Practical exercises:

Measures of Disease Occurrence
Analysis of Epidemiological Data

Nordic Summerschool of Cancer Epidemiology, 2013
Danish Cancer Society, 12–23 August, 2013
http://BendixCarstensen.com/NSCE

Version 3.2

Compiled Sunday 25th August, 2013, 14:04
from: C:/Bendix/undervis/NSCE/2013/pracs/pracs.tex

Bendix Carstensen — Steno Diabetes Center, Gentofte, Denmark
& Department of Biostatistics, University of Copenhagen
bxc@steno.dk
http://BendixCarstensen.com

Esa Läärä — Department of Mathematical Sciences
University of Oulu, Finland
Esa.Laara@oulu.fi
http://stat.oulu.fi/laara/

Contents

Chapter 1

Introduction to exercises

The exercises in this course requires you to do calculations which in principle can be done on a hand-calculator.

However we assume that you use your laptop and use R as a calculator. This will enable you to take the solutions with you home in the form of a file with computer code that does the analyses. It will also enable you to do analyses repeatedly on slightly different sets of data.

At the end of the course you will get a complete set of solution suggestions. Many of these will be quite elaborate, merely as an illustration of how to use the actually existing features in R to produce solutions. They should not be taken as indications of what we assume that you should be able to do.

So here is an indication of how you should use R:

1.1 What is R?

R is free program for data analysis and graphics. It contains all state of the art statistical methods, and has become the preferred analysis tool for most professional statisticians in the world. It can be used as simple calculator and as a very specialized statistical analysis and reporting machinery.

The special thing about R is that you enter commands from the keyboard into a console window, where you also see the results. This is an advantage because you end up with a script that you can use to *reproduce* your analyses—a requirement in any scientific endeavour.

The disadvantage is that you somehow have to find out what to type. The practicals will contain some hints, and you will mostly be using R as a calculator — type an expression, hit the return key and you get the result on your screen.

1.2 Getting R

You can obtain R, which is free, from CRAN (the **C**omprehensive **R** **A**rchive **N**etwork), at http://cran.r-project.org/. Under "`Download and Install R`" click on "`Windows`" and under "`R for Windows`" click on "`base`". Then on "`Download R 3.0.1 for Windows`", which is a self-extracting installer. This means that if you save it to your computer somewhere and click on it, it will install R for you.

Apart from what you have downloaded there are several thousand add-on packages to R dealing with all sorts of problems from ecology to fiance and incidentally, epidemiology. You must download these manually. In this course we shall only need the `Epi` package.

1.2.1 Starting R

You start R by clicking on the icon that the installer has put on your desktop. You should edit the properties of this, so that R starts in the folder that you have created on your computer for this course: Right-click on the R-icon, choose "Properties", and then in the field "Start in", enter the relevant folder-name.

Once you have installed R, start it, and in the menu bar click on Packages→Install package(s)..., chose a mirror (this is just a server where you can get the stuff), and the the `Epi` package.

Once R (hopefully) has told you that it has been installed, you can type:

```
> library( Epi )
```

to get access to the `Epi` package. You can get an overview of the functions and datasets in the package by typing:

```
> library( help=Epi )
```

1.2.2 Quitting R

Type `q()` in the console, and answer "No" when asked whether you want to save workspace image.

1.3 Working with the script editor

If you click on `File→New script`, R will open a window for you which is a text-editor very much like Notepad.

If you write a commands in it you can transfer then to the R console and have them executed by pressing `CTRL-r`. If nothing is highlighted, the line where the cursor is will be transmitted to the console and the cursor will move to the next line. If a part of the screen is highlighted the highlighted part will be transmitted to the console.

1.3.1 Try!

Now open a script by `File→New script`, and type (omit the "`>`" in the beginning of the line):

```
> 5+7
> pi
> 1:10
> N <- c(27,33,81)
> N
```

Run the lines one at a time by pressing `CTRL-r`, and see what happens.

You can also type the commands in the console directly. But then you will not have a record of what you have done. Well, you can press `File→Save History` and save all you typed in the console (including the 73.6% commands with errors).

1.4 Getting a bit more training

If you want to train a bit before the course, there is a nice work-book introduction to R here: http://www.mhills.pwp.blueyonder.co.uk/readme.html.

If you are interested in using R in epidemiology, there is "A short introduction to R", originally written for the European Educational Programme in Epidemiology (and for the IARC summer school in time trends in 2007). A revised version is at: http://pubhealth.ku.dk/~bxc/Epi/R-intro.pdf.

1.5 Changing the looks of R

If you want R to start up with a different font, different colors etc., the go to the folder where R is installed — most likely `Program Files\R\R-3.0.1`, the to the folder `etc`, and open the file `Rconsole` with Notepad. In the file are specifications on how R will look when you start it, pretty self-explanatory, except perhaps for `MDI`.

`MDI` means "`M`ultiple `D`isplay `I`nterface", which means you get a single R-window, and within that sub-windows with the console, the script editor, graphs etc. If this is set to "no", you get `SDI` which means "`S`ingle `D`isplay `I`nterface", which means that R will open the console, script editor etc. in separate windows of their own.

A white background can be trying to look at, so on my (BxC's) computer I use a bold font and the following colors:

```
> background = gray5
> normaltext = yellow2
> usertext = green
> pagerbg = gray5
> pagertext = yellow2
> highlight = red
> dataeditbg = gray5
> dataedittext = red
> datadituser = yellow2
> editorbg = gray5
> editortext = lightblue
```

(If you want to know which colors are available in R, just give the command `colors()`).

1.6 Further reading

On the CRAN web-site the last menu-entry on the left is "Contributed" and will take you to a very long list of various introductions to R, including manuals in esoteric languages such as Danish, Finnish and Hungarian.

Chapter 2

Measures of Disease Occurrence Exercises

2.0 NORDCAN demonstrations

2.0.1 Finding and opening NORDCAN

1. Launch your favourite browser, like *Firefox* or *Internet Explorer*.
2. Enter the website of the *Association of the Nordic Cancer Registries*: www.ancr.nu; when there, click on the link NORDCAN - on the Web.
3. On the next page, choose the English flag, leading you to the actual starting page of The NORDCAN Project: www-dep.iarc.fr/NORDCAN/english/frame.asp

2.0.2 Cancer fact sheet on lung cancer

Create a cancer fact sheet for lung cancer in all the Nordic countries together by appropriate choices from the pertinent menus on the left hand side. Find answers to the following questions:

1. What were the average annual numbers of new cases in men and women during 2006-10?
2. How big were the estimated risks of getting cancer by 75 years of age for the two genders?
3. How many men and women died each year from lung cancer during 2005-2009?
4. What were the numbers of men and women living with lung cancer at the end of 2010, and how big were the corresponding proportions of lung cancer patients out of the whole male and female populations?

2.0.3 Incidence of lung cancer

Learn more about the incidence rates of lung cancer among men in the Nordic Countries during 2006-2010. Go to ONLINE ANALYSIS on the left and click on *Incidence/Mortality*.

Proceed to **Tables** and after text Standardized rates by click Countries. From the pertinent boxes under the heading **Cancer/Years*** select first Lung and then pick up the requires years by simultaneously pushing `Ctrl` key when doing the latter selections year by year.

1. Where was the incidence highest, where lowest? What were the crude rates in these two regions?

2. Compare Finland and Norway. Can you find any real difference in the crude rates? What about the age-standardized rates with different standard populations? (The explanation for the standardized rates and for possible discrepancies between them and the crude rates will be given later on.)

2.0.4 Population size and person-years

Find out data on the population size and person-years, also by age, of all men in Finland in the early 1990s and compare them with the numbers given on lecture slide 22. For that purpose, go first to ONLINE ANALYSIS and click *Incidence/Mortality*. Then proceed down to Population pyramid and select Finland from the pertinent box.

1. Select year 1992 from the scroll-down menu box on the right and execute. Compare the population pyramids of men and women. Check out the total number of men and compare with the person-years given for that year on lecture slide 22.

2. Select years 1993 and 1994 simultaneously by pushing `Ctrl` key when picking the second one of these. Look at the total number on the bottom line of the table and compare with the person-years given for that year on lecture slide 22. Has the population size doubled?

2.0.5 Mortality from lung cancer

Learn more about the mortality rates of lung cancer among men in the Nordic Countries during 2005–2009. Proceed as in task 1.3 above (ONLINE ANALYSIS → *Incidence/Mortality*, *etc.*), but now complete the choices by changing the **Data type** into Mortality and execute.

1. Where was the mortality highest, where lowest? What were the crude rates in these two regions? Are they very different from the corresponding incidence rates in task 1.3 above?

2. Compare Island and Sweden. Can you find any real difference in the crude rates? What about the age-standardized rates with different standard populations?

2.0.6 Prevalence of lung cancer

Learn more about the prevalence of lung cancer among men in the Nordic Countries at the end of 2010. Under ONLINE ANALYSIS now click *Prevalence*. Then continue to **Tables by** and click on Countries. On the next page from the Cancer menu select Lung, and for the year choose 2010 from the pertinent boxes.

1. Where was the total prevalence highest, where lowest? What were the prevalence proportions in these two regions?

2. What was the prevalence proportion of cases diagnosed less than 1 year ago in all Nordic countries jointly?

3. What was the prevalence proportion of cases diagnosed at least 5 years ago in all Nordic countries jointly?

2.0.7 Lung cancer by age, period and cohort

We shall now look at incidence rates by different time scales as exemplified on lecture slides 46 to 50.

1. Create a graph showing the age specific incidence and mortality of lung cancer among men in Denmark during 2006-10. From *Incidence/Mortality*, under **Graphs** choose Age-specific curves. Any comments to the graph?

2. Repeat (a) for Finland and compare the curves between these two countries.

3. Create graphs describing age-incidence curves of lung cancer among males in Denmark for years 1955 and 2000. From *Incidence/Mortality*, under **Graphs** choose Age-specific curves. Select **Cancer/Sex** and **Country** accordingly. Select the years from the pertinent box by pushing `Ctrl` key when making the 2nd selection. Click on Individual years, and execute. Take a look at the graphs first on the linear scale. After that switch to the logarithmic scale by clicking on the gray text Toggle Arithmetic/Logarithmic scale. Compare these curves with the corresponding ones for Finland on lecture slide 47.

4. Create graphs describing trends in the age-specific incidence rates among males in Denmark. From *Incidence/Mortality*, under **Graphs** choose Time-trends by age. For Starting and Ending choose 1955 and 2000, respectively. Under **Age** for From choose 35-, for Interval choose 5, and for Smoothing choose 5 years and execute. When the curves appear, click on the gray text Toggle Arithmetic/Logarithmic scale. Compare these curves with the corresponding ones for Finland found on lecture slide 47.

5. Create graphs describing age-incidence curves by birth cohort of lung cancer among males in Denmark. From *Incidence/Mortality*, under **Graphs** choose Time-trends by cohort. Select **Cancer/Sex** and **Country** as above and **Age** to 84, and execute. When the curves appear, click on the gray text Age/Cohort (3). Compare these curves with the corresponding ones for Finland found on lecture slide 50. You will also notice that a similar table is displayed as on slide 46.

2.0.8 Crude and standardized rates: stomach cancer

Obtain the crude and standardized incidence rates of male stomach cancer in the Nordic countries for 2010.

1. In which country is the incidence highest when measured both by the crude rate and by all the different age-standardized rates?

2. Compare the age-standadized rate based on the World Standard Population of the country in (a) with those of Cali and Birmingham in the 1980s given on lecture slide 61.

3. Why are the standardized rates of type ASR(N) not much different from the crude rates? Why are the ASR(W) and ASR(E) lower when compared to ASR(N)?

2.0.9 Cumulative risk by 75 y: stomach cancer

Obtain the estimated cumulative risks of male stomach cancer by 75 years of age in the Nordic countries for 2010.

1. Where does this measure seem to be highest and where lowest, and how big they are?

2. Compare the figures between these countries with those of Cali and Birmingham on lecture slide 64.

2.0.10 Relative survival

Now we shall have a look at the prognosis of lung cancer patients when compared with the general population. Under ONLINE ANALYSIS proceed to *Survival*. On the next page under **Tables by** click on Country and period. A new page is opened on which under **Cancer** select Lung and under **Survival time** select 5-year.

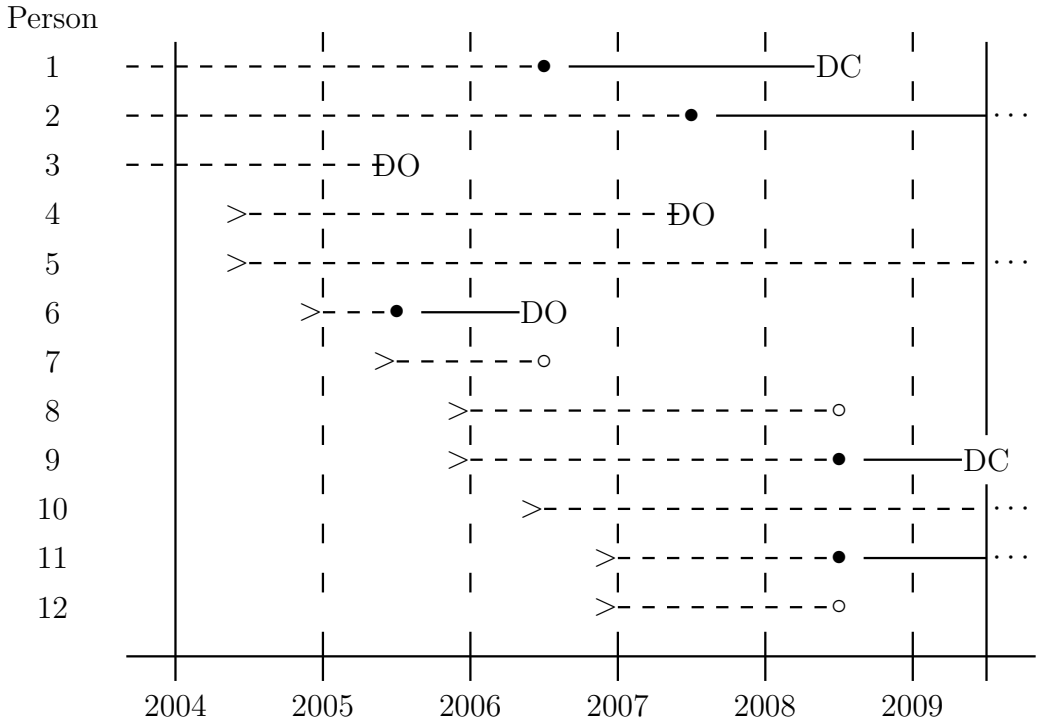
1. In which country was the relative survival poorest and where it was most favourable among male patients diagnosed in 2004–2008? What about female patients? How big where the 5-year relative survival proportions?

2. By how many percent points did the relative survival proportion improve in male patients of Norway during the 40 years since 1964-68?

3. Compare the relative survival between men and women overall. What is your general observation on the direction of the difference?

2.1 Follow-up of a small cohort

The figure below displays the follow-up experience of members of a small study cohort between 1 January 2004 to 30 June 2009 from entry ($>$) to follow-up until death (DC if due to disease C, DO for other causes) or censoring ($\circ$). For those subjects contracting disease C the time of diagnosis is also marked ($\bullet$ = onset of C).



We shall calculate the values of the incidence rate of the disease and of various mortality measures

(a) What is the incidence rate (per 100 y) of disease C during the period from 1 Jan 2004 to 31 Dec 2008? Organize the computations as follows:

0. Find out from the figure, what are the individual contributions (in years) of persons 1, 4, 5, and 12 to the total amount of person-time of follow-up pertinent to this task.
1. The total person-time is 27 years. Assign this to variable `Y.todis` writing and running the following command line: `Y.todis <- 27`
2. What is the total number of new cases of disease C? Assign this to variable `Cases` in the same way.
3. Obtain the incidence rate of C assigning its value into variable `Irate` and printing it as follows: `Irate <- 100*Cases/Y.todis ; Irate`;

(b) What is the mortality rate from disease C during the same period? Proceed with similar steps as above:

1. What is the total person time now? Is it the same as before, or more, or less? Assign this to variable `Y.todth` and run the command.
2. What is the total number of deaths from disease C? Assign this to variable `Dth.C`.

3. Assign the mortality rate of C into variable `Mrate.C` and print

(c) What is the mortality rate from all causes during the same period? Assign the total number of deaths into `Dth.all` and compute the total mortality rate `Mrate.all` applying the same principle as above.

(d) What is the estimated 3-year mortality proportion ("risk" of death for a risk period of 3 years since entry) based on the result in (c) and assuming the constant rate model? Apply the following command: `Mprop3.all <- 1 - exp( - (Mrate.all/100)*3 )` and print the result. – Why division by 100 is necessary here?

(e) What is the mortality rate `Mrate.pts` during the same period from all causes *among the patients with* C after the onset of C ? The person-years for this task can be obtained *e.g.* as follows: `Y.distodth <- Y.todth - Y.todis`; explain why. Count the pertinent number of deaths, compute the rate and print.

(f) What is the estimated 3-year mortality proportion `Mprop3.pts` after the onset of C among the patients with C?

(g) What is the prevalence of C on 30 September 2006, and on 31 December 2008? Find out the sizes of the populations `N1` and `N2` as well as the numbers of prevalent cases `C1 and C2` at the two time points, and compute the correponding prevalence proportions `P1` and `P2`. from these.

Why incidence or mortality proportions for 3-year or any other risk period, calculated by the formula presented on slides 18 and 20, would be problematic in tasks (a) and (b)?

2.2 Infant mortality

During 1978 in Finland 269 boys died at the age of <1 year. The size of this male age group was 33200 on 31 Dec 1977, and on 31 Dec 1978 it was 32500. The number of boys born alive during 1978 was 32800.

(a) Calculate the mortality rate (incidence rate of deaths) in this age group of boys by the usual method, person-years being computed by the mid-population principle; see lecture slides 21 and 22.

(b) In population statistics **infant mortality rate** (IMR) is defined:

$$\text{IMR} = \frac{\text{no. of deaths in age group} < 1 \text{ year during a calendar year}}{\text{no. of live born children during the year}}$$

Calculate the value of this measure for Finnish boys in 1978 from the given data and compare it with the result in item (a)

(c) Is the "infant mortality rate" in item (b) indeed a rate (density) as defined in the lectures – why or why not? Is it a proportion?

2.3 Incidence and mortality of leukaemia in children

In the table below are given the size (in 1000s) of the male population in Finland aged 0-14 years (the age range of "childhood" in pediatrics!) on the 31 December in each year from 1991 to 2000.

year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
population	493	495	496	497	496	495	491	485	481	478

The following numbers of cases describe the incidence and mortality of acute leukaemia in this population for two calendar periods: 5 years 1993 to 1997 (source: NORDCAN), and year 1999 only (source: Finnish Cancer Registry `http://www.cancerregistry.fi/`).

	1993-97	1999
new cases of acute leukaemia	113	26
deaths from acute leukaemia	22	3

(a) Calculate the incidence rates of acute leukaemia in this population for the two periods, person-years again computed from mid-populations.

(b) Calculate similarly the mortality rates of leukaemia.

(c) Is there evidence about any change in the incidence and/or mortality between these two periods?

(d) What would you conclude about the fatality of leukemia in children?

2.4 Rate ratio and rate difference in prevention trial

The Alpha Tocopherol Beta Caroten (ATBC) Prevention Trial (*N Engl J Med 1994*; **330**: 1029-35) addressed among other things the possible benefits of daily intake of vitamin E supplements in reducing the incidence of cancer among male smokers. The study population of 29133 regularly smoking 50-69 years old Finnish men were randomized into two groups: active treatment (vitamin E supplementation), and placebo (no supplementation). The following results were obtained for cancer of the prostate after an average follow-up time of 6 years:

treatment group	number of cases	incidence rate (per 10000 years)
vitamin E supplementation	99	11.6
no supplementation	151	17.8

(a) Calculate the person-years at risk in the two study groups separately.

(b) Estimate the incidence rate ratio ("relative risk") and incidence rate difference ("excess risk") measuring the effect of daily supplementation with vitamin E on the risk prostate cancer.

(c) Estimate either the excess fraction or preventive fraction, whichever more appropriate, to describe the proportional impact of vitamin E supplementation among those exposed to vitamin E.

(d) Discuss the results. What can be concluded from these estimates?

2.5 Rate ratio, rate difference and excess fraction

In the table next page the mortality rates (per 1000 pyrs, age-adjusted) from three important causes of death among life-long non-smokers and regular smokers were observed after 30 years follow-up of a large occupational cohort (men only).

	lung cancer	other lung diseases	cardiovascular diseases
smokers	2.0	3.0	15.0
non-smokers	0.2	1.0	9.0

(a) Calculate for each cause of death the following effect measures for comparison between smokers and non-smokers: rate difference, rate ratio, and excess fraction.

(b) Discuss the results. What can be inferred about the biological strength and the public health impact, respectively, of regular smoking regarding the three diseases.

2.6 Crude and standardized rates

Age specific data on the incidence of colon cancer in male and female populations of Finland during 1999 are given in the following table

	Males				Females				
Age group	Cases	Mid-popul. (1000s)	% of all	Rate ($/10^5$y)	Cases	Mid-popul. (1000s)	% of all	Rate ($/10^5$y)	Rate ratio M/F
0–34	10	1157	46.0	**0.9**	22	1109	41.9	**2.0**	0.44
35–54	76	809	32.0	**9.4**	68	786	29.7	**8.6**	1.09
55–74	305	455	18.0	**67**	288	524	19.8	**55**	1.22
75+	201	102	4.0	**196**	354	229	8.6	**155**	1.27
All	592	2523	100		732	2648	100		

Calculate the following summary measures:

(a) crude incidence rate in both populations and the rate ratio: males *vs.* females,

(b) age-standardized rates and their ratio using the male population as the standard,

(c) age-standardized rates and their ratio using the World Standard Population,

(d) cumulative rates up to 75 years and their ratio,

(e) estimated cumulative "risks" up to 75 years and their ratio.

Compare and comment the results obtained in items (a) to (e).

Hint: Organize the calculations needed for summary measures such that the necessary age-specific quantities are assigned into pertinent vectors, *e.g.* age-specific rates in women:

```
ratesF.a <- c(2.0, 8.6, 55, 155)
```

and weights from the male population:

```
wM <- c(46, 32, 18, 4)
```

and make use of the `sum()` function of R, for example, when computing the age-standardized rate for women in item (b):

```
stdRateF_wM <- sum( wM * ratesF.a ) / sum( wM )
```

2.7 Survival from tongue cancer

The survival experience of males in Finland with cancer of the tongue diagnosed during 1967-74 was studied by Hakulinen *et al.* (1981). Sizes of risk sets, numbers of deaths and losses (censorings) tabulated into 1 year subintervals since the diagnosis are given in the following table.

year	size of risk set	no. of deaths	no. of losses	effect. denom.	prop. deaths	prop. surviv.	cumul. survival
$0- < 1$	130	45	7				0.644
$1- < 2$	78	24	9	73.5		0.673	
$2- < 3$	45	5	7	41.5			0.382
$3- < 4$	33	2	6		0.067		
$4- < 5$	25	1	5				
$5- < 6$	19	-	7	15.5	0.0	1.0	0.340
$6- < 7$	12	-	6				

(a) Complete this table by appropriate figures using the actuarial life table method.

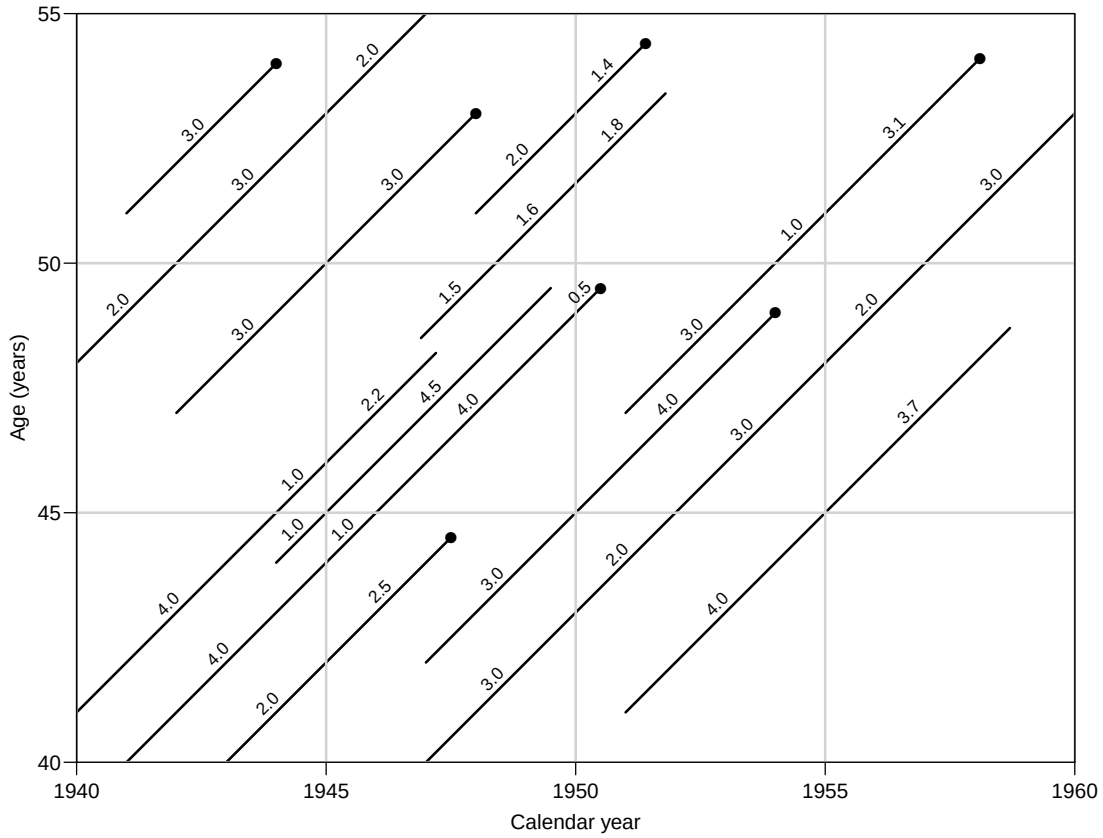
(b) Based on the results obtained above draw a survival curve and estimate graphically the median and the quartiles, if possible, of the survival time distribution.

2.8 Lexis diagram and occupational cohort I

In the Lexis diagram below displayed follow-up times of a small occupational cohort over the years 1940-1959 and the age range 40-54 years (this example is from **B&D**). Each line runs from the entry to follow-up until either the diagnosis of cancer (D), or censoring or withdrawal (W) due to death from other causes or migration.

(a) Calculate the numbers of new cases of cancer, and person-years at risk in all the three 5-year age-bands: 40-44, 45-49, and 50-54 years for each of the 5-year calendar periods 1940-44, 1945-49, and 1950-54 separately. *Advice.* Execute some division of labour in your group, so that not everybody is calculating these items for all periods.

(b) Calculate the numbers of new cases of cancer, person-years at risk in the three 5-year age groups: 40-44, 45-49, and 50-54 years for a *birth cohort* born in 1902-11.



2.9 Lexis diagram and occupational cohort II

Continuing exercise **8. (a)** above. The age-specific incidences (per 100000 person-years) in the three 5-year age-groups during 1940-54 in the whole population of the country were 100, 200, and 400, respectively, so there was no variation between the sub-periods. Assuming that this is an appropriate reference population, calculate the expected number of cases for the index occupational cohort for the same period. Compare the observed and expected number of cases by standardized incidence ratio, SIR. Comment on the result.

Chapter 3

Analysis of Epidemiological Data Exercises

3.1 Single incidence rates

In Kuwait during 1987 six deaths from stomach cancer were registered in males aged 45 to 54 years, and 89 000 men of this age group were living in the country at that time. In Egypt the corresponding figures in the same male age group during 1987 were 53 cases and 1 819 000 men. Calculate for both countries the following quantities:

1. mortality rate,
2. 95% confidence interval of the "true" rate based on SE of the rate (and error margin),
3. 95% confidence interval of the rate based on SE of the log-rate (and error factor). Compare this with the interval obtained in 2.

3.2 Non-significant difference

A cohort of electric engineers, graduated from a certain university of technology during a specified time interval, were followed-up over a period of 50 years. One out of the 10 female graduates and 1 out of the 200 male graduates developed breast cancer during the follow-up. The difference in the incidence between males and females was "not statistically significant" ($P > 0.05$).

How should this result be interpreted? Choose one from the following alternatives:

1. The results provide supporting evidence for the hypothesis no real difference between males and females in the breast cancer risk among electric engineers.
2. The results are consistent with the universal observation that the risk of breast cancer among females is clearly higher than that in males.
3. No conclusion can be made from this result concerning the male/female contrast in breast cancer incidence among graduates of electric engineering.
4. Other conclusion, what?

3.3 Preventive trial

Read the following abstract of the ATBC Cancer Prevention Study and Figure 2 in it (here shown as figure 1), displaying its major results on cancer incidence, and do the following tasks:

1. State the study hypothesis and the corresponding null hypothesis concerning the effect of receiving daily beta carotene supplements vs. not receiving them on the incidence of lung cancer.

2. Calculate the person-years in the group receiving ceta carotene supplements (the "exposed") and in the group receiving placebo ("unexposed").

3. Calculate the point estimate and the 95% confidence interval for the hazard rate ratio $\rho = \lambda_1/\lambda_0$ of lung cancer between the exposed and the unexposed.

4. Calculate the point estimate and the 95% confidence interval for the hazard rate difference $\delta = \lambda_1 - \lambda_0$ of lung cancer between the exposed and the unexposed.

5. Calculate a test statistic and the associated P value corresponding to the null hypothesis stated in item (a).

6. Discuss the results. Can the estimated relative rate be confounded by age and/or smoking, as the analysis was not stratified by these factors?

The Effect of Vitamin E and Beta Carotene on the Incidence of Lung Cancer and Other Cancers in Male Smokers

The Alpha-Tocopherol Beta Carotene Cancer Prevention Study Group

Background: Epidemiologic evidence indicates that diets high in carotenoid-rich fruits and vegetables, as well as high serum levels of vitamin E (alpha-tocopherol) and beta carotene, are associated with a reduced risk of lung cancer.

Methods: We performed a randomized, double-blind, placebo-controlled primary-prevention trial to determine whether daily supplementation with alpha-tocopherol, beta carotene, or both would reduce the incidence of lung cancer and other cancers. A total of 29,133 male smokers 50 to 69 years of age from southwestern Finland were randomly assigned to one of four regimens: alpha-tocopherol (50 mg per day) alone, beta carotene (20 mg per day) alone, both alpha-tocopherol and beta carotene, or placebo. Follow-up continued for five to eight years.

Results: Among the 876 new cases of lung cancer diagnosed during the trial, no reduction in incidence was observed among the men who received alpha-tocopherol (change in incidence as compared with those who did not, −2 percent; 95 percent confidence interval, −14 to 12 percent). Unexpectedly, we observed a higher incidence of lung cancer among the men who received beta carotene than among those who did not (change in incidence, 18 percent; 95 percent confidence interval, 3 to 36 percent). We found no evidence of an interaction between alpha-tocopherol and beta carotene with respect to the incidence of lung cancer. Fewer cases of prostate cancer were diagnosed among those who received alpha-tocopherol than among those who did not. Beta carotene had little or no effect on the incidence of cancer other than lung cancer.

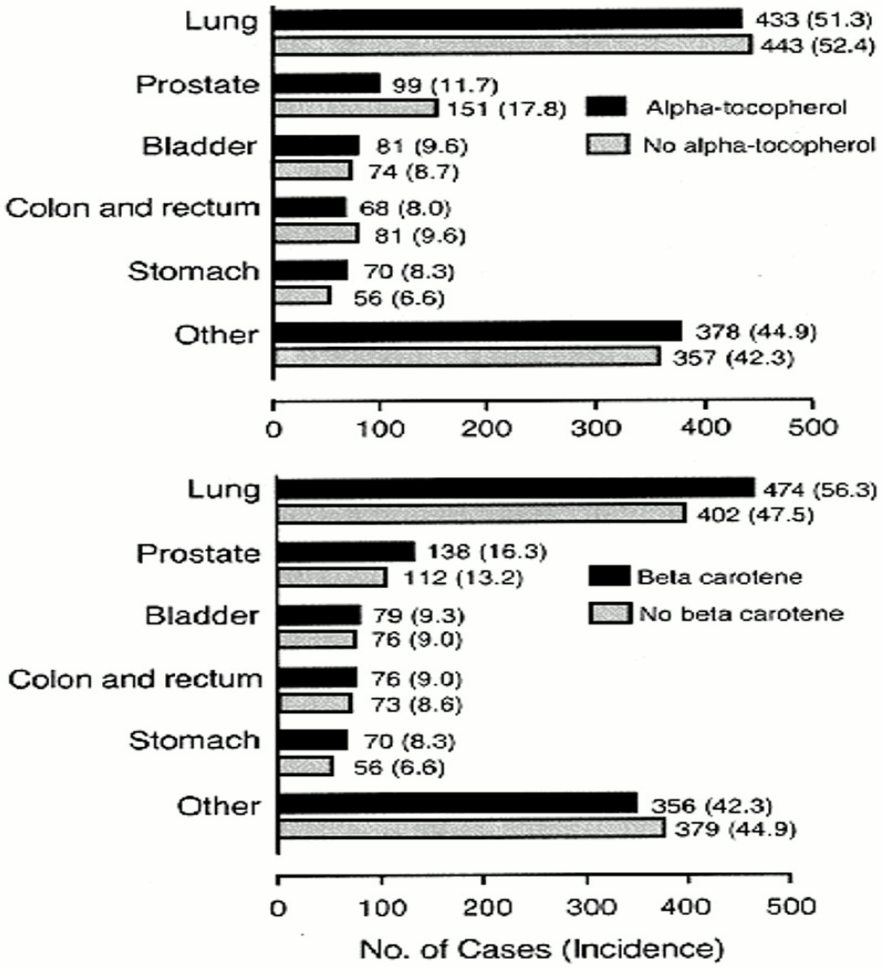


Figure 3.1: *Number and Incidence (per 10 000 Person-Years) of Cancers, According to Site, among Participants Who Received Alpha-Tocopherol Supplements and Those Who Did Not (Upper Panel) and among Participants Who Received Beta Carotene Supplements and Those Who Did Not (Lower Panel).*

Alpha-tocopherol had no apparent effect on total mortality, although more deaths from hemorrhagic stroke were observed among the men who received this supplement than among those who did not. Total mortality was 8 percent higher (95 percent confidence interval, 1 to 16 percent) among the participants who received beta carotene than among those who did not, primarily because there were more deaths from lung cancer and ischemic heart disease.

Conclusions: We found no reduction in the incidence of lung cancer among male smokers after five to eight years of dietary supplementation with alpha-tocopherol or beta carotene. In fact, this trial raises the possibility that these supplements may actually have harmful as well as beneficial effects.

(*New England Journal of Medicine*, Volume 330, pp. 1029–1035, April 14, 1994, Number 15).

3.4 Preventive trial – interpretation

We continue with the ATBC Cancer Prevention Study complementing its results with those of two other randomized trials that addressed the same hypothesis on the possible beneficial effect of beta caroten supplementation on lung cancer incidence.

1. In the ATBC study the observed rate ratio of lung cancer associated with daily intake of beta caroten supplement appeared to be "statistically significantly" different from 1 ($P = 0.01$). However, the direction of the estimated rate ratio was opposite to that of the original study hypothesis, which was based on the observational evidence that motivated the trial. – Do you think that this result provides a sufficient basis to conclude that beta caroten supplementation is actually harmful?

2. In the *Beta Carotene and Retinol Efficacy Trial* conducted in USA, a total of 18314 smokers, former smokers, and workers exposed to asbestos were randomized into two groups: active-treatment group and placebo group (*N Engl J Med* 1996; 334: 1150-1155). The active-treatment group received a combination of 30 mg of beta carotene per day and 25000 IU of retinol (vitamin A) in the form of retinyl palmitate per day. After a follow-up of 4.0 years on average, the active-treatment group had a relative rate of lung cancer of 1.28 (95 % CI, 1.04 to 1.57; $P = 0.02$) as compared with the placebo group. – Taken this result together with that of the ATBC trial, what can we now say about the accumulated evidence on the effects of beta caroten on the incidence of lung cancer among smokers? Would we now be more convinced about the harmfulness of this form of vitamin supplementation?

3. A third beta caroten trial was conducted in a study population of 22071 male American physicians (*N Engl J Med* 1996; 334: 1145-1149). After 13 years follow-up the point estimate of the rate ratio of lung cancer between the beta caroten and the placebo groups among the subset of current smokers in that study population was 0.9, *i.e.* lower than 1 but "non-significant" (95% CI 0.58-1.40, $P = 0.63$). – Is this result in conflict with the results of the two other trials quoted above?

4. In the American physicians' study, among *nonsmokers* the observed rate ratio of lung cancer between beta caroten and placebo groups was 0.78 (95% CI 0.34-1.79, $P = 0.56$). – What can we conclude about the effect of beta caroten supplementation in non-smoking men on the basis of these results? Is it different from that among regular smokers?

3.5 Geographical variation

Geographical variation in the incidence of certain form of cancer D in a country C was mapped using two classifications for dividing the area: (a) by county, and (b) by central hospital district. In the figure 2 the adjusted incidences (per 100,000 person years) of D are given for certain areas according to both divisions.

In addition are given stars indicating that the figure in question is significantly different ($p < 0.01$) from the average incidence of D in the whole country, which was 1 per 100,000 person-years. The two divisions seem to give somewhat contradictory results. How can we explain this apparent paradox?

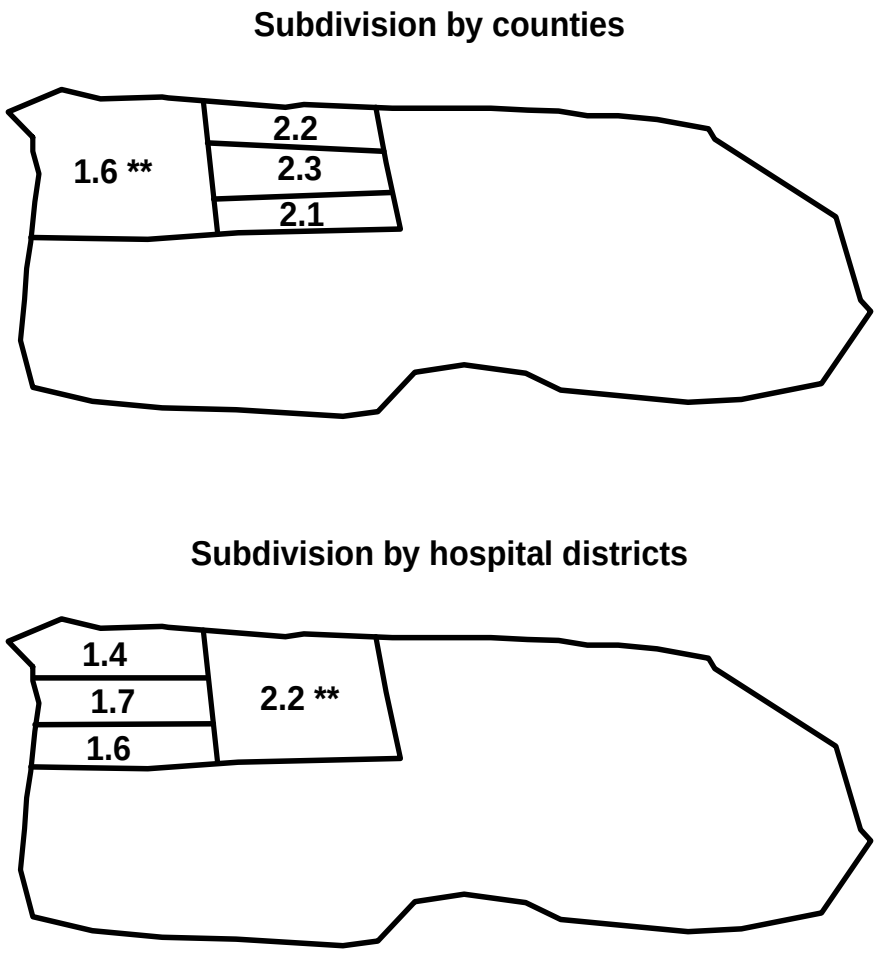


Figure 3.2: *Geographical division by county (top) and hospital district (bottom).*

3.6 Efficiency of study design

You are designing a cohort study to estimate the relative risk associated with a certain exposure factor X. Initially you are planning to recruit 10 000 persons to the cohort, such that 2000 would be exposed and 8000 unexposed to X, and you intend to have a 5 year follow-up period. A statistician points out that the confidence interval of your relative risk estimate is likely to be too wide. You cannot afford to enroll more than 10 000 individuals to the cohort. How could you change your research plan in principle such that the confidence interval would become shorter without increasing the total number of study subjects?

3.7 Case-control study: MI

In the table below are results presented from an unmatched case-control study on the association between physical activity (PA) and risk of myocardial infarction (MI) stratified by gender.

Gender	PA index	Cases	Controls	Total
Men	2500+ kcals	141	208	349
	< 2500 kcals	144	112	256
Total		285	320	605
Women	2500+ kcals	49	58	107
	< 2500 kcals	32	45	77
Total		81	103	184
Both	2500+ kcals	190	266	456
	< 2500 kcals	176	157	333
Total		366	423	789

1. Calculate the point estimate (and the 95% confidence interval) of the rate ratio in both genders separately.
2. What can you say of the possible modification of the effect of PA by gender; is the relative risk different in males than in females?
3. Is gender a confounder for the association between PA and MI; on what grounds?
4. Calculate the crude point estimate of the rate ratio, unadjusted for gender.
5. Calculate the gender-adjusted Mantel-Haenszel summary estimate of the rate ratio (and its 95 % confidence interval). Compare this with the crude one.

3.8 Case-control study: Neonates

Cnattingius *et al.* (*JNCI 1995*; 87 (June 21): 908-914) reported a case-control study on prenatal and neonatal risk factors for childhood lymphatic leukaemia in children. From the National Cancer Register of Sweden they collected all cases of this disease reported in children under 15 years of age from 1973 through 1989. Five controls for each case, matched for age and gender, were obtained from the Medical Birth Register of Sweden. The data on potential risk factors in both cases and controls were obtained from the latter register, too.

One of the findings was that 8 children with leukaemia and 2 of the control children had Down's syndrome.

1. On the basis of this information only, can you obtain any reasonable approximations for the following quantities:
   (a) a crude estimate of the relative hazard of leukemia in children with Down's syndrome as compared with children without this chromosome abnormality,
   (b) an approximate 95% confidence interval for the hazard ratio. What assumptions are needed in order that these approximations would be credible?
2. What additional data would be needed to obtain adequate estimates and confidence intervals?

3.9 Matched case-control study: Chemicals

A certain chemical exposure E was studied as a potential risk factor of cancer D in a case-control study with 20 cases and 20 controls. The following observations were made on the exposure status (+ = exposed, − = nonexposed) of each case and control:

No.	case	control	No.	case	control
1.	+	−	11.	−	+
2.	+	−	12.	+	+
3.	−	−	13.	+	−
4.	+	+	14.	−	−
5.	−	+	15.	+	−
6.	+	−	16.	+	−
7.	+	−	17.	+	−
8.	+	−	18.	+	+
9.	+	+	19.	−	−
10.	−	−	20.	+	−

1. Calculate the point estimate (with the approximate 95% confidence interval) of the hazard rate ratio associated with the exposure, as well as the test statistic and P-value corresponding to the null hypothesis of no effect, assuming that the study subjects have been obtained

   (a) by choosing the control group as a random sample of the source population of the cases without any matching, so that cases and controls labelled with the same ordinal number above are not related to each other,

   (b) by choosing for each case patient an individual control subject with the same age, and gender, such that each control is matched with the case having the same ordinal number above.

2. What appears to be the consequence to the rate ratio estimate here, if matching was applied in collecting the data but ignored in the analysis?

3.10 Cohort study and SMR

An occupational cohort study was started to estimate cancer mortality among male employees having a history of been working in a certain industry I during a certain time period, comparing it with that in a reference population which comprised economically active males at the same socioeconomic level living in the same area but not working in industry I. The results are displayed in the table on the next page. Calculate the following quantities:

1. Age-specific mortality rates in both populations and their ratios between the I-employees and the reference population. Does the rate ratio appear heterogenous over the age groups?

2. Crude mortality rates in the two populations and their ratio.

3. Mantel-Haeszel summary estimate of the rate ratio.

4. Standardised mortality ratio (SMR).

5. Standardised mortality rates in the populations and their ratio using the reference population as the standard.

6. Are the rate ratio estimates sensitive to the choice of standard population? Is age a confounder in these analyses?

	Employees in I		Reference population	
Age group	Deaths	Person-years	Deaths	Person years
30–39	11	10,000	15	30,000
40–49	15	6,000	60	50,000
50–59	10	2,000	150	70,000
Total	36	18,000	225	150,000

3.11 Trial of tolbutamide

The effect of treating middle-aged and elderly diabetic subjects with a drug called tolbutamide vs. placebo as investigated in a famous randomised clinical trial (University Group Diabetes Program 1970). During a fixed follow-up period of 5 years with no losses, 30 out of the 204 patients randomised to tolbutamide died, and 21 out of the 215 patients in the placebo group died, too.

1. Calculate the following quantities:

   (a) Incidence proportions (cumulative incidences) of death in both groups.

   (b) Estimate of the risk ratio with its approximate 95% confidence interval between tolbutamide and placebo.

   (c) Estimate of the risk difference and its approximate 95% confidence interval between tolbutamide and placebo.

2. Is tolbutamide dangerous to diabetics?

Chapter 4

Basic concepts in survival and demography

The following is a *very* condensed overview of concepts and requires some familiarity with probability theory.

The target audience for this is

- mathematicians and statisticians who want to get an overview of how the various concepts in probability translates to epidemiological conepts
- advanced epidemiologists who wants a handy overview of the mathematical relationships between the familiar concepts

This section briefly summarizes relations between various quantities used in analysis of follow-up studies. They are used all the time in the analysis and reporting of results. Hence it is important to be familiar with all of them and the relation between them.

4.1 Probability

Survival function:

$$\begin{aligned} S(t) &= \mathrm{P}\left\{\text{survival at least till } t\right\} \\ &= \mathrm{P}\left\{T > t\right\} = 1 - \mathrm{P}\left\{T \leq t\right\} = 1 - F(t) \end{aligned}$$

Conditional survival function:

$$\begin{aligned} S(t|t_{\text{entry}}) &= \mathrm{P}\left\{\text{survival at least till } t| \text{ alive at } t_{\text{entry}}\right\} \\ &= S(t)/S(t_{\text{entry}}) \end{aligned}$$

Cumulative distribution function of death times (cumulative risk):

$$\begin{aligned} F(t) &= \mathrm{P}\left\{\text{death before } t\right\} \\ &= \mathrm{P}\left\{T \leq t\right\} = 1 - S(t) \end{aligned}$$

Density function of death times:

$$f(t) = \lim_{h\to 0} \mathrm{P}\left\{\text{death in } (t, t+h)\right\}/h = \lim_{h\to 0} \frac{F(t+h)-F(t)}{h} = F'(t)$$

Intensity:

$$\begin{aligned}\lambda(t) &= \lim_{h\to 0} \mathrm{P}\left\{\text{event in } (t, t+h] \mid \text{alive at } t\right\}/h \\ &= \lim_{h\to 0} \frac{F(t+h)-F(t)}{S(t)h} = \frac{f(t)}{S(t)} \\ &= \lim_{h\to 0} -\frac{S(t+h)-S(t)}{S(t)h} = -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t}\end{aligned}$$

The intensity is also known as the hazard function, hazard rate, rate, mortality/morbidity rate.

Relationships between terms:

$$\begin{aligned}-\frac{\mathrm{d}\log S(t)}{\mathrm{d}t} &= \lambda(t) \\ &\Updownarrow \\ S(t) &= \exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right) = \exp\bigl(-\Lambda(t)\bigr)\end{aligned}$$

The quantity $\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s$ is called the *integrated intensity* or the **cumulative rate**. It is *not* an intensity, it is dimensionless.

$$\lambda(t) = -\frac{\mathrm{d}\log(S(t))}{\mathrm{d}t} = -\frac{S'(t)}{S(t)} = \frac{F'(t)}{1-F(t)} = \frac{f(t)}{S(t)}$$

The cumulative *risk* of an event (to time t) is:

$$F(t) = \mathrm{P}\left\{\text{Event before time } t\right\} = \int_0^t \lambda(u)S(u)\,\mathrm{d}u = 1 - S(t) = 1 - \mathrm{e}^{-\Lambda(t)}$$

For small $|x|$ (< 0.05), we have that $1 - \mathrm{e}^{-x} \approx x$, so for small values of the integrated intensity:

$$\text{Cumulative risk to time } t \approx \Lambda(t) = \text{Cumulative rate}$$

4.2 Statistics

Likelihood from one person:

The likelihood from a number of small pieces of follow-up from one individual is a product of conditional probabilities:

$$\begin{aligned}\mathrm{P}\left\{\text{event at } t_4 | \text{entry at } t_0\right\} = {} & \mathrm{P}\left\{\text{event at } t_4 | \text{ alive at } t_3\right\} \times \\ & \mathrm{P}\left\{\text{survive } (t_2, t_3) | \text{ alive at } t_2\right\} \times \\ & \mathrm{P}\left\{\text{survive } (t_1, t_2) | \text{ alive at } t_1\right\} \times \\ & \mathrm{P}\left\{\text{survive } (t_0, t_1) | \text{ alive at } t_0\right\}\end{aligned}$$

Each term in this expression corresponds to one *empirical rate*[1] $(d, y) = (\#\text{deaths}, \#\text{risk time})$, i.e. the data obtained from the follow-up of one person in the interval of length y. Each person can contribute many empirical rates, most with $d = 0$; d can only be 1 for the *last* empirical rate for a person.

Log-likelihood for one empirical rate (d, y):

$$\ell(\lambda) = d \log(\lambda) - \lambda y$$

This is under the assumption that the underlying rate (λ) is constant over the interval that the empirical rate refers to.

Log-likelihood for several persons. Adding log-likelihoods from a group of persons (only contributions with identical rates) gives:

$$D \log(\lambda) - \lambda Y,$$

where Y is the total follow-up time, and D is the total number of failures.

Note: The Poisson log-likelihood for an observation D with mean λY is:

$$D \log(\lambda Y) - \lambda Y = D \log(\lambda) + D \log(Y) - \lambda Y$$

The term $D \log(Y)$ does not involve the parameter λ, so the likelihood for an observed rate can be maximized by pretending that the no. of cases D is Poisson with mean λY. But this does *not* imply that D follows a Poisson-distribution. It is entirely a likelihood based computational convenience. Anything that is not likelihood based is not justified.

A linear model for the log-rate, $\log(\lambda) = X\beta$ implies that

$$\lambda Y = \exp\big(\log(\lambda) + \log(Y)\big) = \exp\big(X\beta + \log(Y)\big)$$

Therefore, in order to get a linear model for λ we must require that $\log(Y)$ appear as a variable in the model for $D \sim (\lambda Y)$ with the regression coefficient fixed to 1, a so-called offset-term in the linear predictor.

4.3 Competing risks

Competing risks: If there is more than one, say 3, causes of death, occurring with (cause-specific) rates λ_1, λ_2, λ_3, that is:

$$\lambda_c(a) = \lim_{h \to 0} \mathrm{P}\left\{\text{death from cause } c \text{ in } (a, a+h] \mid \text{alive at } a\right\}/h, \quad c = 1, 2, 3$$

The survival function is then:

$$S(a) = \exp\left(-\int_0^a \lambda_1(u) + \lambda_2(u) + \lambda_3(u)\,\mathrm{d}u\right)$$

[1] This is a concept coined by BxC, and so is not necessarily generally recognized.

because you have to escape any cause of death. The probability of dying from cause 1 before age a (the cause-specific cumulative risk) is:

$$\mathrm{P}\left\{\text{dead from cause 1 at } a\right\} = \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u \neq 1 - \exp\left(-\int_0^a \lambda_1(u)\,\mathrm{d}u\right)$$

The term $\exp(-\int_0^a \lambda_1(u)\,\mathrm{d}u)$ is sometimes referred to as the "cause-specific survival", but it does not have any probabilistic interpretation in the real world. It is the survival under the assumption that only cause 1 existed an that the mortality rate from this cause was the same as when the other causes were present too.

Together with the survival function, the cause-specific cumulative risks represent a classification of the population at any time in those alive and those dead from causes 1,2 and 3 respectively:

$$1 = S(a) + \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u + \int_0^a \lambda_2(u)S(u)\,\mathrm{d}u + \int_0^a \lambda_3(u)S(u)\,\mathrm{d}u, \quad \forall a$$

Subdistribution hazard Fine and Gray defined models for the so-called subdistribution hazard. Recall the relationship between between the hazard (λ) and the cumulative risk (F):

$$\lambda(a) = -\frac{\mathrm{d}\log\big(S(a)\big)}{\mathrm{d}a} = -\frac{\mathrm{d}\log\big(1-F(a)\big)}{\mathrm{d}a}$$

When more competing causes of death are present the Fine and Gray idea is to use this transformation to the cause-specific cumulative risk for cause 1, say:

$$\tilde{\lambda}_1(a) = -\frac{\mathrm{d}\log\big(1-F_1(a)\big)}{\mathrm{d}a}$$

This is what is called the subdistribution hazard, it depends on the survival function S, which depends on *all* the cause-specific hazards:

$$F_1(a) = \mathrm{P}\left\{\text{dead from cause 1 at } a\right\} = \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u$$

The subdistribution hazard is merely a transformation of the cause-specific cumulative risks. Namely the same transformation which in the single-cause case transforms the cumulative risk to the hazard.

4.4 Demography

Expected residual lifetime: The expected lifetime (at birth) is simply the variable age (a) integrated with respect to the distribution of age at death:

$$\mathrm{EL} = \int_0^\infty a f(a)\,\mathrm{d}a$$

where f is the density of the distribution of lifetimes.

The relation between the density f and the survival function S is $f(a) = -S'(a)$, and so integration by parts gives:

$$\mathrm{EL} = \int_0^\infty a\big(-S'(a)\big)\,\mathrm{d}a = -\Big[aS(a)\Big]_0^\infty + \int_0^\infty S(a)\,\mathrm{d}a$$

The first of the resulting terms is 0 because $S(a)$ is 0 at the upper limit and a by definition is 0 at the lower limit.

Hence the expected lifetime can be computed as the integral of the survival function.

The expected residual lifetime at age a is calculated as the integral of the *conditional* survival function for a person aged a:

$$\mathrm{EL}(a) = \int_a^\infty S(u)/S(a)\,\mathrm{d}u$$

Lifetime lost due to a disease is the difference between the expected residual lifetime for a diseased person and a non-diseased (well) person at the same age. So all that is needed is a(n estimate of the) survival function in each of the two groups.

$$\mathrm{LL}(a) = \int_a^\infty S_{\mathrm{Well}}(u)/S_{\mathrm{Well}}(a) - S_{\mathrm{Diseased}}(u)/S_{\mathrm{Diseased}}(a)\,\mathrm{d}u$$

Note that the definition of the survival function for a non-diseased person requires a decision as to whether one will consider non-diseased persons immune to the disease in question or not. That is whether we will include the possibility of a well person getting ill and subsequently die. This does not show up in the formulae, but is a practical consideration to have in mind when devising an estimate of S_{Well}.

Lifetime lost by cause of death is using the fact that the difference between the survival probabilities is the same as the difference between the death probabilities. If several causes of death (3, say) are considered then:

$$\begin{aligned} S(a) = 1 &- \mathrm{P}\left\{\text{dead from cause 1 at } a\right\} \\ &- \mathrm{P}\left\{\text{dead from cause 2 at } a\right\} \\ &- \mathrm{P}\left\{\text{dead from cause 3 at } a\right\} \end{aligned}$$

and hence:

$$\begin{aligned} S_{\mathrm{Well}}(a) - S_{\mathrm{Diseased}}(a) = &\ \mathrm{P}\left\{\text{dead from cause 1 at } a|\text{Diseased}\right\} \\ &+ \mathrm{P}\left\{\text{dead from cause 2 at } a|\text{Diseased}\right\} \\ &+ \mathrm{P}\left\{\text{dead from cause 3 at } a|\text{Diseased}\right\} \\ &- \mathrm{P}\left\{\text{dead from cause 1 at } a|\text{Well}\right\} \\ &- \mathrm{P}\left\{\text{dead from cause 2 at } a|\text{Well}\right\} \\ &- \mathrm{P}\left\{\text{dead from cause 3 at } a|\text{Well}\right\} \end{aligned}$$

So we can conveniently define the lifetime lost due to cause 2, say, by:

$$\begin{aligned} \mathrm{LL}_2(a) = \int_a^\infty &\ \mathrm{P}\left\{\text{dead from cause 2 at } u|\text{Diseased \& alive at } a\right\} \\ &-\mathrm{P}\left\{\text{dead from cause 2 at } u|\text{Well \& alive at } a\right\}\,\mathrm{d}u \end{aligned}$$

These will have the property that their sum is the years of life lost due to total mortality differences:

$$\mathrm{LL}(a) = \mathrm{LL}_1(a) + \mathrm{LL}_2(a) + \mathrm{LL}_3(a)$$

The term in the integral are computed as (see the section on competing risks):

$$\mathrm{P}\left\{\text{dead from cause 2 at } u | \text{Diseased \& alive at } a\right\} = \int_a^u \lambda_{2,\mathrm{Dis}}(x) S_{\mathrm{Dis}}(x)/S_{\mathrm{Dis}}(a)\,\mathrm{d}x$$

Chapter 5

Measures of Disease Occurrence Solutions

5.1 Follow-up of a small cohort

(a) Person-times for individuals 1, 4, 5 and 12 since entry until exit (onset of disease or censoring) were 2.5, 3, 4.5 and 1.5 years, respectively. Total person-time:

$$2.5 + 3.5 + 1.5 + 3.0 + 4.5 + 0.5 + 1.0 + 2.5 + 2.5 + 2.5 + 1.5 + 1.5 = 27 \text{ years}$$

Number of cases = 5 ⇒ incidence rate of $C = 5/27$ y = 18.5 per 100 y.

(b) Follow-up continued after onset of C for the 5 affected subjects; thus the total person-years will be

$$27 + 2 + 1.5 + 1 + 0.5 + 0.5 = 32.5 \text{ years}$$

Number of cases = 1 ⇒ mortality rate of C: 1/32.5 y = 3.1 per 100 y.

(c) Person-years same as in (b), but now the number of cases = 4.
Hence, the mortality rate is 4/32.5 y = 12.3 per 100 y.

(d) First, the 3-year cumulative rate

$$I \times \Delta = \frac{12.3}{100 \text{ y}} \times 3 \text{ y} = 0.123/\text{y} \times 3 \text{ y} = 0.369,$$

and from that the 3-year mortality proportion (which here is not very close to $I \times \Delta$) is

$$Q = 1 - \exp(-0.123/\text{y} \times 3 \text{ y}) = 1 - \exp(-0.369) = 0.309 = 31\%$$

(e) Person-years since diagnosis: 2 + 1.5 + 1 + 0.5 + 0.5 = 5.5 years, contributed by only those affected. Cases were 2, and the mortality rate = 2/5.5 y = 36.4 per 100 y.

(f) The 3-year cumulative rate is 0.364/y × 3 y = 1.092 (> 1!), and the 3-year mortality proportion

$$Q = 1 - \exp(1.092) = 0.664 = 66.4\%$$

(g) Prevalence of C on 30 September 2006 was $1/7 = 14$ % and on 31 December 2008 it was $3/5 = 60$ %.

The mortality from all other causes should be appropriately allowed for when computing incidence proportions and cause-specific mortality proportions with realistic interpretation.

5.2 Infant mortality

.

(a) Approximate person-years: $\frac{1}{2} \times (33200 + 32500) \times 1$ y $= 32850$ years;
incidence rate of infant deaths $= 269/32850$ y $= 8.19$ per 1000 y

(b) IMR $= 269/32800 = 8.20$ per 1000 liveborn

(c) IMR is not a genuine rate, nor any proportion either. This measure is a **ratio** in which the numerator is not completely included in the denominator. Some infants (< 1 y of age) dying during 1978 were, namely, born in 1977!

5.3 Incidence and mortality of leukaemia in children

It is handy to express the person-times in 100000 years; hence divide the person-times given in 1000 years by 100.

	1993-97	1999
person-years (in 100000s!)	$\frac{1}{2} \times (4.95 + 4.91) \times 5 = 24.65$	$\frac{1}{2} \times (4.85 + 4.81) \times 1 = 4.83$
(a): incidence rate (per 10^5 y)	$113/24.65 = 4.6$	$26/4.83 = 5.4$
(b): mortality rate (per 10^5 y)	$22/24.65 = 0.9$	$3/4.83 = 0.6$

(c) There are too few cases for any conclusions.

(d) It appears that most children survive, but this is not the correct way of calculating survival or fatality measures!

5.4 Rate ratio and rate difference in prevention trial

(a) Person-years in the two groups are obtained from cases and rates:

$$\frac{99}{11.6/10000 \text{ y}} = 85345 \text{ y}, \quad \frac{151}{17.8/10000 \text{ y}} = 84831 \text{ y}$$

(b) Incidence rate ratio $= 11.6/17.8 = 0.652$;
incidence rate difference $= 11.6 - 17.8 = -6.2$ per 10000 y,

(c) Preventive fraction: PF $= (17.8 - 11.6)/17.8 = 0.348 = 35\%$,

(d) The results are promising. However, among other things statistical imprecision in these figures has to be assessed.

5.5 Rate ratio, rate difference and excess fraction

	lung cancer	other lung diseases	cardiovascular diseases
rate difference (per 1000 y)	1.8	2.0	6.0
rate ratio	10.0	3.0	1.7
excess fraction (%)	90	67	40

5.6 Crude and standardized rates

(a) Crude rates and their ratio:
- males: 592/25.23 = 23.5 per 100000 years
- females: 732/26.48 = 27.6 per 100000 years
- ratio M/F: 23.5/27.6 = 0.85.

(b) Standardized rates with the male population as the standard:
- males: $(46 \times 0.9 + \cdots + 4 \times 196)/100 = 23.3$ per 100000 y
- females: $(46 \times 2.0 + \cdots + 4 \times 155)/100 = 19.8$ per 100000 y
- ratio M/F: 23.3/19.8 = 1.18

(c) Standardized rates with World standard population:
- males: $(62 \times 0.9 + \cdots + 2 \times 196)/100 = 15.4$ per 100000 y
- females: $(62 \times 2.0 + \cdots + 2 \times 155)/100 = 13.5$ per 100000 y
- ratio M/F: 15.4/13.5 = 1.14

(d) Cumulative rates up to 75 year (per 100):
- males: $(35 \times 0.9/10^5 + 20 \times 9.4/10^5 + 20 \times 67/10^5) = 0.0156$ or 1.56 per 100
- females: $(35 \times 2.0/10^5 + 20 \times 8.6/10^5 + 20 \times 55/10^5) = 0.0134$ or 1.34 per 100
- ratio M/F: 1.56/1.34 = 1.16

(e) Cumulative risks up to 75 years (per 100):
- males: $1 - \exp(-0.0156) = 0.0155$ or 1.55 %
- females: $1 - \exp(-0.0134) = 0.0133$ or 1.33 %
- ratio M/F = 1.56/1.34 = 1.16

The direction of the M/F contrast is very different for the crude rate than for any kind of standardized rates, showing that age is severely confounding the comparison between the genders. The ratio between males and females varies relatively little across the different standardized measures.

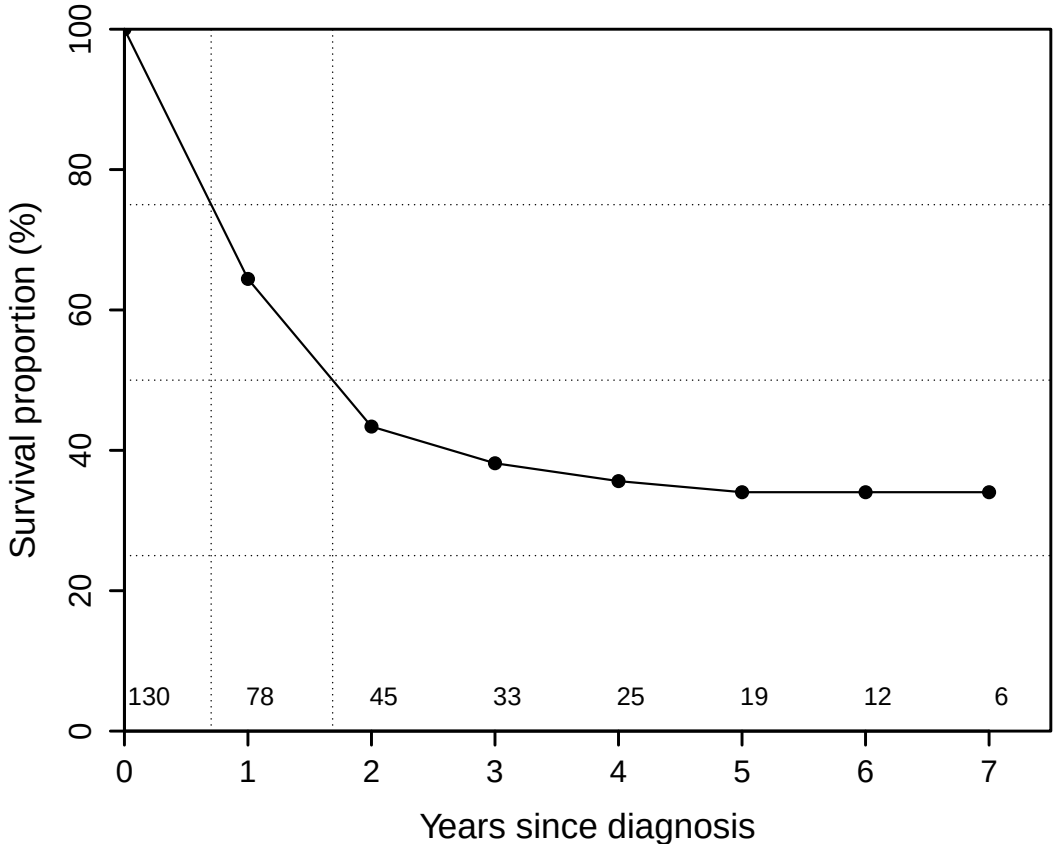
5.7 Survival from tongue cancer patients

(a) Completed life-table:

```
       N deaths losses N.eff prop.d prop.surv cum.surv
[1,] 130     45      7 126.5  0.356     0.644    0.644
[2,]  78     24      9  73.5  0.327     0.673    0.434
```

```
[3,]  45      5      7  41.5  0.120     0.880     0.382
[4,]  33      2      6  30.0  0.067     0.933     0.356
[5,]  25      1      5  22.5  0.044     0.956     0.340
[6,]  19      0      7  15.5  0.000     1.000     0.340
[7,]  12      0      6   9.0  0.000     1.000     0.340
```

(b) Survival curve displayed below, with numbers at risk in the beginning of each interval given above the x-axis. Lower quartile = 0.7; median = 1.69 y; upper quartile unestimable.



5.8 Lexis diagram and occupational cohort I

(a)-(b) Cases and person-years split by age in three periods, and by age in one birth cohort.

	Calendar period 1940-44		Calendar period 1945-49		Calendar period 1950-54		Birth cohort born 1902-11	
Age (y)	Cases	P-years	Cases	P-years	Cases	P-years	Cases	P-years
40-44	-	11	1	9.5	-	6	1	16.5
45-49	-	6	-	12.2	2	10.5	1	15.7
50-54	1	6	1	8.5	1	4.2	1	7.1

5.9 Lexis diagram and occupational cohort II

Expected number of cases

$$E = \frac{100}{10^5\text{y}} \times (11 + 9.5 + 6)\ \text{y} + \frac{200}{10^5\text{y}} \times (6 + 12.2 + 10.5)\ \text{y} + \frac{400}{10^5\text{y}} \times (6 + 8.5 + 4.2)\ \text{y} = 0.1587$$

Observed $O = 6$, standardised incidence ratio $6/0.1587 = 37.8$. – Quite a risky occupation!

Chapter 6

Measures of Disease Occurrence Solutions (R)

This chapter contains solutions to the same practicals as the previous chapter, but with greater detail in the R-solutions, including some extra suggestions you can explore.

```
> options( width=110 )
```

6.1 Basic measures in a cohort

1. sum of individual person-times (years) since entry until exit (onset of disease or censoring) — note we only count follow-up only to the end of 2008:

   ```
   > Y <- 2.5 + 3.5 + 1.5 + 3.0 + 4.5 + 0.5 +
   + 1.0 + 2.5 + 2.5 + 2.5 + 1.5 + 1.5
   > Y
   ```

   ```
   [1] 27
   ```

   Number of cases $= 5 \Rightarrow$ incidence rate of $C : 5/Y$, so if we compute it in cases per 100 years:

   ```
   > 5/Y * 100
   ```

   ```
   [1] 18.51852
   ```

   Similarly the 3-year incidence proportion (cumulative risk) is computed — note that the incidence rate $(5/Y)$ is measured in years^{-1} and the length of the interval (3) in years, so the argument to the exponential is dimensionless:

   ```
   > Q <- 1 - exp( -5/Y * 3 )
   > Q
   ```

   ```
   [1] 0.4262466
   ```

   We can express this in percent, suitably rounded:

   ```
   > round( Q*100, 1 )
   ```

```
[1] 42.6
```

2. Follow-up continued after onset of cancer for the 5 affected subjects; thus the total amount of person-years (until 31.12.2008) will be:

```
> YY <- Y + 2 + 1.5 + 1 + 0.5 + 0.5
```

Number of deaths from cancer was 1 ⇒ (recall we only do follow-up till 31.12.2008) so the mortality rate from cancer is:

```
> 1/YY

[1] 0.03076923
```

3-year incidence proportion can the be computed from this:

```
> Q <- 1 - exp( -1/YY * 3 )
> Q

[1] 0.08817546

> round( Q*100, 1 )

[1] 8.8
```

3. If we look at all cause-mortality, the person-years was same as above, but now the no. of cases is 4. Hence, the mortality rate is:

```
> 4/YY

[1] 0.1230769

> round( 4/YY * 100, 1 )

[1] 12.3
```

the latter is per 100 person-years.

4. The person-years among cancer patients (recall, still only till 31.12.2008):

```
> Yc <- 2 + 1.5 + 1 + 0.5 + 0.5
> Yc

[1] 5.5
```

There were 2 deaths, so the mortality rate is:

```
> 2 / Yc

[1] 0.3636364

> round( 2/Yc*100, 1 )
```

```
[1] 36.4
```

and hence the mortality proportion (i.e. the predicted fraction dead after three years, or 3-year cumulative risk):

```
> Q <- 1 - exp( -2/Yc * 3 )
> Q
```

```
[1] 0.664089
```

```
> round( Q*100, 1 )
```

```
[1] 66.4
```

5. Prevalence of cancer on 30 September 2006 was $1/7 = 14\%$ and on 31 December 2008 it was $3/5 = 60\%$:

```
> 1/7
```

```
[1] 0.1428571
```

```
> 3/5
```

```
[1] 0.6
```

or if we want to add nice labels:

```
> prv <- c(1,3)/c(7,5)
> names(prv) <- c("30sep2006","31dec2008")
> round( prv*100, 1 )
```

```
30sep2006 31dec2008
     14.3      60.0
```

6.1.1 Multistate set-up

The answer to the last questions about drawing boxes can be answered by setting up the the cohort as a `Lexis` object:

```
> library( Epi )
```

First we set up the follow-up of the cohort in a data frame with variables `doe`: date of entry, `dox`: date of exit, `ddx`: date of cancer diagnosis, `xst`: exit status:

```
> coh <- data.frame( doe=c("2004-01-01",
+                          "2004-01-01",
+                          "2004-01-01",
+                          "2004-07-01",
+                          "2004-07-01",
+                          "2005-01-01",
+                          "2005-07-01",
+                          "2006-01-01",
+                          "2006-01-01",
+                          "2006-07-01",
+                          "2007-01-01",
+                          "2007-01-01" ),
+                    dox=c("2008-07-01",
+                          "2009-07-01",
+                          "2005-07-01",
+                          "2007-07-01",
+                          "2009-07-01",
+                          "2006-07-01",
+                          "2006-07-01",
+                          "2008-07-01",
+                          "2009-07-01",
+                          "2009-07-01",
+                          "2009-07-01",
+                          "2008-07-01" ),
+                    ddx=c("2006-07-01",
+                          "2007-07-01",rep(NA,3),
+                          "2005-07-01",rep(NA,2),
+                          "2008-07-01",NA,
+                          "2008-07-01",NA),
+                    xst=factor(c(2,1,3,3,1,3,1,1,2,1,1,1),
+                               labels= c("Well","Dead-Ca","Dead-Oth")),
+                    id=1:12 )
> coh
```

```
          doe        dox        ddx      xst id
1  2004-01-01 2008-07-01 2006-07-01  Dead-Ca  1
2  2004-01-01 2009-07-01 2007-07-01     Well  2
3  2004-01-01 2005-07-01       <NA> Dead-Oth  3
4  2004-07-01 2007-07-01       <NA> Dead-Oth  4
5  2004-07-01 2009-07-01       <NA>     Well  5
6  2005-01-01 2006-07-01 2005-07-01 Dead-Oth  6
7  2005-07-01 2006-07-01       <NA>     Well  7
8  2006-01-01 2008-07-01       <NA>     Well  8
9  2006-01-01 2009-07-01 2008-07-01  Dead-Ca  9
10 2006-07-01 2009-07-01       <NA>     Well 10
11 2007-01-01 2009-07-01 2008-07-01     Well 11
12 2007-01-01 2008-07-01       <NA>     Well 12
```

Once we have the data frame, we can set it up as a `Lexis` object, which is designed to keep track of states and time-scales. In this case we only have one time scale, calendar time, which we call `per` (period), coded as fractions of years[1]:

```
> cL <- Lexis( entry = list(per=cal.yr(doe)),
+               exit = list(per=cal.yr(dox)),
+        exit.status = xst,
+                 id = id,
+               data = coh )
```

```
NOTE: entry.status has been set to "Well" for all.
```

[1] This works because the dates are character strings in ISO-format "yyyy-mm-dd", otherwise a format argument would have to be supplied to `cal.yr`.

Once the data is set up, we can summarize the number of transitions between states and the number of person-years spent in each state:

```
> summary( cL )
```

```
Transitions:
     To
From   Well Dead-Ca Dead-Oth  Records:  Events: Risk time:  Persons:
  Well    7       2        3        12        5      34.97        12
```

But we need to enter the cancer diagnoses, so we cut the follow-up at cancer diagnosis:

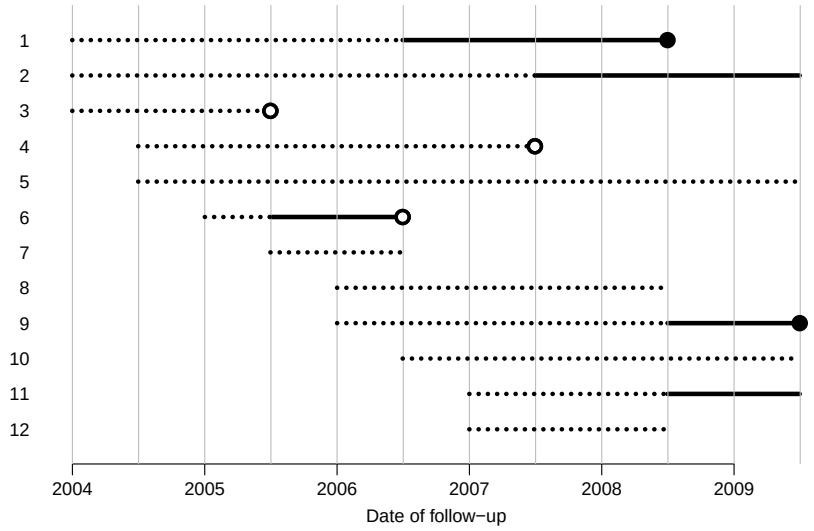
```
> cL <- cutLexis( cL,
+                 cut = cal.yr(cL$ddx),
+           new.state = "Cancer",
+   precursor.states = "Well" )
> summary( cL )
```

```
Transitions:
     To
From     Well Cancer Dead-Ca Dead-Oth  Records:  Events: Risk time:  Persons:
  Well      5      5       0        2        12        7      27.97        12
  Cancer    0      2       2        1         5        3       7.00         5
  Sum       5      7       2        3        17       10      34.97        12
```

In a `Lexis` object, each record represents a piece of follow-up for a person. The variable `lex.Cst` indicates the `state` where the follow-up takes place. So the records with `lex.Cst` equal to `"Well"` corresponds to the broken lines in the figure, and the records with `lex.Cst` equal to `"Cancer"` corresponds to the full lines in the figure, the follow-up after cencer diagnosis:

```
> cL[order(cL$lex.id,cL$per),]
```

```
        per   lex.dur lex.Cst  lex.Xst lex.id        doe        dox        ddx      xst id
1  2003.999 2.4969199    Well   Cancer      1 2004-01-01 2008-07-01 2006-07-01 Dead-Ca  1
13 2006.496 2.0013689  Cancer Dead-Ca      1 2004-01-01 2008-07-01 2006-07-01 Dead-Ca  1
2  2003.999 3.4962355    Well   Cancer      2 2004-01-01 2009-07-01 2007-07-01     Well  2
14 2007.495 2.0013689  Cancer   Cancer      2 2004-01-01 2009-07-01 2007-07-01     Well  2
3  2003.999 1.4976044    Well Dead-Oth      3 2004-01-01 2005-07-01       <NA> Dead-Oth  3
4  2004.497 2.9979466    Well Dead-Oth      4 2004-07-01 2007-07-01       <NA> Dead-Oth  4
5  2004.497 4.9993155    Well     Well      5 2004-07-01 2009-07-01       <NA>     Well  5
6  2005.001 0.4955510    Well   Cancer      6 2005-01-01 2006-07-01 2005-07-01 Dead-Oth  6
18 2005.496 0.9993155  Cancer Dead-Oth      6 2005-01-01 2006-07-01 2005-07-01 Dead-Oth  6
7  2005.496 0.9993155    Well     Well      7 2005-07-01 2006-07-01       <NA>     Well  7
8  2006.000 2.4969199    Well     Well      8 2006-01-01 2008-07-01       <NA>     Well  8
9  2006.000 2.4969199    Well   Cancer      9 2006-01-01 2009-07-01 2008-07-01 Dead-Ca  9
21 2008.497 0.9993155  Cancer Dead-Ca      9 2006-01-01 2009-07-01 2008-07-01 Dead-Ca  9
10 2006.496 3.0006845    Well     Well     10 2006-07-01 2009-07-01       <NA>     Well 10
11 2006.999 1.4976044    Well   Cancer     11 2007-01-01 2009-07-01 2008-07-01     Well 11
23 2008.497 0.9993155  Cancer   Cancer     11 2007-01-01 2009-07-01 2008-07-01     Well 11
12 2006.999 1.4976044    Well     Well     12 2007-01-01 2008-07-01       <NA>     Well 12
```



With this set-up we can draw a 1-dimensional Lexis-diagram (actually the figure above), and add a few bells and whistles to produce the figure used in the exercise text

```
> par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
> plot( cL, ylim=c(12.5,0.5), ylab="", xlab="Date of follow-up",
+             bty="n", las=1, yaxt="n", xlim=c(2004,2009.5),
+             lty=1, col=c("transparent","black")[as.integer(cL$lex.Cst)], lwd=4 )
> abline( v=2003+seq(0,7,0.5), col="gray" )
> lines( cL, lty="11", col=c("black","transparent")[as.integer(cL$lex.Cst)], lwd=4 )
> axis( side=2, labels=1:12, at=1:12, las=1, lty=0 )
> # This is just to get the points of death to look nice
> points( cL, col="white", pch=c(NA,NA,16,16)[as.integer(cL$lex.Xst)], cex=1.6 )
> points( cL, col="black", pch=c(NA,NA,16,1 )[as.integer(cL$lex.Xst)], cex=1.6, lwd=3 )
> points( cL, col="black", pch=c(NA,NA,1 ,1 )[as.integer(cL$lex.Xst)], cex=1.6, lwd=3 )
```

We can also show the states and transitions and person-years in a plot:

```
> boxes( cL, boxpos=T )
```

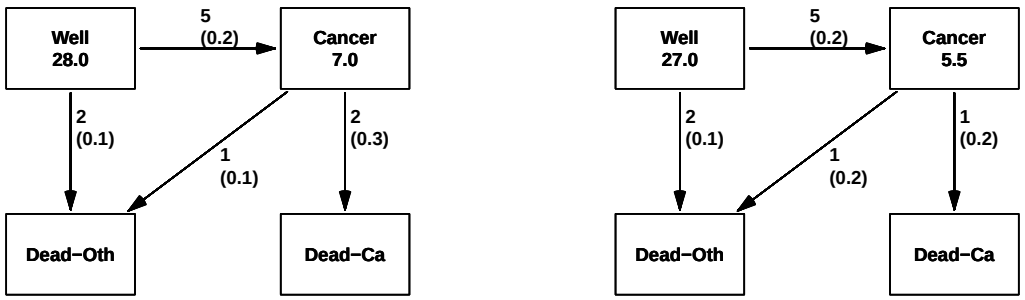


Figure 6.1: *Follow-up of a small cohort across 4 states. The left panel is for the entire follow-up, the right for follow-up censored at 31.12.2008.*

However, this is for the entire follow-up, and we want the follow-up to end at 31.12.2008 (or 1.1.2009), so we split the follow-up of the cohort in order to be able to restrict to the follow-up before that:

```
> cS <- splitLexis( cL, breaks=2009 )
```

Now we can show how the follow-up for each person is split in several intervals; each line in the data frame corresponds to a single follow-up interval, so persons may contribute several lines.

Variables are: `per` is the start of each follow-up interval, `lex.dur` is the length of the interval, `lex.Cst` is the Current `st`ate *i.e.* the state in which the follow-up takes place and `lex.Xst` is the eXit `st`ate, *i.e.* the state to whic the person exits at the end of the interval.

```
> cS[order(cS$lex.id),1:8]
```

```
   lex.id      per   lex.dur lex.Cst  lex.Xst        doe        dox        ddx
1       1 2003.999 2.4969199    Well   Cancer 2004-01-01 2008-07-01 2006-07-01
2       1 2006.496 2.0013689  Cancer  Dead-Ca 2004-01-01 2008-07-01 2006-07-01
3       2 2003.999 3.4962355    Well   Cancer 2004-01-01 2009-07-01 2007-07-01
4       2 2007.495 1.5051335  Cancer   Cancer 2004-01-01 2009-07-01 2007-07-01
5       2 2009.000 0.4962355  Cancer   Cancer 2004-01-01 2009-07-01 2007-07-01
6       3 2003.999 1.4976044    Well Dead-Oth 2004-01-01 2005-07-01       <NA>
7       4 2004.497 2.9979466    Well Dead-Oth 2004-07-01 2007-07-01       <NA>
8       5 2004.497 4.5030801    Well     Well 2004-07-01 2009-07-01       <NA>
9       5 2009.000 0.4962355    Well     Well 2004-07-01 2009-07-01       <NA>
10      6 2005.001 0.4955510    Well   Cancer 2005-01-01 2006-07-01 2005-07-01
11      6 2005.496 0.9993155  Cancer Dead-Oth 2005-01-01 2006-07-01 2005-07-01
12      7 2005.496 0.9993155    Well     Well 2005-07-01 2006-07-01       <NA>
13      8 2006.000 2.4969199    Well     Well 2006-01-01 2008-07-01       <NA>
14      9 2006.000 2.4969199    Well   Cancer 2006-01-01 2009-07-01 2008-07-01
15      9 2008.497 0.5030801  Cancer   Cancer 2006-01-01 2009-07-01 2008-07-01
16      9 2009.000 0.4962355  Cancer  Dead-Ca 2006-01-01 2009-07-01 2008-07-01
17     10 2006.496 2.5044490    Well     Well 2006-07-01 2009-07-01       <NA>
18     10 2009.000 0.4962355    Well     Well 2006-07-01 2009-07-01       <NA>
19     11 2006.999 1.4976044    Well   Cancer 2007-01-01 2009-07-01 2008-07-01
20     11 2008.497 0.5030801  Cancer   Cancer 2007-01-01 2009-07-01 2008-07-01
21     11 2009.000 0.4962355  Cancer   Cancer 2007-01-01 2009-07-01 2008-07-01
22     12 2006.999 1.4976044    Well     Well 2007-01-01 2008-07-01       <NA>
```

```
> boxes( subset(cS,per<2009), boxpos=TRUE )
```

The results of the two different calculations are shown in figure 6.1.

6.2 Infant mortality

1. Approximate person-years: $\frac{1}{2}(33200 + 32500) \times 1$ y $= 32850$ years;
   rate $= 269/32850$ y $= 8.19$ per 1000 y

   ```
   > Y <- (33200 + 32500)/2
   > D <- 269
   > cbind( D, Y, D/Y*100000 )
   ```

   ```
         D     Y
   [1,] 269 32850 818.8737
   ```

2. IMR $= 269/32{,}800 = 8.20$ per 1000 liveborn:

   ```
   > 269/32.800
   ```

   ```
   [1] 8.20122
   ```

3. IMR is not a rate — the denominator is a bad approximation to the person-years in age group 0. In this measure the numerator is not completely included in the denominator, so it is not a proportion either. Some infants (< 1 y of age) dying during 1978 are, namely, born in 1977!

6.3 Incidence and mortality — acute leukaemia

The popualtions size at the *end* of the year:

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Population	493	495	496	497	496	495	491	485	481	478

	1993-97	1999
person-years (in 100000s)	$\frac{1}{2} \times (4.95 + 4.91) \times 5 = 24.65$	$\frac{1}{2} \times (4.85 + 4.81) \times 1 = 4.83$
(a): incidence rate (per 10^5 y)	$113/24.65 = 4.6$	$26/4.83 = 5.4$
(b): mortality rate (per 10^5 y)	$22/24.65 = 0.9$	$3/4.83 = 0.6$

1. The incidence rates requires the person-years. The simple approach is to take the average of the population at each end of the period, i.e. for the first 31.12.1992 and 31.12.1997 and for the second 31.12.1998 and 31.12.1999:

```
> Y <- c( 495+491, 485+481 )/2 * c(5,1)
> names(Y) <- c("1993-97","1999")
> Y
```

```
1993-97    1999
   2465     483
```

   Alternatively, the person-years in the first period could be calculated by computing the person-years for each of the 5 years in the period separately; giving 1/2 for the 1992 and 1997 and 1/1 for the intermediate ones:

```
> Y[1] <- 495/2+496+497+496+495+491/2
> Y
```

```
1993-97    1999
   2477     483
```

   With this, the incidence rates (per 100,000) are (since `Y` is in 1000s):

```
> ir <- c(113,26)/Y*10^2
> round( ir, 2 )
```

```
1993-97    1999
   4.56    5.38
```

2. The mortality rates are computed similarly:

```
> mr <- c(22,3)/Y*10^2
> round( mr, 2 )
```

```
1993-97    1999
   0.89    0.62
```

3. At first glance it looks as if incidence has gome up, but mortality has gone down, but the evidence is way too thin to draw any comclusions.

4. Nothing can be concluded about the fatality; the number of cases in the year 1999 is too small (only 3).

6.4 ATCB-trial — prostate cancer

The Alpha Tocopherol Beta Caroten (ATBC) Prevention Trial (*N Engl J Med 1994*; **330**: 1029-35) addressed among other things the possible benefits of daily intake of vitamin E supplements in reducing the incidence of cancer among male smokers. The study population of 29,133 regularly smoking 50-69 years old Finnish men were randomized into two groups: active treatment (vitamin E supplementation), and placebo (no supplementation). The following results were obtained for cancer of the prostate after an average follow-up time of 6 years:

treatment group	number of cases	incidence rate (per 10,000 years)
vitamin E supplementation	99	11.6
no supplementation	151	17.8

1. Since the rate λ is computed from no. cases, D, and person-years Y as $\lambda = D/Y$ and hence $Y = D/\lambda$, so we compute the person-years in the two groups:

$$\frac{99}{11.6/10000\ \text{y}} = 85345\ \text{y}, \quad \frac{151}{17.8/10000\ \text{y}} = 84831\ \text{y}$$

5-year incidence proportions (cumulative risk):

$$1 - \exp\left(-\frac{11.6}{10^4\text{y}} \times 5\right) = 0.0058, \quad 1 - \exp\left(-\frac{17.8}{10^4\text{y}} \times 5\right) = 0.0089$$

or 5.8 and 8.9 per 1000, respectively.

In R this would go:

```
> rate <- c(11.6,17.8)
> D <- c(99,151)
> names(rate) <- names(D) <- c("VitE","Plc")
> D / (rate/10000)
```

```
    VitE      Plc
85344.83 84831.46
```

```
> cr5 <- 1 - exp( -rate/10000 * 5 )
> round( cr5*100, 2 )
```

```
VitE  Plc
0.58 0.89
```

Thus, the 5-year cumulative risk (incidence proportion) of prostate cancer is 0.6% in the vitamin E group, and 0.9% in the placebo group.

2. The comparative measures:

3. "Relative risk":

   - incidence rate ratio = 11.6/17.8 = 0.652,
   - inc. proportion ratio = 5.8/8.9 = 0.653,

```
> rate[1]/rate[2]

     VitE
0.6516854

> cr5[1]/cr5[2]

    VitE
0.652695
```

4. "Excess risk":

   - rate difference = 11.6 − 17.8 = −6.2 per 10000 y,
   - inc. prop. diff. = 5.8 − 8.9 = −3.1 per 1000.

```
> rate[1]-rate[2]

VitE
-6.2

> round( (cr5[1]-cr5[2])*100, 2 )

 VitE
-0.31
```

5. Preventive fraction:

   - from rates: (17.8 − 11.6)/17.8 = 0.348 = 35%,
   - from proportions: (8.9 − 5.8)/8.9 = 0.347 = 35%.

```
> (rate["Plc"]-rate["VitE"])/rate["Plc"]

      Plc
0.3483146

> (cr5["Plc"]-cr5["VitE"])/cr5["Plc"]

     Plc
0.347305
```

6. The results are promising. However, among other things statistical imprecision in these figures has to be assessed.

6.5 Comparative measures — smokers vs. non-smokers

1. The comparative measures as computed from the mortality rates are:

	lung cancer	other lung diseases	cardiovascular diseases
Mortality rates			
smokers	2.0	3.0	15.0
non-smokers	0.2	1.0	9.0
"excess risk", rate difference (per 1000 y)	1.8	2.0	6.0
"relative risk", rate ratio	10.0	3.0	1.7
excess fraction (%)	90	67	40

These mesaures can be computed from the original table. First we enter the mortality rates in two vectors; one for smokers and one for non-smokers, and annotate them with the causes

```
> sm <- c(2.0,3.0,15.0)
> ns <- c(0.2,1.0,9.0)
> names(sm) <- names(ns) <- c("Lung Ca","Oth lung","CVD")
> rbind( sm, ns )
```

```
   Lung Ca Oth lung CVD
sm     2.0        3  15
ns     0.2        1   9
```

Then we compute the three different measures:

```
> ER <- sm-ns
> RR <- sm/ns
> EF <- (RR-1)/RR*100
> round( rbind( ER, RR, EF ), 1 )
```

```
   Lung Ca Oth lung  CVD
ER     1.8      2.0  6.0
RR    10.0      3.0  1.7
EF    90.0     66.7 40.0
```

2. The strongest biological effect is seen for lung diseases, and particular lung cancer; as is apparent from the large values of RR and EF, but as seen from the ER, the excess mortality rate, the population impact is by far the largest for CVD mortality.

6.6 Standardization: Colon cancer

Age specific data on the incidence of colon cancer in male and female populations of Finland during 1999 are given in the following table

	Males				Females				
Age group	Cases	Mid-popul. (1000s)	% of all	Rate ($/10^5$y)	Cases	Mid-popul. (1000s)	% of all	Rate ($/10^5$y)	Rate ratio M/F
0–34	10	1157	46.0	**0.9**	22	1109	41.9	**2.0**	0.44
35–54	76	809	32.0	**9.4**	68	786	29.7	**8.6**	1.09
55–74	305	455	18.0	**67**	288	524	19.8	**55**	1.22
75+	201	102	4.0	**196**	354	229	8.6	**155**	1.27
All	592	2523	100		732	2648	100		

To be able to manipulate these numbers we put them in a matrix, tun this into a dataframe and give the columns sensible names. In practice this is done by copy-paste from the pdf-document and then in the R-script-editor add the "`c(`" and the commas:

```
> M <- matrix(
+ c(10,1157,46.0,0.9,22,1109,41.9,2.0,0.44
+ ,76,809,32.0,9.4,68,786,29.7,8.6,1.09
+ ,305,455,18.0,67,288,524,19.8,55,1.22
+ ,201,102,4.0,196,354,229,8.6,155,1.27), nrow=4, byrow=T )
> M <- data.frame(M)
> names(M) <- c("mca","mpy","mp","mr",
+               "fca","fpy","fp","fr","rr")
> M
```

```
  mca  mpy mp    mr fca  fpy   fp    fr   rr
1  10 1157 46   0.9  22 1109 41.9   2.0 0.44
2  76  809 32   9.4  68  786 29.7   8.6 1.09
3 305  455 18  67.0 288  524 19.8  55.0 1.22
4 201  102  4 196.0 354  229  8.6 155.0 1.27
```

Once we have the numbers in a dataframe we can do all the calculations using the `with( M, ...)`.

1. Crude incidence rates and M/F RR based on these (rates per 100,000 PY):

```
> rates <-
+ with( M, c( sum(mca)/sum(mpy)*100,
+               sum(fca)/sum(fpy)*100 ) )
> rates[3] <- rates[1]/rates[2]
> names(rates) <- c("M rate","F rate","M/F RR")
> round( rates, 2 )
```

```
M rate F rate M/F RR
 23.46  27.64   0.85
```

2. The age-standardized rates using the male population as standard, is simply the weighted average of the age-specific rates:

```
> wm <- with( M, mpy/sum(mpy) )
> rates <-
+ with( M, c( sum(mca/mpy*wm)*100,
+               sum(fca/fpy*wm)*100 ) )
> rates[3] <- rates[1]/rates[2]
> names(rates) <- c("M rate","F rate","M/F RR")
> round( rates, 2 )
```

```
M rate F rate M/F RR
 23.46  19.85   1.18
```

3. Using the world standardized population (WSP) is just using the same code but defining the weights differently. We can snatch the WSP from the slides:

```
> WSP <- c(96,24,100,90,90,80,80,60,60,60,60,50,40,40,30,20,10,5,3,2)
> WSP
```

```
 [1]  96  24 100  90  90  80  80  60  60  60  60  50  40  40  30  20  10   5   3
[20]   2
```

But our age-classes are wider so the weight we need are the sum of the first 8, the next 4, the next 4 and the last 3. Note that there is no "`[1]`" in the first assignment, because the `wt` is created as a vector of length 1 there, and then later expanded:

```
> wt <- sum(WSP[1:8])
> wt[2] <- sum(WSP[9:12])
> wt[3] <- sum(WSP[13:16])
> wt[4] <- sum(WSP[17:19])
> wt <- wt/sum(wt)
> wt
```

```
[1] 0.62124248 0.23046092 0.13026052 0.01803607
```

```
> rates <-
+ with( M, c( sum(mca/mpy*wt)*100,
+              sum(fca/fpy*wt)*100 ) )
> rates[3] <- rates[1]/rates[2]
> names(rates) <- c("M rate","F rate","M/F RR")
> round( rates, 2 )
```

```
M rate F rate M/F RR
 14.99  13.17   1.14
```

4. The cumulative rates to age 75 are just using weights equal to the length of the intervals, only using the first 3 intervals. But now we also need to use the proper rates, i.e. in units of cases per 1 year:

```
> wy <- c(35,20,20)
> rates <-
+ with( M[1:3,], c( sum(mca/mpy*wy)/1000,
+                    sum(fca/fpy*wy)/1000 ) )
> rates[3] <- rates[1]/rates[2]
> names(rates) <- c("M cum.rate","F cum.rate","M/F RR")
> round( rates, 4 )
```

```
M cum.rate F cum.rate     M/F RR
    0.0156     0.0134     1.1618
```

5. Using cumulative risks amounts to converting to risk before taking the ratio, otherwise the code is the same:

```
> rates <- 1 - exp( -rates )
> rates[3] <- rates[1]/rates[2]
> names(rates) <- c("M cum.risk","F cum.risk","M/F RR")
> round( rates, 4 )
```

```
M cum.risk F cum.risk     M/F RR
    0.0155     0.0133     1.1606
```

It is seen that the comparison based on the crude rates can be quite misleading. But you may equally well say that the comparison based on the standardized rates is equally misleading because it bypasses the important information that the RR varies substantially by age.

6.7 Survival: cancer of the tongue

1. We begin by entering the data in three vectors:

```
> N <- c(130,78,45,33,25,19,12)
> D <- c(45,24,5,2,1,0,0)
> L <- c(7,9,7,6,5,7,6)
```

With these we can now do all the calculations and put it all in a dataframe. Note the use of the function `cumprod` which simply takes the cumulative product of a vector:

```
> res <- data.frame( N=N, D=D, L=L,
+                    eff.den =               N-L/2,
+                   pr.death =             D/(N-L/2),
+                    pr.surv =          1-D/(N-L/2),
+                   cum.surv = cumprod( 1-D/(N-L/2) ) )
> round( res, 3 )
```

```
   N  D L eff.den pr.death pr.surv cum.surv
1 130 45 7   126.5    0.356   0.644    0.644
2  78 24 9    73.5    0.327   0.673    0.434
3  45  5 7    41.5    0.120   0.880    0.382
4  33  2 6    30.0    0.067   0.933    0.356
5  25  1 5    22.5    0.044   0.956    0.340
6  19  0 7    15.5    0.000   1.000    0.340
7  12  0 6     9.0    0.000   1.000    0.340
```

2. The survival curve is simply the last column of the data frame; the x-values being the *end* of the intervals, in this case 1,...,7, but we add the point (0,1) as the start of the curve. Moreover we also add horizontal lines to be able to read off the quartiles of the survival:

```
> plot( 0:7, c(1,res$cum.surv), pch=16, type="b", ylim=0:1,
+       ylab="Survival", xlab="Time since diagnosis" )
> abline( h=c(1:3/4) )
```

From figure 6.2 we see that the lower quartile is 0.7 years, median 1.69 years but that the upper quartile is unestimable from these data.

6.8 Lexis diagram

Cases and person-years split by age in three periods, and by age in one birth cohort.

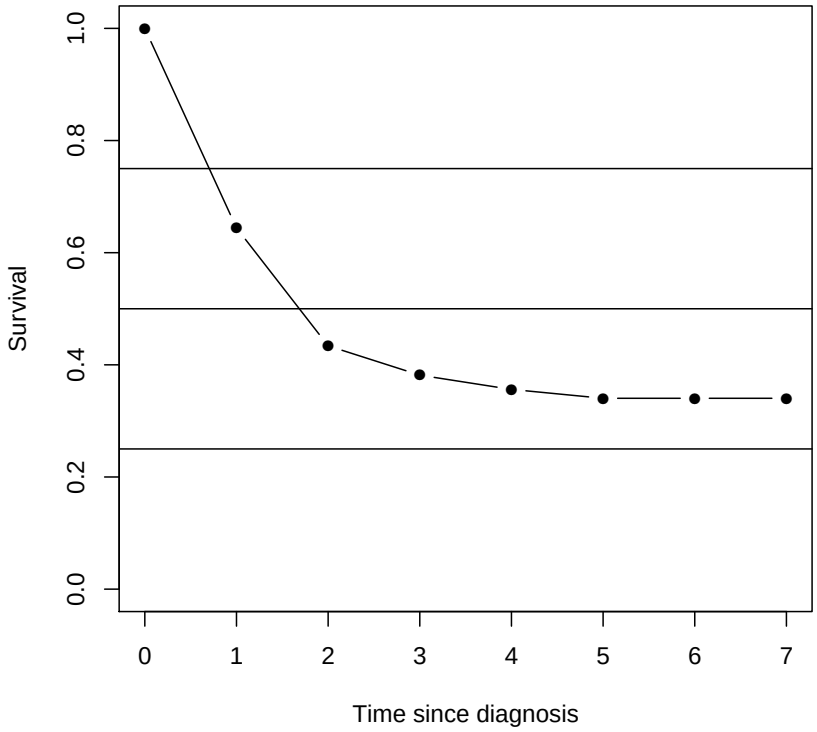


Figure 6.2: *The estimated survival function — lifteable estimator.*

	period 1940-44		period 1945-49		period 1950-54		1902-11 cohort	
Age (y)	Cases	P-years	Cases	P-years	Cases	P-years	Cases	P-years
40-44	-	11	1	9.5	-	6	1	16.5
45-49	-	6	-	12.2	2	10.5	1	15.7
50-54	1	6	1	8.5	1	4.2	1	7.1

1. You can load the dataset from the `Epi` package by:

```
> library( Epi )
> data( occup )
> occup
```

```
    AoE    DoE    DoX Xst
1  51.0 1941.0 1944.0   D
2  48.0 1940.0 1947.0   X
3  47.0 1942.0 1948.0   D
4  51.0 1948.0 1951.4   D
5  48.5 1946.9 1951.8   W
6  41.0 1940.0 1947.2   W
7  44.0 1944.0 1949.5   W
8  40.0 1941.0 1950.5   D
9  40.0 1943.0 1947.5   D
10 47.0 1951.0 1958.1   D
11 42.0 1947.0 1954.0   D
12 40.0 1947.0 1960.0   X
13 41.0 1951.0 1958.7   W
```

In order to compute the cases and person-years we set up a `Lexis` object:

```
> oL <- Lexis( entry = list( age=AoE, per=DoE ),
+               exit = list(          per=DoX ),
+       entry.status = factor( rep("W",nrow(occup)) ),
+        exit.status = factor( Xst ),
+               data = occup )
```

```
Incompatible factor levels in entry.status and exit.status:
 both lex.Cst and lex.Xst now have levels:
 W D X
```

```
> summary( oL )
```

```
Transitions:
     To
From W D X  Records:  Events: Risk time:  Persons:
   W 4 7 2        13        9       85.8        13

Rates:
     To
From W    D    X Total
   W 0 0.08 0.02   0.1
```

Exit status `X` and `W` are synonymous. If we want to classify the follow-up (påerson-years and events) by age and calendar time we must first subdivide by the two timescales, this is done by `splitLexis`:

```
> oL <- splitLexis( oL, time="age", breaks=seq(0,100,5) )
> oL <- splitLexis( oL, time="per", breaks=seq(0,100,5)+1900 )
> oL[order(oL$lex.id,oL$age),]
```

```
   lex.id  age    per lex.dur lex.Cst lex.Xst  AoE    DoE    DoX Xst
1       1 51.0 1941.0     3.0       W       D 51.0 1941.0 1944.0   D
2       2 48.0 1940.0     2.0       W       W 48.0 1940.0 1947.0   X
3       2 50.0 1942.0     3.0       W       W 48.0 1940.0 1947.0   X
4       2 53.0 1945.0     2.0       W       X 48.0 1940.0 1947.0   X
5       3 47.0 1942.0     3.0       W       W 47.0 1942.0 1948.0   D
6       3 50.0 1945.0     3.0       W       D 47.0 1942.0 1948.0   D
7       4 51.0 1948.0     2.0       W       W 51.0 1948.0 1951.4   D
8       4 53.0 1950.0     1.4       W       D 51.0 1948.0 1951.4   D
9       5 48.5 1946.9     1.5       W       W 48.5 1946.9 1951.8   W
10      5 50.0 1948.4     1.6       W       W 48.5 1946.9 1951.8   W
11      5 51.6 1950.0     1.8       W       W 48.5 1946.9 1951.8   W
12      6 41.0 1940.0     4.0       W       W 41.0 1940.0 1947.2   W
13      6 45.0 1944.0     1.0       W       W 41.0 1940.0 1947.2   W
14      6 46.0 1945.0     2.2       W       W 41.0 1940.0 1947.2   W
15      7 44.0 1944.0     1.0       W       W 44.0 1944.0 1949.5   W
16      7 45.0 1945.0     4.5       W       W 44.0 1944.0 1949.5   W
17      8 40.0 1941.0     4.0       W       W 40.0 1941.0 1950.5   D
18      8 44.0 1945.0     1.0       W       W 40.0 1941.0 1950.5   D
19      8 45.0 1946.0     4.0       W       W 40.0 1941.0 1950.5   D
20      8 49.0 1950.0     0.5       W       D 40.0 1941.0 1950.5   D
21      9 40.0 1943.0     2.0       W       W 40.0 1943.0 1947.5   D
22      9 42.0 1945.0     2.5       W       D 40.0 1943.0 1947.5   D
23     10 47.0 1951.0     3.0       W       W 47.0 1951.0 1958.1   D
24     10 50.0 1954.0     1.0       W       W 47.0 1951.0 1958.1   D
25     10 51.0 1955.0     3.1       W       D 47.0 1951.0 1958.1   D
26     11 42.0 1947.0     3.0       W       W 42.0 1947.0 1954.0   D
27     11 45.0 1950.0     4.0       W       D 42.0 1947.0 1954.0   D
28     12 40.0 1947.0     3.0       W       W 40.0 1947.0 1960.0   X
29     12 43.0 1950.0     2.0       W       W 40.0 1947.0 1960.0   X
```

```
30     12 45.0 1952.0     3.0       W       W 40.0 1947.0 1960.0   X
31     12 48.0 1955.0     2.0       W       W 40.0 1947.0 1960.0   X
32     12 50.0 1957.0     3.0       W       X 40.0 1947.0 1960.0   X
33     13 41.0 1951.0     4.0       W       W 41.0 1951.0 1958.7   W
34     13 45.0 1955.0     3.7       W       W 41.0 1951.0 1958.7   W
```

Having split the follow-up we can make a tabulation of the follow-up using the utility function `timeBand`:

```
> table( timeBand(oL,"age","left"), timeBand(oL,"per","left"))
```

```
    1940 1945 1950 1955
40     4    4    2    0
45     3    4    4    2
50     2    4    3    2
```

However we do not want the number of observations (lines) in the dataset, we want the number of person-yeras (`lex.dur`) and the number of deaths (`lex.Xst=="D"`), so we set up a matrix with these as columns, and define the two clssification variables:

```
> FU <- with( oL, cbind(lex.Xst=="D",lex.dur) )
> colnames(FU) <- c("D","Y")
> Age    <- timeBand(oL,"age","left")
> Period <- timeBand(oL,"per","left")
```

This enables us to use `xtabs` to simultaneously tabulate person-years and deaths

```
> FUtab <- xtabs( FU ~ Age + Period )
> ftable(FUtab,col.vars=2:3)
```

```
    Period 1940      1945      1950      1955
              D    Y    D    Y    D    Y    D    Y
Age
40          0.0 11.0  1.0  9.5  0.0  6.0  0.0  0.0
45          0.0  6.0  0.0 12.2  2.0 10.5  0.0  5.7
50          1.0  6.0  1.0  8.6  1.0  4.2  1.0  6.1
```

2. If we want the tabulation by age for the birth cohort 1902–11, we simply restrict the dataset to his group, i.e. the persons where `per` − `age` is betwwen 1029 and 1912:

```
> BC <- subset(oL,per-age>1902 & per-age<1912)
> FU <- with( BC, cbind(lex.Xst=="D",lex.dur) )
> colnames(FU) <- c("D","Y")
> Age   <- timeBand(BC,"age","left")
> FUctab <- xtabs( FU ~ Age )
> FUctab
```

```
Age     D    Y
  40  1.0 16.5
  45  1.0 15.7
  50  1.0  7.1
```

3. The cumulative rate for the cohort:

$$5 \times \left(\frac{1}{16.5} + \frac{1}{15.7} + \frac{1}{7.1} \right) = 1.32, \quad 1 - \exp(-1.32) = 0.73$$

or in terms of the just computed:

```
> sum(FUctab[,1]/FUctab[,2]*5)
```

```
[1] 1.325727
```

4. Occupational cohort. Expected number of cases

$$E = \frac{100}{10^5\text{y}}\times(11+9.5+6+0)\ \text{y}+\frac{200}{10^5\text{y}}\times(6+12.2+10.5+5.7)\ \text{y}+\frac{400}{10^5\text{y}}\times(6+8.5+4.2+6.1)\ \text{y} = 0.1949$$

Observed $O = 7$, standardised incidence ratio $7/0.1949 = 35.9$. Quite a risky occupation!

Note that the point of subdividing the follow-up by age and calendar time is to make it possible to apply population rates to the follow-up — the population rates vary by age and calendar time. So what is done is to match the populatin rates to the follow-up dataset:

```
> p.rates <- data.frame( rate=c(100,200,400), Age=c(40,45,50) )
> oL$Age <- timeBand(oL,"age","left")
> oL <- merge(oL,p.rates)
> oL
```

```
   Age lex.id  age    per lex.dur lex.Cst lex.Xst  AoE    DoE    DoX Xst rate
1   40      8 40.0 1941.0     4.0       W       W 40.0 1941.0 1950.5   D  100
2   40      9 40.0 1943.0     2.0       W       W 40.0 1943.0 1947.5   D  100
3   40      8 44.0 1945.0     1.0       W       W 40.0 1941.0 1950.5   D  100
4   40      6 41.0 1940.0     4.0       W       W 41.0 1940.0 1947.2   W  100
5   40     12 43.0 1950.0     2.0       W       W 40.0 1947.0 1960.0   X  100
6   40      9 42.0 1945.0     2.5       W       D 40.0 1943.0 1947.5   D  100
7   40      7 44.0 1944.0     1.0       W       W 44.0 1944.0 1949.5   W  100
8   40     12 40.0 1947.0     3.0       W       W 40.0 1947.0 1960.0   X  100
9   40     13 41.0 1951.0     4.0       W       W 41.0 1951.0 1958.7   W  100
10  40     11 42.0 1947.0     3.0       W       W 42.0 1947.0 1954.0   D  100
11  45      3 47.0 1942.0     3.0       W       W 47.0 1942.0 1948.0   D  200
12  45      2 48.0 1940.0     2.0       W       W 48.0 1940.0 1947.0   X  200
13  45      5 48.5 1946.9     1.5       W       W 48.5 1946.9 1951.8   W  200
14  45      6 46.0 1945.0     2.2       W       W 41.0 1940.0 1947.2   W  200
15  45      8 45.0 1946.0     4.0       W       W 40.0 1941.0 1950.5   D  200
16  45      6 45.0 1944.0     1.0       W       W 41.0 1940.0 1947.2   W  200
17  45     12 45.0 1952.0     3.0       W       W 40.0 1947.0 1960.0   X  200
18  45     10 47.0 1951.0     3.0       W       W 47.0 1951.0 1958.1   D  200
19  45      7 45.0 1945.0     4.5       W       W 44.0 1944.0 1949.5   W  200
20  45     13 45.0 1955.0     3.7       W       W 41.0 1951.0 1958.7   W  200
21  45     11 45.0 1950.0     4.0       W       D 42.0 1947.0 1954.0   D  200
22  45      8 49.0 1950.0     0.5       W       D 40.0 1941.0 1950.5   D  200
23  45     12 48.0 1955.0     2.0       W       W 40.0 1947.0 1960.0   X  200
24  50      1 51.0 1941.0     3.0       W       D 51.0 1941.0 1944.0   D  400
25  50      3 50.0 1945.0     3.0       W       D 47.0 1942.0 1948.0   D  400
26  50      2 50.0 1942.0     3.0       W       W 48.0 1940.0 1947.0   X  400
27  50      2 53.0 1945.0     2.0       W       X 48.0 1940.0 1947.0   X  400
28  50     10 51.0 1955.0     3.1       W       D 47.0 1951.0 1958.1   D  400
29  50      5 50.0 1948.4     1.6       W       W 48.5 1946.9 1951.8   W  400
30  50      4 51.0 1948.0     2.0       W       W 51.0 1948.0 1951.4   D  400
31  50      4 53.0 1950.0     1.4       W       D 51.0 1948.0 1951.4   D  400
32  50      5 51.6 1950.0     1.8       W       W 48.5 1946.9 1951.8   W  400
33  50     10 50.0 1954.0     1.0       W       W 47.0 1951.0 1958.1   D  400
34  50     12 50.0 1957.0     3.0       W       X 40.0 1947.0 1960.0   X  400
```

With this we can now compute the observed and expected cases:

```
> O <- with( oL, sum( lex.Xst=="D" ) )
> E <- with( oL, sum( lex.dur*rate/10^5 ) )
> c( O, E, O/E )
```

```
[1]  7.00000  0.19490 35.91585
```

Usually, we will use smaller intervals, as well as population rates that actually *do* vary by calendar time, but that would require more complicated computing:

Chapter 7

Analysis of Epidemiological Data Solutions

7.1 Single incidence rates

1. First we enter the numbers of stomach cancer deaths and the number of person-years in two vectors, each of length two representing Kuwait and Egypt respectively

```
> cases <- c(6, 53)
> pyears <- c(0.89, 18.19)
```

We can the divide the two vectors to form the vector of rates. For readability we give names to the vector components using `names() <-`. Finally we print it by just giving the name of the vector:

```
> rates <- cases/pyears
> names(rates) <- c("Kuwait", "Egypt")
> rates
```

```
  Kuwait    Egypt
6.741573 2.913689
```

In order to compute the uncertainty in the empirical mortality rate we use the formula for the standard error of a rate; $\mathrm{SE}(I) = I/\sqrt{D}$, where $I = D/Y$ is the empirical rate and D is the number of deaths:

```
> SE.r <- rates / sqrt(cases)
> CL.low <- rates - 1.96*SE.r
> CL.up <- rates + 1.96*SE.r
> cbind(rates, SE.r, CL.low, CL.up)
```

```
          rates      SE.r   CL.low     CL.up
Kuwait 6.741573 2.7522357 1.347191 12.135955
Egypt  2.913689 0.4002259 2.129246  3.698132
```

Note that we used `cbind()` to collect the results in a matrix

2. It is useful to see if the confidence intervals were substantially different if we used the standard approximation the standard deviation of the log-rate: $\mathrm{SE}(\log I) = 1/\sqrt{D}$:

```
> SE.logr <- sqrt(1/cases)
> CL.low <- rates/exp(1.96*SE.logr)
> CL.up <- rates*exp(1.96*SE.logr)
> cbind(rates, SE.logr, CL.low, CL.up)
```

```
          rates   SE.logr   CL.low     CL.up
Kuwait 6.741573 0.4082483 3.028679 15.006147
Egypt  2.913689 0.1373606 2.225971  3.813878
```

The confidence intervals computed by these two approximate methods are relatively wide and somewhat different, too, for Kuwait with a fairly small number of cases, but they are narrow and quite close to each other for Egypt with a large number of cases.

7.2 Non-significant difference

The possible choices based on a significant finding for the difference in rates based on 1 in 200 man and 1 in 10 women were:

1. The results provide supporting evidence for the hypothesis of no real difference between males and females in the breast cancer risk among electric engineers.
2. The results are consistent with the universal observation that the risk of breast cancer among females is clearly higher than that in males.
3. No conclusion can be made from this result concerning the male/female contrast in breast cancer incidence among graduates of electric engineering.
4. Other conclusion, what?

Out of these alternatives no. 2. appears as the most appropriate interpretation. It takes into account the available external knowledge that is relevant for the question of interest. Alternative 3. is not totally unreasonable, because these data alone do not provide any adequate statistical information as such about the female/male contrast in breast cancer incidence.

A rough comparison in relative terms suggests that females had a 20-fold rate of breast cancer in this small population. However, with only one male case and one female case it is waste of time to try computing any more refined quantitative estimate (and confidence interval) for the relative rate of breast cancer between the two genders.

7.3 Preventive trial

1. The study hypothesis is that Beta Carotene reduces lung cancer incidence among smokers. The corresponding null hypothesis is that the lung cancer incidence is the same in the two treatment arms.
2. First we set up vectors of cases and rates:

   ```
   > cases <- c(474, 402)
   > rates <- c(56.3, 47.5) # per 10000 years
   ```

   For readability, we provide the vector of cases with names:

```
> names( rates ) <- c("BetaCarotene","Placebo")
```

Since the rates are expressed per 10000 person-years, they are computed as

$$\text{rate} = (\text{cases}/Y) \times 10000$$

which is solved for Y to give:

$$Y = (\text{cases}/\text{rate}) \times 10000$$

So the calculation in R is straightforward:

```
> pyears <- (cases/rates)*10000
> pyears
```

```
BetaCarotene      Placebo
    84191.83     84631.58
```

3. The estimate of the theoretical rate ratio $\rho = \lambda_1/\lambda_0$, is simply the ratio of the two empirical rates: $\widehat{\rho} = \text{IR} = I_1/I_0$. The absolute numbers of cases are in turn needed to compute the confidence interval for ρ. The standard error of the log(IR) is $\sqrt{1/D_1 + 1/D_0}$, which is what we compute in the second line:

```
> ratio <- rates[1]/rates[2]
> SE.logr <- sqrt(sum(1/cases))
> ratio.95low <- ratio/exp(1.96*SE.logr)
> ratio.95up  <- ratio*exp(1.96*SE.logr)
> cbind(ratio, SE.logr, ratio.95low, ratio.95up)
```

```
                ratio    SE.logr ratio.95low ratio.95up
BetaCarotene 1.185263 0.06780315    1.037766   1.353724
```

The estimated rate ratio thus suggests an *increase* by about 18-19 percent of lung cancer incidence in beta carotene group as compared with the placebo group. The empirical result is consistent even with the possibility that the rate in the supplementation group would be 35% higher that in the placebo group. A relative rate of this size does not seem impressive as such. Yet, the result is alarming, considering that it was initially hypothesized that beta caroten supplementation would hopefully *reduce* the already high lung cancer incidence among smokers.

4. The estimate of the rate difference $\delta = \lambda_1 - \lambda_0$; *i.e.* the excess (or deficit) rate is just the difference between the two empirical rates: $\widehat{\delta} = \text{ID} = I_1 - I_0$. The standard error of this is computed according to the formula from the lecture notes:

$$\text{SE}(I_1 - I_0) = \sqrt{I_1^2/D_1 + I_0^2/D_0}$$

which is what we do in the second line. Note that the confidence limits are computed on the rate scale:

```
> diff <- rates[1] - rates[2]
> SE.diff <- sqrt(sum(rates^2/cases))
> diff.95low <- diff - 1.96*SE.diff
> diff.95up <- diff + 1.96*SE.diff
> cbind(diff, SE.diff, diff.95low, diff.95up)
```

```
             diff  SE.diff diff.95low diff.95up
BetaCarotene  8.8 3.507089   1.926106  15.67389
```

This result suggests that there would be about 9 excess cases of lung cancer per year in 10000 men, who are on beta caroten as compared with 10000 men without this supplementation.

5. A formal test can be based on the difference between the rates; we take the difference in rates, divide by its standard error, square it and look it up in a χ^2-distribution with 1 d.f.:

```
> Z <- diff/SE.diff
> P <- 1 - pchisq( Z^2, 1 )
> test.diff <- cbind(Z, P)
> test.diff

                    Z          P
BetaCarotene 2.509204 0.01210037
```

Alternatively we can base the test on the difference in the log-rates. The difference in log-rates is the same as the log of the rate.ratio, so the calculation becomes:

```
> Z <- log(ratio)/SE.logr
> P <- 1 - pchisq( Z^2, 1 )
> ( test.ratio <- cbind(Z, P) )

                    Z          P
BetaCarotene 2.506739 0.01218505
```

We can for for easier comparison show the two results underneath each other:

```
> tt <- rbind( test.diff, test.ratio )
> rownames( tt ) <- c("diff","ratio")
> tt

              Z          P
diff  2.509204 0.01210037
ratio 2.506739 0.01218505
```

We see that (with a study of this size) the values of the two test statistics are practically the same regardless of the scale we use for testing (rate or log-rate).

6. The data comes from a randomzed trial, so any difference in age-distribution should be purely incidental. By the size of the study the chance of confounding by an accidental imbalance is therefore remote.

   Neither is there any possibility for confounding by smoking status. All enrolled persons were regular smokers, and the daily amount of cigarettes smoked should have similar distributions in the randomized groups.

   The result provides some evidence against the *null hypothesis* $H_0 : \rho = 1$. However, the direction of the observed rate ratio from the null hypothesis was very surprising given the anticipation that beta carotene would actually reduce the rate of lung cancer among smokers. Thus, one would perhaps not yet "reject" the null hypothesis of no effect in spite of the "significant" P-value obtained in a two-tailed test. However, the result can be viewed to provide more evidence against the initial *research hypothesis* of clinically relevant beneficial effect. – Interpretation of these results combined with those from similar trials will be continued in the next exercise.

7.4 Preventive trial – interpretation

1. Given that the direction of the observed rate ratio was – quite surprisingly – against the research hypothesis and observational evidence, one would perhaps not yet conclude on the basis of this single study that beta carotene supplementation would actually be *harmful*. Yet, as the result was strongly against the hypothesis of a *beneficial* effect, a reasonable practical conclusion would be withhold from recommending this target group to take beta carotene supplementation.

2. These two studies together, in which very similar results were obtained, do now provide more convincing evidence for a harmful effect of beta carotene supplementation in a target population like this.

3. The result from the American Physicians' Study is in no conflict with the two other studies but is actually quite consistent with them. The confidence interval here is wider due to smaller numbers of outcome cases, but is clearly overlapping with those of the other studies.

4. We cannot conclude anything about the effect of beta caroten supplementation in non-smoking men on the basis of the results of this single study with such a wide confidence interval. In particular, there is inadequate evidence concerning the issue whether the effect among non-smokers would be essentially different from that among smokers.

7.5 Geographical variation

There is no paradox. Subdivision by counties implies that the county on the left side probably has a larger population base than any of the smaller counties. Therefore, the chance variation in the incidence rate in that county is smaller, and as a consequence there is a larger propensity to have a "significantly" elevated rate. However, subdivision by hospital districts creates relatively smaller areas within that county, and the individual rates in these districts are affected by larger random variability than those in the remaining big district — or the county containing these small districts.

7.6 Efficiency of study design

In a comaprison of two groups, the limiting factor is the number of cases in the group with the smallest *number of cases* (which is not necessarily the smallest group).

However, if we assume that the anticipated RR associated with the exposure is not extremely large, we can assume that the smaller number of case will occur in the smaller group.

Hence we have two options:

1. Extend the follow-up time to accrue more cases.

2. Change the exposure allocation, such that we get two groups that have similar number of cases. If we anticipate a RR of 2 associated with exposure we should have 1/3 in the exposed group and 2/3 in the unexposed group. Specifically, allocate

exposed and unexposed in the inverse proportion to the anticipated RR to get the maximal precision.

7.6.1 An illustration by simulation

We start by taking the initial proposal and take 2000 exposed and 8000 unexposed, and assume that the cancer incidence rate is 150 per 100,000 person-years and the RR associated with X is 1.85, and finally that the follow-up perid is going to be 1 year.

We put the follow-up time in `T` and then set up vectors of length 2: `G` — ecsposure group, `N` — number of persons, `Y` — person-years, `E` — expected number of cases, `D` — a simulated number of cases

```
> t <- 1
> r <- 150/100000
> rr <- 2
> G <- factor( c("ctr","X") )
> N <- c(8000,2000)
> Y <- N * t
> E <- Y * c(1,rr) * r
> D <- rpois( 2, E )
> # and print the results nicely
> data.frame( G, N, Y, E, D )
```

```
    G    N    Y  E  D
1 ctr 8000 8000 12 14
2   X 2000 2000  6  6
```

With the number of person-years and cases we can now compute the observed rates and the rate-ratio with confidence interval:

```
> rates <- D/Y
> RR <- rates[2]/rates[1]
> erf <- exp(1.96 * sqrt(sum(1/D)) )
> round( c( RR, RR/erf, RR*erf, erf ), 3 )
```

```
[1] 1.714 0.659 4.461 2.602
```

So in this scenario it is clear that we cannot expect to get a precise picture of the RR — the error factor (the last of the 4 numbers) os quite large.

But we could try to do the same again, extending the follow-up to 3 years, say:

```
> t <- 3
> r <- 150/100000
> rr <- 2
> G <- factor( c("ctr","X") )
> N <- c(8000,2000)
> Y <- N * t
> E <- Y * c(1,rr) * r
> D <- rpois( 2, E )
> # and print the results nicely
> data.frame( G, N, Y, E, D )
```

```
    G    N     Y  E  D
1 ctr 8000 24000 36 31
2   X 2000  6000 18 22
```

```
> rates <- D/Y
> RR <- rates[2]/rates[1]
> erf <- exp(1.96 * sqrt(sum(1/D)) )
> round( c( RR, RR/erf, RR*erf, erf ), 3 )
```

```
[1] 2.839 1.644 4.902 1.727
```

We see that we have a somewhat better precision, but the relative unceratinty in the RR is still quite large (the last number, `erf`).

The other possibility would be to balance exposed and unexposed more evenly. Or specifically so that the ratio of unexposed to exposed equals the rate-ratio, thereby creating an approximate equal number of cases in the two groups:

```
> t <- 1
> r <- 150/100000
> rr <- 2
> G <- factor( c("ctr","X") )
> N <- c(6000,4000)
> Y <- N * t
> E <- Y * c(1,rr) * r
> D <- rpois( 2, E )
> # and print the results nicely
> data.frame( G, N, Y, E, D )
```

```
    G    N    Y  E  D
1 ctr 6000 6000  9  8
2   X 4000 4000 12 12
```

```
> rates <- D/Y
> RR <- rates[2]/rates[1]
> erf <- exp(1.96 * sqrt(sum(1/D)) )
> round( c( RR, RR/erf, RR*erf, erf ), 3 )
```

```
[1] 2.250 0.920 5.504 2.446
```

We see that the error-factor is smaller than in the first instance, but what really matters is to increase the follow-up time.

Writing a small R-function

We can of course not really conclude much from a single simulation, so it would be useful to be able to do these calculations with a single command. This is dome by wrapping it all in a function. What we want to be able to hand over as arguments to the function is the follow-up time and the exposure allocation.

So we basically take the code from before

```
> sim <- function( t, N )
+ {
+ r <- 150/100000
+ rr <- 2
+ G <- factor( c("ctr","X") )
+ Y <- N * t
+ E <- Y * c(1,rr) * r
+ D <- rpois( 2, E )
+ rates <- D/Y
+ RR <- rates[2]/rates[1]
+ erf <- exp(1.96 * sqrt(sum(1/D)) )
+ c( RR, RR/erf, RR*erf, erf )
+ }
```

What this function resturns is the value of the *last* expression evaluated, in this case a vector of 4:

```
> sim( t=1, N=c(8000,2000) )
```

```
[1] 3.000000 1.040906 8.646315 2.882105
```

```
> rbind(
+ sim( t=1, N=c(8000,2000) ),
+ sim( t=2, N=c(8000,2000) ),
+ sim( t=3, N=c(8000,2000) ),
+ sim( t=4, N=c(8000,2000) ) )
```

```
         [,1]      [,2]     [,3]     [,4]
[1,] 1.111111 0.4125208 2.992741 2.693467
[2,] 1.875000 1.0154101 3.462271 1.846545
[3,] 3.692308 2.1200420 6.430597 1.741620
[4,] 2.297872 1.4314367 3.688754 1.605291
```

Clarelæy there is a decrease in the uncertainty.

We can also see how the proportion of cases influence the results:

```
> rbind(
+ sim( t=2, N=c(8000,2000) ),
+ sim( t=2, N=c(7000,3000) ),
+ sim( t=2, N=c(6000,4000) ),
+ sim( t=2, N=c(5000,5000) ) )
```

```
         [,1]      [,2]     [,3]     [,4]
[1,] 2.240000 1.1644046 4.309155 1.923730
[2,] 2.592593 1.3714490 4.901047 1.890404
[3,] 1.800000 0.9943830 3.258302 1.810168
[4,] 1.588235 0.8656981 2.913823 1.834630
```

Here the effects on the precison are much smaller.

So if we ant a clear picture of what goes on we must make a lot of simulations to see how the error-factor varies.

7.7 Case-control study: MI

1. First we input the data in 4 vectors, each of length 2, where the first element represents males and the second females:

   ```
   > D1 <- c(141, 49)
   > D0 <- c(144, 32)
   > C1 <- c(208, 58)
   > C0 <- c(112, 45)
   ```

   In order to get nice results we annotate the vectors by a name-vector:

   ```
   > names(D1) <-
   + names(D0) <-
   + names(C1) <-
   + names(C0) <- c("M","F")
   ```

This way we can do all the calculations simultaneously for males and females just using the usual formulae – the ratio of the exposure odds between cases and controls:

```
> EOR <- (D1/D0)/(C1/C0)
> SE.leor <- sqrt(1/D1 + 1/D0 + 1/C1 + 1/C0)
> EOR.95low <- EOR / exp(1.96*SE.leor)
> EOR.95up  <- EOR * exp(1.96*SE.leor)
```

Finally we place the resulting exposure odds ratios together with the confidence intervals for the corresponding hazard ratios below each other:

```
> strata <- cbind(EOR, SE.leor, EOR.95low, EOR.95up)
> round( strata, 3 )
```

```
    EOR SE.leor EOR.95low EOR.95up
M 0.527   0.167     0.380    0.731
F 1.188   0.302     0.657    2.147
```

We see that the exposure odds-ratio for men is much smaller than for women – actually even "significantly" smaller than 1 – suggesting that high physical activity seems to be protective against MI in men.

Whether it is so in women too is difficult to say. Note that the confidence intervals for the hazard ratios overlap, so we cannot base too much on that observation. If the confidence intervals had been clearly non-overlapping we could have inferred that maybe the hazard ratios were different, but we cannot make the opposite conclusion here.

2. One way to test for homogeneity of the true hazard ratios across the genders is to compute the log of the ratio of the two exposure odds-ratios – the difference of the log-odds-ratios – and then compare this with its standard error. The latter is computed using the fact that the two log-odds-ratios are independent.

```
> EOR.ratio <- EOR[1] / EOR[2]
> V.logEOR.ratio <- SE.leor[1]^2 + SE.leor[2]^2
> Wald <- log(EOR.ratio)^2 / V.logEOR.ratio
> P.Wald <- 1 - pchisq(Wald, df=1)
> round( cbind(Wald,P.Wald), 4 )
```

```
   Wald P.Wald
M 5.551 0.0185
```

The Wald statistic is 5.551, which evaluated in a χ^2-distribution with 1 d.f. gives a p-value of 0.018. Hence, these data provide some evidence that physical exercise would have greater effect in men than in women.

3. If it really were so that we had an interaction – the hazard ratio of MI associated with physical activity is *not* the same between males and females – it would really not be meaningful to adjust for confounding by a single summary EOR that assumes homogeneity of hazard ratios.

4. However, for the sake of the exercise, we first compute the crude odds-ratio based on simple sums of each of the two-component vectors.

```
> EOR.crude <- (sum(D1)/sum(D0)) / (sum(C1)/sum(C0))
> SE.lc <- sqrt( 1/sum(D1) + 1/sum(D0) + 1/sum(C1) + 1/sum(C0) )
> EOR.c95low <- EOR.crude / exp(1.96*SE.lc )
> EOR.c95up <- EOR.crude * exp(1.96*SE.lc )
> cbind(EOR.crude, EOR.c95low, EOR.c95up)

     EOR.crude EOR.c95low EOR.c95up
[1,] 0.6371753  0.4793904 0.8468931
```

5. For comparison we can then by hand compute the MH-estimate of the common odds-ratio for men and women:

```
> T <- D1+D0+C1+C0
> A <- (D1 + C0)/T
> B <- (D0 + C1)/T
> P <- D1*C0 / T
> Q <- D0*C1 / T
> EOR.mh <- sum(P) / sum(Q)
> V.lmh <- sum(A*P) / (2*sum(P)^2) +
+          sum(A*Q + B*P) / (2*sum(P)*sum(Q)) +
+          sum(B*Q) / (2*sum(Q)^2)
> EOR.mh95low <- EOR.mh / exp(1.96*sqrt(V.lmh))
> EOR.mh95up <- EOR.mh * exp(1.96*sqrt(V.lmh))
> cbind(EOR.mh, EOR.mh95low, EOR.mh95up)

        EOR.mh EOR.mh95low EOR.mh95up
[1,] 0.6390899    0.481216   0.848758
```

We see that the results are pretty much the same, so the subdivision by sex does not alter the estimate much.

To see if there would be any *a priori* reason to suspect confounding we check whether the potential confounder (sex) is associated with the exposure (physical activity) by computing the prevalence of physically active in each sex. Now recall that we entered the data as vectors with male and female numbers, so we can compute the fraction of exposed by a simple operation:

```
> p.exp <- C1/(C1+C0)
> p.exp

        M         F
0.6500000 0.5631068
```

7.7.1 Statistical modelling

The questions in this exercise can also be answered quite easily using a statistical model called *logistic regression* and fitting it using appropriate statistical functions in R. Analysis of case-control data is done by taking the case-control status as the outcome variable in logistic regression, and other variables are used as explanatory variables or covariates.

Logistic regression in R takes as the response variable a two-column matrix, where the first column contains the "failures" (here: cases) and the second the "non-failures" (here: controls):

```
> y <- cbind( D=c(D0,D1), C=c(C0,C1) )
> sex <- factor( rep(c("M","F"),2) )
> phys <- factor( rep(c("N","Y"),each=2) )
> data.frame( y, sex, phys )
```

```
    D   C sex phys
1 144 112   M    N
2  32  45   F    N
3 141 208   M    Y
4  49  58   F    Y
```

```
> cbind( y, sex, phys )
```

```
    D   C sex phys
M 144 112   2    1
F  32  45   1    1
M 141 208   2    2
F  49  58   1    2
```

With all these items in place, we can now fit a logistic regression model, including separate effects of `phys` for each sex, corresponding to the first question:

```
> mimod <- glm( y ~ sex + sex:phys, family=binomial )
> summary( mimod )
```

```
Call:
glm(formula = y ~ sex + sex:phys, family = binomial)

Deviance Residuals:
[1]  0  0  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3409     0.2312  -1.474 0.140391
sexM          0.5922     0.2633   2.249 0.024512
sexF:physY    0.1723     0.3019   0.571 0.568135
sexM:physY   -0.6401     0.1667  -3.841 0.000123

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.5795e+01  on 3  degrees of freedom
Residual deviance: 5.0848e-14  on 0  degrees of freedom
AIC: 30.149

Number of Fisher Scoring iterations: 2
```

The coefficients reported in the output refer to logarithmis of hazard ratios. If we want the estimated hazard ratios themselves, we use the function `ci.lin` in `Epi` package to extract the coefficients and exponentiate them. When calling this function, we activate the `Exp=TRUE` argument. Finally, using the `subset` argument too, only the relevant components are extracted from the output.

```
> library(Epi)
> ci.lin( mimod, Exp=TRUE )
```

```
              Estimate    StdErr          z            P exp(Est.)      2.5%
(Intercept) -0.3409266 0.2312406 -1.4743372 0.1403908318 0.7111111 0.4519653
sexM         0.5922410 0.2633348  2.2490039 0.0245122487 1.8080357 1.0790859
sexF:physY   0.1723039 0.3018638  0.5708001 0.5681351746 1.1880388 0.6574817
sexM:physY  -0.6400926 0.1666521 -3.8408925 0.0001225878 0.5272436 0.3803267
                97.5%
(Intercept) 1.1188447
sexM        3.0294096
sexF:physY  2.1467306
sexM:physY  0.7309132
```

```
> round( ci.lin( mimod, subset="phys", Exp=TRUE )[,5:7], 3 )


           exp(Est.)  2.5% 97.5%
sexF:physY     1.188 0.657 2.147
sexM:physY     0.527 0.380 0.731
```

To compute the crude odds-ratio, ignoring sex, we just fit the model including only `phys` as an explanatory variable:

```
> mcmod <- glm( y ~ phys, family=binomial )
> summary( mcmod )


Call:
glm(formula = y ~ phys, family = binomial)

Deviance Residuals:
      M        F        M        F
 1.0907  -1.9853  -0.4803   0.8625

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1142     0.1098   1.041   0.2980
physY        -0.4507     0.1452  -3.105   0.0019

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15.7949  on 3  degrees of freedom
Residual deviance:  6.1058  on 2  degrees of freedom
AIC: 32.255

Number of Fisher Scoring iterations: 3


> round( ci.lin( mcmod, subset="phys", Exp=TRUE )[,5:7,drop=F], 3 )


      exp(Est.)  2.5% 97.5%
physY     0.637 0.479 0.847
```

Finally, to fit the model where we assume a homogenous hazard ratio and want to adjust for sex, we add shall the sex factor to this latter model. This can be done using the `update` function:

```
> msmod <- update( mcmod, . ~ . + sex )
> round( ci.lin( msmod, subset="phys", Exp=TRUE )[,5:7,drop=F], 3 )


      exp(Est.)  2.5% 97.5%
physY     0.637 0.479 0.847
```

We see that the estimate where we fit a proper model has a value which is quite close to the value of the MH-estimate.

The modelling approach has the advantage that we get the possibility to estimate the quantities we want, with easily computed confidence intervals. Moreover we have the possibility of comparing the models with likelihood ratio tests:

```
> anova( mimod, msmod, mcmod, test="Chisq" )
```

```
Analysis of Deviance Table

Model 1: y ~ sex + sex:phys
Model 2: y ~ phys + sex
Model 3: y ~ phys
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         0     0.0000
2         1     5.5762 -1  -5.5762  0.01821
3         2     6.1058 -1  -0.5296  0.46678
```

Again we see that there is clear evidence of interaction and of course also of a strong sex-effect. So from a proper statistical point of view, the relevant model for this dataset seems to be the model with interaction, *i.e.* with separate hazard ratios of MI associated with physical activity for males and females.

7.8 Case-control study: Neonates

1. During the 17 years (1973-1989) there must have been hundreds of new cases of leukaemia among children < 15 y in Sweden, as even in Finland the 5-year number of cases among boys only was 113 in 1993-97 (see practical **??**). Let us say that there were 1000 cases and hence 5 x 1000 = 5000 controls. A crude analysis would then be based on these figures.

   (a) We can actually simplify the computations a bit:

   ```
   > EOR.crude <- (8/(1000-8)) / (2/(5000-2))
   > EOR.c <- (8/1)/(2/5)
   > cbind( EOR.crude, EOR.c )

        EOR.crude EOR.c
   [1,]  20.15323    20
   ```

   You can see that there is not much influence by the actual number of cases and controls — all the information we need is that there were 5 times as many controls as cases.

   (b) The same goes for the calculation of the standard deviation of the log odds-ratio:

   ```
   > SE.crude <- sqrt( 1/8 + 1/(1000-8) + 1/2 + 1/(5000-2) )
   > SE.c <- sqrt( 1/8 + 1/2 )
   > cbind( SE.crude, SE.c )

         SE.crude      SE.c
   [1,] 0.7913331 0.7905694
   ```

   So by this token we can compute the confidence intervals based on the approximate figures:

   ```
   > EOR.c95low <- EOR.c / exp(1.96*SE.c)
   > EOR.c95up  <- EOR.c * exp(1.96*SE.c)
   > round( cbind( EOR.c, SE.c, EOR.c95low, EOR.c95up ), 3 )

        EOR.c  SE.c EOR.c95low EOR.c95up
   [1,]    20 0.791      4.247    94.184
   ```

   So based on this computation there is some evidence that Down's syndrome predisposes to leukaemia.

2. In order to be able to produce a more reliable estimate of the effect of Down's syndrome, we would have to have at least data on age and sex (the matching

variables). As a minimum this would require a table classified by case/control status, exposure status (Down's syndrome "yes/no"), age and sex.

We would then fit a logistic regression with case-control status as outcome, and Down's syndrome and age×sex as explanatory variables. The last term, the interaction between age and sex will not be significant (because it is balanced between cases and controls by the very design of the study), but is must be included in the model because the study was designed as stratified on these.

7.9 Matched case-control study: Chemicals

1. We first input the data; this is simply done by entering the exposure status for the case-series and the control-series separately:

```
> library( Epi )
> casexp <- c(1,1,0,1,0,1,1,1,1,0, 0,1,1,0,1,1,1,1,0,1)
> conexp <- c(0,0,0,1,1,0,0,0,1,0, 1,1,0,0,0,0,0,1,0,0)
> cbind(casexp,conexp)
```

```
      casexp conexp
 [1,]      1      0
 [2,]      1      0
 [3,]      0      0
 [4,]      1      1
 [5,]      0      1
 [6,]      1      0
 [7,]      1      0
 [8,]      1      0
 [9,]      1      1
[10,]      0      0
[11,]      0      1
[12,]      1      1
[13,]      1      0
[14,]      0      0
[15,]      1      0
[16,]      1      0
[17,]      1      0
[18,]      1      1
[19,]      0      0
[20,]      1      0
```

   (a) Ignoring the matching, simply mean that we only use the number of exposed and non-exposed cases and controls respectively, so this is a simple tabulation:

```
> D1 <- sum(casexp)
> D0 <- length(casexp) - sum(casexp)
> C1 <- sum(conexp)
> C0 <- length(conexp) - sum(conexp)
> table.u <- rbind(c(D1, D0), c(C1, C0))
> rownames(table.u) <- c("Cases", "Controls")
> colnames(table.u) <- c("Exposed", "Unexposed")
> table.u
```

```
         Exposed Unexposed
Cases         14         6
Controls       6        14
```

   Based on this table we can compute the odds-ratio and associated confidence interval:

```
> EOR.un <- (D1/D0)/(C1/C0)
> SE.lun <- sqrt( 1/D1 + 1/D0 + 1/C1 + 1/C0 )
> EOR.un95low <- EOR.un / exp(1.96*SE.lun)
> EOR.un95up <- EOR.un * exp(1.96*SE.lun)
> round( cbind(EOR.un, SE.lun, EOR.un95low, EOR.un95up), 3 )

     EOR.un SE.lun EOR.un95low EOR.un95up
[1,]  5.444   0.69       1.408     21.055
```

(b) When we do the analysis based on the assumption of matched data collection, we need to tabulate the *matched pairs* by exposure status of the cases and the controls respectively:

```
> ( table.m <- table( casexp, conexp ) )

      conexp
casexp  0  1
     0  4  2
     1 10  4

> EOR.mh  <- table.m[2,1] / table.m[1,2]
> SE.lmh <- sqrt( 1/table.m[2,1] + 1/table.m[1,2] )
> EOR.mh95lo <- EOR.mh / exp(1.96*SE.lmh)
> EOR.mh95up <- EOR.mh * exp(1.96*SE.lmh)
> round( cbind(EOR.mh, SE.lmh, EOR.mh95lo, EOR.mh95up), 3 )

     EOR.mh SE.lmh EOR.mh95lo EOR.mh95up
[1,]      5  0.775      1.096      22.82
```

2. We see that unstratified analysis gives a slightly higher estimate and a lower standard error of the estimate. So the consequence is the that exposure effect is exaggerated if the matching in the study design is ignored in the analysis.

7.9.1 Statistical modelling

1. If we want to do the unmatched analysis of the data by logistic regression, we need to put the exposures into one long vector and create a vector of case-control status:

```
> exp <- c(casexp,conexp)
> cc <- rep(1:0,each=length(casexp))
> mc <- glm( cc ~ factor(exp), family=binomial )
> summary( mc )


Call:
glm(formula = cc ~ factor(exp), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5518  -0.8446   0.0000   0.8446   1.5518

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.8473     0.4879  -1.736   0.0825
factor(exp)1   1.6946     0.6901   2.456   0.0141

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.452  on 39  degrees of freedom
Residual deviance: 48.869  on 38  degrees of freedom
AIC: 52.869

Number of Fisher Scoring iterations: 4
```

```
> library(Epi)
> ci.lin( mc, subset="exp", Exp=TRUE )[,5:7,drop=FALSE]
```

```
             exp(Est.)     2.5%    97.5%
factor(exp)1  5.444444 1.407891 21.05417
```

This analysis may seem a bit of an overkill, since we are just analyzing a two by two table, but the modelling approach is generalizable to instances where more covariates are recorded.

Alternatively, since it is just a two by two table, we could use the `twoby2` command in the `Epi` package:

```
> twoby2(table.u)
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Exposed
Comparing : Cases vs. Controls

         Exposed Unexposed    P(Exposed) 95% conf. interval
Cases         14          6          0.7    0.4728   0.8586
Controls       6         14          0.3    0.1414   0.5272

                                  95% conf. interval
            Relative Risk: 2.3333    1.1263   4.8339
        Sample Odds Ratio: 5.4444    1.4079  21.0542
Conditional MLE Odds Ratio: 5.1912    1.1834  26.2754
    Probability difference: 0.4000    0.0903   0.6185

            Exact P-value: 0.0256
       Asymptotic P-value: 0.0141
------------------------------------------------------
```

We see that all approaches gives the same results.

2. If we want the matched analysis we must use the `clogit` function from the `survival` package, which is one of the built-in packages in R. What is needed is the same data as before but also a vector indication which observations that come from the same matched pair:

```
> mp <- rep(1:length(casexp),2)
> cbind( cc, exp, mp )
```

```
      cc exp mp
 [1,]  1   1  1
 [2,]  1   1  2
 [3,]  1   0  3
 [4,]  1   1  4
 [5,]  1   0  5
 [6,]  1   1  6
 [7,]  1   1  7
 [8,]  1   1  8
 [9,]  1   1  9
[10,]  1   0 10
[11,]  1   0 11
[12,]  1   1 12
[13,]  1   1 13
[14,]  1   0 14
[15,]  1   1 15
```

```
[16,]  1   1 16
[17,]  1   1 17
[18,]  1   1 18
[19,]  1   0 19
[20,]  1   1 20
[21,]  0   0  1
[22,]  0   0  2
[23,]  0   0  3
[24,]  0   1  4
[25,]  0   1  5
[26,]  0   0  6
[27,]  0   0  7
[28,]  0   0  8
[29,]  0   1  9
[30,]  0   0 10
[31,]  0   1 11
[32,]  0   1 12
[33,]  0   0 13
[34,]  0   0 14
[35,]  0   0 15
[36,]  0   0 16
[37,]  0   0 17
[38,]  0   1 18
[39,]  0   0 19
[40,]  0   0 20
```

With this data layout we can do the matched analysis, and use the `ci.lin` function to extract the parameters as before:

```
> library( survival )
> mm <- clogit( cc ~ exp + strata(mp) )
> ci.lin( mm, subset="exp", Exp=TRUE )[,5:7,drop=FALSE]
```

```
    exp(Est.)    2.5%    97.5%
exp         5 1.09555 22.81959
```

As before we get exactly the same results as when we used the "hand calculations", but the point here is that the modelling approach allows you to include further covariates in the analysis.

7.10 Cohort study and SMR

First we enter the number of cases and person-years (in 1000s) in vectors, one for each group, and also put names on the three age-groups

```
> library( Epi )
> D1 <- c(11, 15, 10)
> Y1 <- c(10, 6, 2)
> D0 <- c(15, 60, 150)
> Y0 <- c(30, 50, 70)
> names(D1) <-
+ names(Y1) <-
+ names(D0) <-
+ names(Y0) <-
+ c("30-39", "40-49", "50-59")
> cbind( D1, Y1, D0, Y0 )
```

```
      D1 Y1  D0 Y0
30-39 11 10  15 30
40-49 15  6  60 50
50-59 10  2 150 70
```

1. First we compute the age-specific rates (per 1000 PY) in the workers group and in the population, and then divide them to form the rate-ratio:

```
> I1 <- D1/Y1
> I0 <- D0/Y0
> IR <- I1/I0
> round(cbind(I1, I0, IR),2 )
```

```
      I1   I0   IR
30-39 1.1 0.50 2.20
40-49 2.5 1.20 2.08
50-59 5.0 2.14 2.33
```

   The rate ratio does look reasonably stable across the age range. We could expand with the 95% error factors for the rate ratio, to get a feel for how precisely the rate ratios are estimated:

```
> EF <- exp( 1.96 * sqrt(1/D1+1/D0) )
> round( cbind(I1, I0, IR, EF), 2 )
```

```
      I1   I0   IR   EF
30-39 1.1 0.50 2.20 2.18
40-49 2.5 1.20 2.08 1.76
50-59 5.0 2.14 2.33 1.90
```

   We see that the variation between the rates is very small compared to the statistical uncertainty in the rates themselves.

2. The crude rates and their ratio can be computed:

```
> I1.c <- sum(D1) / sum(Y1)
> I0.c <- sum(D0) / sum(Y0)
> IR.c <- I1.c / I0.c
> round( cbind(I1.c, I0.c, IR.c), 2)
```

```
     I1.c I0.c IR.c
[1,]    2  1.5 1.33
```

3. The Mantel-Haenszel estimate of the rate ratio is computed, according to the formula:

```
> IR.mh <- sum( D1*Y0/(Y1+Y0) ) / sum( D0*Y1/(Y1+Y0) )
> round(IR.mh, 2)
```

```
[1] 2.19
```

4. The SMR is computed as O/E where O is the observed numbers in the workers' group, and E is the expected numbers assuming that the age-specific incidence rates in the reference population would also apply in the workers' group:

```
> Obs <- sum( D1 )
> Exp <- sum( I0 * Y1)
> SMR <- Obs / Exp
> round( cbind(Obs, Exp, SMR), 2)
```

```
     Obs   Exp  SMR
[1,]  36 16.49 2.18
```

5. The directly standardized rates are computed by taking the age-specific rates (`I1` and `I0`) and taking a weighted average. The weights in this case are the distribution of person-years in the population (`Y0`):

```
> I1.s <- sum( Y0*I1 ) / sum( Y0 )
> I0.s <- sum( Y0*I0 ) / sum( Y0 )
> IR.s <- I1.s / I0.s
> round( cbind(I1.s, I0.s, IR.s), 2 )
```

```
     I1.s I0.s IR.s
[1,] 3.39  1.5 2.26
```

6. To see if the standardized rates are sensitive to the choice of standard population, we repeat the calculation using instead the distribution of person-years in the workers' population:

```
> I1.x <- sum( Y1*I1 ) / sum( Y1 )
> I0.x <- sum( Y1*I0 ) / sum( Y1 )
> IR.x <- I1.x / I0.x
> round( cbind(I1.x, I0.x, IR.x), 2)
```

```
     I1.x I0.x IR.x
[1,]    2 0.92 2.18
```

The standardized rates are heavily influenced by the standard chosen, but since the ratio of the rates does not vary appreciably, the rate ratio estimate we get is reasonably stable across the various methods for computing it; that be Mantel-Haenszel, SMR or direct standardization. The ratio of the crude rates is however very misleading as an estimate of the true rate ratio.

7.10.1 Statistical modelling

We cannot reproduce any of these approaches easily with a statistical model, but we can compute the proper maximum likelihood estimate of the rate-ratio, using a Poisson model. We stack the two vectors of events and the two vectors of person-years, and generate two new vectors, one with the age, and one with and indicator of workers or population:

```
> D <- c(D1,D0)
> Y <- c(Y1,Y0)
> A <- factor( rep(names(D0),2) )
> G <- factor( rep(c("Wrk","Pop"),each=3) )
> data.frame( D, Y, A, G )
```

```
    D  Y     A   G
1  11 10 30-39 Wrk
2  15  6 40-49 Wrk
3  10  2 50-59 Wrk
4  15 30 30-39 Pop
5  60 50 40-49 Pop
6 150 70 50-59 Pop
```

Once we have this dataset we can estimate the crude rates as well as their ratio by just ignoring age in a model. We parametrize in two different ways, but the fit is the same:

```
> mc <- glm( D ~ G - 1 + offset(log(Y)), family=poisson )
> round( ci.lin( mc, Exp=TRUE )[,5:7], 2)

     exp(Est.) 2.5% 97.5%
GPop       1.5 1.32  1.71
GWrk       2.0 1.44  2.77
```

Here we recognize the crude rates (& confidence intervals). With a reparametrization we can get the baseline rate in the reference group and the rate ratio:

```
> mc <- glm( D ~ G + offset(log(Y)), family=poisson )
> round( ci.lin( mc, Exp=TRUE )[,5:7], 2)

            exp(Est.) 2.5% 97.5%
(Intercept)       1.50 1.32  1.71
GWrk              1.33 0.94  1.90
```

Likewise we can estimate the age-specific rates and rate-ratios by taking an interaction term into the model; in the first formulation we get the age-specific rates, in the latter we the age-specific rates in one group and the rate-ratios (& confidence intervals):

```
> mi <- glm( D ~ A:G -1 + offset(log(Y)), family=poisson )
> round( ci.lin( mi, Exp=TRUE )[,5:7], 2)

            exp(Est.) 2.5% 97.5%
A30-39:GPop       0.50 0.30  0.83
A40-49:GPop       1.20 0.93  1.55
A50-59:GPop       2.14 1.83  2.51
A30-39:GWrk       1.10 0.61  1.99
A40-49:GWrk       2.50 1.51  4.15
A50-59:GWrk       5.00 2.69  9.29


> mi <- glm( D ~ A-1 + A:G + offset(log(Y)), family=poisson )
> round( ci.lin( mi, Exp=TRUE )[,5:7], 2)

            exp(Est.) 2.5% 97.5%
A30-39            0.50 0.30  0.83
A40-49            1.20 0.93  1.55
A50-59            2.14 1.83  2.51
A30-39:GWrk       2.20 1.01  4.79
A40-49:GWrk       2.08 1.18  3.67
A50-59:GWrk       2.33 1.23  4.43
```

The proper overall rate ratio estimate is from the model where we assume that the rates are proportional between the two populations:

```
> ms <- glm( D ~ A + G + offset(log(Y)), family=poisson )
> ms

Call:  glm(formula = D ~ A + G + offset(log(Y)), family = poisson)

Coefficients:
(Intercept)       A40-49       A50-59         GWrk
    -0.6912       0.8633       1.4572       0.7839

Degrees of Freedom: 5 Total (i.e. Null);  2 Residual
Null Deviance:            61.53
Residual Deviance: 0.06734         AIC: 38.37
```

```
> round( ci.lin( ms, Exp=TRUE )[,5:7], 2 )
```

```
            exp(Est.) 2.5% 97.5%
(Intercept)      0.50 0.33  0.76
A40-49           2.37 1.51  3.73
A50-59           4.29 2.78  6.64
GWrk             2.19 1.51  3.18
```

We can also formally assess whether the model with the proportionality assumption is plausible; *i.e.* test it against the interaction model. We can of course also test the rather uninteresting hypotheses of no age effect or no group effect, but we will leave this out here.

```
> anova( ms, mi, test="Chisq" )
```

```
Analysis of Deviance Table

Model 1: D ~ A + G + offset(log(Y))
Model 2: D ~ A - 1 + A:G + offset(log(Y))
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         2    0.06734
2         0    0.00000  2  0.06734   0.9669
```

We see, as we would suspect from the computed age-specific rate ratios that there is no evidence for heterogeneity of the rate ratios. Since we have tabulated data we could have dispensed with this an just taken the test statistic for interaction from output from the model summary `Residual Deviance:  0.06734` — it is the same as the χ^2-statistic from the `anova` function.

Thus we again see that the most versatile tool in analysis of rates is a proper statistical model fitted in a program that allows to extract the relevant parts of the fit (and leave the irrelevant ones).

7.11 Trial of tolbutamide

1. First we enter the data and give names to the vectors

```
> library( Epi )
> options(digits=3)
> D <- c( 30,  21)
> n <- c(204, 215)
> names( D ) <-
+ names( n ) <- c("Tolbutamide", "Placebo")
> cbind( D, n )
```

```
             D   n
Tolbutamide 30 204
Placebo     21 215
```

   (a) The cumulative risk of death in the groups is just the ratio:

```
> Q <- D/n
> Q
```

```
Tolbutamide     Placebo
     0.1471      0.0977
```

(b) The estimated relative risk is just the ratio of these two numbers, and we use the well-known formula (from the lectures) for the standard error of the log-QR:

```
> QR <- Q[1]/Q[2]
> SE.lqr <- sqrt( 1/D[1]-1/n[1] + 1/D[2]-1/n[2] )
> QR.95lo <- QR / exp(1.96*SE.lqr)
> QR.95up <- QR * exp(1.96*SE.lqr)
> cbind( QR, SE.lqr, QR.95lo, QR.95up)

              QR SE.lqr QR.95lo QR.95up
Tolbutamide 1.51  0.267   0.892    2.54
```

(c) The difference in cumulative death probabilities is also estimated using the traditional formulae:

```
> QD <- Q[1] - Q[2]
> SE.qd <- sqrt( sum( Q*(1-Q)/n ) )
> QD.95low <- QD - 1.96*SE.qd
> QD.95up <- QD + 1.96*SE.qd
> cbind( QD, SE.qd, QD.95low, QD.95up)

                QD SE.qd QD.95low QD.95up
Tolbutamide 0.0494 0.032  -0.0134   0.112
```

All these computations (and a few more) are easily done using the `twoby2` function from the `Epi` package. Note that we need to input the number of survivors in the second column, not the total number:

```
> twoby2( cbind( D, n-D ) )

2 by 2 table analysis:
------------------------------------------------------
Outcome   : D
Comparing : Tolbutamide vs. Placebo

             D        P(D) 95% conf. interval
Tolbutamide 30 174  0.1471    0.1048    0.203
Placebo     21 194  0.0977    0.0645    0.145

                                  95% conf. interval
            Relative Risk: 1.5056    0.8918    2.542
        Sample Odds Ratio: 1.5928    0.8794    2.885
Conditional MLE Odds Ratio: 1.5910    0.8457    3.041
    Probability difference: 0.0494   -0.0137    0.114

            Exact P-value: 0.136
       Asymptotic P-value: 0.125
------------------------------------------------------
```

The estimated relative risk and its confidence interval is exactly reproduced by `twoby2`, but the traditional formula for the confidence interval for a difference of two proportions is not very accurate, so a better one is implemented in `twoby2`, hence the different result.

2. Even though the observed mortality in the Tolbutamide arm was 50% larger than in the placebo arm, there is no sufficient evidence yet for a higher mortality – the lower end of the confidence interval is about 0.9. Likewise is the lower bound for the confidence interval of the risk difference below 0. However, the result is alarming.