Epidemiologic Data Analysis using R
Part 5: Time-splitting and SIR

Janne Pitkäniemi
(Esa Läärä)

Finnish Cancer Registry, Finland, <janne.pitkaniemi@cancer.fi>
(University of Oulu, Finland, <esa.laara@oulu.fi>)

University of Tampere
Faculty of Social Sciences
Feb 26- Apr 9 2018

Contents

1. Special exposure cohorts
2. Observed and expected cases, SIR & SMR
3. Lexis diagram and life lines
4. Splitting follow-up time simultaneously by age and period
5. Merging reference rates with the cohort data and performing SIR/SMR computations

Main R functions to be covered

- `Lexis.diagram()` and other `Lexis` tools in `Epi`
- `merge()`

Special cohorts of exposed subjects

- Occupational cohorts, exposed to potentially hazardous agents
- Cohorts of patients on chronic medication, which may have harmful long-term side-effects

No internal comparison group of unexposed subjects.

Question: Do incidence or mortality rates in the *exposed* target cohort differ from those of a roughly comparable *reference* population?

Reference rates obtained from:

- population statistics (mortality rates)
- disease & hospital discharge registers (incidence)

Accounting for age distribution

- Compare rates in a study cohort with a standard set of age-specific rates from the reference population.
- Reference rates normally based on large numbers of cases, so they are assumed to be "known" without error.
- Calculate **expected** number of cases, E, if the standard age-specific rates had applied in our study cohort.
- Compare this with the **observed** number of cases, D, by the **standardized incidence ratio** SIR (or standardized mortality ratio SMR)

$$\text{SIR} = D/E, \qquad \text{SE}[\log(\text{SIR})] = 1/\sqrt{D}$$

Example: HT and breast ca.

- A cohort of 974 women treated with hormone (replacement) therapy were followed up.
- $D = 15$ incident cases of breast cancer were observed.
- Person-years (Y_a) and reference rates (λ_a^*, per 100000 y) by age group (a) were:

Age	Y_a	λ_a^*	E_a
40–44	975	113	1.10
45–49	1079	162	1.75
50–54	2161	151	3.26
55–59	2793	183	5.11
60–64	3096	179	5.54
$\sum$			16.77

Example: HT use and breast ca. (cont'd)

- "Expected" number of cases at ages 40–44:

$$975 \times \frac{113}{100\,000} = 1.10$$

- Total "expected" cases is $E = 16.77$
- The SIR is $15/16.77 = 0.89$.
- Error factor: $\exp(1.96 \times \sqrt{1/15}) = 1.66$
- 95% confidence interval is:

$$0.89 \overset{\times}{\div} 1.66 = (0.54, 1.48)$$

A statistical model for SIR

- The theoretical rates λ_{ap} by age (a) and calendar period (p) in the cohort are assumed to be proportional to the rates λ_{ap}^* in the reference population:

$$\lambda_{ap} = \rho \times \lambda_{ap}^*$$

  ρ = hazard ratio btw the cohort and the reference pop'n.
- The population rates λ_{ap}^* are assumed to be known.
- Cohort data: numbers of cases D_{ap} and p-years Y_{ap} by *a*ge and *p*eriod are computed.
- It can be shown that the likelihood of ρ is of Poisson type, and the maximum likelihood estimator of ρ is:

$$\widehat{\rho} = \frac{D}{\sum \lambda_{ap}^* Y_{ap}} = \frac{\text{Observed}}{\text{Expected}} = \text{SIR}$$

Example: The Welsh Nickel Workers' Study

- A cohort of 679 men working in nickel smelters in South Wales first employed 1903-25 (for details see **B&D**).
- Outcomes of interest: deaths from nasal (ICD code 160) and lung cancer (ICD 162 and 163) during follow-up 1934-76.
- Outcome event indicator and basic time variables:

`icd`	=	code for cause of death, 0 if not yet dead
`date.bth`	=	date of birth
`date.in`	=	date of starting follow-up
`date.out`	=	date when follow-up ended

Example (cont'd)

- Interesting risk factors in the original data frame:

`expos` = exposure index based on years employed in high-risk areas in the smelter by 1925 $\rightarrow$ categorized version EXP

`date.1st` = date when first employed $\rightarrow$ AFE

- Risk factors to be formed from original variables:

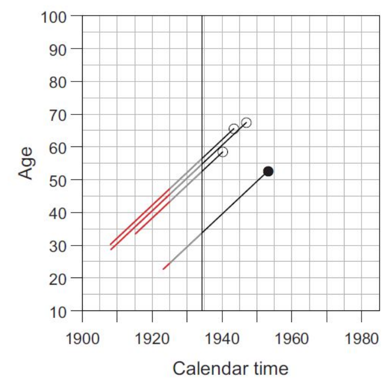
`age.1st` = age when first employed $\rightarrow$ AFE

`year.1st` = year of first employment $\rightarrow$ YFE

`time.1st` = time since first exposure $\rightarrow$ TFE

Lexis diagram & 4 lifelines from the nickel cohort

Diagram invented by *Wilhelm Lexis* (1837-1914), German mathematician and demographer, professor in Tartu 1874-76.

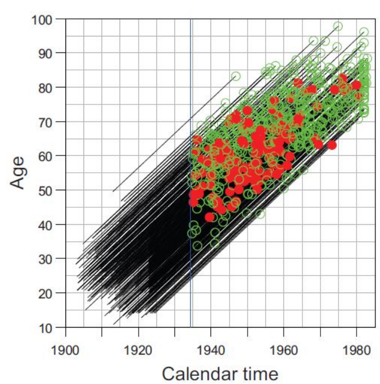


Individual lifelines run diagonally from a given (age, time) starting point to an endpoint.

Here the lines go from start of exposure till the age and time of exit.

Mortality follow-up started in 1934.

Nickel cohort: All lifelines in the Lexis diagram



Follow-up starts not until 1934 for all subjects.

- dot (red) = lung ca. death,
- circle (green) = censoring

Function `splitLexis()` splits individual follow-up times into rectangles defined by agebands and calendar periods.

Splitting follow-up by age & calendar time

from the registration of:

- Entry,
- Exit,
- Failure status

of the individuals in the cohort, and the definition of the scale by:

- **O**rigin
- **S**cale
- **C**utpoints

to the table of:

- D = events,
- Y = person time,

by age and period.

Expected numbers in practice

- From the records of age-period split & expanded cohort data:

  $y_{i,ap}$ = person-time slot in a record defined by

  a = ageband of the record

  p = period of the record

- From the file containing the reference rates:

  λ^*_{ap} = age & period specific rate

  a = ageband of the population rate

  p = period of the population rate

Expected numbers in practice (cont'd)

Population rates are matched up to the expanded cohort data, and expected numbers individually are computed as:

$$e_{i,ap} = \lambda^*_{ap} \times y_{i,ap}$$

and these are eventually summed: $E = \sum e_{i,ap}$

Always two datasets are needed for SIR:

1. the *cohort* data with follow-up information on its individual members. This must be split & expanded to match with
2. the *reference rate* data with age & period specific rates in the chosen reference population.

SMR-calculations in R using Lexis tools:

1. Read in the cohort data (Welsh Nickel Workers)

and convert the dates dd/mm/yyyy into "decimal years"

```
> nick <- read.table( "nickel.txt",
+         header=T, as.is=T )
> for (j in 4:7 ) nick[ , j] <-
+   cal.yr( nick[ , j], format = "%d/%m/%Y" )
```

List the records for the 4 men in a previous Lexis diagram

```
> round(nick[11:14, ],2)
   id icd expos date.bth date.1st date.in date.out
11 19 160   10.0  1881.73  1915.18 1934.25  1940.21
12 21  14    0.0  1877.80  1908.00 1934.25  1943.37
13 22 177    2.5  1879.50  1908.17 1934.25  1946.98
14 23 162    0.0  1900.50  1923.15 1934.25  1953.20
```


2. Reference rates in E & W read in

```
> ewrates <- read.table("ewrates.txt",header=T)
> ewrates[c(1:8, 143:150), ]
```

8 first and last rows checked

```
    year age lung nasal  other
1   1931  10    1     0   1269
2   1931  15    2     0   2201
3   1931  20    6     0   3116
4   1931  25   14     0   3024
5   1931  30   30     1   3188
6   1931  35   68     1   4165
7   1931  40  149     3   5651
143 1976  45  403     3   4311
144 1976  50 1003     9   7687
145 1976  55 1896     9  12544
146 1976  60 3342    15  20787
147 1976  65 4985    17  33729
148 1976  70 6718    20  55480
149 1976  75 8068    38  89199
150 1976  80 7744    33 137360
```

E & W lung ca. death rates by age and period

```
> tapply(ewrates$lung, list("age" = ewrates$age,
         "year" = ewrates$year),sum)
```

age \ year	1931	1936	1941	1946	1951	1956	1961	1966	1971	1976
10	1	1	1	1	0	0	0	0	0	0
15	2	2	2	3	2	2	2	2	1	1
20	6	6	6	8	7	4	5	4	4	2
25	14	14	16	18	13	12	11	10	10	7
30	30	30	34	36	35	35	34	25	24	17
35	68	68	81	94	98	93	90	76	58	56
40	149	149	191	236	248	251	223	216	177	139
45	274	274	384	544	579	590	563	531	503	403
50	431	431	597	954	1224	1248	1221	1160	1070	1003
55	586	586	883	1350	2003	2317	2284	2201	2077	1896
60	646	646	1021	1717	2555	3315	3663	3695	3546	3342
65	636	636	970	1763	2926	3926	4844	5273	5174	4985
70	533	533	748	1400	2624	3878	4977	6210	6820	6718
75	464	464	631	1085	2069	3332	4513	5914	7273	8068
80	324	324	385	765	1416	2258	3417	4563	6089	7744

3. Creating and expanding the Lexis object

The data frame converted to a Lexis object in two time scales: year (calendar time) and age:

```
nickL <- Lexis( entry = list( year = date.in ),
                 exit = list( year = date.out,
                  age = date.out - date.bth ),
          exit.status = as.numeric( nick$icd %in% c(162, 163) ),
                 data = nick )
```

The Lexis object jointly split by age and period. Agebands and period bands are named like in the ewrates file – "left" means the lower cutpoint (1st year) of a band.

```
> nickL.a <- splitLexis(nickL, "age", br=seq(10,85,5) )
> nickL.ap<- splitLexis(nickL.a,"year",br=seq(1931,1981,5))
> nickL.ap$year <- timeBand(nickL.ap, "year", "left")
> nickL.ap$age <- timeBand(nickL.ap, "age", "left")
```


The expanded data frame viewed

```
> dim(nickL.ap)
[1] 6948   13    # 10-fold expansion!
> round( subset( nickL.ap, lex.id %in% 13:14)
+        [ , c(1:4,6,8,10,12,13)] ,2)
```

	lex.id	year	age	lex.dur	lex.Xst	icd	date.bth	date.in	date.out
90	13	1931	50	0.25	0	177	1879.5	1934.25	1946.98
91	13	1931	55	1.50	0	177	1879.5	1934.25	1946.98
92	13	1936	55	3.50	0	177	1879.5	1934.25	1946.98
93	13	1936	60	1.50	0	177	1879.5	1934.25	1946.98
94	13	1941	60	3.50	0	177	1879.5	1934.25	1946.98
95	13	1941	65	1.50	0	177	1879.5	1934.25	1946.98
96	13	1946	65	0.98	0	177	1879.5	1934.25	1946.98
97	14	1931	30	1.25	0	162	1900.5	1934.25	1953.20
98	14	1931	35	0.50	0	162	1900.5	1934.25	1953.20
99	14	1936	35	4.50	0	162	1900.5	1934.25	1953.20
100	14	1936	40	0.50	0	162	1900.5	1934.25	1953.20
101	14	1941	40	4.50	0	162	1900.5	1934.25	1953.20
102	14	1941	45	0.50	0	162	1900.5	1934.25	1953.20
103	14	1946	45	4.50	0	162	1900.5	1934.25	1953.20
104	14	1946	50	0.50	0	162	1900.5	1934.25	1953.20
105	14	1951	50	2.20	1	162	1900.5	1934.25	1953.20

4. Merging the cohort data with E&W rates

```
> nickLew.ap <- merge(nickL.ap, ewrates,
       by = c("age", "year"))  # key columns
> round(nickLew.ap[1:20, c(1:4,6:8,10,12,13,14) ],1)
```

	year	age	lex.id	lex.dur	lex.Xst	id	icd	date.bth	date.in	date.out	lung
1	1931	20	197	0.3	0	273	154	1909.5	1934.2	1965.4	6
2	1931	20	236	1.3	0	325	434	1910.5	1934.2	1953.5	6
3	1931	20	400	0.5	0	574	491	1909.7	1934.2	1980.4	6
4	1931	20	384	0.3	0	546	0	1909.5	1934.2	1982.0	6
5	1931	20	156	0.9	0	213	162	1910.1	1934.2	1973.2	6
6	1931	25	236	0.5	0	325	434	1910.5	1934.2	1953.5	14
7	1931	25	38	0.3	0	56	502	1904.5	1934.2	1956.1	14
8	1931	25	581	1.5	0	842	420	1905.7	1934.2	1973.9	14
9	1931	25	267	0.1	0	369	420	1904.3	1934.2	1974.7	14
10	1931	25	478	1.8	0	690	420	1906.5	1934.2	1961.6	14
11	1931	25	251	1.8	0	344	420	1908.9	1934.2	1977.2	14
12	1931	25	156	0.9	0	213	162	1910.1	1934.2	1973.2	14
13	1931	25	400	1.3	0	574	491	1909.7	1934.2	1980.4	14
14	1931	25	390	1.8	0	556	0	1908.6	1934.2	1982.0	14
15	1931	25	85	1.0	0	111	0	1905.2	1934.2	1982.0	14
16	1931	25	315	1.1	0	443	420	1905.3	1934.2	1971.1	14
17	1931	25	168	0.1	0	227	0	1904.3	1934.2	1982.0	14
18	1931	25	169	1.8	0	228	502	1906.9	1934.2	1978.6	14
19	1931	25	121	1.5	0	157	332	1905.8	1934.2	1980.5	14
20	1931	25	17	1.6	0	28	420	1905.8	1934.2	1967.4	14

5. Calculation of observed and expected

Cases & person-time slots renamed, expectations $\lambda^*_{ap} y_{i,ap}$ of becoming a case computed, and tables by a and p produced.

```
> nickLew.ap <- transform( nickLew.ap,
+           d_iap = lex.Xst, y_iap = lex.dur,
+           e_iap = lex.dur * lung/1.0E6 )

> Obs.lung <- with(nickLew.ap, tapply(d_iap,
+   list("age" = age, "year" = year), sum))

> Exp.lung <- with(nickLew.ap, tapply(e_iap,
+   list("age" = age, "year" = year), sum))

> Obs.lung ; round(Exp.lung,3)
```


Observed and expected numbers printed

```
    year
age  1931 1936 1941 1946 1951 1956 1961 1966 1971 1976
  30    0    0    0   NA   NA   NA   NA   NA   NA   NA
  35    0    0    0    0   NA   NA   NA   NA   NA   NA
  40    0    1    1    0    0   NA   NA   NA   NA   NA
  45    3    2    4    1    0    0   NA   NA   NA   NA
  50    1    5    3    7    6    2    0   NA   NA   NA
  55    0    5    6    6    4    5    1    0   NA   NA
  60    1    4    5    3   11    6    1    1    1   NA
  65    0    0    1    5    4    6    3    1    0    0
  70    0    0    1    3    0    2    2    1    0    0
  75   NA    0    0    0    0    1    1    2    1    3
  80   NA   NA    0    0    0    0    1    0    0    3
    year
age   1931  1936  1941  1946  1951  1956  1961  1966  1971  1976
  30 0.004 0.005 0.001    NA    NA    NA    NA    NA    NA    NA
  35 0.012 0.032 0.015 0.004    NA    NA    NA    NA    NA    NA
  40 0.027 0.075 0.090 0.045 0.011    NA    NA    NA    NA    NA
  45 0.054 0.135 0.184 0.246 0.110 0.025    NA    NA    NA    NA
  50 0.082 0.231 0.281 0.438 0.511 0.220 0.046    NA    NA    NA
  55 0.070 0.263 0.411 0.557 0.790 0.834 0.343 0.069    NA    NA
  60 0.035 0.162 0.362 0.644 0.880 1.108 1.155 0.502 0.104    NA
  65 0.004 0.045 0.178 0.481 0.775 1.015 1.314 1.240 0.539 0.122
  70 0.001 0.004 0.041 0.157 0.486 0.682 0.796 1.173 1.203 0.519
  75    NA 0.001 0.003 0.039 0.136 0.342 0.470 0.498 0.955 0.885
  80    NA    NA 0.001 0.001 0.037 0.098 0.158 0.218 0.293 0.536
```


6. Calculation of SMR

We can sum either over individual time slots:

```
> D <- sum(nickLew.ap$d_iap)
> E <- sum(nickLew.ap$e_iap)
```

or over the newly formed tables:

```
> D <- sum(Obs.lung, na.rm=T)
> E <- sum(Exp.lung, na.rm=T)
```

Either way, the calculation proceeds:

```
> SMR <- D/E; SE <- 1/sqrt(D); EF <- exp(1.96*SE)
> round(c(D, E, SMR, SMR/EF, SMR*EF), 2)
[1] 137.00  26.62   5.15   4.35   6.08
```

SMR = 5.15 [95% CI 4.35 to 6.08]
⇒ substantial excess risk of lung cancer in smelter workers.

Concluding remarks

- If specific exposure factors exist that have variable values within the target cohort, the estimation of rate ratios associated with them may be efficiently adjusted for age and calendar period by taking the age- and period-specific expected number as the baseline in Poisson-modelling.
- Follow-up time could be split yet by another relevant time axis, like time passed since start of exposure, which may be taken as an explanatory variable when modelling the effects of exposure within a cohort.
- The main challenge is to identify a sufficiently comparable reference population. The "general" population is rarely an ideal one.