+\, 0\log(\lambda) - \lambda y_3$



Probability	log-Likelihood
$\mathrm{P}(\text{event at } t_\text{x}\|\text{entry } t_0)$	$1\log(\lambda) - \lambda y$
$= \mathrm{P}(\text{surv } t_0 \rightarrow t_1\|\text{entry } t_0)$	$= 0\log(\lambda) - \lambda y_1$
$\times\, \mathrm{P}(\text{surv } t_1 \rightarrow t_2\|\text{entry } t_1)$	$+\, 0\log(\lambda) - \lambda y_2$
$\times\, \mathrm{P}(\text{event at } t_\text{x}\|\text{entry } t_2)$	$+\, 1\log(\lambda) - \lambda y_3$

Aim of dividing time into bands:

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

$$
\begin{array}{rcr}
0\log(\lambda) - \lambda y_1 & & 0\log(\lambda_1) - \lambda_1 y_1 \\
+\,0\log(\lambda) - \lambda y_2 & \rightarrow & +\,0\log(\lambda_2) - \lambda_2 y_2 \\
+\,d\log(\lambda) - \lambda y_3 & & +\,d\log(\lambda_3) - \lambda_3 y_3
\end{array}
$$

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

Intervals: How should it 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  1
```

- 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 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

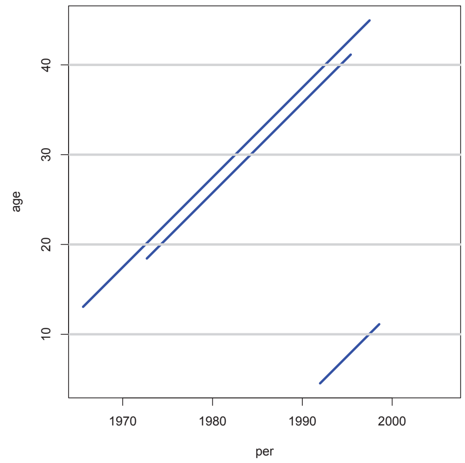
```
stset  Exit,  failure(St==1)  entry(Entry)  origin(Bdate) /*
          */  scale(365.25) id(Id)

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    En
     1 1965.589 13.056   6.943   Alive   Alive  1 1952.533 1965.
     1 1972.533 20.000  10.000   Alive   Alive  1 1952.533 1965.
     1 1982.533 30.000  10.000   Alive   Alive  1 1952.533 1965.
     1 1992.533 40.000   4.952   Alive    Dead  1 1952.533 1965.
     2 1972.686 18.439   1.560   Alive   Alive  2 1954.246 1972.
     2 1974.246 20.000  10.000   Alive   Alive  2 1954.246 1972.
     2 1984.246 30.000  10.000   Alive   Alive  2 1954.246 1972.
     2 1994.246 40.000   1.141   Alive   Alive  2 1954.246 1972.
     3 1991.974  4.536   5.463   Alive   Alive  3 1987.437 1991.
     3 1997.437 10.000   1.121   Alive    Dead  3 1987.437 1991.
```

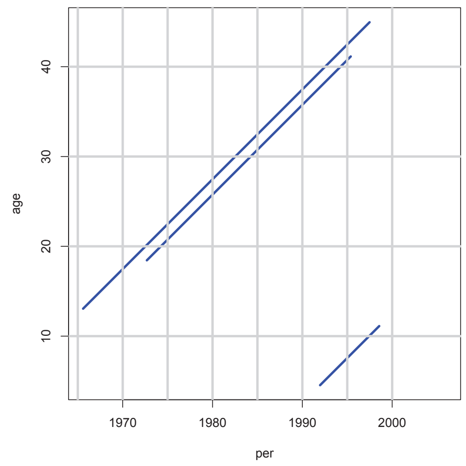
Time-splitting with R Lexis, splitLexis

```
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 )
```



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)
- Possibility of different rates in different intervals.

What about the Cox-model?

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

- It allows rates to vary over time (the baseline)
- — internally in the program, the data is split
- Time-dependent covariates require multiple records per person
- Additional time-scales require multiple records per person

What happens when splitting time?

We are actually mimicking a **continuous** surveillance of the study population.

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 — all derived from the current date.
- Non-deterministically varying covariates. (current smoking habits, occupational exposure, ...)

Models for time-split data

For follow-up data we make linear models for:

$$\eta = \log(\lambda Y) = \log(\lambda) + \log(Y)$$

by telling the software that D is Poisson.

If the model for the rate λ is multiplicative:

$$\begin{aligned} \lambda &= \exp(x_1\beta_1 + x_2\beta_2 + \cdots) \\ \log(\lambda Y) &= x_1\beta_1 + x_2\beta_2 + \cdots + \log(Y) \end{aligned}$$

Among the covariates are some that model the time-effect (in the IHD-example, age).

Models for time-split data

Relation to the Cox-model:

$$\begin{aligned} \lambda(t) &= \lambda_0(t)\exp(x_1\beta_1 + x_2\beta_2 + \cdots) \\ &= \exp(\log(\lambda_0(t)) + x_1\beta_1 + x_2\beta_2 + \cdots) \\ &= \exp(z_1\alpha_1 + z_2\alpha_2 + \cdots + x_1\beta_1 + x_2\beta_2 + \cdots) \end{aligned}$$

Covariates z_1, z_2, ... represent time, time^2, etc.; possibly splines.

“Among the covariates are some that model the time-effect (in the IHD-example, age).”

The baseline hazard — unspecified in the Cox-model — is replaced by a parmetric function, $\exp(z_1\alpha_1 + z_2\alpha_2 + \cdots)$

Independent observations?

When we split data, each individual contributes several observations, which are not independent.

Yet, we treat them as such.

The likelihood contribution from one person is a **product** of **conditional** probabilities.

Because the likelihood is a **product**, we can use the program (`proc genmod`, `glm`, ...) as if they were independent; we are only interested in getting the maximum likelihood estimates.

The offset

Need to take account of the "covariate" $\log(Y)$, which has a regression coefficient fixed to be one:

$$\log(\lambda Y) = x_1\beta_1 + x_2\beta_2 + \cdots + \log(Y)$$

$\log(Y)$ is called an **offset**-variable.

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.

Poisson model for split data

- Each interval contribute λY 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 $\lambda \sum Y$ — the same as from aggregated data.
- The event intervals contribute each $D \log \lambda$.
- The log-likelihood contribution from those with the same lambda is $\sum D \log \lambda$ — the same as from aggregated data.
- The log-likelihood is the same for split data and aggregated data — no need to tabulate first.

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.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 table xx 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 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://www.biostat.ku.dk/~bxc/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 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 with time-split data     */
        entry  = doe,             /* Date of entry                    */
        exit   = dox,             /* Date of exit                     */
        fail   = chd,             /* Event (failure) indicator        */
        breaks = 40 to 70 by 10,  /* Where to split the time scale    */
        origin = dob,             /* Origin of the time-scale         */
        scale  = 365.25,          /* Conversion from days to years    */
        left   = agr );           /* The name of the new age-variable */
```

Run this piece of SAS code.

(In the top of the file `http://www.biostat.ku.dk/~bxc/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://www.biostat.ku.dk/˜bxc/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`