Project 1 : Ida Sandum, Mats Seglem and Gard Strøm

Introduction

In this report we will first model an outbreak of measles in problem 1, and then we will look at insurance claims modelled as a Poisson process in problem 2. The report will consist of discussions and figures from some computer code examples, whereas the code itself will be given as an additional file.

Problem 1

a)

A Markov chain is a stochastic process where the probability of each event only depends on the state attained in the previous event [1]. This is called the Markov property and is fulfilled here. For example, if you were susceptible three days ago and are infected now, then tomorrow you can either still be infected or recovered, regardless of your state three days ago.

The transition probability matrix is given as $P_{ij}^{n,n+1} = Pr\{X_{n+1} = j | X_n = i\}$, i.e the probability that X_{n+1} is in the state j given that X_n is in state i. In this case the Markov chain has stationary transition probabilities, and we therefore use $P_{ij}^{n,n+1} = P_{ij}$.

In order to find the transition probability matrix we may calculate these probabilities using the assumptions in the problem text. Since the probability of becoming infected if you are susceptible is β and the total probability needs to be 1, the probability of remaining susceptible is $1 - \beta$. Starting in the susceptible state we get $P_{00} = Pr\{X_1 = 0 | X_0 = 0\} = 1 - Pr\{I|S\} = 1 - \beta$ and $P_{01} = Pr\{X_1 = 1 | X_0 = 0\} = \beta$. Recovering from a disease you do not have is not possible, so $P_{02} = 0$. Using the same reasoning for the other states we get $P_{10} = 0, P_{11} = 1 - \gamma, P_{12} = \gamma, P_{20} = \alpha, P_{21} = 0, P_{22} = 0$. This results in the matrix $\mathbf{P}$ in the problem text.

b)

We calculate the square of the transition probability matrix

$$\mathbf{P}^2 = \begin{pmatrix} (1-\beta)^2 & \beta(1-\beta) + \beta(1-\gamma) & \beta\gamma \\ \gamma\alpha & (1-\gamma)^2 & \gamma(1-\gamma) + \gamma(1-\alpha) \\ \alpha(1-\beta) + \alpha(1-\alpha) & \alpha\beta & (1-\alpha)^2 \end{pmatrix}.$$

As $0 < \alpha, \beta, \gamma < 1$, the transition probability matrix is regular. In this problem we have a finite state space, so according to theorem 4.1[1] the limiting distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2)$ exists. Inserting the given values for α, β and γ gives the transition probability matrix

$$\mathbf{P} = \begin{pmatrix} 0.99 & 0.01 & 0 \\ 0 & 0.90 & 0.10 \\ 0.005 & 0 & 0.995 \end{pmatrix}.$$

Using $\mathbf{P}$ and theorem 4.1 we get equations for the limiting probabilities,

$$\begin{aligned} \pi_0 &= 0.99\pi_0 + 0.005\pi_2 \\ \pi_1 &= 0.01\pi_0 + 0.90\pi_1 \\ \pi_0 + \pi_1 + \pi_2 &= 1 \end{aligned}$$

The first of these equations yields $2\pi_0 = \pi_2$ and the second yields $\pi_0 = 10\pi_1$. Using these in the last equation gives $\pi_0 = 10/31$. So the limiting distribution is

$$(\pi_0, \pi_1, \pi_2) = \left(\frac{10}{31}, \frac{1}{31}, \frac{20}{31}\right)$$

In the long run, out of 365 days, an individual spends a mean number of 117.74 days in the susceptible state, 11.77 days in the infected state and 235.48 days in the recovered state.

c)

Estimates of each of the long run means from the simulation are 119.56 days in the susceptible state, 11.87 in the infected state and 233.57 days in the recovered state. This seems reasonable compared to 1b).

The central limit theorem states that if you take sufficiently large random samples from the population the distribution will tend towards a normal distribution even if the variables themselves are not normally distributed [2]. In this problem we look at a mean over 10 years of daily observations, and it is therefore natural to use this theorem. Moreover, since we can assume normally distributed variables X_i's and the standard deviation is unknown [3], we may use the Student's t-distribution to compute the confidence interval. We do this by calculating the standard deviation S and the mean $\bar{X}$ of the sample of size $n = 30$ in R. Then $T = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ is a t-distributed random variable [3]. We can calculate the 95% confidence interval (CI) by using $Pr\{Z < t_{0.025,n-1}\} = 0.975$. Then we get the upper and lower bound for μ as $\bar{X} - Z_{0.025}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{0.025}\frac{S}{\sqrt{n}}$, where $Z_{0.025}$ is 2.045[4].

Computed CIs for the susceptible, infected and recovered states are [108.928, 130.200], [10.302, 13.438] [222.280, 244.852] respectively. Our calculated values in 1b) are compatible with these intervals as they are contained in them.

d)

$\{I_n : n = 0, 1, 2...\}$ is the number of infected individuals at time step n. The probability of becoming infected if you are susceptible is $\beta = 0.5I_n/N$. Since you do not know how many people are susceptible, you cannot know how many will become infected. Therefore, the stochastic process does not satisfy the Markov property and I_n is not a Markov chain.

$\{Z_n : n = 0, 1, 2...\}$ is the number of infected individuals and the number of susceptible individuals at time step n. We also know the total number of people, N, so we can calculate the number of recovered individuals, R_n. Therefore we have what we need to compute the transition probabilities from state n to state $n + 1$. Hence Z_n satisfies the Markov property and is a Markov chain.

For the same reasons $\{Y_n : n = 0, 1, 2...\}$ satisfies the Markov property and is also a Markov chain.

e)

Since we are dealing with a regular transition probability matrix, we know that there exists a limiting probability distribution. This means that the transition probabilities converges as $n \to \infty$, and the probability of finding the Markov chain in some state j is approximately π_j regardless of the starting point. We therefore expected the behaviour we see in figure 1.

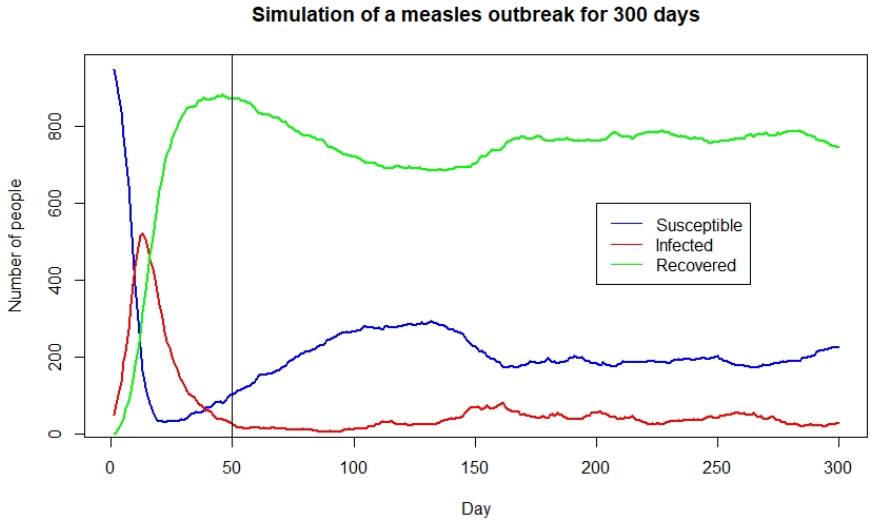

Figure 1: Evolution of susceptible, infected and recovered individuals in a 300 day simulation of a measles outbreak.

In the start there were mostly susceptible individuals and some infected. This means a lot of people were prone to disease. As more and more people got infected, the chance of becoming infected increased as β is dependent on I_n. Meanwhile, the individuals who recovered became immune and the probability $\alpha = 0.005$ of becoming susceptible again was very small. This results in an explosive outbreak of disease early on, but increasing immunity in the population makes sure less and less people get infected, and stabilization occurs.

f)

Our estimate for the expected maximum number of infected individuals during the simulated time steps is 523.442, and the estimate for expected time at which the number of infected individuals first takes its highest value is after 12.862 days. In reality, a non-integer value of these parameters does not make sense, so these are just estimations.

Computed confidence interval for the expected value of the time of most infected individuals is $[12.808, 12.916]$ and for the expected value of the number of most infected people it is $[522.129, 524.755]$. This can be used to asses potential severity of the outbreak because it gives an indication of how many people may be infected simultaneously and also how fast this happens. It is bad for society if many people are ill at once, because there are less people to work and a lot of important societal functions may collapse, or at least struggle to maintain a sufficient flow. An example is health care. Hospitals become over-flood with people and health workers get overworked. If the disease spreads quickly as well, then there is less time to prepare, which makes the situation worse. In this case over half of the population was infected at the same time and it happened after 12-13 days, which is quite severe.

g)

To simulate the outbreak with vaccinated individuals, we used the same simulation as in problem 1e). For the different cases, we started with a different Y_0. Let V be the number of vaccinated people. Then $Y_0 = (950 - V, 50, V)$. Also, when we calculated the number of new susceptible individuals each day, the vaccinated people could not lose

immunity. Therefore, we had to subtract V from the number of trials in that binomial distribution.

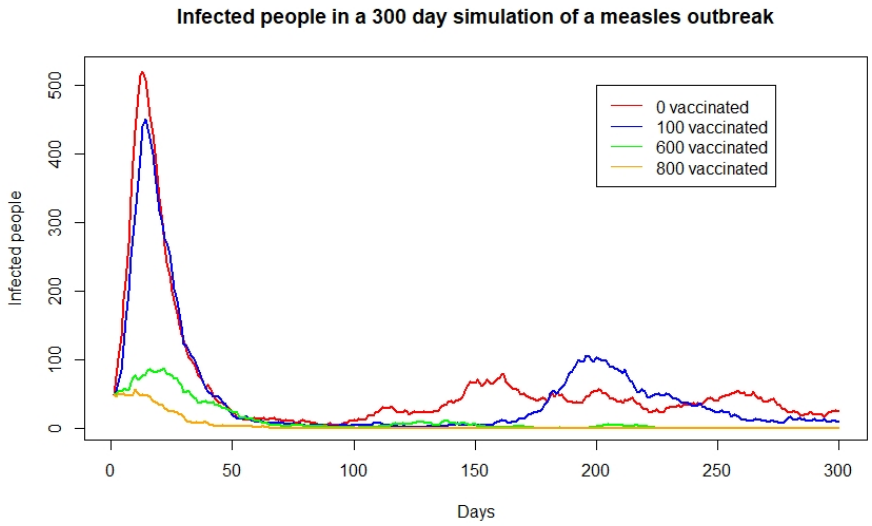

Figure 2: Evolution of infected people in a 300 days simulation of a measles outbreak with different amount of vaccinated people at day 1. Total number of people in the simulation is 1000, and it is always 50 infected at day 1.

As more and more individuals are vaccinated we can see from figure 2 that the peak gets smaller and smaller, and also happens slightly later, except from in the case with 800 vaccinated people.

As we can see from table 1 the expected maximum value of the number of infected individuals is highest for the sample with no vaccinated people, and it decreases for each higher number of vaccinations. This seems reasonable, as the vaccines makes sure less people get infected. We also see that the number of days before the maximum increases for each sample until the sample with 800 vaccinated individuals. It makes sense that it takes more time to reach the peak when there are fewer people who are susceptible. However, we see that when 80% of the population are vaccinated, the peak appears very quickly. This is due to the fact that there were 50 infected individuals at the beginning, and it only took a little over 2 people to get infected to reach the peak. This seems logical, as when so many are vaccinated there are few left to get infected. And the people who do get infected also become immune for some time after. Say we let 950 individuals be vaccinated and 50 be infected. Then the peak would appear at once, as no one were susceptible.

Table 1: Expected maximum values

Vaccinated individuals	$E\left[max\{I_1, ..., I_{300}\}\right]$	$E\left[\arg\max_{n\leq 300}\{I_n\}\right]$
0	523.442	12.862
100	439.463	13.692
600	96.992	16.904
800	51.239	2.042

Problem 2

a)

To compute the probability that there are more than 100 claims after 59 days it is easier to calculate the conjugate and use the law of total probability. Thus, we rather compute a sum from $x = 0$ to $x = 100$ than a sum from $x = 101$ to $x = \infty$.

$$Pr\{X(59) > 100\} = 1 - Pr\{X(59) \leq 100\}$$

$$Pr\{X(59) > 100\} = 1 - \sum_{x=0}^{100} \frac{88.5^x}{x!} e^{-88.5} \approx 0.1028$$

The probability that more than 100 claims have arrived during the first 59 days is approximately 0.103

As the code is pseudo random it will most of the time not get the exact same value as the theoretically calculated one. We use a seed to get the same value for each run, and with this seed the code gives an estimated value of 0.111 for 1000 runs of the simulation. If we run the simulation a million times we get an estimated probability of 0.103, which corresponds to the calculated value. It seems from figure 3 that about 1/10 of the graphs reach 100 claims in 59 days. This coincides with the calculated value and suggest that the code is a good approximation of the problem.

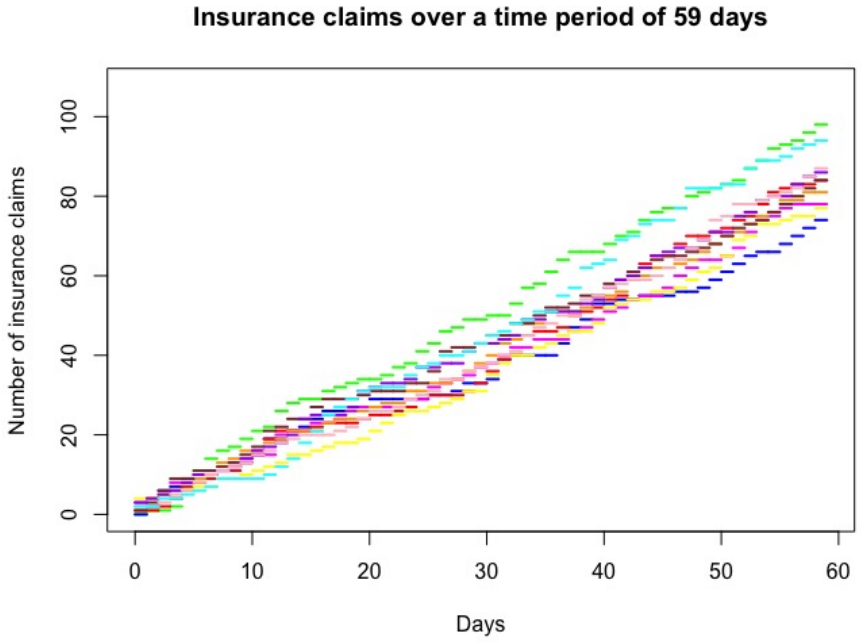

Figure 3: Ten simulations of insurance claims over the period $0 \leq t \leq 59$. This figure is generated with a specified seed.

b)

The code calculates the estimated total value of all claims in a time period $t \in [0, 59]$. When simulating the code 1000 times we get an estimated probability of 0.739 that the total amount of claims exceeds 8 mill. kr.

The graphs do not coincide exactly with the estimated probability. In figure 4, the total value of insurance claims exceeds 8 mill. kr. in 9/10 of the simulations. The result

is still reasonable because the number of simulations is small. Therefore it is not unlikely that the percentage of graphs that exceeds 8 mill. kr. are roughly 90%. On the other hand, if we had 1000 simulations, we would expect 73.9% of the graphs to exceed 8 mill. kr..

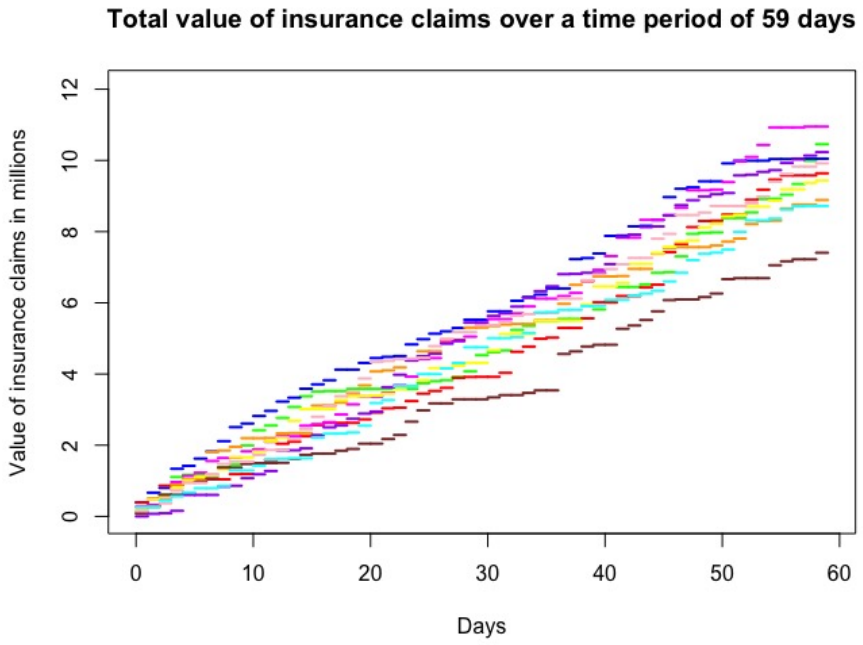

Figure 4: Ten simulations of total value of insurance claims in mill. kr. in the time period $0 \leq t \leq 59$.

c)

According to theorem 5.2 [1], we only need to show that $Y_t|X_t \sim Binomial(X_t, p)$ since it is given that $X_t \sim Poisson(\lambda)$. $Y_t|X_t$ is binomial distributed since X_t is the number of trials that can end in either success or failure, where success is that a claim exceeds 250000kr.. Furthermore, all trials are independent and have the same probability of success. Finding this probability of success p is straight forward, since the value of each claim is independent and exponentially distributed. Therefore $p = P(X > 0.25)$.

$$\begin{aligned} P(X > 0.25) &= 1 - P(X \leq 0.25) \\ P(X > 0.25) &= 1 - \int_0^{0.25} \lambda e^{-\lambda x} dx \\ P(X > 0.25) &\approx 0.0821 \end{aligned}$$

This corresponds to roughly 8.21% of all claims being above 250000 and need to be investigated.

We know that $Y_t|X_t$ is a binomial distribution with probability of success $p = 0.0812$. Therefore, by theorem 5.2, Y_t is Poisson distributed[1] with rate $\lambda p \approx 0.123$.

References

[1] Pinsky, M.A. and Karlin, S., 2011, An Introduction to stochastic modeling, Academic Press/Elsevier. (4th Ed)

[2] *Central limit theorem* (2021, September 21). In Wikipedia. https://en.wikipedia.org/wiki/Central_limit_theorem

[3] *Student's t-distribution.* (2021, August 31). In Wikipedia. https://en.wikipedia.org/wiki/%27s_t-distribution

[4] Institutt for matematiske fag, 2000. Tabeller og formler i statistikk. Fagbokforlaget (10 oppslag)