Practical exercises: Measures of Disease Occurrence Analysis of Epidemiological Data

Nordic Summer School of Cancer Epidemiology
Danish Cancer Society, 12–23 August, 2022
http://BendixCarstensen.com/NSCE/2022
Version 1.5

Compiled Thursday 21st July, 2022, 10:26
from: C:\Bendix\teach\NSCE\2022\pracs/pracs.tex

Bendix Carstensen Steno Diabetes Center Copenhagen, Gentofte, Denmark
& Department of Biostatistics, University of Copenhagen
bendix.carstensen@regionh.dk b@bxc.dk
http://BendixCarstensen.com

Esa Läärä Department of Mathematical Sciences
University of Oulu, Finland
Esa.Laara@oulu.fi

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 "`Download R for Windows`" and then click on "`base`" and further "`Download R 3.4.1 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 data sets 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 a script editor

1.3.1 Built-in editor in R

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.

1.3.2 Rstudio

is a front-end to Rwith many facilities. It is a commercial product but there is a free version which works excellent with many handy facilities; if you go to their website, https://www.rstudio.com/, it is easy to download and install.

It is becoming the de-facto interface to R so it is a good idea to use it; you will find that it is quite easy to get help on.

1.3.3 Try!

Now open a script by `File→New script`, and type:

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

Run the lines one at a time by pressing `CTRL-r` (if you are using RStudio it is `CTRL-Enter`), 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 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://bendixcarstensen.com/Epi/R-intro.pdf.

1.5 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.

A very short (12 pages) and handy introduction found there is "A (very) short Introduction to R" by Paul Torfs and Claudia Brauer
https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf. That will take you a long way.

Chapter 2

Measures of Disease Occurrence — Exercises

2.0 Using NORDCAN

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 Cancer Data, then NORDCAN - on the Web, and finally http://www-dep.iarc.fr/nordcan.htm.

3. On the page you just reached, 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 2012–16?

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 2012–2016?

4. What were the numbers of men and women living with lung cancer at the end of 2016, and how big were the corresponding proportions of lung cancer patients out of the whole male and female populations?

5. Compare the trends of age-standardized incidence and mortality rates in men and women. What kind of observations you make?

2.0.3 Incidence of lung cancer

Learn more about the incidence rates of lung cancer among men in the Nordic Countries during 2012-2016. 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 23. 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 23.

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 23. 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 2012–2016. Proceed as with the incidence of lung cancer 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 2016. 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 2016 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 47 to 51.

1. Create a graph showing the age specific incidence and mortality of lung cancer among men in Denmark during 2012-16. From *Incidence/Mortality*, under **Graphs** choose Age-specific curves. Any comments to the graph?

2. Repeat the previous task 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 48.

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 48.

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 51. You will also notice that a similar table is displayed as on slide 47.

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 2016.

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 62.

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 2016.

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

2. Compare the figures of these countries with those of Cali and Birmingham given on lecture slide 75.

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 2012–2016? What about female patients? How big where the 5-year relative survival proportions?

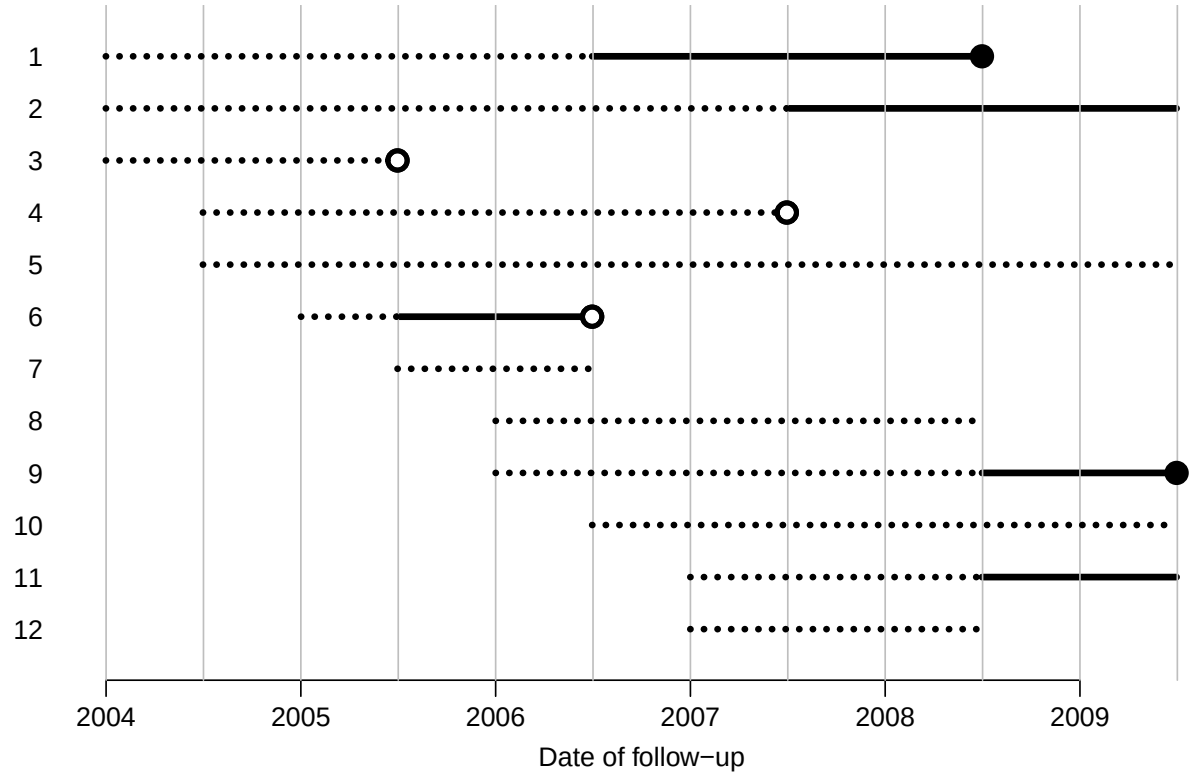
2. By how many percent points did the relative survival proportion improve in male patients of Norway during the 45 years since 1967-71?

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

2.1 Basic measures in a cohort

The figure below shows 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 (• if due to cancer C, ∘ for other causes) or censoring (end of line). Follow-up until the occurrence of cancer C is shown with a broken line. For those subjects contracting cancer C, follow-up after diagnosis is

shown with a solid line.



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

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

   (i) 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.

   (ii) The total person-time is 27 years. Assign this to variable `Y.todis` writing and running the following command line:
   ```
   > Y.todis <- 27
   ```

   (iii) What is the total number of new cases of cancer C? – Assign this number to variable `Cases` in the same way.

   (iv) Obtain the incidence rate of cancer C assigning its value into variable `Irate` and printing it as follows:
   ```
   > Irate <- 100*Cases/Y.todis
   > Irate
   ```

2. What is the mortality rate from cancer C during the same period? – Proceed with similar steps as above:

   (i)-(ii) 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.

   (iii) What is the total number of deaths from disease C? Assign this to variable `Dth.C`.

(iv) Assign the mortality rate from C into variable `Mrate.C` and print

3. 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.

4. What is the estimated 3-year mortality proportion ("risk" of death for a risk period of 3 years since entry) from all causes based on the result in the previous item 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?

5. What is the mortality rate `Mrate.pts` during the same period from all causes *among the patients with cancer* 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 corresponding prevalence proportions `P1` and `P2`. from these.

Why the incidence or mortality proportions for 3-year or any other risk period, calculated by the simple formula presented on slides 16 and 17, would be problematic in tasks 1 and 2?

Difficult: The follow-up of the cohort is an example of a *multistate* set-up where a person can be in each of 4 possible states: "Alive and well", "Alive with cancer", "Dead from cancer" and "Dead from other causes".

1. Draw four boxes, one for each state, and indicate with arrows the possible transitions between them.

2. Indicate for each arrow how many transitions there were in the cohort.

3. Indicate in the boxes, how many person-years was lived in each box.

4. Identify the calculation of rates in this diagram.

2.2 Population life table

Consider the lifetable for the Danish population for the years 1991–95, in table 2.1.

The survival function in the table can be thought of as number of a hypothetical cohort of 100,000 persons starting at age 0, that will still be alive by age a.

1. Calculate the probability that a 40 year old man reaches age 70 / 80 / 90, respectively.

The **M**edian **R**esidual **L**ifetime is the time which half of the (currently living part of the population) will survive and the other half not.

2. Find the MRL for men and women aged 40, respectively.

Table 2.1: *Life table for the Danish population for the period 1991–95.(From: Befolkningens bevægelser 1998, Danmarks Statistik, 2000). $S(a)$: The survival function ($\times 100,000$); $p(a)$: Death probability ($\times 100,000$); $R(a)$: Expected residual life time.*

	Men			Women				Men			Women		
Age	$S(a)$	$p(a)$	$R(a)$	$S(a)$	$p(a)$	$R(a)$	Age	$S(a)$	$p(a)$	$R(a)$	$S(a)$	$p(a)$	$R(a)$
0	100,000	712	72.53	100,000	541	77.84	50	92,470	575	25.72	95,542	400	29.92
1	99,288	59	72.05	99,459	52	77.27	51	91,938	606	24.86	95,159	434	29.03
2	99,230	33	71.09	99,407	32	76.31	52	91,381	642	24.01	94,746	464	28.16
3	99,197	30	70.11	99,375	22	75.33	53	90,795	728	23.16	94,306	506	27.29
4	99,168	26	69.14	99,353	19	74.35	54	90,133	829	22.33	93,829	561	26.42
5	99,142	22	68.15	99,335	15	73.36	55	89,386	909	21.51	93,302	618	25.57
6	99,121	20	67.17	99,319	14	72.37	56	88,573	991	20.70	92,726	683	24.73
7	99,101	23	66.18	99,305	14	71.38	57	87,696	1,136	19.91	92,093	765	23.89
8	99,079	25	65.20	99,291	15	70.39	58	86,700	1,315	19.13	91,388	841	23.07
9	99,055	20	64.21	99,276	14	69.40	59	85,560	1,431	18.38	90,619	940	22.26
10	99,035	18	63.22	99,263	11	68.41	60	84,335	1,595	17.64	89,767	1,052	21.47
11	99,017	17	62.24	99,252	13	67.42	61	82,990	1,804	16.92	88,823	1,132	20.69
12	99,001	20	61.25	99,239	14	66.43	62	81,493	1,924	16.22	87,817	1,215	19.93
13	98,981	24	60.26	99,225	14	65.44	63	79,925	2,070	15.53	86,750	1,326	19.16
14	98,957	26	59.27	99,211	17	64.45	64	78,271	2,290	14.84	85,600	1,461	18.42
15	98,931	36	58.29	99,195	19	63.46	65	76,478	2,494	14.18	84,349	1,596	17.68
16	98,896	49	57.31	99,175	21	62.47	66	74,571	2,780	13.53	83,003	1,711	16.96
17	98,847	61	56.34	99,154	23	61.48	67	72,498	3,045	12.90	81,583	1,848	16.25
18	98,787	76	55.37	99,132	32	60.50	68	70,290	3,336	12.29	80,075	2,015	15.54
19	98,711	95	54.41	99,100	41	59.52	69	67,945	3,752	11.70	78,462	2,187	14.85
20	98,618	93	53.46	99,059	36	58.54	70	65,396	4,058	11.13	76,746	2,361	14.17
21	98,526	87	52.51	99,023	32	57.56	71	62,742	4,420	10.58	74,934	2,621	13.50
22	98,441	90	51.56	98,991	35	56.58	72	59,969	4,864	10.05	72,970	2,873	12.85
23	98,352	87	50.60	98,957	33	55.60	73	57,052	5,291	9.54	70,874	3,078	12.22
24	98,266	91	49.65	98,924	30	54.62	74	54,033	5,778	9.04	68,692	3,316	11.59
25	98,177	102	48.69	98,894	35	53.64	75	50,911	6,271	8.57	66,415	3,676	10.97
26	98,076	106	47.74	98,860	41	52.65	76	47,718	6,783	8.11	63,973	4,074	10.37
27	97,972	105	46.79	98,820	40	51.67	77	44,481	7,346	7.66	61,367	4,370	9.79
28	97,869	112	45.84	98,780	42	50.70	78	41,214	8,030	7.23	58,685	4,818	9.20
29	97,759	119	44.89	98,738	48	49.72	79	37,904	8,710	6.82	55,858	5,365	8.66
30	97,643	125	43.94	98,690	52	48.74	80	34,603	9,471	6.42	52,861	5,925	8.12
31	97,522	134	43.00	98,639	60	47.77	81	31,326	10,389	6.04	49,729	6,610	7.60
32	97,391	150	42.06	98,580	65	46.79	82	28,071	11,293	5.68	46,442	7,451	7.10
33	97,245	159	41.12	98,516	61	45.82	83	24,901	12,149	5.34	42,982	8,337	6.63
34	97,090	158	40.18	98,456	72	44.85	84	21,876	13,043	5.01	39,398	9,230	6.19
35	96,936	168	39.25	98,385	90	43.88	85	19,023	14,200	4.69	35,762	10,137	5.77
36	96,773	187	38.31	98,297	105	42.92	86	16,321	15,642	4.38	32,137	11,407	5.36
37	96,592	210	37.38	98,194	118	41.97	87	13,768	17,076	4.10	28,471	12,688	4.99
38	96,390	228	36.46	98,078	119	41.02	88	11,417	18,402	3.84	24,858	13,835	4.64
39	96,170	251	35.54	97,961	131	40.06	89	9,316	20,246	3.59	21,419	15,391	4.30
40	95,928	283	34.63	97,833	157	39.12	90	7,430	21,659	3.37	18,123	16,864	4.00
41	95,657	296	33.73	97,680	164	38.18	91	5,821	22,775	3.17	15,066	18,541	3.71
42	95,374	293	32.83	97,520	176	37.24	92	4,495	24,923	2.96	12,273	20,439	3.44
43	95,094	304	31.92	97,348	201	36.30	93	3,375	26,578	2.77	9,765	22,521	3.19
44	94,806	323	31.02	97,153	211	35.38	94	2,478	28,725	2.59	7,565	24,601	2.97
45	94,500	347	30.12	96,948	231	34.45	95	1,766	30,641	2.44	5,704	26,453	2.78
46	94,171	383	29.22	96,724	264	33.53	96	1,225	33,252	2.30	4,195	28,752	2.60
47	93,810	431	28.33	96,468	293	32.61	97	818	34,446	2.19	2.989	30,269	2.44
48	93,406	478	27.45	96,186	316	31.71	98	536	33,589	2.08	2,084	31,732	2.29
49	92,959	527	26.58	95,882	355	30.81	99	356	37,944	1.88	1.423	35,125	2.12

2.3 Incidence and mortality – acute leukaemia

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 of and mortality from 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

1. Calculate the incidence rates of acute leukaemia in this population for the two periods.
2. Calculate similarly the mortality rates of leukaemia.
3. Is there evidence about any change in the incidence and/or mortality between these two periods?
4. What would you conclude about the fatality of leukemia in children?

2.4 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 10000 years)
vitamin E supplementation	99	11.6
no supplementation	151	17.8

1. Calculate the person-years at risk in the two study groups separately.
2. Estimate the “relative risk” (using incidence rate ratio) and “excess risk” (using rate difference) for measuring the effect of daily supplementation with vitamin E on the risk prostate cancer.
3. Estimate either the attributable fraction or preventive fraction, whichever more appropriate, to describe the proportional impact of vitamin E supplementation.
4. Discuss the results. What can be concluded from these estimates?

2.5 Comparative measures – smokers vs. non-smokers

In the table below you see the mortality rates (per 1000 person-years, age-adjusted) from three important causes of death among life-long non-smokers and regular smokers as 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

1. Calculate for each cause of death the following effect measures for comparison between smokers and non-smokers:

   (a) "excess risk", *i.e.* rate difference,

   (b) "relative risk", *i.e.* rate ratio,

   (c) attributable fraction.

2. 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 Infant mortality

During 1978 in Finland 269 boys died at the age of <1 year. The size of this male age group was 33,200 on 31 Dec 1977, and on 31 Dec 1978 it was 32,500. The number of boys born alive during 1978 was 32,800.

1. Calculate the mortality rate (per 1000 person-years) in this age group of boys in the year 1978 by the usual method.

2. In national vital statistics the *infant mortality rate* (IMR) is commonly computed as:

$$\text{IMR} = \frac{\text{no. of deaths in age group} < 1 \text{ year during a calendar year}}{\text{no. of live born children during the year}} \times 1000$$

   Calculate the value of this measure for Finnish boys in 1978 from the given data and compare it with the result in item 1.

3. Is the "infant mortality rate" in item 2 indeed a rate as defined in the lectures — why or why not? Is it a proportion?

2.7 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		

Calculate the following summary measures:

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

2. age-standardized rates and their ratio using the male population as the standard,

3. age-standardized rates and their ratio using the World Standard Population,

4. cumulative rates up to 75 years and their ratio,

5. cumulative risks up to 75 years and their ratio.

Compare and comment the results obtained in items 1 to 3.

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:

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

2.8 Standardized rates

Below is the number of cases (D) and the age-specific incidence rates (in cases per 100,000 person-years) from the Danish Cancer Register for the period 1983–87 for colon cancer, rectum cancer and lung cancer, by sex.

	Colon				Rectum				Lung			
	Men		Women		Men		Women		Men		Women	
Age	D	Rate	D	Rate	D	Rate	D	Rate	D	Rate	D	Rate
0- 4	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
5- 9	2	0.25	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
10-14	0	0.00	1	0.11	0	0.00	0	0.00	0	0.00	0	0.00
15-19	3	0.30	7	0.73	1	0.10	0	0.00	1	0.10	0	0.00
20-24	4	0.39	8	0.82	1	0.10	1	0.10	8	0.78	4	0.41
25-29	13	1.36	5	0.55	2	0.21	3	0.33	4	0.42	1	0.11
30-34	18	1.89	27	2.96	11	1.15	4	0.44	7	0.73	14	1.53
35-39	50	4.81	38	3.83	19	1.83	26	2.62	46	4.43	35	3.52
40-44	51	5.42	75	8.29	43	4.57	29	3.21	116	12.32	109	12.05
45-49	94	12.68	124	16.92	81	10.92	75	10.24	262	35.33	209	28.52
50-54	173	26.23	231	34.36	157	23.81	104	15.47	592	89.76	421	62.62
55-59	316	49.31	338	50.22	273	42.60	193	28.67	1089	169.95	650	96.57
60-64	492	78.05	511	73.67	402	63.77	251	36.19	1884	298.86	795	114.62
65-69	737	134.35	695	109.04	533	97.16	369	57.89	2206	402.13	843	132.26
70-74	870	189.61	1006	171.59	601	130.99	430	73.34	2308	503.02	773	131.85
75-79	853	267.27	1081	225.24	539	168.88	427	88.97	1824	571.51	621	129.39
80-84	602	342.50	903	281.20	312	177.51	318	99.03	891	506.93	336	104.63
85-89	279	359.19	522	316.19	180	231.73	184	111.45	305	392.66	135	81.77
90+	95	347.54	174	263.40	67	245.11	79	119.59	62	226.82	40	60.55

The effective population size in the period is 2,521,177 men and 2,596,061 women.

The data are available as the file `std-rates.txt` in the course folder; you can read it into R using:

```
> std <- read.table( "std-rates.txt", header=T )
```

1. How many person-years was accumulated by the Danish men aged 70–79 in the period 1983–87 ?
2. Calculate the crude rates for each sex and site.
3. Calculate the cumulative rates to ages 65, 70, 75 and 80.
4. Calculate the standardized rates, standardized to the world standard population:

Age	Weight (×1000)	Age	Weight (×1000)	Age	Weight (×1000)
0– 4	120	35–39	60	70–74	20
5– 9	100	40–44	60	75–79	10
10–14	90	45–49	60	80–84	5
15–19	90	50–54	50	85–89	3
20–24	80	55–59	40	90+	2
25–29	80	60–64	40		
30–34	60	65–69	30		

5. Calculate the male-female ratios of the crude, the standardized and the cumulative rates. Why are they not the same?
6. Calculate the age-specific male-female rate-ratios. Comment on the results.

2.9 Survival: cancer of the tongue

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

1. Complete this table by appropriate figures using the actuarial life table method.
2. 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.

2.10 Conditional survival

For Danish patients diagnosed with cancer of colon and rectum in the period 1978–87 we found the following probabilities of death (in %):

Year from	Colon		Rectum	
diagnosis	Men	Women	Men	Women
1st	43.44	42.13	36.60	34.29
2nd	22.80	19.11	24.00	21.86
3rd	16.74	14.60	21.02	15.67
4th	13.84	10.62	15.59	13.54
5th	11.00	8.69	14.55	11.40
6th	10.13	7.36	9.95	11.17
7th	8.67	5.65	11.37	8.99
8th	7.97	5.51	8.69	8.55
9th	7.42	5.37	10.07	8.14
10th	7.75	5.94	5.16	7.26
11th	4.91	5.66	7.14	2.57
12th	6.72	5.42	6.06	5.63
13th	6.20	6.25	5.00	2.13

1. Calculate for each of the groups the cumulative probability of surviving 1, 3, and 5 years respectively.

2. Calculate the *conditional* probabilities of surviving 3 and 5 years after diagnosis *given* that a Danish patient already has survived 1 year.

From Young, Ries & Pollack: "Cancer Patient Survival Among Ethnic Groups in the United States", JNCI, vol 73, pp. 341–52, we find that for white anglosaxons the cumulative survival probabilities for colon and rectum cancer patients diagnosed 1973–79 in the SEER areas are (in %):

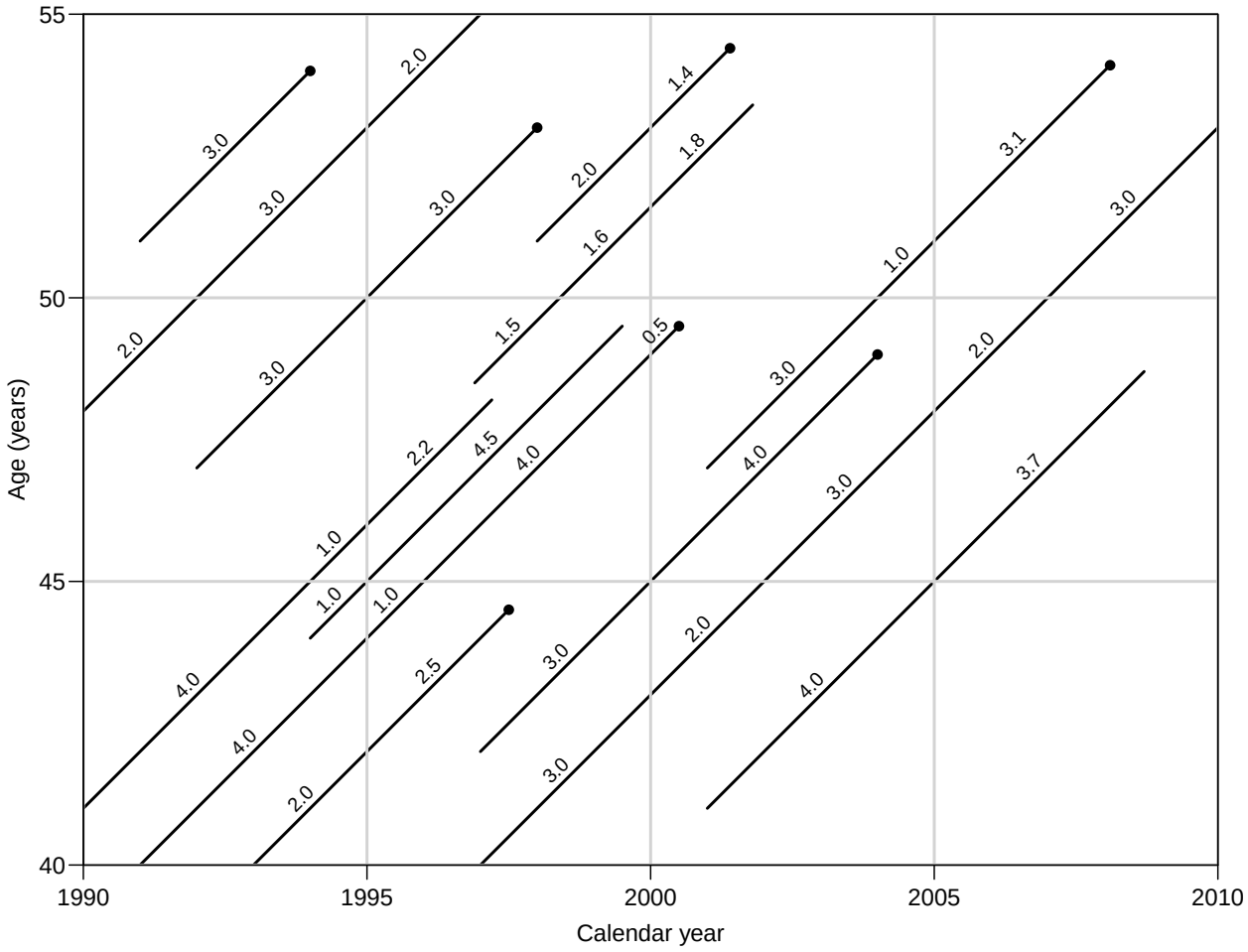
Years from	Colon		Rectum	
diagnosis	Men	Women	Men	Women
1	68	69	74	74
3	46	48	48	50
5	36	39	35	39

3. Calculate the *conditional* probabilities of surviving 3 and 5 years after diagnosis *given* that a U.S. patient already has survived 1 year.

4. Compare the cumulative survival probabilities and the conditional survival probabilities given survival of the first year between Denmark and USA.

2.11 Lexis diagram

In the Lexis diagram below displayed follow-up times of a small occupational cohort over the years 1990-2009 and the age range 40-54 years (this example is modified from a similar one in **B&D**). Each line runs from the entry to follow-up until either the diagnosis of cancer (•), or censoring or withdrawal (no symbol) due to death from other causes or migration.



1. Calculate the numbers of new cases of cancer, and person-years at risk in all the three 5-year agebands: 40-44, 45-49, and 50-54 years for each of the 5-year calendar periods 1990-94, 1995-99, and 2000-04 separately.

   Hint 1: Execute some division of labour in your group, so that not everybody is calculating these items for all periods.

   Hint 2: The data set is available as an example dataset, `occup`, in the `Epi` package. Try:

```
> library( Epi )
> ### data( occup )
> occup <- read.table("http://BendixCarstensen.com/NSCE/R/occup.txt", header=TRUE)
> str( occup )
> occup
> ### example( occup )
```

2. 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 1952-61.

3. Continuing from the previous item, estimate the cumulative rate and the cumulative risk over the whole 15-year age range for the chosen birth cohort.

NB. Estimation of the cumulative risk by the simple formula, presented on lecture slide 63, in which the competing risk of death is ignored, is not so problematic here, because of the relatively young age range covered, in which the mortality is expected to be quite low.

4. The age-specific incidences (per 100,000 person-years) in the three 5-year age-groups during 1990–2010 in the whole population of the country were 100, 200, and 400, respectively, so there was no variation between the subperiods. 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 standardised incidence ratio, SIR.

   Comment on the result.

2.12 Cumulative rates

In the period 1935–47 a large number of persons undergoing cerebral angiography were injected with Thorotrast, a contrast medium containing radioactive Thorium. In order to assess the elevation of the mortality related to the injection of Thorotrast, a control group of patients was selected who had also undergone cerebral angiography on similar indications in the period 1946–63, but with another contrast medium.

Below is a table of deaths and person-years at risk for the two groups, by current age.

	Thorotrast		Controls	
Current age	No. Deaths	Person-years	No. deaths	Person-years
0–19	5	572.1	11	1536.1
20–29	17	1974.2	16	2449.1
30–39	58	3489.0	35	4228.8
40–49	100	4502.2	67	5822.3
50–59	184	4433.5	137	6647.0
60–69	205	2998.1	211	5780.3
70–79	137	1134.4	206	3113.6
80+	45	261.5	114	939.8
Total	751	19365.4	797	30517.6

Calculate the following three things:

1. The estimates of the overall rates in each of the two groups and the rate ratio.

2. A confidence interval for the rate-ratio between the two groups.

3. The cumulative rates to 70 and 80 years in the two groups.

4. The ratio of the cumulative rates.

5. Comment on the results.

2.13 Attributable risk

Consider again the Thorotrast-study material from exercise 2.12 Remember the definition and interpretation of Attributable risk from the lectures.

1. Calculate the attributable risk of Thorotrast exposure on death of patients undergoing cerebral angiography:

   (a) Based on the crude relative risk.

   (b) Based on the relative risk from the cumulative rates to age 70.

   (c) Based on the relative risk from the cumulative rates to age 80.

   Comment on the differences, and calculate the number of deaths attributable to Thorotrast in the three cases.

2. Calculate the attributable risk in each age-group.

3. Calculate the number of deaths attributable to Thorotrast in each group, and compare the sum to the previous results.

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. Alpha-tocopherol had no apparent effect on total mortality, although more deaths from hemorrhagic stroke were observed among the

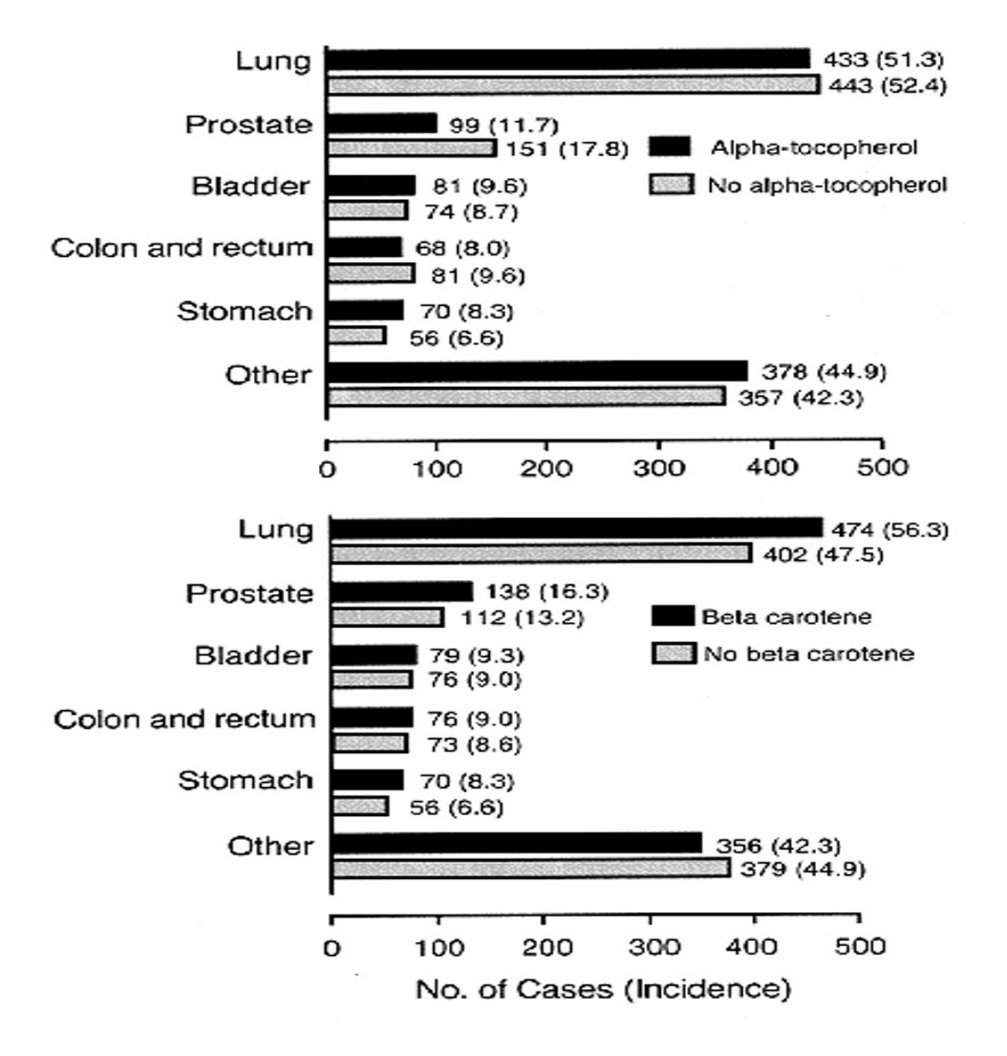


Figure 3.1: *Number and Incidence (per 10 000 Person-Years) of Cancers, According to Site, among Participants Who Received Alpha-Tocopherol Supplements and Those Who Did Not (Upper Panel) and among Participants Who Received Beta Carotene Supplements and Those Who Did Not (Lower Panel).*

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 18 314 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 25 000 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

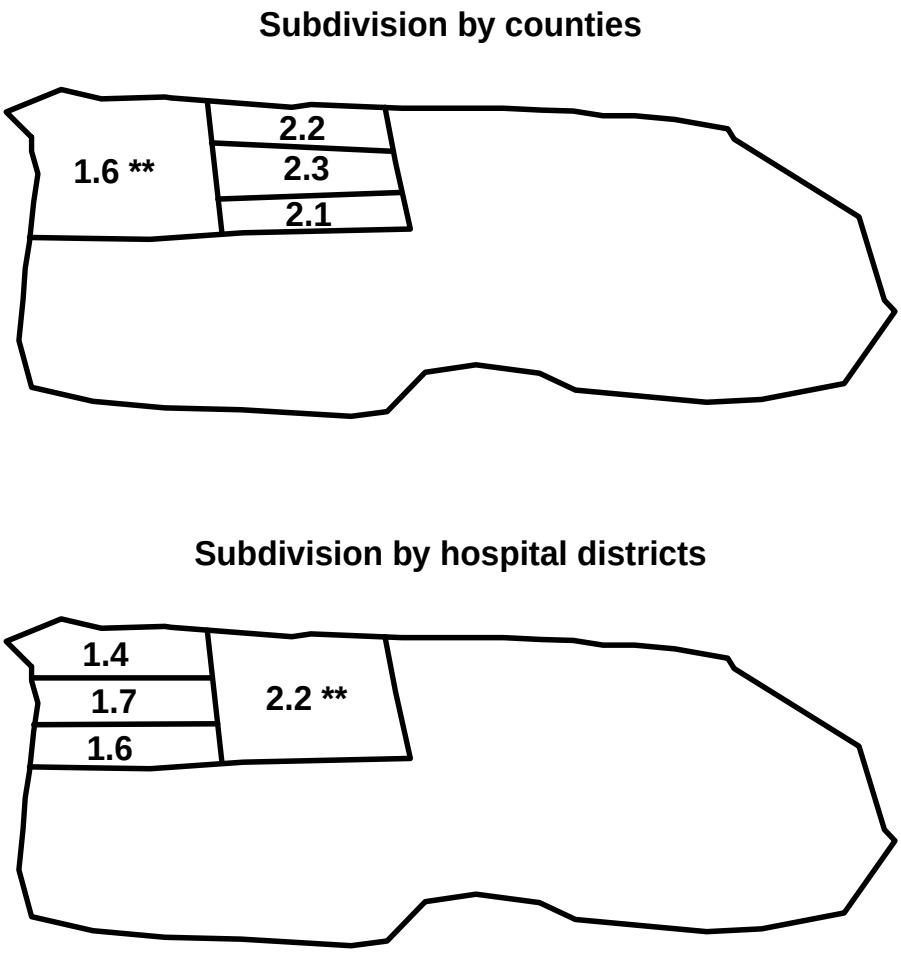


Figure 3.2: *Geographical division by county (top) and hospital district (bottom).*

person-years. The two divisions seem to give somewhat contradictory results. How can we explain this apparent paradox?

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.

Table 3.1: *Table of cases and controls by sex and PA (physical activity) index*

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 summary estimate of the rate ratio (and its 95 % confidence interval), using `glm` with binomial error as indicated in the lecture slides.
6. Compare this with the crude one.
7. Is there effect-modification by sex?
8. How would you report this?

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. Age-adjused summary estimate of the rate ratio, using `glm` with Poisson error as indicated in the lectures.

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?

7. Is there effect modification by age?

8. 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 summary of relations between various quantities used in analysis of follow-up studies. They are ubiquitous 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:

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

where T is the variable "time of death"

Conditional survival function:

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

Cumulative distribution function of death times (cumulative risk):

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

Density function of death times:

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

Intensity:

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

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

Note that f and λ are *scaled* quantities, they have dimension time^{-1}.

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 (rate), it is dimensionless, despite its name.

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

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

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

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 contribution from follow up of one person:

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

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

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):

$$\ell(\lambda) = \log\big(\mathrm{P}\{d \text{ events in } y \text{ follow-up time}\}\big) = d\log(\lambda) - \lambda y$$

This is under the assumption that the 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 ($Y = \sum_i y_i$), and D is the total number of failures ($D = \sum_i d_i$), where the sums are over individuals' contributions with the *same* rate, λ, for example from the same age-class fro all individuals.

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 (D, Y) 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 $\log(\lambda)$ 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 are more than one, say 3, causes of death, occurring with (cause-specific) rates λ_1, λ_2, λ_3, that is:

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

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 all 3 causes of death. The probability of dying from cause 1 before age a (the cause-specific cumulative risk) is:

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

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 and 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, $\tilde{\lambda}_i(a)$. 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:

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

Here, $\tilde{\lambda}_1$ is called the subdistribution hazard; as a function of $F_1(a)$ it depends on the survival function S, which depends on *all* the cause-specific hazards:

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

The subdistribution hazard is merely a transformation of the cause-specific cumulative risk. Namely the same transformation which in the single-cause case transforms the cumulative risk to the hazard. It is a mathematical construct that is not interpretable as a hazard despite its name.

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 lifetime (age at death).

The relation between the density f and the survival function S is $f(a) = -S'(a)$, 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 decision required in order to devise 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:

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

and hence:

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

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

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

These quantities have the property that their sum is the total years of life lost due to the disease:

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

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

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

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

Chapter 5

Measures of Disease Occurrence — Solutions

5.1 Basic measures in a cohort

1. We may obtain the total person-time Y as the sum of individual person-times (in years) since **entry** until **exit**: onset of disease, or death or censoring. Note also that we count follow-up maximally up to the end of 2008. Below is shown how the calculation is done:

```
> options( width=110 )
> Y.todis <- 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.todis
```

```
[1] 27
```

The number of cases of cancer C is $D = 5$. Thus, the incidence rate is $I = D/Y = 5/27$ y. It is computed as follows, expressed as cases per 100 years:

```
> Cases <- 5
> Irate <- 100*Cases/Y.todis
> round(Irate, 1)
```

```
[1] 18.5
```

2. The follow-up continued after onset of cancer for the 5 affected subjects; thus the total amount of person-years (until death, censoring or 31.12.2008) will be:

```
> Y.todth <- Y.todis + 2 + 1.5 + 1 + 0.5 + 0.5
> Y.todth
```

```
[1] 32.5
```

The number of deaths from cancer C was 1 (recall, that we still only do follow-up till 31.12.2008), so the mortality rate (per 100 years) from C in the cohort is:

```
> Dth.C <- 1
> Mrate.C <- 100*Dth.C/Y.todth
> round(Mrate.C, 1)
```

```
[1] 3.1
```

3. If we look at the total mortality, the total person-time is same as above, but now the number of cases is 4. Hence, the mortality rate is found from

```
> Dth.all <- 4
> Mrate.all <- 100*Dth.all/Y.todth
> round( Mrate.all, 1 )
```

```
[1] 12.3
```

the unit for the rate again being per 100 person-years. The 3-year mortality proportion (%) is obtained from

```
> Mprop3.all <- 1 - exp( - (Mrate.all/100)*3 )
> round(100*Mprop3.all, 1)
```

```
[1] 30.9
```

Division by 100 in the formula is necessary, because the mortality rate was expressed as per 100 years but the length of the 3-year risk period was expressed in years.

4. The person-years among the 5 cancer patients (recall, still only up to 31.12.2008):

```
> Y.distodth <- Y.todth - Y.todis
> Y.distodth
```

```
[1] 5.5
```

There were 2 deaths among these patients, so the mortality rate (/100 years) for them is:

```
> D.pts <- 2
> Mrate.pts <- 100*D.pts / Y.distodth
> round( Mrate.pts, 1 )
```

```
[1] 36.4
```

and hence the 3-year mortality proportion (i.e. the predicted fraction dead after three years, or 3-year cumulative risk) in percent is estimated as

```
> Mprop3.pts <- 1 - exp( -(Mrate.pts/100)* 3 )
> round( 100*Mprop3.pts, 1 )
```

```
[1] 66.4
```

5. The prevalence of cancer on 30 September 2006 was $1/7 = 14\%$ and on 31 December 2008 it was $3/5 = 60\%$, obtained as follows:

```
> N1 <- 7 ; N2 <- 5
> D1 <- 1 ; D2 <- 3
> P1 <- 100*D1/N1; P2 <- 100*D2/N2
> round( c(P1, P2), 1)
```

```
[1] 14.3 60.0
```

or if we want to add nice labels:

```
> Prev <- c(D1, D2)/c(N1, N2)
> names(Prev) <- c("30sep2006","31dec2008")
> round( 100*Prev, 1 )
```

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

The simple formula for computing the incidence proportion or the mortality proportion for a given cause of death ignores the incidence of competing events, for instance death before getting cancer when assessing the incidence proportion of cancer, and death from other causes when considering the cause-specific mortality proportion. When, however, the total mortality is estimated, there are no competing events.

5.1.1 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.
```

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:

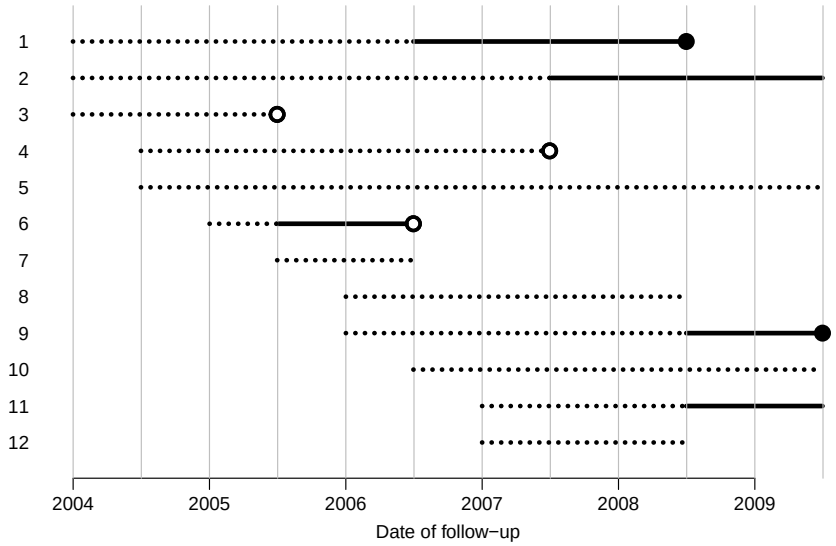
```
> cL <- cutLexis( cL,
+                 cut = cal.yr(cL$ddx),
+           new.state = "Cancer",
+    precursor.states = "Well" )
> summary( cL )
```

[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`.

```
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),]
 lex.id     per lex.dur lex.Cst  lex.Xst        doe        dox        ddx      xst id
      1 2004.00     2.5    Well   Cancer 2004-01-01 2008-07-01 2006-07-01 Dead-Ca  1
      1 2006.50     2.0  Cancer  Dead-Ca 2004-01-01 2008-07-01 2006-07-01 Dead-Ca  1
      2 2004.00     3.5    Well   Cancer 2004-01-01 2009-07-01 2007-07-01     Well  2
      2 2007.49     2.0  Cancer   Cancer 2004-01-01 2009-07-01 2007-07-01     Well  2
      3 2004.00     1.5    Well Dead-Oth 2004-01-01 2005-07-01       <NA> Dead-Oth  3
      4 2004.50     3.0    Well Dead-Oth 2004-07-01 2007-07-01       <NA> Dead-Oth  4
      5 2004.50     5.0    Well     Well 2004-07-01 2009-07-01       <NA>     Well  5
      6 2005.00     0.5    Well   Cancer 2005-01-01 2006-07-01 2005-07-01 Dead-Oth  6
      6 2005.50     1.0  Cancer Dead-Oth 2005-01-01 2006-07-01 2005-07-01 Dead-Oth  6
      7 2005.50     1.0    Well     Well 2005-07-01 2006-07-01       <NA>     Well  7
      8 2006.00     2.5    Well     Well 2006-01-01 2008-07-01       <NA>     Well  8
      9 2006.00     2.5    Well   Cancer 2006-01-01 2009-07-01 2008-07-01 Dead-Ca  9
      9 2008.50     1.0  Cancer  Dead-Ca 2006-01-01 2009-07-01 2008-07-01 Dead-Ca  9
     10 2006.50     3.0    Well     Well 2006-07-01 2009-07-01       <NA>     Well 10
     11 2007.00     1.5    Well   Cancer 2007-01-01 2009-07-01 2008-07-01     Well 11
     11 2008.50     1.0  Cancer   Cancer 2007-01-01 2009-07-01 2008-07-01     Well 11
     12 2007.00     1.5    Well     Well 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 5.1: *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 `state` *i.e.* the state in which the follow-up takes place and `lex.Xst` is the eXit `state`, *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 2004.00    2.50    Well   Cancer 2004-01-01 2008-07-01 2006-07-01
      1 2006.50    2.00  Cancer  Dead-Ca 2004-01-01 2008-07-01 2006-07-01
      2 2004.00    3.50    Well   Cancer 2004-01-01 2009-07-01 2007-07-01
      2 2007.49    1.51  Cancer   Cancer 2004-01-01 2009-07-01 2007-07-01
      2 2009.00    0.50  Cancer   Cancer 2004-01-01 2009-07-01 2007-07-01
      3 2004.00    1.50    Well Dead-Oth 2004-01-01 2005-07-01       <NA>
```

```
 4 2004.50    3.00    Well Dead-Oth 2004-07-01 2007-07-01       <NA>
 5 2004.50    4.50    Well     Well 2004-07-01 2009-07-01       <NA>
 5 2009.00    0.50    Well     Well 2004-07-01 2009-07-01       <NA>
 6 2005.00    0.50    Well   Cancer 2005-01-01 2006-07-01 2005-07-01
 6 2005.50    1.00  Cancer Dead-Oth 2005-01-01 2006-07-01 2005-07-01
 7 2005.50    1.00    Well     Well 2005-07-01 2006-07-01       <NA>
 8 2006.00    2.50    Well     Well 2006-01-01 2008-07-01       <NA>
 9 2006.00    2.50    Well   Cancer 2006-01-01 2009-07-01 2008-07-01
 9 2008.50    0.50  Cancer   Cancer 2006-01-01 2009-07-01 2008-07-01
 9 2009.00    0.50  Cancer  Dead-Ca 2006-01-01 2009-07-01 2008-07-01
10 2006.50    2.50    Well     Well 2006-07-01 2009-07-01       <NA>
10 2009.00    0.50    Well     Well 2006-07-01 2009-07-01       <NA>
11 2007.00    1.50    Well   Cancer 2007-01-01 2009-07-01 2008-07-01
11 2008.50    0.50  Cancer   Cancer 2007-01-01 2009-07-01 2008-07-01
11 2009.00    0.50  Cancer   Cancer 2007-01-01 2009-07-01 2008-07-01
12 2007.00    1.50    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 5.1.

5.2 Population life table

1. The probability that a 40 year old man reaches age 70 is the conditional probability that a man reaches 70 given that he already has reached 40, and this is, using the survival function from the life table:

$$\frac{65,396}{95,928} = 0.6817$$

and for ages 80 and 90 we get:

$$\frac{34,603}{95,928} = 0.3607 \qquad \frac{7,430}{95,928} = 0.0775$$

Or, using R:

```
> num <- c(65396,34603,7430)
> den <- rep(95928,3)
> names( num ) <- c("70","80","90")
> round( num/den, 4 )
    70     80     90
0.6817 0.3607 0.0775
```

2. The median residual lifetime after 40 is the time until half of those alive at 40 have died. Out of a generation of 100,000 men in the table, 95,928 were alive at age 40. Thus we want to know when $95,928/2 = 47,964$ are left alive. This is seen to be somewhere between 75 and 76 years. This age-class has a death probability of 0.06271 ($p(75)$), i.e. for the mortality rate in the group, λ_{75} we have $1 - \exp(-\lambda_{75} \times 1 \text{ year}) = 0.06271$, i.e. $\lambda_{75} = -\ln(1 - 0.06271)/1 \text{ year} = 0.06476/\text{year}$. The time, ℓ, needed for 50,911 alive at 75 to be reduced to 47,964 is the solution to:

$$\exp(-0.06476/\text{year} \times \ell) = \frac{47,964}{50,911} \qquad \Leftrightarrow \qquad \ell = -\log\left(\frac{47,964}{50,911}\right) \Big/ 0.06476 = 0.92$$

```
> -log(47964/50911)/(-log(1-0.06271))
[1] 0.9207217
```

So at age 75.92 the remaining number of people is 47,964. Therefore the median residual lifetime for men at 40 is $75.92 - 40 = 35.92$ years.

For women we find that 97,833 are alive at 40, and that $97,833/2 = 48,916.5$ are left at some point between 81 and 82. As the death probability for this ageclass is 0.06610 the mortality rate is $-\ln(1-0.06610)/\text{ year} = 0.06839/\text{ year}$, and as 49,729 are alive at 81 we solve:

$$\exp(-0.06839/\text{year} \times \ell) = \frac{48,916.5}{49,729} \qquad \Leftrightarrow \qquad \ell = -\log\left(\frac{48,916.5}{49,729}\right) \Big/ 0.06839 = 0.24$$

Thus, the median residual lifetime for women aged 40 is $81.24 - 40 = 41.24$ years.

Thus more than half of the men reaching 40 have lived more than half of their life, whereas less than half of the women reaching 40 have.

5.3 Incidence and mortality – acute leukaemia

The population size (in 1000s) at the *end* of each 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 y) are (since `Y` is expressed in 1000s):

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

1993-97    1999
   4.56    5.38
```

2. The mortality rates (also per 100,000 y) 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 the incidence has come up, but the mortality has gone down. However, the evidence is way too thin to draw any conclusions. More formal assessment requires appropriate statistical analysis of the error margin of the contrast between the observed rates of the two periods. – See the materials of "Analysis of epidemiologic data".

4. As the observed **mortality-to-incidence ratios** (M/I ratio) of leukaemia in these periods, $22/113 = 0.19$ and $3/26 = 0.12$, are quite low, this suggests that the great majority of children contracting leukaemia would have survived. However, a more appropriate assessment of the question requires proper **survival analysis**, based on following-up those children with leukaemia over a sufficiently long period.

5.4 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 incidence rate I is computed from the no. cases, D, and person-years Y as $I = D/Y$, it follows that: $Y = D/I$. Accordingly, we compute the person-years in the two groups:

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

   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
```

2. The comparative measures:

   (i) "Relative risk": incidence rate ratio $= 11.6/17.8 = 0.652$,

```
> rate[1]/rate[2]
     VitE
0.6516854
```

   (ii) "Excess risk": rate difference $= 11.6 - 17.8 = -6.2$ per 10000 y,

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

VitE
-6.2
```

3. As the incidence among the exposed is lower, we compute the prevented fraction: from rates: PF $= (17.8 - 11.6)/17.8 = 0.348 = 35\%$,

```
> PF <- (rate["Plc"]-rate["VitE"])/rate["Plc"]
> 100*round(PF,3)

 Plc
34.8
```

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

5.5 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
rate difference (per 1000 y)	1.8	2.0	6.0
rate ratio	10.0	3.0	1.7
attributable 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:

```
> diff <- sm-ns
> ratio <- sm/ns
> AF <- (ratio - 1)/ratio
> round( rbind( diff, ratio, 100*AF ), 1 )
```

```
      Lung Ca Oth lung  CVD
diff      1.8      2.0  6.0
ratio    10.0      3.0  1.7
         90.0     66.7 40.0
```

2. The strongest biological effect is seen for lung cancer, as is apparent from the large values of the rate ratio and AF. However, as seen from the rate difference, *i.e.* the excess mortality rate, the population impact is by far the largest for CVD mortality.

5.6 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
> I <- 1000* D/Y
> round( c( D, Y, I ), 2)
```

```
[1]   269.00 32850.00     8.19
```

2. IMR = 269/32,800 = 8.20 **per 1000 liveborn**. – Note the unit!

```
> B <- 32800
> IMR <- 1000* D/B
> round( c(D, B, IMR), 2)
```

```
[1]   269.0 32800.0     8.2
```

3. IMR is not a rate, although in many populations the denominator is a close 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! It would be more appropriate to call this measure as **infant motality ratio**.

5.7 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 weights 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:

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

M cum.risk F cum.risk     M/F RR
    0.0155     0.0133     1.1618
```

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 comparing a single standardized rate may also be somewhat misleading, because it conceals the important information that the rate ratio varies by age.

5.8 Standardized rates

1. The rates (λ) are calculated using the number of cases (N)and the accumulated person-years (Y) in each age-category:

$$\lambda = \frac{N}{Y} \quad \Leftrightarrow \quad Y = \frac{N}{\lambda}$$

thus the person-years for men in ages 70–74 is

$$\frac{2308}{503.02} \times 100,000 = 458,829$$

and in ages 75–79:

$$\frac{1824}{571.51} \times 100,000 = 319,155$$

so the total amount of person-years in ages 70–79 is 777,983. This is accumulated over a 5-year period (1983–87), so the effective population size (average number of men aged 70–79) is $777,983/5 = 155,597$.

2. The crude, cumulative and rates standardized rates, and the corresponding male-female rate ratios are can be computed using R as a simple calculator.

   First read the data and see how they look:

```
> std <- read.table( "std-rates.txt", header=T )
> str( std )

'data.frame':        114 obs. of  5 variables:
 $ age : int  0 5 10 15 20 25 30 35 40 45 ...
 $ sex : chr  "M" "M" "M" "M" ...
 $ typ : chr  "Colon" "Colon" "Colon" "Colon" ...
 $ D   : int  0 2 0 3 4 13 18 50 51 94 ...
 $ rate: num  0 0.25 0 0.3 0.39 ...
```

   Since we have given that the number of male-person-years is 2,521,177 and the female is 2,596,061, we just need the total number of cases to compute the crude rates (using the subsetting rules):

```
> raw.colon.m  <- sum( subset( std, sex=="M" & typ=="Colon" )$D ) / 25.21177
> raw.rectum.f <- sum( subset( std, sex=="F" & typ=="Rectum" )$D ) / 25.96061
> c(raw.colon.m,raw.rectum.f)

[1] 184.5170  96.0301
```

   Similar computations are made for the other combinations of sex and type of cancer.

3. The cumulative rates are just the sum of the rates up to a given age, multiplied by the interval length:

```
> cum65.colon.m <- sum( subset( std, sex=="M" & typ=="Colon" & age<66 )[,"rate"] ) * 5
> cum65.colon.m

[1] 1575.2
```

   Similar computations are made for the other combinations of sex, type of cancer and age-limit.

4. The standardized rates are just the observed rates multiplied by the defined weights and then summed over all ages. Therefore we first need to enter a vector of weights for doing the standardizations (and make sure the weight sum to 1 by dividing with the sum)

```
> wt <- c(120,100,90,90,80,80,60,60,60,60,50,40,40,30,20,10,5,3,2)
> wt <- wt / sum(wt )
> wt
```

```
 [1] 0.120 0.100 0.090 0.090 0.080 0.080 0.060 0.060 0.060 0.060 0.050 0.040
[13] 0.040 0.030 0.020 0.010 0.005 0.003 0.002
```

Now we can compute the standardized rates (note that the multiplication with `wt` is *inside* the argument to `sum`:

```
> std.colon.m  <- sum( subset( std, sex=="M" & typ=="Colon" )$D*wt )
> std.rectum.f <- sum( subset( std, sex=="F" & typ=="Rectum" )$D*wt )
> c(std.colon.m,std.rectum.f)
```

```
[1] 107.657  57.560
```

In summary the results for the first parts of the exercise are:

	Colon			Rectum			Lung		
	Men	Women	ratio	Men	Women	ratio	Men	Women	ratio
per 100,000 person-years									
Crude rate	36.90	44.26	0.83	25.56	19.21	1.33	92.06	38.41	2.40
Standardized rate	22.07	20.62	1.07	15.88	9.55	1.66	58.52	23.15	2.53
%									
Cumulative rate to 65	0.90	0.96	0.94	0.75	0.49	1.53	3.06	1.60	1.91
Cumulative rate to 70	1.58	1.51	1.04	1.23	0.72	1.59	5.07	2.26	2.24
Cumulative rate to 75	2.52	2.37	1.07	1.89	1.06	1.65	7.59	2.92	2.60
Cumulative rate to 80	3.86	3.49	1.11	2.73	1.50	1.72	10.45	3.57	2.93

Computing details There are facilities in R for doing these analyses in one go: First we compute the crude rates; to that end we need the person-years in the population as a vector, and then the total number of cases by sex *and* site:

```
> Y <- c( 25.21177, 25.96061 )
> names( Y )  <- c("M","F")
> Y
```

```
       M        F
25.21177 25.96061
```

```
> D <- with( std, tapply( D, list(sex,typ), sum ) )
> D
```

```
  Colon  Lung Rectum
F  5746  4986   2493
M  4652 11605   3222
```

When we compute the rates we need to have the person-years in the right order, then we can divide the table by the person-years:

```
> Y <- Y[2:1]
> Y
```

```
       F        M
25.96061 25.21177
> round( D/Y, 1 )
  Colon  Lung Rectum
F 221.3 192.1   96.0
M 184.5 460.3  127.8
```

The cumulative rates are simply the integral of the rates up to a given age; so the size of the rate times the length of the interval to which it applies, in this case 5 years, and we must recall that the rates are given in cases per 100,000 PY. Also note that the age-classes are coded by their left endpoint; therefore when computing the rate until age 70, say, we use "<70":

```
> c65 <- with( subset(std,age<65), tapply( rate, list(sex,typ), sum )*5/10^5 )
> c70 <- with( subset(std,age<70), tapply( rate, list(sex,typ), sum )*5/10^5 )
> c75 <- with( subset(std,age<75), tapply( rate, list(sex,typ), sum )*5/10^5 )
> c80 <- with( subset(std,age<80), tapply( rate, list(sex,typ), sum )*5/10^5 )
> rbind( c65, c70, c75, c80 )
      Colon      Lung    Rectum
F 0.0096230 0.0159975 0.0048635
M 0.0090345 0.0306340 0.0074530
F 0.0150750 0.0226105 0.0077580
M 0.0157520 0.0507405 0.0123110
F 0.0236545 0.0292030 0.0114250
M 0.0252325 0.0758915 0.0188605
F 0.0349165 0.0356725 0.0158735
M 0.0385960 0.1044670 0.0273045
```

Finally, the standardized rates, standardized to the world standard population, is simply the weighted averages, obtained by multiplying the age-specific rates by the weight (previoulsy calculated):

```
> wst <- with( std, tapply( rate*wt, list(sex,typ), sum ) )
```

We can put all these rates together in an `array`, that is, a multidimensional table. First we define the dimensions in a list

```
> dnam <- list( sex=c("F","M","M/F"),
+               typ=c("colon","Lung","rectum"),
+           measure=c("crude","wst","cum65","cum70","cum75","cum80"))
> res <- array( NA, dim=c(3,3,6), dimnames=dnam )
```

This is now an array with only missing values in it, so we put in the values we just computed — not that we refor to the dimension by names, whcih reduces the possibility of errors:

```
> res[1:2,,"crude"] <- D/Y
> res[1:2,,"wst"]   <- wst
> res[1:2,,"cum65"] <- c65
> res[1:2,,"cum70"] <- c70
> res[1:2,,"cum75"] <- c75
> res[1:2,,"cum80"] <- c80
> ftable( res )
           measure       crude         wst       cum65       cum70       cum75       cum80
sex typ
F   colon          221.3353230  20.6155700   0.0096230   0.0150750   0.0236545   0.0349165
    Lung           192.0602020  23.1456600   0.0159975   0.0226105   0.0292030   0.0356725
    rectum          96.0301010   9.5547800   0.0048635   0.0077580   0.0114250   0.0158735
```

```
M   colon          184.5169935  22.0664500   0.0090345   0.0157520   0.0252325   0.0385960
    Lung           460.3008833  58.5196700   0.0306340   0.0507405   0.0758915   0.1044670
    rectum         127.7974533  15.8836600   0.0074530   0.0123110   0.0188605   0.0273045
M/F colon                   NA          NA          NA          NA          NA          NA
    Lung                    NA          NA          NA          NA          NA          NA
    rectum                  NA          NA          NA          NA          NA          NA
```

```
> ftable( res, row.vars=3 )
        sex           F                                   M                                   M/F
        typ       colon        Lung      rectum       colon        Lung      rectum       colon
measure
crude       221.3353230 192.0602020  96.0301010 184.5169935 460.3008833 127.7974533          NA
wst          20.6155700  23.1456600   9.5547800  22.0664500  58.5196700  15.8836600          NA
cum65         0.0096230   0.0159975   0.0048635   0.0090345   0.0306340   0.0074530          NA
cum70         0.0150750   0.0226105   0.0077580   0.0157520   0.0507405   0.0123110          NA
cum75         0.0236545   0.0292030   0.0114250   0.0252325   0.0758915   0.0188605          NA
cum80         0.0349165   0.0356725   0.0158735   0.0385960   0.1044670   0.0273045          NA
```

Note that the male-female ratio is empty. One advantage of putting it all in an array is that summary measures are easily computed for the entire array:

```
> res["M/F",,] <- res["M",,]/res["F",,]
> round( ftable( res, row.vars=3 ), 3 )
        sex       F                       M                     M/F
        typ   colon    Lung  rectum   colon    Lung  rectum   colon    Lung  rectum
measure
crude       221.335 192.060  96.030 184.517 460.301 127.797   0.834   2.397   1.331
wst          20.616  23.146   9.555  22.066  58.520  15.884   1.070   2.528   1.662
cum65         0.010   0.016   0.005   0.009   0.031   0.007   0.939   1.915   1.532
cum70         0.015   0.023   0.008   0.016   0.051   0.012   1.045   2.244   1.587
cum75         0.024   0.029   0.011   0.025   0.076   0.019   1.067   2.599   1.651
cum80         0.035   0.036   0.016   0.039   0.104   0.027   1.105   2.929   1.720
```

5. The rate-ratios between men and women vary by the measure they are based on. The ratios based on standardized rates or cumulative rates all assume that the incidence rate-ratio is the same throughout the age-span, which it obviously is not. The more this assumption is violated the larger the differences between the ratios based on the various measures.

   For example, the male-female rate-ratio based on cumulative rate to 65 for lung cancer is 1.91, and based on cumulative rates to 80 it is 2.93, 50% larger.

6. In order to get a bit more insight as to how the M/F rate-ratio varies by age we can simply compute these for each site and divide them:

   ```
   > round( with( subset(std,sex=="M"), tapply( rate, list(age,typ), sum ) ) /
   +        with( subset(std,sex=="F"), tapply( rate, list(age,typ), sum ) ), 2 )
   ```

   ```
      Colon Lung Rectum
   0    NaN  NaN    NaN
   5    Inf  NaN    NaN
   10  0.00  NaN    NaN
   15  0.41  Inf    Inf
   20  0.48 1.90   1.00
   25  2.47 3.82   0.64
   ```

```
30  0.64 0.48   2.61
35  1.26 1.26   0.70
40  0.65 1.02   1.42
45  0.75 1.24   1.07
50  0.76 1.43   1.54
55  0.98 1.76   1.49
60  1.06 2.61   1.76
65  1.23 3.04   1.68
70  1.11 3.82   1.79
75  1.19 4.42   1.90
80  1.22 4.84   1.79
85  1.14 4.80   2.08
90  1.32 3.75   2.05
```

It is seen that for all three sites there is an incresing tendency in the male-female rate-ratio, hence the increasing values of the ratios based on the cumulative rates.

5.9 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. In the survival curve, the y-values are the cumulative survival proportions given at the last column of the data frame. The corresponding x-values being the *end* points of the intervals, in this case 1,...,7 (years after diagnosis). However, we need to 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 5.2 we see that the lower quartile is 0.7 years, median 1.69 years but that the upper quartile is unestimable from these data.

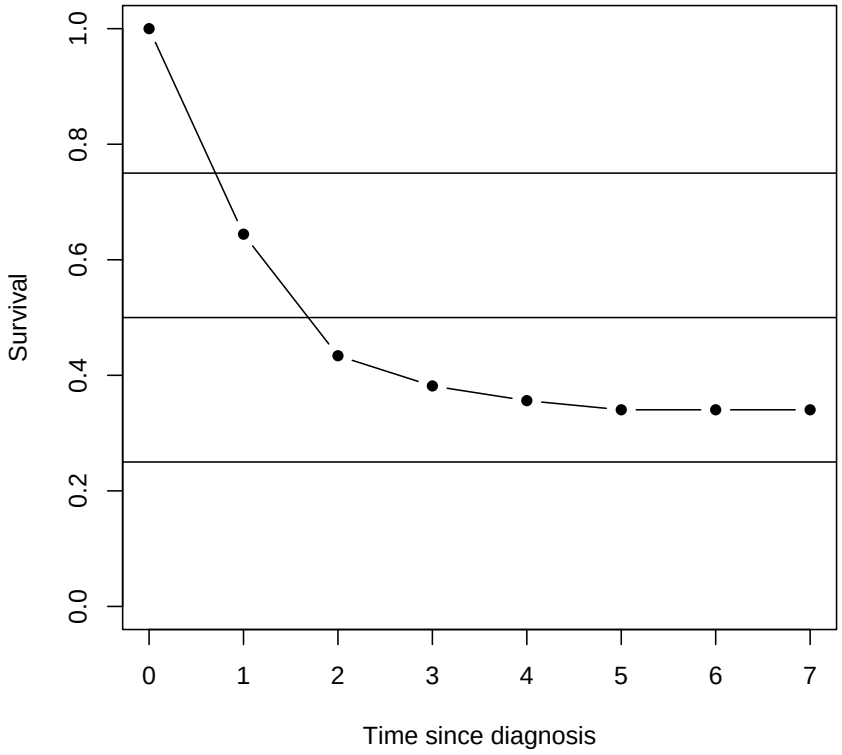


Figure 5.2: *The survival curve – lifetable estimator.*

5.10 Conditional survival

1. For men with colon cancer we find that the cumulative probability of surviving:

   1 year is $1 - 0.4343 = 0.5657$

   3 years is $0.5657 \times (1 - 0.2280) \times (1 - 0.1674) = 0.3636$

   5 years is $0.3636 \times (1 - 0.1384) \times (1 - 0.1100) = 0.2788$

   This can be automated by putting data into R. The actual numbers can be copy-pasted from the .pdf-file of the exercises, equipped with commas, and wrapped into a `matrix()` statement. This will give you the dataset inside the program for later further manipulation.

   Note that we give the data as one long vector (using `c()`) and then just specify the number of columns, and leave to R to find out how many rows. The default is to fill matrices column-wise, so we also need the `byrow=TRUE`:

```
> prob <- matrix( c(
+ 1,43.44,42.13,36.60,34.29,
+ 2,22.80,19.11,24.00,21.86,
+ 3,16.74,14.60,21.02,15.67,
+ 4,13.84,10.62,15.59,13.54,
+ 5,11.00,8.69,14.55,11.40,
```

```
+ 6,10.13,7.36,9.95,11.17,
+ 7,8.67,5.65,11.37,8.99,
+ 8,7.97,5.51,8.69,8.55,
+ 9,7.42,5.37,10.07,8.14,
+ 10,7.75,5.94,5.16,7.26,
+ 11,4.91,5.66,7.14,2.57,
+ 12,6.72,5.42,6.06,5.63,
+ 13,6.20,6.25,5.00,2.13), ncol=5, byrow=T )
> prob
      [,1]  [,2]  [,3]  [,4]  [,5]
 [1,]    1 43.44 42.13 36.60 34.29
 [2,]    2 22.80 19.11 24.00 21.86
 [3,]    3 16.74 14.60 21.02 15.67
 [4,]    4 13.84 10.62 15.59 13.54
 [5,]    5 11.00  8.69 14.55 11.40
 [6,]    6 10.13  7.36  9.95 11.17
 [7,]    7  8.67  5.65 11.37  8.99
 [8,]    8  7.97  5.51  8.69  8.55
 [9,]    9  7.42  5.37 10.07  8.14
[10,]   10  7.75  5.94  5.16  7.26
[11,]   11  4.91  5.66  7.14  2.57
[12,]   12  6.72  5.42  6.06  5.63
[13,]   13  6.20  6.25  5.00  2.13
```

We do not need the first column, so it is dropped by using indexing of columns (negative numbers mean "omit"):

```
> prob <- prob[,-1]
> colnames(prob) <- c("Col.M","Col.F","Rect.M","Rect.F")
> prob
      Col.M Col.F Rect.M Rect.F
 [1,] 43.44 42.13  36.60  34.29
 [2,] 22.80 19.11  24.00  21.86
 [3,] 16.74 14.60  21.02  15.67
 [4,] 13.84 10.62  15.59  13.54
 [5,] 11.00  8.69  14.55  11.40
 [6,] 10.13  7.36   9.95  11.17
 [7,]  8.67  5.65  11.37   8.99
 [8,]  7.97  5.51   8.69   8.55
 [9,]  7.42  5.37  10.07   8.14
[10,]  7.75  5.94   5.16   7.26
[11,]  4.91  5.66   7.14   2.57
[12,]  6.72  5.42   6.06   5.63
[13,]  6.20  6.25   5.00   2.13
```

Having the numbers in a matrix we can just multiply the rows together to get the 1, 3 and 5-year survival. First we must however divide all columns by 100 to get probabilities, and the compute the cumulative survival probabilities — note that we do calculations for both sites and both sexes simultaneously:

```
> prob <- prob/100
> surv1 <- 1-prob[1,]
> surv3 <- surv1 * (1-prob[2,]) * (1-prob[3,])
> surv5 <- surv3 * (1-prob[4,]) * (1-prob[5,])
> ( DK.surv <- rbind( surv1, surv3, surv5 ) )
```

```
          Col.M     Col.F    Rect.M    Rect.F
surv1 0.5656000 0.5787000 0.6340000 0.6571000
surv3 0.3635491 0.3997663 0.3805572 0.4329991
surv5 0.2787782 0.3262608 0.2744896 0.3316927
```

We can get the figures printed in percent, rounded to one decimal:

```
> round( DK.surv*100, 1 )
      Col.M Col.F Rect.M Rect.F
surv1  56.6  57.9   63.4   65.7
surv3  36.4  40.0   38.1   43.3
surv5  27.9  32.6   27.4   33.2
```

Note that these are considerably lower than the corresponding American figures.

2. The conditional probability of surviving 3 years given that a man with colon cancer already has survived one year is:

$$\begin{aligned} \Pr(\text{survive 2nd} \mid \text{alive after 1st}) &\times \Pr(\text{survive 3rd} \mid \text{alive after 2nd}) \\ &= (1 - 0.2280) \times (1 - 0.1674) \\ &= 0.6427 \end{aligned}$$

This figure can also be computed as:

$$\frac{\mathrm{P}\{\text{Survive 3 years}\}}{\mathrm{P}\{\text{Survive 1 year}\}} = \frac{0.3636}{0.5657} = 0.6427$$

Thus, the conditional survival probabilities can be obtained from the table by dividing the last two rows with the first, which goes like this in R:

```
> cond3 <- surv3/surv1
> cond5 <- surv5/surv1
> DK.cond <- rbind( cond3, cond5 )
> round( DK.cond*100, 1 )
      Col.M Col.F Rect.M Rect.F
cond3  64.3  69.1   60.0   65.9
cond5  49.3  56.4   43.3   50.5
```

There is of course a function in R that will do these calculations in one go and return the survival function for all times:

```
> round( (S.DK <- apply( 1-prob, 2, cumprod ) )*100, 1 )
      Col.M Col.F Rect.M Rect.F
 [1,]  56.6  57.9   63.4   65.7
 [2,]  43.7  46.8   48.2   51.3
 [3,]  36.4  40.0   38.1   43.3
 [4,]  31.3  35.7   32.1   37.4
 [5,]  27.9  32.6   27.4   33.2
 [6,]  25.1  30.2   24.7   29.5
 [7,]  22.9  28.5   21.9   26.8
 [8,]  21.1  26.9   20.0   24.5
 [9,]  19.5  25.5   18.0   22.5
[10,]  18.0  24.0   17.1   20.9
[11,]  17.1  22.6   15.8   20.4
[12,]  16.0  21.4   14.9   19.2
[13,]  15.0  20.1   14.1   18.8
```

Likewise, there is a function that will divide the entire array by the first line, so as to produce the condtional survival given one-year survival:

```
> sweep( as.array(S.DK), 2, S.DK[1,], "/" )
```

```
          Col.M     Col.F    Rect.M    Rect.F
 [1,] 1.0000000 1.0000000 1.0000000 1.0000000
 [2,] 0.7720000 0.8089000 0.7600000 0.7814000
 [3,] 0.6427672 0.6908006 0.6002480 0.6589546
 [4,] 0.5538082 0.6174376 0.5066693 0.5697322
 [5,] 0.4928893 0.5637823 0.4329489 0.5047827
 [6,] 0.4429596 0.5222879 0.3898705 0.4483985
 [7,] 0.4045550 0.4927786 0.3455422 0.4080874
 [8,] 0.3723120 0.4656265 0.3155146 0.3731960
 [9,] 0.3446864 0.4406224 0.2837423 0.3428178
[10,] 0.3179732 0.4144494 0.2691012 0.3179292
[11,] 0.3023608 0.3909916 0.2498874 0.3097585
[12,] 0.2820421 0.3697998 0.2347442 0.2923191
[13,] 0.2645555 0.3466873 0.2230070 0.2860927
```

3. From the American material we get the conditional survival probabilities in the same way. We just have to enter the data first:

```
> US.surv <- matrix( c(
+ 1, 68, 69, 74, 74,
+ 3, 46, 48, 48, 50,
+ 5, 36, 39, 35, 39), ncol=5, byrow=TRUE )
> US.surv <- US.surv[,-1]/100
> colnames( US.surv ) <- colnames( DK.surv )
> US.surv
```

```
     Col.M Col.F Rect.M Rect.F
[1,]  0.68  0.69   0.74   0.74
[2,]  0.46  0.48   0.48   0.50
[3,]  0.36  0.39   0.35   0.39
```

Then we can perform the same operation as before now on the American data:

```
> acond3 <- US.surv[2,]/US.surv[1,]
> acond5 <- US.surv[3,]/US.surv[1,]
> US.cond <- rbind(acond3,acond5)
> round( US.cond*100, 1 )
```

```
       Col.M Col.F Rect.M Rect.F
acond3  67.6  69.6   64.9   67.6
acond5  52.9  56.5   47.3   52.7
```

4. We see that there is a considerable difference in the cumulative survival probabilities between Denmark and USA, but conditioning on surviving one year almost removes the differences.

   A quick way of comparing the probabilities between USA and Denmark is to form the ratio of them:

```
> round( US.surv/DK.surv, 2 )
```

```
     Col.M Col.F Rect.M Rect.F
[1,]  1.20  1.19   1.17   1.13
[2,]  1.27  1.20   1.26   1.15
[3,]  1.29  1.20   1.28   1.18
```

```
> round( US.cond/DK.cond, 2 )
```

```
       Col.M Col.F Rect.M Rect.F
acond3  1.05  1.01   1.08   1.03
acond5  1.07  1.00   1.09   1.04
```

Thus most the differences in survival between Denmark and USA is during the first year, where Danish patients apparently have a higher mortality.

This is probably due to differences in registration procedures between Denmark and USA. Danish figures include persons that are initially reported to the registry via death certificates and subsequently traced and equipped with a proper date of diagnosis. These so-called DCI — Death Certificate Initiated — cases obviously have a rather short survival. Such persons are more likely not to be traced in the USA, and are therefore not included in the American figures, that in this way are somewhat optimistic.

5.11 Lexis diagram

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

	period 1990-94		period 1995-99		period 2000-04		1952-61 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	8.6	1	4.2	1	7.1

1. You can load the dataset from the course website by:

```
> library( Epi )
> occup <- read.table("http://BendixCarstensen.com/NSCE/R/occup.txt", header=TRUE)
> occup
```

```
    AoE    DoE    DoX Xst
1  51.0 1991.0 1994.0   D
2  48.0 1990.0 1997.0   X
3  47.0 1992.0 1998.0   D
4  51.0 1998.0 2001.4   D
5  48.5 1996.9 2001.8   W
6  41.0 1990.0 1997.2   W
7  44.0 1994.0 1999.5   W
8  40.0 1991.0 2000.5   D
9  40.0 1993.0 1997.5   D
10 47.0 2001.0 2008.1   D
11 42.0 1997.0 2004.0   D
12 40.0 1997.0 2010.0   X
13 41.0 2001.0 2008.7   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
```

Exit status `X` and `W` are synonymous. If we want to classify the follow-up (person-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(0,110,5)+1900 )
> oL[order(oL$lex.id,oL$age),]

 lex.id  age    per lex.dur lex.Cst lex.Xst  AoE    DoE    DoX Xst
      1 51.0 1991.0     3.0       W       D 51.0 1991.0 1994.0   D
      2 48.0 1990.0     2.0       W       W 48.0 1990.0 1997.0   X
      2 50.0 1992.0     3.0       W       W 48.0 1990.0 1997.0   X
      2 53.0 1995.0     2.0       W       X 48.0 1990.0 1997.0   X
      3 47.0 1992.0     3.0       W       W 47.0 1992.0 1998.0   D
      3 50.0 1995.0     3.0       W       D 47.0 1992.0 1998.0   D
      4 51.0 1998.0     2.0       W       W 51.0 1998.0 2001.4   D
      4 53.0 2000.0     1.4       W       D 51.0 1998.0 2001.4   D
      5 48.5 1996.9     1.5       W       W 48.5 1996.9 2001.8   W
      5 50.0 1998.4     1.6       W       W 48.5 1996.9 2001.8   W
      5 51.6 2000.0     1.8       W       W 48.5 1996.9 2001.8   W
      6 41.0 1990.0     4.0       W       W 41.0 1990.0 1997.2   W
      6 45.0 1994.0     1.0       W       W 41.0 1990.0 1997.2   W
      6 46.0 1995.0     2.2       W       W 41.0 1990.0 1997.2   W
      7 44.0 1994.0     1.0       W       W 44.0 1994.0 1999.5   W
      7 45.0 1995.0     4.5       W       W 44.0 1994.0 1999.5   W
      8 40.0 1991.0     4.0       W       W 40.0 1991.0 2000.5   D
      8 44.0 1995.0     1.0       W       W 40.0 1991.0 2000.5   D
      8 45.0 1996.0     4.0       W       W 40.0 1991.0 2000.5   D
      8 49.0 2000.0     0.5       W       D 40.0 1991.0 2000.5   D
      9 40.0 1993.0     2.0       W       W 40.0 1993.0 1997.5   D
      9 42.0 1995.0     2.5       W       D 40.0 1993.0 1997.5   D
     10 47.0 2001.0     3.0       W       W 47.0 2001.0 2008.1   D
     10 50.0 2004.0     1.0       W       W 47.0 2001.0 2008.1   D
     10 51.0 2005.0     3.1       W       D 47.0 2001.0 2008.1   D
     11 42.0 1997.0     3.0       W       W 42.0 1997.0 2004.0   D
     11 45.0 2000.0     4.0       W       D 42.0 1997.0 2004.0   D
     12 40.0 1997.0     3.0       W       W 40.0 1997.0 2010.0   X
     12 43.0 2000.0     2.0       W       W 40.0 1997.0 2010.0   X
```

```
12 45.0 2002.0     3.0       W       W 40.0 1997.0 2010.0   X
12 48.0 2005.0     2.0       W       W 40.0 1997.0 2010.0   X
12 50.0 2007.0     3.0       W       X 40.0 1997.0 2010.0   X
13 41.0 2001.0     4.0       W       W 41.0 2001.0 2008.7   W
13 45.0 2005.0     3.7       W       W 41.0 2001.0 2008.7   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"))
```

```
    1990 1995 2000 2005
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 1990      1995      2000      2005
              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 1952–61, we simply restrict the dataset to his group, i.e. the persons where `per` − `age` is between 1952 and 1962:

```
> BC <- subset(oL,per-age>1952 & per-age<1962)
> 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 and the cumulative risk for the birth cohort born 1952-1961 from 40 till 55 years of age are computed

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

or in terms of the just computed:

```
> CumRate <- sum(FUctab[,1]/FUctab[,2]*5)
> CumRisk <- 1 - exp( - CumRate )
> round( c(CumRate, CumRisk), 3)
```

```
[1] 1.326 0.734
```

4. The total expected number of cases E in the whole study cohort, based on the age and period-specific rates in the national male population and the person-years in the study cohort split by age and period, is obtained as

$$E = \frac{100}{10^5\text{y}} \times (11 + 9.5 + 6 + 0)\ \text{y} + \frac{200}{10^5\text{y}} \times (6 + 12.2 + 10.5 + 5.7)\ \text{y} + \frac{400}{10^5\text{y}} \times (6 + 8.6 + 4.2 + 6.1)\ \text{y} = 0.1949$$

The total observed number is $O = 7$. Accordingly, the standardised incidence ratio SIR $= O/E$ is SIR $= 7/0.1949 = 35.9$. – 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 population rates to the dataset covering the follow-up of the cohort.

```
> 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
```

```
 lex.id  age    per lex.dur lex.Cst lex.Xst Age  AoE    DoE    DoX Xst rate
      8 40.0 1991.0     4.0       W       W  40 40.0 1991.0 2000.5   D  100
      9 40.0 1993.0     2.0       W       W  40 40.0 1993.0 1997.5   D  100
      8 44.0 1995.0     1.0       W       W  40 40.0 1991.0 2000.5   D  100
      6 41.0 1990.0     4.0       W       W  40 41.0 1990.0 1997.2   W  100
     12 43.0 2000.0     2.0       W       W  40 40.0 1997.0 2010.0   X  100
      9 42.0 1995.0     2.5       W       D  40 40.0 1993.0 1997.5   D  100
      7 44.0 1994.0     1.0       W       W  40 44.0 1994.0 1999.5   W  100
     12 40.0 1997.0     3.0       W       W  40 40.0 1997.0 2010.0   X  100
     13 41.0 2001.0     4.0       W       W  40 41.0 2001.0 2008.7   W  100
     11 42.0 1997.0     3.0       W       W  40 42.0 1997.0 2004.0   D  100
      3 47.0 1992.0     3.0       W       W  45 47.0 1992.0 1998.0   D  200
      2 48.0 1990.0     2.0       W       W  45 48.0 1990.0 1997.0   X  200
      5 48.5 1996.9     1.5       W       W  45 48.5 1996.9 2001.8   W  200
      6 46.0 1995.0     2.2       W       W  45 41.0 1990.0 1997.2   W  200
      8 45.0 1996.0     4.0       W       W  45 40.0 1991.0 2000.5   D  200
      6 45.0 1994.0     1.0       W       W  45 41.0 1990.0 1997.2   W  200
     12 45.0 2002.0     3.0       W       W  45 40.0 1997.0 2010.0   X  200
```

```
10 47.0 2001.0     3.0       W       W  45 47.0 2001.0 2008.1   D  200
 7 45.0 1995.0     4.5       W       W  45 44.0 1994.0 1999.5   W  200
13 45.0 2005.0     3.7       W       W  45 41.0 2001.0 2008.7   W  200
11 45.0 2000.0     4.0       W       D  45 42.0 1997.0 2004.0   D  200
 8 49.0 2000.0     0.5       W       D  45 40.0 1991.0 2000.5   D  200
12 48.0 2005.0     2.0       W       W  45 40.0 1997.0 2010.0   X  200
 1 51.0 1991.0     3.0       W       D  50 51.0 1991.0 1994.0   D  400
 3 50.0 1995.0     3.0       W       D  50 47.0 1992.0 1998.0   D  400
 2 50.0 1992.0     3.0       W       W  50 48.0 1990.0 1997.0   X  400
 2 53.0 1995.0     2.0       W       X  50 48.0 1990.0 1997.0   X  400
10 51.0 2005.0     3.1       W       D  50 47.0 2001.0 2008.1   D  400
 5 50.0 1998.4     1.6       W       W  50 48.5 1996.9 2001.8   W  400
 4 51.0 1998.0     2.0       W       W  50 51.0 1998.0 2001.4   D  400
 4 53.0 2000.0     1.4       W       D  50 51.0 1998.0 2001.4   D  400
 5 51.6 2000.0     1.8       W       W  50 48.5 1996.9 2001.8   W  400
10 50.0 2004.0     1.0       W       W  50 47.0 2001.0 2008.1   D  400
12 50.0 2007.0     3.0       W       X  50 40.0 1997.0 2010.0   X  400
```

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

```
> Obs <- with( oL, sum( lex.Xst=="D" ) )
> Exp <- with( oL, sum( lex.dur*rate/10^5 ) )
> round(c( Observed=Obs, Expected=Exp, SIR=Obs/Exp ), 3)
Observed Expected      SIR
   7.000    0.195   35.916
```

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

5.12 Cumulative rates

1. The estimate of the rates in the two groups are:

   Thorotrast: 751/19365.4 years = 0.03878 years^{-1} = 38.8/1000 years

   Controls: 797/30517.6 years = 0.02612 years^{-1} = 26.1/1000 years

   and hence the estimate of the rate ratio is:

$$\hat{\text{RR}} = \frac{751/19365.4 \text{ years}}{797/30517.6 \text{ years}} = 1.485$$

2. The standard deviation of log(RR) is $\sqrt{1/D_1 + 1/D_0}$, which in this case is:

$$S = \sqrt{\frac{1}{751} + \frac{1}{797}} = 0.0509$$

   leading to a 95% confidence interval of:

$$1.485 \overset{\times}{\div} \exp(1.96 \times 0.0509) = (1.344; 1.641)$$

   In R we would do this calculation as follows:

```
> RR <- (751/19365.4)/(797/30517.6)
> SE <- sqrt( 1/751 + 1/797 )
> erf <- exp( 1.96*SE )
> round( c( RR, RR/erf, RR*erf ), 4 )
```

```
[1] 1.4849 1.3441 1.6406
```

3. Calculation of the cumulative rates requires that we compute the age-specific rates and then make a weighted sum of them, so we start out by entering the data from a copy-paste from the .pdf-file:

```
> th <- matrix( c(
+  0,   5,  572.1,  11, 1536.1,
+ 20,  17, 1974.2,  16, 2449.1,
+ 30,  58, 3489.0,  35, 4228.8,
+ 40, 100, 4502.2,  67, 5822.3,
+ 50, 184, 4433.5, 137, 6647.0,
+ 60, 205, 2998.1, 211, 5780.3,
+ 70, 137, 1134.4, 206, 3113.6,
+ 80,  45,  261.5, 114,  939.8), ncol=5, byrow=TRUE )
> colnames(th) <- c("age","D.th","Y.th","D.ct","Y.ct")
> th
```

```
     age D.th   Y.th D.ct   Y.ct
[1,]   0    5  572.1   11 1536.1
[2,]  20   17 1974.2   16 2449.1
[3,]  30   58 3489.0   35 4228.8
[4,]  40  100 4502.2   67 5822.3
[5,]  50  184 4433.5  137 6647.0
[6,]  60  205 2998.1  211 5780.3
[7,]  70  137 1134.4  206 3113.6
[8,]  80   45  261.5  114  939.8
```

Then we compute the age-specific rates in the two groups — note that we can refer to the columns in the matrix `th` by column names:

```
> R.th <- th[,"D.th"]/th[,"Y.th"]
> R.ct <- th[,"D.ct"]/th[,"Y.ct"]
```

If we want to compute the cumulative rates, we must also have the interval lengths, 20 years, for the first, 10 for the next 6 and undefined for the last:

```
> ell <- c(20,rep(10,6),NA)
> cbind( th[,"age"], ell, R.th, R.ct )
```

```
        ell        R.th        R.ct
[1,]  0  20 0.008739731 0.007160992
[2,] 20  10 0.008611083 0.006533012
[3,] 30  10 0.016623674 0.008276580
[4,] 40  10 0.022211363 0.011507480
[5,] 50  10 0.041502199 0.020610802
[6,] 60  10 0.068376639 0.036503296
[7,] 70  10 0.120768688 0.066161357
[8,] 80  NA 0.172084130 0.121302405
```

The cumulative rates are now computed by multiplying the age-specific rates by the interval length and then adding up; the cumulative rate to age 70 is up to and including the age class 60–69, *i.e.* the first 6 classes; the cumulative rates to age 80 includes one more class:

```
> C70.th <- sum( (R.th*ell)[1:6] )
> C70.ct <- sum( (R.ct*ell)[1:6] )
> C80.th <- sum( (R.th*ell)[1:7] )
> C80.ct <- sum( (R.ct*ell)[1:7] )
```

we can then summarize the cumulative rates and the ratio of them between the two groups:

```
> round( rbind( c( C70.th, C70.ct, C70.th/C70.ct ),
+               c( C80.th, C80.ct, C80.th/C80.ct ) ), 3 )
```

```
      [,1]  [,2]  [,3]
[1,] 1.748 0.978 1.788
[2,] 2.956 1.639 1.803
```

We can produce a readable out put by putting the results in a matrix and giving it row- and column-names:

```
> Cmat <- rbind( c( C70.th, C70.ct, C70.th/C70.ct ),
+                c( C80.th, C80.ct, C80.th/C80.ct ) )
> rownames( Cmat ) <- c("cr.70","cr.80")
> colnames( Cmat ) <- c("Thorotrast","Controls","Ratio")
> round( Cmat, 2 )
```

```
      Thorotrast Controls Ratio
cr.70       1.75     0.98  1.79
cr.80       2.96     1.64  1.80
```

We see that the cumulative rates have approximately the same ratio whether we cumulate rates to age 70 or age 80.

4. The reason that these ratios are larger than the ratio of the crude rates is the assumptions behind the calculation based on the crude rates does not hold.

   Using the crude rates assumes that the rates are constant throughout the lifespan which is clearly not the case. When we use the cumulative rates for comparison, we only make the assumption that the ratio of the rates (RR) is the same in all age classes. This is also the minimal required assumption needed to make the calculation of the RR meaningful.

   This assumption is easily checked when we have the data available in R:

```
> cbind( th[,"age"], R.th/R.ct )
```

```
     [,1]     [,2]
[1,]    0 1.220464
[2,]   20 1.318088
[3,]   30 2.008520
[4,]   40 1.930167
[5,]   50 2.013614
[6,]   60 1.873163
[7,]   70 1.825366
[8,]   80 1.418637
```

We see that with the exception of the two first age-classes with very few cases, the age-specific rate-ratios are consistently larger than the crude rate-ratio.

5.13 Attributable risk

1. From the solution to exercise 2.12 we have that the relative risk based on the crude rates is 1.485, and hence the attributable risk is

$$\text{AR} = \frac{1.485 - 1}{1.485} = 0.327$$

For the relative risks based on the cumulative rates to 70 and 80, respectively, we get:

$$\text{AR}_{70} = \frac{1.82 - 1}{1.82} = 0.441 \qquad \text{AR}_{80} = \frac{1.80 - 1}{1.80} = 0.446$$

2. The differences are merely reflections of the differences in the estimates of the relative risk, where the estimate based on the crude rates is clearly invalid because it rests on the rather strong (and obviously wrong!) assumption that the rates are constant over the total age-span.

   The attributable number of cases (AC) are in the three instances $0.327 \times 751 = 245.6$, $0.441 \times 751 = 331.2$ and $0.446 \times 751 = 334.9$.

3. The age-specific relative risk (rate-ratio) and the nuber of cases in the throtrast grousp can be used to compute the attributable risk and attributable number of cases in each age-group.

   In order to find RRs and the numbers in the Thorotrast group we enter the same data again as in exercise 2.12:

```
> th <- matrix( c(
+  0,   5,  572.1,  11, 1536.1,
+ 20,  17, 1974.2,  16, 2449.1,
+ 30,  58, 3489.0,  35, 4228.8,
+ 40, 100, 4502.2,  67, 5822.3,
+ 50, 184, 4433.5, 137, 6647.0,
+ 60, 205, 2998.1, 211, 5780.3,
+ 70, 137, 1134.4, 206, 3113.6,
+ 80,  45,  261.5, 114,  939.8), ncol=5, byrow=TRUE )
> colnames(th) <- c("age","D.th","Y.th","D.ct","Y.ct")
> th
     age D.th   Y.th D.ct   Y.ct
[1,]   0    5  572.1   11 1536.1
[2,]  20   17 1974.2   16 2449.1
[3,]  30   58 3489.0   35 4228.8
[4,]  40  100 4502.2   67 5822.3
[5,]  50  184 4433.5  137 6647.0
[6,]  60  205 2998.1  211 5780.3
[7,]  70  137 1134.4  206 3113.6
[8,]  80   45  261.5  114  939.8
```

   Now we can compute the relevant figures; first the RR which is merely the ratio of the rates, so what we have here is just the age-specific RRs:

```
> RR <- (th[,"D.th"]/th[,"Y.th"]) / (th[,"D.ct"]/th[,"Y.ct"])
```

The attributable risk is a simple functions of the RR:

```
> AR <- (RR-1)/RR
```

which when multiplied by the total number of cases gives the attributable number of cases in the study population:

```
> D.tot <- th[,"D.th"] + th[,"D.ct"]
> AC <- AR * D.tot
> round( cbind( RR, AR, AC ), 3 )
```

```
        RR    AR      AC
[1,] 1.220 0.181   2.890
[2,] 1.318 0.241   7.964
[3,] 2.009 0.502  46.697
[4,] 1.930 0.482  80.479
[5,] 2.014 0.503 161.585
[6,] 1.873 0.466 193.916
[7,] 1.825 0.452 155.092
[8,] 1.419 0.295  46.921
```

```
> round( sum(AC), 1 )
```

```
[1] 695.5
```

These numbers are given in the nice table below:

Age	RR	AR	AC
0-19	1.220	0.181	0.9
20-29	1.318	0.241	4.1
30-39	2.009	0.502	29.1
40-49	1.930	0.482	48.2
50-59	2.014	0.503	92.6
60-69	1.873	0.466	95.6
70-79	1.825	0.452	61.9
80+	1.419	0.295	13.3
$\sum$			345.7

We see that the attributable number of cases in total is not exactly the same as the number obtained by using the common estimate of the relative risk to calculate an overall attributable risk. This is because the assumption about proportionality of rates, i.e. of constant relative risk over age-classes is not exactly fulfilled.

If the relative risks were exactly the same in all age classes then so would the attributable risks be, and hence the calculation based on the common relative risk estimate and the calculation based on age-specific relative risks would give the same result.

Chapter 6

Analysis of Epidemiological Data — Solutions

6.1 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.

6.2 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 higher 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.

6.3 Preventive trial

1. The study hypothesis is that beta carotene intake 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 $\log(RR)$ 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, but also that it would be only 5% higher. 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 carotene 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 10,000 men on beta carotene as compared with 10,000 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 randomized 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.

6.3.1 Modeling

We may do the calculations simpler and more elegantly by using a modeling approach; first we rename the variables for easier programming, and scale the PY to 10,000 years, and re-order the levels of the exposure factor:

```
> library( Epi )
> D <- cases
> Y <- pyears/10000
> G <- factor(c("Beta","Plc"))
> G <- Relevel( G, 2:1 )
> data.frame( D, Y, G )
              D        Y    G
BetaCarotene 474 8.419183 Beta
Placebo      402 8.463158  Plc
```

either using a multiplicative model providing rate-ratio:

```
> mm <- glm( cbind(D,Y) ~ G, family=poisreg )
> round( ci.exp( mm, pval=TRUE ), 3 )
            exp(Est.)   2.5%  97.5%     P
(Intercept)    47.500 43.076 52.378 0.000
GBeta           1.185  1.038  1.354 0.012
```

. . . or an additive model, providing rate-difference:

```
> ma <- glm( cbind(D,Y) ~ G, family=poisreg(link=identity) )
> round( ci.exp( ma, Exp=FALSE, pval=TRUE ), 3 )
            Estimate   2.5%  97.5%     P
(Intercept)     47.5 42.857 52.143 0.000
GBeta            8.8  1.926 15.674 0.012
```

Again we see that the rate difference is 9 cases per 10,000 PY in favor of the placebo group — and the numerical results are the same as by "hand-calculation".

6.4 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.

6.5 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.

6.6 Efficiency of study design

In a comparison 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.

6.6.1 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 period is going to be 1 year.

We put the follow-up time in `t` and then set up vectors of length 2: `G` — exposure 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 18
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] 1.333 0.529 3.359 2.519
```

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) is quite large.

Note we could also get the same result by Poisson regression

```
> library(Epi)
> round( ci.exp( glm( cbind(D,Y) ~ G, family=poisreg ) ), 3 )
            exp(Est.)  2.5% 97.5%
(Intercept)     0.002 0.001 0.004
GX              1.333 0.529 3.359
```

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 35
2   X 2000  6000 18 18
> 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] 2.057 1.165 3.632 1.766
> round( ci.exp( glm( cbind(D,Y) ~ G, family=poisreg ) ), 3 )
            exp(Est.)  2.5% 97.5%
(Intercept)     0.001 0.001 0.002
GX              2.057 1.165 3.632
```

We see that we have a somewhat better precision, but the relative uncertainty 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 5
2   X 4000 4000 12 7
> 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] 2.100 0.666 6.617 3.151
> round( ci.exp( glm( cbind(D,Y) ~ G, family=poisreg ), subset="G" ), 3 )
   exp(Est.)  2.5% 97.5%
GX       2.1 0.667 6.617
```

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 )
+ erf <- exp(1.96 * sqrt(sum(1/D)) )
+ c( ci.exp( glm( cbind(D,Y) ~ G, family=poisreg ), subset="G" ), erf )
+ }
```

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

```
> sim( t=1, N=c(8000,2000) )
[1] 0.5714286 0.1298716 2.5142573 4.4000713
```

```
> rbind(
+ t1=sim( t=1, N=c(8000,2000) ),
+ t2=sim( t=2, N=c(8000,2000) ),
+ t3=sim( t=3, N=c(8000,2000) ),
+ t4=sim( t=4, N=c(8000,2000) ) )
       [,1]      [,2]     [,3]     [,4]
t1 2.545455 0.9867672 6.566229 2.579635
t2 2.285714 1.1245913 4.645679 2.032511
t3 2.100000 1.2383351 3.561233 1.695842
t4 2.666667 1.6801134 4.232518 1.587208
```

Clearly there is a decrease in the uncertainty, with increasing follow-up time.

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

```
> rbind(
+ N8.2=sim( t=2, N=c(8000,2000) ),
+ N7.3=sim( t=2, N=c(7000,3000) ),
+ N6.4=sim( t=2, N=c(6000,4000) ),
+ N5.5=sim( t=2, N=c(5000,5000) ) )
         [,1]     [,2]     [,3]     [,4]
N8.2 2.947368 1.477824 5.878224 1.994423
N7.3 2.138889 1.199347 3.814446 1.783396
N6.4 2.700000 1.436319 5.075475 1.879827
N5.5 1.307692 0.784777 2.179038 1.666339
```

Here the effects on the precision are much smaller.

So if we want a clear picture of what goes on we must make a lot of simulations to see how the error-factor varies. Note that the *true* value of the rate-ratio is 2, try to run the previous set of simulations a couple of times and see how the estimates vary.

6.7 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
```

6.7.1 Statistical modeling

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):

```
> library(Epi)
> 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
M   144 112   M    N
F    32  45   F    N
M.1 141 208   M    Y
F.1  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
```

5. With all these items in place, we can now fit a logistic regression model, adjusting for sex (note that `y` is now a 2-column matrix):

```
> y

    D   C
M 144 112
F  32  45
M 141 208
F  49  58

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

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

Deviance Residuals:
      M        F        M        F
 0.8619  -1.5686  -0.7455   1.3480

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.01908    0.17083   0.112  0.91109
sexM         0.12387    0.17042   0.727  0.46733
physY       -0.45059    0.14522  -3.103  0.00192 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15.7949  on 3  degrees of freedom
Residual deviance:  5.5762  on 1  degrees of freedom
AIC: 33.725

Number of Fisher Scoring iterations: 3
```

6. It is easy to compare the estimate with the crude, not adjusting for sex:

```
> mcmod <- glm( y ~ phys, family=binomial )
> round( rbind( ci.exp(mamod,subset="phys"),
+               ci.exp(mcmod,subset="phys") ), 3 )
      exp(Est.)  2.5% 97.5%
physY     0.637 0.479 0.847
physY     0.637 0.479 0.847
```

   So we see there is apparently minimal confounding.

7. In order to see if there is effect-modification we include 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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.5795e+01  on 3  degrees of freedom
Residual deviance: 1.7319e-14  on 0  degrees of freedom
AIC: 30.149

Number of Fisher Scoring iterations: 2
```

The coefficients reported in the output refer to logarithms of hazard ratios for the two sexes. If we want the estimated hazard ratios themselves, we use the function `ci.exp` in `Epi` package to extract the coefficients and exponentiate them. By using the `subset` argument too, only the relevant components are extracted from the output.

```
> ci.exp( mimod )

            exp(Est.)      2.5%     97.5%
(Intercept) 0.7111111 0.4519653 1.1188447
sexM        1.8080357 1.0790859 3.0294096
sexF:physY  1.1880388 0.6574817 2.1467306
sexM:physY  0.5272436 0.3803267 0.7309132

> round( ci.exp( mimod, subset="phys" ), 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 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(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.exp( mcmod, subset="phys" ), 3 )
```

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

8. The modeling 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, mamod, mcmod, test="Chisq" )

Analysis of Deviance Table

Model 1: y ~ sex + sex:phys
Model 2: y ~ sex + phys
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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again we see that there is clear evidence of interaction but a very small sex-effect. However the sex-effect is irrelevant, it makes no sense to evaluate a sex effect assuming that the exposure effect is the same for men and women, when we just established that it is not.

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.

6.8 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.

6.9 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(1,1,0,1,0,1,1,1,1,1,0, 0,1,1,0,1,1,1,1,1,0,1)
> conexp <- c(0,0,0,1,1,0,0,0,1,1,0, 1,1,0,0,0,0,0,0,1,0,0)
> 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.

6.9.1 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 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(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.

6.10 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
```

We see that this is substantially different from the rate-ratios we saw across the age-groups

3. We can compute the adjusted RR using `glm`. To that end we first stack the deaths and the person-years, and define age-classes and groups. The `levels=` argument in the definition of `G` makes sure that `Pop` is the first level (and hence the reference level when modeling):

```
> D <- c(D1,D0)
> Y <- c(Y1,Y0)
> A <- factor( rep(c('30-39','40-49','50-59'),2) )
> G <- factor( rep(c("Wrk","Pop"),each=3), levels=c("Pop","Wrk") )
> 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
```

Then we can compute both the crude and the adjusted RR:

```
> mc <- glm( cbind(D,Y) ~     G, family=poisreg )
> ma <- glm( cbind(D,Y) ~ A + G, family=poisreg )
> round( rbind( ci.exp( mc, subset="G" ),
+               ci.exp( ma, subset="G" ) ), 3 )
```

```
     exp(Est.)  2.5% 97.5%
GWrk     1.333 0.938 1.896
GWrk     2.190 1.508 3.181
```

— as expected the adjusted is not far from the age-specific RRs, but the crude is, so we have a massive age-confounding of the crude estimate.

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 the SMR or direct standardization. The ratio of the crude rates is however very misleading as an estimate of the true rate ratio.

6.10.1 Statistical modeling

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, and also the SMR. 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(c('30-39','40-49','50-59'),2) )
> G <- factor( rep(c("Wrk","Pop"),each=3), levels=c("Pop","Wrk") )
> 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( cbind(D,Y) ~ G - 1, family=poisreg )
> round( ci.exp( mc ), 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( cbind(D,Y) ~ G, family=poisreg )
> round( ci.exp( mc ), 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( cbind(D,Y) ~ A:G -1, family=poisreg )
> round( ci.exp( mi ), 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( cbind(D,Y) ~ A-1 + A:G, family=poisreg )
> round( ci.exp( mi ), 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( cbind(D,Y) ~ A + G, family=poisreg )
> ms
Call:  glm(formula = cbind(D, Y) ~ A + G, family = poisreg)

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.exp( ms ), 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: cbind(D, Y) ~ A + G
Model 2: cbind(D, Y) ~ A - 1 + A:G
  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 and 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 out the irrelevant ones).

6.11 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: 0.1363
        Asymptotic P-value: 0.1246
------------------------------------------------------
```

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.