Epidemiology course for PhD students

Department of Biostatistics
Institute of Public Health
Spring 2015
http://BendixCarstensen.com/EpiPhD/F2015
Version 2.0 (preliminary)

Compiled Monday 2nd February, 2015, 20:31
from: `C:/Bendix/undervis/Epi/Epi-PhD/F2015/pracs/pracs.tex`

Bendix Carstensen	Steno Diabetes Center, Gentofte, Denmark & Department of Biostatistics, Institute of Public Health, University of Copenhagen bxc@steno.dk http://BendixCarstensen.com
Elisabeth Wreford Andersen	Department of Informatics and Mathematical Modeling DTU Data Analysis ewan@imm.dtu.dk
Henrik Ravn	Division of Epidemiology, Statens Serum Institut hjn@ssi.dk
Per Kragh Andersen	Department of Biostatistics, Institute of Public Health, University of Copenhagen pkan@sund.ku.dk
Torben Martinussen	Department of Biostatistics, Institute of Public Health, University of Copenhagen tma@sund.ku.dk

Contents

Chapter 1

Introduction

1.1 Preface

This is the collection of exercises for the course in epidemiology for PhD-students in the spring of 2015.

The exercises are based on students using SAS as the computer program for solving the exercises, and the weight of the recap of the exercises during the course will be on SAS too.

Students are however welcome to use other software packages, provided that they bring them on their own computer, and can access the datasets to be used from the net. Some of the teachers will have some expertise in some of the other frequently used computer packages such as Stata, R and (limited) SPSS. Most of the teachers have experience using R, some have experience in Stata, whereas none of the teachers use SPSS. This is reflected in the solutions sections which are only for SAS, Stata and R.

1.2 Website

The course website for the statistics practicals is `http://BendixCarstensen.com/EpiPhD/F2015`. Note that this website is sensitive to upper- and lower-case letters.

There will be links to this document, to the data, to the programs mentioned in this document and to solutions to the practicals as the course proceeds.

Whenever we refer to "from `www`" it means that you should go to the course website or the data website.

1.3 Data

The datasets are all found in the data folders `http://192.38.117.59/~pka/epidata/` resp. `http://192.38.117.59/~pka/spss-stata-data/`

There is also a link to these at the course website: `http://BendixCarstensen.com/EpiPhD/F2015`.

Description of the datasets are in the exercise texts.

Chapter 2

Exercises

2.1 Vaccinations and childhood mortality in Guinea-Bissau

In rural Guinea-Bissau, 5274 children under 7 months of age were visited two times at home with an interval of six months (Kristensen et al., *BMJ*, 2000, [2]). Information about vaccination (BCG, DTP, measles vaccine) was collected at each visit, and at the second visit, death during follow-up was registered. Some children moved away during follow-up, some survived until the next visit. The following variables are found in the data set `bissau.txt`:

`id`	Id number
`fuptime`	Follow-up time in days
`dead`	0 = censored, 1 = dead
`bcg`	1 = Yes, 2 = No
`dtp`	Number of DTP doses (0,1,2,3)
`age`	Age at first visit in days
`agem`	Age at first visit in months

2.1.1 A single risk, odds and rate

Tabulate the number of children, the number of deaths and the number of person-years.

- Do the following by using the formulae from the lectures:

  1. What is the overall risk of death? Make a confidence interval for this proportion.
  2. What is the overall odds of death? Make a confidence interval for this odds.
  3. What is the overall *rate* of death (per year). Make a confidence interval for this rate.

- Do the same by using your statistical package. Do you get the same confidence intervals?

2.1.2 Rates, risks and odds

First, make a table of the number of children, the number of deaths and the number of person-years by BCG vaccination status.

- Based on this do the following calculations by hand (or a suitable program on your computer), by inserting the numbers in the formulae from the lectures:

  1. Estimate the 6-month *risk* of death for children with or without BCG vaccination (`SAS` users may use the program `bissau.sas`).
  2. Compute 95% confidence limits for the two risk parameters.
  3. Estimate the 6-month *odds* of death for children with or without BCG vaccination. Compute also 95% confidence limits for the odds parameters. Compare with the risk parameters.
  4. Estimate (by your preferred statistical software) the *rate (per day)* of death for children with or without BCG vaccination. Compute also 95% confidence limits for the rate parameters.
  5. Estimate the *rate (per year)* of death for children with or without BCG vaccination. Compute also 95% confidence limits for the rate parameters.
  6. Create a new binary variable indicating whether or not the child was DTP vaccinated at first visit and repeat the previous questions for this DTP variable.

- Calculations using a statistical model

  1. Compute the risk with 95% confidence intervals in each of the two groups. You must fit a binomial model (without intercept) with log-link and exponentiate the estimates afterward.
  2. Compute the odds of death 95% confidence intervals in each of the two groups. You must fit a binomial model (without intercept) with logit link and exponentiate the estimates afterward.
  3. Compute the rate of death per *year* in each of the two groups. You must fit a Poisson model (without intercept) with log link and the log-person-years as offset and exponentiate the estimates afterward.
  4. Do the same for the subdivision of data by DTP.

2.1.3 Rate ratio, risk ratio, odds ratio

Continuing from before, calculate relative effects of BCG and DTP on mortality.

1. Calculate (SAS-users may use `proc freq`) the risk ratio and odds ratio and 95% confidence interval (CI) for the effect of BCG on mortality, i.e., compare the risk/odds of dying among BCG-vaccinated vs. BCG-unvaccinated. What do you conclude?

2. Do the same for DTP (any dose vs. none, i.e. as a binary exposure). What do you conclude?

3. Test the association between BCG and DTP-any dose using a Chi-square test. In this mortality is not involved, only test whether the occurrence of the two types of vaccination are related. How would you describe the relationship? What do you conclude?

4. Estimate the DTP effect (risk ratio and odds ratio) separately for each level of BCG. What happened?

5. Until now we have not accounted for the follow-up time. Repeat question 1, 2, and 4 but now by calculating the rate ratio and 95% CI for the BCG and DTP exposure.

2.1.4 Confounder control: stratified analysis of odds ratio and risk ratio.

1. Revisit the previous analyses of this dataset, but now using death (`dead`) as outcome, and estimate the DTP effect for each level of BCG.

2. Use the BCG as a potentially confounding variable and obtain the MH-estimate for the OR and RR. What are they?

3. Do the same, using age in months (`agem`) as control variable in the analysis. Is there any DTP effect?

4. Do the same, but now using both `agem` and `bcg` (that is, the cross-classification) as control variables in the analysis. Is there any DTP effect?

2.1.4.1 Computing hints

In SAS you can make an analysis controlling for confounding by including the confounder variable before the exposure and outcome variables in the table statement, and adding `cmh` as option (`cmh` = Cochran-Mantel-Haenzsel):

```
proc freq data = bissau ;
table agem * dtpany * dead / norow nocol nopct cmh ;
run ;
```

2.1.5 Survival analysis of childhood mortality in Guinea-Bissau

The SAS program `bissau.sas` reads the data from `www` and fits a simple Cox regression model with follow-up time as the time variable and including `bcg` and `agem` as categorical covariates.

1. Fit a simple Cox regression model with follow-up time as the time variable and including `bcg` and `agem` as categorical covariates and re-find estimates from the lectures.

2. Estimate the effect of any dose of DTP, using the created variable `dtpany` adjusted only for age in months as a categorical (`class` in SAS) variable.

3. Now, also adjust for BCG. What happened? Can you explain?

4. Is there an interaction between DTP (`dtpany`) and BCG?

5. Make a Cox regression analysis with DTP (`dtpany`) and BCG, but now with age as time-variable, i.e. with delayed entry.

6. Repeat the Poisson and logistic regression models that you have seen during the lectures, and compare the results:

Cox RR (95%CI)	Poisson RR (95%CI)	Logistic OR (95%CI)
0.71 (0.53-0.94)	0.71 (0.53-0.94)	0.71 (0.53-0.96)

   All models should be *adjusted for age in months as a categorical variable.* In the Cox model, follow-up time was used as the time-variable. In the Poisson model, the follow-up time was used as time at risk. The logistic regression did not take the follow-up time into account.

   What do you conclude?

2.2 Case-control study of renal cancer and trichorehtene

This exercise is based on the paper by Vamvakas *et al.*: Renal cell cancer correlated with occupational exposure to trichloroethene. J Cancer Res Clin Oncol, 1998, pp 374–382. [1].

The paper is available at the course homepage as
http://BendixCarstensen.com/EpiPhD/Papers/Vamvakas.1998.pdf

We will discuss the following points based on the paper:

1. What is the primary aim of the study?

2. How were the cases sampled?

3. How were the controls sampled?

4. Are they comparable; i.e. what assumptions are needed?

5. What is the (actual) study base?

6. What study base is the intended (for generalization)?

7. Is the sampling scheme incidence density sampling?

8. Can the age-effect on the occurrence of renal cancer be estimated?

9. Is age a confounder?

10. Key in the numbers in table 6 (p.380), and verify the analysis using your statistical software (in SAS you could use `proc freq`).

11. Is there any evidence of heterogeneity of the odds-ratio across age-classes? (*Hint:* Use the Breslow-Day-test.)

12. In particular, how does the odds-ratio estimate given by Vamvakas *et al.* compare the the Mantel-Haenszel estimate based on the same data?

13. What is the main result (in plain words)?

2.3 IHD data from Clayton & Hills.

The study is described by Clayton & Hills, Ch. 13. The tabulated data set of counts of IHD cases and person-years is avaliable from `www` in the file `ihd-tab.txt`.

The `SAS` program `ihd-reg.sas` reads the data from `www` and fits a Poisson regression model without interaction between age and exposure.

1. Fit the Poisson model from Clayton & Hills Tables 22.7-8 (p.222) and perform the tests from exercises 24.1 and 24.2 (pp.237–238). `SAS`-users may use the program `ihd-reg.sas` and notice the use of the `ESTIMATE` command to obtain a given reference group and the rate ratios with 95% confidence intervals.

2. Fit the model with interaction and re-find results from Clayton & Hills Table 24.5 (p.242) and the test for no interaction.

2.3.1 Using continuous variables

The IHD-data contains enegy consumption as a continuous variable. The dataset `diet.txt` has the following variables:

`id`	Person id
`doe`	Date of entry
`dox`	Date of exit
`chd`	CHD-status at exit: 0-no, 1-yes
`dob`	Date of birth
`job`	Not used
`month`	Notudes
`energy`	Daily energy intake in MJ
`height`	Height in cm
`weight`	Weight in cm
`fat`	Daily fat intake (g)
`fibre`	Daily fibre intake (g)

1. Read the individual diet data records from the file.

2. Create variables for the person-years, by subtracting entry date from date of exit. Also create a variable with the log-person-years.

3. Use CHD as outcome variable in a Poisson-analysis with the log-person-years as offset, using energy as a linear explanatory variable. Is there an effect on mortality?

4. Is there any evidence of a non-linear effect of energy, when using linear splines with knots at say 2, 2.5 and 3? (approx. the quartiles).

5. How does the non-linear relationship look? Plot the estimated curve together with the estimated linear relationship.

6. Same question for weight and BMI (the latter you have to calculate yourself as weight/height^2).

2.3.2 Splitting the follow-up of the IHD data

The following exercise is designed to illustrate how follow-up time is subdivided in order to produce the table of events and person-years. Furthermore the aim is to show you that tabulated data and time-split data gives the same results if only age and exposure are used as variables.

We will first analyze frequency records as above (these are almost identical to Table 22.6 in C & H). Next, we shall read the individual records and construct the corresponding table of cases and person-years.

The splitting of follow-up along a timescale is quite a technical task, which is handeled somewhat differently in SAS, Stata and R, so the exercise is here given in three different versions, one for each programming language.

2.3.2.1 Using SAS

1. Import the program `ihd-lexis.sas` to the program editor. Run the first part of the program — the part reading the tabulated data and `proc genmod`. Compare with the results from table 24.1 in Clayton & Hills.

2. Next, read the individual records from the file `diet.txt`, including the `proc print` and check on the output that it looks reasonable and that you understand what the data represents.

3. Now you should import the macro `%Lexis` and use it to split into the age intervals 40–50, 50–60 and 60–70 years:

   In order to use this you must first load it from the appropriate folder folder on the net:

   ```
   options source2 ;  * List the included code in the log-window ;
   filename lexispr url "http://BendixCarstensen.com/Lexis/Lexis.sas";
   %inc lexispr ;
   ```

   Once you have specified `%inc lexispr ;` and run that line in SAS, SAS will know the macro `%lexis` and you can use it in the rest of the session.

4. The time-splitting is now done by running the SAS-macro `%Lexis`[1]. A SAS-macro is a piece of SAS-program (normally quite long) where certain small parts of the program can be changed when the program is run. The SAS-convention is that names of such programs start with a “`%`”.

   To use the macro we must specify the follow-information from the input file:

[1] Named after the German demographer, statistician and economist, Wilhelm Lexis, 1837–1914. He wrote the book “Einführung in die Theorie der Bevölkerungsstatistik, (Strassbourg, 1875)”, while he was professor in Dorpat (now Tartu, Estonia), wherein he devised the so called Lexis diagram.

- Date of entry into the study — `doe`
- Date of exit from the study — `dox`
- Status at exit from the study — `chd` (1 if CHD occurred at `dox`, 0 otherwise).

Moreover, we must decide which timescale to split the data on. In this case we want to split along the scale "current age", i.e. time since date of birth. To this end we must specify:

- The origin of the time-scale, i.e. where the time-scale is 0, in this case date of birth — `dob`.
- The intervals where we want the follow-up grouped, here ages 40–50, 50–60 and 60–70.
- As a purely technical thing we need to specify the conversion between the scale in which time is measured in the input dataset (in this case days) and in the specification of the grouping (in this case years) — 365.25.

In the case of `%Lexis` we must supply these 6 parameters in order to specify how to split time.

Finally we must tell the program where the the original data is, where the time-split data has to go, and what the name of the created age-variable should be.

This looks like this (you do not have to write the stuff between the `/*...*/`):

```
%Lexis( data   = ihdindiv,        /* Dataset with original data       */
        out    = ihdsplit,        /* Dataset with time-split data     */
        entry  = doe,             /* Date of entry                    */
        exit   = dox,             /* Date of exit                     */
        fail   = chd,             /* Event (failure) indicator        */
        origin = dob,             /* Origin of the time-scale         */
        scale  = 365.25,          /* Conversion from input scale to breaks-scale
        breaks = 40 to 70 by 10,  /* Where to split the time scale    */
        left   = agr );           /* The name of the new age-variable */
```

Run this piece of SAS code.

(In the top of the file `http://BendixCarstensen.com/Lexis/Lexis.sas` are some more detailed explanations of how to use `%Lexis`).

5. How many records are in the resulting dataset (`ihdsplit`)

6. Take at look at the resulting data file, for example the first 20 records:

   ```
   proc print  data = ihdsplit (obs=20) ;
   run ;
   ```

   How does this compare with the the original dataset?

7. Use `%PYtab` to tabulate ihd-cases and person-years by exposure and age-group. You must first get this from the net as you did with the `%Lexis` macro:

```
filename pytabpr url "http://BendixCarstensen.com/Lexis/PYtab.sas";
%inc pytabpr ;

%PYtab( data  = ihdsplit,
        class = exposure agr,
        fail  = chd,
        risk  = risk,
        scale = 1000 ) ;
```

Compare with the sums from the table given in the first data step in `ihd-lexis.sas`

8. Use `proc genmod` to estimate the effect of age and exposure from the split dataset. How does the estimates compare with those based on the initial tabulated dataset?

9. Add an interaction between age and exposure and check that you get the same test for interaction as with the grouped data.

10. Compare the type 3 likelihood ratio statistic (`Chi-square`) for the interaction with the deviance of the model without interaction for the grouped data.

2.3.2.2 Using **Stata**

1. First we enter the grouped data and make a simple Poisson analysis:

```
input eksp agr pyrs cases
1 0  346.87  2
1 1  979.34 12
1 2  699.14 14
0 0  560.13  4
0 1 1127.70  6
0 2  794.15  8
end

poisson cases i.eksp i.agr , exposure(pyrs)
```

2. Then read the individual data, convert to date variables:

```
infile id str10 doe str10 dox chd str10 dob job month energy height /*
*/ weight fat fibre  in 2/L /*
*/ using "http://BendixCarstensen.com/EpiPhD/F2014/data/diet.txt", clear

*Get the dates into date format

gen date_entry = date(doe,"MDY")
gen date_exit  = date(dox,"MDY")
gen date_birth = date(dob,"MDY")
format date_entry date_exit date_birth %td
```

3. Now tell Stata that this is survival data, that is, when persons enter, exit and whether they are dead or not at exit (`fail`), and finally which scale we are on (`origin`):

   ```
   stset date_exit, failure(chd==1) entry(date_entry) origin(date_birth) /*
   */ scale(365.25) id(id)

   display _N
   ```

   Note that Stata now has generated 4 new variables `_t0`, `_t`, `_d` and `_st`, describing the survival. Read the help page for `stset` and make sure you understand what they mean. (A useful introduction to `stset` is `www.pauldickman.com/survival/stset.pdf`).

4. Then split the data into age groups 40–50, 50–60, 60–70 and generate a new variable called `current_age`:

   ```
   stsplit current_age, at(40(10)70) after(date_birth)

   * How many observations?
   display _N
   ```

5. Now take a look at the data:

   ```
   list in 1/10
   browse
   ```

6. Tabulate IHD cases and person-years by exposure and age group. To this end we use the sytem variables `_t0` and `_t` which hold the left and the right end-points on the "analysis time scale", in this case the current age:

   ```
   gen pyrs=_t-_t0
   gen exposure = (energy < 2.75) + energy-energy

   * Only count CHD cases once
   gen event=_d

   table current_age exposure, c(sum event sum pyrs) format(%9.2f)
   ```

   Note that `current_age` is 0 for all follow up before age 40 (left of first cutpoint).

7. Now use `poisson` (or `glm`) to estimate the effect of age and exposure from the split dataset. How does the estimates compare with those based on the initial tabulated dataset?

   ```
   * drop follow-up before age 40
   keep if current_age>0

   poisson event i.exposure i.current_age, exposure(pyrs)
   ```

— the same result as with the tabulated data.

8. Add an interaction between age and exposure and check that you get the same test for interaction as with the grouped data.

   ```
   poisson event i.exposure i.current_age i.exposure#i.current_age, exposure(pyrs)
   testparm i.exposure#i.current_age
   est store m1

   poisson event i.exposure i.current_age , exposure(pyrs)
   est store m2
   lrtest m1 m2
   ```

9. Compare the type 3 likelihood ratio statistic for the interaction with the deviance of the model without interaction for the grouped data.

   ```
   collapse (sum) pyrs event , by(exposure current_age)

   poisson event i.exposure i.current_age , exposure(pyrs)
   ```

2.3.2.3 Using R

The following instructions are fairly detailed. You should make sure that you know what goes on, and that consult the help-pages for the functions uses, so that you get a bit of a feeling for how the R-machinery works.

1. Load the Epi package and read the (modified) grouped IHD-data from the file ihd-xtab.dta from the data folder http://BendixCarstensen.com/EpiPhD/F2014/data

   ```
   > options( width=90 )
   > library( Epi )
   > library( foreign )
   > ihdt <- read.table("http://BendixCarstensen.com/EpiPhD/F2014/data/ihd-tab.txt", header=T )
   > ihdt
   ```

   Fit a Poisson model to data with exposure and age-effects:

   ```
   > mt <- glm( cases ~ factor(age) + exposure,
   +            offset = log(pyrs), family=poisson, data=ihdt )
   > round( ci.exp( mt ), 3 )
   ```

   Compare with the results from table 24.1 in Clayton & Hills.

2. Next, read the individual records from the file diet.txt; remembering to specify how missing is coded:

   ```
   > ihdi <- read.table( "../data/diet.txt",
   +                     # "http://www.biostat.ku.dk/~pka/epidata/diet.txt",
   +                     header=TRUE, na.strings=".", as.is=TRUE )
   > head( ihdi )
   > str( ihdi )
   > # Turn character variables into dates and then to calendar years:XS
   > for( i in c(2,3,5) ) ihdi[,i] <- cal.yr( as.Date(ihdi[,i],format="%m/%d/%Y") )
   > str( ihdi )
   > head( ihdi )
   ```

Now check that it looks reasonable and that you understand what the data represents.

3. Now you should set up the dataset as a `Lexis` object[2]., so that R will know when persons are at risk etc. `entry` is a named list, the names giving the names of the timescales we want to use, in this case `per` (calendar time, `period`) and age. `exit` is also a named list, with one element with the name of one of the timescales, giving the values of the exit times on this time scale. `exit.status` gives the state that persons are in at exit from the study. If `entry.status` is not specified, it is assumed that everyone starts in the *first* state, and this is noted:

```
> Lx <- Lexis( entry = list( per=doe,
+                            age=doe-dob ),
+               exit = list( per=dox ),
+        exit.status = factor( chd, labels=c("Well","IHD") ),
+               data = ihdi )
> summary( Lx )
```

There is a method for plotting the follow-up in boxes. Not desperately exciting but capturing the essence:

```
> boxes( Lx, boxpos=TRUE, scal.Y=1000, show.BE=TRUE )
```

4. The time-splitting is now done by the function `splitLexis`. To use the function we must specify which timescale to split the data on. In this case we want to split along the scale “current age”, i.e. time since date of birth, here named `age`. We then specify the intervals where we want the follow-up grouped, here ages 40–50, 50–60 and 60–70, so use the breakpoints 40, 50, 60 and 70:

```
> Ls <- splitLexis( Lx, breaks=c(40,50,60,70), time.scale="age" )
> summary( Ls )
> head( Ls )
```

For the fun of it you can try the default `plot` and `points` methods for a `Lexis` object. Note that grid-lines corresponding to the breaks gets inserted:

```
> plot( Ls, col=gray(0.3) )
> points( Ls, col="red", pch=c(NA,16)[Ls$lex.Xst], cex=0.7 )
```

On the diagram it appears that all persons are censored at age 70 and at the end of 1976, whereas some follow-up time is present before age 40.

5. The number of records are in the resulting dataset (`Ls`):

```
> nrow( Ls )
```

6. We now list the first 20 records:

```
> head( Ls, 20 )
```

[2] Named after the German demographer, statistician and economist, Wilhelm Lexis, 1837–1914. He wrote the book “Einführung in die Theorie der Bevölkerungsstatistik, (Strassbourg, 1875)”, while he was professor in Dorpat (now Tartu, Estonia), wherein he devised the so called Lexis diagram.

7. Now reproduce the table in Clayton & Hills:

   First use the function `timeBand` to produce a variable which is equal to the left endpoint of the intervals into which the follow-up have been split:

```
> Ls <- transform( Ls, agr = timeBand( Ls, "age", "factor" ),
+                      eksp = factor( energy<2.75, labels=c("High","Low") ) )
> str( Ls )
```

   Then make a table like the one in C& H:

```
> round( ftable( xtabs( cbind( D=(lex.Xst=="IHD"), Y=lex.dur ) ~
+                       agr + eksp,
+                       data = Ls ),
+                row.vars = 1 ), 2 )
```

   You should see that the data is not quite the same as in the book.

   Now we do the grouped analysis on the slightly modified data that you can get from the data folder (which should be identical to the table you just made):

```
> ihdx <- read.table("http://BendixCarstensen.com/EpiPhD/F2014/data/ihd-xtab.txt", header=T )
> ihdx
> mt <- glm( cases ~ factor(age) + exposure,
+            offset = log(pyrs), family=poisson, data=ihdx )
> round( ci.lin( mt, E=T ), 3 )
```

8. Estimate the effect of age and exposure from the split dataset. Remember to exclude follow-uptime before age 40 — as you saw from the table above:

```
> Ls <- subset( Ls, agr %in% levels(agr)[2:4] )
> Ls$agr <- factor( Ls$agr )
> table( Ls$agr )
> head( Ls )
> mi <- glm( (lex.Xst=="IHD") ~ factor(agr) + eksp,
+            offset = log(lex.dur), family=poisson, data=Ls )
> round( ci.lin( mi, E=T ), 3 )
> round( ci.lin( mt, E=T ), 3 )
> ci.lin( mi, E=T ) / ci.lin( mt, E=T )
```

   We see that the estimates are identical for the two ways of modeling. The point of using the individual data is that individual-level variables could be included in a model too.

9. Add an interaction between age and exposure and check that you get the same test for interaction as with the grouped data.

```
> mix <- update( mi, . ~ . + factor(agr):eksp )
> mtx <- update( mt, . ~ . + factor(age):exposure )
> anova( mi, mix, test="Chisq" )
> anova( mt, mtx, test="Chisq" )
```

10. Compare the type 3 likelihood ratio statistic (`Chi-square`) for the interaction with the deviance of the model without interaction for the grouped data. You can get the deviance from the `summary`:

```
> summary( mt )
```

2.4 Case-control study of BCG vaccination and leprosy.

The study is decribed by Clayton & Hills, p.156. In short, 260 cases of leprosy among individuals aged less than 35 years were ascertained in a study area in Malawi. Subjects were grouped into 7 age intervals and according to absence or presence of a scar after BCG vaccination. Three sets of controls were studied:

1. a population survey of 80,622 persons
2. a random sample of 1000 persons
3. a 4 to 1 age-matched sample

The file `bcgalldata.txt` includes data from this study: for each of the 14 `age` by `scar` combinations, a text variable `status` indicates the type of person in question (`case, conall, con1000, conmatch`) and the numerical variable `n` the number of such persons.

The `SAS` program `bcg-reg.sas` reads these data and fits a logistic regression model with no interaction between `age` and `scar` using all cases and all controls.

1. Fit the model from Clayton & Hills Table 23.5 (p.232). `SAS`-users may use the program `bcg-reg.sas`; what are the reference groups?
2. Estimate odds ratios and confidence intervals with non-exposed and youngest, respectively, as reference groups (in `SAS`: use '`ESTIMATE`' statements).
3. Estimate instead odds ratios and confidence intervals with the age group 20-24 as reference.
4. Test the hypothesis of no interaction between `age` and `scar`.
5. Analyse the data set with only 1000 controls (i.e., use the controls `con1000`: Table 23.6, p.233) and compare the precision of the estimate for `scar` with that based on the entire sample.
6. Analyse the matched data set (i.e., use the controls `conmatch`: Table 23.6, p.233) and compare with the results from Table 23.7.
7. Try (erroneously) to drop `age` from the analysis of the matched data and study the consequences for the estimate of `scar`.

2.5 Case-control study of malignant melanoma.

Anne Østerlind conducted in the middle of the 80's a case-control study of risk factors for malignant melanoma in Denmark.

The review paper "Malignant melanoma in Denmark" from *Acta Oncologica*, 1990 ,[3] is from Anne Østerlind's thesis and gives an overview of the results from the study which included 1400 interviewed persons, 474 cases and 926 controls, cf. table 5 in the article.

In the article incidence changes between 1943 and 1982 are also discussed; that part of the paper will not be touched upon in this exercise.

2.5.1 Discussion of the article.

1. Explain the design, the data base and data collection, particularly how the matching was conducted.

2. How were interviews planned to minimize bias?

3. Explain the drop-out, particularly the analyses in Tables 5-7. What are the consequenses of these results for the subsequent analyses?

4. How are the analyses carried out? Are all variables included in one step or are the analyses conducted in smaller steps? How are the matching variables accounted for? Comments?

5. Explain the analyses presented in Table 9. How many logistic regression models are fitted here?

6. What is the conclusion from the analyses in the table?

7. What is the purpose of Table 11?

8. Which modifiable factors seem to affect the melanoma risk?

2.5.2 Melanoma data

We have access to a subset of the variables from the study. These are found in the file `melanom.txt`. The variables are described in the table below. Based on these data, results from AØ's Tables 9 and 10 can (almost) be reconstructed. Revised versions of those two tables are also found below.

The `SAS` program `melanom.sas` reads the data from `www` and fits a simple logistic regression model including only the variable `skin`.

2.5.3 Simple tabulation analysis

1. Make the two by two table showing the association between case-control status and whether or not the person experienced *any* sunburns before the age of 15. `SAS`-users may use the program `melanom.sas` to read in the data from `www`. Estimate the odds ratio with associated 95% confidence limits and test for no association between the risk factor and case-control status.

2. Conduct similar analyses for the factors `sex, hair, eyes, freckles, acuterea, chronrea`. Compare with Table 9 in the article.

3. The case control study was matched for sex and age and, therefore, analyses of any risk factor should be adjusted for these two variables. Study how much the association between the risk factor "any sunburns before the age of 15" and case-control status is affected by adjustment for sex.

4. Same question for age.

Table 2.1: *Variables in the melanoma data set. Some variables have missing values for some of the persons, these are coded “.”. In the file there is one line for each person in the study. Data are found in the file* `melanom.txt`*.*

`casecon`	— case-control status: 1:case, 0:control
`sex`	— 1:man, 2:woman
`ageint`	— age at interview in years
`agroup`	— grouped age: 10:10–19, 20:20–29, ...
`skin`	— skin colour: 0:dark, 1:medium, 2:light
`hair`	— hair colour: 0:dark brown/black, 1:light brown, 2:blond, 3:red
`eyes`	— eye colour: 0:brown, 1:grey/green, 2:blue
`freckles`	— freckles: 1:many, 2:some, 3:none
`acuterea`	— acute reaction to sunlight: 1:blisters, 2:painful sunburn, 3:mild sunburn, 4:no sunburn
`chronrea`	— chronic reaction to sunlight 1:deep tan, 2:moderate tan, 3:mild tan, 4:no tan
`nvsmall`	— number of naevi < 5mm
`nvlarge`	— number of naevi ≥ 5mm
`nvtot`	— total number of naevi
`burn15`	— number of sunburns before age 15

2.5.4 Simple analysis controlling for age

1. The case-control study was matched for sex and age and, therefore, analyses of any risk factor should be adjusted for these two variables.

   Study how much the association between the risk factor “any sunburns before the age of 15” and case-control status is affected by adjustment for sex.

2. Study how this association is affected by adjustment for age.

3. Study how this association is affected by adjustment for *both* age and sex.

2.5.5 Introductory analyses.

1. Estimate (log-)odds ratios for the variable `skin` (see top left in AØ’s Table 9). `SAS`-users may use the program `melanom.sas`.

2. Estimate also odds ratios (in `SAS`: use `ESTIMATE` statements).

3. Conduct the other analyses in AØ’s Table 9 (*left* part) where the factors `hair, eyes, freckles, acuterea, chronrea` are studied one at a time.

4. Conduct the analysis corresponding to Table 9 (*right* part) where several variables are included simultaneously (see the table footnote).

5. Reconstruct the results from AØ’s Table 10 concerning number of raised naevi.

Table 2.2: *Corrected Table 9. from the paper*

Factor	Category	OR (crude)	OR (adjusted)
Skin colour	Dark	(1.0)	(1.0)
	Medium	1.4 (1.0-1.9)	**1.3 (1.0-1.8)**
	Light	1.7 (1.2-2.3)	1.3 (0.9-**1.9**)
	trend test	$p < 0.01$	$p =$**0.15**
Hair colour	Dark-brown/black	(1.0)	(1.0)
	Light-brown	1.5 (1.2-1.9)	1.5 (1.2-1.9)
	Blond/fair	1.7 (1.0-2.9)	**1.6 (0.9-2.8)**
	Red	**1.7** (1.1-2.7)	1.3 (0.8-2.0)
	trend test	$p < 0.001$	$p = 0.04$
Eye colour	Brown	(1.0)	(1.0)
	Grey/green	0.9 (0.6-1.2)	0.7 (0.5-1.1)
	Blue	1.1 (0.8-1.5)	0.9 (0.6-1.3)
	trend test	$p =$**0.32**	$p =$**0.98**
Freckles	None	(1.0)	(1.0)
	Some	1.5 (1.2-1.9)	1.5 (1.2-2.0)
	Many	3.0 (2.2-4.1)	**3.0** (2.1-4.1)
	trend test	$p < 0.001$	$p < 0.001$
Acute reaction to sunlight	No sunburn	(1.0)	(1.0)
	Mild sunburn	1.3 (1.0-1.6)	1.1 (0.8-1.4)
	Painful sunburn	1.6 (**1.0**-2.6)	1.3 (0.8-2.1)
	Blisters	2.2 (0.9-5.0)	1.6 (0.7-3.9)
	trend test	$p =$**0.005**	$p =$**0.15**
Chronic reaction to sunlight	Deep tan	(1.0)	(1.0)
	Moderate tan	1.4 (1.1-1.8)	1.2 (0.9-1.6)
	Mild tan	1.8 (1.3-2.6)	1.4 (1.0-2.1)
	No tan	2.0 (1.0-3.7)	1.2 (0.6-2.5)
	trend test	$p < 0.001$	$p =$**0.10**

NB: new variables must be defined from the original variables `nvtot, nvsmall, nvlarge`.

2.5.6 Trend tests and interactions.

6. In the analyses so far all variables have been considered as categorical ('`class`' in `SAS`) variables while all tests in Tables 9 and 10 are trend tests. Conduct the analyses which give the P-values in Table 9 (right part) for the variables `skin` and `freckles`.

7. May `freckles` be scored linearly (1, 2, 3), when this variable is studied separately? (Conduct a test for linearity/departures from trend).

8. In AØ's Table 11 `freckles` and the total number of naevi (suitably grouped) are

Table 2.3: *Corrected Table 10.*

Factor	Category	OR (crude)	OR (adjusted)
Number of raised naevi on arms, total	None	(1.0)	(1.0)
	1	1.5 (1.1-2.1)	**1.5 (1.1-2.0)**
	2-4	2.3 (1.6-3.1)	2.2 (1.6-3.1)
	5+	5.4 (3.5-8.1)	**4.9 (3.2-7.5)**
	trend test	$p < 0.001$	
Number of raised naevi on arms, < 5 mm (diameter)	None	(1.0)	(1.0)
	1	1.6 (1.1-2.2)	1.6 (1.1-**2.2**)
	2-4	2.5 (1.8-3.4)	**2.4 (1.7-3.4)**
	5+	5.0 (3.3-7.7)	**4.7 (3.0-7.4)**
	trend test	$p < 0.001$	
Number of raised naevi on arms, $\geq$ 5mm (diameter)	None	(1.0)	(1.0)
	1	1.8 (1.2-2.8)	**1.6 (1.1-2.5)**
	2+	3.6 (1.8-7.2)	**2.7 (1.3-5.5)**
	trend test	$p < 0.001$	

studied. Conduct this analysis. Is there any interaction between these two variables?

9. Study, in a similar vein, interactions between `acuterea` and `skin` and between the grouped version of `nnvtot` from question 5. and `agroup`.

10. All of AØ's analyses are conducted without accounting for the match variable age (`agroup`) (in spite of warnings given by Clayton & Hills!). Repeat some of the previous analyses adjusting for `agroup`. Are there any substantial differences? Explain!

2.6 Testicular cancer risk and maternal parity.

This exercise deals with the article "Testicular cancer risk and maternal parity: a population-based cohort study", by T. Westergaard, P.K. Andersen, J.B. Pedersen, M. Frisch, J.H. Olsen, M. Melbye. *Br. J. Cancer*, **77**,pp. 1180-1185 (1998). [4].

2.6.1 Discussion of the article.

1. What is the authors' argument for the existence of an effect of maternal parity on the risk of testicular cancer in the son?

2. Describe the design of the study:

    - a. which "sons" are included in the study?
    - b. when are they followed?
    - c. how are cases defined and ascertained?

3. Concentrating on all testicular cancers, what do you consider to be the main result reported in Table 1?

4. Explain in words the interpretation of the value RR=0.80 for parity 2+.

5. Compare this value with the corresponding crude RR (and 95 % CI) obtained without any adjustment. Explain the differences between the two results.

6. Draw a Lexis diagram to illustrate the combinations of age and calendar period which contribute person-years to the study. An empty diagram is available as `http://BendixCarstensen.com/EpiPhD/F2014/blank-Lexis.pdf`

7. Explain the meaning of the estimates for "Interval from ..." in the lower part of Table 1.

8. What type of analysis is reported in Table 2?

9. Discuss how, alternatively, a case-control design could have been conducted to address the same question as the cohort study reported in the article.

2.6.2 Practical exercises

The file `testis.txt`, available at `www` contains for each (non-empty) combination of the factors `SON_AGE, SON_KOH, MOTH_AGE, PARITY` the number of person-years at risk `PYRS`, the numbers of non-seminomas and seminomas, respectively `NONSEMI SEMI`, and the total number of testis cancer cases `CASES`. The first line of the file contains the variable names.

The `SAS` program `testis.sas` reads the data from `www`.

10. Compute the crude rate ratio for testis cancer for parity 2+ versus parity 1. Compare with 5. above. `SAS`-users may use the SAS program `testis.sas` (and `PROC GENMOD`).

11. Reconstruct the estimates for "parity of mother at birth of son" from the top of Table 1 in the article both for all testis cancers and for non-seminomas.

12. Reconstruct the estimates from Table 2 in the article concerning mother's age (for all testis cancers). Is there an interaction between parity and mother's age?

13. Same question for birth cohort of the son.

2.7 Mediation analysis

The first exercise is a very simple exercise illustrating the simplest possible setup, based on simulated data.

Suppose there is a relationship between the outcome variable of interest, y, and the exposure variable of interest, x. But that some of the effect is through the mediating variable, m, which is influenced by x and in turn influences y.

```
> library(Epi)
> tmat <- matrix( NA, 3,3 )
> rownames(tmat) <-
+ colnames(tmat) <- c("x","m","y")
> tmat[1,2] <- tmat[1,3] <- tmat[2,3] <- 1
> boxes.Lexis( tmat, boxpos=list(x=c(15,50,85),y=c(20,80,20)), hmult=2, wmult=2 )
```

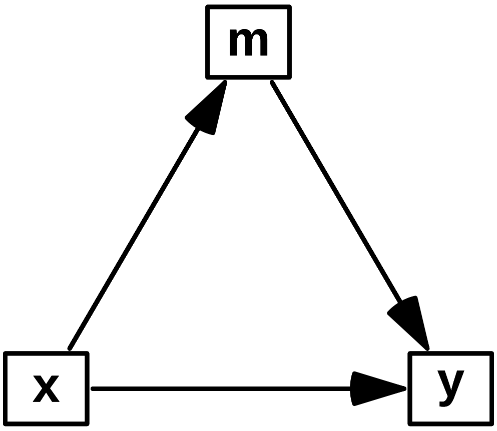


Figure 2.1: *Relationship between variable*

Under this assumption, and under the assumption that the relationships are linear we have for the recorded values of the three variables y_i, x_i and m_i:

$$y_i = \mu + \beta_x x_i + \beta_m m_i + e_i, \quad e_i \sim \mathcal{N}(0, \sigma^2)$$

and for the mediator:

$$m_i = \xi + \theta_x x_i + a_i, \quad a_i \sim \mathcal{N}(0, \tau^2)$$

If we naively insert the latter equation in the first we get:

$$\begin{aligned} y_i &= \mu + \beta_x x_i + \beta_m(\xi + \theta_x x_i + a_i) + e_i \\ &= (\mu + \beta_m \xi) + \beta_x x_i + \beta_m \theta_x x_i + \beta_m a_i + e_i \\ &= (\mu + \beta_m \xi) + (\beta_x + \beta_m \theta_x) x_i + \beta_m a_i + e_i \end{aligned}$$

Thus in a marginal analysis of y on x we would say that $\beta_m \theta_x$ of the effect of x on y were mediated by m.

Also we should see that the residual variance in the marginal analysis is $\beta_m^2 \tau^2 + \sigma^2$.

The point of the following exercise is to estimate the parameters in the two defining relationships and see if the relationship in the marginal analysis is as predicted from the two analyses.

Note that the setup is conditional on the relationship being as postulated, there is no guarantee that other relationships between variables could generate data with a similar joint distribution.

This exercise is based on simulated data in the dataset `medex.txt`. The dataset has 3 variables — the names are in the first line, so you can read it into Rby:

```
> med <- read.table( "http://BendixCarstensen.com/EpiPhD/F2014/data/medex.txt", header=TRUE )
> str( med )
> pairs( med, pch=16, cex=0.8, gap=0 )
```

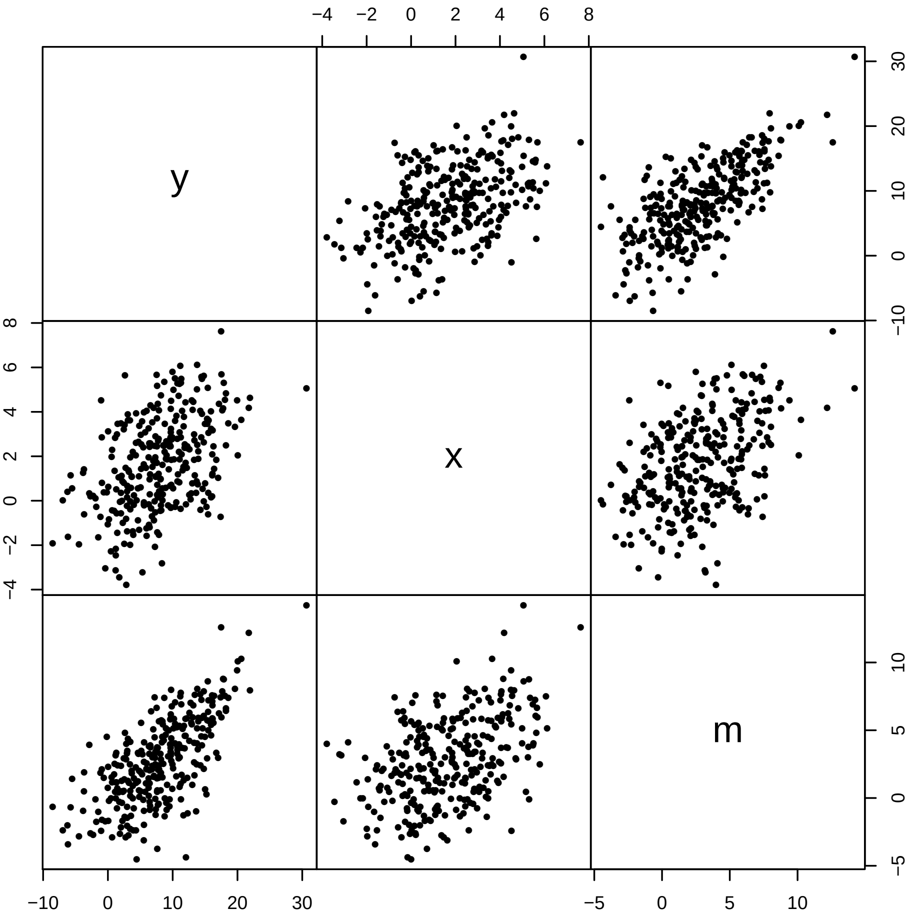


Figure 2.2: *Pairwise relationship between the simulated variables.*

1. There is nothing in the set-up that assumes that the variables y, x and m be measured on the same scale. Now, describe what units (scale) the regression parameters in the equations above have, and verify that all calculations can be done using the correct units — that is that it all fits together in terms of units.

2. First check whether mediation is likely:

   (a) Is there a relationship between y and x?

   ```
   > summary( lm( y ~ x, data=med ) )$coef
   ```

   (b) Is there a relationship between m and x?

   ```
   > summary( lm( m ~ x, data=med ) )$coef
   ```

   (c) Is there a relationship between y and x, when controlling for m?

   ```
   > summary( lm( y ~ x + m, data=med ) )$coef
   ```

3. How much of the effect of x on y is mediated through m ? Look at the marginal relationship of y and x and see if the parameter estimates look as expected.

4. Calculate the variance to expect in the marginal relationship and compare to the actual obtained.

5. What is the conclusion?

References

[1] Vamvakas et al. Renal cell cancer correlated with occupational exposure to trichlorethene. *J Cancer Res Clin Oncol*, pages 374–382, 1998.

[2] I. Kristensen, P. Aaby, and H. Jensen. Routine vaccinations and child survival: follow up study in Guinea-Bissau, West Africa. *BMJ*, 321(7274):1435–1438, Dec 2000.

[3] A. Østerlind. Malignant melanoma in Denmark. Occurrence and risk factors. *Acta Oncol*, 29(7):833–854, 1990.

[4] T. Westergaard, P. K. Andersen, J. B. Pedersen, M. Frisch, J. H. Olsen, and M. Melbye. Testicular cancer risk and maternal parity: a population-based cohort study. *Br. J. Cancer*, 77(7):1180–1185, Apr 1998.

Chapter 3

Solutions with SAS

The SAS-programs are available on the course web site in the folder http://BendixCarstensen.com/EpiPhD/F2015/sas. There is also a link to this on the website.

SAS

SAS is the default programming language in this course. It is a big package with many capabilities, but a bit clumsy for simple calculations. It is expensive, but PhD-students can have a free copy through KU, but only for the duration of your PhD.

The output from SAS comes in two different files, which makes it difficult to make a safe documentation of results, and always makes the documentation hard to follow because two different files have to be read in parallel. The solutions here consists only of the program code, not of listinngs of the `.log` (log window) and `.lst` (ouptput window) as this would be too extensive.

3.1 Vaccinations and childhood mortality in Guinea-Bissau

In rural Guinea-Bissau, 5274 children under 7 months of age were visited two times at home with an interval of six months (Kristensen et al., *BMJ*, 2000, [2]). Information about vaccination (BCG, DTP, measles vaccine) was collected at each visit, and at the second visit, death during follow-up was registered. Some children moved away during follow-up, some survived until the next visit. The following variables are found in the data set `bissau.txt`:

`id`	Id number
`fuptime`	Follow-up time in days
`dead`	0 = censored, 1 = dead
`bcg`	1 = Yes, 2 = No
`dtp`	Number of DTP doses (0,1,2,3)
`age`	Age at first visit in days
`agem`	Age at first visit in months

The following SAS-programs does all the calculations required in the exercises.

3.1.1 A single risk, odds and rate

Tabulate the number of children is 5274, the number of deaths 222 and the number of person-years 2409.8 (namely 880187 days)

- Following the lectures we get

  1. The overall risk of death is 222/5274=4.21%. A naïve 95% confidence inerval for this is:

     $$p \pm 1.96\sqrt{p \times (1-p)/n} = 0.0421 \pm 1.96\sqrt{0.0421 \times 0.9579/5274} = (0.0367; 0.0475),$$

     but a better one is the formula:

     $$\frac{p}{p+(1-p) \overset{\times}{\div} \text{erf}}, \quad \text{erf} = \exp\left(1.96\sqrt{1/x + 1/(n-x)}\right)$$

     Which gives:

     $$\text{erf} = \exp(1.96\sqrt{1/222 + 1/5052} = 1.144$$

     and so the c.i.:

     $$\frac{0.0421}{0.0421 + 0.9579 \overset{\times}{\div} 1.144} = (0.0370; 0.0479)$$

  2. The overall odds of death is simply:

     $$\frac{222}{5274-222} = 0.0439$$

     and the s.e. on the log-scale is is used to compute the 95% c.i.:

     $$\text{erf} = \exp\left(1.96\sqrt{1/222 + 1/5052}\right) = 1.144$$

     so we get:

     $$0.0439 \overset{\times}{\div} 1.144 = c(0.0384, 0.0502)$$

  3. The overall *rate* of death (per year) is

     $$222/2409.8 = 0.0921$$

     and the error factor is $\exp(1.96/sqrtD) = 1.141$ (with $D = 222$), so the confidence interval is:

     $$0.0921 \overset{\times}{\div} 1.141 = (0.0807, 0.1050)$$

- Using your statistical package, you get (almost) the same confidence intervals, the programs are:

The SAS-program is in `http://BendixCarstensen.com/EpiPhD/F2015/sas` as `bissau-sol0.sas`.

```
data bissau;
  *filename bisfile url "http://www.biostat.ku.dk/~pka/epidata/bissau.txt";
  filename bisfile "../data/bissau.txt" ;
  infile bisfile firstobs=2;
  input id fuptime dead bcg dtp age agem;
  lpy = log( fuptime/36525 ) ;
run;

title "All children in study" ;
proc means data=bissau  nway ;
    var dead fuptime ;
    output out=bcgsum
           sum= ;
run;

proc print data=bcgsum ;
run ;

title "Use log-link to produce log-probability" ;
proc genmod  data = bcgsum ;
  model dead/_freq_ = / dist=bin link=log ;
  estimate "prob" intercept 1 / Exp ;
run ;
proc genmod  data = bcgsum ;
  model dead = / dist=bin link=log ;
  estimate "prob" intercept 1 / Exp ;
run ;

title "Use logit-link to produce log-odds" ;
proc genmod  data = bcgsum ;
  model dead/_freq_ = / dist=bin link=logit ;
  estimate "odds" intercept 1 / Exp ;
run ;
proc genmod  data = bissau ;
  model dead = / dist=bin link=logit ;
  estimate "odds" intercept 1 / Exp ;
run ;

data bcgsum ;
  set bcgsum ;
  * We want rates per 100 person-years ;
  lpy = log( fuptime/36525 ) ;
run ;

title "Poisson model to derive rate" ;
proc genmod  data = bcgsum ;
  model dead = / dist=poisson link=log offset=lpy ;
  estimate "rate" intercept 1 / Exp ;
run ;
proc genmod  data = bissau ;
  model dead = / dist=poisson link=log offset=lpy ;
  estimate "rate" intercept 1 / Exp ;
run ;
```

3.1.2 Rates, risks and odds

Program is in `http://BendixCarstensen.com/EpiPhD/F2015/sas`as `bissau-sol1.sas`.

```
data bissau;
  *filename bisfile url "http://www.biostat.ku.dk/~pka/epidata/bissau.txt";
  filename bisfile "../data/bissau.txt" ;
  infile bisfile firstobs=2;
  input id fuptime dead bcg dtp age agem;
  dtpany=1; if dtp>0 then dtpany=2;
run;

title "Bissau data" ;
proc print data = bissau (obs=25) ;
run ;

title "Analysis by BCG groups" ;
proc means data=bissau  nway ;
    class bcg;
    var dead fuptime ;
    output out=bcgsum
           sum= ;
run;

proc print data=bcgsum ;
run ;

data bcgres;
  set bcgsum ;
/* Q1: We take the results from PROC MEANS (BCG=1) */
d=dead;
n=_freq_;
```

```
y=fuptime;

/* Q2: naive CI for pi */
pi = d/n;
sdpi = sqrt(pi*(1-pi)/n);
pilow = pi - 1.96*sdpi;
piup  = pi + 1.96*sdpi;

/* Q3: odds with CI + improved CI for pi*/
omega = pi/(1-pi);
sdlogomega = sqrt(1/d+1/(n-d));
errorfact  = exp(1.96*sdlogomega);
omegalow   = omega / errorfact;
omegaup    = omega * errorfact;
pilow_2    = omegalow/(1+omegalow);
piup_2     = omegaup /(1+omegaup );

/* Q4: rate per day - note that we added fuptime in PROC MEANS above */
lambda=d/y;
errorfact_rate = exp(1.96*sqrt(1/D));
lambda_low     = lambda / errorfact_rate;
lambda_up      = lambda * errorfact_rate;

/* Q5: rate per year is rate per day times 365.25 */
lambda_year     = lambda    *365.25;
lambda_year_low = lambda_low*365.25;
lambda_year_up  = lambda_up *365.25;
run;

proc print data=bcgres; run;

/* Q6: we repeat everything using the DTPANY variable created in the first DATA step */
title "Analaysis by DTP groups" ;
proc means data=bissau  nway ;
    class dtpany;
    var dead fuptime ;
    output out=dtpsum
           sum= ;
run;

proc print data=dtpsum ;
run ;

data dtpres;
  set dtpsum ;
/* Q1: We take the results from PROC MEANS (BCG=1) */
d=dead;
n=_freq_;
y=fuptime;

/* Q2: naive CI for pi */
pi = d/n;
sdpi = sqrt(pi*(1-pi)/n);
pilow = pi - 1.96*sdpi;
piup  = pi + 1.96*sdpi;

/* Q3: odds with CI + improved CI for pi*/
omega = pi/(1-pi);
sdlogomega = sqrt(1/d+1/(n-d));
errorfact  = exp(1.96*sdlogomega);
omegalow   = omega / errorfact;
omegaup    = omega * errorfact;
pilow_2    = omegalow/(1+omegalow);
piup_2     = omegaup /(1+omegaup );

/* Q4: rate per day - note that we added fuptime in PROC MEANS above */
lambda=d/y;
errorfact_rate = exp(1.96*sqrt(1/D));
lambda_low     = lambda / errorfact_rate;
lambda_up      = lambda * errorfact_rate;

/* Q5: rate per year is rate per day times 365.25 */
lambda_year     = lambda    *365.25;
lambda_year_low = lambda_low*365.25;
lambda_year_up  = lambda_up *365.25;
run;

proc print data=dtpres; run;
```

3.1.3 Rate ratio, risk ratio, odds ratio

Program is in http://BendixCarstensen.com/EpiPhD/F2015/sasas bissau-sol2.sas.

```
data bissau;
  filename bisfile url "http://www.biostat.ku.dk/~pka/epidata/bissau.txt";
  filename bisfile "../data/bissau.txt" ;
  infile bisfile firstobs=2;
  input id fuptime dead bcg dtp age agem;
  dead2=dead;
  if dead2=0 then dead2=2;
  if dtp=0 then dtpany=2;
  if dtp>0 then dtpany=1;
run;

/***********************************************************************************/
/* Q1 Effect of BCG on mortality */
title "BCG EFFECT";
proc freq data=bissau;
  table bcg*dead2 / nocol nopercent relrisk;
run;

/***********************************************************************************/
/* Q2 Effect of DTP on mortality */
title "DTP EFFECT";
proc freq data=bissau;
  table dtpany*dead2 / nocol nopercent relrisk;
run;

/***********************************************************************************/
/* Q3 Association between BCG and DTP */
title "ASSOCIATION BETWEEN DTP AND BCG";
proc freq data=bissau;
  table dtpany*bcg / chisq;
run;

/***********************************************************************************/
/* Q4 Effect of DTP on mortality for each value (level) of BCG */
title "DTP EFFECT among BCG vaccinated";
proc freq data=bissau;
  where bcg=1;
  table dtpany*dead2 / nocol nopercent relrisk;
run;

title "DTP EFFECT among BCG UN-vaccinated";
proc freq data=bissau;
  where bcg=2;
  table dtpany*dead2 / nocol nopercent relrisk;
run;

/* MUCH EASIER - YOU CAN ADD BCG IN THE THIS WAY: */
title "DTP EFFECT";
proc freq data=bissau;
  table bcg*dtpany*dead2 / nocol nopercent relrisk;
run;

/***********************************************************************************/
/* Q5 RATES */
/* We need no. of deaths and the sum of fullow-up time for each value of BCG: */
title "RATE RATIOS: BCG EFFECT";
proc means data=bissau sum;
  class bcg;
  var dead fuptime;
run;
/* Now we calculate by hand the rate ratio with 95%-CI: */
data rr;
  rate1 = (125/554929);
  rate2 = ( 97/325258);
  rr = rate1/rate2;
  sd = sqrt((1/125) + (1/97));
  error_factor = exp(1.96*sd);
  lower = rr/error_factor;
  upper = rr*error_factor;
run;
proc print data=rr;
  var rr lower upper error_factor ;
run;

title "RATE RATIOS: DTP EFFECT";
proc means data=bissau sum;
  class dtpany;
  var dead fuptime;
run;

/* Now we calculate by hand the rate ratio with 95%-CI: */
data rr;
  rate1 = (94/364012);
  rate2 = (128/516175);
  rr = rate1/rate2;
  sd = sqrt((1/94) + (1/128));
  error_factor = exp(1.96*sd);
  lower = rr/error_factor;
  upper = rr*error_factor;
run;
proc print data=rr;
  var rr lower upper error_factor ;
```

```
run;

title "RATE RATIOS: DTP EFFECT BY BCG STATUS";
proc means data=bissau sum;
  class bcg dtpany;
  var dead fuptime;
run;

/* Now we calculate by hand the rate ratio with 95%-CI: */
data rr;
  bcg="YES";
  rate1 = (92/358571);
  rate2 = (33/196358);
  rr = rate1/rate2;
  sd = sqrt((1/92) + (1/33));
  error_factor = exp(1.96*sd);
  lower = rr/error_factor;
  upper = rr*error_factor;
  output;
  BCG="NO";
  rate1 = (2/5441);
  rate2 = (95/319817);
  rr = rate1/rate2;
  sd = sqrt((1/2) + (1/95));
  error_factor = exp(1.96*sd);
  lower = rr/error_factor;
  upper = rr*error_factor;
  output;
run;

proc print data=rr;
  var bcg rr lower upper error_factor ;
run;
```

3.1.4 Confounder control: stratified analysis of odds ratio and risk ratio.

Program is in http://BendixCarstensen.com/EpiPhD/F2015/sasas bissau-sol3.sas.

```
/*************************************************************************************/
/* Read the data and transform */

data bissau;
* filename bisfile url "http://www.biostat.ku.dk/~pka/epidata/bissau.txt" ;
  filename bisfile "../data/bissau.txt" ;
  infile bisfile firstobs=2;
  input id fuptime dead bcg dtp age agem;
  dead2=dead;
  if dead2=0 then dead2=2;
  if dtp=0 then dtpany=2;
  if dtp>0 then dtpany=1;
run;

/*************************************************************************************/
/* Q1 Effect of DTP on mortality by bcg */
title "DTP EFFECT for each level of bcg";
proc freq data=bissau;
  where bcg eq 1 ;
  table dtpany*dead2 / nocol nopercent relrisk;
run;
proc freq data=bissau;
  where bcg eq 2 ;
  table dtpany*dead2 / nocol nopercent relrisk;
run;

/*************************************************************************************/
/* Q2 Effect of DTP on mortality controlled for BCG */
title "DTP EFFECT controlled for BCG";
proc freq data=bissau;
  table bcg*dtpany*dead2 / nocol nopercent relrisk cmh ;
run;

/* Compare with bcg-UNadjusted effect of DTP */
title "DTP EFFECT NOT controlled for BCG";
proc freq data=bissau;
  table dtpany*dead2 / nocol nopercent relrisk ;
run;

/*************************************************************************************/
/* Q3 Effect of DTP on mortality controlled for age in months */
title "DTP EFFECT controlled for AGE";
proc freq data=bissau;
  table agem*dtpany*dead2 / nocol nopercent relrisk cmh ;
```

```
run;

/**************************************************************************************/
/* Q4 Effect of DTP on mortality controlled for age in months AND bcg */
title "DTP EFFECT controlled for AGE and BCG";
proc freq data=bissau;
  table bcg*agem*dtpany*dead2 / nocol nopercent relrisk cmh ;
run;

/**************************************************************************************/
/* Qx Effect of DTP on mortality controlled for age in months AND bcg
   using logistic regression */
title "DTP EFFECT controlled for AGE and BCG - logistic regression";
proc genmod data=bissau;
  class bcg agem dtpany ;
  model dead2 = bcg agem dtpany
        / dist=binomial  link=logit ;
  estimate "OR by dtnm / agecont" dtpany 1 -1 / exp ;
run;

/* Using agem as continuoius (linear) */
title "DTP EFFECT controlled for AGE (continuous) and BCG - logistic regression";
proc genmod data=bissau;
  class bcg dtpany ;
  model dead2 = bcg agem dtpany
        / dist=binomial  link=logit ;
  estimate "OR by dtnm / agecont" dtpany 1 -1 / exp ;
run;

/* Computing the RR effect by using the log-link */
title "DTP EFFECT controlled for AGE and BCG - logistic regression";
proc genmod data=bissau;
  class bcg agem dtpany ;
  model dead2 = bcg agem dtpany
        / dist=binomial  link=log ;
  estimate "OR by dtnm / agecont" dtpany 1 -1 / exp ;
run;
```

3.1.5 Survival analysis of childhood mortality in Guinea-Bissau

The SAS program `bissau.sas` in http://BendixCarstensen.com/EpiPhD/F2015/sas reads the data from the web and defines / recodes relevant variables and fits a simple Cox regression model with follow-up time as the time variable and including `bcg` and `agem` as categorical covariates.

1. Fit a simple Cox regression model with follow-up time as the time variable and including `bcg` and `agem` as categorical covariates. In the program we have just recoded `bcg` to a 0/1 variable.

2. Estimate the effect of any dose of DTP, using the created variable `dtpany` adjusted only for age in months as a categorical (`class` in SAS) variable.

3. Now, also adjust for BCG. We then find a positive risk associated with DTP.

4. Is there an interaction between DTP (`dtpany`) and BCG?

   No, there is no interaction. From the table made by `proc tabulate` we see that there are virtually no children with a DTP vaccination who is without BCG vaccination. And we see that the mortality is very low among the second smallest group, those with BCG and no DTP. So the strange results hinges on the fact that almost 4000 of the 5000 in the study eiter are vaccinated by both or by none of the two.

   We see that there is an excess risk associated with DTP of 1.3 both for BCG=0 and for BCG=1, but neither are significant.

5. Make a Cox regression analysis with DTP (`dtpany`) and BCG, but now with age as time-variable, i.e. with delayed entry.

We see that we get pretty much the same results as when using the age at entry as control variable.

6. Repeat the Poisson and logistic regression models that you have seen during the lectures.

   We see that the three types of analysis gives witually the same rsults for the protective effect of BCG-vaccination namely a mortality RR of 0.71, just significant.

   With rather short follow-up time, and hence little scope for censoring and variation of rates over time, the three different models are virtually the same.

   The Poisson model where we use the log of the persons-years as offset is a model where we assume that the mortality is constant throughout follow-up (as opposed to the Cox-model where it is allowed to vary witout any restrictions). The Logistic regression model is a further sinmplification of the Poisson model where we ignore censoring, so essentially assume that everyone is followd or the same time.

Program is in `http://BendixCarstensen.com/EpiPhD/F2015/sas` as `bissau-solcox.sas`.

```
options ps=200 nocenter ;

data bissau;
* filename bisfile url "http://www.biostat.ku.dk/~pka/epidata/bissau.txt";
  filename bisfile "../data/bissau.txt";
  infile bisfile firstobs=2;
  input id fuptime dead bcg dtp age agem;
  * DTP - indicator ;
  dtpany = dtp>0 ;
  * BCG-indicator ;
  bcg = 2 - bcg ;
  outage = age + fuptime ;
  lfup = log(fuptime) ;
run;

proc print data=bissau (obs=10) ;
run ;

proc tabulate  data=bissau  noseps formchar="           ";
  class agem ;
  var fuptime dead ;
  table agem, n*f=5. dead*f=5. fuptime*f=comma10. ;
run ;

title "Q1: Simple analysis of bcg effect" ;
proc phreg  data = bissau ;
  class agem ;
  model fuptime * dead(0) = bcg agem / rl;
run;

title "Q2: Simple analysis of dtp effect" ;
proc phreg  data = bissau ;
  class agem ;
  model fuptime * dead(0) = dtpany agem / rl ;
run;

title "Q3: Analysis of dtp and bcg effect" ;
proc phreg  data = bissau ;
  class agem ;
  model fuptime * dead(0) = dtpany bcg agem / rl type3 ;
run;

title "Q4: Analysis of dtp and bcg effect with interaction" ;
proc phreg  data = bissau;
  class agem ;
  model fuptime * dead(0) = dtpany bcg dtpany*bcg agem / rl type3 ;
run;
title2 "Explaining the missing interaction and showing possible confounding" ;
proc tabulate data = bissau  noseps  missing  formchar=" ----   ---" ;
  class bcg dtpany dead ;
  table ( bcg all ) * dtpany,
        ( n * f=6. pctn<dead all> * f=6.1 ) * (dead all)
        / rts=15 ;
run ;

title "Q4a: Analysis of dtp effect separately for bcg Y/N" ;
```

```
proc sort data= bissau ; by bcg ; run ;
proc phreg  data = bissau;
  by bcg ;
  class agem ;
  model fuptime * dead(0) = dtpany agem / rl type3 ;
run;

title "Q5: Analysis of dtp and bcg effect using current age as timescale" ;
proc phreg  data = bissau;
  class agem ;
  model (age,outage) * dead(0) = dtpany bcg / rl type3 ;
run;

title "Q6: Analysis of bcg effect using age at entry - Cox-model" ;
proc phreg  data = bissau ;
  class agem ;
  model fuptime * dead(0) = bcg agem / rl ;
run;
title "Q6: Analysis of bcg effect using age at entry - Poisson-model" ;
proc genmod  data = bissau ;
  class agem ;
  model dead = bcg agem / dist = poisson  offset = lfup ;
  estimate "RR bcg" bcg 1 / Exp ;
run;

title "Q6: Analysis of bcg effect using age at entry - Poisson-model with the same FU for all" ;
proc genmod  data = bissau ;
  class agem ;
  model dead = bcg agem / dist = poisson ;
run;

title "Q6: Analysis of bcg effect using age at entry - logistic regression-model" ;
proc genmod  data = bissau  descending ;
  class agem ;
  model dead = bcg agem / dist = binomial ;
  estimate "RR bcg" bcg 1 / Exp ;
run;

title "QX: Logistic and Poission regression - the real cheat" ;
data xx ;
  set bissau ;
  if dead eq 0 then output ;
  if dead eq 1 then do ;
     fuptime = fuptime - 1 ; dead = 0 ; output ;
 fuptime =           1 ; dead = 1 ; output ;
 end ;
run ;

proc print data=xx (obs=10) ;
run ;

proc genmod  data = xx ;
  class agem ;
  model dead/fuptime = bcg agem / dist = binomial ;
run;
proc genmod  data = xx ;
  class agem ;
  model dead/fuptime = bcg agem / dist = binomial  link = cll;
run;
proc genmod  data = bissau ;
  class agem ;
  model dead = bcg agem / dist = poisson  offset=lfup ;
run;
```

Chapter 4

Solutions with Stata

The Stata-programs are available on the course web site in the folder http://BendixCarstensen.com/EpiPhD/F2015/stata. There is also a link to this on the website.

Stata

Stata is a commercial statistical package which is renowned for it speed. It is not as expensive as SAS. It has good capabilities for documenting analyses through log-files (derived from do-files) that provides good and readable documentation of analyses in a readable format.

4.1 Vaccinations and childhood mortality in Guinea-Bissau

In rural Guinea-Bissau, 5274 children under 7 months of age were visited two times at home with an interval of six months (Kristensen et al., *BMJ*, 2000, [2]). Information about vaccination (BCG, DTP, measles vaccine) was collected at each visit, and at the second visit, death during follow-up was registered. Some children moved away during follow-up, some survived until the next visit. The following variables are found in the data set `bissau.txt`:

`id`	Id number
`fuptime`	Follow-up time in days
`dead`	0 = censored, 1 = dead
`bcg`	1 = Yes, 2 = No
`dtp`	Number of DTP doses (0,1,2,3)
`age`	Age at first visit in days
`agem`	Age at first visit in months

4.1.1 A single risk, odds and rate

Tabulate the number of children is 5274, the number of deaths 222 and the number of person-years 2409.8 (namely 880187 days)

- Following the lectures we get

  1. The overall risk of death is 222/5274=4.21%. A naïve 95% confidence inerval for this is:

     $$p \pm 1.96\sqrt{p \times (1-p)/n} = 0.0421 \pm 1.96\sqrt{0.0421 \times 0.9579/5274} = (0.0367; 0.0475),$$

     but a better one is the formula:

     $$\frac{p}{p + (1-p) \overset{\times}{\div} \mathrm{erf}}, \quad \mathrm{erf} = \exp\left(1.96\sqrt{1/x + 1/(n-x)}\right)$$

     Which gives:

     $$\mathrm{erf} = \exp(1.96\sqrt{1/222 + 1/5052} = 1.144$$

     and so the c.i.:

     $$\frac{0.0421}{0.0421 + 0.9579 \overset{\times}{\div} 1.144} = (0.0370; 0.0479)$$

  2. The overall odds of death is simply:

     $$\frac{222}{5274 - 222} = 0.0439$$

     and the s.e. on the log-scale is is used to compute the 95% c.i.:

     $$\mathrm{erf} = \exp\left(1.96\sqrt{1/222 + 1/5052}\right) = 1.144$$

     so we get:

     $$0.0439 \overset{\times}{\div} 1.144 = c(0.0384, 0.0502)$$

  3. The overall *rate* of death (per year) is

     $$222/2409.8 = 0.0921$$

     and the error factor is $\exp(1.96/sqrtD) = 1.141$ (with $D = 222$), so the confidence interval is:

     $$0.0921 \overset{\times}{\div} 1.141 = (0.0807, 0.1050)$$

- Using your statistical package, you get (almost) the same confidence intervals, the programs are:

The Stata-program is in `http://BendixCarstensen.com/EpiPhD/F2015/stata` as `bissau-sol0.do`.

The following `Stata`-programs do all the calculations required in the questions for the other exercsies on the Bissau-data:

4.1.2 Rates, risks and odds

4.1.3 Rate ratio, risk ratio, odds ratio

```
use "E:\Epidemiologi\bissau.dta", replace

*The risk of death

* TO USE THE EPITAB COMMANDS WE NEED EXPOSURE VARIABLES CODED AS 0 or 1 AND
* OUTCOME CODED AS 0 or 1

gen bcg01=bcg
replace bcg01=0 if bcg==2

*QUESTION 2.1.1. 1 and 2
ci dead if bcg==1, binomial wald

ci dead if bcg==2, binomial wald


*QUESTION 2.1.1.3

tabodds dead bcg01

*QUESTION 2.1.1.4

ci dead if bcg==1, exposure(fuptime)
ci dead if bcg==2, exposure(fuptime)

*QUESTION 2.1.1.5
gen year=fuptime/365.25

ci dead if bcg==1, exposure(year)
ci dead if bcg==2, exposure(year)

*QUESTION 2.1.1.6
gen dtp_any=1 if dtp>0
replace dtp_any=0 if dtp==0

*RISK
ci dead if dtp_any==1, binomial wald

ci dead if dtp_any==0, binomial wald

*ODDS
tabodds dead dtp_any

*RATES

ci dead if dtp_any==1, exposure(fuptime)
ci dead if dtp_any==0, exposure(fuptime)

ci dead if dtp_any==1, exposure(year)
ci dead if dtp_any==0, exposure(year)


***************************
* 2.1.2                   *
***************************

*RISK RATIO
*VARIABLES MUST BE 0/1

cs dead bcg01

cc dead bcg01

*QUESTION 2.1.2.2

cs dead dtp_any

cc dead dtp_any

*QUESTION 2.1.2.3

tab bcg01 dtp_any, chi2 expected

*QUESTION 2.1.2.4

cs dead dtp_any if bcg01==1
cs dead dtp_any if bcg01==0

cc dead dtp_any if bcg01==1
cc dead dtp_any if bcg01==0

cc dead dtp_any, by(bcg01)
mhodds dead dtp_any bcg01
```

```
*QUESTION 2.1.2.5

ir dead bcg01 fuptime

ir dead dtp_any fuptime

ir dead dtp_any  fuptime if bcg01==1
ir dead dtp_any  fuptime if bcg01==0
```

4.1.4 Confounder control: stratified analysis of odds ratio and risk ratio.

4.1.5 Survival analysis of childhood mortality in Guinea-Bissau

The Stata program `bissau-cox.do` in `http://BendixCarstensen.com/EpiPhD/F2015/stata` reads the data from the web and defines / recodes relevant variables and fits Cox regression models with follow-up time as the time variable and including `bcg` and `agem` as categorical covariates.

```
use "C:\ewan\Epidemiologi\Data\bissau.dta", clear


*QUESTION 1
*GETTING READY TO DO A SURVLVAL ANALYSIS

stset fuptime, failure(dead==1)
*NEW VARIABLES ARE CREATED

*COX MODEL

stcox i.bcg i.agem

*THE HAZARD OF DEATH IS 42% HIGHER IN SUBJECTS WITHOUT A BCG VACCINE COMPARED TO
*SUBJECTS WITH THE VACCINE ADJUSTED FOR AGE.


*GETTING THE PARAMETRISATION FROM THE SLIDES:

stcox b2.bcg b6.agem, nohr

stcox b2.bcg b6.agem

*QUESTION 2
gen dtpany=0 if dtp==0
replace dtpany=1 if dtp>0

stcox i.dtpany i.agem

*QUESTION 3
*ADJUST FOR BCG

stcox i.dtpany i.agem b2.bcg
*THE EFFECT OF DTAANY INCREASES

*QUESTION 4
*INTERACTION BETWEEN BCG AND DTPANY?

stcox i.dtpany i.agem b2.bcg i.dtpany#b2.bcg
*THE WALD TEST FOR THE INTERACTION IS NOT SIGNIFICANT
est store m1

stcox i.dtpany i.agem b2.bcg
est store m2
lrtest m1 m2
*LIKELIHOOD RATIO NOT SIGNIFICANT

*QUESTION 5
*USE AGE AS TIME SCALE


gen outage=age+fuptime

stset outage, failure(dead==1) enter(age)

stcox i.dtpany b2.bcg

*QUESTION 6
```

```
*REPEAT ANALYSES FROM LECTURES
stset fuptime, failure(dead==1)

stcox b2.bcg i.agem

poisson dead b2.bcg i.agem, exposure(fuptime) ir

logistic dead b2.bcg i.agem
```

Chapter 5

Solutions with R

The R-programs are available on the course web site in the folder http://BendixCarstensen.com/EpiPhD/F2015/r. There is also a link to this on the website.

R

R is a free statistics package, which has become the default computing tool a large part of the statisticians of the world. It is dominant in bioinformatics. It is particularly useful for its excellent and versatile graphics. As can be seen from the first few solutions, it can also be used as a mere desk top calculatior.

R comes with a versatil documentation system for analyses and results, Rweave, which is used for these solutions. It provides a way of documention reproducible research, which is widely used, particularly in bioinformatics.

R can be expanded by downloading additional packages, of which there are currently about 3000. The relevant site is CRAN, the Comprehensive R Archive Network, http://cran.r-project.org/.

5.1 Vaccinations and childhood mortality in Guinea-Bissau

In rural Guinea-Bissau, 5274 children under 7 months of age were visited two times at home with an interval of six months (Kristensen et al., *BMJ*, 2000, [2]). Information about vaccination (BCG, DTP, measles vaccine) was collected at each visit, and at the second visit, death during follow-up was registered. Some children moved away during follow-up, some survived until the next visit. The following variables are found in the data set `bissau.txt`:

`id`	Id number
`fuptime`	Follow-up time in days
`dead`	0 = censored, 1 = dead
`bcg`	1 = Yes, 2 = No
`dtp`	Number of DTP doses (0,1,2,3)
`age`	Age at first visit in days
`agem`	Age at first visit in months

5.1.1 A single risk, odds and rate

Tabulate the number of children is 5274, the number of deaths 222 and the number of person-years 2409.8 (namely 880187 days)

- Following the lectures we get

  1. The overall risk of death is 222/5274=4.21%. A naïve 95% confidence inerval for this is:

     $p \pm 1.96\sqrt{p \times (1-p)/n} = 0.0421 \pm 1.96\sqrt{0.0421 \times 0.9579/5274} = (0.0367; 0.0475),$

     but a better one is the formula:

     $$\frac{p}{p + (1-p) \overset{\times}{\div} \mathrm{erf}}, \quad \mathrm{erf} = \exp\left(1.96\sqrt{1/x + 1/(n-x)}\right)$$

     Which gives:

     $$\mathrm{erf} = \exp(1.96\sqrt{1/222 + 1/5052} = 1.144$$

     and so the c.i.:

     $$\frac{0.0421}{0.0421 + 0.9579 \overset{\times}{\div} 1.144} = (0.0370; 0.0479)$$

  2. The overall odds of death is simply:

     $$\frac{222}{5274 - 222} = 0.0439$$

     and the s.e. on the log-scale is is used to compute the 95% c.i.:

     $$\mathrm{erf} = \exp\left(1.96\sqrt{1/222 + 1/5052}\right) = 1.144$$

     so we get:

     $$0.0439 \overset{\times}{\div} 1.144 = c(0.0384, 0.0502)$$

  3. The overall *rate* of death (per year) is

     $$222/2409.8 = 0.0921$$

     and the error factor is $\exp(1.96/sqrtD) = 1.141$ (with $D = 222$), so the confidence interval is:

     $$0.0921 \overset{\times}{\div} 1.141 = (0.0807, 0.1050)$$

- Using your statistical package, you get (almost) the same confidence intervals, the programs are:

The R-program is in http://BendixCarstensen.com/EpiPhD/F2015/r as `bissau-sol0.R`.

Reading and Tabulating the dataset:

```
> bis <- read.table( "../data/bissau.txt", header=TRUE )
> N <- nrow( bis )
> D <- sum( bis$dead )
> Y <- sum( bis$fuptime/365.25 )
> cbind( N, D, Y )
          N   D        Y
  [1,] 5274 222 2409.821
```

shows the number of children is 5274, the number of deaths 222 and the number of person-years 2409.8 (namely 880187 days)

- Following the lectures we get

1. The overall risk of death is 222/5274=4.21%. A naive 95% confidence interval for this is:

$$p \pm 1.96\sqrt{p \times (1-p)/n} = 0.0421 \pm 1.96\sqrt{0.0421 \times 0.9579/5274} = (0.0367; 0.0475),$$

```
> p <- D/N
> se <- sqrt( p*(1-p)/N )
> round( cbind( p, lo=p-1.96*se, hi=p+1.96*se ), 4 )
            p     lo     hi
  [1,] 0.0421 0.0367 0.0475
```

But a better one is the formula:

$$\frac{p}{p+(1-p)\overset{\times}{\div}\text{erf}}, \quad \text{erf} = \exp\left(1.96\sqrt{1/x+1/(n-x)}\right)$$

Which gives:

$$\text{erf} = \exp(1.96\sqrt{1/222+1/5052} = 1.144$$

and so the c.i.:

$$\frac{0.0421}{0.0421+0.9579 \overset{\times}{\div} 1.144} = (0.0370; 0.0479)$$

```
> erf <- exp( 1.96*sqrt(1/D+1/(N-D)) )
> round( cbind( p, lo=p/(p+(1-p)*erf),
+                  hi=p/(p+(1-p)/erf) ), 4 )
            p    lo     hi
  [1,] 0.0421 0.037 0.0479
```

2. The overall odds of death is simply:

$$\frac{222}{5274-222} = 0.0439$$

and the s.e. on the log-scale is is used to compute the 95% c.i., it is the same error factor as before:

$$\text{erf} = \exp\left(1.96\sqrt{1/222 + 1/5052}\right) = 1.144$$

so we get:

$$0.0439 \overset{\times}{\div} 1.144 = c(0.0384, 0.0503)$$

```
> odds <- D/(N-D)
> round( cbind( odds, lo=odds/erf, hi=odds*erf ), 4 )
       odds     lo     hi
[1,] 0.0439 0.0384 0.0503
```

3. The overall *rate* of death (per year) is

$$222/2409.8 = 0.0921$$

and the error factor is $\exp(1.96/sqrtD) = 1.141$ (with $D = 222$), so the confidence interval is:

$$0.0921 \overset{\times}{\div} 1.141 = (0.0807, 0.1050)$$

```
> rate <- D/Y
> erf <- exp(1.96/sqrt(D))
> round( cbind( rate, lo=rate/erf, hi=rate*erf ), 4 )
       rate     lo     hi
[1,] 0.0921 0.0808 0.1051
```

- Using the modelling we can get the same. Although it seems like bringing coal to Newcastle, there is sense in this, because we get some code which is generalizable:

  1. A single proportion can be modelled in a binomial model with log-link and subsequently using `ci.exp` to fish out and exponentiate the result. We can either use the tabulated numbers or the entire data set. Note that when we use tabulated data we must put the response in as a two-column matrix with dead and non-dead, using `cbind`:

```
> library( Epi )
> summary( m0 <-glm( cbind(D,N-D) ~ 1, family=binomial(link=log) ) )
Call:
glm(formula = cbind(D, N - D) ~ 1, family = binomial(link = log))

Deviance Residuals:
[1]  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.16787    0.06569  -48.23   <2e-16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 0  degrees of freedom
Residual deviance: 9.8588e-14  on 0  degrees of freedom
AIC: 9.1983

Number of Fisher Scoring iterations: 3
```

```
> round( ci.exp( m0 ), 4 )
              exp(Est.)  2.5%  97.5%
  (Intercept)    0.0421 0.037 0.0479
> summary( l0 <-glm( dead ~ 1, family=binomial(link=log), data=bis ) )
  Call:
  glm(formula = dead ~ 1, family = binomial(link = log), data = bis)

  Deviance Residuals: 
      Min       1Q   Median       3Q      Max  
  -0.2933  -0.2933  -0.2933  -0.2933   2.5171  

  Coefficients:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept) -3.16787    0.06568  -48.23   <2e-16

  (Dispersion parameter for binomial family taken to be 1)

      Null deviance: 1841.1  on 5273  degrees of freedom
  Residual deviance: 1841.1  on 5273  degrees of freedom
  AIC: 1843.1

  Number of Fisher Scoring iterations: 6
> round( ci.exp( l0 ), 4 )
              exp(Est.)  2.5%  97.5%
  (Intercept)    0.0421 0.037 0.0479
```

2. The same goes for the odds, now we just use the default link function which is the logit, and so exponentiation of the estimate (the intercept) will be the odds:

```
> summary( m1 <-glm( cbind(D,N-D) ~ 1, family=binomial ) )
  Call:
  glm(formula = cbind(D, N - D) ~ 1, family = binomial)

  Deviance Residuals: 
  [1]  0

  Coefficients:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept) -3.12486    0.06857  -45.57   <2e-16

  (Dispersion parameter for binomial family taken to be 1)

      Null deviance: 0.0000e+00  on 0  degrees of freedom
  Residual deviance: 4.3165e-13  on 0  degrees of freedom
  AIC: 9.1983

  Number of Fisher Scoring iterations: 2
> round( ci.exp( m1 ), 4 )
              exp(Est.)   2.5%  97.5%
  (Intercept)    0.0439 0.0384 0.0503
> summary( l1 <-glm( dead ~ 1, family=binomial, data=bis ) )
  Call:
  glm(formula = dead ~ 1, family = binomial, data = bis)

  Deviance Residuals: 
      Min       1Q   Median       3Q      Max  
  -0.2933  -0.2933  -0.2933  -0.2933   2.5171  

  Coefficients:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept) -3.12486    0.06857  -45.57   <2e-16
  (Dispersion parameter for binomial family taken to be 1)

      Null deviance: 1841.1  on 5273  degrees of freedom
  Residual deviance: 1841.1  on 5273  degrees of freedom
  AIC: 1843.1

  Number of Fisher Scoring iterations: 6
> round( ci.exp( l1 ), 4 )
              exp(Est.)   2.5%  97.5%
  (Intercept)    0.0439 0.0384 0.0503
```

3. The likelihood for a constant rate looks like a likelihood for a poisson variate, so we can use the Poisson family to estimate a single rate:

```
> summary( m2 <-glm( D ~ 1, family=poisson, offset=log(Y) ) )
  Call:
  glm(formula = D ~ 1, family = poisson, offset = log(Y))

  Deviance Residuals:
  [1]  0

  Coefficients:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept) -2.38463    0.06712  -35.53   <2e-16

  (Dispersion parameter for poisson family taken to be 1)

      Null deviance: 1.3767e-14  on 0  degrees of freedom
  Residual deviance: 1.3767e-14  on 0  degrees of freedom
  AIC: 9.2413

  Number of Fisher Scoring iterations: 2
> round( ci.exp( m2 ), 4 )
              exp(Est.)   2.5%  97.5%
  (Intercept)    0.0921 0.0808 0.1051
> summary( l2 <-glm( dead ~ 1, family=poisson,
+                             offset=log(fuptime/365.25), data=bis ) )
  Call:
  glm(formula = dead ~ 1, family = poisson, data = bis, offset = log(fuptime/365.25))

  Deviance Residuals:
      Min       1Q   Median       3Q      Max
  -0.3038  -0.3038  -0.3030  -0.2850   3.3151

  Coefficients:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept) -2.38463    0.06712  -35.53   <2e-16

  (Dispersion parameter for poisson family taken to be 1)

      Null deviance: 1759.2  on 5273  degrees of freedom
  Residual deviance: 1759.2  on 5273  degrees of freedom
  AIC: 2205.2

  Number of Fisher Scoring iterations: 6
> round( ci.exp( l2 ), 4 )
              exp(Est.)   2.5%  97.5%
  (Intercept)    0.0921 0.0808 0.1051
```

So we see we get the same by using the `glm` function and the classical formulae. But the `glm` machinery is easier to generalize than the classical formulae.

In the following is shown the R-commands to do all the calculations required in the questions.

5.1.2 Rates, risks and odds

1. For convenience we first load the Epi package, and the read the data — including the variable names from the first line:

```
> library( Epi )
> library( epitools )
> bissau <- read.table( "../data/bissau.txt",
+                       header=TRUE )
> str( bissau )
```

```
  'data.frame':        5274 obs. of  7 variables:
   $ id     : int  1 2 3 4 5 6 7 8 9 10 ...
   $ fuptime: int  65 161 166 166 161 161 166 166 166 166 ...
   $ dead   : int  1 0 0 0 0 0 0 0 0 0 ...
   $ bcg    : int  1 1 2 1 1 1 1 1 1 1 ...
   $ dtp    : int  1 2 0 0 0 0 2 1 2 2 ...
   $ age    : int  182 125 69 96 131 26 129 90 119 146 ...
   $ agem   : int  5 4 2 3 4 0 4 2 3 4 ...
```

   Then we need the no. onservation, no. of deaths and no. person-years for the two groups defined by `bcg`:

```
> with( subset( bissau, bcg==1 ), c(sum(dead),sum(fuptime),length(dead)) )
```

```
  [1]    125 554929   3301
```

```
> with( subset( bissau, bcg==2 ), c(sum(dead),sum(fuptime),length(dead)) )
```

```
  [1]     97 325258   1973
```

   Alternatively — some would say simpler, some would say more convoluted — we could use `xtabs`:

```
> ( xx <- xtabs( cbind(dead,fuptime,n=1) ~ bcg, data=bissau ) )
```

```
  bcg   dead fuptime      n
    1    125  554929   3301
    2     97  325258   1973
```

```
> str( xx )
```

```
   xtabs [1:2, 1:3] 125 97 554929 325258 3301 ...
   - attr(*, "dimnames")=List of 2
    ..$ bcg: chr [1:2] "1" "2"
    ..$    : chr [1:3] "dead" "fuptime" "n"
   - attr(*, "class")= chr [1:2] "xtabs" "table"
   - attr(*, "call")= language xtabs(formula = cbind(dead, fuptime, n = 1) ~ bcg, data = bissa
```

2. We then compute the fraction of dead and the confidence interval:

```
> d <- xx[,"dead"]
> n <- xx[,"n"]
> pi    <- d / n
> sdpi  <- sqrt(pi*(1-pi)/n)
> pilow <- pi - 1.96*sdpi
> piup  <- pi + 1.96*sdpi
> cbind( pi, pilow, piup )
```

```
          pi      pilow       piup
1 0.03786731 0.03135578 0.04437884
2 0.04916371 0.03962328 0.05870414
```

If we prefer the result in % and rounded, we just do:

```
> round( cbind( pi, pilow, piup )*100, 2 )

    pi pilow piup
1 3.79  3.14 4.44
2 4.92  3.96 5.87
```

3. We now compute odds with c.i. and backtransform to improved c.i. for the proportions

```
> omega      <- pi/(1-pi)
> sdlogomega <- sqrt(1/d+1/(n-d))
> errorfact  <- exp(1.96*sdlogomega)
> omegalow   <- omega / errorfact
> omegaup    <- omega * errorfact
> pilow_2    <- omegalow/(1+omegalow)
> piup_2     <- omegaup /(1+omegaup )
> round( cbind( omega, omegalow, omegaup ), 3 )

  omega omegalow omegaup
1 0.039    0.033   0.047
2 0.052    0.042   0.063

> round( cbind( pi, pilow, piup, pilow_2, piup_2 )*100, 2 )

    pi pilow piup pilow_2 piup_2
1 3.79  3.14 4.44    3.19   4.49
2 4.92  3.96 5.87    4.05   5.96
```

4. Now we compute rate per day - note that we added fuptime in `xtabs` above:

```
> y <- xx[,"fuptime"]
> lambda          <- d/y;
> errorfact_rate <- exp(1.96*sqrt(1/d));
> lambda_low      <- lambda / errorfact_rate;
> lambda_up       <- lambda * errorfact_rate;
> cbind( lambda, lambda_low, lambda_up )

        lambda   lambda_low    lambda_up
1 0.0002252540 0.0001890329 0.0002684156
2 0.0002982248 0.0002444082 0.0003638914
```

5. Rates per day are not interesting, so we convert them to rates per year:

```
> round( cbind( lambda, lambda_low, lambda_up )*365.25, 3 )

  lambda lambda_low lambda_up
1  0.082      0.069     0.098
2  0.109      0.089     0.133
```

6. We now repeat everything using the indicator of whether any DTP was received:

```
> ( xx <- xtabs( cbind(dead,fuptime,n=1) ~ bcg, data=bissau ) )
```

```
  bcg   dead fuptime      n
    1    125  554929   3301
    2     97  325258   1973
```

```
> str( xx )

   xtabs [1:2, 1:3] 125 97 554929 325258 3301 ...
   - attr(*, "dimnames")=List of 2
    ..$ bcg: chr [1:2] "1" "2"
    ..$    : chr [1:3] "dead" "fuptime" "n"
   - attr(*, "class")= chr [1:2] "xtabs" "table"
   - attr(*, "call")= language xtabs(formula = cbind(dead, fuptime, n = 1) ~ bcg, data = bissa
```

7. We then compute the fraction of dead and the confidence interval:

```
> d <- xx[,"dead"]
> n <- xx[,"n"]
> pi    <- d / n
> sdpi  <- sqrt(pi*(1-pi)/n)
> pilow <- pi - 1.96*sdpi
> piup  <- pi + 1.96*sdpi
> cbind( pi, pilow, piup )

          pi      pilow       piup
  1 0.03786731 0.03135578 0.04437884
  2 0.04916371 0.03962328 0.05870414
```

If we prefer the result in % and rounded, we just do:

```
> round( cbind( pi, pilow, piup )*100, 2 )

      pi pilow piup
  1 3.79  3.14 4.44
  2 4.92  3.96 5.87
```

8. We now compute odds with c.i. and backtransform to improved c.i. for the proportions

```
> omega      <- pi/(1-pi)
> sdlogomega <- sqrt(1/d+1/(n-d))
> errorfact  <- exp(1.96*sdlogomega)
> omegalow   <- omega / errorfact
> omegaup    <- omega * errorfact
> pilow_2    <- omegalow/(1+omegalow)
> piup_2     <- omegaup /(1+omegaup )
> round( cbind( omega, omegalow, omegaup ), 3 )

    omega omegalow omegaup
  1 0.039    0.033   0.047
  2 0.052    0.042   0.063
```

```
> round( cbind( pi, pilow, piup, pilow_2, piup_2 )*100, 2 )

      pi pilow piup pilow_2 piup_2
  1 3.79  3.14 4.44    3.19   4.49
  2 4.92  3.96 5.87    4.05   5.96
```

9. Now we compute rate per day - note that we added fuptime in `xtabs` above:

```
> y <- xx[,"fuptime"]
> lambda         <- d/y;
> errorfact_rate <- exp(1.96*sqrt(1/d));
> lambda_low     <- lambda / errorfact_rate;
> lambda_up      <- lambda * errorfact_rate;
> cbind( lambda, lambda_low, lambda_up )

        lambda   lambda_low    lambda_up
1 0.0002252540 0.0001890329 0.0002684156
2 0.0002982248 0.0002444082 0.0003638914
```

10. Rates per day are not interesting, we convert them to rates per year:

```
> round( cbind( lambda, lambda_low, lambda_up )*365.25, 3 )

  lambda lambda_low lambda_up
1  0.082      0.069     0.098
2  0.109      0.089     0.133
```

11. Finally we repeat it al but now subdividing persons by wheter they have received any DTP dose or not:

```
> ( xx <- xtabs( cbind(dead,fuptime,n=1) ~ (dtp>0), data=bissau ) )

dtp > 0   dead fuptime      n
  FALSE    128  516175   3101
  TRUE      94  364012   2173

> str( xx )

 xtabs [1:2, 1:3] 128 94 516175 364012 3101 ...
 - attr(*, "dimnames")=List of 2
  ..$ dtp > 0: chr [1:2] "FALSE" "TRUE"
  ..$        : chr [1:3] "dead" "fuptime" "n"
 - attr(*, "class")= chr [1:2] "xtabs" "table"
 - attr(*, "call")= language xtabs(formula = cbind(dead, fuptime, n = 1) ~ (dtp > 0), data =

> d <- xx[,"dead"]
> n <- xx[,"n"]
> pi    <- d / n
> sdpi  <- sqrt(pi*(1-pi)/xx[,"n"])
> pilow <- pi - 1.96*sdpi
> piup  <- pi + 1.96*sdpi
> cbind( pi, pilow, piup )

              pi      pilow       piup
FALSE 0.04127701 0.03427527 0.04827875
TRUE  0.04325817 0.03470440 0.05181194
```

If we prefer the result in % and rounded, we just do:

```
> round( cbind( pi, pilow, piup )*100, 2 )

        pi pilow piup
FALSE 4.13  3.43 4.83
TRUE  4.33  3.47 5.18
```

Compute odds with c.i. and backtransform to improved c.i. for the proportions

```
> omega      <- pi/(1-pi)
> sdlogomega <- sqrt(1/d+1/(n-d))
> errorfact  <- exp(1.96*sdlogomega)
> omegalow   <- omega / errorfact
> omegaup    <- omega * errorfact
> pilow_2    <- omegalow/(1+omegalow)
> piup_2     <- omegaup /(1+omegaup )
> round( cbind( omega, omegalow, omegaup ), 3 )

      omega omegalow omegaup
FALSE 0.043    0.036   0.051
TRUE  0.045    0.037   0.056

> round( cbind( pi, pilow, piup, pilow_2, piup_2 )*100, 2 )

        pi pilow piup pilow_2 piup_2
FALSE 4.13  3.43 4.83    3.48   4.89
TRUE  4.33  3.47 5.18    3.55   5.27
```

Rate per day:

```
> y <- xx[,"fuptime"]
> lambda          <- d/y;
> errorfact_rate <- exp(1.96*sqrt(1/d));
> lambda_low      <- lambda / errorfact_rate;
> lambda_up       <- lambda * errorfact_rate;
> cbind( lambda, lambda_low, lambda_up )

            lambda   lambda_low    lambda_up
FALSE 0.0002479779 0.0002085333 0.0002948836
TRUE  0.0002582332 0.0002109676 0.0003160884
```

Rates per year:

```
> round( cbind( lambda, lambda_low, lambda_up )*365.25, 3 )

      lambda lambda_low lambda_up
FALSE  0.091      0.076     0.108
TRUE   0.094      0.077     0.115
```

5.1.3 Rate ratio, risk ratio, odds ratio

1. In order to compute odds-ratios and risk ratios, we use the twoby2 from the Epi package:

```
> ( tt <- with( bissau, table(BCG=bcg,dead) ) )

   dead
BCG    0    1
  1 3176  125
  2 1876   97

> twoby2( tt )
```

```
  2 by 2 table analysis:
  ------------------------------------------------------
  Outcome   : 0
  Comparing : 1 vs. 2

       0   1    P(0) 95% conf. interval
  1 3176 125  0.9621    0.9551   0.9681
  2 1876  97  0.9508    0.9404   0.9595

                                 95% conf. interval
               Relative Risk: 1.0119    0.9997   1.0242
           Sample Odds Ratio: 1.3137    1.0016   1.7232
  Conditional MLE Odds Ratio: 1.3136    0.9908   1.7376
      Probability difference: 0.0113    0.0001   0.0233

               Exact P-value: 0.0555
          Asymptotic P-value: 0.0487
  ------------------------------------------------------
```

— but this is the wrong way round, so we swap the outcome categories:

```
> twoby2( tt[,2:1] )

  2 by 2 table analysis:
  ------------------------------------------------------
  Outcome   : 1
  Comparing : 1 vs. 2

      1    0    P(1) 95% conf. interval
  1 125 3176  0.0379    0.0319   0.0449
  2  97 1876  0.0492    0.0405   0.0596

                                  95% conf. interval
               Relative Risk:  0.7702    0.5943   0.9982
           Sample Odds Ratio:  0.7612    0.5803   0.9984
  Conditional MLE Odds Ratio:  0.7612    0.5755   1.0093
      Probability difference: -0.0113   -0.0233  -0.0001

               Exact P-value: 0.0555
          Asymptotic P-value: 0.0487
  ------------------------------------------------------
```

So we see that the mortality is smaller among those BCG vaccinated.

But it is always more convenient to annotate variables correctly, so we turn `bcg` `dtpany` and `Dead` into factors. Note that we let the first level of the factor be the exposed:

```
> bissau <- transform( bissau, bcg = factor(bcg,levels=1:2,
+                                          labels=c("BCG","no BCG")),
+                              dtpany = factor(dtp>0,levels=c(TRUE,FALSE),
+                                            labels=c("1+ DTP","no DTP")),
+                                Dead = factor(dead,levels=0:1,
+                                           labels=c("Alive","Dead") )
+                      )
```

2. The same analysis for DTP (any dose):

```
> ( tt <- with( bissau, table(dtpany,Dead) ) )
```

```
          Dead
  dtpany   Alive Dead
    1+ DTP  2079   94
    no DTP  2973  128

> twoby2( tt )

  2 by 2 table analysis:
  ------------------------------------------------------
  Outcome   : Alive
  Comparing : 1+ DTP vs. no DTP

         Alive Dead    P(Alive) 95% conf. interval
  1+ DTP  2079   94      0.9567    0.9473   0.9645
  no DTP  2973  128      0.9587    0.9511   0.9652

                                      95% conf. interval
               Relative Risk:  0.9979    0.9865   1.0095
           Sample Odds Ratio:  0.9522    0.7254   1.2500
  Conditional MLE Odds Ratio:  0.9523    0.7195   1.2640
      Probability difference: -0.0020   -0.0134   0.0089

               Exact P-value: 0.7281
          Asymptotic P-value: 0.7244
  ------------------------------------------------------

> twoby2( tt[,2:1] )

  2 by 2 table analysis:
  ------------------------------------------------------
  Outcome   : Dead
  Comparing : 1+ DTP vs. no DTP

         Dead Alive    P(Dead) 95% conf. interval
  1+ DTP   94  2079     0.0433    0.0355   0.0527
  no DTP  128  2973     0.0413    0.0348   0.0489

                                     95% conf. interval
               Relative Risk: 1.0480    0.8076   1.3599
           Sample Odds Ratio: 1.0502    0.8000   1.3785
  Conditional MLE Odds Ratio: 1.0501    0.7911   1.3899
      Probability difference: 0.0020   -0.0089   0.0134

               Exact P-value: 0.7281
          Asymptotic P-value: 0.7244
  ------------------------------------------------------
```

We see that there is no effect of DTP on mortality; the RR is 1.05 and the c.i. is reasonably narrow: (0.81,1.36).

3. Now we look at the association of the two exposures:

```
> with( bissau, table(dtpany,bcg) )

          bcg
  dtpany    BCG no BCG
    1+ DTP 2142     31
    no DTP 1159   1942
```

We see that DTP vaccination is largely confined to those who are BCG-vaccinated. Thus it is only relevant to evaluate the DTP effect among those BCG-vaccinated, because there is no information on the DTP-effect among the non-BCG-vaccinated:

```
> ( tt <- with( subset(bissau,bcg=="no BCG"), table(dtpany,Dead) ) )

          Dead
  dtpany   Alive Dead
    1+ DTP    29    2
    no DTP  1847   95

> twoby2( tt[,2:1] )

  2 by 2 table analysis:
  ------------------------------------------------------
  Outcome   : Dead
  Comparing : 1+ DTP vs. no DTP

         Dead Alive    P(Dead) 95% conf. interval
  1+ DTP    2    29     0.0645    0.0162   0.2242
  no DTP   95  1847     0.0489    0.0402   0.0595

                                     95% conf. interval
               Relative Risk: 1.3188    0.3403   5.1114
           Sample Odds Ratio: 1.3408    0.3153   5.7028
  Conditional MLE Odds Ratio: 1.3406    0.1528   5.4347
      Probability difference: 0.0156   -0.0322   0.1585

               Exact P-value: 0.6628
          Asymptotic P-value: 0.6913
  ------------------------------------------------------
```

But among those with a BCG-vaccination there is information:

```
> ( tt <- with( subset(bissau,bcg=="BCG"), table(dtpany,Dead) ) )

          Dead
  dtpany   Alive Dead
    1+ DTP  2050   92
    no DTP  1126   33

> twoby2( tt[,2:1] )

  2 by 2 table analysis:
  ------------------------------------------------------
  Outcome   : Dead
  Comparing : 1+ DTP vs. no DTP

         Dead Alive    P(Dead) 95% conf. interval
  1+ DTP   92  2050     0.0430    0.0351   0.0524
  no DTP   33  1126     0.0285    0.0203   0.0398

                                     95% conf. interval
               Relative Risk: 1.5085    1.0201   2.2307
           Sample Odds Ratio: 1.5313    1.0221   2.2942
  Conditional MLE Odds Ratio: 1.5311    1.0110   2.3697
      Probability difference: 0.0145    0.0008   0.0269

               Exact P-value: 0.0444
          Asymptotic P-value: 0.0389
  ------------------------------------------------------
```

and we see that is a borderline significant RR=1.5 associated with DTP.

```
> ( ff <- with( bissau, ftable( dtpany, bcg, dead ) ) )
```

```
              dead    0    1
dtpany bcg
1+ DTP BCG         2050   92
       no BCG        29    2
no DTP BCG         1126   33
       no BCG      1847   95
```

```
> round( cbind( ff, ff[,2]/ff[,1]*100 ), 1 )
```

```
     [,1] [,2] [,3]
[1,] 2050   92  4.5
[2,]   29    2  6.9
[3,] 1126   33  2.9
[4,] 1847   95  5.1
```

This table shows that BCG alone is protective, but that either absence of BCG or addition of DTP increases mortality.

This analysis can be made (but only for the OR) by the `effx` function:

```
> effx( dead, type="binary", exposure=dtpany, strata=bcg, data=bissau )
```

```
---------------------------------------------------------------------------
response      :  dead
type          :  binary
exposure      :  dtpany
stratified by :  bcg

dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  1+ DTP
bcg is a factor with levels: BCG/no BCG
effects are measured as odds ratios
---------------------------------------------------------------------------

effect of dtpany on dead
stratified by bcg

number of observations  5274

                                     Effect  2.5% 97.5%
strata BCG level no DTP vs 1+ DTP     0.653 0.436 0.978
strata no BCG level no DTP vs 1+ DTP  0.746 0.175 3.170

Test for effect modification on 1 df: p-value= 0.86
```

So we see there is no evidence that DTP has differential effect, so we could BCG as a confounder instead (controlling for it:

```
> effx( dead, type="binary", exposure=dtpany, control=bcg, data=bissau )
```

```
---------------------------------------------------------------------------
response      :  dead
type          :  binary
exposure      :  dtpany
control vars  :  bcg

dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  1+ DTP
effects are measured as odds ratios
---------------------------------------------------------------------------

effect of dtpany on dead
controlled for bcg
```

```
  number of observations  5274

  Effect   2.5%  97.5%
   0.660  0.448  0.971

  Test for no effects of exposure on 1 df: p-value= 0.0313
```

4. The `effx` function allows calculation of the rate-ratios etc. very easily:

```
> effx( dead, type="failure", fup=fuptime/365.25, exposure=bcg, data=bissau )

  ---------------------------------------------------------------------------
  response      :  dead
  type          :  failure
  exposure      :  bcg

  bcg is a factor with levels: BCG / no BCG
  baseline is  BCG
  effects are measured as rate ratios
  ---------------------------------------------------------------------------

  effect of bcg on dead
  number of observations  5274

  Effect   2.5%  97.5%
    1.32   1.02   1.73

  Test for no effects of exposure on 1 df: p-value= 0.0395

> effx( dead, type="failure", fup=fuptime/365.25, exposure=dtpany, base=2, data=bissau )

  ---------------------------------------------------------------------------
  response      :  dead
  type          :  failure
  exposure      :  dtpany

  dtpany is a factor with levels: 1+ DTP / no DTP
  baseline is  no DTP
  effects are measured as rate ratios
  ---------------------------------------------------------------------------

  effect of dtpany on dead
  number of observations  5274

  Effect   2.5%  97.5%
   1.040  0.798  1.360

  Test for no effects of exposure on 1 df: p-value= 0.766

> effx( dead, type="failure", fup=fuptime/365.25, exposure=dtpany,
+             strata=bcg, base=2, data=bissau )

  ---------------------------------------------------------------------------
  response      :  dead
  type          :  failure
  exposure      :  dtpany
  stratified by :  bcg

  dtpany is a factor with levels: 1+ DTP / no DTP
  baseline is  no DTP
  bcg is a factor with levels: BCG/no BCG
  effects are measured as rate ratios
  ---------------------------------------------------------------------------
  effect of dtpany on dead
  stratified by bcg

  number of observations  5274

                                       Effect  2.5% 97.5%
  strata BCG level 1+ DTP vs no DTP      1.53 1.030  2.27
  strata no BCG level 1+ DTP vs no DTP   1.24 0.305  5.02

  Test for effect modification on 5272 df: p-value= 0.685
```

```
> effx( dead, type="failure", fup=fuptime/365.25, exposure=dtpany,
+             control=bcg, base=2, data=bissau )

  ---------------------------------------------------------------------------
  response      :  dead
  type          :  failure
  exposure      :  dtpany
  control vars  :  bcg

  dtpany is a factor with levels: 1+ DTP / no DTP
  baseline is  no DTP
  effects are measured as rate ratios
  ---------------------------------------------------------------------------

  effect of dtpany on dead
  controlled for bcg

  number of observations  5274

  Effect   2.5%  97.5%
    1.50   1.03   2.19

  Test for no effects of exposure on 1 df: p-value= 0.0316
```

The results are virtually identical to those for the relative risk, where we ignored the follow-up time.

5.1.4 Confounder control: stratified analysis of odds ratio and risk ratio

We will continue using the data from Guinea-Bissau for this third part of the exercise.

1. Revisit the analysis from previously, using just death (dead) as outcome, and estimate the DTP effect for each level of BCG.

   When we use only dead/alive as outcome in the analysis-function `effx`, we must use “Dead” as outcome:

```
> effx( (Dead=="Dead")*1, type="binary", exposure=dtpany, strata=bcg, data=bissau )

  ---------------------------------------------------------------------------
  response      :  (Dead == "Dead") * 1
  type          :  binary
  exposure      :  dtpany
  stratified by :  bcg

  dtpany is a factor with levels: 1+ DTP / no DTP
  baseline is  1+ DTP
  bcg is a factor with levels: BCG/no BCG
  effects are measured as odds ratios
```

```
---------------------------------------------------------------------------

effect of dtpany on (Dead == "Dead") * 1
stratified by bcg

number of observations  5274

                                     Effect  2.5% 97.5%
strata BCG level no DTP vs 1+ DTP     0.653 0.436 0.978
strata no BCG level no DTP vs 1+ DTP  0.746 0.175 3.170

Test for effect modification on 1 df: p-value= 0.86
```

We see that there is no interaction by the test for effct modification (the likelihood-ratio counterpart of the Breslow-Day-test). So we conclude that there is the same effect of DTP for both levels of BCG vaccination. It is clear that because of the vary sparse data, the effect of DTP in the "no BCG" stratum is largely undetermined.

But we also see that the reference levels used is those exposed to dtp, so we must use the `base` argument to get the right comparison:

```
> effx( (Dead=="Dead")*1,
+       type = "binary",
+   exposure = dtpany,
+       base = "no DTP",
+     strata = bcg,
+       data = bissau )
```

```
---------------------------------------------------------------------------
response      :  (Dead == "Dead") * 1
type          :  binary
exposure      :  dtpany
stratified by :  bcg

dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  no DTP
bcg is a factor with levels: BCG/no BCG
effects are measured as odds ratios
---------------------------------------------------------------------------

effect of dtpany on (Dead == "Dead") * 1
stratified by bcg

number of observations  5274

                                     Effect  2.5% 97.5%
strata BCG level 1+ DTP vs no DTP      1.53 1.020  2.29
strata no BCG level 1+ DTP vs no DTP   1.34 0.315  5.70

Test for effect modification on 1 df: p-value= 0.86
```

2. Use the BCG as a potentially confounding variable and obtain the MH-estimate for the OR and RR. What are they?

   We can use `effx` with a slight modification to compute the common effect, by simply replacing `strata=` with `control=`:

```
> effx( (Dead=="Dead")*1, type="binary", exposure=dtpany, base="no DTP",
+       control=bcg, data=bissau )
```

```
  ---------------------------------------------------------------------------
  response      :  (Dead == "Dead") * 1
  type          :  binary
  exposure      :  dtpany
  control vars  :  bcg

  dtpany is a factor with levels: 1+ DTP / no DTP
  baseline is  no DTP
  effects are measured as odds ratios
  ---------------------------------------------------------------------------

  effect of dtpany on (Dead == "Dead") * 1
  controlled for bcg

  number of observations  5274

  Effect   2.5%  97.5%
    1.52   1.03   2.23

  Test for no effects of exposure on 1 df: p-value= 0.0313
```

It is also possible to estimate the relative risk, using the argument `eff="RR"` — but only from version 1.1.40 of the Epi package, where it is also possible to use just a logical as a binary response. You can check your version of the `Epi`-package by:

```
> installed.packages()["Epi",c("Version","Built"),drop=FALSE]

      Version  Built
  Epi "1.1.40" "2.15.1"

> effx( Dead=="Dead", type="binary", exposure=dtpany, base="no DTP",
+       control=bcg, data=bissau, eff="RR" )

  ---------------------------------------------------------------------------
  response      :  Dead == "Dead"
  type          :  binary
  exposure      :  dtpany
  control vars  :  bcg

  dtpany is a factor with levels: 1+ DTP / no DTP
  baseline is  no DTP
  effects are measured as relative risk
  ---------------------------------------------------------------------------

  effect of dtpany on Dead == "Dead"
  controlled for bcg

  number of observations  5274

  Effect   2.5%  97.5%
    1.49   1.03   2.17

  Test for no effects of exposure on 1 df: p-value= 0.0314
```

3. Do the same, using age in months (`agem`) as control variable in the analysis. Is there any DTP effect?

   We first get an overview of how the data are distributed by age, `agem`:

```
> ( tt <- with( bissau, table(agem,dtp) ) )
```

```
      dtp
agem   0   1   2   3
   0 867   7   0   0
   1 808  81   0   0
   2 559 326  32   2
   3 339 328 122  18
   4 256 267 160  76
   5 196 209 181 109
   6  76 100  96  59

> pctab( tt )

      dtp
agem      0     1     2     3   All     N
   0   99.2   0.8   0.0   0.0 100.0 874.0
   1   90.9   9.1   0.0   0.0 100.0 889.0
   2   60.8  35.5   3.5   0.2 100.0 919.0
   3   42.0  40.6  15.1   2.2 100.0 807.0
   4   33.7  35.2  21.1  10.0 100.0 759.0
   5   28.2  30.1  26.0  15.7 100.0 695.0
   6   23.0  30.2  29.0  17.8 100.0 331.0

> ( tt <- with( bissau, table(agem,bcg) ) )

      bcg
agem BCG no BCG
   0 237    637
   1 468    421
   2 598    321
   3 589    218
   4 581    178
   5 554    141
   6 274     57

> pctab( tt )

      bcg
agem   BCG no BCG   All     N
   0   27.1   72.9 100.0 874.0
   1   52.6   47.4 100.0 889.0
   2   65.1   34.9 100.0 919.0
   3   73.0   27.0 100.0 807.0
   4   76.5   23.5 100.0 759.0
   5   79.7   20.3 100.0 695.0
   6   82.8   17.2 100.0 331.0
```

We see that the distribution of DTP and BCG vaccinations are highly dependent on age. So we should perhaps expect hat some of the effect would disappar when we control for age. Note that we can control for age in two different ways; we can either include it as a continuous variable (with a linear effect) or as a factor:

```
> effx( Dead=="Dead", type="binary", exposure=dtpany, base="no DTP", control=agem, data=bissau
  ---------------------------------------------------------------------------
  response      :  Dead == "Dead"
  type          :  binary
  exposure      :  dtpany
  control vars  :  agem

  dtpany is a factor with levels: 1+ DTP / no DTP
  baseline is  no DTP
  effects are measured as odds ratios
  ---------------------------------------------------------------------------
  effect of dtpany on Dead == "Dead"
  controlled for agem

  number of observations  5274

  Effect   2.5%  97.5%
   1.010  0.731  1.400

  Test for no effects of exposure on 1 df: p-value= 0.947
```

```
> effx( Dead=="Dead", type="binary", exposure=dtpany, base="no DTP", control=factor(agem), dat
  ---------------------------------------------------------------------------
  response      :  Dead == "Dead"
  type          :  binary
  exposure      :  dtpany
  control vars  :  agem

  dtpany is a factor with levels: 1+ DTP / no DTP
  baseline is  no DTP
  effects are measured as odds ratios
  ---------------------------------------------------------------------------

  effect of dtpany on Dead == "Dead"
  controlled for agem

  number of observations  5274

  Effect   2.5%  97.5%
   1.020  0.729  1.420

  Test for no effects of exposure on 1 df: p-value= 0.915
```

We see that with this control there is no effect of DTP, irrespective of how we control for age.

4. Do the same, but now using both `agem` and `bcg` (that is, the cross-classification) as control variables in the analysis. Is there any DTP effect?

   If we control for both, it means that we insert both variables as confounders, in the `effx` function this means that we should insert the two in a `list` used as the `control` argument:

```
> effx( Dead=="Dead", type="binary", exposure=dtpany, base="no DTP", control=list(bcg,factor(a
  ---------------------------------------------------------------------------
  response      :  Dead == "Dead"
  type          :  binary
  exposure      :  dtpany
  control vars  :  bcg factor(agem)

  dtpany is a factor with levels: 1+ DTP / no DTP
  baseline is  no DTP
  effects are measured as odds ratios
  ---------------------------------------------------------------------------

  effect of dtpany on Dead == "Dead"
  controlled for bcg factor(agem)

  number of observations  5274

  Effect   2.5%  97.5%
   1.470  0.954  2.280

  Test for no effects of exposure on 1 df: p-value= 0.0754
```

```
> effx( Dead=="Dead", type="binary", exposure=dtpany, base="no DTP", control=list(bcg,agem), d

  ---------------------------------------------------------------------------
  response      :  Dead == "Dead"
  type          :  binary
  exposure      :  dtpany
  control vars  :  bcg agem

  dtpany is a factor with levels: 1+ DTP / no DTP
  baseline is  no DTP
  effects are measured as odds ratios
  ---------------------------------------------------------------------------

  effect of dtpany on Dead == "Dead"
  controlled for bcg agem

  number of observations  5274

  Effect   2.5%  97.5%
   1.460  0.954  2.230

  Test for no effects of exposure on 1 df: p-value= 0.077
```

Again we see that the addition of `bcg` as a confounding variable reveals some effect of DTP, regardless of whether we use `agem` as a quantitative variable (assuming that the effect of age is linear on the log-odds-scale).

Formally speaking the effect is non-significant as opposed to what it was when we only controlled for BCG and not age. However, the difference between a p-value of 3 and one of 7% is very small, the actual effect is in both cases an odd-ratio of 1.46, and depending on how we control we have and lower limit of the confidnece interval for the OR of 1.03 (signif.) or of 0.95 (non-sign.).

However in both cases we see an indication of elevated risk of about 50%, but we cannot really tell wheter it is a few percent or a doubling of mortality.

Since the absolute mortality is so small, it does not matter wheter we use OR or RR, the two measures are virtually identical when the probabilities of putcome are small:

```
> effx( Dead=="Dead", type="binary", exposure=dtpany, base="no DTP", eff="RR",
+       control=list(bcg,agem), data=bissau )

  ---------------------------------------------------------------------------
  response      :  Dead == "Dead"
  type          :  binary
  exposure      :  dtpany
  control vars  :  bcg agem

  dtpany is a factor with levels: 1+ DTP / no DTP
  baseline is  no DTP
  effects are measured as relative risk
  ---------------------------------------------------------------------------

  effect of dtpany on Dead == "Dead"
  controlled for bcg agem

  number of observations  5274

  Effect   2.5%  97.5%
   1.440  0.956  2.160

  Test for no effects of exposure on 1 df: p-value= 0.0761
```

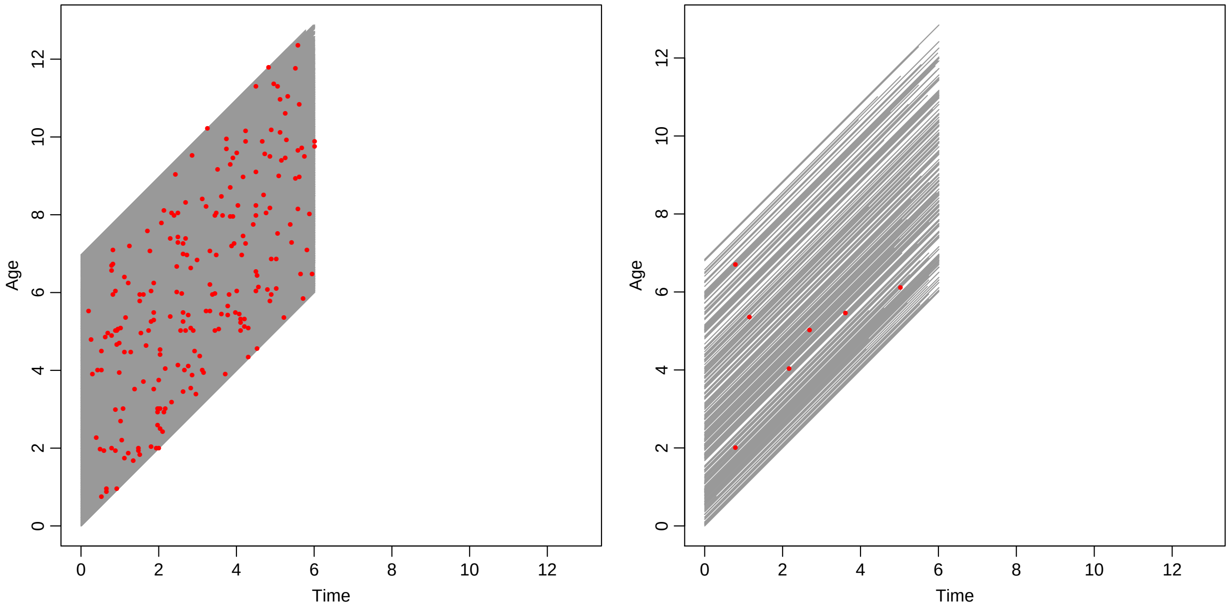
5.1.5 Survival analysis of childhood mortality in Guinea-Bissau

1. We start by reading the data and transforming data as before In order to do survival analysis we set up data as a `Lexis` object with two time-scales, time since visit and current age (time since birth), note we enter the time-scales in months:

```
> Lb <- Lexis( entry = list( Time = 0,
+                           Age = age/(365.25/12) ),
+              exit = list( Time = fuptime/(365.25/12) ),
+       exit.status = Dead,
+              data = bissau )

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

> summary( Lb )

  Transitions:
       To
  From    Alive Dead  Records:  Events: Risk time:  Persons:
    Alive  5052  222      5274       222   28917.85      5274

  Rates:
       To
  From    Alive Dead Total
    Alive     0 0.01  0.01
```

We can show the follow-up in a Lexis-diagram, both for all and for a 5% random sample:

```
> par( mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
> plot( Lb, col=gray(0.6) )
> points( Lb, pch=c(NA,16)[Lb$"lex.Xst"], col="red", cex=0.6 )
> wh <- runif(nrow(Lb))<0.05
> plot( Lb[wh,], col=gray(0.6) )
> points( Lb[wh,], pch=c(NA,16)[Lb[wh,"lex.Xst"]], col="red", cex=0.6 )
```



Note that both `age` and `fuptime` are measured in days, so the time-variables we have in `Lb` are measured in months. A simple Cox-model with bcg and age at entry (in months) at entry is set up using time since first visit (`Time`) as time-scale:

```
> library( survival )
> m1 <- coxph( Surv( Time, Time+lex.dur, lex.Xst=="Dead" ) ~
+              bcg + factor(agem),
+              data=Lb )
> ci.exp( m1 )
```

```
                exp(Est.)      2.5%    97.5%
bcgno BCG       1.4150942 1.0628313 1.884110
factor(agem)1 1.1218691 0.7118992 1.767933
factor(agem)2 0.7734721 0.4659289 1.284014
factor(agem)3 1.2201058 0.7574272 1.965414
factor(agem)4 1.3944726 0.8680841 2.240052
factor(agem)5 1.3918874 0.8534351 2.270062
factor(agem)6 0.9895328 0.4950160 1.978068
```

We see that persons without BCG-vaccination have a higher mortality.

2. We can evaluate the effect of DTP vaccination by changing the covariate `bcg` wirth `dtpany`

```
> m2 <- update( m1, . ~ . - bcg + dtpany )
> ci.exp( m2 )
```

```
                exp(Est.)      2.5%    97.5%
factor(agem)1 1.0293268 0.6558760 1.615418
factor(agem)2 0.6788407 0.4068749 1.132595
factor(agem)3 1.0393116 0.6327681 1.707053
factor(agem)4 1.1729539 0.7104716 1.936490
factor(agem)5 1.1584998 0.6882826 1.949958
factor(agem)6 0.8117511 0.3957607 1.664996
dtpanyno DTP  0.9979789 0.7197838 1.383696
```

and we see there is no mrginal effect of DTP om mortality.

3. But if we enter both variables we see an effect of both:

```
> m3 <- update( m1, . ~ . + dtpany )
> ci.exp( m3 )
```

```
                exp(Est.)      2.5%    97.5%
bcgno BCG       1.7376437 1.1871318 2.543446
factor(agem)1 1.1420990 0.7243778 1.800704
factor(agem)2 0.7302230 0.4365875 1.221349
factor(agem)3 1.0978956 0.6678785 1.804781
factor(agem)4 1.2286509 0.7444014 2.027915
factor(agem)5 1.2090204 0.7186231 2.034071
factor(agem)6 0.8495861 0.4143648 1.741935
dtpanyno DTP  0.6915217 0.4524274 1.056970
```

we see a protective effect of BCD, but potential harmful effect of DTP.

4. We can then try to insert the interaction:

```
> m4 <- update( m3, . ~ . + dtpany:bcg )
> ci.exp( m4 )
```

```
                       exp(Est.)      2.5%    97.5%
bcgno BCG              1.4305868 0.3522739 5.809624
factor(agem)1          1.1453270 0.7261539 1.806468
factor(agem)2          0.7319001 0.4374948 1.224421
factor(agem)3          1.0989575 0.6684984 1.806598
factor(agem)4          1.2299067 0.7451548 2.030008
factor(agem)5          1.2120327 0.7202963 2.039471
factor(agem)6          0.8509784 0.4150298 1.744849
dtpanyno DTP           0.6796851 0.4369438 1.057280
bcgno BCG:dtpanyno DTP 1.2359351 0.2877469 5.308609
```

```
> anova( m3, m4, test="Chisq" )

  Analysis of Deviance Table
   Cox model: response is  Surv(Time, Time + lex.dur, lex.Xst == "Dead")
   Model 1: ~ bcg + factor(agem) + dtpany
   Model 2: ~ bcg + factor(agem) + dtpany + bcg:dtpany
     loglik Chisq Df P(>|Chi|)
  1 -1875.2
  2 -1875.2 0.086  1    0.7693
```

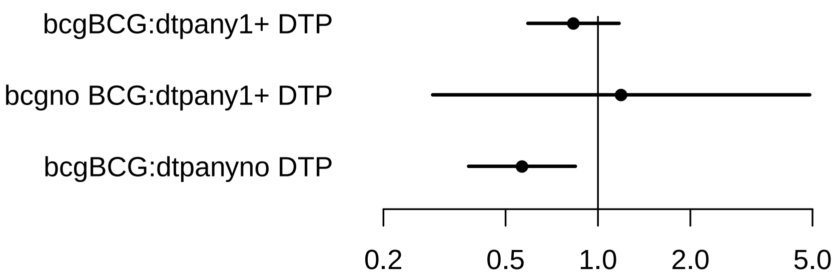
which, although not significant, is not very informative; we want the RR for each combination of the two variables:

```
> m5 <- update( m1, . ~ . - bcg + dtpany:bcg )
> ci.exp( m5 )

                         exp(Est.)      2.5%     97.5%
  factor(agem)1          1.1453270 0.7261539 1.8064679
  factor(agem)2          0.7319001 0.4374948 1.2244210
  factor(agem)3          1.0989575 0.6684984 1.8065977
  factor(agem)4          1.2299067 0.7451548 2.0300083
  factor(agem)5          1.2120327 0.7202963 2.0394708
  factor(agem)6          0.8509784 0.4150298 1.7448486
  bcgBCG:dtpany1+ DTP    0.8321130 0.5908951 1.1718020
  bcgno BCG:dtpany1+ DTP 1.1904099 0.2893258 4.8978542
  bcgBCG:dtpanyno DTP    0.5655749 0.3788463 0.8443395
```

coxph produces a warning, because the interaction generated contains the intercept (namely the sum of the 4 columns), and so automatically exclude the last one.

We can also see the estimates graphically

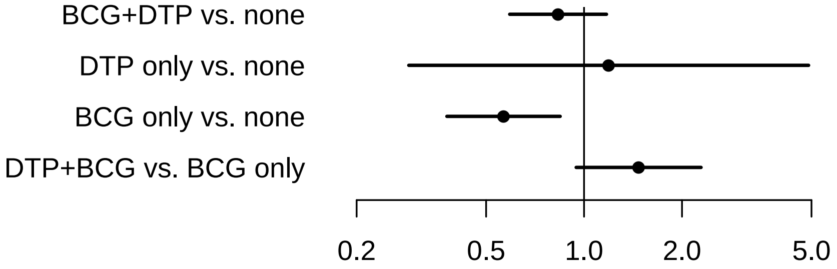
```
> plotEst( ci.exp(m5,subset="bcg"), xlog=T, vref=1 )
```



The reference is the no BCG, no DTP group, so we see that the protective effect is smaller in the DTP-vaccinated group than in the non-DTP-vaccinated group. It is of course also of interest to see the DTP-effect within the BCG-group, and that can be teased out:

```
> CM <- rbind( diag(3), c(1,0,-1) )
> rownames( CM ) <- c("BCG+DTP vs. none",
+                     "DTP only vs. none",
+                     "BCG only vs. none",
+                     "DTP+BCG vs. BCG only")
> CM

                      [,1] [,2] [,3]
  BCG+DTP vs. none       1    0    0
  DTP only vs. none      0    1    0
  BCG only vs. none      0    0    1
  DTP+BCG vs. BCG only   1    0   -1
```

```
> plotEst( ci.exp(m5,subset="bcg",ctr.mat=CM), xlog=T, vref=1 )
```



There are very few persons and

```
> ftable( xtabs( cbind(dead,N=1) ~ bcg + dtpany, data=Lb ) )

                  dead    N
  bcg    dtpany
  BCG    1+ DTP     92 2142
         no DTP     33 1159
  no BCG 1+ DTP      2   31
         no DTP     95 1942
```

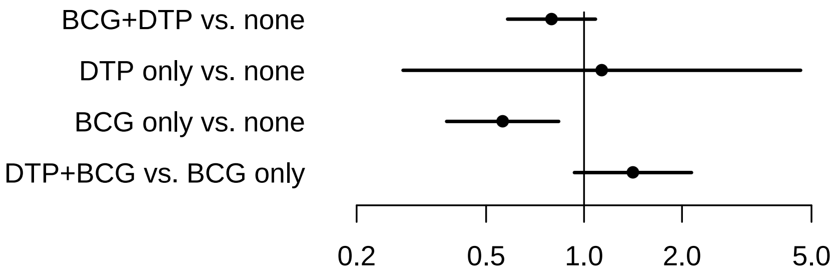
so essentially we can only estimate a BCG-effect among non-DTP vaccinated and a DTP-effect among BCG-vaccinated, which are the effects we see in the main effects model.

5. If we instead use current age as time-scale, we just change the outcome variable using `update`, and then re-use the code to generate the graph:

```
> m5a <- update( m5, Surv(Age,Age+lex.dur,lex.Xst=="Dead") ~ . - factor(agem) )
```

```
> plotEst( ci.exp(m5a,subset="bcg",ctr.mat=CM), xlog=T, vref=1 )
```

We see that the analysis with current age as time scale gives pretty much the same estimates as the analysis with time since entry and age at entry:



6. Finally we compare the results from a Poisson model where we assume constant rates, and a logistic regression model where we altogether ignore censoring:

```
> ci.exp(m3,subset="no")

             exp(Est.)      2.5%    97.5%
bcgno BCG    1.7376437 1.1871318 2.543446
dtpanyno DTP 0.6915217 0.4524274 1.056970

> p3 <- glm( (lex.Xst=="Dead") ~ bcg + dtpany + factor(agem),
+            offset=log(fuptime), family=poisson, data=Lb )
> ci.exp(p3,subset="no")

             exp(Est.)      2.5%    97.5%
bcgno BCG    1.7349813 1.1851585 2.539880
dtpanyno DTP 0.6912244 0.4521804 1.056638

> l3 <- glm( (lex.Xst=="Dead") ~ bcg + dtpany + factor(agem),
+            family=binomial, data=Lb )
> ci.exp(l3,subset="no")

             exp(Est.)      2.5%    97.5%
bcgno BCG    1.7397389 1.1772134 2.571064
dtpanyno DTP 0.6784886 0.4393344 1.047828
```

We see that the three approaches produce virtually identical results. For the Poisson model it is because the mortality varies very little with age and follow-up, and the Poisson model *is* a model that assumes constant mortality. For the logistic model it is because the amount of censring is quite limited