Short course in epidemiology
Advanced stream
Exercises & practicals

SDC / CDC
December 2015
http://BendixCarstensen.com/Epi/Courses/IDEG2015/
Version 3.1

Compiled Monday 7th December, 2015, 07:51
from: /home/bendix/teach/Epi/IDEG2015/pracs/pracs.tex

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

Ed Gregg — Epidemiology and Statistics Branch
Chief — Division of Diabetes Translation
Centers for Disease Control, Atlanta, USA

Contents

Program for advanced stream

Please note the details of the computing requirements on the course web-site, http://bendixcarstensen.com/Epi/Courses/IDEG2015/, including download of datasets and programs for the practicals.

The practicals will be possible to do both with Stata and with R. There will be a wrap-up of the practicals at the end.

Monday 7 December 2015

08:00 – 09:00	Registration
09:00 – 09:15	Welcome & Course Overview
	Point Grey Room — 3rd Floor
Advanced stream:	**Pinnacle II, 3rd floor**
09:15 – 10:15	Population Surveillance and Monitoring (EG)
10:15 – 10:30	Practical: Diabetes Monitoring (EG)
10:30 – 10:45	Practical: Prevalence from a register (BC)
10:45 – 11:05	Morning Refreshment Break
11:05 – 11:35	Case examples: US surveillance (EG)
11:35 – 12:35	Demographic concepts (BxC)
	Practical: Diabetes incidence from a register
12:35 – 13:30	Lunch

Chapter 1

Practicals

This set of practicals will very briefly introduce you to the classical concepts of incidence, mortality and prevalence (with which you are presumably familiar), and then introduce you to:

- data structures from population surveys and registers
- theoretical concepts of rates
- practical use of concepts on data

The main example will be a dataset that resembles the Danish National Diabetes Register.

The section with solutions contain subsections that are numbered in parallel to the exercises, so the solutions corresponding to section 1.2 is in section 3.2 etc.

1.0 Diabetes monitoring

Scenario: You are the NCD Unit Leader of a small developing country, or a province/state of a large developing country.) A wealthy foundation has donated a 3 million USD start-up grant, with 5 years continued funding of 1 million USD per year to enhance diabetes monitoring in your country.

- What type of system would you propose? (*i.e.*., surveys, health systems data; registries) Why?
- Describe the general architecture of your system.
- What types of data would you collect?
- What would be your primary indicators/definitions for risk factors, DM cases, complications, covariates?

1.1 Classical concepts

The following is a brief overview of the basic concepts, amended with exercises in derivation of the measures from the National Danish Diabetes Register. The exercises are given first in general terms, and then in more technical terms for those who wish to pursue the calculations in practice.

1.2 Prevalence

Some use the word prevalence for the *number* of affected people, and specifically refer to the prevalence *proportion* when talking about the fraction of persons affected. Here we shall use the term "prevalence" for the *fraction* of persons affected.

Prevalence always refers to a specified *point* in time — a specific date.

empirical prevalence of a disease in a population is the fraction of the population that suffers from the disease at the specified date.

theoretical prevalence of a disease in a population is the *probability* that a randomly chosen person from the population suffers from the disease at the specified date.

At first glance these two look pretty much the same, but when we qualify the concepts by, say, age, differences emerge:

The *empirical* prevalence necessarily requires that the population be divided in age-*classes* to enable the calculation of fractions for each age-class.

The *theoretical* prevalence lends itself to statistical modeling; it is possible to specify mathematically how the probability of being diseased depends on age, so that we have an expression for the probability (that is the prevalence) for any age, say 63.7 or 71.3 years.

1.2.1 Practical

We will use a simulated version of the Danish National Diabetes Register (all dates are randomly moved ±7 days, so no persons exist in reality).

Dates are coded in years, so that 1 January 2006 is coded 2006.0, 1 July 2006 is coded 2006.5 and 31 December 2006 as 2006.997. This is how the first few of the almost 500,000 records look:

```
   sex    doBth     doDM    doIns    doDth
1    F 1899.984 1990.052       NA 1991.475
2    F 2000.006 2005.738 2005.773       NA
3    F 2000.002 2008.628 2008.679       NA
4    F 1900.985 1993.489       NA 1994.130
5    M 2001.011 2001.019       NA       NA
6    M 2001.990 2005.763 2005.865       NA
7    M 1903.009 1992.683       NA 1994.454
8    M 1902.997 1993.209       NA 2001.495
9    M 1903.016 1990.517       NA 1991.185
10   F 1902.988 2002.438       NA 2003.621
```

1. How would you go about estimating the *number* of prevalent cases in Denmark as of 1 January 2005 if you had access to this dataset?

2. The dataset `dr.dta` is a Stata dataset with a modified version of the Danish National Diabetes Register which is also available as R-dataset, `dr.Rda`. Both are available in the folder http://bendixcarstensen.com/Epi/Courses/IDEG2015/data/.

   Read the dataset either with Stata or with R; with R it looks like this:

```
> library( Epi )
> clear()
> # load( url("http://bendixcarstensen.com/Epi/Courses/IDEG2015/data/dr.Rda") )
> # save( dr, file="../data/dr.Rda" )
> load( file="../data/dr.Rda" )
> str( dr )
> summary( dr )
```

3. How many prevalent cases of diabetes were there in Denmark as of 1 January 2005?

   If you do not use a computer for this, indicate how you would use the data to obtain the number. Do similarly for the remaining questions.

4. How many men and women?

5. How many in each 5-year age-class?

6. How many in each 1-year age-class?

7. The size of the Danish population as of 1 January 1971–2013 by sex and 1-year age-classes is in the dataset `Ndk` available at the course website; the first few lines look like this:

```
> # load( file=url("http://bendixcarstensen.com/Epi/Courses/IDEG2015/data/Ndk.Rda") )
> ### The local vsrions on this computer:
> load( file="../data/Ndk.Rda" )
> head( Ndk )
```

```
  sex A    P     N
1   M 0 1971 35839
2   F 0 1971 34108
3   M 1 1971 36302
4   F 1 1971 34153
5   M 2 1971 37855
6   F 2 1971 35609
```

   Supposing you have access to population data from Denmark how would you compute the prevalence — that is the *proportion* of the population affected?

8. What are the age-specific prevalences in, say, 5-year classes?

9. How do the prevalence look as a function of age?

10. How does the prevalences look if we use 1-year age-classes?

11. How would you go about modeling prevalence as a smooth function of age?

    What would the analysis data set look like? And what kind of statistical model would be applicable and relevant?

The *modeling* of prevalences also illustrates the contrast between the *empirical* and *theoretical* prevalences; the former are necessarily tied to a particular grouping of the population; for example by sex and/or age, whereas the latter refer to *any* combination of sex and age; after modelling we can in principle refer to the prevalence of DM in women aged 68.3 years or 73.6 years.

1.3 Incidence

The incidence (rate) of DM is defined as the number of new cases of DM that occur in a population in a predefined period of time. Of course the number of new DM cases is approximately twice as large if the population you look at is twice as large, but also if you look at the same population for two instead of one year; so the relevant denominator must

be proportional *both* to the number of persons considered *and* the length of time considered. This is the population follow-up time — the person-years.

The total number of person-years in the population may be approximated from population counts at fixed dates, normally by taking averages of population counts at two time points multiplied by the distance between the time points.

As in the case of prevalence we distinguish between the empirical and theoretical incidence rates:

empirical incidence rate refers to a given time-period (and age-interval), and is defined as the number of new cases relative to the population risk time (person-years) in the time-period.

theoretical incidence rate is defined at any point in time (and age) as the probability of seeing an event (DM diagnosis, for example) in a susceptible person in a small period of time *relative* to the length of this period.

Note that both empirical and theoretical incidence rates have a dimension of time^{-1}, namely events, respectively probability *per* time. While empirical incidence rates necessarily refer to a specific time-period, the theoretical incidence rate is defined for any point in time and can vary continuously by time.

1.3.1 Practical

1. How would you find the number of newly diagnosed cases in age 60–64 (incl.) in the year 2006 from the Danish National Diabetes Register?

2. In order to compute the (empirical) incidence rate we also need the person-years in the Danish population. This is available in the dataset `Ydk` from the folder ("http://bendixcarstensen.com/Epi/Courses/IDEG2015/data"). The first few lines look like this:

```
  sex A    P        Y
1   M 0 1971 37139.17
2   F 0 1971 35128.83
3   M 1 1971 36133.67
4   F 1 1971 34223.00
5   M 2 1971 37113.00
6   F 2 1971 34926.33
```

3. How would you go about deriving the age-specific rates in 2006, in 5-year age-classes and by sex?

4. There is no particular reason to choose 5-year intervals; we could as before use 1-year intervals, as population figures are actually available for these.

   How do you think a graph of age-specific rates would look?

5. How would you go about fitting a model with a smooth age-effect for the incidence rates? Specifically, what kind of data would be needed?

1.3.2 Caveat: people only get DM once

In the calculations above we have used the total population risk time as denominator, even though more than 10% of the population over 60 years of age have diabetes. This means that the rates of diabetes are underestimated because persons with diabetes are not at risk of getting diabetes; and we shoudl only include the susceptibles in th edenominayor. Thus the person-years should only be computed for persons without diabetes. One way of doing this is to compute the person-years among diabetes patients and subtract it from to total population person-years.

6. As an example we used the incidence in 2006. How would you compute the person-years among all diabetes patients contributing during 2006, and subdivide it by age class?

7. How large a percentage of the population risk time is among persons with DM.

8. Now re-estimate the the age-specific incidence rates using the correct denominator and compare the two.

1.4 Mortality and survival

When we are talking about mortality rates, we have the same considerations as before regarding empirical and theoretical rates, but as a special feature of mortality we might also be interested in survival.

Survival is defined as the probability, $S(t)$ of being alive after some specified length of time, t. This is a *cumulative* measure that requires an *origin*, that is, t must be defined as time *since* some origin.

In the case of diabetes it will normally be time since diagnosis of diabetes. The survival is a function of the mortality rates, so in order to compute the survival function at different times after diagnosis, we must know the mortality rates as a function of time since diagnosis.

Mortality rates however, is naturally also dependent on age — possibly both on age at diagnosis of DM as well as current age. The latter is the sum of age at diagnosis and the time since diagnosis (duration). So we are facing the problem of describing mortality by time since diagnosis of DM, age at diagnosis of DM as well by the sum of the two. The linear effects of the three variables cannot be separated, but the non-linear effects can.

9. As a start, compute mortality rates among diabetes patients, say during the year 2006. Above we computed the person-years among diabetes patients by age and sex in 2006 in 1-year intervals, in order to subtract these from the total population person-years. But the person-years among DM patients will also be the denominator (person-years) for the mortality. So we just need the number of deaths among diabetes patients classified by age (at death) and sex — how would you compute that from the dataset?

   We can then show the number of cases, person-years, and rates per 1000 PY:

10. Plot the mortality rates as a function of age.

11. How would you make a model that showed mortality rates as a smooth function of age?

12. We could also look at mortality as a function of duration of DM. However this would really only make sense if we controlled for age in some way. So for the sake of the argument do the calculation of duration-specific mortality for persons diagnosed in age 60 in the entire period after 1995.

    How would you extract data (deaths and person-years) for this? How do the mortality rates look as a function of DM duration?

1.4.1 Survival

We can devise a so called life-table survival curve from mortality rates; if the mortality in an interval is λ and the interval length is ℓ the probability of dying in the interval is approximately $\lambda\ell$ — provided that the death probability is not too large (the correct expression is $1 - \exp(-\lambda\ell)$). Thus, the probability of surviving the interval is $1 - \lambda\ell$.

So the probability of surviving the first interval (that starts at time 0) is $1 - \lambda_0\ell$. The probability or surviving the next is $1 - \lambda_1\ell$ — or more precisely, the *conditional* probability of surviving the second interval *given* that the person already survived the first one. Hence the probability of surviving till the end of the second interval is $(1 - \lambda_0\ell) \times (1 - \lambda_1\ell)$. So we have $S(0) = 1$, $S(1) = 1 - \lambda_0\ell$, $S(2) = (1 - \lambda_0\ell) \times (1 - \lambda_1\ell)$, etc.

13. Based on the mortality rates for 1-year intervals of DM duration, how would you calculate the (actuarial) survival curve? In particular indicate at what values of duration you compute the survival probability.

14. An alternative way of computing the survival function(s) is to use the Kaplan-Meier estimator, which requires that we define an observed survival time for each person, as well as an indicator of whether follow-up (the survival time) ended by censoring or death.

    How would you construct a dataset for this type of analysis?

15. What we did was to compute the mortality in 1-year interval of diabetes duration for patients diagnosed in age 60 (that is between their 60^{th} and 61^{st} birthdays). We could of course repeat the exercise for persons diagnosed in ages 50, 51, ..., 99 to get an impression of how mortality and survival depend on age at diagnosis.

    Can you think of a more comprehensive way to address this type of question?

    And of what types of questions on mortality rates and survival you *really* would like to address?

The practical implementation of this is out of the scope of this stream, but in a special section on "Mortality, age at diagnosis, duration and current age" in the solutions chapter, some of these issues are addressed.

Chapter 2

Basic concepts in survival and demography

The following is a condensed overview of concepts central to handling follow-up data; the target audience for this section is

- epidemiologists who wants a handy overview of the mathematical relationships between the theoretical concepts
- statisticians (and probabilists, mathematicians) who want to get an overview of how the various concepts in probability translates to epidemiological concepts

The following is a summary of relations between various quantities used in analysis of follow-up studies. They are ubiquitous in the analysis and reporting of results. Hence it is important to be familiar with all of them and the relation between them.

2.1 Probability

Survival function:

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

Conditional survival function:

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

Cumulative distribution function of death times (cumulative risk):

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

Density function of death times:

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

Intensity:

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

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

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

Relationships between terms:

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

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

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

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

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

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

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

2.2 Statistics

Likelihood contribution from follow up of one person:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.3 Competing risks

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

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

The survival function is then:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.4 Demography

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

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

where f is the density of the distribution of lifetimes.

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

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

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

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

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

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

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

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

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

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

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

and hence:

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

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

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

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

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

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

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

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

Chapter 3

Solutions

3.2 Prevalence

Some use the word prevalence for the *number* of affected people, and specifically refer to the prevalence *proportion* when talking about the fraction affected. Here we shall use the term "prevalence" for the fraction affected.

Prevalence always refers to a specified *point* in time:

empirical prevalence of a disease in a population is the fraction of the population that suffers from the disease

theoretical prevalence of a disease in a population is the *probability* that a randomly chosen person from the population suffers from the disease

At first glance these two look pretty much the same, but when we qualify the concepts by, say, age, differences emerge.

The *empirical* prevalence necessarily requires that the population be divided in age-*classes* to enable the calculation of fractions.

The *theoretical* prevalence lends itself to statistical modeling; it is possible to specify mathematically how the probability of being diseased depends on age, so that we have a probability (that is the prevalence) for any age, say 63.7 years.

3.2.1 Practical

The dataset `dr.dta` is a Stata dataset with a modified version of the Danish National Diabetes Register (all dates are randomly moved ±7 days, so no persons exist in reality). It is also available as R-dataset, `dr.Rda`. Both are available in the folder http://bendixcarstensen.com/Epi/Courses/IDEG2015/data/dr.dta.

Dates are coded in years, so that 1 January 2006 is coded 2006.0, 1 July 2006 is coded 2006.5 and 31 December 2006 as 2006.997.

1. How would you go about estimating the number of prevalent cases in Denmark as of 1 January 2005 if you had access to this dataset?

   You will need all persons that both have a date of diagnosis before 1.1.2005 and who is not dead at that date.

2. We read the dataset either with Stata or with R using:

```
> library( Epi )
> # load(           url("http://bendixcarstensen.com/Epi/Courses/IDEG2015/data/dr.Rda") )
> ### The local version on this computer
> load( file="../data/dr.Rda" )
> str( dr )
'data.frame':        497232 obs. of  5 variables:
 $ sex  : Factor w/ 2 levels "M","F": 2 2 2 2 1 1 1 1 1 2 ...
 $ doBth:Class 'cal.yr'  num [1:497232] 1900 2000 2000 1901 2001 ...
 $ doDM :Class 'cal.yr'  num [1:497232] 1990 2006 2009 1993 2001 ...
 $ doIns:Class 'cal.yr'  num [1:497232] NA 2006 2009 NA NA ...
 $ doDth:Class 'cal.yr'  num [1:497232] 1991 NA NA 1994 NA ...

> head( dr )

  sex    doBth     doDM    doIns    doDth
1   F 1899.984 1990.052       NA 1991.475
2   F 2000.006 2005.738 2005.773       NA
3   F 2000.002 2008.628 2008.679       NA
4   F 1900.985 1993.489       NA 1994.130
5   M 2001.011 2001.019       NA       NA
6   M 2001.990 2005.763 2005.865       NA

> summary( dr )

 sex            doBth           doDM          doIns           doDth
 M:257840   Min.   :1889   Min.   :1942   Min.   :1994     Min.   :1990
 F:239392   1st Qu.:1927   1st Qu.:1995   1st Qu.:1995     1st Qu.:1998
            Median :1939   Median :2002   Median :2002     Median :2003
            Mean   :1940   Mean   :2001   Mean   :2002     Mean   :2003
            3rd Qu.:1951   3rd Qu.:2008   3rd Qu.:2007     3rd Qu.:2008
            Max.   :2011   Max.   :2012   Max.   :2012     Max.   :2012
                                          NA's   :375954   NA's   :310870
```

3. The prevalent cases at 1 January 2005 are those diagnosed before 2005, and who died later than 2005 (or did not die). The second form of the calculation here computes the exit date using `pmin`:

```
> with( dr, table( doDM<2005 & (doDth>2005|is.na(doDth)), exclude=NULL ) )

 FALSE   TRUE   <NA>
292757 204475      0
```

4. How many men and women?

   The further calculations is best made by selecting only those persons that were alive with diabetes at the 1 January 2005, (the data frame `pr2005`):

```
> pr2005 <- subset( dr, doDM<2005 & (doDth>2005|is.na(doDth)) )
> ( ptt <- with( pr2005, table(sex) ) )

sex
     M      F
104171 100304
```

5. How many in each age-class?

   Here we use the function `floor` that throws away decimals — when we divide the age at 2005 (`2005-doBth`) by 5 and remove the decimals and subsequently multiply by 5 we get numbers 0, 5, 10, ... indicating the lower end of each age category:

```
> with( pr2005, table( floor((2005-doBth)/5)*5, sex ) )
```

```
    sex
        M     F
  0    48    60
  5   231   232
  10  503   480
  15  675   596
  20  760   817
  25 1291  1652
  30 1914  2813
  35 3055  3954
  40 4706  4567
  45 6725  5452
  50 9263  6807
  55 14363 9903
  60 15521 11054
  65 14007 11274
  70 11923 11596
  75 9446  11032
  80 6155  9697
  85 2675  5489
  90 779   2320
  95 119   477
  100 12   31
  105 0    1
```

6. In the Epi package is the dataset `N.dk` with the size of the Danish population as of 1 January 1971–2013 by sex and 1-year age-classes. The coding of sex is numeric, so we change it to factor as in the register dataset:

```
> data( N.dk )
> head( N.dk )
```

```
  sex A    P     N
1   1 0 1971 35839
2   2 0 1971 34108
3   1 1 1971 36302
4   2 1 1971 34153
5   1 2 1971 37855
6   2 2 1971 35609
```

```
> str( N.dk )
```

```
'data.frame':        8600 obs. of  4 variables:
 $ sex: num  1 2 1 2 1 2 1 2 1 2 ...
 $ A  : num  0 0 1 1 2 2 3 3 4 4 ...
 $ P  : num  1971 1971 1971 1971 1971 ...
 $ N  : num  35839 34108 36302 34153 37855 ...
 - attr(*, "Contents")= chr "Population size as of 1 January in Denmark"
```

```
> N.dk <- transform( N.dk,
+                    sex = factor( sex, labels=c("M","F") ) )
> xtabs( N ~ sex, data=subset( N.dk, P==2005 ) )
```

```
sex
      M       F
2677292 2734113
```

so there are 2,677,292 men in Denmark as of 1 January 2005.

The overall prevalence of diabetes among men and women is computed by taking the number of men and women with diabetes and dividing it by the total number of persons in the population.

```
> ( pop <- xtabs( N ~ sex, data=subset( N.dk, P==2005 ) ) )
sex
      M       F
2677292 2734113
> round( ptt / pop * 100, 1 )
sex
  M   F
3.9 3.7
```

so the prevalence of diabetes overall was 3.9 and 3.7 percent respectively in men and and women.

7. What are the age-specific prevalences in, say, 5-year classes?

   We make a tabulation of the number of persons by age and sex, and do the same with the number of DM patients from the register, but we only take the first 20 age-classes (0–4,5–9,...,95–99) as these are the ones that are represented in the population figures.

   Note that we compute the persons' ages at the 1 January 2005 (which is coded as 2005.0).

```
> pop <- xtabs( N ~ I(floor(A/5)*5) + sex, data=subset( N.dk, abs(P-2005)<0.1 ) )
> ptt <- with( pr2005, table( floor((2005-doBth)/5)*5, sex ) )[1:20,]
> cbind( ptt, pop )
       M     F      M      F
0     48    60 167882 160174
5    231   232 176410 167652
10   503   480 177531 168497
15   675   596 156371 148211
20   760   817 147943 144598
25  1291  1652 173681 172033
30  1914  2813 193537 190643
35  3055  3954 210636 203290
40  4706  4567 204212 197524
45  6725  5452 187173 182720
50  9263  6807 180774 179027
55 14363  9903 195417 193559
60 15521 11054 158478 160929
65 14007 11274 116440 124845
70 11923 11596  88207 103568
75  9446 11032  68065  90507
80  6155  9697  45263  75487
85  2675  5489  20839  44530
90   779  2320   7147  20756
95   119   477   1286   5563
> round( (ptt / pop) * 100, 2 )
    sex
        M    F
  0  0.03 0.04
  5  0.13 0.14
  10 0.28 0.28
  15 0.43 0.40
  20 0.51 0.57
  25 0.74 0.96
  30 0.99 1.48
  35 1.45 1.95
  40 2.30 2.31
  45 3.59 2.98
```

```
50  5.12  3.80
55  7.35  5.12
60  9.79  6.87
65 12.03  9.03
70 13.52 11.20
75 13.88 12.19
80 13.60 12.85
85 12.84 12.33
90 10.90 11.18
95  9.25  8.57
```

8. How do the prevalence look as a function of time?

   We have the two column matrices `ptt` and `pop` with diabetes cases and population size as of 1 January 2006, so we can plot the ratio of these against the mid-point of the age-intervals. But formally what is assumed is that age-specific prevalences are constant in 5-year age-classes:

```
> par( mfrow=c(1,2), bty="n", las=1 )
> matplot( seq(2.5,97.5,5),  (ptt/pop)*100,
+          type="l", lty=1, lwd=3, col=c("blue","red"),
+          xlab="Age (years)", ylab="Prevalence (%)", las=1, yaxs="i", ylim=c(0,15) )
> matplot( seq(0,100,5),  ((ptt/pop)*100)[c(1:20,20),],
+          type="s", lty=1, lwd=3, col=c("blue","red"),
+          xlab="Age (years)", ylab="Prevalence (%)", las=1, yaxs="i", ylim=c(0,15) )
```

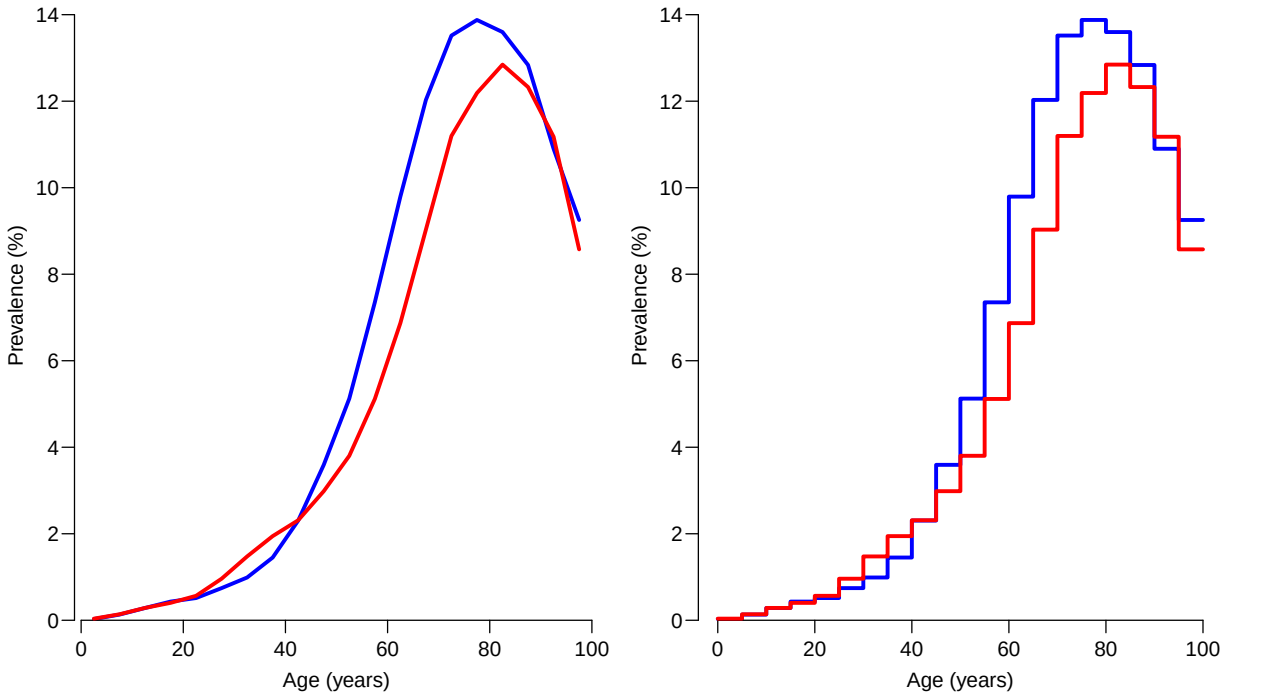

Figure 3.1: *Age-specific prevalence of diabetes at 1 January 2005 in 5-year age-classes in Denmark. The left plot is just connecting the midpoints of the age-classes; the right hand plot shows the formally assumed model with constant prevalence in each 5-year class.*

9. How does the prevalences look if we use 1-year age-classes?

This is just the same calculations, replacing 5 by 1 (leaving it a bit superfluous, though) and almost the same code for the plot:

```
> pop <- xtabs( N ~ floor(A) + sex, data=subset( N.dk, abs(P-2005)<0.1 ) )
> ptt <- with( pr2005, table( floor(2005-doBth), sex ) )[1:100,]
> par( mfrow=c(1,2), bty="n", las=1 )
> matplot( seq(0.5,99.5,1),  (ptt/pop)*100,
+          type="l", lty=1, lwd=3, col=c("red","blue"),
+          xlab="Age (years)", ylab="Prevalence (%)", las=1, yaxs="i", ylim=c(0,15) )
> matplot( seq(0,100,1), ((ptt/pop)*100)[c(1:100,100),],
+          type="s", lty=1, lwd=3, col=c("red","blue"),
+          xlab="Age (years)", ylab="Prevalence (%)", las=1, yaxs="i", ylim=c(0,15) )
```

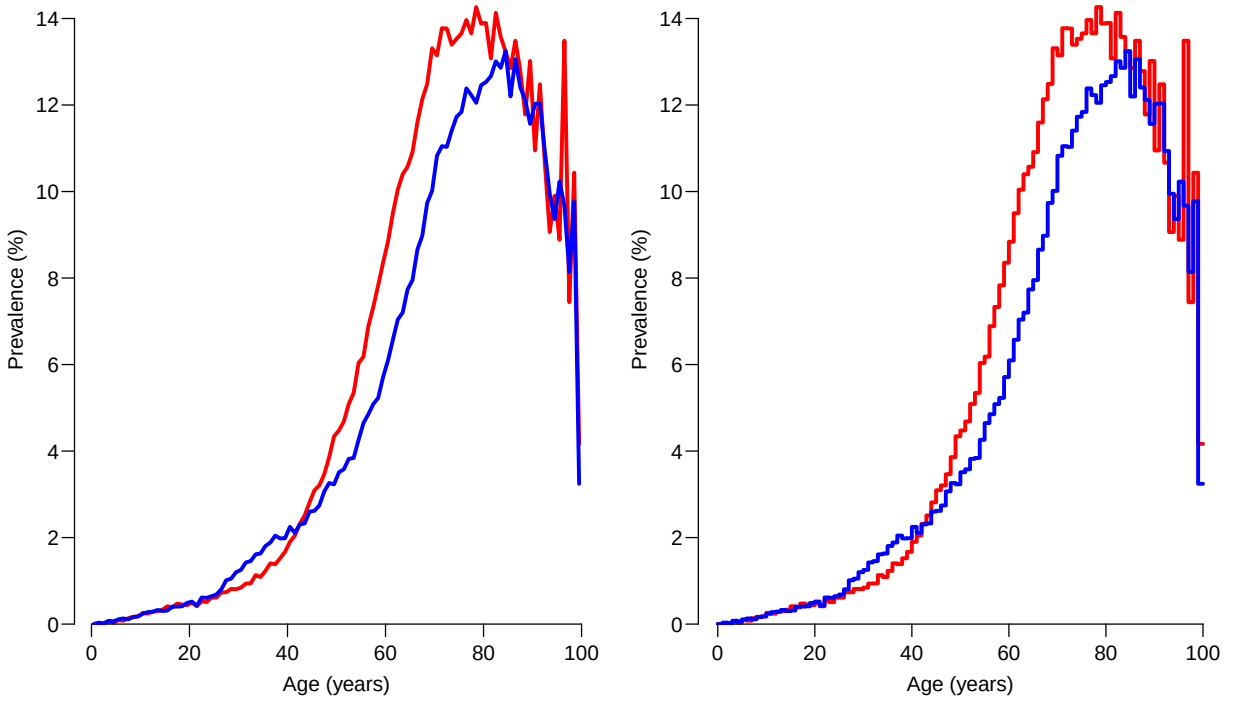

Figure 3.2: *Age-specific prevalence of diabetes at 1 January 2005 in 1-year age-classes in Denmark.*

From figure 3.2 we get broadly the same picture as from 3.1, but the curves are not "credible".

This is illustrates the differences between the empirical prevalences and the theoretical prevalences. From a biological/clinical point of view we would of course expect that the prevalence were a smooth function of time, pretty much as approximated by the left hand curve in figure 3.1.

10. How would you go about showing prevalence as a smooth function of age?

    It would be more logical to describe the original data by a smooth curve. Formally, this would require that we knew the exact ages for every person in the Danish population as of 1 January 2005 as well as the diabetes status; we could then model the 2.5 mill. 0/1 variables for men by a binomial model with some smooth age-effect. But we do not have access to these data, so we use the 1-year age classified data for

the register and the population. We are then formally making an assumption that prevalences are constant in 1-year age-classes, but we impose restrictions on relationship between the prevalences in the different age-classes.

The advantage of this is that we get a more credible relationship between (estimated theoretical) prevalence and age, and in particular one that we can reasonably use for *any* age, not only the midpoints of the intervals.

In practice this is done by fitting a binomial model with a smooth effect of age to the table of prevalent cases and total population using the age-midpoints. In R we need two-column matrix of affected and unaffected as response variable, so the second column must be computed as the population size *minus* the number of patients:

```
> A <- 0:99+0.5
> prM <- cbind(ptt[,"M"],pop[,"M"]-ptt[,"M"])
> prF <- cbind(ptt[,"F"],pop[,"F"]-ptt[,"F"])
> m.pr <- glm( prM ~ Ns(A,knots=seq(10,95,,9)), family=binomial )
> f.pr <- glm( prF ~ Ns(A,knots=seq(10,95,,9)), family=binomial )
```

`Ns` is a so called natural spline (restricted cubic spline) that specifies a smooth function of `A`.

From this model we can make predictions; in principle for *any* point on the age-scale, but in this case it suffices to do it at the midpoint of the age-categories in order to get a smoothly looking curve.

```
> nd <- data.frame( A=0:99+0.5 )
> par( mfrow=c(1,2), bty="n", las=1 )
> matplot( nd$A, cbind( ci.pred(m.pr,nd)[,1],
+                       ci.pred(f.pr,nd)[,1] )*100,
+          type="l", lty=1, lwd=3, col=c("blue","red"),
+          xlab="Age (years)", ylab="Prevalence (%)", las=1, yaxs="i", ylim=c(0,15) )
> matplot( nd$A, cbind( ci.pred(m.pr,nd)[,1],
+                       ci.pred(f.pr,nd)[,1] )*100,
+          type="s", lty=1, lwd=3, col=c("blue","red"),
+          xlab="Age (years)", ylab="Prevalence (%)", las=1, yaxs="i", ylim=c(0,15) )
```

The *modeling* of prevalences also illustrates the contrast between the *empirical* and *theoretical* prevalences; the former are necessarily tied to a particular grouping of the population; for example by sex and/or age, whereas the latter refer to *any* combination of sex and age; we can in principle refer to the prevalence of DM in women aged 68.3 years:

```
> ci.pred( f.pr, data.frame(A=68.3) )
    Estimate       2.5%      97.5%
1 0.09386903 0.09283319 0.09491521
```

This number cannot be derived as an empirical fraction from data; it is a *prediction* from a statistical model. It is our best guess at the probability that a woman aged 68.3 evaluated on 1 January 2005 has diabetes. The model is biologically plausible because the prediction for ages 68.2 and 68.4 are quite similar:

```
> ci.pred( f.pr, data.frame(A=c(68.2,68.3,68.4)) )
    Estimate       2.5%      97.5%
1 0.09344671 0.09241412 0.09448963
2 0.09386903 0.09283319 0.09491521
3 0.09429069 0.09325122 0.09534053
```

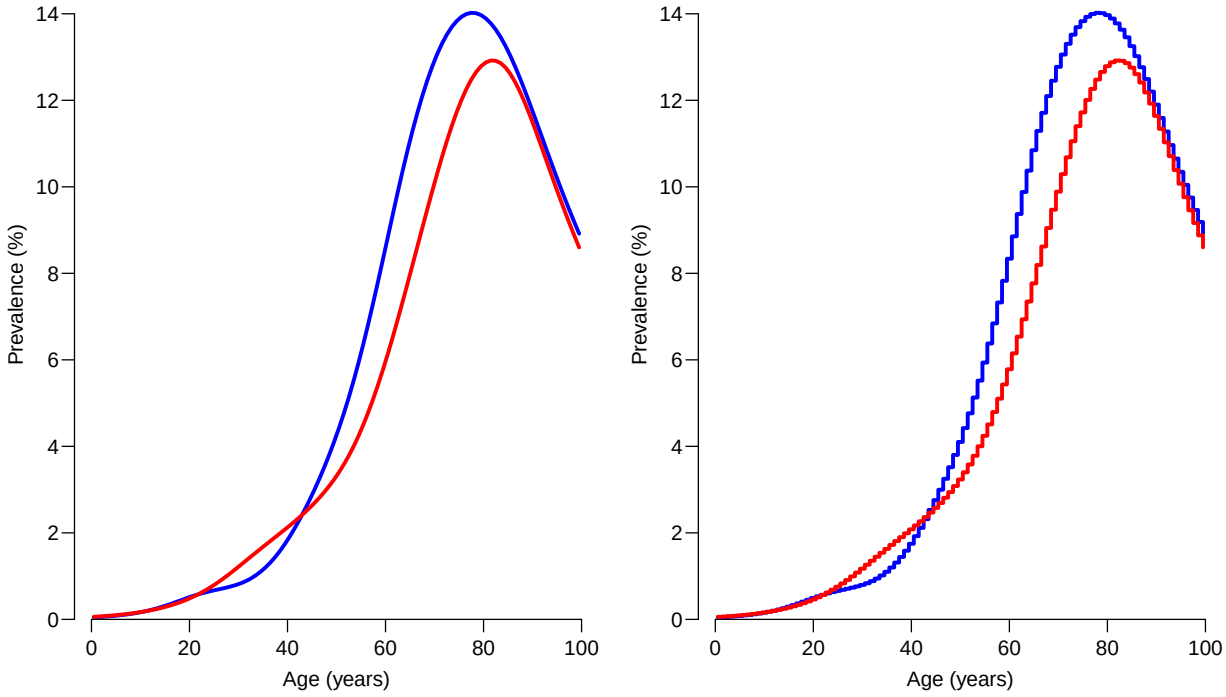

Figure 3.3: *Fitted age-specific prevalences from a binomial model with restricted cubic splines. The left panel is the predicted theoretical prevalence, the right hand plot is the formally fitted model with constant prevalence in each 1-year category and restrictions on the relationship between these.*

We see that we expect that women slightly older has a prevalence (*i.e.* probability of being affected) that is slightly higher too.

The *modeling* of prevalences also illustrates the contrast between the *empirical* and *theoretical* prevalences; the former are necessarily tied to a particular grouping of the population; for example by sex and/or age, whereas the latter refer to *any* combination of sex and age; after modelling we can in principle refer to the prevalence of DM in women aged 68.3 years or 73.6 years.

3.3 Incidence

The incidence (rate) of DM is defined as the number of new cases of DM that occur in a population in a predefined period of time. Of course the number of new DM cases is approximately twice as large if the population you look at is twice as large, but also if you look at the same population for two instead of one year; so the relevant denominator must be proportional *both* to the number of persons considered *and* the length of time considered. This is the population follow-up time — the person-years.

Enumeration of person-years among diabetes patients is a non-trivial task, but the total number of person-years in the population may be approximated from population counts at fixed dates, normally by taking averages of population size between two time points multiplied by the distance between the time points.

As in the case of prevalence we distinguish between the empirical and theoretical

incidence rates:

empirical incidence rate refers to a given time-period (and possibly also age-interval), and is defined as the number of new cases relative to the population risk time (person-years) in the time-period.

theoretical incidence rate is defined at any point in time as the probability of seeing an event (DM diagnosis, for example) in a small period of time *relative* to the length of this period.

3.3.1 Practical

1. How would you find number of newly diagnosed cases in age 60–64 (incl.) in the year 2006 from the Danish National Diabetes Register.

   We load the diabetes register as before, and compute the number of newly diagnosed cases in age 60–64 (incl.) in the year 2006:

```
> load( file="../data/dr.Rda" )
> nrow( subset( dr, floor(doDM)==2006 &
+                   (doDM-doBth)>=60 &
+                   (doDM-doBth)< 65 ) )
[1] 3480
```

2. The person-years in the Danish population is available in the dataset `Ydk`:

```
> # load( file=url("http://bendixcarstensen.com/Epi/Courses/IDEG2015/data/Ydk.Rda" )
> load( file="../data/Ydk.Rda" )
> str( Ydk )
'data.frame':        8400 obs. of  4 variables:
 $ sex: Factor w/ 2 levels "M","F": 1 2 1 2 1 2 1 2 1 2 ...
 $ A  : num  0 0 1 1 2 2 3 3 4 4 ...
 $ P  : num  1971 1971 1971 1971 1971 ...
 $ Y  : num  37139 35129 36134 34223 37113 ...
> head( Ydk )
  sex A    P        Y
1   M 0 1971 37139.17
2   F 0 1971 35128.83
3   M 1 1971 36133.67
4   F 1 1971 34223.00
5   M 2 1971 37113.00
6   F 2 1971 34926.33
```

   The person-years data is actually classified by single years and sex, but we just add them up:

```
> subset( Ydk, A>=60 & A<65 & P==2006 )
     sex  A    P        Y
7121   M 60 2006 40160.17
7122   F 60 2006 39678.17
7123   M 61 2006 38069.33
7124   F 61 2006 37952.83
7125   M 62 2006 35100.50
7126   F 62 2006 35562.67
7127   M 63 2006 32311.67
7128   F 63 2006 32980.00
7129   M 64 2006 29321.67
7130   F 64 2006 30069.83
```

```
> sum( subset( Ydk, A>59 & A<65 & P==2006 )$Y )
[1] 351206.8
```

Thus the incidence rate of diabetes among persons aged 60–64 is

```
> 3480 / 351206.8
[1] 0.009908692
```

per 1 person-year, or, if we want it per 1000 person-years

```
> 3480 / 351.2068
[1] 9.908692
```

so roughly speaking 1% per year.

3. This figure is for a single 5-year age-class and for both sexes. If we want the age-specific rates in 2006, in 5-year age-classes and by sex, we need a table of cases and person-years. Note that the count of cases is a table of how many records we have, whereas the person-years is a summation of the variable `Y`:

```
> ( D <- with( subset(dr,floor(doDM)==2006),
+              table( floor((doDM-doBth)/5)*5, sex ) ) )
     sex
        M    F
  0    26   29
  5    41   38
  10   79   85
  15   55  101
  20   73  101
  25  107  158
  30  178  275
  35  352  317
  40  597  608
  45  882  614
  50 1307  894
  55 1659 1152
  60 2038 1442
  65 1583 1342
  70 1195 1175
  75  950 1096
  80  634  834
  85  245  470
  90   70  153
  95   12   31
  100   0    2
```

```
> ( Y <- xtabs( Y ~ I(floor(A/5)*5) + sex, data=subset(Ydk,P==2006) ) )
                   sex
I(floor(A/5) * 5)          M          F
                0  166333.333 158679.833
                5  173061.833 164956.500
                10 180623.833 171379.833
                15 163695.000 154952.667
                20 149068.000 144681.333
                25 164809.667 163738.500
                30 191372.167 189887.167
                35 200951.833 194950.833
                40 212268.000 205365.000
                45 188801.833 184550.167
```

```
50 181232.333 179168.000
55 186422.833 186106.500
60 174963.333 176243.500
65 121788.167 129678.000
70  91038.500 105248.833
75  68313.833  88990.167
80  45502.167  73541.667
85  22305.833  47119.833
90   7295.500  20757.833
95   1413.667   6093.667
```

The register data has a few incident cases over 100 years, so we must cut those off before we look at the incidence rates. We multiply by 1000 in order to get rates per 1000 PY:

```
> D <- D[1:20,]
> round( inc <- D/Y * 1000, 1 )
```

```
    sex
       M    F
0    0.2  0.2
5    0.2  0.2
10   0.4  0.5
15   0.3  0.7
20   0.5  0.7
25   0.6  1.0
30   0.9  1.4
35   1.8  1.6
40   2.8  3.0
45   4.7  3.3
50   7.2  5.0
55   8.9  6.2
60  11.6  8.2
65  13.0 10.3
70  13.1 11.2
75  13.9 12.3
80  13.9 11.3
85  11.0 10.0
90   9.6  7.4
95   8.5  5.1
```

We can then plot the incidence rates, using both the interval midpoints and, for the sake of illustration, the formally fitted constant rates in each interval:

```
> par( mfrow=c(1,2), bty="n", las=1 )
> matplot( seq(2.5,97.5,5), inc,
+          type="l", lty=1, lwd=3, col=c("blue","red"),
+          log="y", xlab="Age (years)",
+          ylab="Incidence rate of DM 2006 (per 1000 PY)" )
> matplot( seq(0,100,5), inc[c(1:20,20),],
+          type="s", lty=1, lwd=3, col=c("blue","red"),
+          log="y", xlab="Age (years)",
+          ylab="Incidence rate of DM 2006 (per 1000 PY)" )
```

4. There is however no particular reason to choose 5-year intervals; we could as before use 1-year intervals, as population figures are actually available for these:

```
> D <- with( subset(dr,floor(doDM)==2006),
+            table( floor(doDM-doBth), sex ) )
> Y <- xtabs( Y ~ floor(A) + sex, data=subset(Ydk,P==2006) )
> D <- D[1:100,]
> round( cbind( D, Y, inc <- D/Y * 1000), 2 )
```

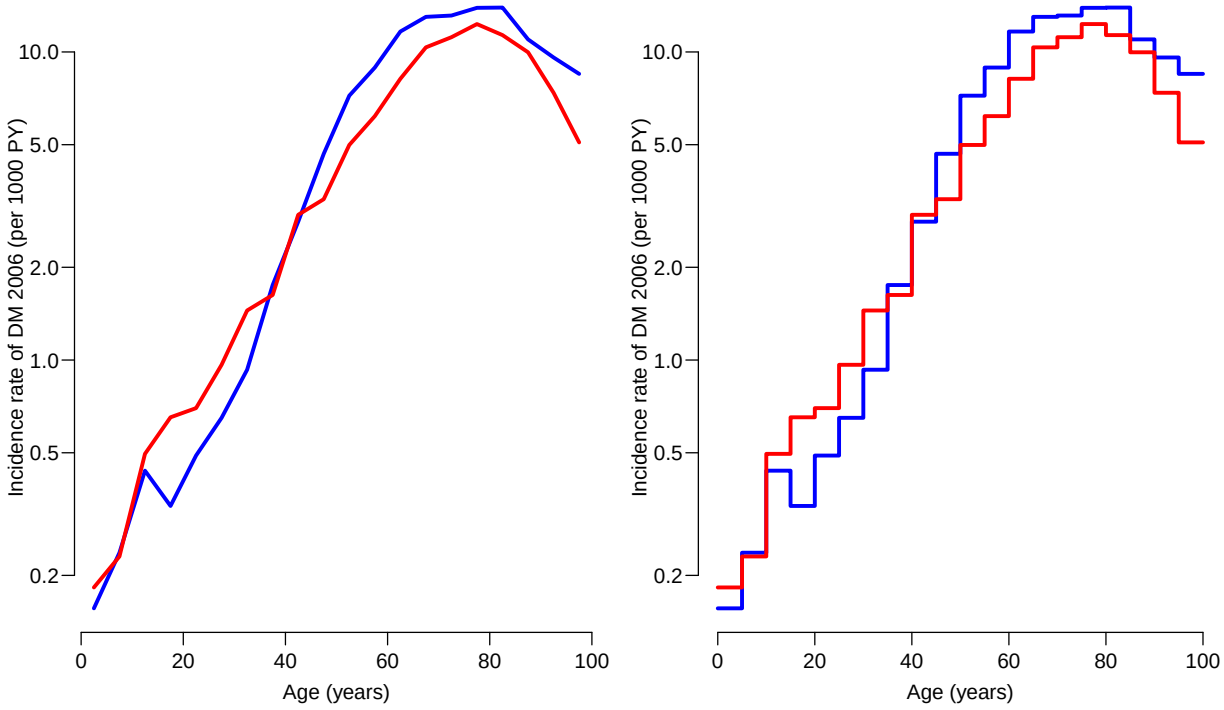

Figure 3.4: *Empirical incidence rates of DM in Denmark for 2006 in 5-year age classes. Left panel is the midpoint of the age-classes connected, right panel is the model formally fitted using constant incidence rates in 5-year intervals.*

	M	F	M	F	M	F
0	3	4	33241.00	31659.83	0.09	0.13
1	1	9	33124.50	31747.33	0.03	0.28
2	5	4	33302.00	31816.83	0.15	0.13
3	9	6	33277.17	31637.00	0.27	0.19
4	8	6	33388.67	31818.83	0.24	0.19
5	3	8	34074.67	32635.83	0.09	0.25
6	8	4	34384.67	32917.33	0.23	0.12
7	5	10	34339.50	32736.00	0.15	0.31
8	10	3	34876.00	33093.33	0.29	0.09
9	15	13	35387.00	33574.00	0.42	0.39
10	12	21	36174.83	34256.00	0.33	0.61
11	22	13	36841.83	35012.67	0.60	0.37
12	5	25	36306.33	34582.83	0.14	0.72
13	24	15	35958.33	34055.00	0.67	0.44
14	16	11	35342.50	33473.33	0.45	0.33
15	10	19	34474.83	32697.17	0.29	0.58
16	12	18	33893.50	32146.83	0.35	0.56
17	12	14	32854.83	31138.17	0.37	0.45
18	15	24	31633.00	29848.50	0.47	0.80
19	6	26	30838.83	29122.00	0.19	0.89
20	8	13	30429.17	29061.83	0.26	0.45
21	18	23	29754.83	28763.33	0.60	0.80
22	8	17	29232.50	28514.50	0.27	0.60
23	20	23	29630.17	28893.33	0.67	0.80
24	19	25	30021.33	29448.33	0.63	0.85
25	16	22	30851.17	30663.00	0.52	0.72
26	21	28	32235.33	32096.00	0.65	0.87
27	24	30	33258.83	33115.83	0.72	0.91
28	24	32	33840.67	33637.83	0.71	0.95
29	22	46	34623.67	34225.83	0.64	1.34

```
30  26  53 36809.50 36572.00  0.71  1.45
31  37  55 38003.17 37998.17  0.97  1.45
32  30  65 37945.67 37683.83  0.79  1.72
33  41  47 38888.17 38574.67  1.05  1.22
34  44  55 39725.67 39058.50  1.11  1.41
35  53  51 38852.83 37949.83  1.36  1.34
36  51  51 38024.17 37125.50  1.34  1.37
37  75  57 38910.83 37584.67  1.93  1.52
38  80  79 41073.50 39666.83  1.95  1.99
39  93  79 44090.50 42624.00  2.11  1.85
40  85 135 44961.33 43465.67  1.89  3.11
41 135 115 43806.83 42233.00  3.08  2.72
42 116 135 42871.83 41363.67  2.71  3.26
43 107 101 41174.83 39942.50  2.60  2.53
44 154 122 39453.17 38360.17  3.90  3.18
45 163 101 38890.00 37886.50  4.19  2.67
46 164 136 38074.17 37173.17  4.31  3.66
47 157 112 37292.00 36515.83  4.21  3.07
48 189 125 37264.17 36451.83  5.07  3.43
49 209 140 37281.50 36522.83  5.61  3.83
50 224 153 37160.17 36407.83  6.03  4.20
51 264 171 36385.33 35929.00  7.26  4.76
52 237 166 36201.83 35931.00  6.55  4.62
53 291 203 36085.33 35818.50  8.06  5.67
54 291 201 35399.67 35081.67  8.22  5.73
55 290 192 35495.00 35338.17  8.17  5.43
56 262 223 35652.50 35865.17  7.35  6.22
57 338 248 36450.00 36519.00  9.27  6.79
58 389 215 38503.00 38410.00 10.10  5.60
59 380 274 40322.33 39974.17  9.42  6.85
60 388 293 40160.17 39678.17  9.66  7.38
61 451 298 38069.33 37952.83 11.85  7.85
62 429 286 35100.50 35562.67 12.22  8.04
63 408 295 32311.67 32980.00 12.63  8.94
64 362 270 29321.67 30069.83 12.35  8.98
65 348 280 26801.33 27891.00 12.98 10.04
66 312 258 25460.33 26660.83 12.25  9.68
67 328 266 24371.83 25764.50 13.46 10.32
68 305 254 23235.33 25220.17 13.13 10.07
69 290 284 21919.33 24141.50 13.23 11.76
70 250 243 20640.17 22868.83 12.11 10.63
71 239 243 19323.50 21878.33 12.37 11.11
72 267 210 18106.17 20842.50 14.75 10.08
73 218 244 16935.33 20092.00 12.87 12.14
74 221 235 16033.33 19567.17 13.78 12.01
75 220 246 15281.67 18938.33 14.40 12.99
76 214 221 14468.83 18292.33 14.79 12.08
77 198 215 13816.00 17803.17 14.33 12.08
78 166 221 12874.33 17306.17 12.89 12.77
79 152 193 11873.00 16650.17 12.80 11.59
80 162 218 11074.00 16252.50 14.63 13.41
81 152 172 10139.67 15703.50 14.99 10.95
82 118 155  9121.17 14918.50 12.94 10.39
83 118 147  7989.00 13699.67 14.77 10.73
84  84 142  7178.33 12967.50 11.70 10.95
85  72 127  6433.33 12529.67 11.19 10.14
86  65 125  5267.33 10708.50 12.34 11.67
87  53  95  4246.00  8992.33 12.48 10.56
88  33  67  3525.17  7991.67  9.36  8.38
89  22  56  2834.00  6897.67  7.76  8.12
90  23  40  2265.83  5842.83 10.15  6.85
91  15  55  1801.67  4958.33  8.33 11.09
92   9  35  1417.33  4143.83  6.35  8.45
93  12  13  1046.67  3279.00 11.46  3.96
94  11  10   764.00  2533.83 14.40  3.95
95   3  14   518.17  1944.67  5.79  7.20
```

```
96   6  11   350.67  1421.00 17.11  7.74
97   2   3   224.50   984.50  8.91  3.05
98   1   1   149.83   743.50  6.67  1.34
99   0   2   170.50  1000.00  0.00  2.00
```

```
> par( mfrow=c(1,2), bty="n", las=1 )
> matplot( seq(0.5,99.5,1), inc,
+          type="l", lty=1, lwd=3, col=c("blue","red"),
+          log="y", xlab="Age (years)",
+          ylab="Incidence rate of DM 2006 (per 1000PY)" )
> matplot( seq(0,100,1), inc[c(1:100,100),],
+          type="s", lty=1, lwd=3, col=c("blue","red"),
+          log="y", xlab="Age (years)",
+          ylab="Incidence rate of DM 2006 (per 1000PY)" )
```

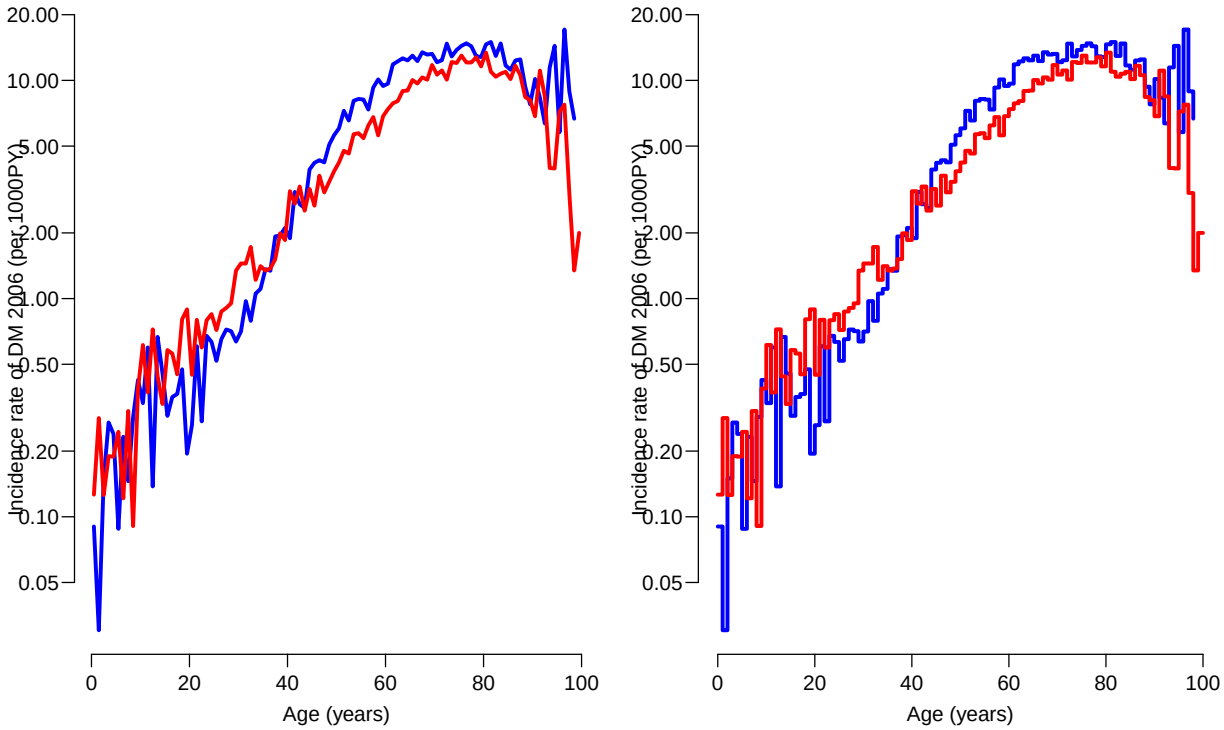

Figure 3.5: *Empirical incidence rates of DM in Denmark for 2006 in 1-year age classes. Left panel is the midpoint of the age-classes connected, right panel is the model formally fitted using constant incidence rates in 1-year intervals.*

Clearly the empirical rates in 1-year classes gives a very poor approximation to the age-specific rates; one would assume that these were a smooth function of age.

5. How would you go about fitting a model with a smooth age-effect for the incidence rates?

   We can fit a model that smooths the incidence rates. Unlike the prevalence data which were simple binomial (DM yes/no), the incidence rates are rate data. Under the assumption that rates are constant in intervals the model is a Poisson model, with the number of incident cases as outcome, and the log of the person-years as

offset[1]. For the splines and predictions we need some functions from the `Epi` package, so we must load this first:

```
> library( Epi )
> A <- 0:99+0.5
> d <- D[,"M"] ; y <- Y[,"M"] ;
> m.inc <- glm( d ~ Ns(A,knots=seq(10,95,,9)), offset=log(y), family=poisson )
> d <- D[,"F"] ; y <- Y[,"F"] ;
> f.inc <- glm( d ~ Ns(A,knots=seq(10,95,,9)), offset=log(y), family=poisson )
```

As before, `Ns` is a so called natural spline (restricted cubic spline) that specifies a smooth function of `A`.

From this model we can make predictions; in principle for *any* point on the age-scale, but in this case it suffices to do it at the midpoint of the age-categories in order to get a smoothly looking curve. Note that we also need to specify `y` as a variable in the prediction frame in order to get the rates in prespecified units (in this case per 1000 PY).

```
> nd <- data.frame( A=0:99+0.5, y=1000 )
> par( mfrow=c(1,2), bty="n", las=1 )
> matplot( nd$A, cbind( ci.pred(m.inc,nd)[,1],
+                       ci.pred(f.inc,nd)[,1] ),
+          type="l", lty=1, lwd=3, col=c("blue","red"),
+          xlab="Age (years)", ylab="Incidence of DM (per 1000 PY)", las=1, log="y" )
> matplot( nd$A, cbind( ci.pred(m.inc,nd)[,1],
+                       ci.pred(f.inc,nd)[,1] ),
+          type="s", lty=1, lwd=3, col=c("blue","red"),
+          xlab="Age (years)", ylab="Incidence of DM (per 1000 PY)", las=1, log="y" )
```

The data points used for fitting the models has one observation per one-year age-class, and hence must necessarily assume that the rates are constant in 1-year classes, but the model places restrictions on the relationship between the rates in each interval. The left graph in figure 3.6 shows the *theoretical* rates that on would infer from the model, whereas the right hand graph shows the formally fitted rates as being constant in each age-class.

3.3.2 Caveat: people only get DM once

In the calculations above we have used the total population risk time as denominator, even though more than 10% of the population over 60 years of age have diabetes. This mean that the rates of diabetes are underestimated because the person with diabetes are not at risk of getting diabetes. Thus the person-years should only be computed for persons without diabetes. One way of doing this is to compute the person-years among diabetes patients and subtract it from to total population person-years.

6. As an example we used the incidence in 2006; so we should compute the person-years among all diabetes patients contributing during 2006, and subdivide it by age class.

[1] A formally correct expression is that the *likelihood* for the rate parameter λ based on data (`D`,`Y`) is proportional to a likelihood for a Poisson variate `D` as observation and a mean equal to the rate (λ) multiplied by the person-years (`Y`). Note in particular that this does not imply an assumption that the data are Poisson distributed; there is not a one-to-one correspondence between models and likelihoods; two different models may have the same likelihood.

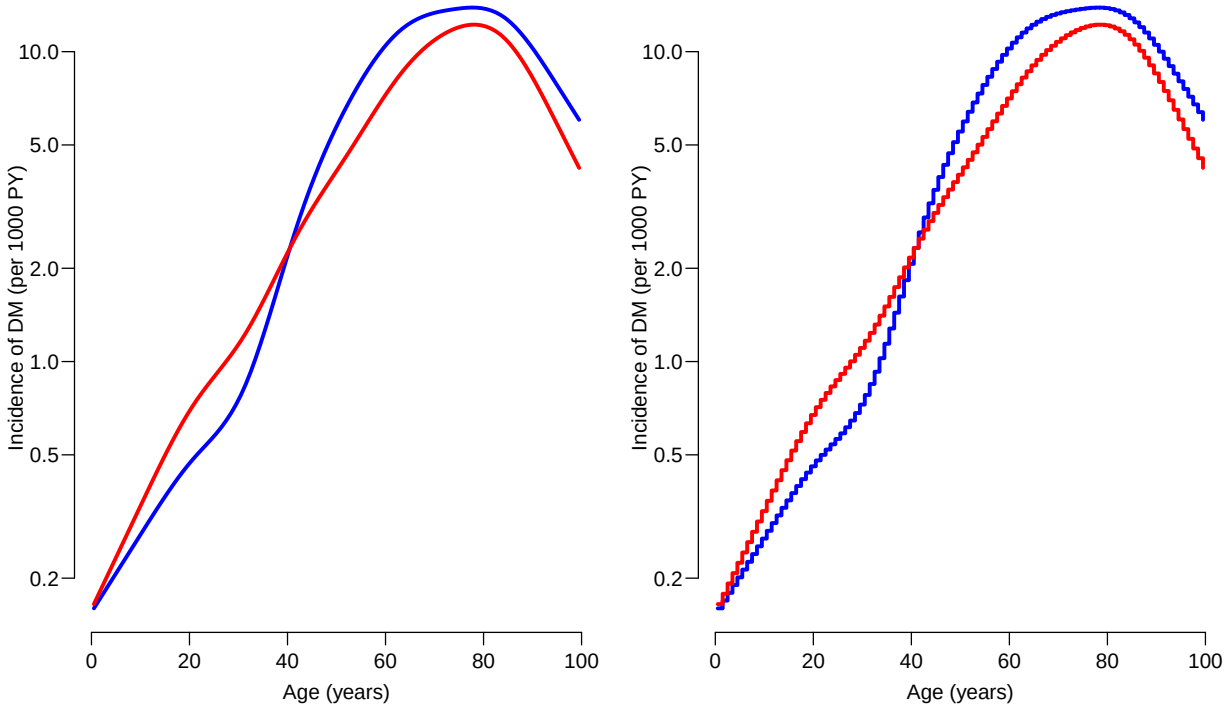

Figure 3.6: *Fitted age-specific incidence rates from a Poisson model with restricted cubic splines. The left panel is the predicted theoretical incidence rates, the right hand plot is the formally fitted model with constant incidence rate in each 1-year category and restrictions on the relationship between these.*

Programming-wise this is done by using a loop over sex and over age-classes. For each age-class we compute the last date of observation and subtract the first date of observation, but only *within* the calendar year 2006, that is from the date coded 2006.0 to the date coded 2007.0:

```
> dmY <- Y * 0
> for( sx in c("M","F") )
+ for( aa in 1:100 )
+ dmY[aa,sx] <- with( subset(dr,sex==sx),
+                     sum( pmax( 0, pmin(2007,doBth+aa  ,doDth,na.rm=TRUE) -
+                                   pmax(2006,doBth+aa-1,doDM ) ) ) )
```

We save this for later use:

```
> save( dmY, file="../data/dmY.Rda" )
```

7. We can see how large a percentage of the population risk time is among persons with DM:

```
> round( 100*t(dmY/Y), 1 )
   floor(A)
sex   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
  M 0.0 0.0 0.0 0.1 0.1 0.1 0.1 0.1 0.2 0.2 0.2 0.3 0.3 0.3 0.4 0.4 0.4
  F 0.0 0.0 0.0 0.1 0.1 0.1 0.1 0.2 0.2 0.2 0.2 0.3 0.3 0.4 0.3 0.4 0.4
   floor(A)
sex  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33
```

```
  M  0.4  0.5  0.5  0.5  0.5  0.5  0.5  0.6  0.6  0.7  0.7  0.9  0.8  0.9  1.0  1.0  1.0
  F  0.4  0.5  0.5  0.5  0.6  0.5  0.6  0.7  0.7  0.8  0.9  1.1  1.2  1.3  1.4  1.5  1.7
   floor(A)
sex   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48   49   50
  M  1.2  1.3  1.4  1.6  1.6  1.8  1.9  2.2  2.4  2.5  2.9  3.2  3.4  3.8  4.0  4.4  4.9
  F  1.7  1.8  1.9  2.1  2.2  2.2  2.3  2.5  2.6  2.5  2.7  2.8  3.1  3.1  3.4  3.6  3.8
   floor(A)
sex   51   52   53   54   55   56   57   58   59   60   61   62   63   64   65   66   67
  M  5.2  5.4  5.9  6.2  6.6  7.1  7.7  8.3  8.8  9.5 10.1 10.8 11.3 11.7 11.8 12.6 12.6
  F  3.9  4.3  4.4  4.5  4.8  5.2  5.6  5.8  6.1  6.4  6.9  7.3  7.9  8.3  8.8  9.0  9.6
   floor(A)
sex   68   69   70   71   72   73   74   75   76   77   78   79   80   81   82   83   84
  M 13.5 13.9 14.2 15.1 15.1 15.2 15.0 14.8 15.0 15.3 15.0 15.2 15.3 15.1 14.5 14.5 15.0
  F 10.2 10.7 11.3 11.8 12.4 12.2 12.8 13.2 13.1 13.6 13.5 13.5 13.5 13.8 13.5 13.7 13.9
   floor(A)
sex   85   86   87   88   89   90   91   92   93   94   95   96   97   98   99
  M 14.7 14.3 13.5 13.6 12.5 13.1 12.6 11.8 11.2  9.7 10.6 10.2  9.5  8.3  4.4
  F 13.9 14.0 13.5 13.3 12.7 12.2 12.4 12.6 12.1 10.5  9.6  9.4  9.6  7.6  3.6
```

So this is not at all a negligible fraction — and these fractions are quite close to the age-specific prevalences at the midpoint of 2006.

The moral here is that the risk time should only be computed among those who are at risk of the event. In many cancer studies, the fraction of the population alive with a given cancer is quite small so this correction is of little practical importance; but as we saw for diabetes, the correction is substantial.

8. We therefore re-estimate the the age-specific rate using the correct denominator:

```
> A <- 0:99+0.5
> d <- D[,"M"] ; y <- Y[,"M"] - dmY[,"M"]
> M.inc <- glm( d ~ Ns(A,knots=seq(10,95,,9)), offset=log(y), family=poisson )
> d <- D[,"F"] ; y <- Y[,"F"] - dmY[,"F"]
> F.inc <- glm( d ~ Ns(A,knots=seq(10,95,,9)), offset=log(y), family=poisson )
```

. . . and make a plot of the correctly estimated incidence rates (as well as the old ones for comparison.)

```
> nd <- data.frame( A=0:99+0.5, y=1000 )
> par( mfrow=c(1,2), bty="n", las=1 )
> matplot( nd$A, cbind( ci.pred(M.inc,nd)[,1],
+                       ci.pred(F.inc,nd)[,1],
+                       ci.pred(m.inc,nd)[,1],
+                       ci.pred(f.inc,nd)[,1] ),
+          type="l", lty=1, lwd=c(3,3,1,1), col=c("blue","red"),
+          xlab="Age (years)", ylab="Incidence of DM (per 1000 PY)", las=1, log="y" )
> matplot( nd$A, cbind( ci.pred(M.inc,nd)[,1],
+                       ci.pred(F.inc,nd)[,1],
+                       ci.pred(m.inc,nd)[,1],
+                       ci.pred(f.inc,nd)[,1] ),
+          type="l", lty=1, lwd=c(3,3,1,1), col=c("blue","red"),
+          xlab="Age (years)", ylab="Incidence of DM (per 1000 PY)", las=1 )
```

From figure 3.7 we see that the correction of the rates is quite substantial; it is largely in the order of magnitude of the age-specific prevalences, that is at the peak some 15%.

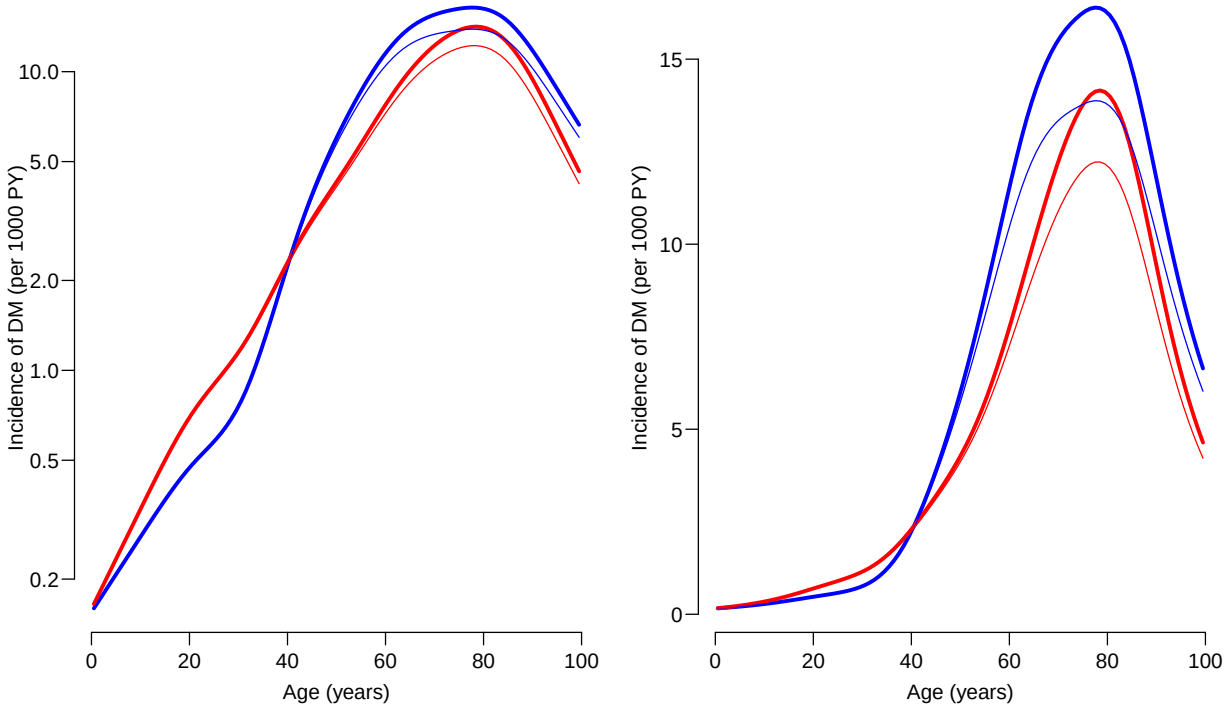

Figure 3.7: *Fitted age-specific incidence rates from a Poisson model with restricted cubic splines. The thick lines are estimates based on the correct follow-up time among persons without diabetes, the thin lines are based on the person-years for the entire population (which is wrong). The left panel is with a logarithmic y-axis; the right hand panel shows the same curves but on an untransformed scale.*

3.4 Mortality and survival

When we are talking about mortality rates, we have the same considerations as before regarding empirical and theoretical rates, but as a special feature of mortality we might also be interested in survival.

Survival is defined as the probability, $S(t)$ of being alive after some specified length of time, t. This is a *cumulative* measure that requires an *origin*, that is, t must be defined as time *since* some origin.

In the case of diabetes it will normally be time since diagnosis of diabetes. The survival is a function of the mortality rates, so in order to compute the survival function at different times after diagnosis, we must know the mortality rates as a function of time since diagnosis.

Mortality rates however, is naturally also dependent on age — possibly both on age at diagnosis of DM as well as current age. The latter is the sum of age at diagnosis and the time since diagnosis (duration). So we are facing the problem of describing mortality by time since diagnosis of DM, age at diagnosis of DM as well by the sum of the two. The linear effects of the three variables cannot be separated, but the non-linear effects can.

1. As a start we can compute mortality rates among diabetes patients, say during the year 2006. Above we computed the person-years among diabetes patients by age and sex in 2006 in 1-year intervals, in order to subtract these from the total population

person-years. But this will also be the denominator (person-years) for the mortality. So we just need the number of deaths among diabetes patients classified by age (at death) and sex:

```
> library( Epi )
> # load( url("http://bendixcarstensen.com/Epi/Courses/IDEG2015/data/dr.Rda") )
> ### The local version on this computer
> load( file="../data/dr.Rda" )
> load( file="../data/dmY.Rda" )
> dd <- with( subset( dr, floor(doDth)==2006 ),
+             table( floor(doDth-doBth), sex ) )
> str( dd )
 'table' int [1:88, 1:2] 0 1 1 0 2 0 1 1 0 2 ...
 - attr(*, "dimnames")=List of 2
  ..$    : chr [1:88] "0" "12" "18" "20" ...
  ..$ sex: chr [1:2] "M" "F"

> dmD <- dmY * 0 # devise a table of 0s with same structure as the person-years
> for( aa in intersect( dimnames(dd )[[1]], # fill in deaths where they are
+                       dimnames(dmD)[[1]] ) )
+      dmD[aa,] <- dd[aa,]
```

We can then show the number of cases, person-years, and rates per 1000 PY:

```
> cbind( dmD, round( dmY, 1 ), round( 1000*dmD/dmY, 2 ) )
     M   F      M      F      M       F
0    0   2    0.6    0.9   0.00 2263.80
1    0   0    2.2    5.2   0.00    0.00
2    0   0   10.0   14.7   0.00    0.00
3    0   0   17.1   16.4   0.00    0.00
4    0   0   31.4   26.4   0.00    0.00
5    0   0   25.4   36.5   0.00    0.00
6    0   0   34.7   40.8   0.00    0.00
7    0   0   41.3   51.1   0.00    0.00
8    0   0   52.6   57.0   0.00    0.00
9    0   0   72.6   60.8   0.00    0.00
10   0   0   85.4   85.3   0.00    0.00
11   0   0   97.3   92.0   0.00    0.00
12   1   0  120.0  111.8   8.33    0.00
13   0   0  115.8  129.1   0.00    0.00
14   0   0  129.6  114.7   0.00    0.00
15   0   0  131.9  122.0   0.00    0.00
16   0   0  137.7  121.6   0.00    0.00
17   0   0  146.4  137.9   0.00    0.00
18   1   0  157.8  142.4   6.34    0.00
19   0   0  159.4  148.5   0.00    0.00
20   0   1  149.7  154.7   0.00    6.46
21   2   0  159.8  167.8  12.51    0.00
22   0   1  142.5  147.8   0.00    6.76
23   1   0  158.2  181.1   6.32    0.00
24   1   0  166.4  201.8   6.01    0.00
25   0   1  198.0  219.7   0.00    4.55
26   2   0  217.1  241.8   9.21    0.00
27   3   0  238.0  292.7  12.60    0.00
28   1   1  289.0  354.5   3.46    2.82
29   1   1  291.2  420.7   3.43    2.38
30   0   0  337.3  465.1   0.00    0.00
31   2   1  369.9  535.5   5.41    1.87
32   2   0  388.3  575.7   5.15    0.00
33   1   0  399.5  644.9   2.50    0.00
34   2   2  463.3  670.0   4.32    2.98
35   2   0  503.5  697.2   3.97    0.00
36   3   2  519.7  691.1   5.77    2.89
37   6   1  608.6  777.8   9.86    1.29
```

```
38   5   5  675.9  891.3   7.40    5.61
39   6   2  776.7  953.1   7.72    2.10
40   6   5  851.7 1005.1   7.04    4.97
41   8   3  943.1 1043.2   8.48    2.88
42  12   4 1014.2 1074.5  11.83    3.72
43  13   5 1032.9  993.8  12.59    5.03
44  15   9 1143.2 1051.5  13.12    8.56
45  14   8 1236.8 1072.0  11.32    7.46
46  11   7 1311.8 1154.3   8.39    6.06
47  16   8 1408.1 1116.4  11.36    7.17
48  19  10 1486.5 1230.9  12.78    8.12
49  19  11 1638.4 1326.3  11.60    8.29
50  29  10 1806.6 1398.7  16.05    7.15
51  25   9 1889.1 1401.1  13.23    6.42
52  34  17 1965.9 1530.4  17.29   11.11
53  25  14 2145.0 1591.8  11.66    8.80
54  26  20 2198.5 1573.0  11.83   12.71
55  38  21 2356.7 1693.0  16.12   12.40
56  43  27 2547.2 1881.0  16.88   14.35
57  44  29 2794.7 2036.3  15.74   14.24
58  63  37 3185.5 2237.9  19.78   16.53
59  89  39 3534.6 2443.9  25.18   15.96
60  82  39 3830.6 2543.2  21.41   15.34
61  90  45 3853.9 2623.8  23.35   17.15
62 136  71 3782.6 2608.7  35.95   27.22
63 119  57 3657.9 2619.4  32.53   21.76
64 103  39 3444.8 2499.6  29.90   15.60
65  97  63 3175.0 2444.4  30.55   25.77
66 104  60 3218.1 2411.2  32.32   24.88
67 147  74 3067.7 2472.0  47.92   29.94
68 136  67 3135.5 2565.0  43.37   26.12
69 131  84 3043.9 2588.7  43.04   32.45
70 141  90 2935.7 2583.9  48.03   34.83
71 145  93 2919.2 2572.7  49.67   36.15
72 150 101 2729.5 2578.2  54.96   39.18
73 147 123 2572.7 2454.8  57.14   50.11
74 169 101 2408.7 2506.7  70.16   40.29
75 180 129 2268.5 2500.5  79.35   51.59
76 175 127 2177.3 2400.9  80.37   52.90
77 197 141 2113.8 2421.6  93.20   58.23
78 183 172 1935.1 2343.8  94.57   73.38
79 185 174 1799.5 2248.9 102.81   77.37
80 212 196 1689.5 2199.2 125.48   89.13
81 195 154 1533.7 2165.9 127.15   71.10
82 197 203 1326.6 2017.5 148.50  100.62
83 186 209 1157.2 1880.7 160.73  111.13
84 149 201 1080.3 1803.6 137.92  111.44
85 161 227  944.8 1736.7 170.41  130.71
86 145 195  753.9 1501.5 192.34  129.87
87 123 195  571.6 1213.5 215.17  160.69
88 118 190  479.9 1059.3 245.90  179.37
89  96 182  353.2  876.9 271.78  207.55
90  76 148  297.9  710.0 255.11  208.44
91  63 152  226.4  616.4 278.31  246.61
92  72 136  167.5  521.1 429.72  260.98
93  50  94  117.0  395.9 427.22  237.40
94  29  87   74.1  265.9 391.14  327.25
95  19  58   54.9  186.9 346.22  310.37
96  17  51   35.9  134.1 473.73  380.27
97   7  44   21.3   94.5 328.77  465.58
98   9  22   12.4   56.3 727.79  390.66
99   3  20    7.5   36.0 399.59  555.37
```

The above table shows that the mortality rates are very variable, particularly in the younger ages, due to the small number of deaths.

2. We can plot the mortality rates in two different ways as we did for the incidence rates:

```
> par( mfrow=c(1,2), bty="n", las=1 )
> matplot( 0:99+0.5, 1000*dmD/dmY,
+          type="l", lty=1, lwd=c(3,3,1,1), col=c("blue","red"),
+          xlab="Age (years)", ylab="Mortality of DM patients (per 1000 PY)",
+          ylim=c( 1, 500), las=1, log="y" )
> matplot( 0:99+0.5, 1000*dmD/dmY,
+          type="l", lty=1, lwd=c(3,3,1,1), col=c("blue","red"),
+          xlab="Age (years)", ylab="Mortality of DM patients (per 1000 PY)",
+          ylim=c( 0, 500), las=1 )
```

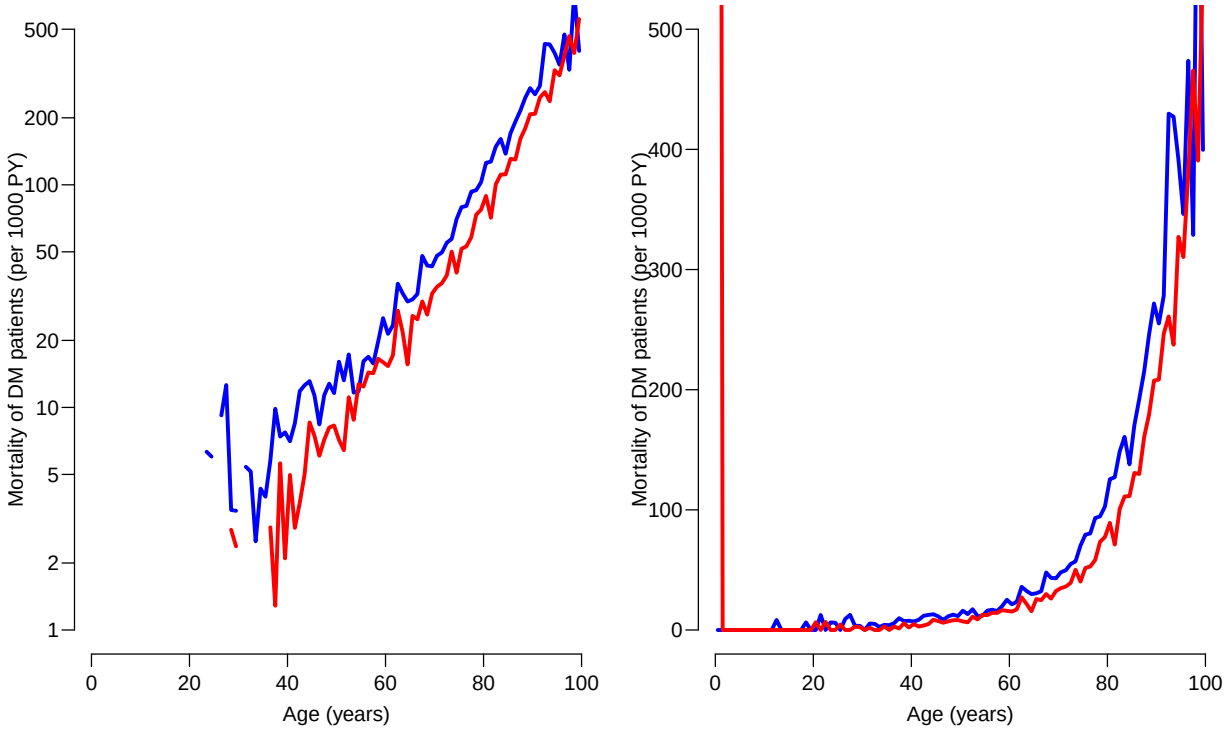

Figure 3.8: *Age-specific mortality rates in Danish DM patients in 2006. The plot on the log-scale is leaving out rates that are numerically equal to 0.*

3. As we did for the incidence rates, it is also possible to make a smooth model for how mortality depends on age:

```
> A <- 0:99+0.5
> d <- dmD[,"M"] ; y <- dmY[,"M"] ;
> m.mort <- glm( d ~ Ns(A,knots=seq(10,95,,9)), offset=log(y), family=poisson )
> d <- dmD[,"F"] ; y <- dmY[,"F"] ;
> f.mort <- glm( d ~ Ns(A,knots=seq(10,95,,9)), offset=log(y), family=poisson )
> nd <- data.frame( A=0:99+0.5, y=1000 )
> par( mfrow=c(1,2), bty="n", las=1 )
> matplot( nd$A, cbind( ci.pred(m.mort,nd)[,1],
+                       ci.pred(f.mort,nd)[,1] ),
+          type="l", lty=1, lwd=3, col=c("blue","red"), ylim=c(0.5,500),
+          xlab="Age (years)", ylab="Mortality among DM patients (per 1000 PY)", las=1, log="y
> matplot( nd$A, cbind( ci.pred(m.mort,nd)[,1],
+                       ci.pred(f.mort,nd)[,1] ),
+          type="s", lty=1, lwd=3, col=c("blue","red"), ylim=c(0.5,500),
+          xlab="Age (years)", ylab="Mortality among DM patients (per 1000 PY)", las=1, log="y
```

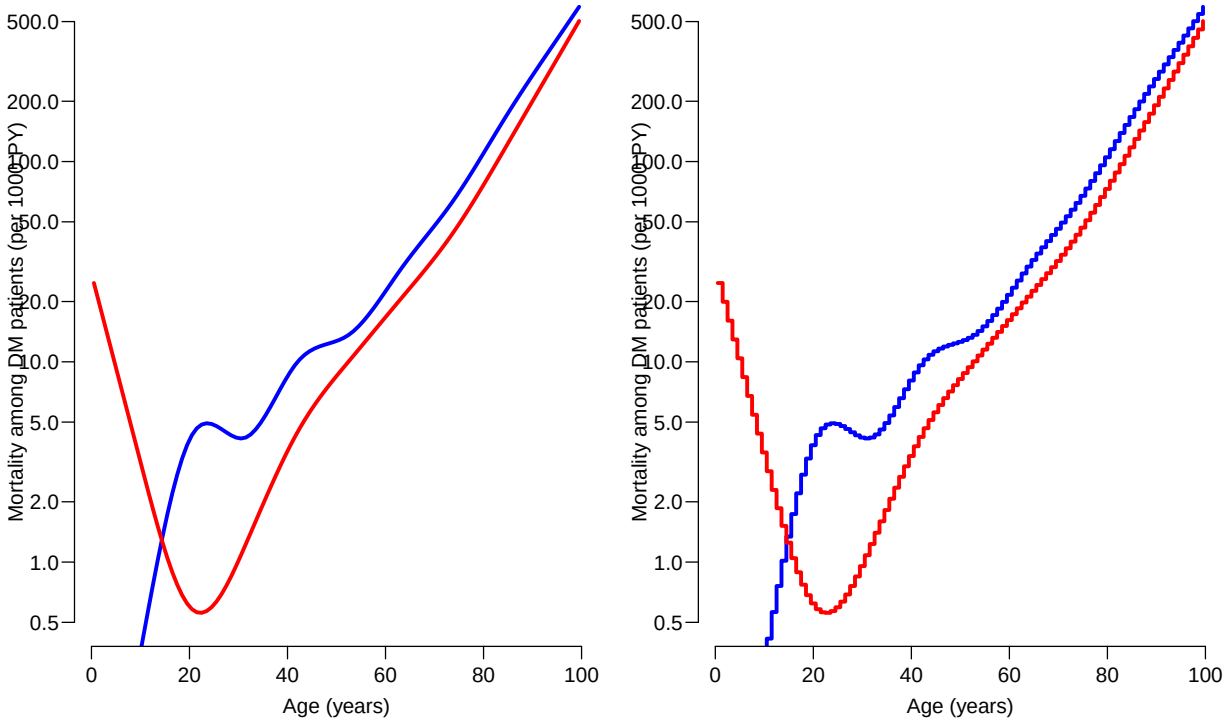

Figure 3.9: *Fitted age-specific mortality rates from a Poisson model with restricted cubic splines. The left panel is the predicted theoretical incidence rates, the right hand plot is the formally fitted model with constant incidence rate in each 1-year category and restrictions on the relationship between these.*

```
> par( mfrow=c(1,2), bty="n", las=1 )
> matplot( nd$A, cbind( ci.pred(m.mort,nd),
+                       ci.pred(f.mort,nd) ),
+          type="l", lty=1, lwd=c(3,1,1),
+          col=rep(c("blue","red"),each=3), ylim=c(0.5,500),
+          xlab="Age (years)", ylab="Mortality among DM patients (per 1000 PY)", las=1, log="y
> matplot( nd$A, cbind( ci.pred(m.mort,nd),
+                       ci.pred(f.mort,nd) ),
+          type="s", lty=1, lwd=c(3,1,1),
+          col=rep(c("blue","red"),each=3), ylim=c(0.5,500),
+          xlab="Age (years)", ylab="Mortality among DM patients (per 1000 PY)", las=1, log="y
```

From figure 3.10 it is clear that modeling may also produce unrealistic results; the mortality curves for women in the youngest ages are based on very few deaths below age 25 (see above for a listing of deaths observed). This also means that the approximations underlying the calculations of the confidence intervals are not valid, so that the confidence intervals shown in figure 3.10 are invalid for ages under 40. So there is no basis for claiming that women have higher mortality in the very young ages — it is based on two deaths among 0-old girls.

4. We could also look at mortality as a function of duration of DM. However this would really only make sense if we controlled for age in some way. So for the sake of the argument we do the calculation for persons diagnosed in age 60 in the period after 1995:

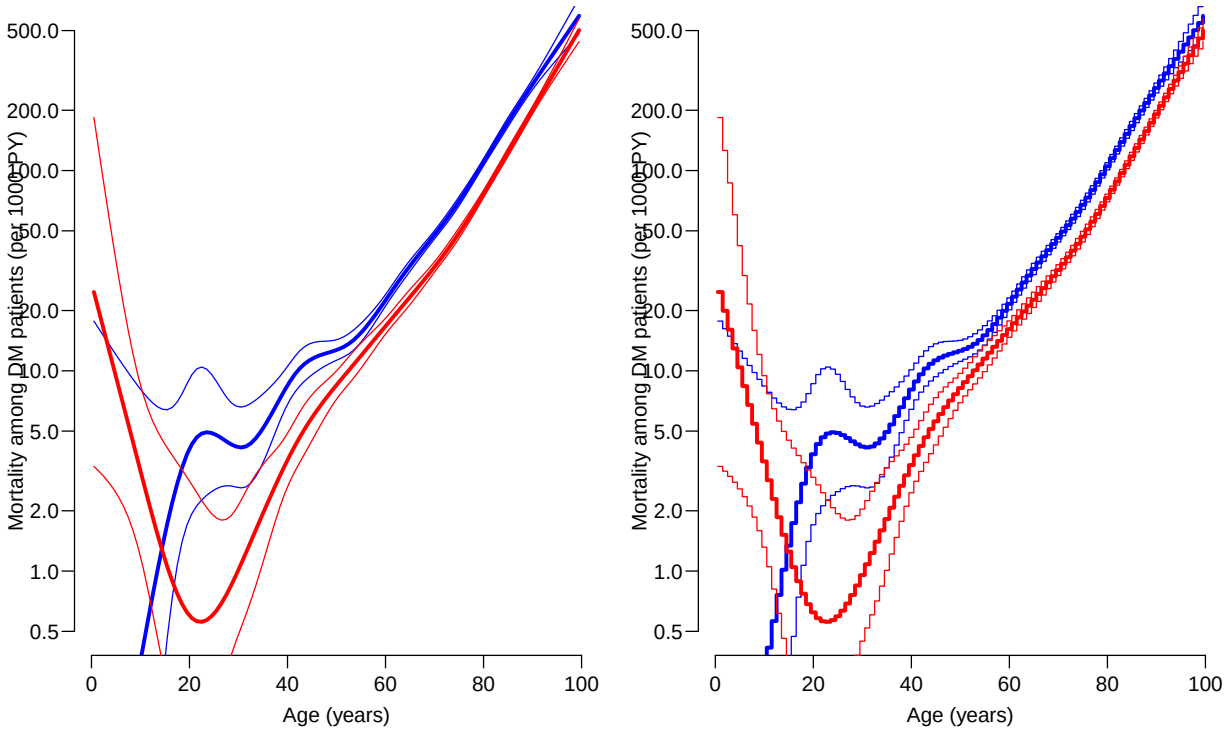

Figure 3.10: *Fitted age-specific mortality rates from a Poisson model with restricted cubic splines. Thin lines indicate 95% confidence intervals. The left panel is the predicted theoretical incidence rates, the right hand plot is the formally fitted model with constant incidence rate in each 1-year category and restrictions on the relationship between these.*

```
> dr60 <- subset( dr, floor(doDM-doBth)== 60 &
+                           doDM       >1995 )
> D60 <- with( dr60, table( floor(doDth-doDM), sex ) )
```

Thus we have the number of deaths among persons diagnosed in age 60 by single year of follow-up. It then remains to enumerate the person-years in these duration classes:

```
> Y60 <- D60 * 0
> for( sx in dimnames(Y60)[[2]] )
+ for( dd in dimnames(Y60)[[1]] )
+ Y60[dd,sx] <- with( subset( dr60, sex==sx ),
+                     sum( pmax( pmin(2012,   # end of FU
+                                     doDth,  # in duration dd
+                                     doDM+as.numeric(dd)+1,
+                                     na.rm=TRUE) -
+                                (doDM+as.numeric(dd)), # start of FU
+                                0 ) ) ) # discard negative FU
> cbind( D60, round(Y60,1), round(1000*D60/Y60,1) )
     M   F      M      F    M    F
0  178 117 5323.4 3821.6 33.4 30.6
1  121  61 4749.5 3446.6 25.5 17.7
2   99  59 4220.9 3109.0 23.5 19.0
3   97  48 3744.7 2773.2 25.9 17.3
4   81  43 3280.8 2453.7 24.7 17.5
5   96  39 2832.6 2137.8 33.9 18.2
6   72  33 2402.2 1861.3 30.0 17.7
7   89  41 1950.7 1560.2 45.6 26.3
```

```
8   58  31 1572.6 1261.4 36.9 24.6
9   37  26 1221.6 1004.3 30.3 25.9
10  39  15  957.8  797.7 40.7 18.8
11  43  18  740.2  640.4 58.1 28.1
12  25  19  552.6  487.0 45.2 39.0
13  28  18  383.2  335.6 73.1 53.6
14  14  13  244.1  207.0 57.3 62.8
15   6   2  137.2  118.0 43.7 17.0
16   4   2   40.6   37.2 98.4 53.7
```

We can illustrate the mortality as a function of diabetes duration:

```
> par( mfrow=c(1,2), bty="n", las=1 )
> matplot( 0:16+0.5, 1000*D60/Y60,
+          type="l", lty=1, lwd=3, col=c("blue","red"),
+          xlab="Diabetes duration (years)", ylab="Mortality of DM patients (per 1000 PY)",
+          ylim=c( 10, 120), las=1, log="y" )
> matplot( 0:16+0.5, 1000*D60/Y60,
+          type="l", lty=1, lwd=3, col=c("blue","red"),
+          xlab="Diabetes duration (years)", ylab="Mortality of DM patients (per 1000 PY)",
+          ylim=c( 0, 120), las=1 )
```

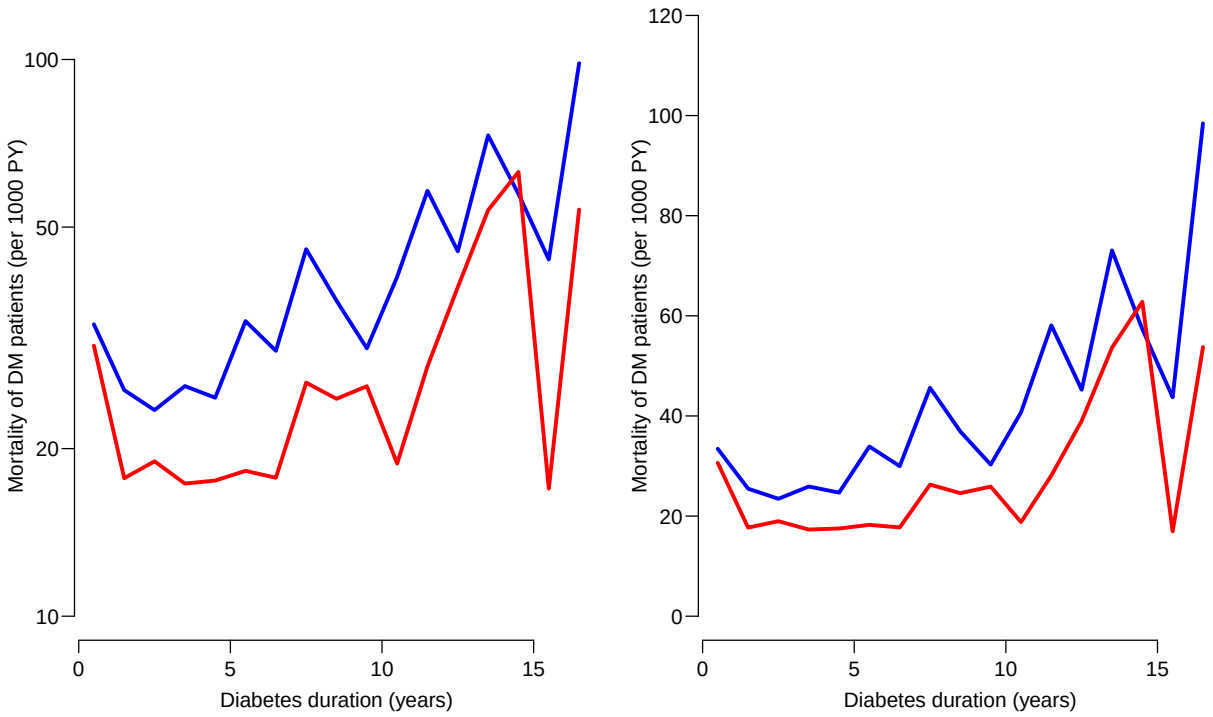

Figure 3.11: *Mortality among Danish 60 year old diabetes patients diagnosed 1995–2011 as a function of duration of diabetes. Left panel is with a logarithmic y-axis, right panel with untransformed y-axis.*

We see in figure 3.11 that the mortality rates are very variable for longer durations of diabetes, due to the very small number of deaths.

3.4.1 Survival

We can devise a so called life-table survival curve from the mortality rates; if the mortality in an interval is λ and the interval length is ℓ the probability of dying in the interval is

approximately $\lambda\ell$ — provided that the death probability is not too large (the correct expression is $1 - \exp(-\lambda\ell)$). Thus, the probability of surviving the interval is $1 - \lambda\ell$.

So the probability of surviving the first interval (that starts at time 0) is $1 - \lambda_0\ell$. The probability or surviving the next is $1 - \lambda_1\ell$ — or more precisely, the *conditional* probability of surviving the second interval *given* that the person already survived the first one. Hence the probability of surviving till the end of the second interval is $(1 - \lambda_0\ell) \times (1 - \lambda_1\ell)$. So we have $S(0) = 1$, $S(1) = 1 - \lambda_0\ell$, $S(2) = (1 - \lambda_0\ell) \times (1 - \lambda_1\ell)$, etc.

5. So we have the mortality rates as `D60/Y60` in units of deaths per 1 person-year, and since the intervals we have been using are 1-year intervals, the numbers can also be taken as the 1-year death probabilities for each interval. Thus we can compute the (conditional) survival probabilities and the survival function as:

```
> ( p60 <- 1 - D60/Y60 )
    sex
           M         F
0  0.9665625 0.9693842
1  0.9745237 0.9823014
2  0.9765453 0.9810227
3  0.9740970 0.9826913
4  0.9753110 0.9824754
5  0.9661094 0.9817572
6  0.9700272 0.9822702
7  0.9543746 0.9737208
8  0.9631180 0.9754235
9  0.9697122 0.9741110
10 0.9592825 0.9811961
11 0.9419072 0.9718911
12 0.9547620 0.9609833
13 0.9269371 0.9463654
14 0.9426528 0.9371992
15 0.9562704 0.9830491
16 0.9015634 0.9462515
> ( S60 <- rbind( 1, apply( p60, 2, cumprod ) ) )
           M         F
   1.0000000 1.0000000
0  0.9665625 0.9693842
1  0.9419381 0.9522275
2  0.9198452 0.9341568
3  0.8960185 0.9179877
4  0.8738967 0.9019004
5  0.8442798 0.8854472
6  0.8189743 0.8697484
7  0.7816083 0.8468921
8  0.7527811 0.8260784
9  0.7299810 0.8046921
10 0.7002580 0.7895607
11 0.6595780 0.7673671
12 0.6297401 0.7374269
13 0.5837294 0.6978753
14 0.5502542 0.6540482
15 0.5261918 0.6429616
16 0.4743952 0.6084034
```

6. An alternative way of computing the survival function(s) is to use the Kaplan-Meier estimator, which requires that we define an observed survival time for each person, as well as an indicator of whether follow-up (the survival time) ended by censoring or death. For illustration we plot the two approaches next to each other:

```
> par( mfrow=c(1,2) )
> matplot( 0:17, S60,
+          type="l", lty=1, lwd=3, col=c("blue","red"),
+          xlab="Diabetes duration (years)",
+          ylab="Survival of 60 year old DM patients",
+          ylim=c(0,1), las=1, yaxs="i", xaxs="i" )
> library( survival )
> dr60 <- transform( dr60, st = pmin( doDth-doDM,
+                                     2012-doDM,
+                                     na.rm=TRUE ),
+                          dd = !is.na( doDth ) )
> km <- survfit( Surv( st, dd ) ~ sex, data=dr60 )
> plot( km, col=c("blue","red"), mark.time=FALSE, lwd=3,
+           xlab="Diabetes duration (years)",
+           ylim=c(0,1), las=1, yaxs="i" )
> matlines( 0:17, S60, type="l", lty=1, lwd=3, col=c("blue","red") )
> matlines( 0:17, S60, type="l", lty=1, lwd=1, col="white" )
```

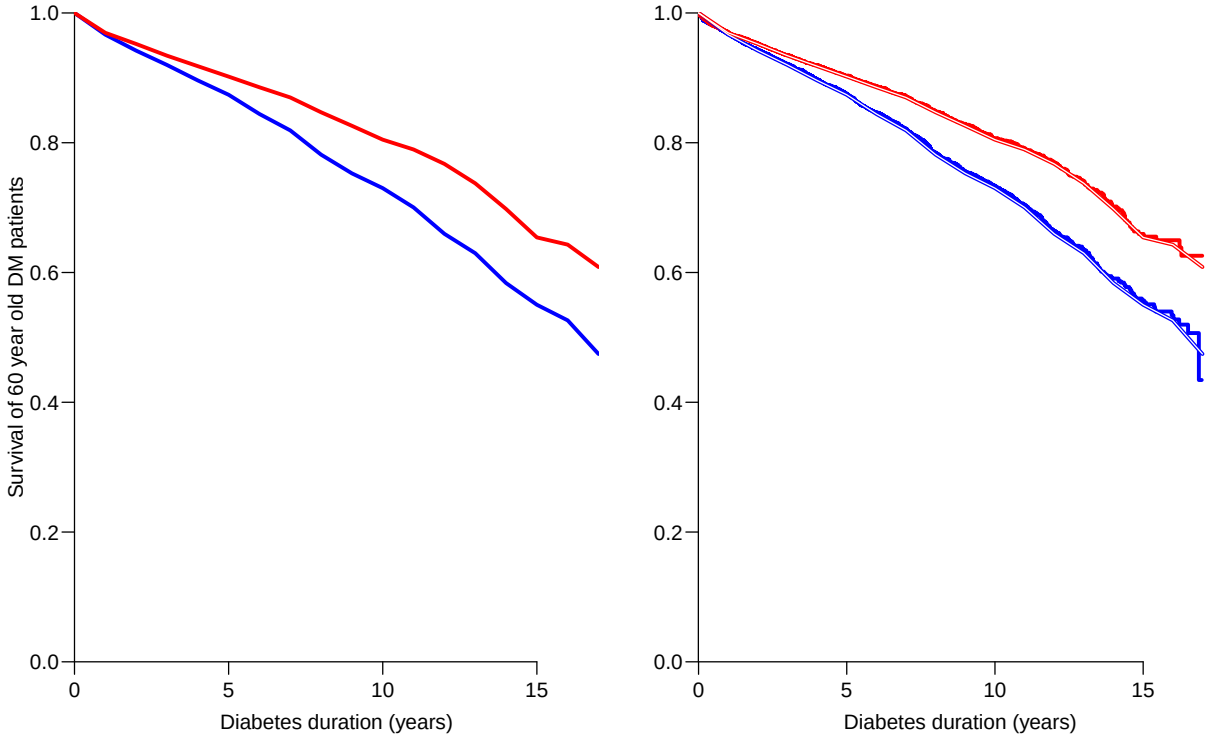

Figure 3.12: *Left: Actuarial survival curve for Danish diabetes patients diagnosed in age 60. Right: Kaplan-Meier survival curves overlaid with the actuarial curves.*

From figure **??** it is evident that the two methods in large data sets like this gives the same results. Even if *mortality* rates are very variable as a function of time since DM diagnosis, the survival curves seem more stable — this is a consequence of the fact that the survival function is a *cumulative* measure.

7. What we did was to compute the mortality in 1-year interval of diabetes duration for patients diagnosed in age 60 (that is between their 60^{th} and 61^{st} birthdays). We could of course repeat the exercise for persons diagnosed in ages 50, 51, ..., 99 to get an impression of how mortality and survival depend on age at diagnosis.

   To illustrate how age at diagnosis and time since diagnosis *simultaneously* influence mortality we need a proper model for the mortality. However it would be prudent

first to contemplate how to report the mortality of DM patients *both* as a function of age and duration of diabetes.

One possibility would be to show the mortality as a function of the patients' current age, but draw a separate curve for each age at diagnosis. So for persons diagnosed at age 50 we would show the mortality as a curve that starts at age 50, and gives the mortality by increasing duration of diabetes and hence also by increasing age. Similar curves could then be drawn for persons diagnosed at age 55, 60 etc. to give an impression of how age at diagnosis and duration of diabetes influence mortality.

The practical implementation of this is out of the scope of this stream, but in the next section is shown how it can be done. The main purpose being to illustrate the type of results achieved.

3.5 Mortality, age at diagnosis, duration and current age

In order to manipulate follow-up of DM patients we set up a Lexis object to handle it. A Lexis object is merely a data frame for follow-up data that allows us easily to keep track of multiple timescales (and multiple states)

```
> load( file="../data/dr.Rda" )
> library( Epi )
> Lx <- Lexis( entry = list( per = doDM,
+                            age = doDM-doBth,
+                            dur = 0 ),
+              exit = list( per = pmin(doDth,2012,na.rm=TRUE) ),
+       exit.status = factor( !is.na(doDth), labels=c("Alive","Dead") ),
+              data = subset( dr, doDM>1995 ) )
NOTE: entry.status has been set to "Alive" for all.
```

Each record in this Lexis object represents the follow-up of a single person; person no, 8 has been followed 7.5 years from 1996.97 or age 88.98.

```
> subset( Lx, lex.id==8 )
       per      age dur lex.dur lex.Cst lex.Xst lex.id sex    doBth    doDM    doIns
36 1996.97 88.97916   0 7.54202   Alive    Dead      8   F 1907.991 1996.97 1998.725
      doDth
36 2004.512
```

We can also summarize how much follow-up time is available in total:

```
> summary( Lx )
Transitions:
     To
From     Alive  Dead  Records:  Events: Risk time:  Persons:
  Alive 275868 95614    371482    95614    2198768    371482
```

In order to model mortality by varying age and duration, we must subdivide follow-up of persons in small intervals and assign an age, a date and a duration to each interval. We can then fit a model for mortality as a function of the variables.

We subdivide data using `splitLexis`:

```
> system.time(
+ Sx <- splitLexis( Lx, #[1:50000,],
+                    breaks=c(0:12/4,4:20),
+                    time.scale="dur" ) )
```

```
   user  system elapsed
 41.338   0.505  41.837
> summary( Sx )
Transitions:
     To
From      Alive  Dead  Records:  Events: Risk time:  Persons:
  Alive 5022194 95614   5117808    95614    2198768    371482
```

Thus we see that the number of events and the total risk time is the same as before, but the number of records has increased from 371,482 (one record per person) to 5,117,808 (one record per follow-up interval).

```
> addmargins( table( table( Sx$lex.id ) ) )
     1      2      3      4      5      6      7      8      9     10     11     12
 15658  12690  12198  10730   9634   9014   9569   9385   8654   8204   8466   8847
    13     14     15     16     17     18     19     20     21     22     23     24
 32232  28805  26019  23664  23219  21612  18667  15671  13595  11863  10396   8597
    25     26    Sum
  7639   6454 371482
```

The 5 mill. records in the dataset represent the follow-up of the 371,482 persons with a diagnosis of diabetes after 1995; each person has a differing no of records, for example 32,232 persons have 13 records, 6,454 have 26 records and 12,198 have 3 records.

We can illustrate this by listing the records belonging to individual no. 8; we see that the three time-scales as well as the interval lengths (`lex.dur`) vary during follow-up.

```
> subset( Sx, lex.id==8 )[,-16]
    lex.id     per      age  dur   lex.dur lex.Cst lex.Xst sex    doBth    doDM    doIns
91       8 1996.97 88.97916 0.00 0.2500000   Alive   Alive   F 1907.991 1996.97 1998.725
92       8 1997.22 89.22916 0.25 0.2500000   Alive   Alive   F 1907.991 1996.97 1998.725
93       8 1997.47 89.47916 0.50 0.2500000   Alive   Alive   F 1907.991 1996.97 1998.725
94       8 1997.72 89.72916 0.75 0.2500000   Alive   Alive   F 1907.991 1996.97 1998.725
95       8 1997.97 89.97916 1.00 0.2500000   Alive   Alive   F 1907.991 1996.97 1998.725
96       8 1998.22 90.22916 1.25 0.2500000   Alive   Alive   F 1907.991 1996.97 1998.725
97       8 1998.47 90.47916 1.50 0.2500000   Alive   Alive   F 1907.991 1996.97 1998.725
98       8 1998.72 90.72916 1.75 0.2500000   Alive   Alive   F 1907.991 1996.97 1998.725
99       8 1998.97 90.97916 2.00 0.2500000   Alive   Alive   F 1907.991 1996.97 1998.725
100      8 1999.22 91.22916 2.25 0.2500000   Alive   Alive   F 1907.991 1996.97 1998.725
101      8 1999.47 91.47916 2.50 0.2500000   Alive   Alive   F 1907.991 1996.97 1998.725
102      8 1999.72 91.72916 2.75 0.2500000   Alive   Alive   F 1907.991 1996.97 1998.725
103      8 1999.97 91.97916 3.00 1.0000000   Alive   Alive   F 1907.991 1996.97 1998.725
104      8 2000.97 92.97916 4.00 1.0000000   Alive   Alive   F 1907.991 1996.97 1998.725
105      8 2001.97 93.97916 5.00 1.0000000   Alive   Alive   F 1907.991 1996.97 1998.725
106      8 2002.97 94.97916 6.00 1.0000000   Alive   Alive   F 1907.991 1996.97 1998.725
107      8 2003.97 95.97916 7.00 0.5420202   Alive    Dead   F 1907.991 1996.97 1998.725
       doDth
91  2004.512
92  2004.512
93  2004.512
94  2004.512
95  2004.512
96  2004.512
97  2004.512
98  2004.512
99  2004.512
100 2004.512
101 2004.512
102 2004.512
103 2004.512
104 2004.512
105 2004.512
106 2004.512
107 2004.512
```

Each record can be made to represent a term in the total likelihood for a model of mortality for patients as a function of age at diagnosis (`age`−`dur`), current age (`age`), duration `dur` and calendar time `per`. The model assumes that mortality is constant in each of the small intervals, but places a restriction on the *size* of the mortality in each interval; it is a continuous function of age, duration and age at diagnosis.

As a small utility we load a function that shrinks the size of the glm objects without influencing the ability to predict from the model.

```
> source( "shrink.glm.R" )
> system.time(
+ mm1 <- glm( lex.Xst=="Dead" ~ Ns( age, knots=seq(10,90,,5)) +
+                                Ns( dur, knots=c(0,1,3,10)) +
+                                Ns( I(age-dur), knots=seq(40,90,,5) ),
+                   offset = log(lex.dur),
+                   family = poisson, model=FALSE, y=FALSE,
+                     data = subset( Sx, sex=="M" ) ) )
> mf1 <- update( mm1, data = subset( Sx, sex=="F" ) )
> mm1 <- shrink.glm( mm1 )
> mf1 <- shrink.glm( mf1 )
> save( Sx, mm1, mf1, file="tmp.Rda" )

  name       mode     class            lg/dim     size(K)
1 dr         list     data.frame       497232 5    17483.2
2 lls        function function         1              18.9
3 Lx         list     Lexis data.frame 371482 12   30477.4
4 mf1        list     glm lm           22         539252.8
5 mm1        list     glm lm           22         580391.0
6 shrink.glm function function         1              10.1
7 Sx         list     Lexis data.frame 5117808 12 399833.3
```

Once these models have been fitted separately for men and woman we can predict the mortality rates (per 1000 PY) for persons diagnosed at ages 40, 45, ..., 75 years of age for durations 0–16 years (which is the range of duration in the dataset).

Note that we do not bother too much about the parametrization — the model is overparametrized because of the linear relationship between the variables. We are only interested in the prediction (and they are correct, despite the warnings):

```
> nd <- data.frame( dur = rep(c(NA,seq(0,16,,50)),8),
+                   adg = rep(8:15*5,each=51),
+               lex.dur = 1000 )[-1,]
> nd$age <- nd$adg + nd$dur
> head( nd )
        dur adg lex.dur      age
2 0.0000000  40    1000 40.00000
3 0.3265306  40    1000 40.32653
4 0.6530612  40    1000 40.65306
5 0.9795918  40    1000 40.97959
6 1.3061224  40    1000 41.30612
7 1.6326531  40    1000 41.63265
> prm <- ci.pred( mm1, nd )
> prf <- ci.pred( mf1, nd )
> par( mfrow=c(1,2), bty="n", las=1 )
> matplot( nd$age, cbind( prm, prf ),
+          lwd=c(3,1,1), lty=1,
+          col=rep(c("blue","red"),each=3), type="l",
+          log="y",
+          xlab="Age at follow-up",
+          ylab="Mortality among DM patients" )
> matplot( nd$age, cbind( prm, prf ),
+          lwd=c(3,1,1), lty=1,
+          col=rep(c("blue","red"),each=3), type="l",
+          xlab="Age at follow-up",
+          ylab="Mortality among DM patients" )
```

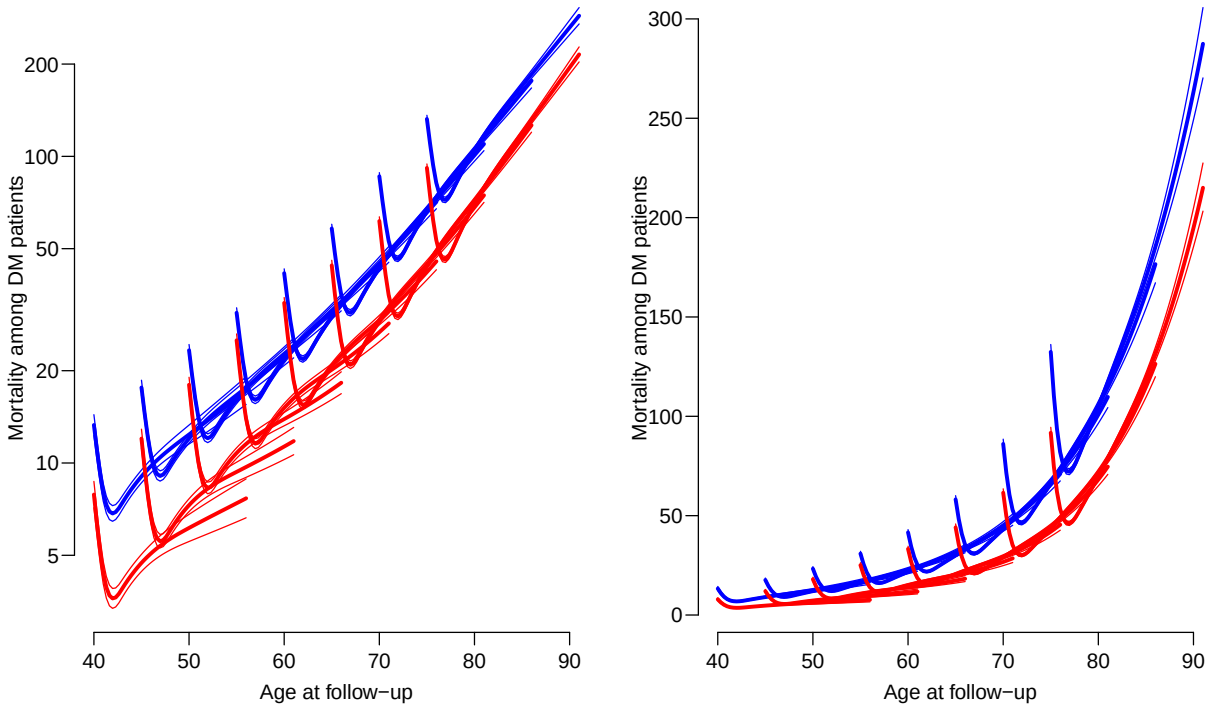

Figure 3.13: *Predicted mortality rates among Danish diabetes patients diagnosed 1995–2011 in different ages. Estimates are from a model with smooth effects of current age, duration and age at diagnosis. Blue curves are for men, red curves for women.*

From the figure 3.13 it is seen that duration has a dramatic effect on mortality, but only during the first two years; mortality drops by a factor of almost 2 during these first years, and then picks up at the usual age-pace, although there is an indication that women diagnosed at younger ages (below 60) seem to have a smaller mortality than women diagnosed later in life (at comparable ages, that is).

3.5.0.1 Interaction

The modeling can be used to explore:

- whether the duration effect is age-dependent and
- whether the effect of age at diagnosis is confounded by calendar time.

Hence we expand the model with calendar time, using 2005 as reference point, and with a simple interaction between duration and age at diagnosis:

```
> system.time(
+ mm2 <- update( mm1, . ~ . + Ns( per, knots=1995+seq(2,15,,4), ref=2005 )
+                            + Ns( dur, knots=c(0,1,10)):Ns( I(age-dur), knots=seq(40,90,,3) ) ) )
   user  system elapsed
 78.425   4.109  92.319
> system.time(
+ mf2 <- update( mf1, . ~ . + Ns( per, knots=1995+seq(2,15,,4), ref=2005 )
+                            + Ns( dur, knots=c(0,1,10)):Ns( I(age-dur), knots=seq(40,90,,3) ) ) )
   user  system elapsed
 67.780   1.076  73.805
```

```
> # shrink the objects
> mm2 <- shrink.glm( mm2 )
> mf2 <- shrink.glm( mf2 )
> nd <- cbind( nd, per=2005 )
```

We can show the calendar time effect as estimated relative to 2005:

```
> p.pt <- seq(1995,2012,,50)
> Cp <- Ns( p.pt, knots=1995+seq(2,15,,4), ref=2005 )
> RRm <- ci.exp( mm2, subset="per", ctr.mat=Cp )
> RRf <- ci.exp( mf2, subset="per", ctr.mat=Cp )
> matplot( p.pt, cbind(RRm,RRf),
+          lwd=c(3,1,1), lty=1,
+          col=rep(c("blue","red"),each=3), type="l",
+          xlab="Date of follow-up", ylim=c(0.5,2), log="y",
+          ylab="Mortality RR (reltive to 2005) among DM patients" )
> abline( h=1 )
> abline( h=c(5:15/10,2), v=1995:2012, col=gray(0.8) )
> matlines( p.pt, cbind(RRm,RRf),
+           lwd=c(3,1,1), lty=1,
+           col=rep(c("blue","red"),each=3), type="l" )
> points( 2005, 1, pch=1, cex=1.3, lwd=5 )
```

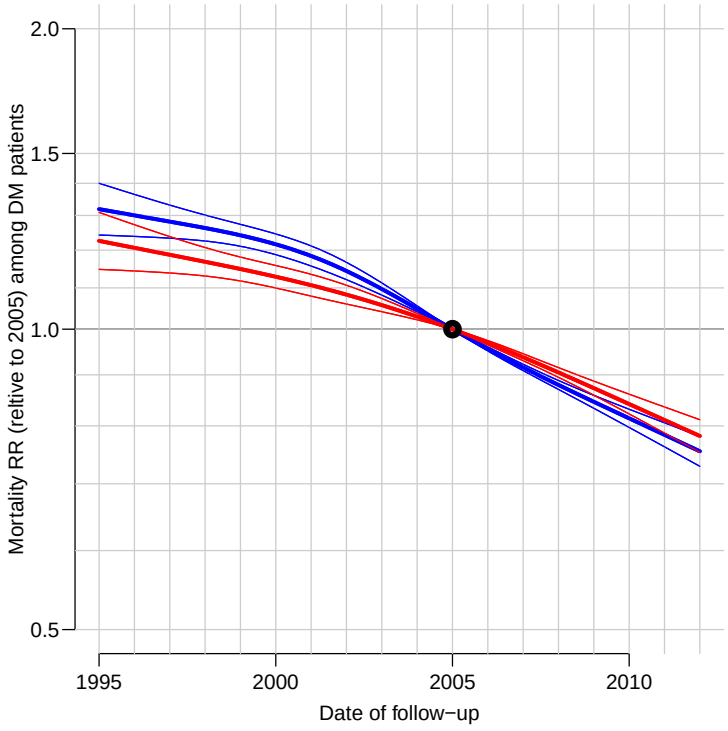

Figure 3.14: *Mortality rate-ratio for men and women respectively, relative to 2005*

From figure 3.14 we see that there is a reduction in mortality among diabetes patients of some 40% over the period, from 1.3 to 0.75 for men and from 1.2 to 0.8 for women.

When we re-do the prediction of the mortality as a function of age, we can do it in a simplified way by fixing the date to 2005, by including calendar time in the prediction, by making a prediction for persons diagnosed in different ages at year 1998 (say) as in figure 3.15:

```
> nd$per <- 2005
> prm <- ci.pred( mm2, nd )
> prf <- ci.pred( mf2, nd )
> par( mfrow=c(2,2) )
> matplot( nd$age, cbind( prm, prf ),
+          lwd=c(3,1,1), lty=1,
+          col=rep(c("blue","red"),each=3), type="l",
+          log="y",
+          xlab="Age at follow-up",
+          ylab="Mortality among DM patients (2005)" )
> matplot( nd$age, cbind( prm, prf ),
+          lwd=c(3,1,1), lty=1,
+          col=rep(c("blue","red"),each=3), type="l",
+          xlab="Age at follow-up",
+          ylab="Mortality among DM patients (2005)" )
> nd$per <- 1998+nd$dur
> prm <- ci.pred( mm2, nd )
> prf <- ci.pred( mf2, nd )
> matplot( nd$age, cbind( prm, prf ),
+          lwd=c(3,1,1), lty=1,
+          col=rep(c("blue","red"),each=3), type="l",
+          log="y",
+          xlab="Age at follow-up",
+          ylab="Mortality among DM patients (diag 1998)" )
> matplot( nd$age, cbind( prm, prf ),
+          lwd=c(3,1,1), lty=1,
+          col=rep(c("blue","red"),each=3), type="l",
+          xlab="Age at follow-up",
+          ylab="Mortality among DM patients (diag 1998)" )
```

In figure 3.15 we see that the conclusion about the effect of age at diagnosis depends on whether we evaluate it with or without a varying period effect.

It is however very clear that there is a markedly higher mortality in the first year or so after diagnosis — presumably an artifact because some very ill persons are diagnosed with diabetes as consequence of other illness, and therefore over-represented among newly diagnosed patients.

If we fix the calendar time, we see that the long-term effect of age at diagnosis among women is negligible; the mortality in different ages is virtually the same regardless of the age at diagnosis. Men, however have higher mortality the younger they are diagnosed.

If we evaluate the joint effect of age, duration *and* calendar time we see no effect of age at diagnosis for men, but that women diagnosed with DM in young ages have smaller mortality than women diagnosed at older age — when compared at the same age.

The calendar time effect we saw in figure 3.14 was roughly log-linearly decreasing by calendar time, slightly steeper for men than for women. Therefore, the difference from the upper to the lower panels in figure 3.15 is that the curves are tilted a bit downward; slightly more for men that for women.

Thus if we are willing to accept an overall decrease in mortality unrelated to diabetes, we will base conclusion on the top panel and conclude that the younger men are at diabetes diagnosis, the higher their mortality at a given age, whereas age at diagnosis has very little effect for women.

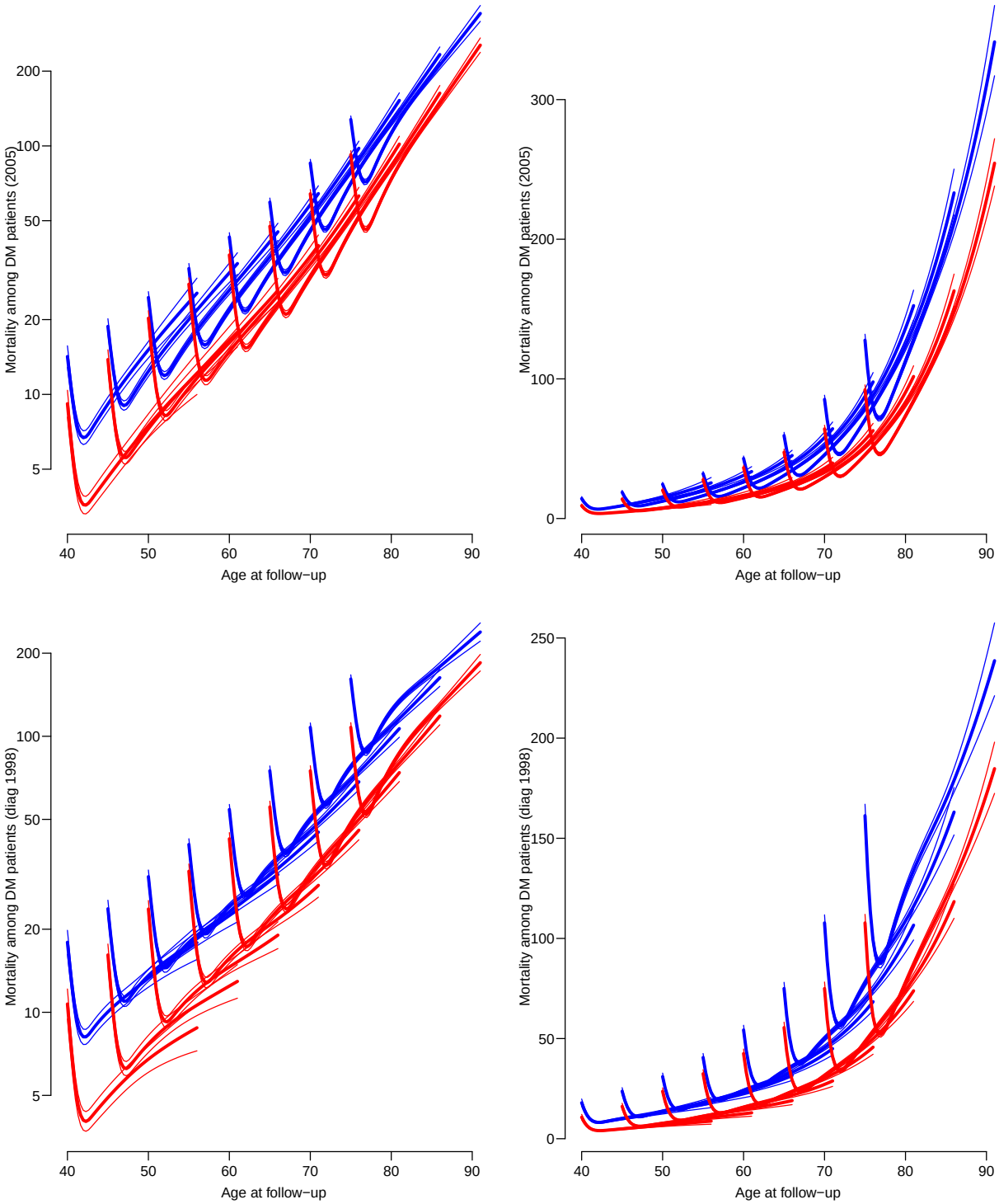

Figure 3.15: *Mortality of diabetes patients diagnosed in ages 40, 45, . . . , 75. The top panels are using 2005 as fixed calendar time, the lower panels showing patients diagnosed in at 1.1.1998 following patients over calendar time.*

References