Practical exercises: Measures of Disease Occurrence Analysis of Epidemiological Data

Nordic Summerschool of Cancer Epidemiology
Danish Cancer Society, 17–28 August, 2015
http://BendixCarstensen.com/NSCE/2015
Version 3.4

Compiled Sunday 16th August, 2015, 21:27
from: /home/bendix/teach/NSCE/2015/pracs/pracs

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

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

Contents

Chapter 1

Introduction to exercises

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

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

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

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

1.1 What is R?

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

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

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

1.2 Getting R

You can obtain R, which is free, from CRAN (the **C**omprehensive **R** **A**rchive **N**etwork), at http://cran.r-project.org/. Under "`Download and Install R`" click on "`Windows`" and under "`R for Windows`" click on "`base`". Then on "`Download R 3.2.1 for Windows`", which is a self-extracting installer. This means that if you save it to your computer somewhere and click on it, it will install R for you.

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

1.2.1 Starting R

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

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

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

```
> library( Epi )
```

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

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

1.2.2 Quitting R

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

1.3 Working with the script editor

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

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

1.3.1 Try!

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

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

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

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

1.4 Getting a bit more training

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

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

1.5 Changing the looks of R

If you want R to start up with a different font, different colors etc., the go to the folder where R is installed — most likely `Program Files\R\R-3.2.1`, the to the folder `etc`, and open the file `Rconsole` with Notepad. In the file are specifications on how R will look when you start it, pretty self-explanatory, except perhaps for `MDI`.

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

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

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

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

1.6 Further reading

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

Chapter 2

Measures of Disease Occurrence Exercises

2.1 Using NORDCAN

2.1.1 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

2.1.2 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:

4. What were the average annual numbers of new cases in men and women during 2009–13?

5. How big were the estimated risks of getting cancer by 75 years of age for the two genders?

6. How many men and women died each year from lung cancer during 2006–2010?

7. What were the numbers of men and women living with lung cancer at the end of 2013, and how big were the corresponding proportions of lung cancer patients out of the whole male and female populations?

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

2.1.3 Incidence of lung cancer

Learn more about the incidence rates of lung cancer among men in the Nordic Countries during 2009-2013. 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.

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

10. 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.)

2.1.4 Population size and person-years

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

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

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

2.1.5 Mortality from lung cancer

Learn more about the mortality rates of lung cancer among men in the Nordic Countries during 2009–2013. Proceed as in task 1.3 above (ONLINE ANALYSIS → *Incidence/Mortality*, *etc.*), but now complete the choices by changing the **Data type** into Mortality and execute.

13. 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?

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

2.1.6 Prevalence of lung cancer

Learn more about the prevalence of lung cancer among men in the Nordic Countries at the end of 2013. 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 2012 from the pertinent boxes.

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

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

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

2.1.7 Lung cancer by age, period and cohort

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

18. Create a graph showing the age specific incidence and mortality of lung cancer among men in Denmark during 2009-13. From *Incidence/Mortality*, under **Graphs** choose Age-specific curves. Any comments to the graph?

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

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

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

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

2.1.8 Crude and standardized rates: stomach cancer

Obtain the crude and standardized incidence rates of male stomach cancer in the Nordic countries for 2013.

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

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

25. 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)?

2.1.9 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 2013.

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

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

2.1.10 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.

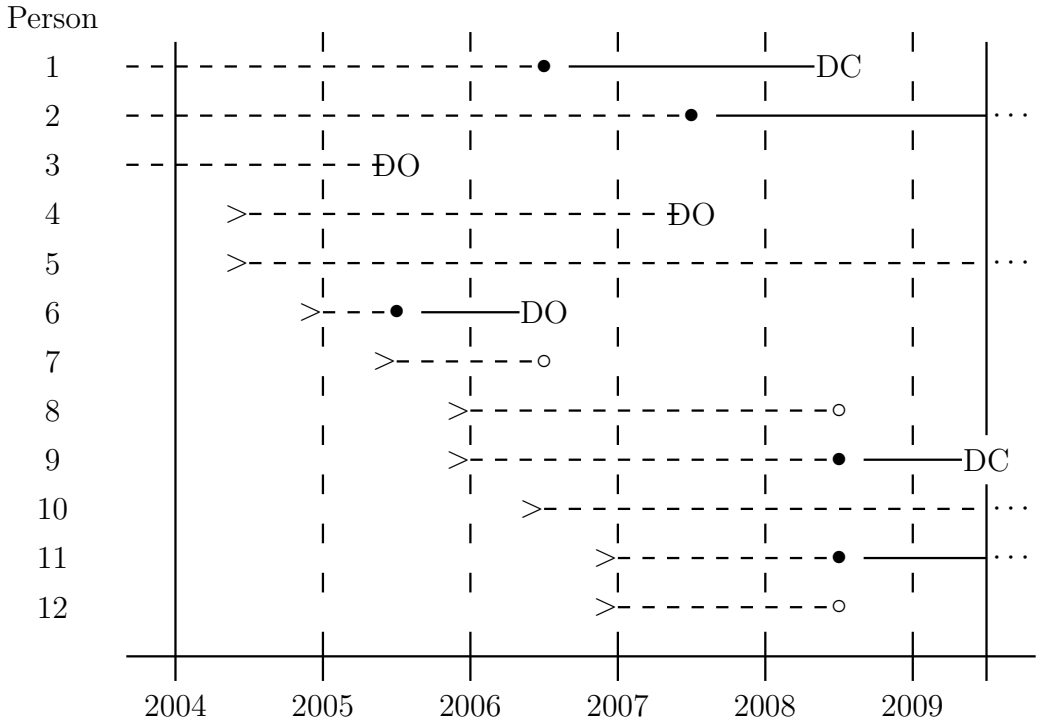
28. In which country was the relative survival poorest and where it was most favourable among male patients diagnosed in 2009–2013? What about female patients? How big where the 5-year relative survival proportions?

29. By how many percent points did the relative survival proportion improve in male patients of Norway during the 45 years since 1964-68?

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

2.2 Follow-up of a small cohort

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



• = disease onset ○ = censoring or withdrawal

DC = death from disease C DO = death from other cause

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

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

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

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

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

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

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

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

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

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

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

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

2.3 Infant mortality

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

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

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

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

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

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

2.4 Incidence and mortality of leukaemia in children

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

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

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

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

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

(b) Calculate similarly the mortality rates of leukaemia.

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

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

2.5 Rate ratio and rate difference in prevention trial

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

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

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

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

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

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

2.6 Rate ratio, rate difference and excess fraction

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

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

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

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

2.7 Crude and standardized rates

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

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

Calculate the following summary measures:

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

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

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

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

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

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

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

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

and weights from the male population:

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

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

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

2.8 Survival from tongue cancer

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

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

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

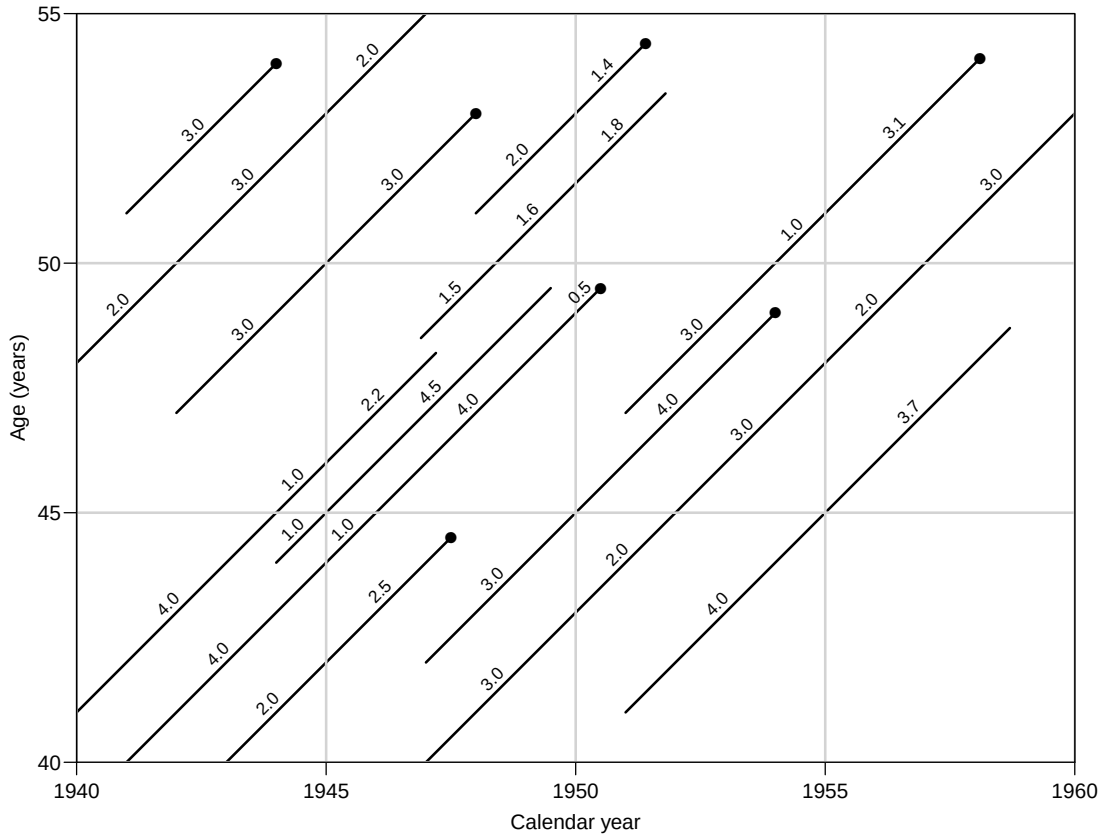
(b) Based on the results obtained above draw a survival curve and estimate graphically the median and the quartiles, if possible, of the survival time distribution.

2.9 Lexis diagram and occupational cohort I

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

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

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



2.10 Lexis diagram and occupational cohort II

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

Chapter 3

Analysis of Epidemiological Data Exercises

3.1 Single incidence rates

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

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

3.2 Non-significant difference

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

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

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

3.3 Preventive trial

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

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

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

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

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

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

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

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

The Alpha-Tocopherol Beta Carotene Cancer Prevention Study Group

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

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

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

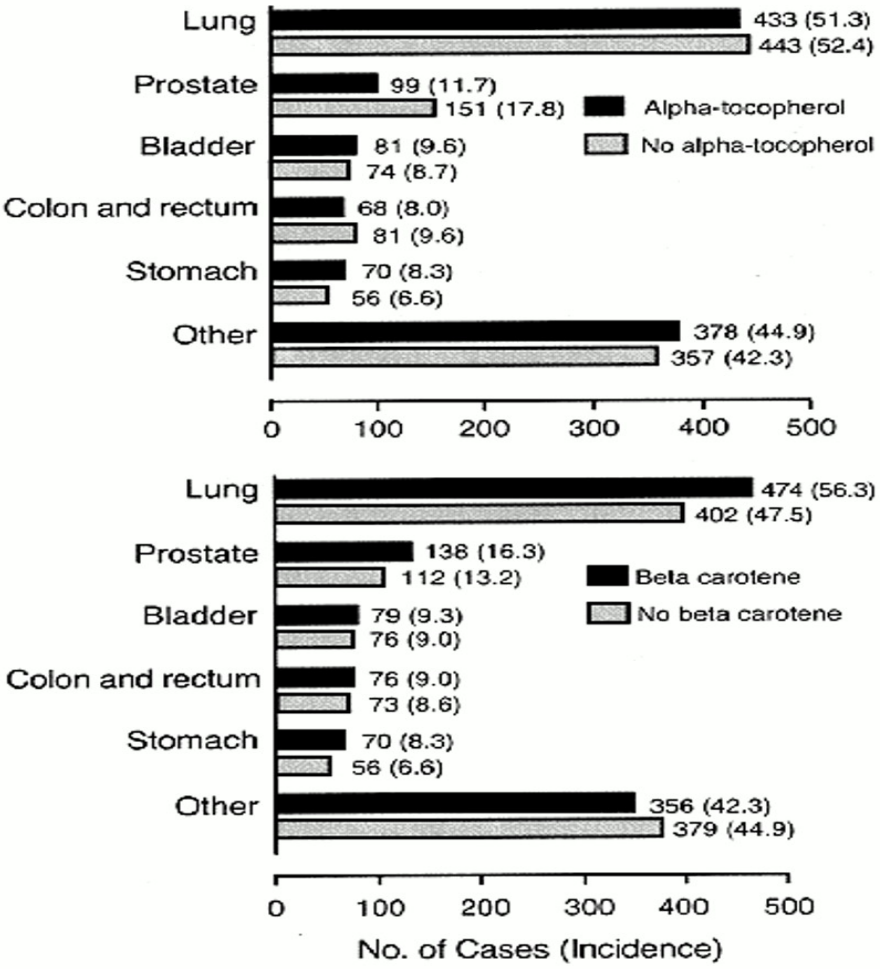


Figure 3.1: *Number and Incidence (per 10 000 Person-Years) of Cancers, According to Site, among Participants Who Received Alpha-Tocopherol Supplements and Those Who Did Not (Upper Panel) and among Participants Who Received Beta Carotene Supplements and Those Who Did Not (Lower Panel).*

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

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

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

3.4 Preventive trial – interpretation

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

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

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

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

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

3.5 Geographical variation

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

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

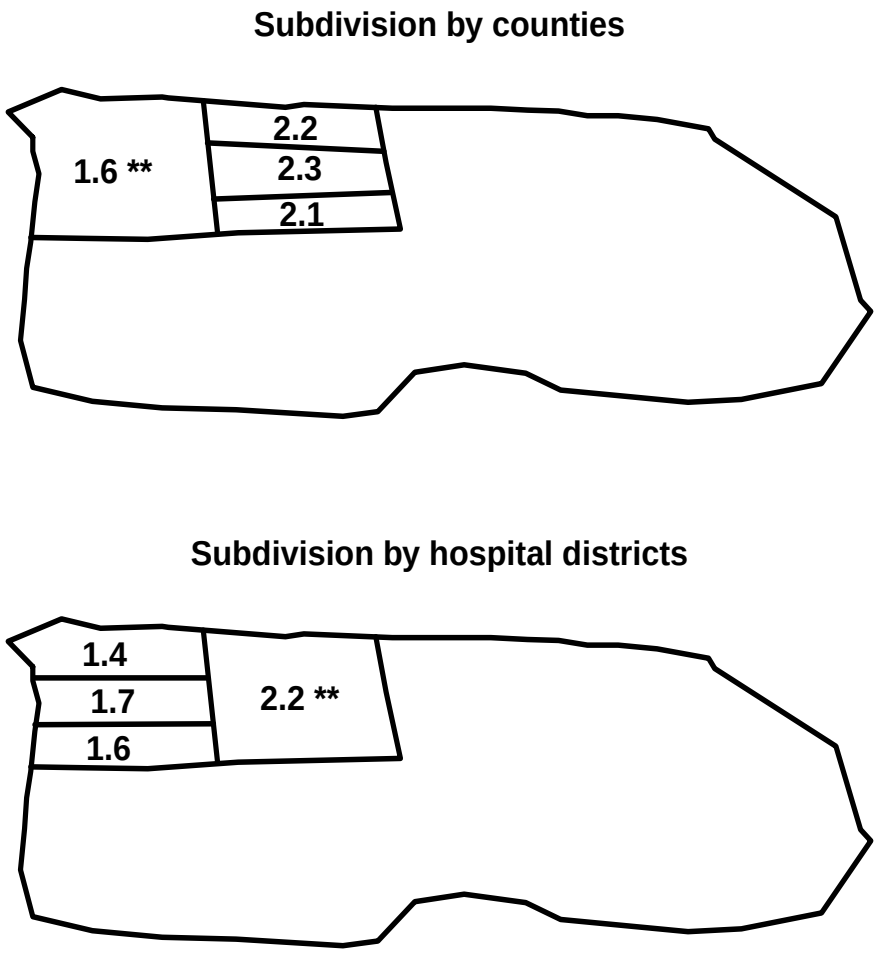


Figure 3.2: *Geographical division by county (top) and hospital district (bottom).*

3.6 Efficiency of study design

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

3.7 Case-control study: MI

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

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

1. Calculate the point estimate (and the 95% confidence interval) of the rate ratio in both genders separately.

2. What can you say of the possible modification of the effect of PA by gender; is the relative risk different in males than in females?

3. Is gender a confounder for the association between PA and MI; on what grounds?

4. Calculate the crude point estimate of the rate ratio, unadjusted for gender.

5. Calculate the gender-adjusted 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 ther effect-modification by sex?

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 *very* condensed overview of concepts and requires some familiarity with probability theory.

The target audience for this is

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

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:

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

Conditional survival function:

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

Cumulative distribution function of death times (cumulative risk):

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

Density function of death times:

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

Intensity:

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

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

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.

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

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

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

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

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

4.2 Statistics

Likelihood from one person:

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

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

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):

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

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

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

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

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

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

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

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

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

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

Therefore, in order to get a linear model for $\log(\lambda)$ 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.

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

4.3 Competing risks

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

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

The survival function is then:

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

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:

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

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

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

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

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

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

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

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

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

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

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

4.4 Demography

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

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

where f is the density of the distribution of lifetimes.

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

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

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

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

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

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

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

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

Note that the definition of the survival function for a non-diseased person requires a decision as to whether one will consider non-diseased persons immune to the disease in question or not. That is whether we will include the possibility of a well person getting ill and subsequently die. This does not show up in the formulae, but is a 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:

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

and hence:

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

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

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

These 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):

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

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