

Computer exercise 3

Failure Intensities and the Poisson Distribution

First we will investigate some demographic data from Norway. Then we will learn how to simulate a Poisson process by means of exponentially distributed random numbers. At the conclusion, we will estimate accident rates (constant ones) from data where accidents in British coal mines have been recorded for a long succession of years.

1 Preparatory exercises

Read the instructions for the computer exercise and chapter 2.6, 7.1 and 7.4 in the textbook.

Question 1: For a Poisson stream of events (a Poisson process on the line) with constant intensity λ , the number of events in an interval $[s, s + t]$ is $N \in \text{Po}(\lambda t)$. Use this to show that the time T_1 of occurring of the first event is exponentially distributed with expectation $1/\lambda$. *Hint:* First explain that $T_1 > t$ if and only if $N(t) = 0$ and then compute the cdf of T_1 .

Question 2: Given a death-rate function $\lambda(t) = \alpha + \beta \exp((t - t_0)/c)$, with parameters as on page (iii) of this exercise, compute the probability that a person alive on her 30th birthday will reach the age of 65.

2 Reliability, failure rate, and expectation

Let T denote a non-negative random variable¹, and let the corresponding distribution function and density function be $F(t) = \text{P}(T \leq t)$ and $f(t)$ respectively. Typically, T is the lifetime of some system. The word lifetime is generic and should be understood in a wide sense: it could be a lifetime of a light bulb, a distance covered by a motor car, an amount of goods produced by a machine, etc, before the light bulb, motor car, or production machine, etc, breaks down. The function $R(t) = 1 - F(t)$ is called the *reliability* of the system considered. The *failure rate function* $\lambda(t)$ (occasionally called *hazard-rate function*) is defined as

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{R(t)}, \quad t > 0, R(t) \neq 0.$$

¹that is to say that $\text{P}(T < 0) = 0$.

If $\lambda(t)$ is given, but not $R(t)$ or $F(t)$, then use

$$R(t) = \exp\left(-\int_0^t \lambda(\tau) d\tau\right), \quad t > 0. \quad (1)$$

Note that a constant failure-rate function $\lambda(t) = \lambda_0$ implies that T is exponentially distributed: $T \in \text{Exp}(1/\lambda_0)$. The expectation $E(T)$ can be determined from the equality

$$E(T) = \int_0^\infty R(\tau) d\tau. \quad (2)$$

3 Norwegian demographic data

When T is the lifetime of a living creature, $R(t)$ and $\lambda(t)$ are often called the *survival function* and the *death-rate function* respectively. In the two data files² `NorwayFemale1846-2009.txt`, `NorwayMale1846-2009.txt`, the Norwegian annual so-called *life tables* for women and men from 1846 until 2009 are included. Please, use **Import Data** option in **File** > pull down menu at the left-hand side corner of Matlab's window to get the two data files into the program. By its means create three variables: `FemaleColheaders`, `MaleColheaders` (for the names of 10 columns of the numerical data), `Female`, `Male` (numerical data of size 18204x10), `FemaleTextdata`, `MaleTextdata` (for the short description of the data). Our interest is in `Female` and `Male` matrices, so type in `size(Female)`, `size(Male)` and `Female(2,6)`, `Male(2,6)` to see if you succeeded in reading data in.

Question 3: Write down specific values the above commands have returned.

We will not analyze these extensive data in their full complexity but we will focus on, let say, year 2000, and limit ourselves to look at the second and sixth columns of both the matrices. The following code should extract corresponding values into the matrix `norway`.

```
size(Female), size(Male)
Female(2,6), Male(2,6)

F=Female(Female(:,1)==2000,[2 6])
M=Male(Male(:,1)==2000,[2 6])
norway=[ M F(:,2)]
```

Please, check if the following values are residing in matrix `norway`.

²The data used here are accessible from the website Human Mortality Databas - an excellent source of various demographic data across the globe. If you are curious you may try to run a similar analysis for the data from Sweden.

Age $\frac{x}{\text{year}}$	Survivors at age x	
	Males	Females
–		
0	100 000	100 000
1	99 575	99 672
2	99 516	99 644
3	99 493	99 614
4	99 477	99 600
5	99 455	99 584
6	99 437	99 580
⋮	⋮	⋮
106	10	39
107	5	19
108	2	9
109	1	4
110	0	2

This table is to be interpreted in the following way. If 100 000 hypothetical, Norwegian men are alive at age 0 (when born in 2000), then, on average, 99 575 of them will reach the age of one (1) year, 99 516 will reach the age of two (2) years, . . . , and 10 will reach the age of 106 years etc.; the same goes for the figures for women in the third column. Now, if we divide the elements of columns 2 and 3 by $N = 100\,000$, we will obtain the survival function R for a newborn male (and female, respectively) Norwegian, valid for the year 2000. Assuming that this is a correct interpretation, let us plot the survival function:

```
>> N=100000;
>> t=norway(:,1);
>> R_male=norway(:,2)/N;
>> R_female=norway(:,3)/N;
>> plot(t,R_male,'b',t,R_female,'r'), grid on
```

Question 4: Are there any differences between the genders? Interpret them.

Now, obtain the death-rate function λ by numerical differentiation; let us use *central difference approximation*:

```
>> n=length(t)
>> lambda_male=zeros(size(t));
>> lambda_male(1)=-(R_male(2)-R_male(1))/R_male(1);
>> for i=2:(n-1), lambda_male(i)=-(R_male(i+1)-R_male(i-1))/(2*R_male(i)); end
>> lambda_male(n)=-(R_male(n)-R_male(n-1))/R_male(n);
```

Do the same for women's data. Plot both death-rates:

```
>> plot(t,lambda_male,'b',t,lambda_female,'r'), grid on
```

Question 5: Again, are there any differences between the genders? Interpret them.

In the field of life insurance one utilizes standardized death-rates. In Norway, dating back to 1963, the insurance companies used the following one, called the N-1963 standard.

$$\lambda_M(t) = \alpha + \beta e^{t/c}, \quad t > 0.$$

This was valid for males (M); for females (F) they have used

$$\lambda_F(t) = \alpha + \beta e^{(t-t_0)/c}, \quad t > 0.$$

In both cases, $\alpha \approx 9 \cdot 10^{-4} \text{ year}^{-1}$, $\beta \approx 4,4 \cdot 10^{-5} \text{ year}^{-1}$, $c \approx 10,34 \text{ year}$, and $t_0 \approx 3 \text{ year}$. In the same figure as above, plot these two death-rates:

```
>> alpha=9e-4; beta=4.4e-5; c=10.34; t0=3;
>> t=0:0.1:100;
>> lambdaM=alpha+beta*exp(t/c);           % Males
>> lambdaF=alpha+beta*exp((t-t0)/c);      % Females
>> hold on
>> plot(t,lambdaM,'k',t,lambdaF,'g')
>> hold off
```

Question 6: Compare the two sets of curves. Do you think that the standards from 1963 still applies to year 2000 data? Judging from the plot, would you say that the death-rates are IFR, DFR, or CFR?³.

The integral in Equation 1 is easy to compute:

$$\int_0^t \lambda_M(\tau) d\tau = \alpha t + c\beta(e^{t/c} - 1), \quad t > 0.$$

So, for males we have

$$R(t) = \exp(-(\alpha t + c\beta(e^{t/c} - 1))), \quad t > 0.$$

Analogously for females

$$R(t) = \exp(-(\alpha t + c\beta(e^{(t-t_0)/c} - e^{-t_0/c}))), \quad t > 0.$$

Now, use Matlab to answer the following questions according to the N-1963 standard:

Question 7: What is the probability that a certain male/female person will reach the age of at least 65? The probability asked for is $P(T > 65 \text{ year})$; in Matlab

```
>> PM=exp(-(alpha*65+c*beta*(exp(65/c)-1)))           % Males
>> PF=exp(-(alpha*65+c*beta*(exp((65-t0)/c)-exp(-t0/c)))) % Females
```

Question 8: A certain person is alive on the day he is 30. What is the conditional probability that the person will live to be 65? In Matlab:

```
>> PcondM=exp(-(alpha*(65-30)+c*beta*(exp(65/c)-exp(30/c)))) % Males
>> PcondF=exp(-(alpha*(65-30)+c*beta*(exp((65-t0)/c)-exp((30-t0)/c)))) % Females
```

Can you write mathematical expressions for what is evaluated here? Compare the probabilities obtained in Questions 7 and 8. Comments?

Question 9: What is the expected lifetime for males and females respectively? We use, of course, Equation 2. We will integrate numerically by means of the trapezium rule, implemented in Matlab as the routine `trapz`:

```
>> t=0:0.01:120;
>> RM=exp(-(alpha*t+c*beta*(exp(t/c)-1))); % Males
>> RF=exp(-(alpha*t+c*beta*(exp((t-t0)/c)-exp(-t0/c)))); % Females
>> figure, plot(t,RM,'b',t,RF,'r'), grid on
>> ETM=trapz(t,RM) % Males
>> ETF=trapz(t,RF) % Females
```

Also, write down the meaning of “ETM” and “ETF” (use mathematical symbols).

4 The Poisson process

We know that in a Poisson process the time distance between two consecutive events is exponentially distributed with parameter $1/\lambda$, where λ is the intensity of the Poisson process. In this section we will first use this property to simulate a Poisson process, and then use the simulations to estimate λ .

4.1 Simulation of a Poisson process

Let the intensity in the Poisson process be $\lambda = 0,5$ (for the sake of simplicity, let λ be physically dimensionless, i.e. choose the unit 1). We will simulate, say, 100 exponentially-distributed random numbers, that is, observations of T with distribution function

$$F_T(t) = 1 - e^{-\lambda t}, \quad t > 0$$

The commands are

```
>> N=100;
>> lambda=0.5;
>> timedistances=exprnd(1/lambda,1,N); % Alternative 1
>> tiemdistances=-log(1-rand(1,N))/lambda; % Alternative 2
```

Question 10: Can you explain what is the difference between the two alternatives (Hint: see Computer Exercise 2)?

We now have a vector `timedistances` with the time distances, but more interesting are the instants where events are occurring. By adding up all distances, we obtain the instants of the occurrences. The command `cumsum` in Matlab will be used. First try it on a small vector to see how this routine works:

```
>> v=[4 7 5 2];
>> cumsum(v)
```

Thus, for our vector with simulated time distances, we write analogously

```
>> instants=cumsum(timedistances);
```

to get the instants. The value of the Poisson process at time t is equal to the number of events up to time t . To draw a plot of the Poisson process, we will use `stairs` in this way:

```
>> stairs(instants,1:N)
>> grid on
```

Now, choose a fixed instant t_1 , e.g. $t_1 = 150$ (also physically dimensionless). From the plot, you immediately find the number of events which occurred up to time t_1 . Use the `zoom` command if it is hard to read. To get a feeling for the randomness, we can for instance simulate values from a Poisson distribution, since the number of events, N_A occurring in the interval $A = [0; t_1]$ is Poisson distributed with expectation λt_1 , i.e. $N_A \in \text{Po}(\lambda t_1)$. Simulate, say, 10 such Poisson variables:

```
>> t1=150;
>> poissrnd(t1*lambda,1,10)
```

Question 11: How much do they seem to vary? What is the theoretical value of the variance $V(N_A)$ and thus the standard deviation? Does the standard deviation seem to be in agreement with the variation of the simulated data?

4.2 Estimation of intensity

The intensity λ of a Poisson process is the expectation for the number of events Y occurring in the time interval, e.g., $[0; t_1]$, *divided by* the length of the same interval, i.e. t_1 . Thus, the parameter λ can be estimated by dividing the number of events in a certain time interval by the length of the interval.

Question 12: Estimate this value using

```
>> lambdahat=sum(instants<=t1)/t1
```

and compare with the true value of λ . Are they similar?

5 Accidents in coal mines in the United Kingdom

The file `coal.dat` contains information about accidents in coal mines in the United Kingdom from 1851 to 1918. Load the data by

```
>> coal=load('coal.dat')
```

Type `size(coal)` and you will find that the data are stored as a 153×6 matrix. A description of the six columns:

Column	Content
1	Day of month (DD)
2	Month (MM)
3	Year (CCYY)
4	Ordinal day of year (DDD), i.e. number of days passed of the year
5	Number of days since previous accident
6	Number of perished

Plot the process by

```
>> N=153;
>> t=cumsum(coal(:,5));
>> stairs(t,1:N)
>> grid on
```

Question 13: Does it look like a sample from a Poisson process? Comment on the shape comparing to what you have seen in the simulated data.

If this is the case of a Poisson process, we are interested in estimating the intensity. However, the intensity does not seem to be constant during the whole period. Divide the data in two periods and estimate two intensities λ_1 and λ_2 ; around instant $1,5 \cdot 10^4$ days — i.e. after 127 accidents — there seems to be a change in intensities from λ_1 to λ_2 . The estimates of λ_1 and λ_2 will now be

```
>> lambda1hat=127/1.5e4 % unit: 1 day(-1)
>> lambda2hat=(N-127)/(t(end)-1.5e4) % unit: 1 day(-1)
```

Plot the process with years on the abscissa (x-axis):

```
>> stairs(coal(:,3)+coal(:,4)/365.25,1:N), grid on
```

Question 14: When does the change occur? There was a legislation effort to improve the safety in the UK coal mines enacted in 1880's. Can the data provide argument that it succeeded?