DIPLOMARBEIT

Modellrisiko bei der Schätzung von epidemiologischen Parametern bei Covid-19

ausgeführt am

Institut für Statistik und Wirtschaftsmathematik
Technische Universität Wien

unter der Anleitung von

Univ.Prof. Dipl.-Math. Dr.rer.nat. Uwe Schmock

eingereicht durch

Daniel Wimmer, BSc

Matrikelnummer: 01612971

Studienkennzahl: E 066 405

Wien, am 23. Oktober 2021

Autor | Betreuer

DIPLOMA THESIS

Model Risk of the Estimation of Epidemiological Parameters of Covid-19

written at the

Institute of Statistics and Mathematical Methods in Economics
Vienna University of Technology

supervised by

Univ.Prof. Dipl.-Math. Dr.rer.nat. Uwe Schmock

submitted by

Daniel Wimmer, BSc

Matriculation number: 01612971

Study Code: E 066 405

Vienna, October 23, 2021

Author

Supervisor

Zusammenfassung

Das Ziel dieser Arbeit ist es, Modelle zu hinterfragen, die die Dynamik der Covid-19 Pandemie beschreiben. Diese Modelle sind höchst relevant, da politische und ökonomische Einschränkungen auf ihren Ergebnissen basieren. Wir wollen herausarbeiten, wo und wie bestimmte Annahmen einen signifikanten Einfluss auf das Resultat dieser Modelle haben. Genauso geben wir Verbesserungsvorschläge, wo es notwendig ist. In der gesamten Arbeit liegt ein spezieller Fokus auf den mathematischen Hintergründen der Schätzmethoden, der Infektionsdynamik und versicherungsmathematischen Themen.

Nach einer kurzen Einleitung in Kapitel 1, beginnen wir mit der Schätzung relevanter Zeitspannen wie dem seriellen Intervall und der Inkubationszeit in Kapitel 2. Wir untersuchen mehrere Verteilungsfamilien und Schätzmethoden und vergleichen diese mit anderen wissenschaftlichen Papers und der offiziellen Schätzung durch österreichische Behörden. Wir legen einen starken Fokus darauf, welche Daten beobachtet werden können und welche Probleme dabei entstehen. Schlussendlich diskutieren wir Schwächen und Verzerrungen der Resultate.

Um den Reproduktionsfaktor in Kapitel 3 zu schätzen, brauchen wir die Verteilung vom seriellen Intervall aus dem vorherigen Kapitel. Wir diskutieren die exakte Definition eines zeitabhängigen Reproduktionsfaktors und wie er beobachtet werden kann. Dann analysieren wir das Modell der österreichischen Behörden und die zugrunde liegenden Annahmen. Wir untersuchen, welche Schwächen dadurch entstehen und welche Aspekte ignoriert werden.

Danach betrachten wir den Galton–Watson Prozess in Kapitel 4, der die Dynamik einer Pandemie stochastisch beschreiben kann. Wir leiten einfache Eigenschaften ab und diskutieren die Aussterbewahrscheinlichkeit. Schlussendlich zeigen wir die Verbindung zum Modell für den Reproduktionsfaktor.

In Kapitel 5 stellen wir mehrere Konzepte zur Schätzung der Fallsterblichkeit vor. Wir analysieren wie die Sterbewahrscheinlichkeit von Alter und Geschlecht abhängt und vergleichen sie für mehrere Länder.

Dann betrachten wir den Einfluss von Sterblichkeitsschocks auf Versicherungen in Kapitel 6. Wir diskutieren, ob Covid-19 ein paralleler Schock ist und besprechen die Folgerungen für Lebenserwartung und Rentenpreise.

Weil es viel mehr interessante Probleme gibt als in dieser Arbeit behandelt werden können, stellen wir zum Schluss offene Fragen, um den interessierten Leser zu weiterführender Forschung zu ermutigen.

Stichworte: Covid-19, Serielles Intervall, Inkubationszeit, Reproduktionsfaktor, Modellrisiko, Galton–Watson Prozess, Fallsterblichkeit, Versicherungen, Sterblichkeitsschock

Abstract

The aim of this thesis is to question the models used to describe the dynamic of the Covid-19 pandemic. These models are highly relevant because political and economical restrictions are based on their results. We want to point out where and how certain assumptions have a significant impact on the outcome of these models. At the same time, we give suggestions for improvement when necessary. Throughout this thesis, a special focus is on the mathematical background of estimation methods, infection dynamics and actuarial topics.

After a short introduction in Chapter 1, we start with the estimation of relevant time spans like the serial interval and the incubation time in Chapter 2. We look at several distribution families and estimation methods and compare these to other scientific papers and to the official estimation by the Austrian authority. We also put a strong focus on which data can be observed and which problems occur thereby. Finally, we discuss limitations and potential biases of the results.

To estimate the reproduction number in Chapter 3, we need the result for the serial interval from the previous chapter. We start with a discussion about the definition of a time-varying reproduction number and how it can be observed. Then, we take a look at the model used by the Austrian authority and the underlying assumptions. We discuss which limitations arise therefrom and which effects are ignored by the model.

Further, we study the Galton–Watson process in Chapter 4, which stochastically describes the dynamics of a pandemic. We derive some basic properties and discuss the extinction probability. Finally, we outline the connection to the model for the reproduction number.

In Chapter 5, we provide several concepts for the estimation of the case fatality rate. We analyse briefly how the dying probability depends on age and sex and compare it for different countries.

We then study the impact of mortality shocks on insurance companies in Chapter 6. We discuss whether Covid-19 is a parallel shock and outline the implications for life expectancy and annuity prices.

As there are way more interesting problems than can be dealt with in this thesis, we finally raise open questions to encourage the interested reader to do further research.

Keywords: Covid-19, Serial Interval, Incubation time, Reproduction Number, Model Risk, Galton–Watson Process, Case Fatality Rate, Insurance, Mortality Shock

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisor, Uwe Schmock. He not only supported me with his deep knowledge, but also with his unlimited curiosity. During our collaboration, I experienced the difficulties in building my own model and how to put the theory into practice. Professor Schmock always offered me his time and a helping hand to train the methods of scientific and mathematical work.

Besides, I wish to acknowledge the professional round-the-clock service of Sandra Trenovatz, who always helped me with organizational matters.

Moreover, I would like to thank Ronald Nemec for giving me the necessary freedom at work to finish my thesis in time.

Last but not least, I will never forget the loving and unconditional support of my family, who always believe in me and my ideas and help me to reach my goals.

Thank you!

Daniel Wimmer

Statement of Originality

I hereby declare that I have authored the present master thesis independently and did not use any sources other than those specified. I have not yet submitted the work to any other examining authority in the same or comparable form. It has not been published yet.

Vienna, October 23, 2021

Daniel Wimmer

Contents

1 Introduction

In the beginning of 2020, Covid-19 turned into a global pandemic. The consequences were restrictions on the economy and freedom in order to save as many lives as possible until a significant portion of people have access to vaccination. The drastic political decisions were supported by epidemiological and mathematical models, which tried to predict the future development of the number of infections or other key figures in specific scenarios. To get meaningful results, these models rely on certain assumptions, which are not always trivial. We want to discuss what happens to the models if some assumptions do not hold. Also, we examine how the results and their certainty would be affected. In other words: What is the underlying model risk?

We start with building our own model in Chapter 2 in order to be able to define the serial interval, the incubation time and the generation time precisely. We examine the connection between these time spans, calculate basic properties and try to estimate their distributions. In this thesis, when we build a model or when we want to derive results, we will put a strong focus on the underlying assumptions and the resulting limitations.

However, in Chapter 3, we have already given a model by the Austrian authority. We start with a discussion about the implicit assumptions made. In the following sections, we examine what happens if these assumptions do not hold. The aim of this thesis is to show that the presented results are not as certain as insinuated and that the interpretation can be entirely different if we change an assumption.

Nevertheless, not only the model itself can be subject to inaccuracies, but also the estimation process. The variables used in a model are often easy to define and to calculate with, however, it is sometimes hard or even impossible to observe them. That leads to additional uncertainty, as these variables have to be approximated or estimated itself. Apart from Chapter 2 and 3, we face these problems also in Chapter 5, where we present several approaches to estimate the case fatality rate. Therefore, throughout this thesis, we will state precisely which variables of the model are necessary for a certain calculation and how they are approximated or estimated with the data we have.

However, we not only want to examine the models itself but also look behind the curtain and study its mathematical justification. Therefore, we will study the theory behind the Galton–Watson process in Chapter 4. After showing some basic properties, we will discuss the extinction probability and proof associated theorems. In Section 4.3, we finally get the formula we have already introduced in Section 3.2.

In Chapter 6, we want to provide some practical implications of the Covid-19

pandemic for insurance companies. We begin with a short introduction to actuarial mathematics and discuss mortality shock in general. Then, we discuss the mortality shock of the pandemic and the additional shock at an infection. Further, we study how life expectancy changed with the pandemic and how many years of life were lost due to Covid-19 as additional measure beside the total number of deaths. We also outline the difference for women and men. Finally, we examine the effect on life expectancy and annuity prices.

To conclude, it is a clear aim of this thesis to combine the theoretical background (Chapter 4), the model and estimation process (Chapter 2, 3 and 5) and the practical implications of the results (Chapter 6). This thesis will try to make an objective contribution to an important discussion that is often driven by emotions and ideology.

As the Covid-19 pandemic offers enormous possibilities to do research on and poses innumerable questions, this thesis is far away from covering all areas and therefore does not claim to be complete. Moreover, the level of relevance for certain areas has changed in the course of this pandemic. In order to give the interested reader the opportunity to continue working on the topic, we raise open questions and ideas for further research in Chapter 7.

2 The Estimation of the Serial Interval

The aim of this chapter is to examine several time periods related to the spread of an infectious disease. We will begin defining these time frames and displaying them in a mathematical model. As we are interested in estimating the length of these periods, we are required to discuss which of the occurring variables can actually be observed. Moreover, the mathematical methods will be explained and their adequacy will be discussed.

For the estimation of the corresponding distributions, we will use the data from [12] as well as from [6]. We will critically review the methodology for the estimation of the serial interval used by the AGES[1], described in [22], and discuss possible improvements. This is of major importance, because it is the foundation for the computation of the basic reproduction number in the model of the Austrian authority.

The relevance of this chapter, on the one hand, lies in the necessity of having a best-estimate for further computations in the following chapters. On the other hand, those time spans are epidemiologic key figures, which have to be known to determine the appropriateness and efficiency of policies.

2.1 Set Up of the Model

We start with discussing viral infections, transmissions and symptoms. After a viral transmission from person A to person B, there are the following three possibilities how the body of person B can react:

(1) Virus transmission is instantly fended off by the immune system. It is not possible to infect a third person.

(2) Virus transmission does not lead to symptoms, but it is possible to infect a third person.

(3) Virus transmission leads to symptoms.

In the epidemiological literature, an infection begins "when infectious virus particles (virions) attach to and enter susceptible cells"[2]. However, here we say person B is

[1] Agentur für Gesundheit und Ernährungssicherheit (agency for health and food security)
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3479527/, accessed on August 25, 2020

infected if this person gets infectious itself, so if (2) or (3) is the case. Possibility (2) is called an *asymptomatic case*.

In this chapter, we take a look at pairs of consecutive infections with both persons developing symptoms. Thus, we only consider case (3), because this can be observed more easily than asymptomatic transmissions. In Section 3.5, we get back to this problem of omitting asymptomatic transmissions and discuss the consequences for the estimation of the reproduction number.

Having such a pair of consecutive infections, we define some important points in time. Let i_0, i_1 denote the time of infection and s_0, s_1 the corresponding clinical onset[3] of these two cases. For s_0 and s_1 being well-defined, it is important that both persons get symptoms, that is why we only consider case (3). We treat all four of them as random variables for a certain pair of consecutive cases.

These four variables help us expressing the following time frames. We define the *generation time* G as the duration from the infection of one person to the point in time, when this person infects the next one, so $G := i_1 - i_0$. The focus lies on this number, because it contributes to the spreading speed of the pandemic. However, it is difficult to measure, because, in general, the time of infection cannot be observed.

The length of the *serial interval* S is the time between the clinical onset of consecutive cases, thus $S := s_1 - s_0$. To be in line with the epidemiological literature, we say *serial interval* instead of *serial interval length* in the following, although it is not an interval but a real number. The advantage of this period is that it can be observed explicitly in most cases. However, it may not be clear if someone suffers from two different infections at the same time.

What distinguishes the serial interval from the generation time is the *incubation time* for both cases I_0, I_1, which denotes the duration from the infection to the clinical onset for both cases and are defined as $I_0 = s_0 - i_0$ and $I_1 = s_1 - i_1$, respectively. That makes it difficult to deduce the serial interval from the generation time, as we have additional uncertainty, which comes from the incubation time not being a fixed number. The incubation time is of concrete practical use, as it defines how long somebody should be put under quarantine after a possible contact with an infected person, for example in a foreign country.

Last but not least, we are interested in the time between the symptom onset of the first person and the transmission. We call this *transmission time* and define it as $T = i_1 - s_0$. We notice that $T < 0$ is possible, because a transmission can also happen before symptoms occur. It helps to decide how long an infected person should be put in quarantine after clinical onset.

[3]The term *clinical onset* is the technical term for the beginning of the symptoms.

We summarize the definitions above:

$$\begin{array}{r|l} \text{generation time} & G := i_1 - i_0 \\ \text{serial interval} & S := s_1 - s_0 \\ \text{incubation time} & I_0 := s_0 - i_0, I_1 := s_1 - i_1 \\ \text{transmission time} & T := i_1 - s_0 \end{array} \tag{2.1}$$

We make the following (trivial) assumptions on the time points. Firstly, symptoms cannot occur before this person is infected, so $i_0 \leq s_0$ and $i_1 \leq s_1$. Per definition person 0 is the first who is infected, which leads to $i_0 \leq i_1$. Hence, there remain three possible orders: (i_0, i_1, s_0, s_1), (i_0, i_1, s_1, s_0), (i_0, s_0, i_1, s_1). We see that the serial interval can be negative whereas the generation time is non-negative. In the following i_0, i_1, s_0, s_1 are dates, thus all periods are integers. In detail, G, I_0, I_1 take values in $\mathbb{N}$ and S, T take values in $\mathbb{Z}$. Furthermore, we notice that the following equalities hold by definition:[4]

$$S = G + (I_1 - I_0), \qquad T = S - I_1, \qquad T = G - I_0. \tag{2.2}$$

Additionally, we make a non-trivial assumption for the whole chapter. Let $J := \{1, \ldots, N\}$ be an index set of $N \in \mathbb{N}^*$ pairs of infected persons, then we assume that the random vectors $((G, S, I_0, I_1, T)^i)_{i \in J}$ are identically distributed. We need this assumption to estimate the distributions, or at least mean and variance, of these random variables. Although there might be some interesting differences in the distributions of these periods for criteria like age or sex, we estimate one distribution for all people disregarding these characteristics, because there are too few data points for separate estimations according to these criteria. If the estimation is not independent from a certain criteria like age, it would be imprecise to use the estimation for one population also for another one with a different age structure.

In addition, in Section 2.5, we need to make some further assumptions on the correlation and dependence of these five time periods in order to get some meaningful results.

2.2 Data Issues

When it comes to the estimation of the distribution of those random variables, the first question is: Which data is available? If we consider only pairs (and not chains) of consecutive cases, it is impossible to observe i_0. Also i_1 is difficult to identify, but it can be restricted to an interval where the two individuals could have had contact. Therefore, we introduce two new points in time a, b, which determine the exposure time. Thus, it follows that $i_0 \leq a \leq i_1 \leq b \leq s_1$. We additionally assume that an

[4] These equations in combination with $I_0 \overset{\mathrm{d}}{=} I_1$ do not imply $G \overset{\mathrm{d}}{=} S$, because T is not independent from I_0 or I_1.

infected person knows when symptoms have started, so that we can observe s_0, s_1 exactly. To sum up, we got four variables for each pair: (s_0^i, s_1^i, a^i, b^i) for $i \in J$. Thus, we can work with the data from [12] in our model, which contains all four variables for $N = 35$ observations. For the cases where b is unknown, we set $b = s_1$, as this is the upper bound for exposure time.

For the estimation of the serial interval, we will use the data from [6] because we have $N = 468$ data points, but only (s_0^i, s_1^i) is available.

Most papers use data from publicly available reports, where many cases have to be excluded due to missing information. The question whether these two steps would lead to a selection bias was addressed by [16, p. 11], but they found that the distribution of age and sex is similar to the original cases for their data. Another problem comes with transmission clusters, which are often the major part of the data. As the correct order is more difficult to distinguish, the transmission chain was build up by the order of symptom onset, which ignores that incubation time is not deterministic. As the order is crucial for the estimation of the epidemiological key figures, this leads to a problem, which is made even worse by asymptomatic transmissions. They are hard to find and therefore pose a high risk, because there are no appropriate control measures, as [16, p. 4] suggests.

For different families of distributions, we estimate the necessary parameters with parametric methods like the method of moments (MOM) and maximum-likelihood estimation (MLE). More details about these methods can be found in [13, p. 326-334]. For all our estimations, we need independent and identically distributed observations. Therefore, we add this assumption for both datasets, which could be problematic due to superspreading events.

2.3 Serial Interval

We start with the serial interval S because it can be extracted directly from our data. Let J be an index set, then we get $\hat{S}^i = s_1^i - s_0^i$ for $i \in J$ and collect them in a vector $\hat{S} = (\hat{S}^i)_{i \in J}$. To get a first feeling for the data from [6], we see a box plot in Figure 2.1.

Although the data is discrete, we want to approximate it with a continuous distribution. Due to the fact that also negative values are possible, we do not want to use some of the common distributions like gamma, Weibull or log-normal. The problem that many estimations of the serial interval use these distributions, ignoring the negative values, for example in [22, p. 1] and [18, p. 285], is also mentioned by [6, p. 1341] and [16, p. 7]. In addition, in the appendix of [6], the fits for shifted, truncated and the original data are compared for gamma, Weibull and log-normal distribution for the positive datasets and normal distribution for all datasets. They found that the normal distribution fits best for the original, as well as for the shifted data.

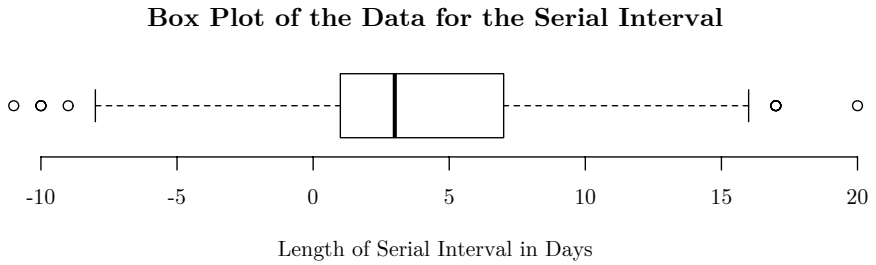

Figure 2.1: We see a box plot for the length of the serial interval with the data from [6]. How box plots are defined can be found in [13, p. 81].

If we disregarded negative values, we would neglect an important part of the distribution, because transmission is also possible before clinical onset. Therefore, we use a generalisation of the normal distribution which can deal with skewness, namely the skew-normal distribution, in order to have more freedom than with a classical normal distribution. Given a location parameter $\xi \in \mathbb{R}$, a scale parameter $\omega > 0$ and a shape parameter $\alpha \in \mathbb{R}$, the *skew-normal distribution* is defined by its density:

$$f(x|\xi,\omega,\alpha) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\frac{x-\xi}{\omega}\right), \qquad x \in \mathbb{R}, \tag{2.3}$$

where ϕ and Φ are the density and distribution function of the standard normal distribution, respectively. For $\xi = 0$ and $\omega = 1$, we call this distribution *standard skew-normal*. For these parameters, it is easy to see that f is really a density. We see that f is positive and that

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x &= \int_{-\infty}^{\infty} 2\phi(x)\int_{-\infty}^{\alpha x}\phi(y)\,\mathrm{d}y\,\mathrm{d}x = 2\int_{-\infty}^{\infty}\int_{-\infty}^{\alpha x}\phi(x)\phi(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= 2\left(\int_{-\infty}^{0}\int_{-\infty}^{\alpha x}\phi(x)\phi(y)\,\mathrm{d}y\,\mathrm{d}x + \int_{0}^{\infty}\int_{-\infty}^{\alpha x}\phi(x)\phi(y)\,\mathrm{d}y\,\mathrm{d}x\right) \\
&= 2\left(\int_{0}^{\infty}\int_{-\infty}^{-\alpha x}\phi(-x)\phi(y)\,\mathrm{d}y\,\mathrm{d}x + \int_{0}^{\infty}\int_{-\alpha x}^{\infty}\phi(x)\phi(-y)\,\mathrm{d}y\,\mathrm{d}x\right) \\
&= 2\int_{0}^{\infty}\int_{-\infty}^{\infty}\phi(x)\phi(y)\,\mathrm{d}y\,\mathrm{d}x = 2\int_{0}^{\infty}\phi(x)\,\mathrm{d}x = 1.
\end{aligned}$$

Therefore, f is a density, where we can add a location and scale parameter. For $\alpha = 0$, we have a normal distribution with mean ξ and variance ω^2. For fitting this distribution to the data from [12], we try two different methods, namely the method of moments and maximum-likelihood estimation. For both methods, we need independent and identically distributed random variables.

Firstly, we try the method of moments. For a random variable X that is standard skew-normal distributed, we get the formulas for the first moment and the second

and third central moment from [3, p. 342]. So, with

$$m = \frac{\alpha}{\sqrt{1+\alpha^2}}\sqrt{\tfrac{2}{\pi}},$$

we have

$$\begin{aligned}\mathbb{E}[X] &= m,\\ \mathbb{E}[(X-\mathbb{E}[X])^2] &= 1-m^2,\\ \mathbb{E}[(X-\mathbb{E}[X])^3] &= \frac{4-\pi}{2}m^3.\end{aligned}$$

For a skew-normal distributed random variable Y, we can adjust these formulas with the location parameter ξ and scale parameter ω, so we get

$$\begin{aligned}m_1 &:= \mathbb{E}[Y] = \xi + \omega m,\\ m_2 &:= \mathbb{E}[(Y-\mathbb{E}[Y])^2] = \omega^2(1-m^2),\\ m_3 &:= \mathbb{E}[(Y-\mathbb{E}[Y])^3] = \frac{4-\pi}{2}\omega^3 m^3.\end{aligned}$$

If we now estimate the moments for our data, we can solve these equations, and calculate the parameters ξ, ω, α.

Secondly, we conduct a maximum-likelihood estimation. Therefore, we need to maximize the likelihood function:

$$L(\xi,\omega,\alpha|\hat{S}) := \prod_{i\in J} f(\hat{S}^i|\xi,\omega,\alpha), \tag{2.4}$$

which is defined on $\mathbb{R}\times\mathbb{R}_+\times\mathbb{R}$ As we cannot do this explicitly, we maximize it numerically.

These estimation processes were implemented in R using the function `sn.mple()` from the package `sn`, where all necessary functions regarding the skew-normal distribution are implemented, and lead to the following results:

	ξ	ω	α
MOM	−0.60	6.58	1.75
MLE	−0.49	6.50	1.69

(2.5)

Further, we are interested in the mean and variance as well as in the 2.5%- and 97.5%-quantile of the data and our estimations, which are printed in the following table:

	$\mathbb{E}$	Var	$q_{2.5\%}$	$q_{97.5\%}$
data	3.96	22.5	−5	15
MOM-fit	3.96	22.5	−4.5	14.1
MLE-fit	3.97	22.3	−4.5	14.1

(2.6)

Trivially, the mean and variance for the method of moments parameters is the same as for the data. We observe that both estimations lead to similar results.

We want to display our data using a kernel density function, which is the average over the density functions of normal distributions around each data point with a certain standard deviation $\sigma \in \mathbb{R}_+$. So, we define the kernel density function as

$$k(x) := \frac{1}{N} \sum_{i=1}^{N} \phi\left(\frac{x - \hat{S}^i}{\sigma}\right), \qquad x \in \mathbb{R}. \tag{2.7}$$

Here and in the following, the kernel density is computed using the function `density()` from the package `stats` in R.

In Figure 2.2, we see a comparison between the kernel density and the two fitted skew-normal distributions. We observe that the two estimated distributions are so similar that it is not possible to visually distinguish between them.

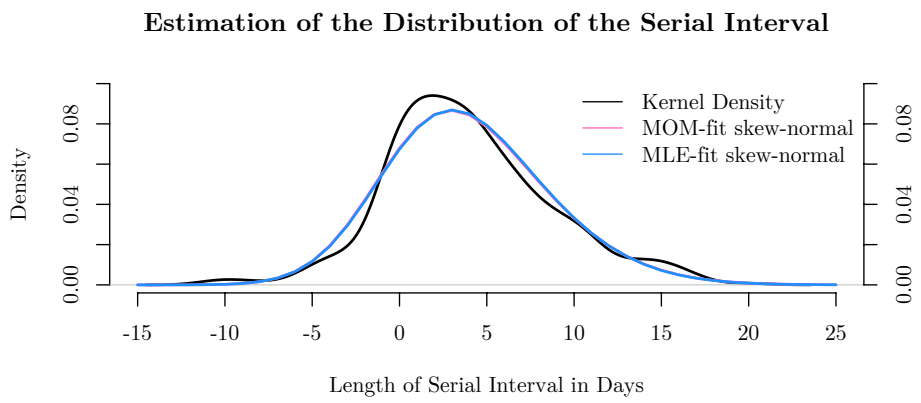

Figure 2.2: We see the kernel density with the underlying kernel density function defined in (2.7) and the skew-normal fit with the parameters from MLE and MOM for the length of the serial interval from (2.5).

In [12, p. 674], they mention the possible problem that the estimation of the serial interval relies on whether the patient can recall the date of clinical onset correctly. But if these biases do not differ systematically within patients, this inaccuracy have almost no influence. Furthermore, a potential selection bias brought up by [30, p. 14] is that transmission pairs with a shorter serial interval can be found easier and earlier. That implies that the estimated serial interval is too small. Additionally, [30, p. 10] proposes that possibly leaving out an intermediary in the transmission chain would imply an overestimation of the serial interval. This problem occurs especially when working with cluster data.

Besides, [6, p. 1342] argues that the serial interval changes over time: The more people are infected, the shorter will the serial interval be, as “a susceptible person is

likely to become infected more quickly if they are surrounded by 2 infected persons instead of 1". However, this statement can be doubted. It is just more likely that a susceptible person is infected, but the serial interval measures the time till another one gets infected and that does not change as long as the contact intensity to susceptible people does not change. Further, the serial interval reacts to policies. If people are immediately isolated after symptom onset, the serial interval gets smaller, as no further people can be infected, who would lead to a greater serial interval. In contrast, social distancing measures have no influence on the length of the serial interval, but on the number of transmissions per person.

To conclude, our calculations suggest a much higher variance of the serial interval than in [22, p. 1] and the presence of negative values. We will examine the impact of the serial interval on the estimation of the basic reproduction number in Section 3.7.

2.4 Incubation Time

The incubation time is more difficult to estimate, as we know neither i_0 nor i_1. But if we use x, y, we can at least estimate the incubation time for the second person I_1. If we assume that every infected individual has exactly one infector[5], there is a bijection between the pairs and the infected persons. Therefore, a random choice of pairs is also a random choice of infected persons.[6] Thus, if I denotes the incubation time for a random case, we have[7] $I_1 \overset{\mathrm{d}}{=} I$. We will try to fit a gamma, log-normal and Weibull distribution, since the data is non-negative and those are frequently used in literature. However, we will also estimate a negative binomial distribution, because the underlying data from [12] is discrete.

For all estimations, we want to use the maximum-likelihood approach, so we need independent and identically distributed random variables. The problem is that we cannot observe i_1 directly, so we cannot use the standard method. However, we know that $i_1 \in \{a, a+1, \dots, b\}$, and therefore, I_1 conditional on s_1, a and b,[8] takes values in $K := \{s_1 - b, s_1 - b + 1, \dots, s_1 - a\}$.

For the discrete case, we notice that

$$\mathbb{P}(I_1 \in K|\theta) = \sum_{k=a}^{b} f(s_1 - k|\theta),$$

[5] This should be true for most of the cases. If someone was exposed to viral transmission from more than one person, we say that the last one is the infector.

[6] In general, this is not true for I_0, because people who infect more other people and are therefore overrepresented in the first cases of a random sample of pairs, can have a different incubation time than those who infect less other people.

[7] For two random variables X, Y, $X \overset{\mathrm{d}}{=} Y$ means that X and Y are equal in distribution.

[8] That is a reasonable approach, as we know these variables in the specific case.

where f is the probability mass function of the chosen distribution with parameter vector θ. So, we can expand the concept of maximum-likelihood estimation by changing the likelihood function into a product of sums of probabilities given θ, what is also mentioned in [33]:

$$L(\theta|(s_1, a, b)) = \prod_{i=1}^{n} \sum_{k=a^i}^{b^i} f(s_1^i - k|\theta). \tag{2.8}$$

This means that we do not have to assume a certain probability distribution of I_1 (conditional on s_1, a and b) on the value set K.

For the continuous case, we need an integral instead of the sum and we have to integrate over $[a - 0.5, b + 0.5]$, so that the time frame has the same length as in the discrete case. Then, we get

$$L(\theta|(s_1, a, b))) = \prod_{i=1}^{n} \int_{a^i-0.5}^{b^i+0.5} f(s_1^i - z|\theta)dz, \tag{2.9}$$

where f is a density function of the chosen distribution with parameter vector θ. We found the maximum of all likelihood functions numerically using the functions `integrate()` and `optim()` from the package `stats` in R.

As comparison for these estimated distributions, we want to build a probability mass function similar to the kernel density function from Section 2.3. Therefore, we assume that i_1 is equally distributed in $\{a, a + 1, \ldots, b\}$ and so I_1 (conditional on s_1, a and b) is equally distributed in K. Now, we can take the average over these probability mass functions of discrete uniform distributions. So, we have

$$p(k) := \frac{1}{N} \sum_{i=1}^{N} \frac{1}{b^i - a^i + 1} \cdot \mathbb{1}_{k \in K^i}, \qquad k \in \mathbb{N}. \tag{2.10}$$

In the following table, we see the results for the estimated parameters for the different distributions:

negative-binomial	size= 4.772, prob= 0.438
gamma	shape= 2.45, scale= 2.50
log-normal	$\mu = 1.58$, $\sigma = 0.72$
Weibull	shape= 1.78, scale= 6.94

(2.11)

Further, we are interested in the mean and variance as well as in the 95%-quantile of p and our estimations. These numbers are printed in the following table:

	$\mathbb{E}$	Var	$q_{95\%}$
p	6.17	16.0	13
negative-binomial	6.12	14.0	13
gamma	6.12	15.3	13.6
log-normal	6.29	26.8	15.8
Weibull	6.17	12.8	12.8

The estimated distributions and p are plotted in Figure 2.3. We observe that a quarantine of 14 days is appropriate.

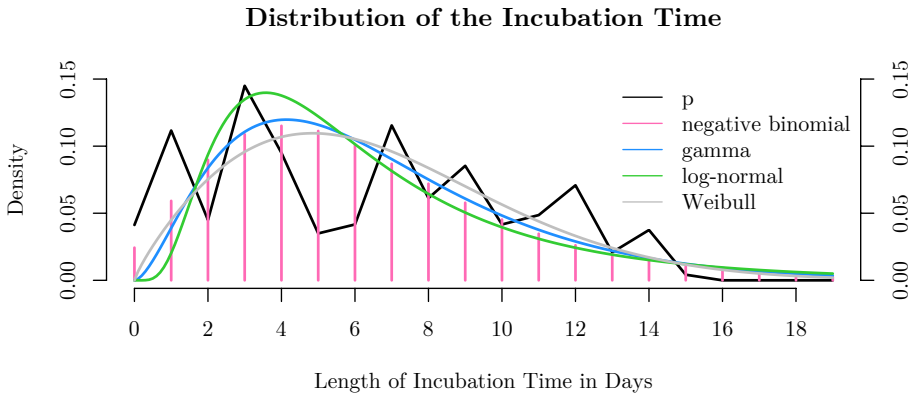

Figure 2.3: For the incubation time, we see p as defined in (2.10) and estimated negative binomial, gamma, log-normal and Weibull distribution with the parameters from (2.11).

A potential bias of the estimation of the incubation time, mentioned by [12, p. 674], is that there might be a delay in recognizing first symptoms. That would lead to an overestimated incubation period. Similar as for the serial interval, [30, p. 10] pointed out that skipping a possible intermediary in the chain of transmissions would lead to a further overestimation of the incubation time. Additionally, [16, p. 2] proposes that the incubation time is longer for cases with a less severe course of the disease, which would lead to an underestimation if the severe cases are overrepresented, what is suggested. Further, an asymptomatic infection cannot be captured by these considerations since only symptomatic cases have an incubation time.

To summarize, our results confirm that it is reasonable to put people under quarantine for two weeks if they might have had contact with infected subjects, for example if they come from countries that have many cases or are in near social environment of positive tested people. The long and highly-variable incubation time is a fact that sets Covid-19 apart from influenza and makes it hard to contain the pandemic, which is also stated by [7, p. 1].

2.5 Generation Time

As we do not have any information about i_0, we cannot estimate the distribution, mean or variance of the generation time without additional non-trivial assumptions. Within this section, we assume the following:

(1) $I_0 \stackrel{\mathrm{d}}{=} I_1$,

(2) I_0 and I_1 are uncorrelated,

(3) G and I_1 are uncorrelated.

Before we take a look at the results under these assumptions, we want to discuss their limitations and which possible cases or scenarios are excluded by them. As already mentioned in Section 2.4, assumption (1) is not straightforward. Thereby, we exclude the possibility that the number of infection an individual generates is dependent on the incubation time. Although this dependency is not proven yet, it seems plausible. Nevertheless, we want to see which information about the generation time can be deduced with this assumption.

With assumption (2), we assume that the second incubation time is not influenced by the first one, so we disregard the following: It could be possible that the length of the incubation time has an impact on the size of the viral load with which the second person is infected, which can influence the length of the second incubation time.

A similar problem comes up with assumption (3), because the size of the viral load could also depend on the length of the generation time. However, we disregard this possibility.

To conclude, all three assumptions are based on the idea that every infection is equivalent in the sense that it is not distinguished between a small or big viral load or a slight or heavy infection. Also, the course of the disease is only determined by properties of the body of the infected person and not by characteristics of the infection. Although we would even get the independence for assumption (2) and (3) with this idea, the uncorrelatedness is enough for our propose.

Using these assumptions, we can calculate some values regarding the distribution of the generation time. Using equation (2.2), assumption (1) and the results from Section 2.3, we obtain that

$$\mathbb{E}[G] = \mathbb{E}[S] = 4.0.$$

If we transform equation (2.2) to $G + I_1 = S + I_0$ and calculate the variance of both sides, we get

$$\mathrm{Var}(G) + \mathrm{Var}(I_1) + 2\,\mathrm{Cov}(G, I_1) = \mathrm{Var}(S) + \mathrm{Var}(I_0) + 2\,\mathrm{Cov}(S, I_0).$$

Subtracting $\mathrm{Var}(I_1) = \mathrm{Var}(I_0)$ due to assumption (1) and using assumption (3), leads to

$$\mathrm{Var}(G) = \mathrm{Var}(S) + 2\,\mathrm{Cov}(S, I_0). \tag{2.12}$$

With equation (2.2) and assumption (2), we additionally get the following relation

$$\mathrm{Cov}(S, I_0) = \mathrm{Cov}(I_1, I_0) + \mathrm{Cov}(T, I_0) = \mathrm{Cov}(T, I_0) = \mathrm{Cov}(G, I_0) - \mathrm{Var}(I_0). \tag{2.13}$$

To get a more meaningful result, we can take a look at two possible scenarios about the infectiousness. At first, we assume that the infectiousness of a person depends

solely on the date of the clinical onset and not on the date of the own infection. This implies that T and I_0 are uncorrelated, because the length of the first incubation time I_0 has no impact on the infectiousness and therefore on the infection of the second person i_1. Combining this information with the equations (2.12) and (2.13), leads to

$$\operatorname{Var}(G) = \operatorname{Var}(S).$$

The second possible case we take a look at is that the infectiousness of a person depends only on the date of the own infection and not on the beginning of the symptoms, therefore the lengths of the periods from i_0 to s_0 and from i_0 to i_1 are uncorrelated. That implies that $\operatorname{Cov}(G, I_0) = 0$ and together with the equations (2.12) and (2.13), we have

$$\operatorname{Var}(G) = \operatorname{Var}(S) - 2\operatorname{Var}(I_0) < \operatorname{Var}(S).$$

The relation $\operatorname{Var}(G) < \operatorname{Var}(S)$ is also proposed by [10, p. 2].

Looking at the results from the previous sections, we see that the second assumption has to be rejected in our setting with the data from [12], as the indicated variance would be negative. Thereby, we can deduce that the date of the symptom onset is positively correlated to the date of the infection of the next person. Additionally, this justifies the definition of the transmission time T as something useful. We will elaborate more on that in Section 2.6.

However, it seems pretty unlikely that $\operatorname{Cov}(G, I_0) < 0$ or that $\operatorname{Cov}(T, I_0) > 0$. Thus, if $\operatorname{Cov}(G, I_0) \geq 0$ and $\operatorname{Cov}(T, I_0) \leq 0$ hold true, we have a lower and upper bound for the variance of the generation time, namely $\operatorname{Var}(G) \in [\operatorname{Var}(S) - 2\operatorname{Var}(I_0), \operatorname{Var}(S)]$. For our results from the previous sections for the skew-normal and negative binomial MLE, we get $\operatorname{Var}(G) \in [-5.7, 22.3]$.

To conclude this section, it is impossible to estimate the distribution of the generation time given only transmission pairs. Even the variance is not straightforward and requires additional assumptions. Thus, whenever an estimation for the generation time is stated, there are significant limitations due to critical assumptions. Also, using the serial interval as a proxy for the generation time is difficult, as it requires the incubation time to be constant or at least equal within the transmission pairs.

2.6 Transmission Time

Now, we want to examine the infectiousness with respect to the clinical onset. By identifying at which stage of the disease other people are infected, it is possible to optimize isolation policies. According to the definition in (2.1), and as we have only a and b to estimate i_1 in the data from [12], we conduct an extended maximum-likelihood estimation as for the incubation time in Section 2.4. It might also be interesting which portion of transmissions occur pre-symptomatic, in order to evaluate the accuracy and the success of current policies.

For estimating the distribution of T, we choose again the skew-normal distribution (compare Section 2.3), because we want to model skewed data and might have (theoretically unlimited) negative values. As our data is discrete, we need a continuity correction. So, we have the likelihood function

$$L(\theta|(s_0,a,b))) = \prod_{i=1}^{n} \int_{a^i-0.5}^{b^i+0.5} f(z - s_0^i|\theta)dz, \tag{2.14}$$

where f is the density of the skew-normal distribution and $\theta = (\xi, \omega, \alpha)$, compare equation (2.3). We maximized this function numerically with the functions `integrate()` and `optim()` from the package `stats` in R and got the results $\xi = -4.45$, $\omega = 5.93$, $\alpha = 1.83$.

As in equation (2.10), we define a naive estimation by taking the average over the probability mass functions of uniform distributions in $K^i := \{a^i - s_0^i, a^i - s_0^i + 1, \ldots, b^i - s_0^i\}$ for $i \in \{1, \ldots, N\}$, so

$$p(k) := \frac{1}{N} \sum_{i=1}^{N} \frac{1}{b^i - a^i + 1} \cdot \mathbb{1}_{k \in K^i}, \qquad k \in \mathbb{Z}. \tag{2.15}$$

For the portion of pre-symptomatic transmissions, we look at $\mathbb{P}(T < 0)$. In the following table, we see the results:

	$\mathbb{E}$	Var	$q_{2.5\%}$	$q_{97.5\%}$	$\mathbb{P}(T<0)$
p	-0.30	20.6	-10	11	59.7%
skew-normal MLE	-0.30	17.9	-7.8	8.8	55.7%

In Figure 2.4, we see the comparison of the density function of the estimated skew-normal distribution and the naive estimation p, which is not smooth because of too few data points.

We see that more than half of all transmissions are earlier than the clinical onset. However, there are not only virological reasons for the significant ratio of pre-symptomatic transmissions like the size of the viral load, but also behavioural ones, because the major part of the people changes their behaviour after the beginning of the symptoms. In [12, p. 672], [30, p. 1] and [18, p. 285], it is proposed that $\mathbb{E}[S] < \mathbb{E}[I]$ is a simple indicator for a significant portion of pre-symptomatic transmissions. This is consistent with our model, as $\mathbb{E}[S] < \mathbb{E}[I]$ is with equation (2.2) equivalent to $\mathbb{E}[T] < 0$, which means that on average a transmission takes place before infector's symptom onset. For our estimates, $\mathbb{E}[S] < \mathbb{E}[I]$ is true, because of $4.0 < 6.1$, so we are in line with literature.

The potential bias of overestimating the incubation time due to a delay in recognizing first symptoms mentioned in [12, p. 674] would imply an underestimated transmission time, which means that the portion of pre-symptomatic transmissions is actually smaller.

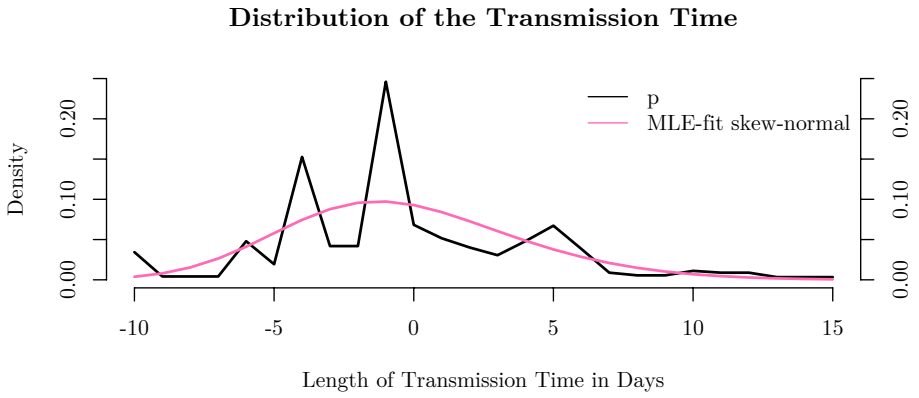

Figure 2.4: For the transmission time, we see p as defined in (2.15) and a density of the skew-normal distribution with the parameters estimated with MLE.

We want to conclude this section talking about the practical implications. As [10, p. 7] suggests, physical distancing is the most important policy, because also pre-symptomatic transmissions can be avoided. Compared to contact tracing and quarantine measures, it is therefore much more efficient. This inefficiency of the isolation policies is also mentioned by [12, p. 672] and [18, p. 286]. However, it can also be argued that post-symptomatic transmissions are so low just because of these isolation measures. This argument is supported by [30, p. 7], which found that the transmission time decreases in the course of the pandemic.

2.7 Discussion of the AGES Methodology

In this section, we want to discuss [22], a paper on the estimation of the serial interval published by the AGES, which is part of the Austrian Ministry of Health. It is of specific interest to check the model risk of this estimation, because it contributes to the model risk of the reproduction number, which uses the serial interval in its calculation.

In [22, p. 3], two other papers were compared to their results, namely [6] and [18]. In the following table we analyse the results.

	distribution	mean	sd	$q_{2.5\%}$	$q_{97.5\%}$	$S \in$
here	skew-normal($\xi = -0.49$, $\omega = 6.50$, $\alpha = 1.69$)	4.0	4.7	-4.5	14.1	$\mathbb{R}$
[22]	gamma(shape $= 2.88$, scale $= 1.55$)	4.5	2.6	0.9	10.9	$\mathbb{R}_+$
[6]	normal($\mu = 3.96$, $\sigma^2 = 22.56$)	4.0	4.8	-5.3	13.3	$\mathbb{R}$
[18]	log-normal($\mu = 1.39$, $\sigma^2 = 0.32$)	4.7	2.9	1.3	12.2	$\mathbb{R}_+$

It can be observed that those estimations where S is constrained on $\mathbb{R}_+$ have a significantly lower standard deviation, which is supported by Figure 2.5, where the

densities of the estimated distributions are plotted. This implies that dropping negative values has a strong impact and might lead to wrong results as negative serial intervals are treated as a measurement error, but are naturally occurring. Additionally, longer serial intervals are underrepresented, because they are more difficult to find. These effects lead to a smaller standard deviation. Further, the mean is larger for the constrained distributions. That is important for the measurement of the speed of the pandemic spread, which we will discuss in Section 3.1.

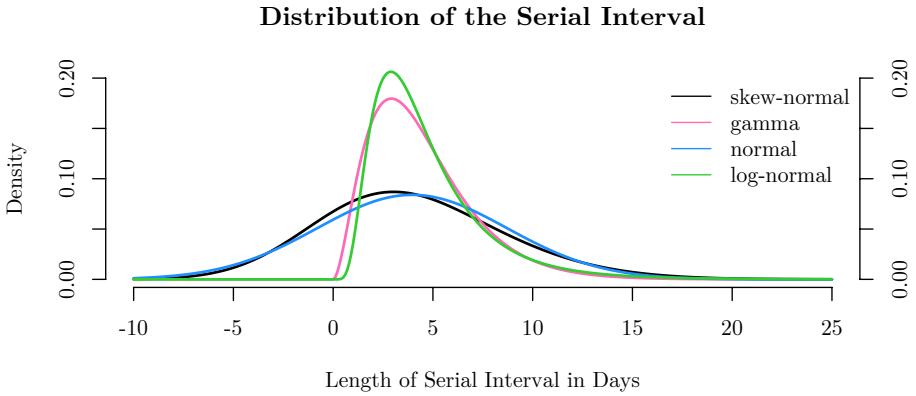

Figure 2.5: For the length of the serial interval, we see a comparison of the results for the estimated densities from the different papers and our approach.

To conclude, distribution families only defined on $\mathbb{R}_+$ are not suitable for the serial interval, as it may take negative values due to a not constant incubation time, which is confirmed by [6, p. 1343]. Besides, one have to be cautious with a low standard deviation, because there might be more uncertainty in the serial interval than estimated. We discuss the implications of these mistakes in Section 3.7.

3 The Estimation of the Reproduction Number

We want to continue our estimation of epidemiological key figures with the *reproduction number* R. According to [21, p. 18], the basic reproductive rate is defined as "the number of infections produced, on average, by an infected individual in the early stages of a pandemic, when virtually all contacts are susceptible". However, R is dimensionless and therefore formally not a rate. It is also controversial, how many people have already been infected or had antibodies before the outbreak. Furthermore, it is not scientifically proven that surviving an infection makes someone immune against a second infection. Additionally, from the definition above, it is not clear if "the number of infections produced" counts only those who were directly infected by the individual or if it counts all infections that took place because of this individual, so subsequent infections in the transmission chains. Besides, it varies in scientific literature whether the reproduction number is fixed for a certain disease or time-dependent.

It is difficult to define a time-dependent reproduction number so that it is mathematically well defined. We can define it as the ratio between two consecutive generations. However, as the generation time is not constant, we cannot observe the size of both generations and need an estimation for one them. One of these two generations can be fixed as the number of new infection at a day. We can either fix the first generation and estimate the size of the second or vice-versa. If we fix the first generation, we can theoretically define the consecutive generations as the number of people that were infected by the first generation, but we cannot observe that exactly. If we fix the second generation, we see that it is not possible to get a well defined first generation: It is clear that all infectors of the second generation have to be part of the first generation. However, we do not know how to deal with the cases that have not generated an infection on that day. The advantage of fixing the second generation is that we get a more current number, which is important for the handling of a pandemic.

That is the reason, why we will use the number of incident cases of a day as the second generation and an estimation for the first described in Section 3.2, whenever we work with real data. We can interpret the reproduction number R_t as the average number of infections generated by one infected individual at day t. That is mathematically not well-defined, because we cannot specify from which sample we take the average, but it helps to understand the meaning of this number.

Allowing the reproduction number to vary over time, gives us the chance to model the course of a pandemic spread in a certain population. It mainly depends on the number of people who are immune, the number of physical contacts and the infectiousness of each individual.

The reproduction number is not only of scientific interest, but also important for political decisions like physical distancing, exit restrictions or the closing of schools and stores. In Spring 2020, after the shutdown of the Austrian economy, the reproduction number was the most important measure of the development of the pandemic and easing measures were promised conditioned on it.

The estimation of this important number is not straightforward and requires the construction of a model. Therefore, the result is exposed to model risk. There are a lot of assumptions which have to be made in order to be able to estimate the reproduction number and corresponding confidence bounds. Some of those might be unrealistic or too restrictive. We will discuss which assumptions are required to conduct the calculation and which consequences they have on the result. Furthermore, we will explain which of them can be softened or changed to get a more precise result.

However, it is not only a matter of precision, but also a question of time. To minimize the number of infected people, it is important to identify an upward trend in the reproduction number as early as possible. On the other hand, easing measures can be implemented earlier, if we find a way to let the reproduction number react faster on the infection dynamic.

This chapter puts a special emphasis on the estimation used by the AGES, which is a part of the Austrian Ministry of Health and is responsible for the monitoring of the pandemic and the estimation of the reproduction number. The methodology is described in [23].

3.1 The Interpretation of the Reproduction Number and its Connection to the Serial Interval

As the reproduction number is the ratio between two generations, it follows that $R_t > 1$ for $t \in \{t_0, t_0 + 1, \ldots, t_1\}$ means that the pandemic expands between t_0 and t_1 and $R_t < 1$ that it goes back. However, R_t alone does not tell us, how fast the virus spreads. This can only be observed in combination with the serial interval or the generation time, which were discussed in the previous chapter. As the generation time is the time difference between the infections of a transmission pair, the next generation of infected people arises on average after $\mathbb{E}[G]$ days. From Section 2.5, we remember that $\mathbb{E}[G] = \mathbb{E}[S]$, which is important because S can be observed, but G not.

We want to combine the reproduction number and the generation time to get a measure for the speed of the spread, namely the doubling time. We first assume

that the reproduction number is constant over time. If $N_1, N_2 \in \mathbb{N}$ are the sizes of two consecutive generations of infected people and $R \in \mathbb{R}_+$ the reproduction number between these generations, then

$$N_2 = N_1 \cdot R.$$

Thus, if we want to know how many generations it takes for the size of the infected population to double, we calculate $\tau \in \mathbb{R}$ from

$$2 = R^\tau \Leftrightarrow \tau = \frac{\log(2)}{\log(R)}.$$

So, after τ generation times the generation size has doubled. If we multiply this with the average generation time, we get

$$T := \frac{\log(2)}{\log(R)} \mathbb{E}[G], \tag{3.1}$$

which gives us the doubling time in days. For a constant reproduction number, we get an exponential process with growth factor $R^{1/\mathbb{E}[G]}$.

If we drop the assumption of R being constant, we can use the R_t for one day and calculate the instantaneous doubling time, but we can also take the geometric mean for the last days in order to get a smoother result.

The doubling time does not additionally depend on the time to recovery. This effect is already incorporated in the generation time. However, a problem arises when the generation time changes in the course of the pandemic. This could happen due to isolation and contact tracing measures, as already mentioned in Section 2.3. That makes the resulting reproduction numbers incomparable, because if we assume that the doubling time is fixed, we observe that a shorter generation time and serial interval leads to a lower R, which is also mentioned by [30, p. 2] and [9, p. 6].

3.2 Replication of the Official Methodology

We want to start our calculations with understanding the model in [23] and reconstructing the official results. In the following, we want to emphasize the assumptions that were made.

For each $t \in \mathbb{N}^*$, let Y_t be a random variable describing the number of incident cases on day t. We assume that Y_t conditional on $(Y_s)_{s<t}$ is Poisson distributed with parameter λ_t, compare equation (3.2). Further, for a fixed $T \in \mathbb{N}^*$ let $S = \{1, 2, \ldots, T\}$ and for each $s \in S$ let w_s be the probability that a new case is generated s days after the infection of the infector. So, $\sum_{s \in S} w_s = 1$ if we assume that an infector can only generate a case within T days after the own infection. The $(w_s)_{s \in S}$ are deterministic weights and are supposed to represent the distribution of the serial interval. It follows

that the number of infected people at time $t-s$ contributes with $R_t \cdot w_s$ to the number of infected people at time t. This leads to

$$\mathbb{E}[Y_t | Y_{t-s}, s \in S] = \lambda_t = R_t \cdot \sum_{s=1}^{T} Y_{t-s} w_s, \qquad t \in \mathbb{N}^*, \tag{3.2}$$

where we assume that $Y_k = 0$ for $k \leq 0$. In Section 4.3, we get the same equation using an extended Galton–Watson process. We notice that this sum is the estimation for the size of the previous generation mentioned in the beginning of this chapter.

Given $\tau = 13$ and a $t \geq \tau$, suppose that R_t is constant for the period $\{t-\tau+1, \ldots, t\}$. Now, we use Bayes inference to estimate R_t. Therefore, we take a Gamma(a, b) as the a-priori-distribution, which has the density

$$p(x) = \frac{x^{a-1} \mathrm{e}^{-x/b}}{b^a \Gamma(a)}, \qquad x > 0.$$

Further, we need the likelihood function for the Poisson distribution, which is given by

$$L(y; R_t) = \prod_{i=t-\tau+1}^{t} \left(R_t \sum_{s=1}^{T} y_{i-s} w_s \right)^{y_i} \exp\left(-R_t \sum_{s=1}^{T} y_{i-s} w_s \right) \frac{1}{y_i!}$$

given that $Y_k = y_k$ for $k \leq t$ and $y_k = 0$ for $k \leq 0$. To get the resulting a-posteriori-density of R_t, we multiply the a-priori-density with the likelihood function and scale it in order to get a function with integral 1:

$$\begin{aligned} p(R_t; y) &= \frac{p(R_t) L(y; R_t)}{\int p(x) L(y; x) dx} \\ &= k(y, t, a, b) \cdot R_t^{a-1} \mathrm{e}^{-R_t/b} \cdot \prod_{i=t-\tau+1}^{t} R_t^{y_i} \exp\left(- R_t \sum_{s=1}^{T} y_{i-s} w_s \right) \\ &= k(y, t, a, b) \cdot R_t^{a-1+\sum_{i=t-\tau+1}^{t} y_i} \exp\left(- R_t \left(\frac{1}{b} + \sum_{i=t-\tau+1}^{t} \sum_{s=1}^{T} y_{i-s} w_s \right) \right), \end{aligned} \tag{3.3}$$

where $k(y, t, a, b)$ combines all factors that do not depend on R_t. Thus, the a-posteriori-distribution of R_t given that $Y_k = y_k$ for $k \leq t$ is

$$\text{Gamma}\left(\tilde{a} = a + \sum_{i=t-\tau+1}^{t} y_i, \tilde{b} = \frac{1}{\frac{1}{b} + \sum_{i=t-\tau+1}^{t} \sum_{s=1}^{T} y_{i-s} w_s} \right). \tag{3.4}$$

The mean of the gamma distribution of R_t is given by $\tilde{a} \cdot \tilde{b}$, therefore our estimator for the reproduction number is

$$\hat{R}_t = \frac{a + \sum_{i=t-\tau+1}^{t} y_i}{\frac{1}{b} + \sum_{i=t-\tau+1}^{t} \sum_{s=1}^{T} y_{i-s} w_s} \tag{3.5}$$

The 95%-confidence interval is calculated by the 2.5%- and the 97.5%-percentile. In [23, p. 3], they suggest a Gamma$(a = 1, b = 5)$ as a-priori-distribution.

In [23, p. 2], the weights $(w_s)_{s \in S}$ are chosen as

$$w_s = \frac{f(s)}{\sum_{r \in S} f(r)}, \qquad f(s) = \frac{s^{a-1} \mathrm{e}^{-s/b}}{b^a \Gamma(a)}, \qquad s \in S,$$

with f being a density function of a Gamma$(a = 2.88, b = 1.55)$ distribution estimated in [22, p. 1].

To sum up, the following assumptions were made:

(1) Y_t is conditionally Poisson distributed,

(2) y_t is the correct number of incident cases,

(3) the serial interval is a good proxy for the weights w_s,

(4) the estimation of the serial interval is reasonable,

(5) the reproduction number is constant in $\{t - \tau + 1, \dots, t\}$,

(6) the a-priori-distribution from R_t is a gamma distribution with coefficients a, b.

The result is only correct if all of these assumptions are fulfilled. We will discuss the appropriateness and relevance of those in the following sections.

The official reproduction number for Austria for each day is published on the website of the AGES[1]. It is accompanied by the two-sided 95%-confidence bound. Those values are plotted in Figure 3.1. We observe that the uncertainty increases if we have fewer cases, which can be seen in the beginning of the crises and since May. Moreover, this connection is a direct implication of the posterior distribution in (3.4).

In Figure 3.2, we see the comparison of the official results and the replication using formula (3.5) and the statistic of incident cases from [1]. The deviation might come from a change of the incident cases afterwards, which can occur if the symptom onset date emerges later, as stated in the explanations of the dashboard.[2] This slight difference can also be observed in the replication of the confidence bounds, which are not plotted.

These replicated results will be used in the following sections to compare the AGES approach with modifications of this model.

[1] https://www.ages.at/

[2] https://www.sozialministerium.at/Informationen-zum-Coronavirus/Dashboard.html, accessed on December 12, 2020.

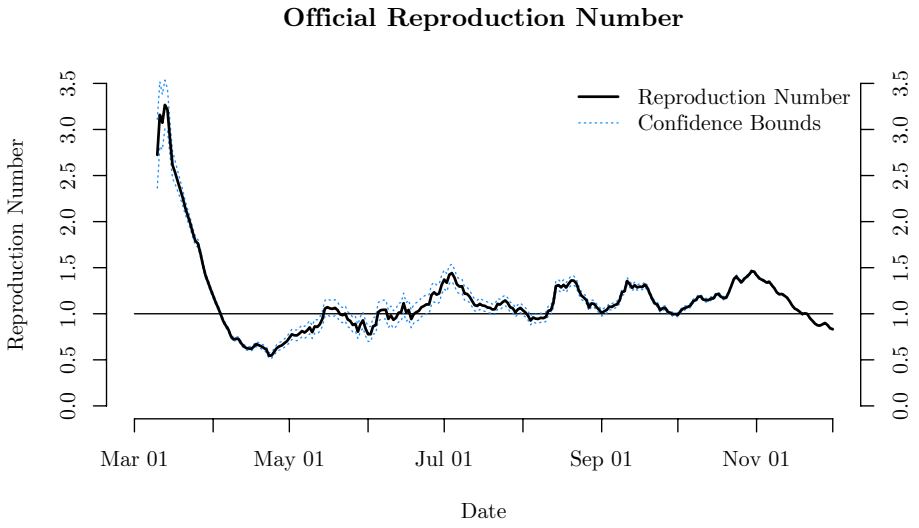

Figure 3.1: We see the official reproduction numbers published by the AGES together with the 95%-confidence bounds from March to November 2020.

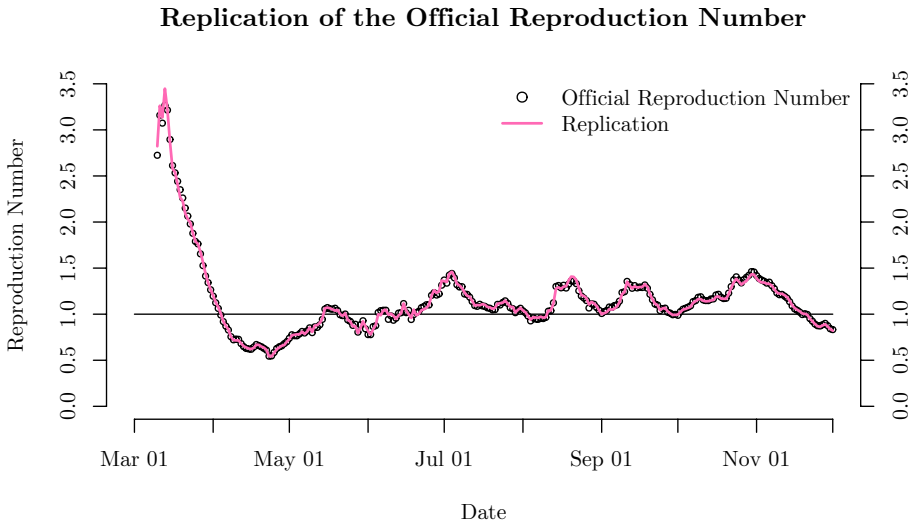

Figure 3.2: By comparing the official reproduction number and our replication, we see that we managed to replicate the published reproduction number with the methods described from March to November 2020.

3.3 Superspreaders

The aim of this section is to question assumption (1), which supposes a conditional Poisson distribution for the number of incident cases. The advantage of assuming that is that the Poisson distribution needs only one parameter, but therefore it is rather inflexible. If X is Poisson(λ) distributed, then we know that $\mathbb{E}[X] = \lambda$ and $\mathrm{Var}(X) = \lambda$. So, we see that we cannot choose mean and variance separately. If we assume that each individual of the previous generation infects other people independently, we see that the mean value as well as the variance for a single person is R_t. We now have to discuss whether R_t is also an appropriate value for the variance.

An important characteristic of the Covid-19 pandemic is the significance of so-called superspreaders. We therefore observe that a variance of around 1 is unlikely. Choosing a variance that is significantly too low, leads to confidence bounds of the estimation of the reproduction number that are too narrow. In Section 3.8, we will see what happens if we follow a deterministic approach, which means setting the variance to zero. We then can calculate the reproduction number directly, but we do not know anything about accuracy of the calculation. When using a Poisson distribution, we get confidence bounds, but the high accuracy, which is stated by the confidence bounds, is misleading.

To overcome this problem, we can assume that Y_t conditional on $\{Y_{t-1}, \dots, Y_{t-T}\}$ is negative binomial distributed. Then, we have the possibility to choose mean and variance separately. This distribution family might be more suitable, because the variance is likely to be significantly higher than the mean value. The probability mass function of a negative binomial distribution with parameters $(\alpha, p) \in \mathbb{R}_+ \times (0, 1)$ is given by

$$f(k) = \binom{k+\alpha-1}{k}(1-p)^\alpha p^k, \qquad k \in \mathbb{N}.$$

The parameter α is referred to as the dispersion parameter. If $X \sim \mathrm{NB}(\alpha_1, p)$ and $Y \sim \mathrm{NB}(\alpha_2, p)$, then $X+Y \sim \mathrm{NB}(\alpha_1+\alpha_2, p)$. Thus, we suppose that every individual of the previous generation infects other people on day $t \in \mathbb{N}^*$ according to a negative binomial distribution $\mathrm{NB}(\alpha, p_t)$ with mean value

$$\frac{\alpha p_t}{1-p_t} = R_t \qquad \Leftrightarrow \qquad p_t = \frac{R_t}{R_t+\alpha}.$$

Thereby, R_t is the mean of infected people by one case, which is the interpretation that we want. The size of the previous generation is, like in Section 3.2, given by

$$Z_t = \sum_{s=1}^{T} Y_{t-s} w_s, \qquad t \in \mathbb{N}^*.$$

If we assume that these individuals infect others independently, the sum of these negative binomial distribution is conditional on $\{Y_{t-1}, \dots, Y_{t-T}\}$ as well negative

binomial distributed, so

$$Y_t|\{Y_{t-1},\dots,Y_{t-T}\} \sim \mathrm{NB}(\alpha Z_t, p_t), \qquad t \in \mathbb{N}^*.$$

For a day $t \in \mathbb{N}^*$, we can now proceed with a Bayes estimation. We take the same a-priori distribution as in Section 3.2, namely

$$p(x) = \frac{x^{a-1}\mathrm{e}^{-x/b}}{b^a\Gamma(a)}, \qquad x > 0.$$

To make the results comparable, we set $\tau = 13$ and assume that R_t is constant in $\{t-\tau+1,\dots,t\}$. So, we get for the likelihood function

$$L(R_t) = \prod_{s=t-\tau+1}^{t} \binom{Y_s + \alpha Z_s - 1}{Y_s} \left(\frac{R_t}{R_t+\alpha}\right)^{\alpha Z_s} \left(\frac{\alpha}{R_t+\alpha}\right)^{Y_s}.$$

A density of the a-posteriori distribution is given by the product of the a-priori density and the likelihood function normed to an integral of 1. Combining all factors that are not driven by R_t, given that $\sum_{s=t-\tau+1}^{t} Z_s = z_t$ and $\sum_{s=t-\tau+1}^{t} Y_s = y_t$, we get

$$p(R_t; y_t, z_t) = (R_t+\alpha)^{-(\alpha z_t + y_t)} R_t^{a-1+\alpha z_t} \mathrm{e}^{-R_t/b} k(y_t, z_t, \alpha, a, b)$$

for a density of the a-posteriori distribution.

We calculated the mean, which will be our estimate for the reproduction number, and confidence bounds numerically in R. In Figure 3.3, we see the comparison between the Poisson based estimate from [23, p. 2] and the negative binomial based estimate described above. We can observe that the negative binomial distribution gives us a slightly higher estimate for the reproduction number than the Poisson distribution. However, there is a big difference when it comes to confidence bounds. The wider confidence bounds of the negative binomial estimate reflect the higher uncertainty of the reproduction number because of the higher variance of the number of new infections per infector. We observe that the confidence bounds are wider with low infection numbers. In particular, the upper bound is important because policymakers tend to be cautious, as described in [29, p. 9].

Let $X \sim \mathrm{NB}(\alpha, p_t)$ and let R_t be its mean, then

$$\mathrm{Var}(X) = \frac{\alpha p_t}{(1-p_t)^2} = R_t\left(\frac{R_t}{\alpha}+1\right),$$

which strictly dominates the variance of the Poisson model, which is R_t. So, a lower α leads to a higher variance. Therefore, the correct choice of α requires data about the variance of new infections generated by one case. Here, we want to show the results for different values of α. In Figure 3.4, we observe that a higher variance in

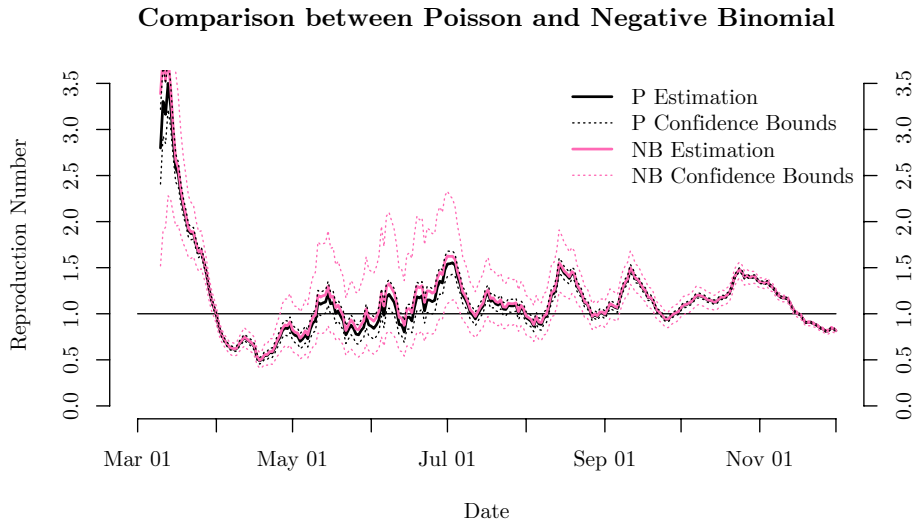

Figure 3.3: We see the comparison between the reproduction number using a Poisson approach (P) and a negative binomial approach (NB) with $\alpha = 0.1$ with the corresponding confidence bounds from March to November 2020.

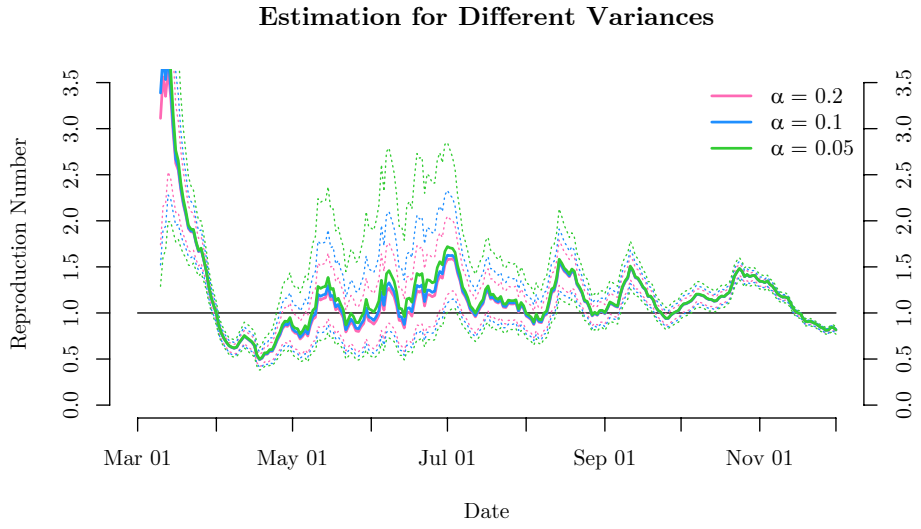

Figure 3.4: We see the reproduction number and corresponding confidence bounds for $\alpha = 0.2, 0.1, 0.05$ from March to November 2020. Under the negative binomial approach, the corresponding variance of new infections generated by one case for $R_t = 1$ is $6, 11, 21$ compared to a variance of 1 at the Poisson model.

the number of new infections by one case (a lower α), leads also to higher uncertainty in the estimation of the reproduction number and to a slightly higher estimate.

However, a high heterogeneity in infectiousness, and therefore a higher variance, is actually easier to control, because there are many people who do not infect anybody at all and most infections arise from superspreading events. This argument is also explained in [11, p. 2]. Nevertheless, they state that the majority of transmissions do not emerge from superspreading events, so Covid-19 has a low dispersion, which makes this pandemic hard to control. In [15, p. 4], they additionally find that the heterogeneity in infectiousness increases with the age. Thus, additional to the fact that young people are often infected asymptomatically, they are a significant risk factor of this pandemic. The authors recommend that young people should be screened more.

3.4 The Number of Incident Cases

In this section, we want to discuss the assumption (3) and parts of (2). Therefore, we first have to specify what the incident cases Y_t are exactly. We distinguish between two possibilities.

Firstly, we can define the incident cases as the number of new infections on a certain day. In this case, the weights (w_s) should be the generation time. Unfortunately, it is much more difficult to estimate the generation time instead of the serial interval, so we would have to use the serial interval as a proxy for the generation time. As already stated in Section 2.5, there are several reasons why this could be problematic. The two most obvious are that the serial interval can be negative, whereas the generation time is strictly positive, and the serial interval is likely to have a higher variance. Additionally, there are some biases in the estimation of the serial interval discussed in Section 2.3.

In practice, we have the problem that we would have a time delay, because usually infections can only be observed, when symptoms are developed. Even if we consider to assign an infection to a past day, it would be difficult to find out the exact day of the infection. Additionally, we usually can only find a small part of the significant number of asymptomatic cases. We will discuss this problem in Section 3.5.

Secondly, we can define the incident cases as the number of symptom onsets on a certain day. Then the serial interval would be the correct choice for the weights $(w_s)_{s\in S}$. However, we then have to extend S by some negative integers, as the serial interval can also be negative. Although including weights with a non-positive index is not a problem for formula (3.5), where we estimate R, we cannot define Y_t as conditional Poisson distributed anymore, because we would have to condition Y_t on itself and the future, which is not possible. Besides, we could only estimate R with a delay of $-\min(S)$ days, which will be problematic, when the pandemic has to be analysed urgently because of political decisions.

When it comes to practice, it is not difficult to extract the symptomatic out of all arising cases and date them back to the symptom onset date. However, as the incubation time is not deterministic, this would lead to additional uncertainty, as it

is also stated in [29, p. 9]. We can also assume that the significant majority of people with symptoms would make a test, so they can be observed. However, here we ignore asymptomatic cases completely, which will also be discussed in Section 3.5.

To conclude, both options have advantages and downturns and need assumptions, which do not necessarily hold. How much the approaches differ depends mainly on the number of tests and asymptomatic cases.

3.5 Unreported and Asymptomatic Cases and the Number of Tests

A significant problem when trying to measure and control pandemic dynamics are unreported and asymptomatic cases. We define an unreported case as an infected person with symptoms that is not recorded by the number of incident cases for any day. Unreported and asymptomatic cases, both contribute to y_t not being the correct number of incident cases and thus causing a conflict with assumption (2).

Whereas asymptomatic cases exist naturally, unreported cases are subject to human behaviour. As the nature of the non-mutated virus[3] is not expected to have changed over time, the assumption that the portion of asymptomatic cases remains constant is reasonable. For the portion of unreported cases, it is more difficult. One might assume, that it is likely to have decreased fast in the beginning of the pandemic as the awareness and fear increased. Afterwards it might have remained at a quite constant level, and maybe increased with the relaxation of the lockdown as the fear decreased.

We start with examining how the reproduction number is influenced by the number of undetected cases, which we define as all cases that are not recorded by the number of incident cases for any day. Therefore, let $p_1, p_2 \in [0, 1]$ be the portions of cases that are detected in two consecutive generations. The detected and undetected cases might have different reproduction numbers, because people who do not know that they are infected might act differently, probably less cautious. So, let R^a, R^b be the actual reproduction numbers of detected and undetected cases between those two generations respectively. If N_1, N_2 denote the number of infections in the corresponding generation, the following equation is implied:

$$N_2 = R^a N_1 p_1 + R^b N_1 (1 - p_1).$$

The observed reproduction number $\hat{R}$ is then given by

$$\hat{R} = \frac{N_2 p_2}{N_1 p_1} = \frac{(R^a p_1 + R^b (1 - p_1)) \cdot p_2}{p_1},$$

while the actual reproduction number R is calculated by

$$R = \frac{N_2}{N_1} = R^a p_1 + R^b (1 - p_1) = \hat{R} \cdot \frac{p_1}{p_2}.$$

[3] This variant of the virus was dominant in Austria till December 2020.

We can conclude that the reproduction number is not influenced by the portion of detected cases and is therefore calculated correctly, $R = \hat{R}$, if and only if the portion stays constant. That is also mentioned by [29, p. 9]. If the portion of detected cases increases, then $R < \hat{R}$, so the actual reproduction number is overestimated, and respectively for a decreasing portion. Furthermore, a different reproduction number of the two groups does not have an impact at all.

If we reconsider the thought from above, that the portion of detected cases might have increased in the beginning of the pandemic, this would imply that the reproduction number was overestimated in the beginning. Ex post, this raises the question if the success of the lockdown policies is overestimated as well. Maybe the rapid rise of incident cases in the beginning is also significantly driven by an increase in awareness and detection, and thus not as steep as it looks like. This problem is also addressed by [14, p. 11] and will be explained below.

An interesting role is played by the number of tests, which has increased on average, especially in the beginning. It is clear that more tests lead to a higher ratio of detected cases and that the number of tests needed for a further detection increase with the number of tests.

Upon these considerations, we try to build a model. Let $y_t \in \mathbb{N}$ be the true number of new infections at time t and $z_t \in \mathbb{N}$ the number of detected cases. For the undetected cases, we distinguish between those who have symptoms, but do not make a test, and those who are asymptomatic, $s_t \in \mathbb{N}$ and $a_t \in \mathbb{N}$ respectively. Obviously the following equation holds:

$$y_t = z_t + s_t + a_t. \tag{3.6}$$

We assume that the portion of asymptomatic cases $q \in [0, 1]$ remains constant over time and that s_t, a_t are linearly dependent on y_t. Moreover, s_t is dependent on the awareness of the population, so let $f(t) \in [0, 1]$ be the portion of unaware people at day t. By assuming that everybody who shows symptoms can get a test, it is clear that s_t does not depend on the number of tests. However, a_t does, because the more tests are conducted, the more asymptomatic cases can be found, for example in the social environment of a detected case. Therefore, let $g(x) \in [0, 1]$ be the portion of undetected asymptomatic from all asymptomatic cases for $x \in \mathbb{N}$ tests. Let b_t denote the number of tests at day t. That leads to

$$s_t = y_t(1 - q)f(t), \qquad a_t = y_t q g(b_t). \tag{3.7}$$

Inserting the formulas (3.7) in equation (3.6), we find the following formula to estimate the true number of infections from the observed number:

$$y_t = z_t \cdot \frac{1}{1 - (1 - q)f(t) - qg(b_t)}. \tag{3.8}$$

Now, we want to give an example for the functions f and g. For f, we therefore assume that there is a constant portion of people who will never test themselves, and

that the rest can be modelled by a exponential decline, due to more information, more news and therefore a higher awareness. For g, according to the comments above, we need a decreasing, convex function with values in $[0, 1]$. Thus, we define

$$f(t) = c_1 + c_2 e^{-c_3 t}, \qquad g(x) = \min\left(1, \frac{c_4}{x^{c_5}}\right), \tag{3.9}$$

with suitable constants $c_1, \ldots, c_5$, which fulfil $c_1, c_2, c_5 \in [0, 1)$, $c_3, c_4 \in \mathbb{R}_+$, $c_1 + c_2 \in [0, 1)$.

In Figure 3.5, we see the comparison between the calculated R from the observed numbers of infections and the result of R when we estimate the true number of infections using the functions above. We observe what we already stated before: In the beginning the reproduction number was overestimated[4]. Since April 2020, the difference is no longer visible.

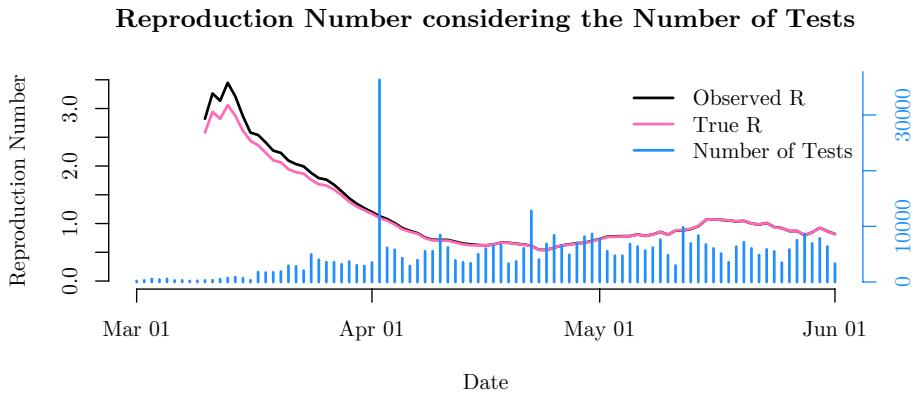

Figure 3.5: We see the number of tests on each day (right axis) from March to May 2020. We compare the official reproduction number to an estimate considering the number of tests (left axis). We used the functions defined in formula (3.9) with the following parameters: $q = 0.5, c_1 = 0.05, c_2 = 0.65, c_3 = 0.1, c_4 = 2.5, c_5 = 0.2$.

When using the number of tests to recalculate the incident cases on a daily basis, there are some problems, which will also be explained in Section 3.9. Firstly, some tests are registered in the EMS[5] a few days later, than they were conducted. This can be observed on April 02, 2020, where a lot of tests were added. Secondly, if it is possible, the incident cases are backdated to the date of symptom onset. We furthermore observe that there are less tests on Mondays.

In [14], it is stated that the growth rate of incident cases is overestimated, because of a simultaneous increase of tests. It is claimed that more tests lead to a higher portion of detected cases. Similar to our argument, this increase in detection leads

[4]The size of the overestimation depends on the choice of the functions and parameters.
[5]Epidemiologisches Meldesystem (epidemiological reporting system)

to overestimating the speed at which the pandemic spreads. Therefore, the number of incident cases is normalized by the number of tests, which has the underlying assumption that there is a direct proportionality between the number of tests and the number of detected infections. However, a problem with this assumption is that it would only hold if people are tested randomly. But that is not the case: Apart from the symptomatic cases, people at high risk of infection, for example in the social environment of a case, are tested preferentially. Thus, the probability of finding an infected person is decreasing with the number of tests and is not, as assumed, constant.

3.6 Imported Cases

A similar problem addressing assumption (2) arises from imported cases. When we want to talk about this issue, we first have to split our population into sub-populations. When looking at one sub-population, we can define an imported case as an individual of our sub-population that has been infected by an individual from another sub-population. If we assign a territory to each sub-population, we can distinguish whether this infection has taken place on the own territory or abroad.

In the setting of Section 3.2, an imported case at time t is recorded by y_t, but the infector of this individual is not by any $y_s, s \in \mathbb{N}^*$. This is inconsistent with our model, especially with equation (3.2), because there are cases that are generated out of nowhere.

So, if y_t denotes the total number of incident cases and p_t the number of imported cases on day t, then the modified reproduction number $R_t^{(p)}$ can be approximated by

$$R_t^{(p)} = R_t \cdot \frac{y_t - p_t}{y_t} = R_t \cdot \left(1 - \frac{p_t}{y_t}\right),$$

so, for $p_t > 0$, the reproduction number is smaller when we take imported cases into account. With respect to measures, it follows that if $\frac{p_t}{y_t}$ is small, social distancing is necessary, whereas if $\frac{p_t}{y_t}$ is high, travel restrictions are more important to reduce the spread.

However, if we do not adapt the model and ignore this mistake, this leads to an overestimation of the reproduction number, which is also described in [29, p. 3]. Alternatively, we could change the definition of the reproduction number from *average number of new infections that are directly infected by one case* to *ratio of the number cases in two consecutive generations*. This might even be the more interesting number.

As stated by [9, p. 6], especially in the beginning of the pandemic in our population, imported cases play a crucial role as they are a significant part of the incident cases. For this reason, the estimation of the reproduction number has a higher uncertainty in the beginning.

Of course, also exported cases play an important role in the global setting. However, they have no influence on the local reproduction number, no matter how it is defined.

3.7 Impact of the Weights

As the weights w_s have a significant impact on the estimation of the reproduction number, we focus in this section on the assumptions (3) and (4). In Section 2.7, we already pointed out the problems with the estimation of the serial interval in [22, p. 1], which is used in [23, p. 2]. This addresses both assumptions. Thus, we want to determine how the reproduction number is influenced by a change in the weights w_s and therefore by a possible estimation mistake. Subsequently, we will discuss the impact of weights with a negative index and look at the reproduction number resulting when using our estimation for the serial interval from Section 2.3.

We remember that w_s is the probability mass at s in the (discrete) distribution of the serial interval. For two different scenarios, let S^a, S^b describe the serial interval and R^a, R^b the corresponding reproduction numbers between two generations. From equation (3.1), assuming that the doubling time is fix, we get

$$\frac{\log(2)}{\log(R^a)}\,\mathbb{E}[S^a] = \frac{\log(2)}{\log(R^b)}\,\mathbb{E}[S^b],$$

which simplifies to

$$R^b = (R^a)^{\mathbb{E}[S^b]/\,\mathbb{E}[S^a]}.$$

This would imply that the reproduction number is dependent on the mean of the serial interval. In Figure 3.6, we see the comparison of the resulting reproduction number between a serial interval with the gamma distribution from [22, p. 1] and a gamma distribution with a different mean. Let the quotient of the means be $c := \mathbb{E}[S^b]/\,\mathbb{E}[S^a]$. We also see a third line, namely $(R^b)^{1/c}$, which is very close to the line of R^a. For those time points where we observed a fast changing doubling time, these lines do not coincide perfectly, as our assumption of a constant reproduction number between two generations does not hold.

The accuracy of the assumption of a constant doubling time is dependent on the variation of the rate of spread and the variance of the serial interval. For example, if we have a deterministic serial interval the assumption is perfectly correct. In Figure 3.7, the underlying serial intervals follow a gamma distribution, which all have the same mean but different variances. We observe that they only lead to a different reproduction number in the beginning of the pandemic, where the doubling time changed rapidly with a clear upward trend.

For the four different distributions for the serial interval in Section 2.7, the resulting reproduction number is plotted in Figure 3.8. The graphs with the serial interval from [22, p. 1] (gamma) and [18, p. 285] (log-normal) are nearly identical, as well as

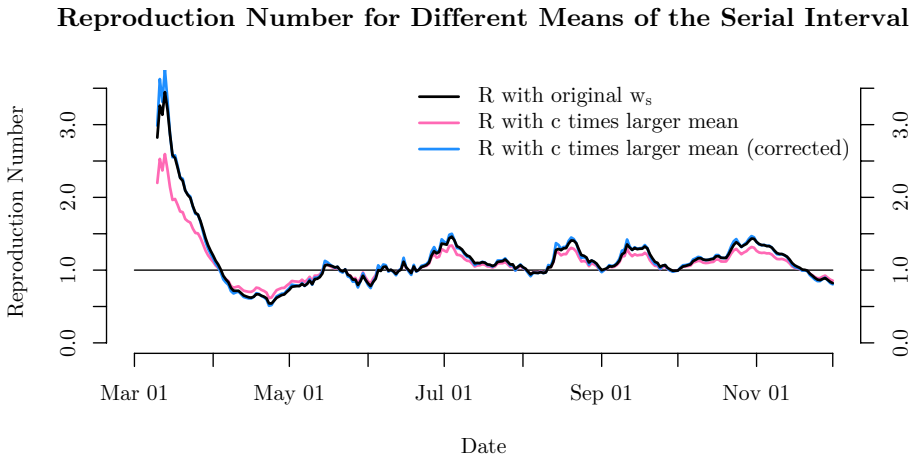

Figure 3.6: We see the original reproduction number from March to November 2020. Further, we see an estimate for the reproduction number using weights that come from a gamma distribution with a $c = 0.72$ times higher mean and a corrected version where we take the new reproduction number to the power of $\frac{1}{c}$.

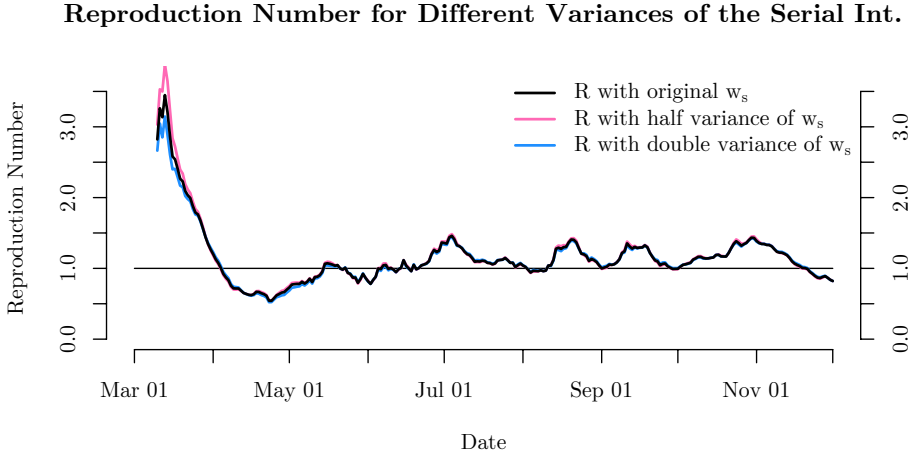

Figure 3.7: We see the original reproduction number from March to November 2020. Further, we see an estimate for the reproduction number using weights that come from a gamma distribution with half of the variance and the doubled variance.

our estimate (skew-normal) and [6, p. 1341] (normal). So, we see a clear difference between the distributions that include negative values and those only defined on $\mathbb{R}_+$. Possible explanations are the slightly lower mean (compare Figure 3.6) and the significantly higher variance (compare Figure 3.7). In the density plot, Figure 2.5, we observe that the normal and skew-normal distribution have considerably much

mass on $\mathbb{R}_-$, which is another possible explanation for the different shape of the reproduction number.

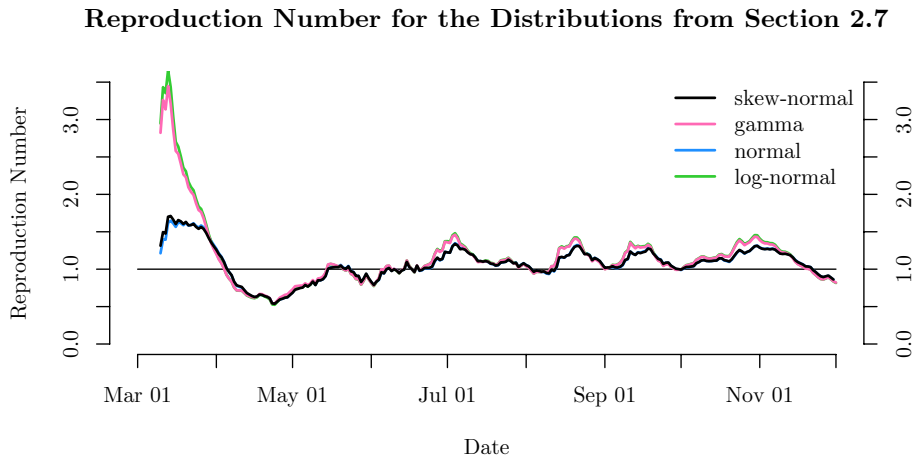

Figure 3.8: We see several estimations for the reproduction number based on the four different weighting schemes following the four distributions from Section 2.7 from March to November 2020.

However, we must not forget that the distribution of the serial interval is estimated itself and that the uncertainty from this estimation process adds to the uncertainty arising from the model for the reproduction number. In [29, p. 8], they estimate both numbers together to get an accurate quantification of the uncertainty, but that requires a lot of data that is hard to observe. Nevertheless, as the uncertainty coming from the serial interval decreases over time, the uncertainty for the reproduction number declines as well.

3.8 Reproduction Number for a Single Day

We continue with questioning assumption (5), which is necessary for the Bayes approach. If we do not consider the number of incident cases as a random variable, but follow a deterministic approach, equation (3.2) can be transformed so that the reproduction number can be calculated directly. It follows that

$$R_t = \frac{y_t}{\sum_{s=1}^{T} y_{t-s} w_s},$$

which also makes assumption (1) unnecessary.

We recognize that this formula is quite similar to equation (3.5) for $\tau = 1$. In order

to smooth those values, we can calculate a moving average using the last $n \in \mathbb{N}^*$ days:

$$R_t^{MA} = \sum_{i=1}^{n} R_{t-i+1} w_i.$$

But now we are free to set the weights as we want to, they do not have to be constant anymore. For example, we can assign linear weights:

$$w_i = \frac{n-i+1}{\frac{n(n+1)}{2}}, \qquad i \in \{1, \dots, n\}.$$

In Figure 3.9, we can see the results for the single-day reproduction number and a moving average with linear weights for the last 7 days in order to account for weekday specific differences, compared to the official reproduction number.

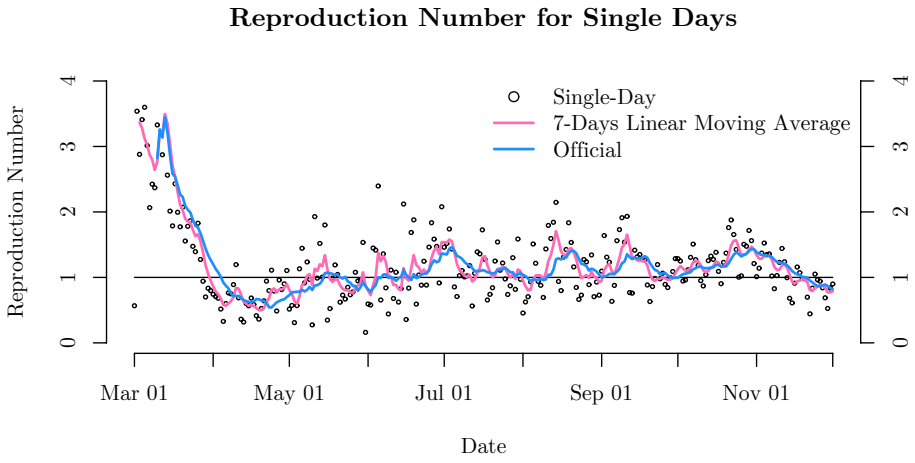

Figure 3.9: We see the official reproduction number in comparison with the single-day approach and a corresponding 7-days linear moving average.

Taking fewer days into account has got advantages as well as drawbacks. On the one hand, the curve is less smooth and more vulnerable to spikes or clusters. But, trends are visible earlier, on the other hand. This can help to react more quickly in case of a sudden increase.

3.9 Delay in Reporting

Another problem that gets in conflict with assumption (2) is the delay in reporting, which was especially significant in the beginning of the pandemic. So, y_t is not the number of symptom onsets on date t, but the number of reported cases on that

day. This implies that y_t is lagged. For example, if every case is reported five days after symptom onset, y_t gives the symptom onsets for $t - 5$. Therefore, also the reproduction number is lagged.

We take a look at the German data from the RKI ([24]). For every case, they have the reporting date (Meldedatum) and the reference date (Referenzdatum). The difference of these is the delay in reporting, which is a non-negative integer for those people who are tested because of symptoms. However, in some cases this difference is negative, for example if someone is tested positive before that person shows symptoms.

To analyse this delay, we start looking at its empirical distribution. We see the ratio of delays smaller or equal to 0 for the months March to July 2020. We observe that it is increasing every month, especially from April to May, which is an indicator for a constantly better handling of the pandemic. The kernel density estimate of those cases that have a positive delay is plotted in Figure 3.10, where we can also see that the delay is decreasing significantly from March to May and remains approximately constant since then.

	March	April	May	June	July
Ratio of Delay ≤ 0	0.18	0.25	0.41	0.45	0.46

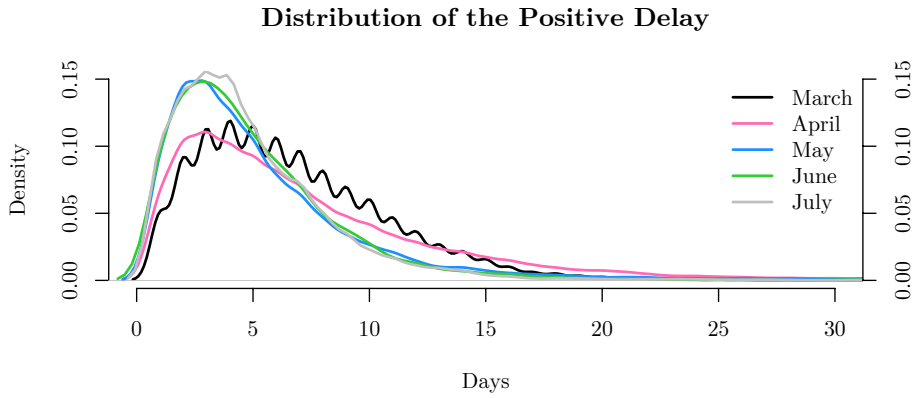

Figure 3.10: We see a kernel density estimate as approximation for the distribution of the delay given that the delay is positive for March to July 2020.

Focussing on the development of the delay, we see the daily and weekly averages in Figure 3.11. One can observe a strong positive correlation between the delay and the number of incident cases, which can be explained with limited test and organisational capacities in the beginning of the pandemic.

In [25, p. 2726-2737], it is stated that the delay in detection leads to more infections and that infectious people with a delay are the main driver of the pandemic spread. They conclude that it would be a big mistake to only test suspected cases.

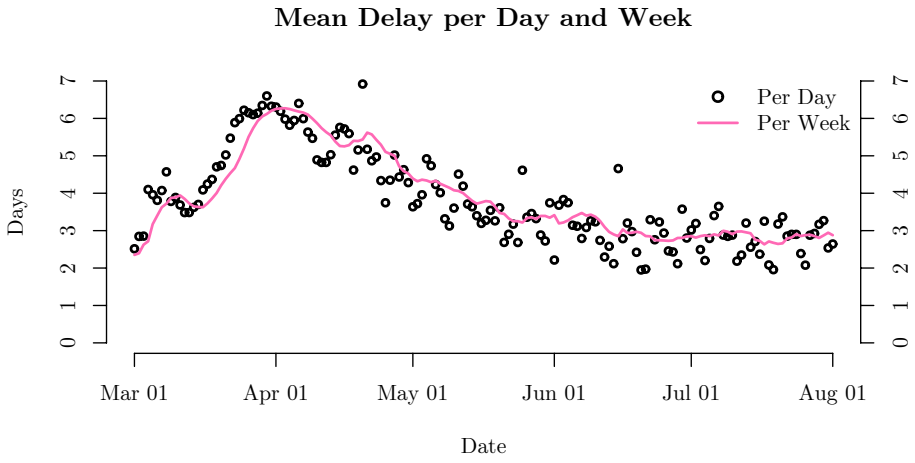

Figure 3.11: We see the mean of the delay per day and per week for March to July 2020.

Finally, when it comes to the estimation of the growth rate and the reproduction number, it turns out that the delay in reporting itself only leads to a time-lagged result, but does not imply a misestimation. However, if the delay changes over time, we observe a bias. That is an underestimation of the current trend (i.e. the correct reproduction number is further away from 1 than the estimated) if the delay decreases and an overestimation if the delay increases. If we obtain a more or less constant number of incident cases, this change of delay does not have an impact.

4 Galton–Watson Process

An important characteristic of the Covid-19 pandemic is the high variance of the number of infections that are directly infected by a certain case. Individuals that infect a high number of other people are so-called superspreaders and play a significant role in this pandemic as already explained in Section 3.3. To provide a mathematical model that describes the dynamic of the spread of a pandemic, we study the *Galton–Watson process*. After calculating some basic properties of this process, where we also proof general probabilistic theorems, we focus on the extinction probability. Therefore, we will not only proof the corresponding theorems, but also give a formula for an upper and lower bound and give examples for certain distributions that are interesting for us like the negative binomial. In Section 4.3, we extend the Galton–Watson process in a way that it can describe the Covid-19 pandemic. We finally get the formula that is used in [23] and in Section 3.2.

4.1 Basic Properties

Let Y_n be an $\mathbb{N}$-valued random variable that describes the number of cases in generation $n \in \mathbb{N}$. For $i \in \{1, 2, \dots, Y_{n-1}\}$, $n \in \mathbb{N}^*$, let X_i^n be an $\mathbb{N}$-valued random variable that denotes the offspring that is the number of infections generated by the i-th individual in the $(n-1)$-th generation. Starting with $Y_0 = y_0 \in \mathbb{N}^*$, we can recursively define

$$Y_n := \sum_{i=1}^{Y_{n-1}} X_i^n, \qquad n \in \mathbb{N}^*.$$

We assume that $(X_i^n)_{i,n\in\mathbb{N}^*}$ are independent and for every fixed $n \in \mathbb{N}^*$ that $(X_i^n)_{i\in\mathbb{N}^*}$ are identically distributed. We will refer to that as the *offspring distribution*. If the first moment of the offspring distribution exists for $n \in \mathbb{N}^*$, then we have $\mathbb{E}[X_i^n] = R_n < \infty$ for all $i \in \mathbb{N}^*$ with R_n denoting the deterministic reproduction number from generation $n-1$ to generation n according to the definition of the reproduction number discussed in Chapter 3. If also the second moment exists for $n \in \mathbb{N}^*$, we define $\sigma_n^2 := \mathrm{Var}(X_i^n) < \infty$ for all $i \in \mathbb{N}^*$. Since the distribution of Y_n is determined only by the last value Y_{n-1} and not dependent on the entire past, Y_n is a Markov process.

For an $\mathbb{N}$-valued random variable Z, we define the *probability generating function*

as

$$g_Z(t) = \mathbb{E}[t^Z] = \sum_{n=0}^{\infty} t^n \mathbb{P}(Z = n), \qquad t \in \mathbb{C},$$

which converges absolutely for $|t| \leq r \in \mathbb{R}_+$. The radius of convergence r is at least 1, because for $|t| \leq 1$ we have

$$\left| \sum_{n=0}^{\infty} t^n \mathbb{P}(Z = n) \right| \leq \sum_{n=0}^{\infty} |t^n \mathbb{P}(Z = n)| \leq \sum_{n=0}^{\infty} \mathbb{P}(Z = n) = 1. \tag{4.1}$$

As the probability generating function is a power series, it inherits all properties a the power series within the radius of convergence. So, it is continuous on $[0, r]$ by Abel's theorem and infinitely often continuously differentiable on $[0, r)$ by dominated convergence theorem. As the coefficients are all non-negative, it is also monotone increasing and convex. With monotone convergence, we have

$$\begin{aligned} g'_Z(1) &= \mathbb{E}[Z], \\ g''_Z(1) &= \mathbb{E}[Z(Z-1)], \end{aligned} \tag{4.2}$$

where $g'_Z(1)$ and $g''_Z(1)$ may be the left-sided limit in case g_Z is not differentiable at 1. If either the first or second moment does not exist, we have $g'_Z(1) = \mathbb{E}[Z] = \infty$ or $g''_Z(1) = \mathbb{E}[Z(Z-1)] = \infty$, respectively. If the first and second moment exist, we can express the variance of Z by

$$\operatorname{Var}(Z) = g''_Z(1) + g'_Z(1) - (g'_Z(1))^2. \tag{4.3}$$

Before we can do further calculations, we need an auxiliary theorem for random sums.

Theorem 4.1. *Let $(X_k)_{k\in\mathbb{N}^*}$ be a sequence of real-valued, independent and identically distributed random variables and let N be an $\mathbb{N}$-valued random variable that is independent from $(X_k)_{k\in\mathbb{N}^*}$. We define the random sum $S := \sum_{k=1}^{N} X_k$. Further, let g_S, g_N and g_X denote the probability generating function for S, N and $(X_k)_{k\in\mathbb{N}^*}$, respectively. For (b) and (c), we assume the existence of the first and second moment of N and $(X_k)_{k\in\mathbb{N}^*}$ respectively. Then we have*

(a) $g_S(t) = g_N(g_X(t))$ *for* $t \in \mathbb{C}, |t| \leq 1$,

(b) $\mathbb{E}[S] = \mathbb{E}[N]\,\mathbb{E}[X_1]$, (Wald's equation)

(c) $\operatorname{Var}(S) = \mathbb{E}[N]\operatorname{Var}(X_1) + \operatorname{Var}(N)(\mathbb{E}[X_1])^2$. (Blackwell-Girshick equation)

Proof. (a) Using the assumption that the random variables $(X_k)_{k\in\mathbb{N}^*}$ are independent and identically distributed, we have, for an $n \in \mathbb{N}$, that

$$\mathbb{E}[t^{\sum_{k=1}^{n} X_k}] = \prod_{k=1}^{n} \mathbb{E}[t^{X_k}] = (\mathbb{E}[t^{X_1}])^n = (g_X(t))^n,$$

for $|t| \leq 1$. Using the dominated convergence theorem and the independence of $\sum_{k=1}^{n} X_k$ from the event $\{N = n\}$, we get

$$g_S(t) = \mathbb{E}[t^{\sum_{k=1}^{N} X_k}] = \sum_{n=0}^{\infty} \mathbb{E}[t^{\sum_{k=1}^{n} X_k} 1_{\{N=n\}}] = \sum_{n=0}^{\infty} \mathbb{E}[t^{\sum_{k=1}^{n} X_k}]\, \mathbb{E}[1_{\{N=n\}}]$$
$$= \sum_{n=0}^{\infty} (g_X(t))^n \mathbb{P}(N = n) = g_N(g_X(t)),$$

for $|t| \leq 1$, which concludes this part.

(b) We derive the equation from (a), then we have

$$g_S'(t) = g_N'(g_X(t)) g_X'(t), \qquad |t| < 1.$$

We take the left-sided limit at 1 and since $g_X(1) = 1$, we have

$$g_S'(1) = g_N'(1) g_X'(1).$$

With equation (4.2), we have the desired result.

(c) We derive the equation from (a) a second time, which leads to

$$g_S''(t) = g_N'(g_X(t)) g_X''(t) + g_N''(g_X(t))(g_X'(t))^2, \qquad |t| < 1.$$

Again, we take the left-sided limit at 1 and since $g_X(1) = 1$, we have

$$g_S''(1) = g_N'(1) g_X''(1) + g_N''(1)(g_X'(1))^2.$$

With equation (4.2) and (4.3) and those for $g_S'(1)$ and $g_S''(1)$, we get

$$\begin{aligned}
\operatorname{Var}(S) &= g_S''(1) + g_S'(1) - (g_S'(1))^2 \\
&= g_N'(1) g_X''(1) + g_N''(1)(g_X'(1))^2 + g_N'(1) g_X'(1) - (g_N'(1) g_X'(1))^2 \\
&= g_N'(1)\big(g_X''(1) + g_X'(1) - (g_X'(1))^2\big) + \big(g_N'(1) + g_N''(1) - (g_N'(1))^2\big)(g_X'(1))^2 \\
&= \mathbb{E}[N] \operatorname{Var}(X_1) + \operatorname{Var}(N)(\mathbb{E}[X_1])^2,
\end{aligned}$$

which concludes the proof. □

Now, we can easily compute the expected value of the Galton–Watson process Y_n for $n \in \mathbb{N}^*$. Therefore, we assume that the first moment of the offspring distribution exists. With Theorem 4.1(b), we get

$$\mathbb{E}[Y_n] = \mathbb{E}[Y_{n-1}] R_n,$$

and with mathematical induction, it follows that

$$\mathbb{E}[Y_n] = y_0 \prod_{i=1}^{n} R_i. \tag{4.4}$$

In case of a constant reproduction number R, this simplifies to

$$\mathbb{E}[Y_n] = y_0 R^n. \tag{4.5}$$

We continue with looking at the variance of Y_n for $n \in \mathbb{N}^*$. Therefore, we assume that the first and second moment of the offspring distribution exist. Using equation (4.4) and Theorem 4.1(c), we get

$$\operatorname{Var}(Y_n) = \sigma_n^2 y_0 \prod_{i=1}^{n-1} R_i + R_n^2 \operatorname{Var}(Y_{n-1}).$$

If we have a recursively defined real-valued sequence $z_m = a_m + b_m z_{m-1}$ for $m \in \mathbb{N}^*$, we get $z_m = z_0 \prod_{k=1}^{m} b_k + \sum_{k=1}^{m} a_k \prod_{j=k+1}^{m} b_j$ by repeated substitution. With $\operatorname{Var}(Y_0) = 0$, we therefore get

$$\operatorname{Var}(Y_n) = y_0 \sum_{k=1}^{n} \sigma_k^2 \Bigg(\prod_{i=1}^{k-1} R_i \Bigg) \Bigg(\prod_{j=k+1}^{n} R_j^2 \Bigg).$$

If there exists a $k \in \{1, \dots, n\}$ such that $R_k = 0$, then we have $\mathbb{E}[Y_n] = 0$ and $\operatorname{Var}(Y_n) = 0$. If $R_k > 0$ for all $k \in \{1, \dots, n\}$, then we can simplify the formula for the variance further to

$$\operatorname{Var}(Y_n) = \mathbb{E}[Y_n] \sum_{k=1}^{n} \frac{\sigma_k^2}{R_k} \prod_{j=k+1}^{n} R_j.$$

For a constant reproduction number $R_k = R$ for all $k \in \{1, \dots, n\}$, this simplifies to

$$\operatorname{Var}(Y_n) = y_0 R^{n-1} \sum_{k=1}^{n} \sigma_k^2 R^{n-1-k}.$$

If we additionally assume that $\sigma_k^2 = \sigma^2$ for all $k \in \{1, \dots, n\}$, we get

$$\operatorname{Var}(Y_n) = \begin{cases} y_0 \sigma^2 R^{n-1} \frac{R^n - 1}{R - 1}, & \text{if } R \neq 1, \\ y_0 \sigma^2 n, & \text{if } R = 1. \end{cases}$$

We now want to look at the covariance. With the existence of the second moment and Cauchy–Schwarz equation, we know that $\mathbb{E}[Y_m Y_n]$ exists for $m, n \in \mathbb{N}, m < n$. As $(Y_n)_{n \in \mathbb{N}}$ is a Markov process, we know that Y_m is Y_{n-1}-measurable. With a similar argument like in the proof of Theorem 4.1(a), we have

$$\mathbb{E}[Y_m Y_n | Y_{n-1}] = Y_m \, \mathbb{E}\left[\sum_{i=1}^{Y_{n-1}} X_i^n \middle| Y_{n-1} \right] = Y_m Y_{n-1} R_n.$$

With the defining property of the conditional expectation, it follows that

$$\mathbb{E}[Y_m Y_n] = \mathbb{E}[Y_m Y_{n-1}] R_n,$$

which we iterate till we have

$$\mathbb{E}[Y_m Y_n] = \mathbb{E}[Y_m^2] \prod_{k=m+1}^{n} R_k.$$

Using this equation and the recursive formula for the expected value, we can compute the covariance by

$$\begin{aligned}\operatorname{Cov}(Y_m, Y_n) &= \mathbb{E}[Y_m Y_n] - \mathbb{E}[Y_m]\,\mathbb{E}[Y_n] = \mathbb{E}[Y_m^2] \prod_{k=m+1}^{n} R_k - (\mathbb{E}[Y_m])^2 \prod_{k=m+1}^{n} R_k \\ &= \operatorname{Var}(Y_m) \prod_{k=m+1}^{n} R_k.\end{aligned}$$

For an $n \in \mathbb{N}^*$ and each $m \in \{1, \dots, n\}$, let g_{Y_n} and g_{X^m} denote the probability generating functions for Y_n and $(X_k^m)_{k\in\mathbb{N}^*}$, respectively. Inductively applying Theorem 4.1(a) and using $g_{Y_0}(t) = t^{y_0}$, shows that

$$g_{Y_n} = g_{Y_0} \circ g_{X^1} \circ g_{X^2} \circ \cdots \circ g_{X^n} = (g_{X^1} \circ g_{X^2} \circ \cdots \circ g_{X^n})^{y_0}. \tag{4.6}$$

We know that a probability generating function uniquely defines a random variable. We use this for the following theorem.

Theorem 4.2. *The sum of two independent Galton–Watson processes $\hat{Y}$ and $\tilde{Y}$ with the same family of offspring distributions $(X_k^m)_{k,m\in\mathbb{N}^*}$ and starting values $\hat{y}_0, \tilde{y}_0 \in \mathbb{N}$ is again a Galton–Watson process Y with this offspring distribution and starting value $y_0 = \hat{y}_0 + \tilde{y}_0$. Also, a Galton–Watson process Z with starting value z_0 can be split into a sum of z_0 independent Galton–Watson processes $(Z^{(i)})_{i\in\{1,\dots,z_0\}}$ with the same offspring distribution and starting value 1.*

Proof. For arbitrary $m, n \in \mathbb{N}^*$, let $g_{\hat{Y}_n}$, $g_{\tilde{Y}_n}$, g_{Y_n} and g_{X^m} denote the probability generating functions of $\hat{Y}_n$, $\tilde{Y}_n$, Y_n and $(X_k^m)_{k\in\mathbb{N}^*}$ respectively. Using the independence of $\hat{Y}_n$ and $\tilde{Y}_n$ and equation (4.6), we have

$$\begin{aligned}g_{Y_n}(t) &= \mathbb{E}[t^{Y_n}] = \mathbb{E}[t^{\hat{Y}_n + \tilde{Y}_n}] = \mathbb{E}[t^{\hat{Y}_n}]\,\mathbb{E}[t^{\tilde{Y}_n}] \\ &= g_{\hat{Y}_n}(t) g_{\tilde{Y}_n}(t) = ((g_{X^1} \circ g_{X^2} \circ \cdots \circ g_{X^n})(t))^{\hat{y}_0 + \tilde{y}_0}.\end{aligned}$$

Thus, Y is a Galton–Watson process with the same family of offspring distributions and starting value $y_0 = \hat{y}_0 + \tilde{y}_0$, because the probability generating function of a random variable uniquely defines its distribution.

With equation (4.6), for an arbitrary $n \in \mathbb{N}^*$, we have

$$g_{Z_n}(t) = ((g_{X^1} \circ g_{X^2} \circ \cdots \circ g_{X^n})(t))^{z_0} = \prod_{i=1}^{z_0} (g_{X^1} \circ g_{X^2} \circ \cdots \circ g_{X^n})(t) = \prod_{i=1}^{z_0} g_{Z_n^{(i)}}(t),$$

where $(Z^{(i)})_{i\in\{1,\dots,z_0\}}$ is a family of independent Galton–Watson processes with the same offspring distributions and starting value 1. Since a probability generating function of a random variable uniquely defines its distribution, it follows that $Z = \sum_{i=1}^{z_0} Z^{(i)}$. □

In the following theorem, we see that a normed Galton–Watson process is a martingale.

Theorem 4.3. *Let* $I := \{n \in \mathbb{N} \mid \forall k \in \{1,\dots,n\} : R_k \in \mathbb{R}_+\}$ *be an index set. Then, the stochastic process*

$$M_n := Y_n \prod_{k=1}^{n} R_k^{-1}, \qquad n \in I,$$

is a martingale with respect to the generated filtration $(\mathcal{F}_n)_{n\in I}$.

Proof. $(M_n)_{n\in I}$ is trivially adapted to the generated filtration. Let $n \in I$, then

$$\mathbb{E}[|M_n|] = \mathbb{E}[Y_n] \prod_{k=1}^{n} R_k^{-1} = y_0 < \infty$$

shows the integrability. Further, since $(Y_n)_{n\in I}$ is a Markov process, we have

$$\mathbb{E}[Y_n|\mathcal{F}_{n-1}] = \mathbb{E}[Y_n|Y_{n-1}] \qquad n \in I, n \geq 1. \tag{4.7}$$

For each $k \in \mathbb{N}$, with $\mathbb{P}(Y_{n-1} = k) > 0$, we have

$$\mathbb{E}[Y_n|Y_{n-1} = k] = \mathbb{E}\left[\sum_{i=1}^{k} X_i^n \middle| Y_{n-1} = k\right] = \sum_{i=1}^{k} \mathbb{E}[X_i^n|Y_{n-1} = k] = k\,\mathbb{E}[X_1^n], \tag{4.8}$$

where the last equation holds, because the random variables $(X_k)_{k\in\mathbb{N}^*}$ are independent and identically distributed and independent from the event $\{Y_{n-1} = k\}$. It follows that

$$\mathbb{E}[Y_n|Y_{n-1}] = Y_{n-1}\,\mathbb{E}[X_1^n] = Y_{n-1}R_n. \tag{4.9}$$

Combining the equations (4.7) and (4.9), and multiplying with $\prod_{k=1}^{n} R_k^{-1}$, leads to

$$\mathbb{E}[M_n|\mathcal{F}_{n-1}] = M_{n-1},$$

which proves the martingale property. □

For $I = \mathbb{N}$, as $(M_n)_{n\in\mathbb{N}}$ is a non-negative martingale, we know with Doob's almost sure convergence theorem that there exists an $\mathcal{F}_\infty$-measurable random variable M_∞ such that $M_\infty = \lim_{n\to\infty} M_n$ almost surely, compare [32, p. 109]. With Fatou's lemma, we get

$$\mathbb{E}[M_\infty] = \mathbb{E}\left[\lim_{n\to\infty} M_n\right] \leq \lim_{n\to\infty} \mathbb{E}[M_n] = \mathbb{E}[M_0] = y_0.$$

We have equality if and only if the martingale is uniformly integrable, according to the Vitali convergence theorem, compare [32, p. 131-132].

If we now assume that $R_k \geq 1$ for all $k \in \mathbb{N}^*$, then $(Y_n)_{n\in\mathbb{N}}$ is a submartingale, because, for $n \in \mathbb{N}^*$, we get

$$\mathbb{E}[Y_n|\mathcal{F}_{n-1}] = Y_{n-1}\,\mathbb{E}[X_1^n|\mathcal{F}_{n-1}] = Y_{n-1}\,\mathbb{E}[X_1^n] = Y_{n-1}R_n \geq Y_{n-1}.$$

Similarly, if $R_k \leq 1$ for all $k \in \mathbb{N}^*$, then $(Y_n)_{n\in\mathbb{N}}$ is a supermartingale.

4.2 Extinction Probability

The Galton–Watson process was originally intended to compute the extinction probability of family names in the 19th century. We can transfer this result to compute the probability that a pandemic vanishes. Throughout this section, we assume that $(X_i^n)_{i,n\in\mathbb{N}^*}$ are independent and identically distributed, and call the corresponding Galton–Watson process *time-homogeneous*. Note that this is a stricter definition than in Section 4.1. So, for each $i, n \in \mathbb{N}^*$, we can set $p_k := \mathbb{P}(X_i^n = k)$ for $k \in \mathbb{N}$ and $R := \mathbb{E}[X_i^n] \geq 0$ for the offspring. Note that it is not necessary that the first moment exists, so $R = \infty$ is possible. This section is inspired by [2].

The following theorem helps us to find the extinction probability, when we know the distribution of the offspring.

Theorem 4.4. *Let $(Y_n)_{n\in\mathbb{N}}$ be a time-homogeneous Galton–Watson process with $y_0 = 1$, and $(X_i^n)_{i,n\in\mathbb{N}^*}$ its offspring. Then, the extinction probability q is the smallest fixed point of the probability generating function g_X of X_i^n in $[0,1]$ and*

$$\begin{aligned} q &= 0, && \text{if } p_0 = 0,\\ q &\in (0,1), && \text{if } p_0 > 0 \text{ and } R > 1,\\ q &= 1, && \text{if } p_0 > 0 \text{ and } R \leq 1. \end{aligned}$$

Proof. For every $\mathbb{N}$-valued random variable Z, the probability generating function g_Z exists, is continuous in $[0,1]$, compare equation (4.1), and has at least one fixed point at $1 = \mathbb{E}[1^Z] = g_Z(1)$. So, there is a smallest fixed point. Besides, we notice that $g_Z(0) = \mathbb{P}(Z = 0)$ and that g_Z is strictly increasing as long as $\mathbb{P}(Z = 0) < 1$.

For each $n \in \mathbb{N}$, let g_{Y_n} be the probability generating function of Y_n. We define the extinction probability q as

$$\mathbb{P}\Big(\bigcup_{k\in\mathbb{N}}\{Y_k = 0\}\Big).$$

If the process is extinct at time $n \in \mathbb{N}$, then it is certain that it is extinct at time $n+1$, so we know that $\{Y_n = 0\} \subset \{Y_{n+1} = 0\}$ for each n. As the probability measure $\mathbb{P}$ is continuous from below, it follows that

$$q = \mathbb{P}\Big(\bigcup_{k\in\mathbb{N}}\{Y_k = 0\}\Big) = \lim_{n\to\infty}\mathbb{P}\Big(\bigcup_{k=1}^{n}\{Y_k = 0\}\Big) = \lim_{n\to\infty}\mathbb{P}(Y_n = 0) = \lim_{n\to\infty} g_{Y_n}(0).$$

As the sequence $(g_{Y_n}(0))_{n\in\mathbb{N}}$ is growing and limited, and since $\mathbb{R}$ is complete, we know that the limit exists. For $n \in \mathbb{N}$, let g_{Y_n} denote the probability generating function of Y_n. From Theorem 4.1(a), we know that $g_{Y_{n+1}} = g_{Y_n} \circ g_X$. Since $y_0 = 1$ and the offspring distribution is the same for every generation, we see with equation (4.6) that also $g_{Y_{n+1}} = g_X \circ g_{Y_n}$ is true. As g_{Y_n} is continuous, it follows that

$$g_X(q) = g_X\left(\lim_{n\to\infty} g_{Y_n}(0)\right) = \lim_{n\to\infty} g_X(g_{Y_n}(0)) = \lim_{n\to\infty} g_{Y_{n+1}}(0) = q,$$

so q is a fixed point. For an arbitrary fixed point $a \in [0,1]$, we have

$$a = g_X(a) = g_{Y_n}(a) \geq g_{Y_n}(0),$$

because g_{Y_n} is monotone increasing. Taking the limit on the right-hand side for $n \to \infty$, we get $a \geq q$, so q is the smallest fixed point.

If $p_0 = g_X(0) = 0$, the smallest fixed point is $q = 0$, because it is impossible to get extinct when every individual has at least one offspring.

Assume from now on that $p_0 > 0$. If $p_0 + p_1 = 1$, we have $R \leq 1$ and we know that $g_X(t) = p_0 + tp_1$, so g_X is an affine function. Since $p_0 > 0$, we know that g_X has exactly one fixed point, so $q = 1$.

So, we can assume from now on that $p_0 + p_1 < 1$. Since we then have $g_X''(t) = \mathbb{E}[X(X-1)t^{X-2}] > 0$ for $t \in (0,1)$, it follows that g_X' is strictly increasing and g_X is strictly convex on $[0,1]$. If g_X had three fixed points $t_1 < t_2 < t_3$ in $[0,1]$, then for $\lambda = \frac{t_2 - t_1}{t_3 - t_1}$, we would have

$$g_X(\lambda t_3 + (1-\lambda)t_1) = g_X(t_2) = t_2 = \lambda t_3 + (1-\lambda)t_1 = \lambda g_X(t_3) + (1-\lambda)g_X(t_1),$$

which is not possible for a strictly convex function. Hence, g_X has at most two fixed points. We know that $g_X(0) > 0$, $g_X(1) = 1$ and $g_X'(1^-) = R$ (which denotes the left-hand limit). If $R > 1$, there exists $s \in (0,1)$ with $g_X(s) < s$. So, by the intermediate value theorem and because $g_X(0) = p_0 > 0$, there is a second fixed point in $(p_0, 1)$ and $q < 1$. Otherwise, if $R \leq 1$, then $g_X(t) > t$ for $t \in (0,1)$, so $q = 1$. □

For $y_0 > 1$, we see that every starting individual reproduces independently, because we can split this process into y_0 independent processes, according to Theorem 4.2. Therefore, the extinction probability with starting value y_0 is the probability that each of the y_0 families get extinct, so it is q^{y_0}.

To continue, we can set bounds for q with the following theorem, which is an exercise in [2, p. 12].

Theorem 4.5. *Let q be the extinction probability of a time-homogeneous Galton–Watson process $(Y_n)_{n\in\mathbb{N}}$ with offspring $(X_i^n)_{i,n\in\mathbb{N}^*}$ and $y_0 = 1$. Assume that $q < 1$ and $p_1 \neq 1$ (otherwise q is trivial), then we get the following bounds*

$$b_1 := \frac{p_0}{1-p_1} \leq q \leq \frac{p_0}{1-p_0-p_1} =: b_2.$$

Proof. Let g_X denote the probability generating function of the offspring distribution. With the formula for finite geometric sums and the definition of the probability generating function, we get

$$\begin{aligned}\sum_{n=0}^{\infty}\mathbb{P}(X_1^1 > n)s^n &= \sum_{n=0}^{\infty}\sum_{k=n+1}^{\infty} p_k s^n = \sum_{k=1}^{\infty}\sum_{n=0}^{k-1} p_k s^n = \sum_{k=1}^{\infty} p_k \frac{1-s^k}{1-s} \\ &= \frac{1}{1-s}\left(\sum_{k=1}^{\infty} p_k - \sum_{k=1}^{\infty} p_k s^k\right) = \frac{1}{1-s}\big(1-p_0-(g_X(s)-p_0)\big) \\ &= \frac{1-g_X(s)}{1-s},\end{aligned}$$

for $|s| < 1$. Note that the sum is absolute convergent and we can therefore change the order of summation. From Theorem 4.4, we know that $g_X(q) = q$, so we have

$$\sum_{n=0}^{\infty}\mathbb{P}(X_1^1 > n)q^n = 1. \tag{4.10}$$

By just considering the first two summands, it follows that

$$(1-p_0) + (1-p_0-p_1)q \leq 1,$$

because all summands are non-negative. As $q < 1$ and $p_1 \neq 1$ guarantee that $1-p_0-p_1 > 0$, we get

$$q \leq \frac{p_0}{1-p_0-p_1},$$

which proves the upper bound. Using equation (4.10) and $\mathbb{P}(X_1^1 > n) \leq \mathbb{P}(X_1^1 > 1)$ for all $n \in \mathbb{N}^*$, it follows that

$$\begin{aligned}1 &= \sum_{n=0}^{\infty}\mathbb{P}(X_1^1 > n)q^n \leq (1-p_0) + \sum_{n=1}^{\infty}\mathbb{P}(X_1^1 > 1)q^n \\ &= (1-p_0) + (1-p_0-p_1)\sum_{n=1}^{\infty} q^n = (1-p_0) + \frac{(1-p_0-p_1)q}{1-q}.\end{aligned}$$

By multiplying with $(1-q)$ and simplifying, we get

$$1-q \leq 1-p_0-p_1 q \qquad \Leftrightarrow \qquad q \geq \frac{p_0}{1-p_1},$$

which finishes the proof. □

We want to calculate this extinction probability explicitly for the case that the offspring has a geometric distribution and numerically for a negative binomial distribution in the following two examples.

Example 4.6. We parametrize the geometric distribution by its mean $R \in \mathbb{R}_+$. Then, the probability mass function is given by

$$p_k = \frac{1}{R+1}\left(\frac{R}{R+1}\right)^k, \qquad k \in \mathbb{N}.$$

At first, we calculate the probability generating function for $|t| < 1 + \frac{1}{R}$ by

$$\begin{aligned} g_X(t) &= \sum_{k=0}^{\infty} t^k \frac{1}{R+1}\left(\frac{R}{R+1}\right)^k = \frac{1}{R+1}\sum_{k=0}^{\infty}\left(\frac{tR}{R+1}\right)^k \\ &= \frac{1}{R+1}\cdot\frac{1}{1-\frac{tR}{R+1}} = \frac{1}{R+1-tR}. \end{aligned} \tag{4.11}$$

Then, we want to compute the fixed points and therefore look at $g_X(t) = t$, which leads to the quadratic equation

$$\begin{aligned} &\frac{1}{R+1-tR} = t \\ \Leftrightarrow \quad & t^2R + t(-R-1) + 1 = 0. \end{aligned} \tag{4.12}$$

That gives us the fixed points $t_1 = 1$ and $t_2 = \frac{1}{R}$, so the extinction probability is $q = \min(1, \frac{1}{R})$ according to Theorem 4.4. For the bounds b_1, b_2 from Theorem 4.5, we get

$$b_1 = \frac{\frac{1}{R+1}}{1-\frac{R}{(R+1)^2}} = \frac{R+1}{(R^2+2R+1)-R} = \frac{R+1}{R^2+R+1} < q,$$

$$b_2 = \frac{\frac{1}{R+1}}{1-\frac{1}{R+1}-\frac{R}{(R+1)^2}} = \frac{R+1}{(R^2+2R+1)-(R+1)-R} = \frac{R+1}{R^2} > q.$$

Example 4.7. As explained in Section 3.3, the negative binomial distribution for the offspring is especially interesting in this case. For $\alpha > 0$ and $p \in (0,1)$, its probability mass function is given by

$$p_k = \binom{k+\alpha-1}{k}(1-p)^\alpha p^k, \qquad k \in \mathbb{N}.$$

For the negative binomial distribution, the probability generating function for $t \in [0, \frac{1}{p})$ is

$$\begin{aligned} g_X(t) &= \sum_{k=0}^{\infty} t^k \binom{k+\alpha-1}{k}(1-p)^\alpha p^k \\ &= \frac{(1-p)^\alpha}{(1-tp)^\alpha}\sum_{k=0}^{\infty}\binom{k+\alpha-1}{k}(1-tp)^\alpha (tp)^k \\ &= \left(\frac{1-p}{1-tp}\right)^\alpha, \end{aligned}$$

because the last sum is over another probability mass function and therefore equals 1. In general, solving the equation

$$\left(\frac{1-p}{1-tp}\right)^{\alpha} = t$$

is only possible numerically. However, for $\alpha \in \{\frac{1}{2}, 1, 2\}$, we can calculate the extinction probability algebraically. For $\alpha = \frac{1}{2}$, we get the equation

$$t^3 p - t^2 + (1-p) = 0.$$

Since we know that $t_1 = 1$ is a solution, we can factorize the left side by polynomial division, so we get

$$(t-1)\big(t^2 p - t(1-p) - (1-p)\big) = 0.$$

Thus, the remaining solutions are

$$t_{2,3} = \frac{1-p \pm \sqrt{1+2p-3p^2}}{2p},$$

where $t_2 > 0$ and $t_3 < 0$ for each $p \in (0,1)$. Note that $t_2 < 1$ if and only if $p > \frac{2}{3}$, so

$$q = \begin{cases} 1, & \text{if } p \leq \frac{2}{3}. \\ \frac{1-p+\sqrt{1+2p-3p^2}}{2p}, & \text{if } p > \frac{2}{3}, \end{cases}$$

For $\alpha = 1$, this is the geometric distribution from Example 4.6. For $\alpha = 2$, we get with similar considerations that

$$q = \begin{cases} 1, & \text{if } p \leq \frac{1}{3}. \\ \frac{2-p-\sqrt{4p-3p^2}}{2p}, & \text{if } p > \frac{1}{3}, \end{cases}$$

For all $\alpha > 0$, we can calculate the bounds from Theorem 4.5, so

$$b_1 = \frac{(1-p)^{\alpha}}{1-\alpha(1-p)^{\alpha}p}, \qquad b_2 = \frac{(1-p)^{\alpha}}{1-(1-p)^{\alpha}-\alpha(1-p)^{\alpha}p}.$$

In Figure 4.1, we see the extinction probability for a negative binomial distribution plotted in a heatmap depending on mean and variance. We observe that it decreases for increasing mean and decreasing variance. For the numerical computation of q, we used the iteration from Theorem 4.4, where we build the sequence $(q_n)_{n\in\mathbb{N}}$ with $q_0 = 0$ and $q_{n+1} = g_X(q_n)$ for $n \in \mathbb{N}$, and iterate it until $|g_X(q_n) - q_n| < 10^{-7}$. Note that a classic bisection does not work for all parameter pairs, because we might have $g_X(t) \geq t$ for all $t \in [0, \frac{1}{p})$. This is the case when $R = g_X'(1) = 1 \Leftrightarrow \alpha = \frac{1-p}{p}$.

We continue with the extinction–explosion principle.

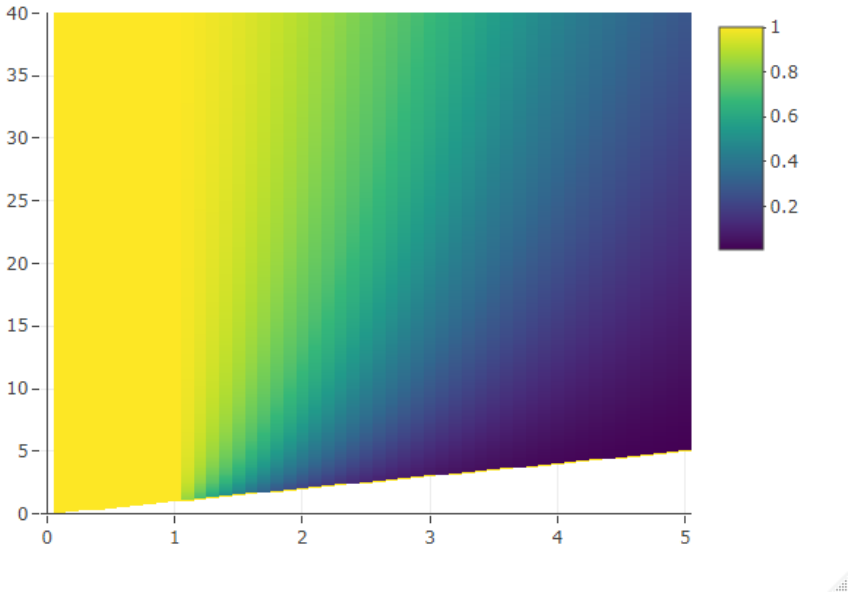

Figure 4.1: We see a heatmap showing the extinction probability for a negative binomial distribution depending on the mean (x-axis) and on the variance (y-axis).

Theorem 4.8 (Extinction–explosion principle)**.** *Let $(Y_n)_{n\in\mathbb{N}}$ be a Galton–Watson process and let $(X_i^n)_{i,n\in\mathbb{N}^*}$ be its offspring. Additionally, assume that $p_1 \neq 1$. Then, we have almost surely extinction or explosion, so*

$$\mathbb{P}\left(\bigcup_{k\in\mathbb{N}}\{Y_k = 0\}\right) + \mathbb{P}\left(Y_n \to \infty\right) = 1.$$

Proof. As the transition probabilities only depend on the current state and not on the entire history, the Galton–Watson process is a homogeneous Markov chain. Therefore, we consider this problem in the framework of Markov chains. We observe that 0 is an absorbing state and show that every $k \in \mathbb{N}^*$ is a transient state. We fix a $k \in \mathbb{N}^*$ and assume that $y_0 = k$, so it follows that

$$\bigcap_{j=1}^{k}\{X_j^1 = 0\} \subset \bigcap_{n\in\mathbb{N}^*}\{Y_n \neq k\}.$$

However, for $p_0 = 0$, the left side is just the empty set. We see that for this case, we have additionally

$$\left(\bigcap_{j=1}^{k}\{X_j^1 = 1\}\right)^{\mathrm{C}} \subset \bigcap_{n\in\mathbb{N}^*}\{Y_n \neq k\}.$$

Combined, we know that

$$\mathbb{P}\Big(\bigcap_{n\in\mathbb{N}^*}\{Y_n \neq k\}\Big) \geq \left\{ \begin{array}{ll} p_0^k, & \text{if } p_0 > 0 \\ 1-p_1^k, & \text{if } p_0 = 0 \end{array} \right\} > 0,$$

which proves that if the process starts in state k there is a positive chance that we will never return. So, we have proven the transience of state k. Thus, the process either gets absorbed by the state 0 or goes to infinity. □

Figure 4.2 illustrates Theorem 4.4 and 4.8. It shows 30 Monte Carlo simulations for different offspring means. The population starts with $y_0 = 100$ and we consider 50 generations. On the basis of Section 3.3, we chose a negative binomial distribution with $\alpha = 0.1$ and mean value R for the offspring. At the top, we see that $R = 0.9$ implies extinction as it was stated by Theorem 4.4. In the middle, for $R = 1$, we have also almost surely extinction, but not as fast as in the first plot. At the bottom with $R = 1.1$, we observe a positive chance of survival according to Theorem 4.4 and the explosion for those that survived like it was described in Theorem 4.8. Besides, we recognize an approximately exponential behaviour of the trajectories, which we have also seen in formula (4.5), where we computed the mean value of such a process.

For the next theorem, we remember the martingale $(M_n)_{n\in\mathbb{N}}$ defined in Theorem 4.3, which simplifies in this section to $M_n = Y_n R^{-n}$ for $n \in \mathbb{N}$.

Theorem 4.9. *Let $(Y_n)_{n\in\mathbb{N}}$ be a time-homogeneous Galton–Watson process with $p_1 \neq 1$, $R \in \mathbb{R}_+$ and $y_0 \in \mathbb{N}^*$. Let M_∞ be the limit of the martingale from Theorem 4.3. Then, it holds true that either $M_\infty = 0$ almost surely or*

$$\mathbb{P}(M_\infty > 0) = \mathbb{P}(Y_n \to \infty).$$

Proof. For $R \leq 1$, using $p_1 \neq 1$, we get with Theorem 4.4 that $q = 1$. Using $\{Y_n = 0\} = \{M_n = 0\}$ for all $n \in \mathbb{N}$ and $\bigcup_{n\in\mathbb{N}}\{M_n = 0\} \subset \{M_\infty = 0\}$ implies that

$$1 = q = \mathbb{P}\Big(\bigcup_{n\in\mathbb{N}}\{Y_n = 0\}\Big) = \mathbb{P}\Big(\bigcup_{n\in\mathbb{N}}\{M_n = 0\}\Big) \leq \mathbb{P}(M_\infty = 0) \leq 1,$$

so the first alternative, $M_\infty = 0$ almost surely, is true.

From now on, assume that $R > 1$. It is clear that $\{M_\infty > 0\} \subset \{Y_n \to \infty\}$. Let $\pi_{y_0} := \mathbb{P}(M_\infty = 0)$ for $M_0 = y_0$. With Theorem 4.8, it remains to show that π_{y_0} either equals $\mathbb{P}\left(\bigcup_{n\in\mathbb{N}}\{Y_n = 0\}\right) = q^{y_0}$ or 1. Using Theorem 4.2, we have $\pi_{y_0} = \pi_1^{y_0}$. Therefore, without loss of generality, we can assume that $y_0 = 1$ and show that $\pi_1 \in \{q, 1\}$.

We notice that there are $Y_1 = X_1^1$ individuals in the first generation and each of them starts a new Galton–Watson process. For $k \in \mathbb{N}^*$, let $(Y_n^{(k)})_{n\in\mathbb{N}}$ be an independent

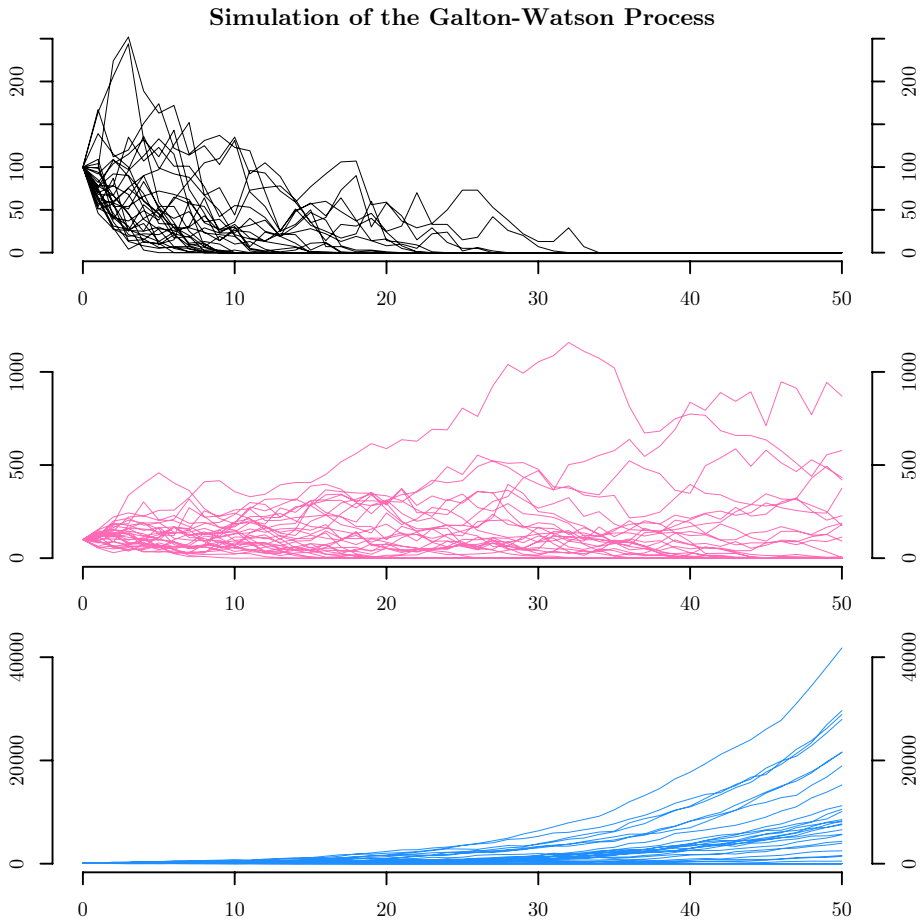

Figure 4.2: We see 30 simulations of the size of the population for $R = 0.9, 1, 1.1$, $y_0 = 100$ and $\alpha = 0.1$.

copy of $(Y_n)_{n\in\mathbb{N}}$, so a Galton–Watson process with $y_0^{(k)} = 1$ and the same offspring distribution. Let these processes be independent, then it follows that

$$Y_{n+1} = \sum_{k=1}^{X_1^1} Y_n^{(k)}, \qquad n \in \mathbb{N}.$$

We normalize this equation by dividing by $R^{-(n+1)}$, so we get

$$M_{n+1} = \frac{1}{R} \sum_{k=1}^{X_1^1} M_n^{(k)}, \qquad n \in \mathbb{N},$$

where $(M_n^{(k)})_{n\in\mathbb{N}}$ denotes the martingale generated by $(Y_n^{(k)})_{n\in\mathbb{N}}$ for $k \in \mathbb{N}^*$. We take

the a.s.-limit on both sides and get

$$M_\infty = \frac{1}{R} \sum_{k=1}^{X_1^1} M_\infty^{(k)},$$

where the random variables $(M_\infty^{(k)})_{k\in\mathbb{N}^*}$ are independent copies of M_∞. Since $M_\infty^{(k)}$ is independent from X_1^1 for all $k \in \mathbb{N}^*$, it follows that

$$\begin{aligned}\pi_1 &= \mathbb{P}(M_\infty = 0) = \mathbb{P}(M_\infty^{(k)} = 0 \text{ for all } k \in \{1, 2, \dots, X_1^1\}) \\ &= \sum_{n=0}^{\infty} \mathbb{P}(X_1^1 = n)\mathbb{P}(M_\infty^{(k)} = 0 \text{ for all } k \in \{1, 2, \dots, n\}) \\ &= \sum_{n=0}^{\infty} p_n \mathbb{P}(M_\infty = 0)^n = \sum_{n=0}^{\infty} p_n \pi_1^n = g_X(\pi_1),\end{aligned}$$

where g_X denotes the probability generating function of the offspring distribution. We see that π_1 is a fixed point and with Theorem 4.4, it follows that $\pi_1 \in \{q, 1\}$, which concludes the proof. □

4.3 Extended Galton–Watson Process

If we want to use the Galton–Watson process as defined in Section 4.1 for modelling the Covid-19 pandemic, there is one main problem. The difficulty when mathematically defining the reproduction number is that the generation time is not deterministic and therefore, it is not easy to clearly define what a generation is (compare the remarks in the beginning of Chapter 3). However, we can overcome this problem by extending the Galton–Watson process.

Therefore, let $(Y_n)_{n\in\mathbb{N}}$ be a stochastic process denoting the number of new infections on each day. For $k, m \in \mathbb{N}$ with $k < m$, let $(X_i^{k,m})_{i\in\mathbb{N}^*}$ denote the number of people that are infected on day m by the i-th new case from day k. Let $Y_0 = y_0$ and set $Y_j = 0$ for $j \in \mathbb{Z}_-$, then it follows that

$$Y_n = \sum_{k=1}^{T} \sum_{i=1}^{Y_{n-k}} X_i^{n-k,n},$$

where T is the maximum length of the generation time. We assume that all infections happen independently, so the random variables $(X_i^{k,m})_{k,m\in\mathbb{N},k<m,i\in\mathbb{N}^*}$ are independent. For fixed $k, m \in \mathbb{N}, k < m$, the set $(X_i^{k,m})_{i\in\mathbb{N}^*}$ is assumed to be identically distributed. Further, we assume that these random variables are integrable with $\mathbb{E}[X_i^{k,m}] = R_m w_{m-k}$, where R_m is the reproduction number on day m. For $j \in \{1, 2, \dots, T\}$, the

weight w_j denotes the probability that the generation time is j days. Further, define $\sigma_{k,m} := \mathrm{Var}(X_i^{k,m})$ for all $i \in \mathbb{N}^*$.

With similar considerations as in equation (4.8), we get

$$\mathbb{E}[Y_n | Y_{n-1}, \dots, Y_{n-T}] = R_n \sum_{k=1}^{T} Y_{n-k} w_k, \tag{4.13}$$

which is exactly formula (3.2) used for the estimation of the reproduction number in Section 3.2. With this equation, we get

$$\mathbb{E}[Y_n] = R_n \sum_{k=1}^{T} \mathbb{E}[Y_{n-k}] w_k,$$

which is a recursive formula for the expectation that cannot be well displayed explicitly.

From [31, p. 385-386], we know that

$$\mathrm{Var}(S) = \mathbb{E}[\mathrm{Var}(S|N)] + \mathrm{Var}(\mathbb{E}[S|N]),$$

which is called the law of total variance. With equation (4.13) and the fact that $(X_i^{k,m})_{i \in \mathbb{N}^*}$ are independent and identically distributed for fixed $k, m \in \mathbb{N}$ with $k < m$, we have

$$\begin{aligned}
\mathrm{Var}(Y_n) &= \mathrm{Var}(\mathbb{E}[Y_n | Y_{n-1}, \dots, Y_{n-T}]) + \mathbb{E}[\mathrm{Var}(Y_n | Y_{n-1}, \dots, Y_{n-T})] \\
&= \mathrm{Var}\left(R_n \sum_{k=1}^{T} Y_{n-k} w_k \right) + \mathbb{E}\left[\sum_{k=1}^{T} Y_{n-k} \sigma_{n-k,n} \right] \\
&= R_n^2 \, \mathrm{Var}\left(\sum_{k=1}^{T} Y_{n-k} w_k \right) + \sum_{k=1}^{T} \mathbb{E}[Y_{n-k}] \sigma_{n-k,n},
\end{aligned}$$

which cannot be simplified further, because the set $(Y_{n-k})_{k=1}^{T}$ is not uncorrelated.

5 Case Fatality Rate

One of the major epidemological key figures is mortality, so we want to raise the question, which portion has died after an infection with the coronavirus. We examine several approaches how to measure this number and whether it has changed over the course of the pandemic. Further, we investigate the impact of some characteristics on the probability of dying. We find that men are significantly more likely to die and that the case fatality rate increase heavily with age. For all these computations, we use Austrian data from [4], where we have the age, sex, date of infection and date of death or recovery for each infected individual. After that, we take a look at the global situation and investigate why the case fatality rate differs so much across countries.

5.1 Estimation Approaches

To begin, we want to examine how to estimate the case fatality rate in the beginning of a pandemic. The major problem we have to deal with is that deaths come up later than infections. Therefore, it is not correct to compare the total number of infections with the total number of deaths at a certain point in time. Thus, as long as the pandemic is going on, it is not possible to tell the correct case fatality rate of a population. For this section, we assume that the case fatality rate is constant over the course of the pandemic in a certain population.

For $t \in \mathbb{N}^*$, let $Y_t \in \mathbb{N}$ be the total number of infections, $D_t \in \mathbb{N}$ the total number of deaths and $C_t \in \mathbb{N}$ the total number of recoveries up to time t. At time t, let $K_t \in \mathbb{N}$ be the number of current infections and $G_t \in \mathbb{N}$ be a random variable denoting the number of current infections that will end with death. Thus, G_t takes values in $\{0, 1, \dots, K_t\}$. It can be easily seen that $Y_t = C_t + D_t + K_t$. At time t, the current estimate of the case fatality rate $F_t \in [0, 1]$ fulfils

$$F_t = \frac{D_t + \mathbb{E}[G_t]}{Y_t}.$$

Using that G_t is bounded, we can calculate a lower and an upper naive bound $l_t, u_t \in [0, 1]$ for the case fatality rate by

$$l_t = \frac{D_t}{Y_t}, \qquad u_t = \frac{D_t + K_t}{Y_t}. \tag{5.1}$$

For the Austrian infection data, these bounds are plotted in Figure 5.1. We see a high level of uncertainty in March, since the range for the case fatality rate is nearly

the entire $[0, 1]$-interval. When the number of incident cases began to decrease in the end of March, the upper bound dropped fast, whereas the lower bound increased constantly till the end of April. The width of the bounded region is given by $\frac{K_t}{Y_t}$ and is therefore determined by the total number of infections and the number of currently infected people. That explains why the two bounds were very close in June, when there were few incident but a lot of old cases, and widened again in autumn, when the number of incident cases was very high.

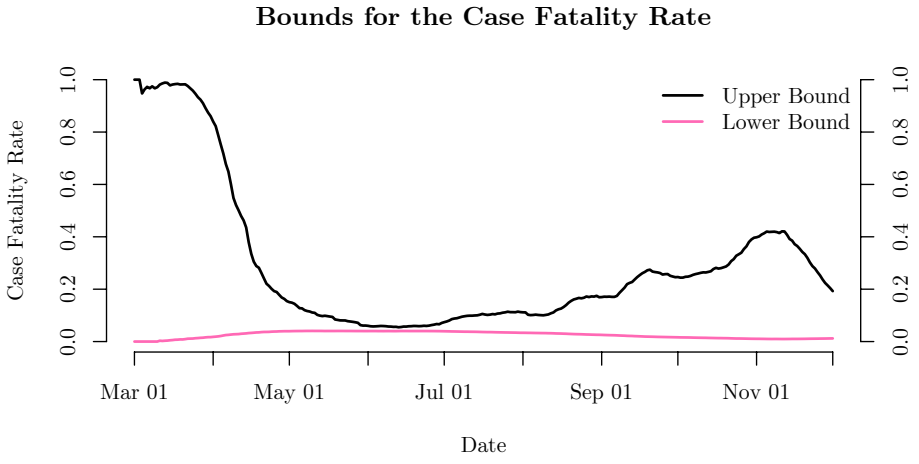

Figure 5.1: According to equation (5.1), we see the upper and lower bound for the case fatality rate from March to November 2020.

As a next step, we try to estimate the case fatality rate using two different approaches. For the first one, we assume that G_t is a sum of K_t random variables that are independent and identically Bernoulli distributed. We suppose that the probability of dying is given by F_t. Then, G_t is binomial distributed with parameters $(n = K_t, p = F_t)$, so we have $\mathbb{E}[G_t] = K_t F_t$. Substituting that into the definition of the case fatality rate, gives us

$$F_t = \frac{D_t + K_t F_t}{Y_t} \quad \Leftrightarrow \quad F_t = \frac{D_t}{Y_t - K_t} = \frac{D_t}{D_t + C_t}. \tag{5.2}$$

We can extend this approach by dropping the assumption of identical distributions. Instead, we can assign an individual probability of dying to every person that is currently infected. This parameter can depend on the age and sex of the person, pre-existing illnesses or the time since the infection. For a time t and a individual $i \in \{1, 2, \ldots, K_t\}$, let $q_t^{(i)}$ be the probability that person i will die from this infection. So, we get

$$F_t = \frac{D_t + \sum_{i=1}^{K_t} q_t^{(i)}}{Y_t}. \tag{5.3}$$

However, this extension is not very useful, because it requires detailed data over a longer period of time. However, here, we are interested in estimating the case fatality rate at an early stage of a pandemic, where these data are not available.

So, we return to formula (5.2). It is quite natural to simply consider only those people for whom it is already clear whether they die or recover. However, the reason why this is not perfectly correct is that the distributions of the length of the period from infection to recovery and from infection to death do not have to be the same. In fact, for Covid-19, the data from [4] show that the mean is approximately the same, but the standard deviation of the time to death is twice as large as the standard deviation of the time to recovery.

The second approach deals with this problem. We build up on the first approach, but, instead of D_t+C_t, we take a different basic set over which we want to calculate the probability of dying, namely the portion of infected people that would have already died conditioned on that their death is certain. For an infected person i, let X_i be a random variable that can take values in $\{\text{recover}\} \cup \{\text{die after } k \text{ days} \mid k \in \{0, 1, \dots, k_{max}\}\}$, and let $(X_i)_i$ be independent and identically distributed. For $k \in \{0, 1, \dots, k_{max}\}$ define $p_k := \mathbb{P}(X_1 \text{ is dead after } k \text{ days} \mid X_1 \text{ dies})$. For $k > k_{max}$, we set $p_k = 1$. It is easy to estimate p_k with the data from [4], as we just need a distribution of the duration from infection to death. Let y_t be the number of incident cases on day t, so $Y_t = \sum_{s=1}^{t} y_s$. At time t, the above described portion is then $\sum_{k=0}^{t-1} p_k y_{t-k} \leq Y_t$. Thus, we have

$$F_t = \frac{D_t}{\sum_{k=0}^{t-1} p_k y_{t-k}}. \tag{5.4}$$

We use all past values for the estimation of p_k, so these weights can change a bit over time.

In Figure 5.2, we see the naive bounds as well as the two methods to find the case fatality rate. For each day, the plot shows the average case fatality rate up to that day. We observe that our estimates differ a lot in the first month of the Covid-19 pandemic, but are very close since mid April. We see that the case fatality rate is decreasing from July to November, which is probably because of better medical treatment. According to [8, p. 2], it can be observed in most countries that the case fatality rate reduced from the first to the second wave.

5.2 Case Fatality Rate in the Course of the Pandemic

If we analyse a pandemic ex-post, we can examine whether the case fatality rate has changed over time. This could be due to a limited number of intensive care beds, better medication or vaccination. When we want to measure the case fatality rate for a certain day, there are at least three different approaches.

Let t be an arbitrary day. We have to assign a certain subset of cases to day t. Firstly, we can take all incident cases on day t as our sample. Secondly, we can

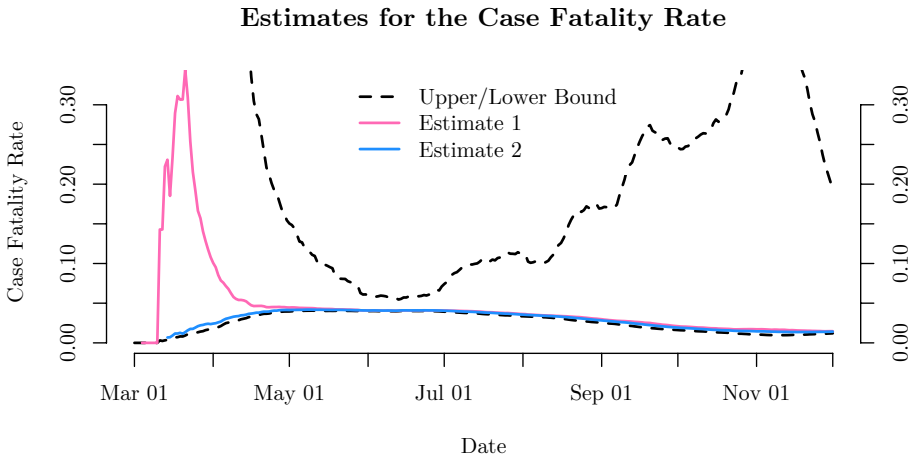

Figure 5.2: In addition to the bounds from Figure 5.1, we see the results of the two approaches described in equation (5.2) and (5.4) with $k_{max} = 60$.

consider all cases that ended on day t, either with death or recovery. Thirdly, we can take all cases into account that are infected on day t. For each approach, we then calculate the proportion of cases that ended with death from the corresponding sample.

If the duration of the infections were constant, then the first and the second approach would give the same graph, but just lagged by this duration. The third approach would then be a moving average of the first one. However, this is not the case. That rises a problem for the third method: Cases with a longer infection duration contribute more the case fatality rate than cases with a shorter infection duration, whereas for the first and second method every case is used exactly once. If longer infections have a different dying probability than shorter ones, we get a biased estimate with the third approach.

With the data from [4], the courses of the case fatality rate for the three approaches are plotted in Figure 5.3. Although the estimated case fatality rates are quite different, we observe a common significant decline in July 2020, which could be due to better medication and a lower average age of the infected people. In October 2020, we then see a slight increase as the average age increases again. In Section 5.3, we will take a look this change of the average age of infected people. The high case fatality rate in the beginning might also come from a smaller number of tests and a therefore higher portion of deaths, because many infections were undetected. However, in [8, p. 3], it is stated that the age structure has changed because young people have a higher mobility and are therefore more affected in the second wave than in the first one.

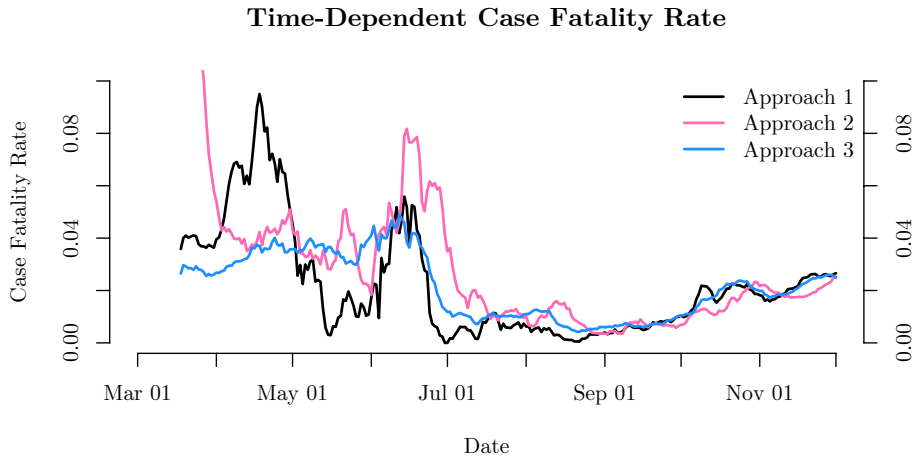

Figure 5.3: For the three approaches described, we see the case fatality rate for every single day in the course of the pandemic from March to November 2020.

5.3 Differences for Age and Sex

We start with taking a look at the impact of the age on the case fatality rate. It is a well-known fact that older people have a higher risk of dying from a Covid-19 infection. The case fatality rate for each age is plotted in Figure 5.4, where we considered all cases that are either dead or recovered but ignore ongoing infections. Below the age of 50, the absolute number of deaths is near 0 for each age. At the age of 62 the case fatality rate is the first time greater than 1% and at the age of 79 the first time greater than 10%. From that point, it rises more or less linearly up to 45% at the age of 100. Thereafter, there are too few infections to calculate a reliable case fatality rate.

Besides, we want to investigate whether the age of the deaths vary over time. The results plotted in Figure 5.5 show no long-term trend. We only observe higher fluctuations at lower numbers of new infections.

Then, we want to examine whether the sex influences the case fatality rate. So, we conducted a two-sample t-test to see whether the difference in the case fatality rate is statistical significant. The null hypothesis is that women have a higher or equal case fatality rate. We consider only cases that have already died or recovered till November 2020. We can then reject the null hypothesis with a p-value of approximately 10^{-12}. From the beginning of the pandemic up till November, we get a case fatality rate for men of 2.16% and for women of 1.75% considering all infections that ended. That is interesting, because at high ages, where the most people die, the portion of women is higher.

Therefore, we want to combine these two characteristics. In Figure 5.6, we see that

Figure 5.4: Considering all deaths and recoveries till November 2020, we see the case fatality rate for each age.

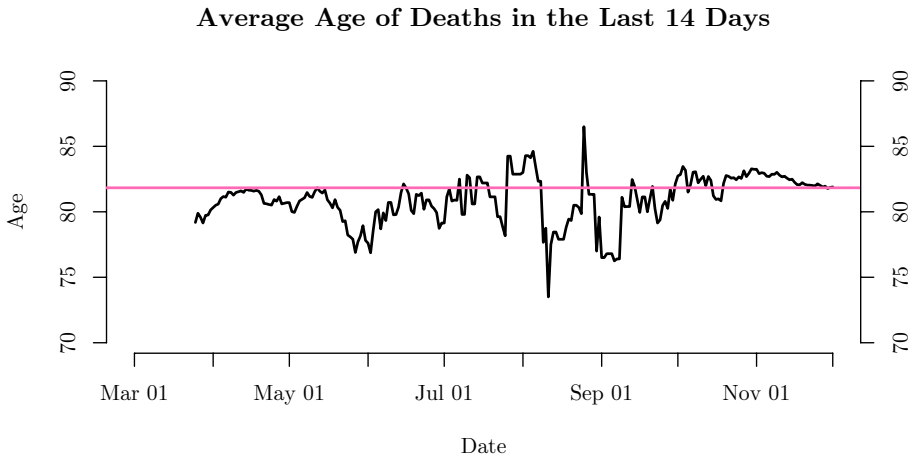

Figure 5.5: We see the average age of people died in the last 14 days from March to November 2020. In pink, we see the overall average at an age of 81.8.

for all ages from 65, men have a higher case fatality rate. We can also observe that till the age of 83 the portion of male and female is approximately equal. From that point the female portion increases steadily.

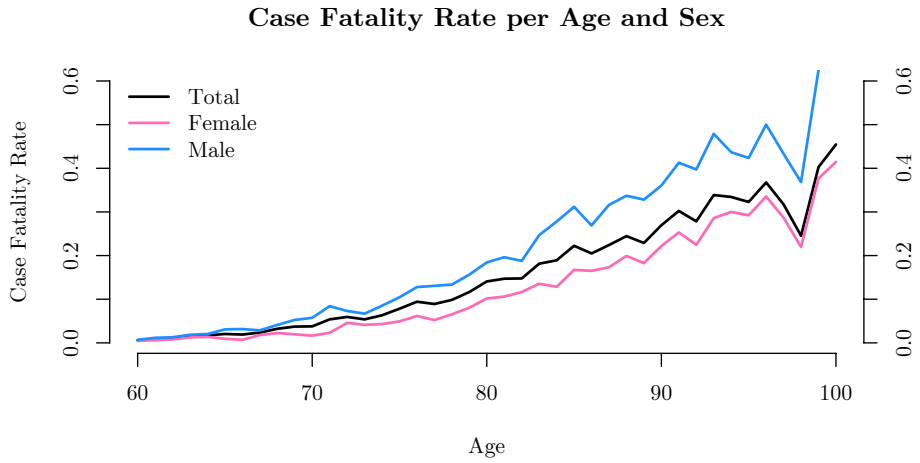

Figure 5.6: For each age from 60 to 100, we see the case fatality rate for male and female separately considering all deaths and recoveries till November 2020.

5.4 Comparison over Different Countries

In this section, we want to examine the case fatality rate in different countries to find its main drivers. We will run a linear regression on the most important factors to show a significant dependency of the case fatality rate on these factors. Finally, we will discuss the limitations of this analysis. In Figure 5.7, we see a map showing the case fatality rate in each country till November 2020.

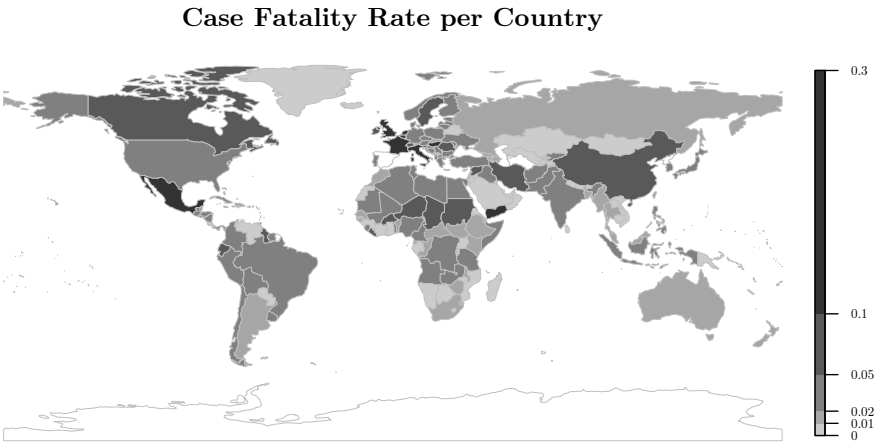

Figure 5.7: We see a map showing the case fatality rate for each country considering all deaths and recoveries till November 2020.

Let I be the set of all countries with Covid-19 cases where the necessary data

points are available ($n = 158$). For $i \in I$, let F_i be the case fatality rate, A_i be the ratio of people who are older than 65 years and B_i be the number of hospital beds per thousand people in country i. Let $\alpha, \beta_1, \beta_2 \in \mathbb{R}$ and let ϵ_i be the error term for country i. Then, we can set up the following equation:

$$F_i = \alpha + \beta_1 A_i + \beta_2 B_i + \epsilon_i, \qquad i \in I. \tag{5.5}$$

Using the data from [20] and the function `lm()` from the package `stats`, we can conduct a linear regression on these factors in R to see if a dependency is empirically valid. We observe for both factors a highly significant dependency. Firstly, the ratio of people older than 65 years has a positive influence on the case fatality rate, which is also confirmed by the results in Section 5.3. Secondly, a higher number of hospital beds leads on average to a lower case fatality rate.

	Estimate	Std. Error	t value	Pr(>|t|)
α (Intercept)	0.01839	0.00453	4.06332	0.00008
β_1 (ratio of age 65 or older)	0.00385	0.00075	5.16341	0.00000
β_2 (hospital beds per thousand)	−0.00408	0.00150	−2.72564	0.00716

However, we face some problems, when trying to compare the case fatality rates in countries all over the world. Firstly, a uniform definition of a Covid-19 death poses difficulties. To solve this problem, the WHO has implemented a definition in [19, p. 15]:

> *COVID-19 death is defined for surveillance purposes as a death resulting from a clinically compatible illness in a probable or confirmed COVID-19 case, unless there is a clear alternative cause of death that cannot be related to COVID disease (e.g. trauma). There should be no period of complete recovery between the illness and death.*

Additionally, also the number of tests is important: If the relative number of tests in a country is lower, the number of cases is more underestimated, which leads to a higher ratio of deaths.

Last but not least, the progress in the course of the pandemic is also relevant, especially in the beginning. As the number of deaths is lagged a couple of days behind the number of infections, countries have a lower case fatality rate in the beginning of the spread than later. However, this effect disappears gradually.

6 Covid-19 as Mortality Shock

As this pandemic leads to a significant disruption of mortality rates all over the world, we want to examine how they are affected. Therefore, we will introduce the term structure of mortality, which is a name that is related to finance, where the term structure of interest rates describes the yield curve. What happens if an event like a pandemic suddenly changes the term structure of mortality? One example for a shock is a parallel shock. We further analyse empirically if the Covid-19 pandemic leads to such a parallel shock in the term structure of mortality and discuss consequences for life expectancy and annuity prices.

As in Chapter 5, Covid-19 death rates play a major role. In addition to Section 5.3, we will discuss whether the probability of death has increased proportionally for every age.

We follow the approach from [17] and build up on it. We support our theoretical analysis by some empirical calculations on Austrian data. We will use the life table from [28], the age distribution of the Austrian population from [26, 27] and the Austrian Covid-19 data from [4].

6.1 Mortality Shocks

We start with a short introduction to actuarial notation. Let $\mu_x \geq 0$ be the *mortality rate* of a person at age $x \in \mathbb{R}_+$. For a person at age x, let ${}_tp_x$ be the chance of surviving t years and ${}_tq_x$ the probability of dying within the next t years. It holds that ${}_tp_x + {}_tq_x = 1$. For $t = 1$, we simply write q_x and p_x. We know that

$$ {}_tp_x = \exp\left(-\int_0^t \mu_{x+s}\mathrm{d}s\right). $$

If we distinguish between men and women, we write ${}_tq_x^{(m)}, {}_tp_x^{(m)}, \mu_x^{(m)}, {}_tq_x^{(f)}, {}_tp_x^{(f)}, \mu_x^{(f)}$ respectively. Furthermore, it is important to distinguish between *cohort life tables* and *period life tables*. The cohort life table measures how many people die from a generation, so it tracks the whole life of fixed individuals that are born during the same specific time frame. However, a period life table considers people that life during the same specific time frame and measures the survival ratios within this time frame, it is also called a generation life table. Therefore, a period life table is easier to observe, whereas the cohort life table gives the true survival ratios for the selected

generation. They are in particular different if the mortality changes over time. Our life table [28] is a period life table.

For a person at age x, we have an *accidental mortality rate* $\lambda_x \geq 0$ and a *natural mortality rate* $\theta_x \geq 0$ with $\lambda_x + \theta_x = \mu_x$. From that, the *term structure of mortality* is defined as $(\log(\theta_x))_x$. An event that suddenly changes the mortality rate from μ_x to $\tilde{\mu}_x$, is called a *mortality shock*, in contrast to a long-term trend. When we observe such a shock, we want to find the mortality risk-adjusted age, that means, for an age x, we want to find the age y, so that $\tilde{\mu}_x = \mu_y$. If there is no such y, we take $\arg\min_y |\tilde{\mu}_x - \mu_y|$.

Furthermore, we want to discuss shocks in a certain model for mortality. In 1825, Benjamin Gompertz developed a law of human mortality, which assumes that the natural mortality rate increases exponentially with age. If we set the accidental mortality rate $\lambda_x = \lambda$ for all x, we get the *Gompertz–Makeham law of mortality*, which states that the mortality rate is given by

$$\mu_x = \lambda + h\mathrm{e}^{gx}, \qquad x \in \mathbb{R}_+,$$

with growth rate $g > 0$ and starting value $h > 0$. With the definitions and equations above, for $x, t \in \mathbb{R}_+$, it follows that

$$\begin{aligned} {}_tq_x &= 1 - {}_tp_x = 1 - \exp\left(-\int_0^t \mu_{x+s}\mathrm{d}s\right) = 1 - \exp\left(-\int_0^t \lambda + h\mathrm{e}^{g(x+s)}\mathrm{d}s\right) \\ &= 1 - \exp\left(-\lambda t - h\mathrm{e}^{gx}\frac{\mathrm{e}^{gt} - 1}{g}\right), \end{aligned}$$

under Gompertz–Makeham. So, for the one year death probabilities, we get

$$q_x = 1 - \exp\left(-\lambda - h\mathrm{e}^{gx}\frac{\mathrm{e}^{g} - 1}{g}\right). \tag{6.1}$$

This law does not work for $x < 35$, but it is a good approximation for $x \geq 35$, which we denote as Gompertzian age range. This is also supported by Figure 6.1, where we see the observed dying probabilities from the period life table [28] for male and female, and a maximum-likelihood estimation of the Gompertz–Makeham parameters for $\lambda = 0$. The maximum-likelihood estimation works as described in the following. For an age x, let N_x be the size of the population with age x and let D_x be the number of observed deaths at age x within one year. Then, the dying probabilities in the period life table are estimated by $\hat{q}_x = \frac{D_x}{N_x}$. If we assume that D_x is Poisson-distributed with parameter $q_x N_x$, we get

$$L(g, h, \lambda | D_x) = \prod_{x=35}^{99} \mathrm{e}^{-q_x N_x} \frac{(q_x N_x)^{D_x}}{D_x!}$$

for the likelihood function. If we substitute q_x from equation (6.1) and take the logarithm, we get the log-likelihood function

$$l(g,h,\lambda|D_x) = \sum_{x=35}^{99} -N_x\big(1-\exp(-\lambda - h\mathrm{e}^{gx}\tfrac{\mathrm{e}^g-1}{g})\big) + D_x \log\Big(N_x(1-\exp(-\lambda - h\mathrm{e}^{gx}\tfrac{\mathrm{e}^g-1}{g}))\Big) - \log(D_x!). \tag{6.2}$$

As we cannot compute its maximum directly, we approximate the result numerically using the function `optim()` from the package `stats` in R.

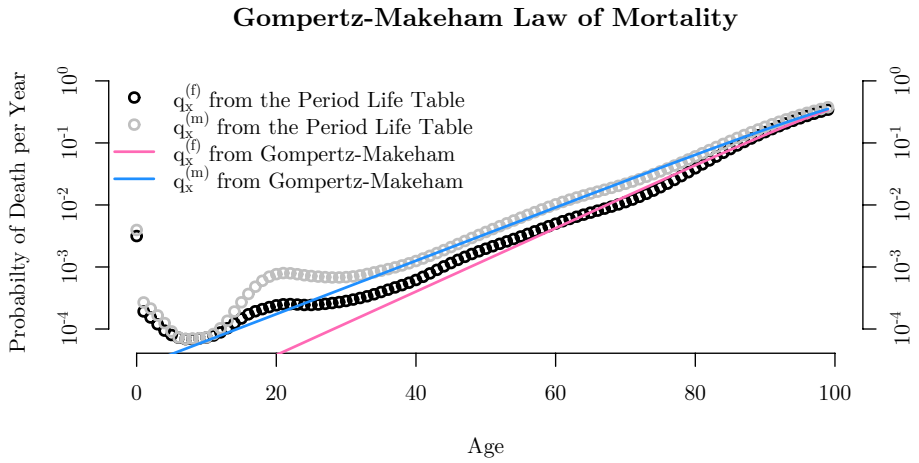

Figure 6.1: We see the dying probability from [28] and the estimation by Gompertz–Makeham law with $\lambda = 0$. We conducted a maximum-likelihood estimation for $x \geq 35$. The resulting parameters are $g = 0.1178, h = 3.405 \cdot 10^{-6}$ for female and $g = 0.0990, h = 2.280 \cdot 10^{-5}$ for male.

In the Gompertz–Makeham framework, the term structure of mortality for the Gompertzian age range is

$$\log(\mu_x - \lambda) = \log(h) + gx. \tag{6.3}$$

If an event like a pandemic impacts death rates, the term structure changes. One example for such a change is a *parallel shock* of the log mortality rate, which is defined as an additive increase (or decrease) by a constant $u \in \mathbb{R}$ for all x in the Gompertzian age range, so equation (6.3) changes to

$$\log(\tilde{\mu}_x - \tilde{\lambda}) = \log(h) + gx + u. \tag{6.4}$$

Assuming an unchanged accidental death rate $\tilde{\lambda} = \lambda$, we observe that with a parallel shock the new natural mortality is proportional to the old, so the natural death rate

is multiplied with the same factor, e^u, for all x in the Gompertzian age range, but the mortality growth rate stays the same.

The interpretation of a parallel shock in the term structure of mortality is that people at every age are equally affected in the sense that the probability of death has increased for everyone by the same factor. This was, for example, not the case for the Spanish flue pandemic in 1918, where people at younger ages were affected disproportionately higher.

An alternative to a parallel shock is a constant shock, where we assume that the new mortality rate must be equal to the old at some age x^*, so we need $\tilde{\mu}_{x^*} = \mu_{x^*}$. Thus, a shock $u > 0$ requires a decline in the mortality growth rate g in order to compensate for the increased mortality. We define g_u as the adjusted slope and substitute in our assumption $\tilde{\mu}_{x^*} = \mu_{x^*}$. This gives us

$$\log(h) + g_u x^* + u = \log(h) + g x^*,$$

so we need $g_u = g - \frac{u}{x^*} < g$ to fulfil the condition. The comparison of a parallel and a constant shock to the term structure of mortality is plotted in Figure 6.2.

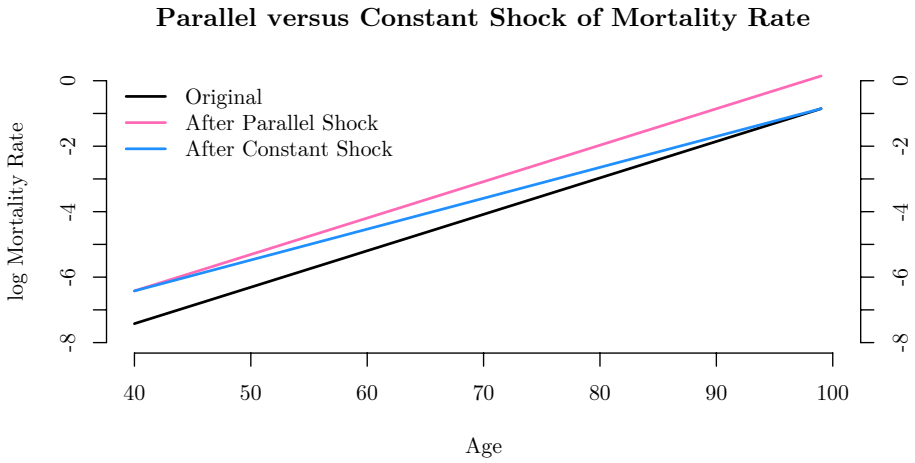

Figure 6.2: For the ages 40 to 100, we compare the original log mortality rate to the log mortality rate after a parallel shock and after a constant shock with $x^* = 99$.

6.2 Is Covid-19 a Parallel Shock?

Now, we want to determine whether Covid-19 was a parallel shock to the Austrian term structure of mortality. Given $x \in \{0, 1, 2, \ldots, 99\}$, let $q_x \in [0, 1]$ be the chance of death for an Austrian individual within the next year at age x before Covid-19, using

the data from [28]. We assume that the mortality rate is constant within one year. So, for $t \in [0, 1)$, we assume $\mu_{x+t} = \mu_x$. Thus, we have $q_x = 1 - \mathrm{e}^{-\mu_x}$. It follows that

$$\mu_x = -\log(1 - q_x).$$

Let μ_x^C be the Covid-19 mortality rate and suppose that the non-Covid-19 mortality rate μ_x is unchanged. Then, the mortality rate in the presence of this disease is $\mu_x^C + \mu_x$. Suppose that the time frame under consideration is $t \in \mathbb{N}^*, t \leq 365$ days long, let $D_x^t \in \mathbb{N}$ be the number of Covid-19 deaths at the age x in this time frame, using the data from [4], and $N_x \in \mathbb{N}^*$ the size of the population with age x, using the data from [26, 27]. So, the probability of dying from Covid-19 in this time frame is

$$\left(1 - \exp(-\tfrac{t}{365}(\mu_x^C + \mu_x))\right) \cdot \frac{\mu_x^C}{\mu_x + \mu_x^C}.$$

Now, we estimate this probability by observing which portion of the population actually died from Covid-19. Then, we can find the Covid-19 mortality rate μ_x^C, by solving the equation

$$\left(1 - \exp(-\tfrac{t}{365}(\mu_x^C + \mu_x))\right) \cdot \frac{\mu_x^C}{\mu_x + \mu_x^C} = \frac{D_x^t}{N_x} \tag{6.5}$$

numerically for each x.

Since we know that $\tilde{\mu}_x = \mu_x^C + \mu_x$ is the total mortality rate in the presence of Covid-19, we can now find the mortality risk-adjusted age as described in Section 6.1. As we assumed a constant mortality rate within one year, we approximate the results by linear interpolation. The results of this mapping are plotted in Figure 6.3: We see how much the mortality risk-adjusted age is higher than the original age for female and male separately.

Furthermore, we want to discuss whether Covid-19 is a parallel shock. If we assume that the accidental death rate is 0, we have to check if the excess factor

$$\log\left(\frac{\mu_x^C + \mu_x}{\mu_x}\right) \tag{6.6}$$

is approximately equal for all x. In Figure 6.4, we see this term for each x for female and male separately. We observe that the shock is not parallel, since it rises with increasing age, and that the shock size is $u \in [0, 0.16]$. Only for the ages 80 to 100, it is approximately constant. It can be concluded that not everyone is affected proportionally to the dying probability, but that older people are affected disproportionally higher. Further, we can see that the dying probability of men has increased disproportionally more than for women, although they already had a higher probability of death before Covid-19. This extends the results from Section 5.3, where we showed that men have a higher case fatality rate than women.

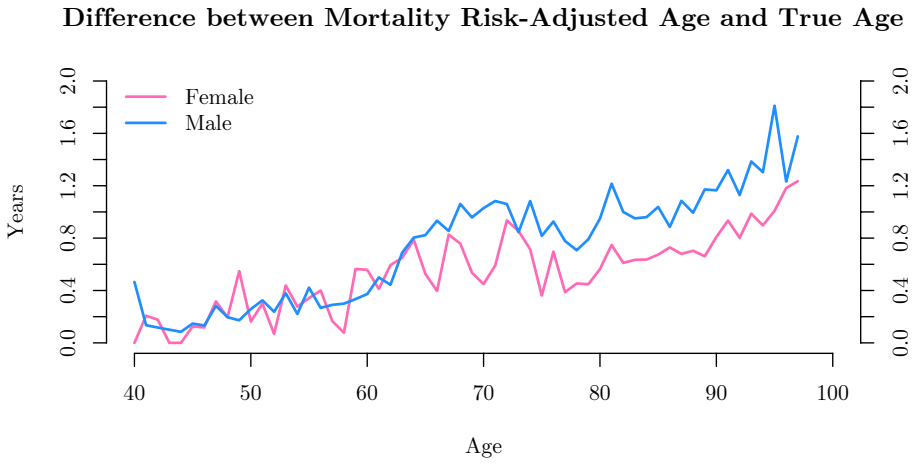

Figure 6.3: We see the absolute difference between the mortality-risk-adjusted age and the true age for Austrian data considering all deaths till November 2020 with population data from January 01, 2020.

Figure 6.4: For the age 40 to 100, we see the excess factor of mortality rates from formula (6.6) to check whether Covid-19 is a parallel shock.

6.3 Shock at an Infection

Another interesting question is: How much does the mortality rate increase, if a person is infected? Is the factor of this shock equal for all ages? Therefore, we need to estimate the mortality rate for an infected individual. Analogous to the procedure in Section 6.2, consider a time frame with a length of $t \in \mathbb{N}^*, t \leq 365$ days. For each age x, let $I_x^t \in \mathbb{N}$ and $D_x^t \in \mathbb{N}$ be the number of Covid-19 deaths and infections

respectively, at the age x in this time frame. Let $\tilde{\mu}_x$ and μ_x^I be the mortality rate in general in the presence of Covid-19 and for infected people respectively. Analogous to equation (6.5), we can estimate the mortality rate for infected people by solving the equation

$$\left(1 - \exp(-\tfrac{t}{365}(\mu_x^I + \tilde{\mu}_x))\right) \cdot \frac{\mu_x^I}{\tilde{\mu}_x + \mu_x^I} = \frac{D_x^t}{I_x^t}, \tag{6.7}$$

and the size of the shock for each age, that is the relative increase of the mortality rate, is given by

$$\log\left(\frac{\mu_x^I + \tilde{\mu}_x}{\tilde{\mu}_x}\right). \tag{6.8}$$

In Figure 6.5, we see the difference between the true age and the mortality risk-adjusted age after an infection. That means, for example, that if a 70-year-old man is infected, he suddenly has a biological age of 79 according to the mortality rate. We can observe that women are less affected than men.

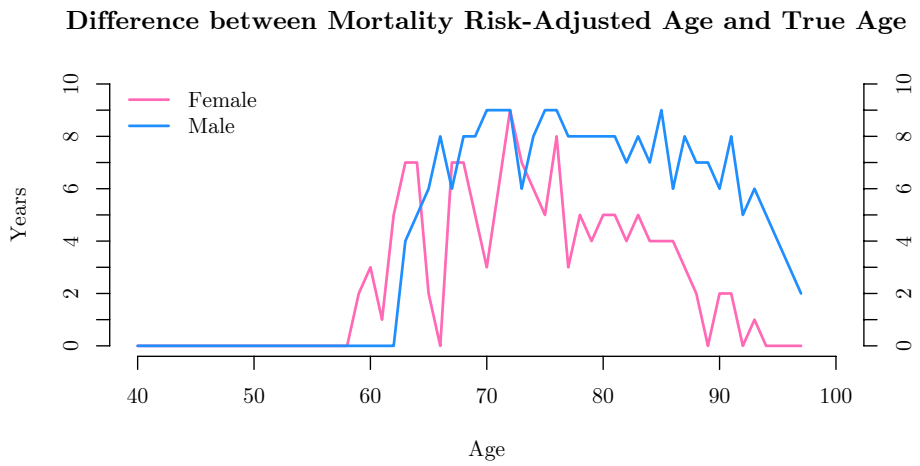

Figure 6.5: For the age 40 to 100, we see the absolute difference between the mortality-risk-adjusted age at an infection and the true age separately for male and female.

The excess factor is plotted in Figure 6.6. We see that the rise in mortality rate increases till the age of 70 where it peaks. After this point it declines.

To be precise, since the true infection numbers cannot be observed, but only the number of positive tests, the resulting mortality rates refer to a positive test. Additionally, we know from Chapter 5 that $\frac{D_x^t}{I_x^t}$ is just an approximation for the case-fatality-rate, because deaths are time-lagged.

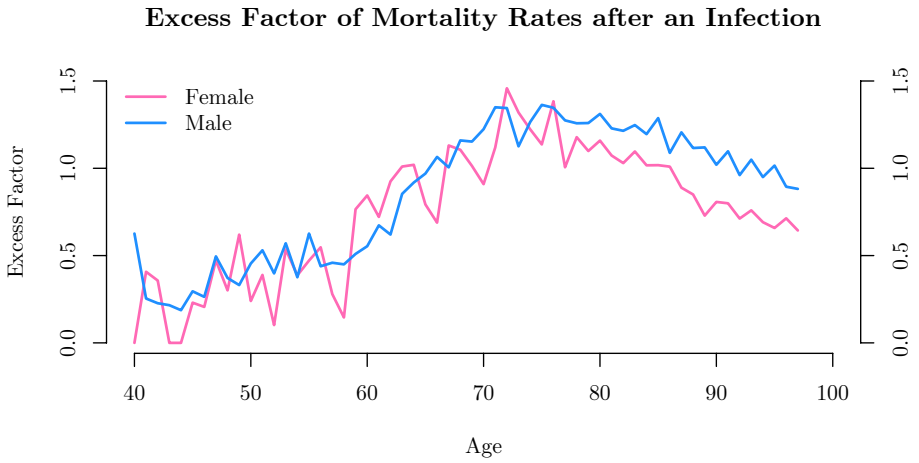

Figure 6.6: For the age 40 to 100, separately for male and female, we see the additional excess factor of mortality rates from formula (6.8) after a person gets infected.

6.4 Effect on Life Expectancy

In this section, we want to find empirical answers to some questions related to life expectancy.

One of the main arguments of policy makers for Covid-19 measures, especially lockdowns, is saving lives by reducing the number of incident cases or keeping the intensive care beds free. Critics argue that we protect people who would have died soon anyway. In any case, it would be beneficial for the discussion to only talk about the number of people who died, but also take the number of years of life lost into account. A bias of the following considerations is that people who die from Covid-19 are likely to have a significant lower life expectancy than other people of that age because of pre-existing illnesses that favour death after an infection, compare [5, p. 7]. Nevertheless, we assume that people who died from Covid-19 have on average the same life expectancy like the whole population.

For an age x, let $d_x \in \mathbb{N}$ be the number of Covid-19 deaths and $e_x \in \mathbb{R}_+$ be the truncated life expectancy at age x. If we multiply these two and sum over all x, we get the total number of lost years of life

$$L := \sum_{x=0}^{99} d_x e_x.$$

Let T_x be a random variable denoting the number of years a person at the age of x is still alive. Then, the truncated life expectancy can be calculated in a discrete setting

with the final age $\omega = 100$ by

$$e_x = \mathbb{E}[T_x] = \sum_{n=1}^{\omega-x} n \, {}_np_x \, q_{x+n} = \sum_{n=1}^{\omega-x} \sum_{k=1}^{n} {}_np_x \, q_{x+n} = \sum_{k=1}^{\omega-x} \sum_{n=k}^{\omega-x} {}_np_x \, q_{x+n} = \sum_{k=1}^{\omega-x} {}_kp_x.$$

A similar formula works for the continuous case, where we need an integral instead of the sum, and can calculate the true life expectancy instead of the truncated.

We can also add a time dependence to see how the loss of years of life develop over the different phases of the pandemic. Let $d_{x,t}$ be the number of people died at age x on day t of the pandemic. Then

$$L_t := \sum_{x=0}^{99} d_{x,t} e_x$$

is the number of years of life lost on day t. This time series is plotted in Figure 6.7 together with the daily number of incident cases. Comparing the first and the second wave, we observe that within the first wave many more years of life were lost in relation to the number of incident cases than during the second wave. That supports the thesis that medical treatment was improved and older people were protected better after the first wave. We also see that the time lag of the peaks between the two curves increases from 10 in the first wave to 20 in the second wave.

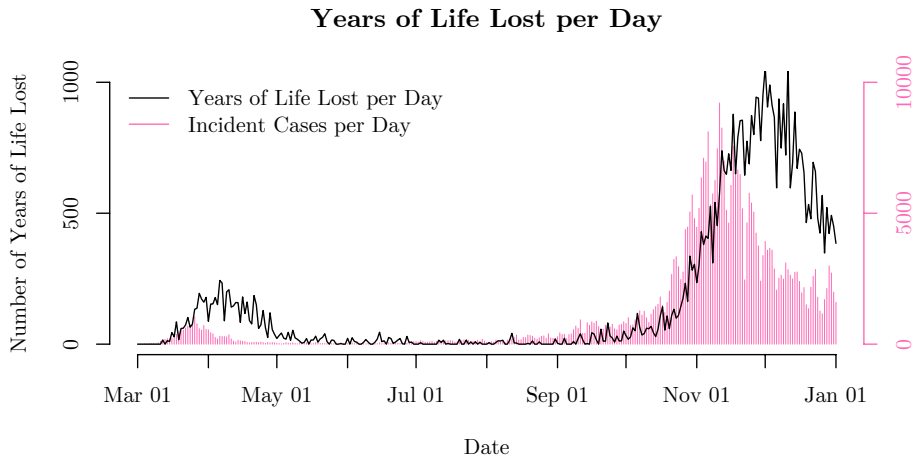

Figure 6.7: We see the number of years of life lost per day (left axis) compared to the number of incident cases (right axis) from March to December 2020.

However, the number of deaths in other areas decreased due to lockdowns and other Covid-19 measures. On the one hand, fewer deaths caused by the flu lowered

the natural mortality rate.[1] On the other hand, there have been avoided a lot of accidental deaths because of less traffic and limited possibilities of spending free time. Thus, the accidental death rate decreased, so $\tilde{\lambda} < \lambda$. In the UK, for example, the accidental death rate declined by 50-70%. However, after lockdowns the accidental death rate increased almost up to its normal level.[2] Similar results were observed in other countries all over the world. Besides, there might have also been deaths which are indirectly related to the pandemic like suicides or because there are less intensive care beds available. However, it seems that there are actually less suicides since the pandemic outbreak in March.[3] An analysis by the Cambridge University shows that the death rates of 20 to 24 year old men declined during the lockdown. The mortality rate of this group is usually higher than the age suggests, because of risk-taking behaviour. The total death rates are suggested to have reduced by 30%.[4] To sum up, there are a lot of direct and indirect effects, which work in different directions. In our analysis, we want to focus on the increase in death probability because of a Covid-19 infection and assume that the other effects play a minor role or equalize.

In [5, p. 7], it is stated that those who die from a Covid-19 infection have already a lower life expectancy than the average. This would indicate that the number of years of life lost is actually smaller. It also implies that those who survive this pandemic are healthier and will probably live longer. This will lead to lower death rates in the next years compared to pre-Covid-19. Thus, one should be cautious when using life tables without adjusting them, as prices of annuities might be too low.

6.5 Effect on Annuity Prices

We want to use the results from the previous sections to derive some implications an important insurance product, the annuity. We distinguish between a temporary (one year) and a permanent shock, but only consider positive shocks. We will discuss parallel shocks in general in the Gompertz–Makeham framework, look particular at the Covid-19 shock and examine its impact on annuity prices. Further, we will outline the role of interest rates.

Let $r > -1$ be the interest rate, then the discount factor is denoted by $v = \frac{1}{1+r}$. We define an *annuity* as an insurance product that pays 1 unit each year as long as

[1] https://www.tagesspiegel.de/wissen/wenige-influenza-tote-haben-die-corona-massnahmen-die-grippewelle-beendet/25782934.html, accessed on November 08, 2020

[2] https://fleetworld.co.uk/car-accident-rates-climb-sharply-after-decline-in-lockdown/, accessed on November 08, 2020

[3] https://www.aerzteblatt.de/nachrichten/117216/Moeglicherweise-weniger-Suizide-seit-Corona, accessed on November 08, 2020

[4] https://www.telegraph.co.uk/news/2020/05/28/deaths-young-men-have-fallen-lockdown-cambridge-study-shows/, accessed on November 08, 2020

the underlying person is alive. Then, the present value of an annuity a_x is given by

$$a_x := \sum_{k=1}^{\omega-x} v^k {}_kp_x,$$

where ${}_kp_x$ is probability to survive for k years at age x. For a positive shock of mortality rates, the adjusted survival probability ${}_k\tilde{p}_x$ is smaller than the original one, so ${}_k\tilde{p}_x \leq {}_kp_x$ for all $k, x \in \mathbb{N}$. Therefore, the new price of an annuity $\tilde{a}_x$ is cheaper, so $\tilde{a}_x \leq a_x$ for each age x. However, if the shock is only temporary (one year), we can directly derive that

$$\tilde{a}_x = \sum_{k=1}^{\omega-x} v^k \tilde{p}_x \cdot {}_{k-1}p_{x+1} = \tilde{p}_x v \sum_{k=0}^{\omega-x-1} v^k {}_kp_{x+1} = \tilde{p}_x v(1 + a_{x+1}).$$

Furthermore, mortality shocks, which make annuities cheaper, can be compensated by a decline in interest rates. Thus, we are interested in finding a $\tilde{v} = \frac{1}{1+\tilde{r}}$ that solves the equation

$$\sum_{k=1}^{\omega-x} v^k {}_kp_x = \sum_{k=1}^{\omega-x} \tilde{v}^k {}_k\tilde{p}_x$$

for permanent shocks. However, this equation has no algebraic solution, so we will compute it numerically using the function `uniroot()` from the package `stats` in R.

Now, we want to take a look at the results for permanent parallel shocks in the Gompertz–Makeham framework in general, firstly. For the Gompertz–Makeham parameters, we take the results presented in Figure 6.1. As these parameters were estimated from a period life table, we have to assume in the following that mortality rates will not change in the future. The annuity prices for women are plotted in Figure 6.8. We remember the size of the parallel shock u from equation (6.4). We see a decline in annuity prices for positive parallel shocks, which vanishes towards the end of life. For each further shock, we observe approximately the same absolute decline in price. For comparison, we remember from Figure 6.4 that Covid-19 has a shock size of $u \in [0, 0.16]$.

For different shock sizes u, we see how the interest rates have to be modified to get the same price for the annuity for women with $r = 3\%$. We observe that $\tilde{r}$ declines faster with higher age.

u	$\tilde{r}$		
	$x = 40$	$x = 60$	$x = 80$
0	3.00%	3.00%	3.00%
0.2	2.88%	2.60%	1.22%
0.4	2.75%	2.15%	−0.85%
0.6	2.60%	1.63%	−3.27%

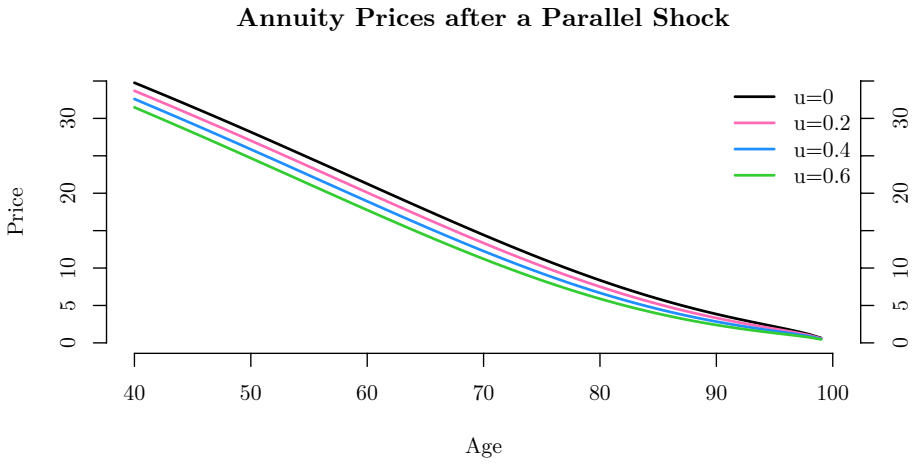

Figure 6.8: For female, for age 40 to 100, we see the annuity prices after a parallel shocks of size $u = 0, 0.2, 0.4, 0.6$ with an interest rate of r=1%.

Secondly, we want to compare the annuity prices before Covid-19 and for the adjusted mortality rates under the assumption that the shock is permanent. We again have to assume that these mortality rates will not change in the future. In Figure 6.9, we see the size of the relative decline of the prices for annuities, separately for male and female. We observe that the drop for men is more significant than for women. Further, we notice that annuities for older people are relatively more affected.

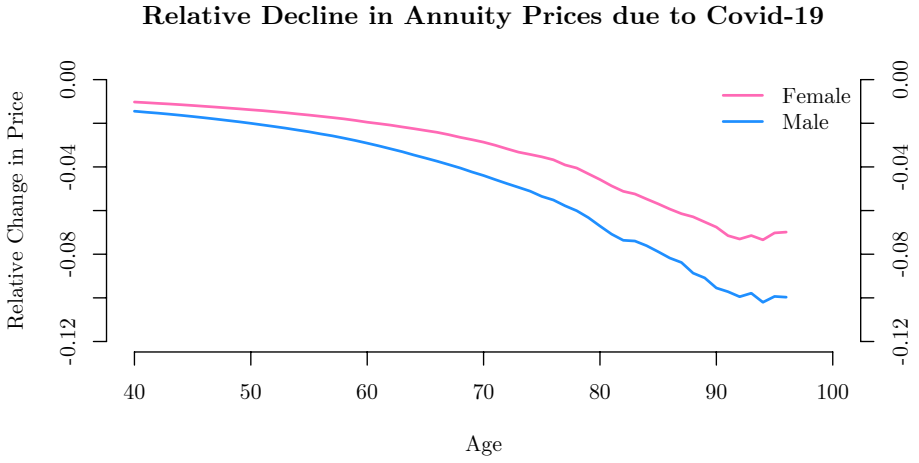

Figure 6.9: For the age 40 to 100, separately for male and female, we see the relative decline in annuity prices because of Covid-19 with an interest rate of r=1%.

Additionally, we see how the interest rates need to be modified after the Covid-19

shock, so that there is no change in annuity prices. Starting with $r = 3\%$, the new interest rates are calculated for several ages and separately for male and female using the function `uniroot()` from the package `stats` in R. We notice that the interest rate has to be cut more for men than for women and that it decreases faster with higher ages, what we already observed above.

x	40	50	60	70	80
$\tilde{r}^{(f)}$	2.96%	2.94%	2.87%	2.72%	2.27%
$\tilde{r}^{(m)}$	2.94%	2.90%	2.79%	2.52%	1.79%

7 Open Questions

We want to close the thesis with an outlook on further questions concerning Covid-19. As there are way more important problems can be addressed in this thesis, we want to encourage the interested reader to further research on open questions. However, this chapter does not claim to be complete.

In Section 2.1, we stated that the epidemiological time frames do not have to be equal for each age and sex. If enough data for infection pairs is available, one can estimate these distributions separately and test whether there are differences. Further, it would be interesting to examine the impact of certain scenarios, for example a completely lockdown or mask requirements, on the serial interval.

From the list of assumptions in Section 3.2, we have not discussed the a-priori distribution of the Bayes approach (assumption (6)). Besides, if the necessary data is available, we need to estimate the variance of the generated cases by one infector in order to show that the Poisson based approach from [23] is inappropriate and to justify the negative binomial approach from Section 3.3. Analogous to the serial interval, one can try to determine the impact of Covid-19 measures like lockdowns or mask requirements on the reproduction number to evaluate the costs and benefits of these measure objectively in order to find the most efficient.

With respect to the role of tests, described in Section 3.5, it would be interesting to examine how the number of tests is related to the number of infections, under the assumption that people are tested according to the probability of infection. If we estimate a function for this relation, especially the slope and curvature will be important to answer questions like: What is the optimal number of tests (if we have an underlying utility function)? Another point that we do not discuss in this thesis is the role of wrong-positive and wrong-negative tests and how long a test result should be valid.

For the Galton–Watson process, there are further possibilities to extend the model or to calculate properties as well. Firstly, one can assume a stochastic starting value Y_0 and discuss the consequences for the presented theorems. Secondly, one can study the expected time to extinction and calculate it explicitly or at least find bounds. Further, it might be interesting how it depends on the mean of the offspring distribution. So, in the setting of a pandemic, we would like to answer the question: How long does it take to defeat the pandemic given a certain reproduction number? Then, given all necessary parameters, the duration and amount of contact reduction can be optimized. Thirdly, one can incorporate the possibility of immigration. So, the new

process would be defined as

$$Y_{n+1} := \sum_{i=1}^{Y_n} X_i^n + J_{n+1},$$

where J_{n+1} denotes a random variable describing the number of imported cases in generation $n+1$. Analogous to Section 4.1, one can calculate some basic properties of this process like the mean, variance, covariance or the probability generating function. As it reflects the risk of imported cases, this extension is also a resumption of Section 3.6.

For the country comparison in Section 5.4, it would also be interesting to study the impact of test capacities and lockdown policies on the case fatality rate. Besides, in Chapter 6, one can discuss the impact on other insurance products like life insurance or health insurance. Moreover, there are a lot of economical implications due to Covid-19 from other business sectors that can be further studied. However, there are of course also a lot of scientific problems from other areas, for example biological questions regarding immunity or asymptomatic transmission probability. Another interesting topic is the dynamic of viral loads indoor and outdoor: What is the risk of infection indoor dependent on the volume of the room and the amount of viral particles submitted from an infected person? And how much time does a room need to be virus-free again?

A topic that is completely left out in this thesis is vaccination. As there are many questions that want to be answered, we only outline some short ideas. Firstly, one can take a look at the vaccination protection dependent on the time since complete vaccination. Secondly, one can calculate the optimal duration between the two vaccination parts when vaccines are in short supply. That depends, among other things, on future supply and on the vaccination protection after the first and after the second vaccination part. Thirdly, as also vaccinated people can be infected, it is possible to calculate all the epidemiological key figures described in this thesis for the subset of vaccinated and unvaccinated people respectively. Fourthly, all topics above can be analysed and compared for different vaccines or virus variants.

I hope that this thesis can be a starting point for further research and critical analysis of models and underlying assumptions.

Appendix: R Codes

In this chapter, you can find exemplary R codes, which were used to process the data, implement the estimations, calculate the results and generate the plots. All codes were written and run under R version 3.6.3. The following packages were used apart from the standard packages:

- We used `readxl` and `lubridate` to read and process data.
- We used `fitdistrplus` and `sn` to work with distributions.
- We used `extrafont`, `xtable`, `plotly` and `rworldmap` to generate plots and tables.

Estimation of the Serial Interval

```
library(readxl)
library(sn)

data <- read_excel("data468.xlsx", range = "Q2:Q470")$diff

# Maximum Likelihood
start.sn <- c(-1,7,3)
par.sn.mle <- cp2dp(sn.mple(y=data, cp=dp2cp(start.sn, "SN"))$cp, "SN")

# Method of Moments
m1 <- mean(data)
m2 <- var(data)
m3 <- mean((data-m1)^3)

delta <- uniroot(function(d) (4-pi)/2*(d*sqrt(2/pi))^3/(1-2*d^2/pi)^(3/2)-m3/m2^(3/
    2), interval=c(-1,1))$root
alpha <- delta/sqrt(1-delta^2)
omega <- uniroot(function(w) w^2*(1-2*delta^2/pi)-m2, interval=c(0,100))$root
xsi <- uniroot(function(e) e+omega*delta*sqrt(2/pi)-m1, interval=c(-10,10))$root

# Basic Properties

sort(data)[length(data)*0.025]
sort(data)[length(data)*0.975]
qsn(c(0.025,0.975), xsi, omega, alpha)
qsn(c(0.025,0.975), par.sn.mle[1], par.sn.mle[2], par.sn.mle[3])

par.sn.mle[1]+par.sn.mle[2]*par.sn.mle[3]/sqrt(1+par.sn.mle[3]^2)*sqrt(2/pi)
(par.sn.mle[2]*sqrt(1-2/pi*par.sn.mle[3]^2/(1+par.sn.mle[3]^2)))^2

# Discussion of AGES Methodology

```

```
qgamma(c(0.025,0.975), 2.88, scale=1.55)
qnorm(c(0.025,0.975), 3.96, sqrt(22.56))
qlnorm(c(0.025,0.975), log(4.7)-log(1+2.9^2/4.7^2)/2, sqrt(log(1+2.9^2/4.7^2)))
```

Replication of the Official Reproduction Number

```
download.file("https://info.gesundheitsministerium.at/data/data.zip",
              "data.zip", mode = "wb") # downloaded on Jan 04, 2021

download.file("https://www.ages.at/fileadmin/AGES2015/Wissen-Aktuell/COVID19/R_eff.csv",
              "R_ages.csv", mode = "wb") # downloaded on Jan 04, 2021

R.off <- read.csv2("R_ages.csv") # Official Reproduction Number
posTest <- read.csv2(unz("data.zip", "Epikurve.csv"))[,1:2]

n <- length(posTest[,1])
y <- posTest[,2]

a <- 1
b <- 5
tau <- 13
m <- 21

w0 <- dgamma(1:m, shape=2.88, scale=1.55)
w <- w0/sum(w0)

d <- 1:n*NA
for(i in 2:n) {
  d[i] <- sum(w[1:min(i-1,m)] * y[(i-1):(max(1,i-m))])
}

R.rep <- R.rep.lq <- R.rep.uq <- 1:n*NA # Replication of the Reproduction Number & Bounds
for(t in 13:n) {
  R.rep[t] <- (a+sum(y[(t-tau+1):t])) / (1/b+sum(d[(t-tau+1):t]))
  R.rep.lq[t] <- qgamma(0.025, shape=(a+sum(y[(t-tau+1):t])), scale=1/(1/b+sum(d[(t-tau+1):t])))
  R.rep.uq[t] <- qgamma(0.975, shape=(a+sum(y[(t-tau+1):t])), scale=1/(1/b+sum(d[(t-tau+1):t])))
}
```

Galton–Watson Process

```
## Simulation

X0 <- 100
R0 <- c(0.9,1,1.1)
r <- 0.1
N <- 30 # Number of Simulations
M <- 51 # Length of Simulations

MC1 <- MC2 <- MC3 <- matrix(NA, ncol=N, nrow=M)
MC1[1,] <- MC2[1,] <- MC3[1,] <- X0
for(n in 1:N) {
```

```
  for(m in 2:M) {
    MC1[m,n] <- sum(rnbinom(MC1[m-1,n], size=r, mu=R0[1]))
    MC2[m,n] <- sum(rnbinom(MC2[m-1,n], size=r, mu=R0[2]))
    MC3[m,n] <- sum(rnbinom(MC3[m-1,n], size=r, mu=R0[3]))
  }
}

## Extinction Probability for NB

library(plotly)

m1 <- 0:50/10 # mean
m2 <- 0:400/10 # variance

f <- function(t,r,p) ((1-p)/(1-t*p))^t(r)-t
g <- function(t,r,p) ((1-p)/(1-t*p))^r

q <- matrix(NA, nrow=length(m1), ncol=length(m2))
for(i in 2:length(m1)) {
  for(j in 2:length(m2)) {
    a <- m1[i]^2/(m2[j]-m1[i])
    p <- 1-m1[i]/m2[j]
    if(a>=0 & p>=0) {
      count <- 0
      x <- 0
      while(abs(f(x,a,p))>10^-7) {
        x <- g(x,a,p)
        count <- count+1
      }
      q[i,j] <- x
    }
  }
}

plot_ly(x=m1, y = m2, z = t(q)) %>% add_heatmap()
```

Case Fatality Rate

```
download.file("https://info.gesundheitsministerium.at/data/data.zip",
              "data.zip", mode = "wb") # downloaded on Jan, 4th 2021

posTest <- read.csv2(unz("data.zip", "Epikurve.csv"))[,1:2]
genesen.kum <- read.csv2(unz("data.zip", "GenesenTimeline.csv"))[,1:2]
tod.kum <- read.csv2(unz("data.zip", "TodesfaelleTimeline.csv"))[,1:2]

n <- min(c(length(posTest[,1]), length(genesen.kum[,1]), length(tod.kum[,1])))

M <- cbind(posTest[1:n,], cumsum(posTest[1:n,2]), genesen.kum[1:n,2], tod.kum[1:n
    ,2])
M <- cbind(M, M[,3]-M[,4]-M[,5])
names(M) <- c("date", "posTest", "posTest.kum", "genesen.kum", "tod.kum", "krank")

M[,1] <- as.Date(M[,1], "%d.%m.%y")

## Estimation Approaches

library(readxl)
plattform <- read_excel("../../Datenplattform/Datenplattform_Antrag_37_#3.xlsx",
```

```
                              col_types = c("numeric", "date", "date", "date", "date", "
                                  date", "date", "text", "date", "date", "text"))
plattform$TodesDatum <- as.POSIXct(plattform$TodesDatum, tz="UTC")

max(plattform$DatumGeheilt, na.rm=TRUE)

plattform.d <- plattform[!is.na(plattform$TodesDatum),]

InfToDeath.rh <- matrix(NA, nrow=n, ncol=61)
for(j in 18:(n-1)) {
  plattform.d1 <- plattform.d[as.Date(plattform.d$TodesDatum)<=as.Date("2020-02-25")
      +j,]

  InfToDeath <- as.numeric(na.omit((plattform.d1$TodesDatum-plattform.d1$
      DiagnoseDatum)/60/60/24))
  InfToDeath.ah <- 0:60*NA
  for(i in 0:60) {
    InfToDeath.ah[i+1] <- sum(InfToDeath==i)
  }
  InfToDeath.rh[j+1,] <- InfToDeath.ah/sum(InfToDeath.ah)
  print(length(plattform.d1$TodesDatum)-sum(InfToDeath.ah)) #excluded
}
plot(0:60,InfToDeath.rh[n,], type="h")

plattform.r <- plattform[!is.na(plattform$DatumGeheilt),]
InfToRecover <- as.numeric(na.omit((plattform.r$DatumGeheilt-plattform.r$
    DiagnoseDatum)/60/60/24))
InfToRecover.ah <- 0:60*NA
for(i in 0:60) {
  InfToRecover.ah[i+1] <- sum(InfToRecover==i)
}
InfToRecover.rh <- InfToRecover.ah/sum(InfToRecover.ah)
plot(0:60,InfToRecover.rh, type="h")
length(plattform.r$DatumGeheilt)-sum(InfToRecover.ah) #excluded

# Different SD
sd(InfToDeath)
sd(InfToRecover)

InfToRecover[InfToRecover>50]

plattform$DatumGeheilt[max(plattform$DatumGeheilt, na.rm=TRUE)]

# Bounds
lowerbound <- M$tod.kum/M$posTest.kum
upperbound <- (M$tod.kum+M$krank)/M$posTest.kum

# Estimate 1
est1 <- M$tod.kum/(M$genesen.kum+M$tod.kum)

# Estimate 2
Y <- 1:n*NA
for(i in 18:n) {
  pk <- cumsum(InfToDeath.rh[i,])
  Y[i] <- sum(M$posTest[1:i] * c(rep(1,max(0,i-61)),pk[min(i,61):1]))
}
est2 <- M$tod.kum/Y

## Course of the Pandemic

CFR.course1 <- CFR.course2 <- CFR.course3 <- 1:n*NA
for(j in 18:(n-1)) {
```

```
  plattform1 <- plattform[as.Date(plattform$DiagnoseDatum)<=as.Date("2020-03-01")+j
      & as.Date(plattform$DiagnoseDatum)>as.Date("2020-03-01")+j-7,]
  CFR.course1[j] <- sum(!is.na(plattform1$TodesDatum))/sum(!is.na(plattform1$
      DatumGeheilt) | !is.na(plattform1$TodesDatum))

  plattform1 <- plattform[(as.Date(plattform$DatumGeheilt)<=as.Date("2020-03-01")+j
      | as.Date(plattform$TodesDatum)<=as.Date("2020-03-01")+j) & (as.Date(plattform
      $DatumGeheilt)>as.Date("2020-03-01")+j-7 | as.Date(plattform$TodesDatum)>as.
      Date("2020-03-01")+j-7),]
  CFR.course2[j] <- sum(!is.na(plattform1$TodesDatum))/sum(!is.na(plattform1$
      DatumGeheilt) | !is.na(plattform1$TodesDatum))

  plattform1 <- plattform[( as.Date(plattform$DatumGeheilt)>=as.Date("2020-03-01")+j
      | as.Date(plattform$TodesDatum)>=as.Date("2020-03-01")+j ) & as.Date(
      plattform$DiagnoseDatum)<=as.Date("2020-03-01")+j,]
  CFR.course3[j] <- sum(!is.na(plattform1$TodesDatum))/sum(!is.na(plattform1$
      DatumGeheilt) | !is.na(plattform1$TodesDatum))
}

## Sex and Age

t.test(as.numeric(!is.na(plattform.m1$TodesDatum)), as.numeric(!is.na(plattform.f1$
    TodesDatum)), alternative="greater")

plattform.x <- plattform[!is.na(plattform$TodesDatum) | !is.na(plattform$
    DatumGeheilt),]

CFR.age <- CFR.age.m <- CFR.age.f <- 1:100*NA
for(i in 1:100) {
  CFR.age[i] <- mean(as.numeric(!is.na(plattform.x$TodesDatum[plattform.x$Alter==i])
      ))
  CFR.age.m[i] <- mean(as.numeric(!is.na(plattform.x$TodesDatum[plattform.x$Alter==i
      & plattform.x$Geschl=="M"])))
  CFR.age.f[i] <- mean(as.numeric(!is.na(plattform.x$TodesDatum[plattform.x$Alter==i
      & plattform.x$Geschl=="W"])))
}

altersschnitt <- 1:n*NA
for(i in 25:n) {
  altersschnitt[i] <- mean(plattform.d$Alter[as.Date(plattform.d$TodesDatum)<=as.
      Date("2020-03-01")+i & as.Date(plattform.d$TodesDatum)>as.Date("2020-03-01")+i
      -14])
}
```

Mortality Shock

```
library(readxl)
library(lubridate)

data <- read_excel("sterblichkeit.xlsx", col_names = FALSE, skip = 4)
names(data) <- c("age", "q_m", "q_f", "q", "n_m", "n_f", "n")

plattform <- read_excel("../../Datenplattform/2021_02_01_Antrag_37_#4.xlsx",
                        col_types = c("text", "date", "date", "date", "date", "date
                            ", "date", "date", "numeric", "date"))

data$ci_m <- data$ci_f <- data$cd_m <- data$cd_f <- data$age*NA
for(i in 1:length(data$age)) {
  data$ci_m[i] <- sum(plattform$Alter==data$age[i] & plattform$Geschl=="M")
```

```
13    data$ci_f[i] <- sum(plattform$Alter==data$age[i] & plattform$Geschl=="W")
14    data$cd_m[i] <- sum(plattform$Alter==data$age[i] & plattform$Geschl=="M" & !is.na(
          plattform$TodesDatum))
15    data$cd_f[i] <- sum(plattform$Alter==data$age[i] & plattform$Geschl=="W" & !is.na(
          plattform$TodesDatum))
16  }
17  data$ci <- data$ci_m+data$ci_f
18  data$cd <- data$cd_m+data$cd_f
19
20  ## Estimation of Gompertz
21
22  qx.gompertz <- function(x,g,h, lambda=0) {
23    1-exp(-lambda+h*exp(g*x)*(1-exp(g))/g)
24  }
25
26  mux.gompertz <- function(x,g,h) {
27    h*exp(g*x)
28  }
29
30  mux2.gompertz <- function(x,g,h) {
31    log(h)+g*x
32  }
33
34  ols <- function(gh) {
35    sum((data$q[35:99+1]-qx.gompertz(35:99,gh[1],gh[2]))^2)
36  }
37  gh.opt <- optim(c(0.1,0.00001), ols)$par
38
39  mle.gompertz.poisson.f <- function(par, lambda=0.0000) { #c(lambda, g, h) male
        :10^-5, female:2*10^-5
40    -sum(data$n_f[35:99+1]*(-1+exp(-lambda-par[2]*exp(par[1]*35:99)*(exp(par[1])-1)/
          par[1])+data$q_f[35:99+1]*log(data$n_f[35:99+1]*(1-exp((-lambda-par[2]*exp(par
          [1]*35:99)*(exp(par[1])-1)/par[1])))))
41  }
42  mle.gompertz.binom.f <- function(par, lambda=0.000) { #c(lambda, g, h) male:10^-5,
        female:2*10^-5
43    -sum(data$n_f[35:99+1]*((1-data$q_f[35:99+1])*(-lambda-par[2]*exp(par[1]*35:99)*(
          exp(par[1])-1)/par[1])+data$q_f[35:99+1]*log(1-exp((-lambda-par[2]*exp(par[1]*
          35:99)*(exp(par[1])-1)/par[1])))))
44  }
45  mle.gompertz.poisson.m <- function(par, lambda=0.0000) { #c(lambda, g, h) male
        :10^-5, female:2*10^-5
46    -sum(data$n_m[35:99+1]*(-1+exp(-lambda-par[2]*exp(par[1]*35:99)*(exp(par[1])-1)/
          par[1])+data$q_m[35:99+1]*log(data$n_m[35:99+1]*(1-exp((-lambda-par[2]*exp(par
          [1]*35:99)*(exp(par[1])-1)/par[1])))))
47  }
48  mle.gompertz.binom.m <- function(par, lambda=0.000) { #c(lambda, g, h) male:10^-5,
        female:2*10^-5
49    -sum(data$n_m[35:99+1]*((1-data$q_m[35:99+1])*(-lambda-par[2]*exp(par[1]*35:99)*(
          exp(par[1])-1)/par[1])+data$q_m[35:99+1]*log(1-exp((-lambda-par[2]*exp(par[1]*
          35:99)*(exp(par[1])-1)/par[1])))))
50  }
51
52  gh.m <- optim(c(0.1,0.00001), mle.gompertz.poisson.m)$par
53  gh.f <- optim(c(0.1,0.00001), mle.gompertz.poisson.f)$par
54
55  ## Parallel & Constant Shock
56
57  v <- 1
58  x0 <- 99
59  # daphie4eva
60
61  ## Mortality Risk-adjusted Age & Excess Factor
```

```
62
63 t <- as.numeric(as.Date("2021-01-29")-as.Date("2020-02-28"))
64
65 mux.f <- -log(1-data$q_f)[0:99+1]
66 mux.m <- -log(1-data$q_m)[0:99+1]
67 mux.f.C <- mux.m.C <- 0:99*NA
68 for(i in 0:99+1) {
69   mux.f.C[i] <- uniroot(function(x) (1-exp(-t/365*(x+mux.f[i])))*x/(x+mux.f[i])-data
       $cd_f[i]/data$n_f[i], c(0,1), tol=10^-12)$root
70   mux.m.C[i] <- uniroot(function(x) (1-exp(-t/365*(x+mux.m[i])))*x/(x+mux.m[i])-data
       $cd_m[i]/data$n_m[i], c(0,1), tol=10^-12)$root
71 }
72
73 adj.age.f <- adj.age.m <- 0:99*NA
74 for(i in 0:99+1) {
75   adj.age.f[i] <- (mux.f[i]+mux.f.C[i]-mux.f[i])/(mux.f[i+1]-mux.f[i])
76   adj.age.m[i] <- (mux.m[i]+mux.m.C[i]-mux.m[i])/(mux.m[i+1]-mux.m[i])
77 }
78
79 ux.m <- log((mux.m+mux.m.C)/mux.m)
80 ux.f <- log((mux.f+mux.f.C)/mux.f)
81
82 ## Shock at Infection
83
84 mux.tilde.f <- mux.f+mux.f.C
85 mux.tilde.m <- mux.m+mux.m.C
86 mux.f.I <- mux.m.I <- 0:99*NA
87 for(i in 0:99+1) {
88   mux.f.I[i] <- uniroot(function(x) (1-exp(-t/365*(x+mux.tilde.f[i])))*x/(x+mux.
       tilde.f[i])-data$cd_f[i]/data$ci_f[i], c(0,1), tol=10^-12)$root
89   mux.m.I[i] <- uniroot(function(x) (1-exp(-t/365*(x+mux.tilde.m[i])))*x/(x+mux.
       tilde.m[i])-data$cd_m[i]/data$ci_m[i], c(0,1), tol=10^-12)$root
90 }
91
92 adj.age.I.f <- adj.age.I.m <- 0:99*NA
93 for(i in 0:99+1) {
94   adj.age.I.f[i] <- which.min(abs(mux.f.I[i]-mux.tilde.f[i:100]))-1
95   adj.age.I.m[i] <- which.min(abs(mux.m.I[i]-mux.tilde.m[i:100]))-1
96 }
97
98 ux.I.m <- log((mux.tilde.m+mux.m.I)/mux.tilde.m)
99 ux.I.f <- log((mux.tilde.f+mux.f.I)/mux.tilde.f)
100
101 ## Life Expectancy
102
103 D <- matrix(NA, ncol=100, nrow=366) #d_{x,t} Tote pro Tag pro Alter
104
105 plattform.d <- plattform[!is.na(plattform$TodesDatum),]
106
107 for(i in 1:366) {
108   plattform.day <- plattform.d[as.Date(plattform.d$TodesDatum)==as.Date("2020-03-01"
       )+i-1,]
109   for(j in 1:100) {
110     D[i,j] <- sum(plattform.day$Alter==data$age[j]) # & plattform$Geschl=="M"
111   }
112 }
113
114 e <- 0:99*NA
115 for(i in 0:99) {
116   kpx <- 1:99*0
117   for(j in 1:(99-i)) {
118     kpx[j] <- prod(1-data$q[i:(i+j-1)+1])
119   }
```

```
  e[i+1] <- sum(kpx)
}
L <- as.vector(D%*%e)

## Annuity Prices

tpx <- function(t,x,g,h,u) {
  exp(-h*exp(u)*exp(g*x)*(exp(g*t)-1)/g)
}

AP.ps <- function(u,r,g,h) {
  ap <- 1:100*NA
  for(x in 1:99) {
    ap[x] <- sum((1/(1+r))^(1:(100-x)) * tpx(1:(100-x),x,g,h,u))
  }
  ap
}

AP.real <- function(r,q) {
  ap <- 1:100*NA
  for(x in 1:99) {
    ap[x] <- sum((1/(1+r))^(1:(100-x)) * cumprod(1-q[x:99]))
  }
  ap
}

qx.C.f <- 1-exp(-(mux.tilde.f))
qx.C.m <- 1-exp(-(mux.tilde.m))

r.ps <- matrix(NA, ncol=3, nrow=4)
for(i in 1:4) {
  for(j in 1:3) {
    r.ps[i,j] <- uniroot(function(x) AP.ps(0,0.03,gh.f[1], gh.f[2])[20*(j+1)]-AP.ps
        (0.2*(i-1),x,gh.f[1], gh.f[2])[20*(j+1)], c(-1,1), tol=10^-10)$root
  }
}
r.ps*100

r.covid.f <- r.covid.m <- 1:5*NA
for(i in 1:5) {
  r.covid.f[i] <- uniroot(function(x) AP.real(0.03,data$q_f)[30+10*i]-AP.real(x,qx.C
      .f)[30+10*i], c(-1,1), tol=10^-10)$root
  r.covid.m[i] <- uniroot(function(x) AP.real(0.03,data$q_m)[30+10*i]-AP.real(x,qx.C
      .m)[30+10*i], c(-1,1), tol=10^-10)$root
}
r.covid.f*100
r.covid.m*100
```

List of Figures

Conventions & Abbreviations

Conventions

- $\mathbb{N} = \{0, 1, 2, \dots\}$ are the natural numbers.
- $\mathbb{N}^* = \mathbb{N} \backslash \{0\}$ are the natural numbers without 0.
- The empty sum is 0.
- The empty product is 1.
- $0^0 = 1$
- $\log(\cdot)$ denotes the natural logarithm (with basis e).
- The subset relation "$\subset$" includes the equality of both sets.

Abbreviations

Bibliography

[1] AGES, *Dashboard COVID19*, https://info.gesundheitsministerium.at/data/data.zip, January 2021. 23

[2] Gerold Alsmeyer, *Part I, The simple Galton-Watson process: Classical approach*, Uni Münster, https://www.uni-muenster.de/Stochastik/lehre/WS1011/SpezielleStochastischeProzesse/Ch_1.pdf, January 2011. 45, 46

[3] Samir K. Ashour and Mahmood A. Abdel-hameed, *Approximate skew normal distribution*, Journal of Advanced Research **1** (2010), no. 4, 341–350. 8

[4] Austrian National Public Health Institute (Gesundheit Österreich GmbH), *Datenplattform Covid-19*, https://datenplattform-covid.goeg.at/, December 2020. 55, 57, 58, 63, 67

[5] Andrew J.G. Cairns, David Blake, Amy R. Kessler, and Marsha Kessler, *The Impact of Covid-19 on Future Higher-Age Mortality*, http://www.pensions-institute.org/wp-content/uploads/wp2007.pdf, May 2020. 70, 72

[6] Zhanwei Du, Xiaoke Xu, Ye Wu, Lin Wang, Benjamin Cowling, and Lauren Ancel Meyers, *Serial Interval of COVID-19 among Publicly Reported Confirmed Cases*, Emerging Infectious Disease journal **26** (2020), no. 6, 1341. 3, 6, 7, 9, 16, 17, 34

[7] Keisuke Ejima, Kwang Su Kim, Christina Ludema, Ana I. Bento, Shoya Iwanami, Yasuhisa Fujita, Hirofumi Ohashi, Yoshiki Koizumi, Koichi Watashi, Kazuyuki Aihara, Hiroshi Nishiura, and Shingo Iwami, *Estimation of the incubation period of COVID-19 using viral load data*, Epidemics **35** (2021), 100454. 12

[8] Guihong Fan, Zhichun Yang, Qianying Lin, Shi Zhao, Lin Yang, and Daihai He, *Decreased Case Fatality Rate of COVID-19 in the Second Wave: A study in 53 countries or regions*, Transboundary and Emerging Diseases **68** (2020). 57, 58

[9] Seth Flaxman, Swapnil Mishra, Axel Gandy, H. Juliette T. Unwin, Thomas A. Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W. Eaton, Mélodie Monod, Pablo N. Perez-Guzman, Nora Schmit, Lucia Cilloni, Kylie E. C. Ainslie, Marc Baguelin, Adhiratha Boonyasiri, Olivia Boyd, Lorenzo Cattarino, Laura V. Cooper, Zulma Cucunubá, Gina Cuomo-Dannenburg, Amy Dighe, Bimandra Djaafara, Ilaria Dorigatti, Sabine L. van

Elsland, Richard G. FitzJohn, Katy A. M. Gaythorpe, Lily Geidelberg, Nicholas C. Grassly, William D. Green, Timothy Hallett, Arran Hamlet, Wes Hinsley, Ben Jeffrey, Edward Knock, Daniel J. Laydon, Gemma Nedjati-Gilani, Pierre Nouvellet, Kris V. Parag, Igor Siveroni, Hayley A. Thompson, Robert Verity, Erik Volz, Caroline E. Walters, Haowei Wang, Yuanrong Wang, Oliver J. Watson, Peter Winskill, Xiaoyue Xi, Patrick G. T. Walker, Azra C. Ghani, Christl A. Donnelly, Steven Riley, Michaela A. C. Vollmer, Neil M. Ferguson, Lucy C. Okell, Samir Bhatt, and Imperial College C. O. V. I. D.-19 Response Team, *Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe*, Nature **584** (2020), no. 7820, 257–261. 21, 32

[10] Tapiwa Ganyani, Cécile Kremer, Dongxuan Chen, Andrea Torneri, Christel Faes, Jacco Wallinga, and Niel Hens, *Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020*, Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin **25** (2020), no. 32372755, 2000257 (eng). 14, 16

[11] Daihai He, Shi Zhao, Xiaoke Xu, Qiangying Lin, Zian Zhuang, Peihua Cao, Maggie H. Wang, Yijun Lou, Li Xiao, Ye Wu, and Lin Yang, *Low dispersion in the infectiousness of COVID-19 cases implies difficulty in control*, BMC Public Health **20** (2020), no. 1, 1558. 28

[12] Xi He, Eric H. Y. Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y. Wong, Yujuan Guan, Xinghua Tan, Xiaoneng Mo, Yanqing Chen, Baolin Liao, Weilie Chen, Fengyu Hu, Qing Zhang, Mingqiu Zhong, Yanrong Wu, Lingzhai Zhao, Fuchun Zhang, Benjamin J. Cowling, Fang Li, and Gabriel M. Leung, *Temporal dynamics in viral shedding and transmissibility of COVID-19*, Nature Medicine **26** (2020), no. 5, 672–675. 3, 6, 7, 9, 10, 12, 14, 15, 16

[13] Jürgen Hedderich and Lothar Sachs, *Angewandte Statistik : Methodensammlung mit R*, 15., überarbeitete und erweiterte auflage. ed., Springer Spektrum, Berlin, 2016 (ger). 6, 7

[14] Christof Kuhbandner, *The Scenario of a Pandemic Spread of the Coronavirus SARS-CoV-2 is Based on a Statistical Fallacy*, April 2020. 30, 31

[15] Anthony Kyriakopoulos and Shi Zhao, *A computational analysis on Covid-19 transmission raises immuno-epidemiology concerns*, (2020). 28

[16] Shujuan Ma, Jiayue Zhang, Minyan Zeng, Qingping Yun, Wei Guo, Yixiang Zheng, Shi Zhao, Maggie H. Wang, and Zuyao Yang, *Epidemiological parameters of coronavirus disease 2019: a pooled analysis of publicly reported individual data of 1155 cases from seven countries*, medRxiv (2020), 2020.03.21.20040329. 6, 12

[17] Moshe A. Milevsky, *Is Covid-19 a Parallel Shockto the Term Structure of Mortality?*, https://moshemilevsky.com/wp-content/uploads/2020/05/MILEVSKY_20MAY2020_AMAZON.pdf, May 2020. 63

[18] Hiroshi Nishiura, Natalie M. Linton, and Andrei R. Akhmetzhanov, *Serial interval of novel coronavirus (COVID-19) infections*, International Journal of Infectious Diseases **93** (2020), 284–286. 6, 15, 16, 33

[19] World Health Organization, *Coronavirus disease (COVID-19), Situation Report - 102*, https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200501-covid-19-sitrep.pdf, November 2020. 62

[20] Our World in Data, *Covid-19 dataset*, https://github.com/owid/covid-19-data/tree/master/public/data, July 2020. 62

[21] Miquel Porta, *A Dictionary of Epidemiology*, vol. Sixth edition, Oxford University Press, Oxford, 2014 (English). 19

[22] Lukas Richter, Daniela Schmid, Ali Chakeri, Sabine Maritschnik, Sabine Pfeiffer, and Ernst Stadlober, *Schätzung des seriellen Intervalles von COVID19, Österreich*, AGES, https://www.ages.at/download/0/0/068cb5fb9f2256d267e1a3dc8d464623760fcc30/fileadmin/AGES2015/Wissen-Aktuell/COVID19/Sch%C3%A4tzung_des_seriellen_Intervalles_von_COVID19_2020-04-08.pdf, June 2020. 3, 6, 10, 16, 23, 33

[23] Lukas Richter, Daniela Schmid, and Ernst Stadlober, *Methodenbeschreibung für die Schätzung von epidemiologischen Parametern des COVID19 Ausbruchs, Österreich*, AGES, https://www.ages.at/download/0/0/e03842347d92e5922e76993df9ac8e9b28635caa/fileadmin/AGES2015/Wissen-Aktuell/COVID19/Methoden_zur_Sch%C3%A4tzung_der_epi_Parameter.pdf, June 2020. 20, 21, 23, 26, 33, 39, 77

[24] RKI, *Covid-19 Data*, https://prod-hub-indexer.s3.amazonaws.com/files/dd4580c810204019a7b8eb3e0b329dd6/0/full/4326/dd4580c810204019a7b8eb3e0b329dd6_0_full_4326.csv, August 2020. 37

[25] Xinmiao Rong, Liu Yang, Huidi Chu, and Meng Fan, *Effect of delay in diagnosis on transmission of COVID-19*, Mathematical Biosciences and Engineering **17** (2020), 2725–2740. 37

[26] Statistik Austria, *Bevölkerung am 1.1.2020 nach Alter und Bundesland - Frauen*, https://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/bevoelkerung/bevoelkerungsstruktur/bevoelkerung_nach_alter_geschlecht/index.html, February 2021. 63, 67

[27] _____, *Bevölkerung am 1.1.2020 nach Alter und Bundesland - Männer*, https://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/bevoelkerung/bevoelkerungsstruktur/bevoelkerung_nach_alter_geschlecht/index.html, February 2021. 63, 67

[28] _____, *Leibrententafel: Barwerte einer lebenslang vorschüssigen Rente vom Betrag 1 nach der Sterbetafel 2010/2012*, https://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/bevoelkerung/sterbetafeln/index.html, February 2021. 63, 64, 65, 67

[29] R. N. Thompson, J. E. Stockwin, R. D. van Gaalen, J. A. Polonsky, Z. N. Kamvar, P. A. Demarsh, E. Dahlqwist, S. Li, E. Miguel, T. Jombart, J. Lessler, S. Cauchemez, and A. Cori, *Improved inference of time-varying reproduction numbers during infectious disease outbreaks*, Epidemics **29** (2019), 100356. 26, 29, 30, 32, 35

[30] Lauren C. Tindale, Jessica E. Stockdale, Michelle Coombe, Emma S. Garlock, Wing Yin Venus Lau, Manu Saraswat, Louxin Zhang, Dongxuan Chen, Jacco Wallinga, Caroline Colijn, Eduardo Franco, Marc Lipsitch, Marc Lipsitch, Joel Miller, and Virginia E. Pitzer, *Evidence for transmission of COVID-19 prior to symptom onset*, eLife **9** (2020), e57149. 9, 12, 15, 16, 21

[31] N.A. Weiss, P.T. Holmes, and M. Hardy, *A Course in Probability*, Pearson Addison Wesley, 2005. 54

[32] David Williams, *Probability with martingales*, 9. print.. ed., Cambridge mathematical textbooks, Cambridge Univ. Press, Cambridge [u.a.], 2005 (eng). 44, 45

[33] Zhigang Zhang and Jianguo Sun, *Interval censoring*, Statistical methods in medical research **19** (2010), no. 19654168, 53–70 (eng). 11