University of Zurich

Epidemiology

Data Analysis Course 2016 - Project n° 3

$\mathcal{H}$annes Imboden | $\mathcal{L}$inus Leo Stöckli | $\mathcal{M}$iloš Ivanović
$\mathcal{N}$icola Gadola | $\mathcal{P}$hilipp A. Huber

Contents

Introduction

Epidemiology is the study of the cause, effect and the spread of health concerning events within a population.[1]As this usually involves a lot of people, statistics is a key instrument in processing any information: Mathematical models can be used to describe various processes, e.g. the contagiousness of a certain disease. Furthermore they allow to make predictions about the future, i.e. estimate a possible outcome of a setup, when given the initial conditions.[2]
The goal of this project is to apply statistical methods to the following initial situation and therewith resolve the later specified tasks:

"A dangerous and highly contagious viral disease, similar to the smallpox, has been observed in a few towns. Once an individual has been infected, the first symptoms appear very quickly, and the health of the individual degenerates rapidly, leading to his or her death"

Two datasets were provided to us - one containing the number of hours each (out of 88) individual survived after experiencing the first symptoms, the other one containing information about the contagiousness of the disease, i.e. the percentage of healthy people that incurred the disease versus the number of already infected individuals they have been in contact with for one day.
The tasks provided by the assistants are as follows:

1. Extract from the provided data the mean expected survival time of an infected individual. You can use an unbinned Maximum Likelihood fit. Compute the uncertainty on this result.

2. Extract from the provided data the mean expected survival time of an infected individual. You can use an unbinned Maximum Likelihood fit. Compute the uncertainty on this result.

3. Now consider a small village of 100 inhabitants. Keep your model simple: assume that each individual is in daily contact with all the others. Assume that initially one single inhabitant gets infected. Develop a simulation showing the evolution in time of:

    - the number of still healthy inhabitants
    - the number of infected inhabitants
    - the number of deceased inhabitants

[1]https://de.wikipedia.org/wiki/Epidemiologie
[2]Epidemiology - Instruction sheet, Elena Graverini and Andreas Weiden

4. Use your simulation to answer important questions such as:
    - after how many days does the epidemic stop (i.e. after how many days do people stop getting infected)?
    - how many people will survive the epidemic?
    - how much time do authorities have in order to put protection measures in place (e.g. quarantine, vaccination campaigns etc.) before 30% of the population is either infected or dead?
5. Estimate the uncertainties on your predictions, for example by varying the values of the parameters that you extracted from the given data within their uncertainties. Use your simulation to determine how the previous result changes when accounting for these uncertainties.

Evaluation

Task 1

The first task was to determine the mean expected survival time using an unbinned Maximum Likelihood fit, given the dataset described earlier.
To get an idea of the underlying probability distribution, we put our data into a histogram and obtained the following figure:

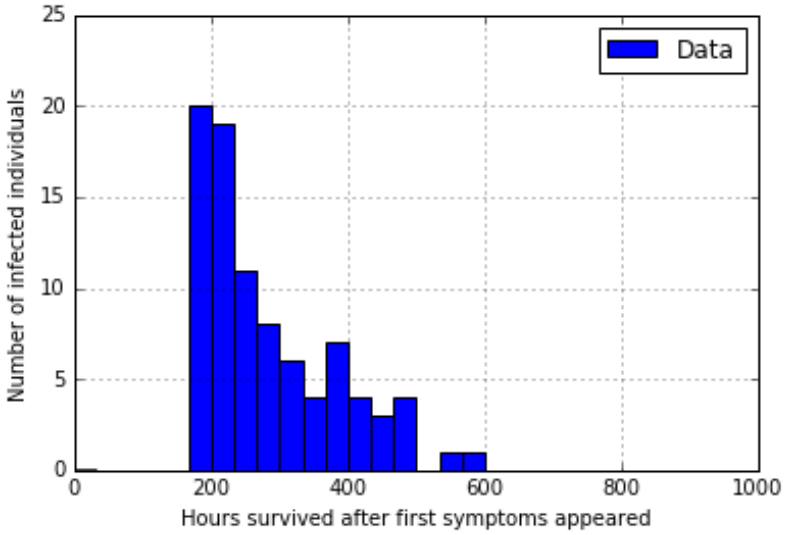

Figure 1: First Dataset

From there we concluded, that the probability distribution follows an exponential distribution, which is given by

$$f(t|\tau) = \frac{1}{\tau} \cdot e^{-\frac{t}{\tau}} \tag{1}$$

where τ is our wanted value for the expected survival time. Furthermore, we can observe a right shift, which we will be discussing later on.
The integral of the probability density over the interval has to be equal to one (100%). As our interval was bounded by the values 169 and 596 hours, $f(t|\tau)$ had to be normalized as follows:

$$\int_{t_{min}}^{t_{max}} f(t|\tau)\mathrm{d}t = \int_{169}^{596} \frac{1}{\tau} \cdot e^{-\frac{t}{\tau}}\mathrm{d}t = -e^{-\frac{t}{\tau}}\Big|_{169}^{596} = \underbrace{-e^{-\frac{596}{\tau}} + e^{-\frac{196}{\tau}}}_{\text{normalization constant}} \stackrel{!}{=} 1 \tag{2}$$

$$\Rightarrow F(t|\tau) = \frac{1}{-e^{-\frac{596}{\tau}} + e^{-\frac{196}{\tau}}} \cdot \frac{1}{\tau} \cdot e^{-\frac{t}{\tau}}\mathrm{d}t \tag{3}$$

We then proceeded to do the Maximum Likelihood fit for an exponential distribution. The log-likelihood function $\ln L(\tau)$ is given by:

$$\ln L(\tau) = \sum_{i=1}^{N} \ln F(t_i|\tau) \tag{4}$$

By then deriving the log-likelihood function and setting it equal to zero, we could determine the maximum value of the function, that is $\hat{\tau}$. Alternatively, we could plot the log-likelihood function for a set of possible values for τ and therewith obtain our $\hat{\tau}$. Using *Python*, the latter can be done by running the following:

```
def maxvalue(l, i):
        key = itemgetter(1)
    return max(enumerate(sub[i] for sub in l), key = key)


def maximum_likelihood(data, dist, tau_array):
    LL = []
    indexlist = []
    for tau in range(len(tau_array)):
        lnL = 0.
        for i in range(len(data)):
            lnL += np.log(dist(data[i], tau_array[tau],data))
        indexlist.append([tau, lnL])
        LL.append(lnL)
    maxval = list(maxvalue(indexlist, 1))
    tau_hat = tau_array[maxval[0]]
    fig = plt.figure(dpi = 500)
    plt.plot(tau_array, LL, label = "log-likelihood")
    return LL, tau_hat, fig
```

This produces the following figure:

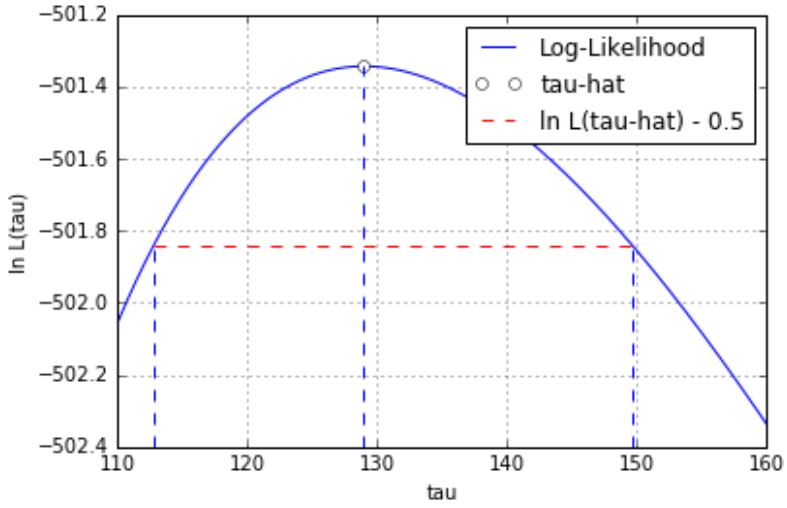

Figure 2: Log-Likelihood Function

Clearly visible in the graph, the parabola is not symmetrical. This has to be considered when determining the uncertainty on $\hat{\tau}$. As an anti-symmetrical curve indicates a relatively small sample size $N \not\to \infty$, a special ansatz can be used to determine the uncertainty on $\hat{\tau}$, namely:[3]

$$\ln L(\hat{\tau} \pm n \cdot \sigma_{\hat{\tau}}) \equiv \ln L(\hat{\tau}) - \frac{n^2}{2} \tag{5}$$

This gives us the following errors on $\hat{\tau}$:

$$\hat{\tau} = (128.98 + 20.75/ - 16.24)\ \text{h}$$

As of now, we did not include the right shift of our dataset mentioned earlier. So we proceeded by adding the lower limit $t_{min} = 169$ h to our $\hat{\tau}$, giving us:

$$\hat{\tau}_{shift} = \hat{\tau} + 169\ \text{h} = (128.98 + 169)\ \text{h} = 297.98\ \text{h}$$

$$\Rightarrow \hat{\tau}_{shift} = (297.98 + 20.75/ - 16.24)\ \text{h}$$

[3]This was discussed in the "Datenanalyse" lecture of O. Steinkamp

Including all this in our initial histogram, we end up with the figure below:

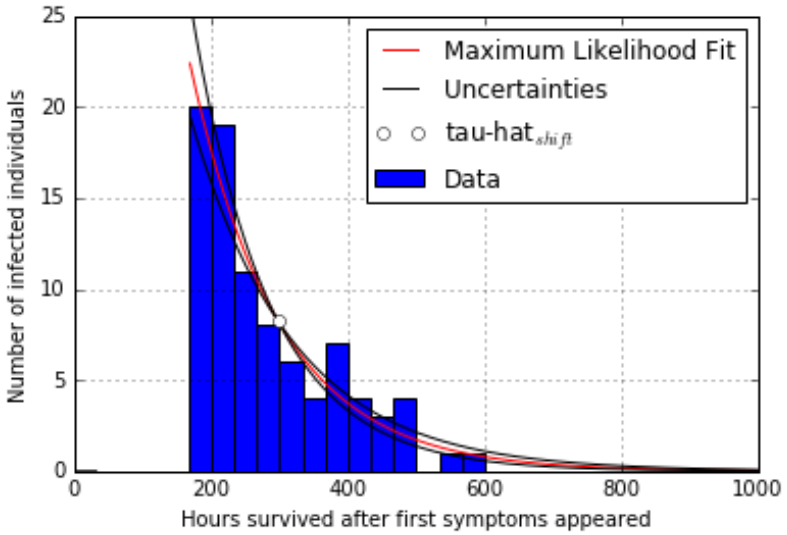

Figure 3: Histogram and Exponential Distributions

So that the exponential curves intersect at $\hat{\tau}_{shift}$, correctly corresponding to the histogram, the only thing left to do was to normalize them as follows:

```
y = 8.23*np.e**(-(x-297.98)/128.98)
sigma_1 = 8.23*np.e**(-(x-297.98)/149.73)
sigma_2 = 8.23*np.e**(-(x-297.98)/112.74)
```

Task 2

The second step in the evaluation of the data was to extract the probability of getting infected, as a function of the number of already infected people around. To get an idea of the contagiousness, we used old archives. In these documents, the percentage of people getting infected daily was listed. In addition to that, we know with how many already infected people they had been in contact with.
This allowed us to estimate the contagiousness of the disease.

The data was then plotted to determine the type of function these values follow. We first tried this with a histogram. This however did not provide clear indications, so we used a scatterplot instead. On the graph, the x-Axis indicates the number of sick people around and the y-Axis shows the probability to get infected:

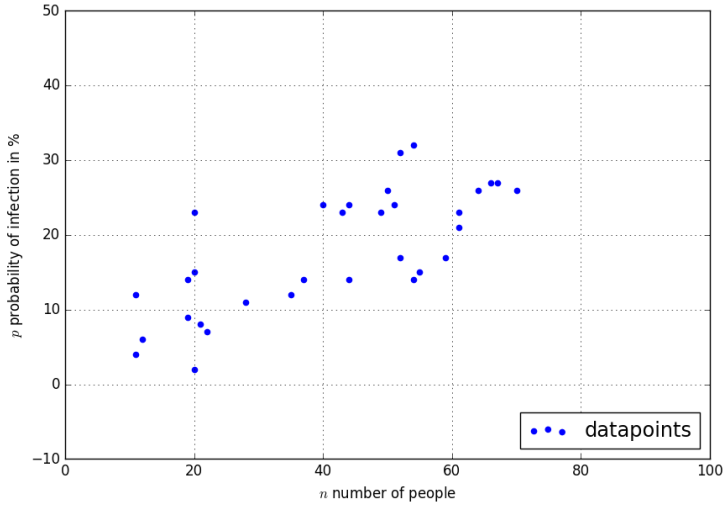

Figure 4: Scatterplot for the data

The plot indicates a linear growth, so we determined the parameters of this straight line $(a \cdot x + b)$ via the maximum likelihood method. From the lecture, these are found as follows:

$$\hat{a} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \overline{x}^2} \qquad \hat{b} = \frac{\overline{x^2}\overline{y} - \bar{x}\overline{xy}}{\overline{x^2} - \overline{x}^2} \tag{6}$$

The bar over the variables indicates that we need the average value. The percentage of healthy people having caught the disease corresponds to our variable x and the number of infected people he/she had been in contact with is y. The two parameters can then be computed quite easily in *Python* :

```
data=read_from_file("")
x=data[:,1]
y=data[:,0]
ahat_0 = (mean(x**2)*mean(y)-mean(x)*mean(x*y))
         /float(mean(x**2)-mean(x)*mean(x))
ahat_1 = (mean(x*y)-mean(x)*mean(y))
         /float(mean(x**2)-mean(x)*mean(x))
print "ahat_0:", ahat_0
print "ahat_1:", ahat_1
funcx=np.linspace(0,100,1000)
funcy= ahat_0 + ahat_1*funcx
```

In the example code, "ahat_0" corresponds to our "b", the intersection with the y-Axis, and "ahat_1" is "a", the slope.

Also, the covariance matrix for the errors on these two parameters was derived in class:

$$\mathrm{cov}(\hat{b},\hat{a}) = \frac{\sigma^2}{N\cdot(\overline{x^2}-\overline{x}^2)}\cdot\begin{pmatrix}\overline{x^2} & -\overline{x}\\ -\overline{x} & 1\end{pmatrix} \tag{7}$$

Note that the variance, σ^2, is equal to $(\overline{x^2}-\overline{x}^2)$, so these cancel out.
In *Python*, this looks like follows:

```
m=np.array([[mean(x**2), -mean(x)],[-mean(x), 1]])
cov=1/float(len(x))*m

print "uncertainty on ahat_0 : ", cov[0][0]**0.5
print "uncertainty on ahat_1 : ", cov[1][1]**0.5
```

Therefore, the error on $\hat{a}$ is $\sqrt{\overline{x^2}}/N$
The one on $\hat{b}$ is $\sqrt{1}/N$
This gives us the following values for $\hat{a}$ and $\hat{b}$:

$$\hat{a} = (0.325 \pm 0.177) \cdot \frac{1}{100} \cdot \text{people}^{-1}$$

$$\hat{b} = (4.51 \pm 7.93) \cdot \frac{1}{100}$$

Thus, the function linking the probability to get infected with the number of sick people he/she had been in contact is:

$$\mathrm{f}(x) = 0.325x + 4.51$$

This means that the probability of an individual to get infected, having been in contact with only one infected person is 4.84%.

The following plot shows the best fit found with the maximum likelihood method as well as the steepest and flattest slopes possible within the errors:

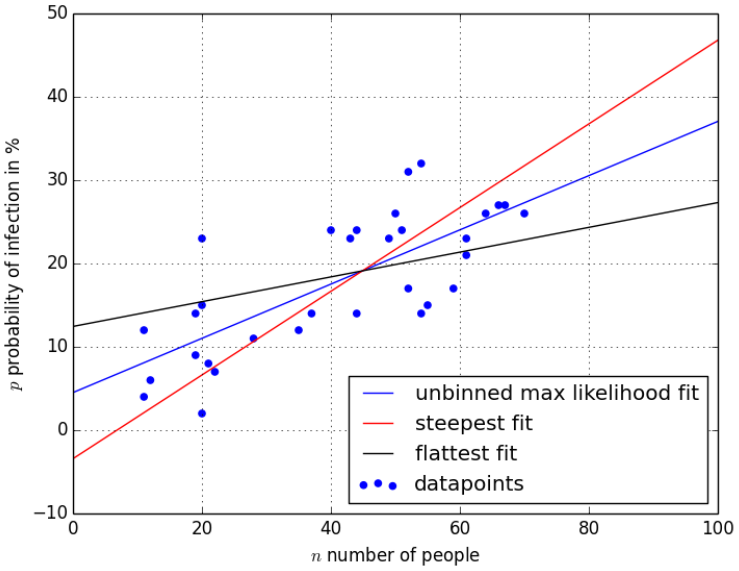

Figure 5: Best fit and errors on the slope

Task 3

In the third step we simulated the spread of the disease in a small village. Assuming there are 100 inhabitants and they are all in contact with each other daily, we can plot the evolution of healthy, infected and deceased individuals.

First, from the function calculated in the second task, we can determine how the disease spreads in the village. We did have to round the amount of infected people, because a person can either be infected or helathy, nothing in between, thus no more people get infected once 99 are infected, because the probabilty of infection, when 99 are infected is (as one can read from task 2) 35% and because this is less than 50%, the last person does never get infected.
Not accounting for the deaths, this looks as shown below:

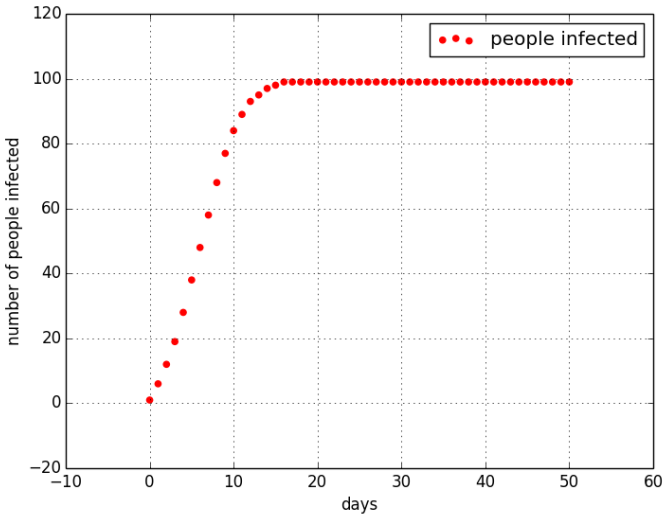

Figure 6: Spread of the disease in the village

Note that initially the spread is exponential but then rapidly decreases. This decrease is due to the fact that the sample (inhabitants of the village) is finite and thus has to stop sometime.

Second, the findings of the first and second tasks, namely the contagiousness and deadliness of the disease, were combined to plot the complete evolution in the village. The following graph illustrates this progression:

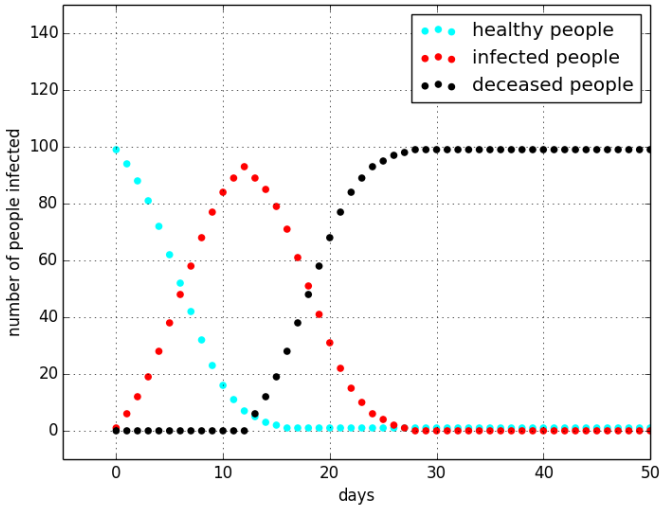

Figure 7: Evolution of the disease in the village

As a last step, all these findings were combined in an animation to better illustrate how the disease evolves.
Initially all inhabitants except one are healthy:

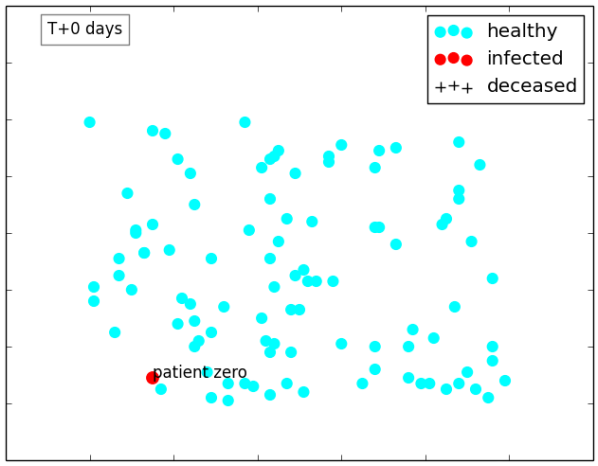

Figure 8: The village on the day the epidemic starts

After on average 12 days, the first people start to die. At day 15, most of the population is infected with some casualties:

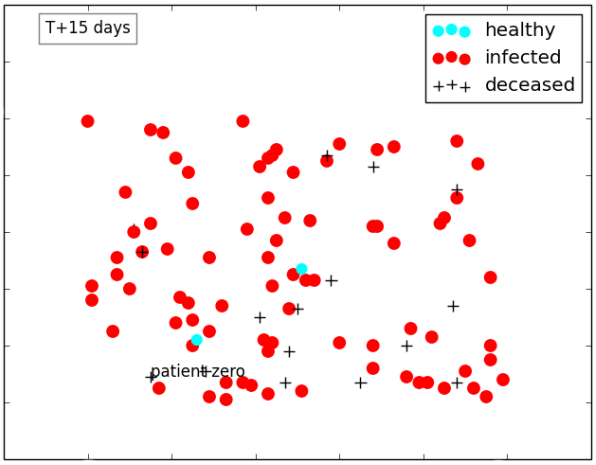

Figure 9: The village on the 15th day

After 28 days, all infected have died and only one individual survives:

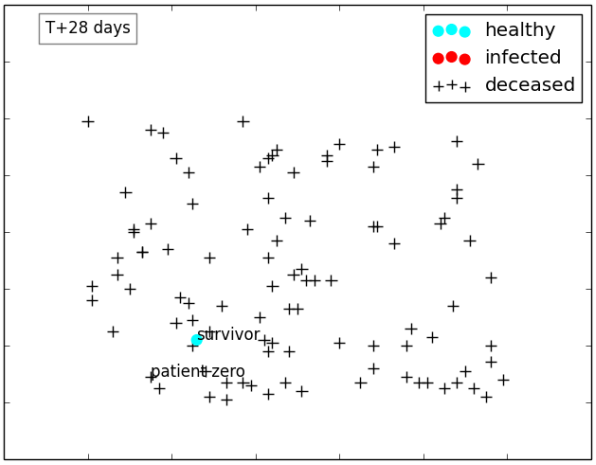

Figure 10: The village on the 28th day

The full animation can be found at: www.physik.uzh.ch/~linsto/DA

Task 4 & 5

Now that the data has been analyzed and plotted, we can answer the questions from the instructions.

- After how many days does the epidemic stop ?

If we don't consider the disease deadly, the epidemic "stops" after 16 days. This means that after 16 days, everyone in the village is infected.
However, it is important to note that we are dealing with probabilities to get infected. In our calculation, we have to round the number of infected people, as noted before in task 3. If we would not round our numbers, the number of infections would converge to 100, and after 18 days 99.5 out of 100 people are infected, we can assume that everyone is infected.
If we now take into account that the disease is deadly, this looks slightly different. After 12 days on average, the first infected start to die, and since we assume that the virus living inside the hosts body dies with the host, less people catch the sickness. This results in a "stop" of the epidemic after 16 days. If we calculate with the uncertainty of our infection probability, we get that the epidemic stops after 16 ± 1 days.

- How many people will survive the epidemic ?

With the data provided, new people get infected until day 16. As an infected person dies after an average of 12 days, the last one dies 28 days after the outbreak. In our simplified setup, only one person survives the epidemic.
As previously shown however, both the probability function and the average survival time have uncertainties. Taking the smallest probability to get infected and the shortest average survival time, we can estimate the maximum number of survivors. In that case, 2 people survive.
On the other hand, if we consider the highest probability function, no one other than the first one gets infected. Varying the expected survival time does not change that. This is due to the fact that even though the probability to get infected with more sick people around is higher than before, the initial patient doesn't infect anyone. The probability with one infected in the village is less than 0.5% and is thus not registered as an infection by the program.

- How much time do authorities have to put out protective measures ?

In the most likely case, 30% of the population is infected after 5 days only. If the epidemic wants to be stopped or contained (with quarantines or vaccinations), authorities have to act very quickly.
In the case where we take the flattest probability function and the shortest average survival time, they have even less time. 30% of the inhabitants are infected after only 3 days. This is caused by the initially higher chance to get infected having been in contact with only a few sick people.

Discussion

The model from task 3 is of course heavily simplified and thus only allows us to approximate predictions. For instance, we didn't factor in any possible resistances to the virus and assumed that it dies with the host.
Also, the fact that everyone is in daily contact with everyone else is unlikely.

Another problem is that the program has to round the percentages to plot them. This can lead to errors in the plots if the percentage is never quite high enough to infect someone, as it was the case in our animation. This issue has been discussed in the first question of task 4.

In conclusion, it has to be noted that a real evolution would be much more complex. However, with the data provided and the assumptions made, we can get a good idea of what the evolution of a disease would look like.

Appendix

All the python scripts, animations, pictures, etc. can be downloaded at:
www.physik.uzh.ch/∼linsto/DA