University of Mannheim
Chair of Econometrics
Prof. Dr. Christoph Rothe

Bachelor Thesis

Estimation of Covid-19 Death Probabilities in German Regions and Evaluating future Results using Bayesian Statistics

Shreyas Prakash Gupta
B2 6-7
68259, Mannheim
Matr.-Nr. 1636008

Date of submission: 03.08.2022

Contents

List of Figures **1**

1 Introduction **2**

2 Literature Review **3**

3 Data **4**

4 Bayesian Models and Estimation **5**

4.1 Introduction to Bayesian Estimation 5

4.2 Choice of Prior 6

5 Estimation of death probabilities in German Counties **7**

5.1 Procedure of Estimation 7

5.1.1 Estimation Analysis and Concerns 10

5.2 Analysis of Age-Group Death Estimation 12

5.3 Different Priors and Likelihoods - Metropolis-Hastings Algorithm . . 13

5.3.1 Modeling with different Likelihoods: Poisson 16

5.4 Posterior quantiles and intervals - Credibility Interval 18

5.5 Comparison of Bayes Estimate, MH Algorithm (and Bootstrap) . . . 19

5.5.1 Excursion: Comparison Bootstrapping 21

5.6 Evaluation of Procedures 22

6 Interaction between political measures and Estimation **23**

6.1 Period Division and Events 24

6.2 Analysis of Events based on Estimations 25

7 Conclusion and Comments **27**

A Appendix **29**

A.1 Point Estimation Bayes 29

A.2 Calculate Posterior 29

A.3 Weighted Probability Estimator 30

B Output results **32**

B.0.1 Age Group Estimation 33

B.0.2 Credible Intervals 35

C Code **38**

Bibliography **63**

List of Figures

1 *Scatter plot depicts the comparison between the Bayes estimate and the actual ratio. The second plot depicts the data points which relatively experienced a higher deviation from the ratio.* 9

2 Age-Groups death rates in First Wave 13

3 Comparison of Likelihoods 17

4 These figures show: 1. The posterior vector of the MCMC process, 2. The Bayesian estimation posterior, 3. The comparison between the estimates, 4. The Mixing for the MH algorithm 21

5 Comparison to Bootstrap 22

6 Robert-Koch-Institut (2021), Rheinland-Pfalz Timeline 24

7 Course of death probabilities 26

8 Posterior vector with a Uniform(0,0.5) Prior distribution 27

9 Age-Groups death rates in First Down phase 33

10 Age-Groups death rates in second Wave 34

11 Credibility Intervals 37

1 Introduction

The novel Coronavirus affected the whole world and led to many casualties during this process. Every city in the world suffered from the virus. This thesis focuses on German counties, cities, and states, tries to implement Bayesian estimation techniques, usage of MCMC algorithms, and will discuss probabilities of death in each county and/or state in Germany.

The data is directly drawn from the RKI (Robert Koch Institut) as each district's health department has to report each Covid-case immediately to the RKI. Thus, accuracy and reliability will be given. I also want to tackle the issue of evaluating different probabilities in each state against each other. Chapter six will also discuss different measures taken by various regions, depending on political and health-related issues regarding the outcome of the death-probability estimation.

The Bayes estimation technique bares many advantages, as it tackles the issue of sample size. This thesis will provide information about the different death-rate distributions and show discrepancies in handling the pandemic in different regions. As counties and states vary in size, handling a Bayesian framework is suitable. This topic will be addressed in Chapter five.

I want to show that Bayesian statistics can be appropriately used to estimate mortality rates and can establish viable credibility intervals. The structure of this thesis is as follows: the second chapter will deal with the previously explored literature. The third chapter will present the data and discuss data handling and issues to overcome. The fourth chapter will give a short introduction and overview of the mechanics and intuition of the general procedure. As mentioned before, Chapter five will present the actual death rate estimation of all counties in Germany and introduce other issues and topics regarding inference. The sixth chapter will integrate the estimation technique into political decisions and measurements taken against the pandemic and possible correlations between the outcomes and significant events. The chapter explores different time frames and outlines the pandemic based on the Bayesian model estimation.

It is difficult to incorporate every county estimation and present it insightfully. Therefore, the Code-Section of the Appendix will provide a function from which one can estimate the posterior distributions with estimates and quantiles. This code, alongside the given Output, can also be found on GitHub (`https://github.com/spg2143/BA_Covid-Death-Estimation`). This analysis will provide information about the plausibility and usefulness of Bayesian estimation regarding the framework, data, and possible other modeling approaches, which will be dealt within chapter seven.

2 Literature Review

Plenty of papers on how to model and forecast deadly viruses regarding humans and/or animals like Matt Keeling (2008), which show the mortality of various diseases and infections using Bayesian Models are already in existence. Papers like Marshall (1991) used a standard Bayes Estimator estimated by moments and developed a shrinkage estimator, where the author shrunk the estimator towards a local neighborhood death rate. Also, Schmertmann and Gonzaga (2018) introduced estimation of death rates with different likelihoods and modeling, which is also incorporated into this thesis (Chapter 5.1.3). The Bayesian framework is not new to estimating death and infection growth rates. Guzmán-Rincón et al. (2022), actually estimated Covid growth rates by implementing a Bayesian hierarchical model with Gaussian Processes in England.

However, as the pandemic sets in, in 2020, detailed literature is not vastly available. Nevertheless, as seen in Matt Keeling (2008) one can access papers on other similar pandemic-type diseases to evaluate proper measures.
Additionally, crucial works in mortality estimation were done in the paper written by Ronald Lee and Lawrence R. Carter (1992), as the paper explained the Lee-Carter Model for age-related classification. Karim Barigou et al. (2022) introduced the mortality forecast methodology with the idea to express the mortality setup with regards to the pandemic by taking into account that exposed individuals will be more likely to perish. This setup is explained and applied in Cairns et al. (2020). Like the latter cited, many papers use a Poisson or a negative Binomial model to represent the data at hand. In all of these papers, the Bayesian framework is set. These papers have shown that Bayesian Modeling will lead to good results and advantages in contrast to the frequentistic approach. There are many approaches how to handle estimation. However, it proved to be logical to incorporate the Bayesian technique. As mentioned, the literature for pandemic-related papers, especially regarding Covid, is extremely scarce. The paper of Karim Barigou et al. (2022) introduced the age-specific death rate estimation with the Lee-Carter Model, a regression estimated via the Bayesian Estimation technique. This paper tried to implement the pandemic

into this model (as did Cairns et al. (2020)).

3 Data

As already stated, the data is drawn directly from the RKI. Data compression shows each day as an entry for every city ID of Germany. The data is depicted as a time series. The appended data shows the counties with the individual case and death count and the death ratio. Furthermore, the statistics about males and females are also attached to this data.

The data is depicted by a time series representing the death counts and the number of infections. The time frame will be generated weekly and is described as $T = \{1, 2, 3, ..., 120, 119\}$ which portrays the weeks of 2020-10 - 2022-35. The partitioning of age will happen by 10 years: $A = \{0 - 9, 10 - 19, ..., 80 - 89, 90+\}$. For this estimation to coherently run, I will accumulate the infections and deaths for a known sequence. The partitioning will follow the graph of Robert-Koch-Institut (2021) and covers major political decisions and important events (e.g., waves and down-phases) for the pandemic in Germany. The analysis is driven for counties and states but can also be applied to the whole country. The data itself is first described as a country-wide accumulation of cases, and additionally, another data frame will be provided, in which the specific keys for counties are located.

As stated above, the country-wide data will be stored in the frame "*Altersverteilung.xlsx*" and "*Sterberate.xlsx*" which can be found in the repository on GitHub, which is linked in the Introduction. The statistical coding in this thesis was created in the IDE "RStudio," and the code generated can also be found there alongside the plots and tables, which are also included in the section B.0.2. These Data-Sets are for estimating the age-specific death rates, as they only describe the Infections and deaths across Germany.

Therefore, I opted to include another data set, with all cases from each county and state drawn from RKI only with different formatting. This Data-Set also allows selecting a time frame, for time-specific analysis, regarding the periods of the pandemic. The data shows the age partitioning as follows: {A00-A04, A05-A14, A15-A34, A35-A59, A60-A79, A80+}, genders as: {Male, Female, unknown} and the different counties, characterized by county ID and location. I transformed the loaded data by characterizing the state ID and extracting the specific state from the county, allowing a more profound analysis of accumulated cases in each state. These characteristics are also the parameters in the function which is added to the GitHub repository.

4 Bayesian Models and Estimation

4.1 Introduction to Bayesian Estimation

The Bayes Estimation technique is elegant while using data skewed in size. As the prior belief influences the outcome, the method critics argue that the estimation is too subjective as the interpretation of likelihoods and a priori distributions can be perceived differently. However, by closer examination, one can find the usefulness of the estimation technique. Unlike the frequentistic approach, the Bayesian handling desires a posterior distribution and not a point estimation. One way of capturing uncertainty in a frequentistic practice is through bootstrapping, which will be discussed later. While analyzing the Bayesian technique, one can see the intuition firstly through the Bayes' theorem (Gelman (2014)):

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int_\Theta p(D|\theta)p(\theta)d\theta}$$

In estimating the posterior distribution, the Bayesian framework incorporates uncertainty of estimation. The likelihood of the data is represented by $p(D|\theta)$. The prior $p(\theta)$ includes the past knowledge of the desired parameter θ and shows accuracy of information. The denominator is a normalization constant, the marginal distribution of data, assuring that the posterior distribution will accumulate to 1. Therefore, the estimation procedure is referred to as a proportional equation:

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

The posterior consists of elements from the a priori distribution and evidence given by the data. One can think of an updating system in which the posterior is an amalgam of prior and likelihood. If the prior is highly uncertain, e.g., high variance prior, then the input of data will proceed to skew the posterior from the prior enormously. Updating is vastly important in the Bayesian setting as it is used in Bayesian filtering, Kalman filters as de JONG and TICKLE (2006) did, among others. In the Covid setting, we can interpret the parameter θ as the death probability and constitute the likelihood of the respective data.

As I want to conduct it in the next section, Bayesian point estimation will be conducted to minimize the least squares estimation error of the posterior distribution. The error can be written as the mean squared error like in Dr. Ing. Gerd Ehret (1711):

$$\epsilon(\hat{\theta}) := E[p(\hat{\theta} - \theta)^2 | D]$$

and finally, one can see the intuitive result (see Appendix in A.1):

$$\hat{\theta} = E[p(\theta|D)]$$

that the expected value of the posterior estimated distribution is the best point predictor in a Bayesian setting.

Now, regarding the setting in this paper. I will use a very simplistic approach, the binomial distribution, a well-known and identified distribution for modeling death rates. The parameter to be estimated is the probability of success, which in this case would be the probability of death. Thus, the number of Infections will postulate the baseline for the model, with the number of successes as the accumulated death count for a specific region, gender, or age group. [1]

4.2 Choice of Prior

One can observe that the estimation in small datasets depends on the prior as the posterior calculates from both the likelihood and the a priori distribution. The choice of prior can be made with three different methods

[1] I will assume the binomial distribution over the infection as a sum of i.i.d. Bernoulli(p) random variables.

- Non-Informative Priors such as uniform distribution with no bias nor viable information
- Subjective Priors, which include idiosyncratically perceived information
- Information-based Priors, which include past proven data

Also, conjugate priors can circumvent long-during MCMC algorithms, which might take time. This means that if the a priori distribution is in the same 'family of distributions' as the likelihood of the data, one can easily extract the posterior by its similarity. The following example will show this conjugation between the binomial and the beta distribution (Gelman (2014))
I will select the a beta prior, which is the natural choice for estimating the probability parameter in the model.

5 Estimation of death probabilities in German Counties

As explained in the latter section, one can conduct the Bayesian analysis with the standard Bayes point prediction. This estimation will use the Binomial distribution, as the time series data suggests that a binary representation of all individuals exposed to the virus is a plausible model for the death count in the respective counties. It is also possible to include a negative binomial or a Poisson model. Still, my analysis proved that all distributions would yield the exact outcome of estimation (5.5). This example will follow to inspect the death ratios in regional counties. The case numbers will be relatively low compared to an accumulated amount in each state, where the likelihood estimation formation in the Bayesian setting can be problematic.

5.1 Procedure of Estimation

Given the reasoning, I will formulate a prior that the probability of death is small, therefore, incorporating a prior with an expected value in the region of $0.05 - 0.15$ and compare it with a right skewed prior, inducing a death probability of $0.85 - 0.95$. Step-by-step estimation is as follows:

i. Understand the data and determine the distribution of given observations. In this case, a binomial distribution with the parameter in question, which is the probability of death $\theta = p_i$

ii. State distribution of data given and state prior distribution[2] as initial thinking for distribution.

iii. Perform Bayesian Estimation: Incorporate data with the given prior and source the posterior distribution for θ

Step iii. shows the establishment of the posterior distribution. The calculation is simplified as the distributions are conjugate. Thus the posterior will be stated in terms of the prior and the data. Therefore, there is no need for tedious MCMC computation (Gelman (2014)) A.2:

$$p_{\text{death}}|\text{Data} \sim Beta(D + \alpha_0, I - D + \beta_0)$$

As the posterior is a beta distribution, one can find out the expected value with:

$$E_{\text{Beta}} = \frac{\alpha}{\alpha + \beta}$$

$$\hat{p_d} = \frac{\sum^N d_i + \alpha_0}{\alpha_0 + \beta_0 + I}$$

The estimate for the parameter can be expressed in a weighted order to show the trade-off between the importance of data and the prior depending on the uncertainty and size of the sample. From the equation of $\hat{p_d}$, one can derive:

$$= (\frac{I}{\alpha_0 + \beta_0 + I})\bar{D} + (\frac{\alpha_0 + \beta_0}{\alpha_0 + \beta_0 + I})(\frac{\alpha_0}{\alpha_0 + \beta_0})$$

Derivation of this intuitive identity can be found in the Appendix(A.3). As the specific case numbers are considerably high, the estimation for the ratio will instead tend to the likelihood, which means that we can expect that the ratios of death will give an adequate representation of the grand scheme of things. The parameters

[2] Generally, the distribution can be estimated through the method of moments or Maximum Likelihood estimation, but as this approach is as viable and with the R package "fitdistrplus" by Delignette-Muller and Dutang (2015) easily executable, it is as effective and fast.

in the posterior will predominantly be given by the data, as the amount of cases in each period and region of Germany is relatively high. However, regarding the importance of prior, the weighted probability estimator can also explain differences in these previous assumptions.

Regarding the actual estimation of all counties, one can plot the Bayes Estimate against the given data ratio with the use of two different priors (Prior values in expected value: {0.15; 0.9}).

Scatter plot comparison of estimates with all data points and selected with higher deviation

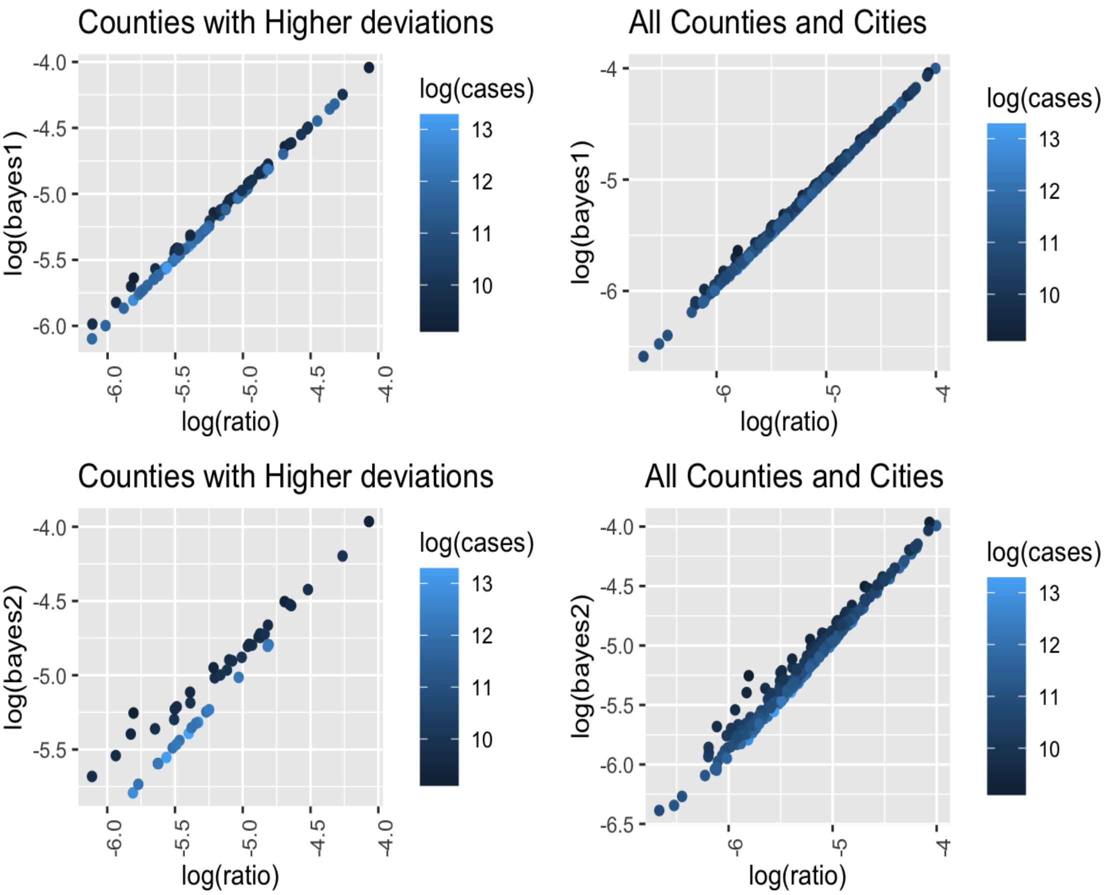


Figure 1: *Scatter plot depicts the comparison between the Bayes estimate and the actual ratio. The second plot depicts the data points which relatively experienced a higher deviation from the ratio.*

The first graph in each row in Figure 1. shows the counties which follow the set heuristic only to select the following states that suffice this condition with prior 1:

$$\text{Counties with higher diversion} = \{-10 > \log(|\text{bayes}_{estimate} - \text{ratio}_{\text{frequentistic}}|) > -8.5\}^{\complement}$$

and Prior 2 (big):

$$\text{Counties with higher diversion} = \{-9 > \log(|\text{bayes}_{estimate} - \text{ratio}_{\text{frequentistic}}|) > -7\}^{\complement}$$

The estimation is done with two different priors to show the influence. The first row incorporates the prior, with low expected value, close to the data ratio, whereas the second row prior is postulated with a high expected value. The bounds were chosen after plotting the histogram of all differences in both cases of prior distribution (B). The heuristic ensures that only counties with more significant derivation will be presented. As apparent, only the complementary set of values which are close to the ratio estimation, will be considered. The further analysis will regard the counties with the highest absolute log differences between the data ratio and Bayes estimate. Only the prior with higher expected probability will be shown for identifiability due to the apparent more significant discrepancy between the estimates. Even outliers will more likely tend towards the ratio, as the scatter plots in Figure 1 present, as the number of cases in each county is comparatively high. Generally, the posterior is mainly built by the likelihood, as several thousands of observations are given, which then primarily define the posterior. This approach generally generates a Bayes ratio that does not deviate much from the data ratio as the prior includes information from the evidence. Thus, if including priors skewed away from the data ratio, the estimate would change significantly from the frequentistic estimate, as seen in Figure 1. The given evidence can not back up counties with less observation. They would be expected to have a higher Bayes estimate as the prior distributions set out a high probability of death. On the other hand, counties with many observations are expected to have little to no deviation from the original ratio, which is intuitive looking at the setup of the posterior distribution 5.1.

5.1.1 Estimation Analysis and Concerns

The next table will present the counties with the highest absolute log differences and summarizes the characteristics of the estimation:

Only the log evaluation will yield presentable differences in death probabilities, as margins are small. Nonetheless, one can inspect these differences, as mentioned previously. Starting by introducing the weights of, either the evidence, or the prior depending on the formulas given the weighted estimate formula (Chapter 5.1):

County	Bayes Estimate	Death Ratio	Alpha Bayes	Beta Bayes
Zweibrücken	0.0052	0.0030	47	8995
Pirmasens	0.0111	0.0092	116	10485
Suhl	0.0190	0.0171	199	10490
Landau in der Pfalz	0.0071	0.0054	85	11992
Lüchow-Dannenberg	0.0045	0.0029	57	12571

County	Evidence-Importance	Prior Importance	Cases	Deaths
Zweibrücken	0.57	0.43	8973	27
Pirmasens	0.83	0.17	10463	96
Suhl	0.90	0.10	10468	179
Landau in der Pfalz	0.76	0.24	11970	65
Lüchow-Dannenberg	0.65	0.35	12549	37

This table proves that the number of infections and the "quality" of data are decisive for the posterior estimate weights. The intuition holds for "Zweibrücken", as the county has the lowest number of cases and the highest percentage of prior importance for estimating the posterior distribution. "Lüchow - Danneberg" counters the intuition, as the county has the most case numbers but accounts for the penultimate percentage in importance. This case is not as intuitive as "Zweibrücken". However, examining the weighted point estimate and the actual importance of the death count, the term for the importance of the evidence (data) is split up into another trade-off. Recalling the term to be:

$$(\frac{I}{\alpha_0 + \beta_0 + I})\bar{D},$$

the interaction between the ratio of deaths and the case number performs a valuable role. The lower the ratio, the less the estimation will be. Additionally, the importance will decrease simultaneously if the prior values are not set to be of high importance. With the beta prior, it is clear that higher values for the shape parameter α_0 and β_0 induce less variation. Intuitively, the uncertainty of the prior reduces, and the subjective perception believes that the prior is already a decent estimator. Thus, the prior parameters will demean the evidence's importance and shift the focus mainly towards the prior [3]. Therefore, the already stated criticism is partly shown in this example. The belief that a parameter might be in an uncertain range can deflect the actual value. However, the cost of uncertainty can also be circumvented by choosing prior distributions with little-to-none explainable capability,

[3] Figure 1

e.g., the uninformative uniform distribution. Moreover, the usual case, as it is in "Suhl", shows that, if a sufficient amount of data is provided, the estimation will naturally lead to the evidence rather than to the prior.
The framework also presents the geographic closeness of the counties in question. Three out of five counties share the same state and are all in the vicinity of each other, which makes it apparent that the inference provides evidence for geographic spillovers. According to the Bayesian paradigm, this means that if a cumulative region is affected, the probability will be high that an adjacent county will also be affected with a similar death estimate. As mentioned in the Introduction, Chapter 6 will deal with that fact, among the influences of political measures taken in that region.

Another issue is that I assume the same prior for every county. It is possible to issue this by employing an ensemble building technique, which selects which prior will be set for an individual county. However, this exceeds the limits of this thesis, because the computation requires complex Learners. However, given the dataset, it is possible to deepen this analysis, regarding the age groups over the whole pandemic.

5.2 Analysis of Age-Group Death Estimation

This section will deal with all of Germany's Covid cases and presents the age specific death-rate estimation. The second part of the data is parted in ten different age groups. The framework can now be applied to exactly such distribution, so that one can examine different time periods in the pandemic and evaluate death rates against each other.

It is possible to regard the data in periods such infection waves or down-phases (As seen in Chapter Six). Firstly, considering the first wave one will yield the estimates as in Figure 2.

Following the intuition, Age Groups with more case numbers wont deviate from the evidence. The prior almost has no importance and the point estimation converges to an actual frequentistic estimate, given the fact that hyper-parameters α_0 and β_0 explain little of the posterior. The cascading color scheme represents the number of cases. If an Age Group has a relatively low amount of cases, the prior will push the Bayes estimate away from the evidence, which is shown with the first two age groups {0-9} and {10-19}.

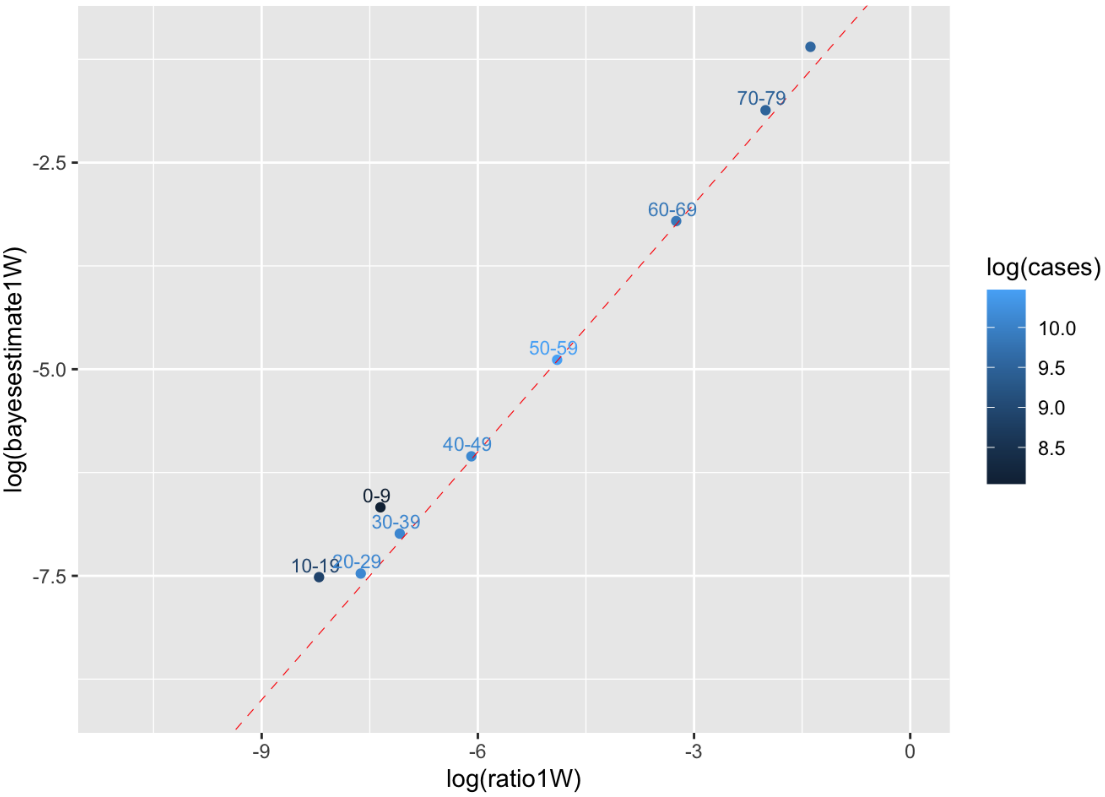


Figure 2: Age-Groups death rates in First Wave

The appendix (Appendix A.4) will show the different wave estimates, and captures the basic gist of the estimation as already stated in the latter example.

5.3 Different Priors and Likelihoods - Metropolis-Hastings Algorithm

Furthering this analysis, one can implement different priors, which explain the gathered information more eloquently. Of course, representing θ can only be considered if the a priori distribution characteristics span from $\{0, 1\}$, consequently assuring that the death probability is an actual probability. Thus, one can implement a uniform distribution spanning from $\{0, 1\}$ or, for example, a normal distribution that fixes the characteristics between said region. Of course, it is possible to compute the posterior with various distributions. However, based on the lack of information, I will incorporate the priors in question to the Metropolis-Hastings algorithm. Especially the uniform distribution presents itself as potentially valuable, as there is no influence of prior (mis-)information. The normal prior influences the result to a degree where to prior is too powerful, meaning that, in order to assure the characteristics of the distribution to be bounded by $0, 1$, the variance has to be significantly low, which per assumption reduces the chance to learn from the given data, as sampling will only lead to draw from values adhering to the prior, which makes the calculation of the posterior redundant.

I will present and use the Metropolis-Hastings algorithm to find posterior distributions. The algorithm was brought up by Metropolis et al. (1953) and was refined for non-symmetric prior distributions by Hastings (1970). Together, the scheme of the algorithm sets up as follows:

```
Input: DeathData depending on Age and Time Period
Set number of simulations $N_{sim}$
Create posterior vector: $\theta = p = []$
Initialize posterior vector with one draw from the prior p($\theta$ = p)
for t in sample runs do
    1. Sample candidate from previously set prior
    2. Use the t-th value of the posterior for comparison $\theta_t = p_t$
    3. Calculate the Metropolis Hastings Ratio
    4. Use the ratio as a probability $min(1, MHRatio)$
    5. Sample between the proposal value $\theta^* = p^*$ and baseline value $\theta_t = p_t$
    6. Store value in posterior vector
    7. increment t and run process again
end
Plot histogram of posterior vector and evaluate mixing by plotting scatter plot
  of posterior vector.
```

In words, the algorithm starts by initializing a vector in which all values of the posterior are stored. Saying that we have information about our posterior, we draw one sample from the prior and store it in the posterior vector as the initializing value. Now coming to the actual Monte-Carlo simulation, which simulates the Markov-Chain, as the posterior will iterate n times. In the Monte-Carlo simulation, one will draw a proposal value from the prior and will evaluate this draw against the previously accepted value for the posterior (baseline for next iteration). The baseline value, which is stored in the posterior vector as θ_t, will be evaluated against the proposal value θ^* as a likelihood ratio process. The data's Likelihood will then be evaluated with the given proposal and the baseline value. This ratio will then function as a probability for sampling the next value in the posterior vector. The probability is defined with: $\min(1, \text{MH Ratio})$, as the Likelihood with the proposal value can be higher than the previously accepted value, which makes the probability > 1. Therefore, the proposal value θ^* will be accepted with a probability of 100%. Continuing, θ^* will be accepted as the next value for the posterior vector. In this case, θ equals the probability in the binomial distribution. Therefore, the data serves as evidence to influence the a priori distribution. Iterating this process will create a Markov Chain and eventually converge to the steady-state posterior distribution. The MCMC simulation will create this chain which starts in a random initial state drawn from the prior distribution. The chains have Markov properties, which allows us to assume convergence to the stationary distribution in question, as the next entry in the posterior vector only depends on the previously accepted value Gelman (2014), however, normally one incorporates a burn in phase for the Markov-Chain to reach its steady state distribution. Thus, sampling from the proposal and baseline value will be equivalent to sampling from the true posterior distribution. Yielding the posterior vector with the length $n + 1$ as we simulate this process n times. Intuitively, the accepted values will converge to the expected values as the probabilities will be higher in the areas where the probability mass for the respective value for θ is high. This defines the posterior probability distribution.

In this analysis, I will define the Likelihood as a Log-Likelihood. It is commonly known that handling the sum of probabilities is more accessible than the product of all likelihoods. Furthermore, dealing with small death probabilities, the computation quickly fails, as the Likelihood is too small. Therefore I adjusted the Metropolis-Hastings algorithm to the extent that the calculation will be done with the inverse of the log-likelihood, which deviates from the standard calculation of the MH-Ratio:

$$MH\_Ratio = \frac{\frac{1}{\mathcal{L}(D|\theta^*)}}{\frac{1}{\mathcal{L}(D|\theta_t)}}.$$

$$= \frac{\frac{1}{\log((\binom{I}{D})+D\log(\theta^*)+(I-D)\log(\theta^*))}}{\frac{1}{\log((\binom{I}{D})+D\log(\theta_t)+(I-D)\log(\theta_t))}}$$

$$= \frac{\log(\binom{I}{D}) + D\log(\theta_t) + (I-D)\log(\theta_t)}{\log(\binom{I}{D}) + D\log(\theta^*) + (I-D)\log(\theta^*)}$$

This allows incorporating Log-Likelihoods into the MH- algorithm. As the property of Markov-Chains still holds up, one can generally still assume convergence to the actual posterior distribution. The sum of all log probabilities will be negative, as the range of probabilities will be between {0,1}. Therefore the sign cancels out. If the log-likelihood value for a proposal has a small absolute value, the probability of acceptance should be high, due to summation of low log-values (high ratios and thus probabilities). Thus, taking the inverse of both likelihoods can penalize low likelihoods for a proposal value θ^*. The rest of the algorithm remains equal to Metropolis-Hastings.

5.3.1 Modeling with different Likelihoods: Poisson

The modeling in this paper is handled with a binomial data-set distribution and a classic beta prior. This is computable with the Kernel method explained in the introduction to Bayesian Estimation and can also be computed via the MH algorithm. However, if the data is interpreted differently, e.g., a Poisson distribution, it is necessary to implement the introduced Metropolis-Hastings algorithm to render the posterior distribution. Modeling the data with Poisson is set up as follows (Schmertmann and Gonzaga (2018)):

$$D_i \sim Poisson(I_i \times p_i).$$

The probability p_i still represents the death probability, or the mortality ratio for each state (region), depending on age, gender, and time window. Naturally, the prior choice will not deviate from the previous selected beta prior, as the characteristics of the parameter did not change. I will present the posteriors for both MH simulations, with one: the binomial likelihood. And two: the Poisson likelihood. The data has following specifics: {State: ”NRW”, Gender: ”W”, Age Group: ”A35-A59”, Time Stamp: ”2021-01-21” } with a beta prior and parameter $\{\alpha_0 = 4, \beta_0 = 50\}$.

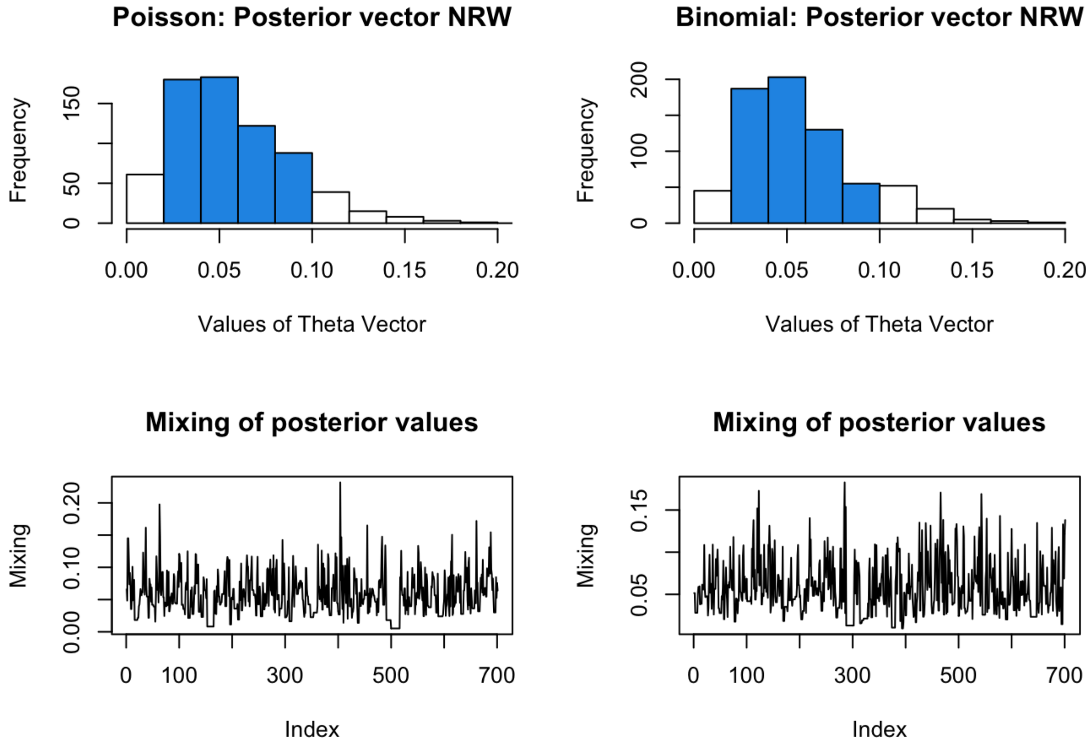


Figure 3: Comparison of Likelihoods

Mean Poisson LH	0.05728825
Mean Binomial LH	0.0575546
Expected value of Prior	0.07407407
Death Ratio Data	0.0005428436

As apparent, the posterior is very similar in both likelihood cases, which underlines the plausibility of the distribution. Further, the graphs under the respective posteriors represent the mixing of the posterior values. As there is a high fluctuation in values, I assume that the characteristics of the posterior are met and thus, also provide further evidence for the legitimacy of the distribution. The variation of the Poisson estimate is higher, due to the definition of the likelihood. The Binomial representation strictly bounds the results to be binary, whereas the Poisson distribution, can not guarantee these bounds. Therefore, the Poisson-Likelihood approach is bound to present a higher mixing value, as more it is more probable to sample from higher valued outcomes (Deaths presented as binary → Death = 1), which can be a positive remark, because the probability of explaining every characteristic will be higher.

5.4 Posterior quantiles and intervals - Credibility Interval

The Bayes estimation provides a credibility interval which can give out specific intervals for estimation accuracy. This thesis will provide credibility intervals of 80%. The calculation of a higher interval is too inconclusive, as the quantile variation is too large, with almost no factor explainability. The credibility intervals can be derived from the posterior distribution, commensurate to Maximum Likelihood or Least Squares estimation. In this framework, I will present 16 different posteriors (all German states) depending on specific age groups, which can be evaluated over the course and significant incidents in the pandemic. Main political and pandemic-related events mark thresholds for the span of the data in question (See Chapter 6). From there on, one can calculate the empirical quantiles for the distribution and produce the credibility interval. The credibility interval is generally equal to the confidence interval. The branding is only different between the two approaches of frequentistic and Bayesian modeling to distinguish both. Using the conjugate approach, where one is not dependent on running the Metropolis-Hastings Algorithm, one can extract the theoretical quantiles from the posterior to coherently set up the credibility interval. The credibility intervals in the MH-Algorithm are calculated by the "empirical" quantiles of the posterior vector of accepted values.

During the conjugation method computation, capturing logical posterior quantiles in the estimation was challenging. The table below will show that the output presents too inaccurate quantiles to allow any form of interpretation. The interval below shows only the credibility interval of 30%, and moreover has almost no explainable value. However, using the MH algorithm, the results were intuitive and reasonable, and thus, will be used to estimate the posteriors and present credibility intervals.

	0.35%	0.65%
Zweibrücken	2.087998e-13	1.035415e-02
Pirmasens	4.250337e-07	8.311239e-01
Suhl	2.172771e-28	4.400588e-03
Landau in der Pfalz	1.775373e-13	1.127270e-02
Lüchow-Dannenberg	8.473039e-22	2.222896e-05

Unlike the previous data choice, the following representation of the intervals will be shown in histograms in the Appendix B.0.2 of all German states to state differences in estimation. The data offers a variety of options to assemble. I will show the credibility intervals for each state depending on the gender over the period up until the beginning of the second wave in late October 2020 (Robert-Koch-Institut (2021)). The data is now set up to cover all periods and includes all deaths and Infections in the respective state. Running the code provided in the GitHub repository will also show the numerical intervals of each state. In some cases, the posterior is too spread out to state conclusions over the accuracy of the estimation. Therefore, the blue area in the respective histogram represents 80% of the data, thus representing the credibility interval of 80%. One would expect that smaller states, like ”Bremen” or ”Hamburg,” present a larger credibility interval, as these states do not have abundant evidence. However, the posterior credibility interval varies through each state and each data adjustment (Gender, Age, etc...). Examining ”Bremen”, one can see that the credibility interval is, in fact, the smallest in each state. Nevertheless, some states agree with the latter assumption, e.g., ”Baden-Württemberg” or ”Bayern”. The MH algorithm provided a close approximation to the actual posterior distribution. Thus, the credibility intervals are reasonable. Also, regarding the respective ratios of the specific data, all posteriors include the point unbiased by the prior and suggested by the data, which underlines the plausibility of the estimated distribution. There is no evidence that regional factors in the size class of actual states influenced the estimation, as adjacent states in this size classification do not provide proof for this identity. Chapter seven will analyze deeper levels of county data. Here, the accumulated data region’s margins are too big for a reasonable deduction of regional influence on death rates and their credibility interval. Generally, the intervals only differ marginally, as the same prior will also induce a similar posterior, which makes it difficult to extract significant variations of credibility intervals.

5.5 Comparison of Bayes Estimate, MH Algorithm (and Bootstrap)

We can inspect the closeness of the estimation by introducing a conjugate prior to the MH Algorithm. Doing so would expect the exact distribution outcome as in the conjugation method. I want to check, ceteris paribus, which technique fits better to the given framework and data. The package ”fitdistrplus” (Delignette-Muller and Dutang (2015)) offers users to empirically calculate the distribution over the data and shows the accuracy of the presented distribution. It is possible to inspect different cases.

Showcasing the results for state "'Rheinland-Pfalz' and age group 'A35 - A59' of all males. As mentioned, it is possible to plot the distributions of both estimation techniques. The results show that the conjugate method is more consistent and accurate to what the actual belief and rational assumptions propose. The Metropolis-Hastings algorithm is dependent on the prior deviation, making it hard to compute reasonable conclusions about the posterior without credible prior information. In theory, the simulation will hold up. Also, empirically, it is shown that the evidence, in the form of the death data, is not as influential as initially expected. The function encoded for this thesis outputs the difference between Bayes estimates and MH estimates and will show the estimation results. It will also measure differences in estimation. As in Figure 4., one can see that the MH algorithm is influenced by the a priori distribution. The reasoning could be that the likelihood is skewed too far to the zero-bound, so the a priori distribution will be unlikely to identify the actual characteristics of the posterior distribution. Plotting all the posterior's accepted data points will show outliers in maximum and minimum areas of the probability (proposals of $\hat{p} \approx 0$). The problem, however, is that a bad prior will influence the drawings to that point until the Monte-Carlo simulation eventually stops. The case of convergence still proves that even if the prior is not well equipped to showcase information about the posterior, the chain will ultimately converge to the steady-state posterior distribution Gelman (2014). E.g., the MH-posterior is similar to the standard Bayesian estimate. The posterior distribution of the MH algorithm is also approximately credible, as the mixing in Figure 4. shows that a sufficient amount of the characteristics of the posterior distribution is covered, alongside a few outliers. Suppose one scatters the different points of ratio, Bayes estimate, MH estimate, and the prior expected value. In that case, one can see that the variation of prior influences the posterior. The influence is more significant in the MH algorithm than in the Bayesian estimation, depending on the number of simulations. As mentioned, the amalgam of prior and evidence is shown again (Figure 4. bottom left), where both estimates take weights from each input characteristic. Intuitively thinking, if the prior variance is relatively large, the MH-Algorithm would learn more from the data as an effect of uncertainty.

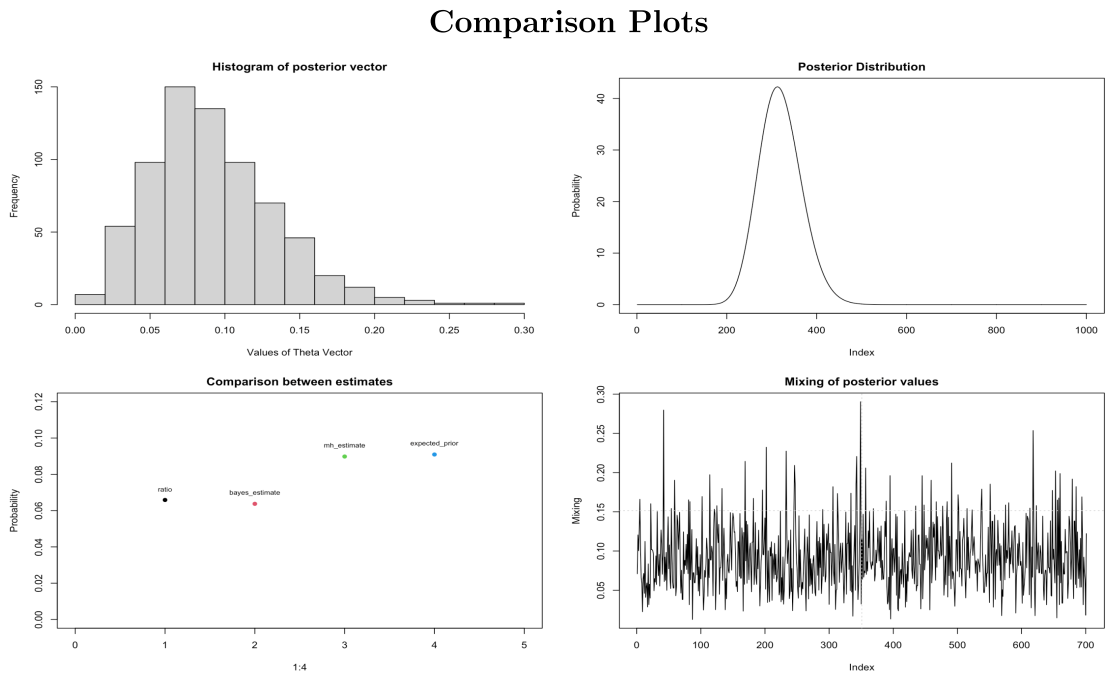


Figure 4: These figures show: 1. The posterior vector of the MCMC process, 2. The Bayesian estimation posterior, 3. The comparison between the estimates, 4. The Mixing for the MH algorithm

5.5.1 Excursion: Comparison Bootstrapping

A way of building a proper distribution about an estimate in frequentistic statistics would be Bootstrapping. Bootstrapping, samples from the data as many times and will estimate the values as often as the user requires it to Efron and Tibshirani (1993):

- For b = 1,2,...,B:
  - Creating a new data-set from the initial data-set, by sampling with replacement
  - From there, it is possible to estimate the parameter again. Iterate this procedure B times.

Obviously, bootstrapping's disadvantage is that it is not possible to implement prior beliefs. Thus, sample size does matter. Now, considering, one the MH algorithm and, two the Bootstrapping technique with data characterizations of {state = Hamburg, age = A35-A59, Date: 2020-10-27}:

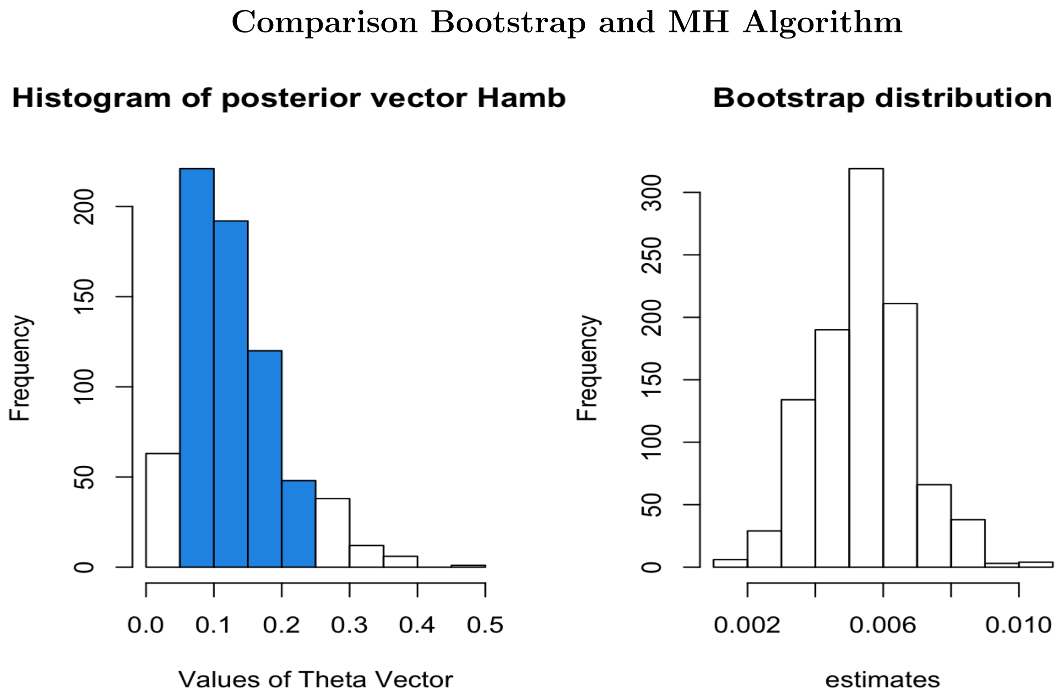


Figure 5: Comparison to Bootstrap

Evidently, the bootstrapping will only yield a posterior distribution depending on the initial distribution. It wraps around the mean of the data and postulates the parameter distribution from there. This identity is one major disadvantage, as one can not include any prior information whatsoever. Thus, fully dependent on the data quality and size, because the size of the data set is the driving factor in accuracy for the distribution.

5.6 Evaluation of Procedures

Moving on from the comparison, the evaluation of procedures will be addressed. I will opt not to include the frequentistic approach because the case of subjective bias will not be deeper analyzed. The Bayesian framework always suggests a prior, which in many cases has many advantages. However, for example, the plausibility of priors for all data points has to be addressed. To specify the model setup, one can implement many subjective priors with lengthy computations. However, instead of implementing this, with little information about long-term Covid effects, I will disregard that issue to the extent that I will not follow up on the frequentistic paradigm and choose to differ between the similar conjugate prior and Metropolis-

Hastings estimation. The advantages of both evaluation techniques are similar, as both require the same framework in Bayesian Statistics. However, for the general estimation of Covid-death rates, the conjugation procedure will be more effective and saves computational time, as MCMC algorithms require tedious runs and simulations. As introduced, the setup in this paper is based on a simple case with a binomial likelihood and a beta prior distribution. The modeling allows using the more coherent and stringent approach, as both distributions root from the same family.

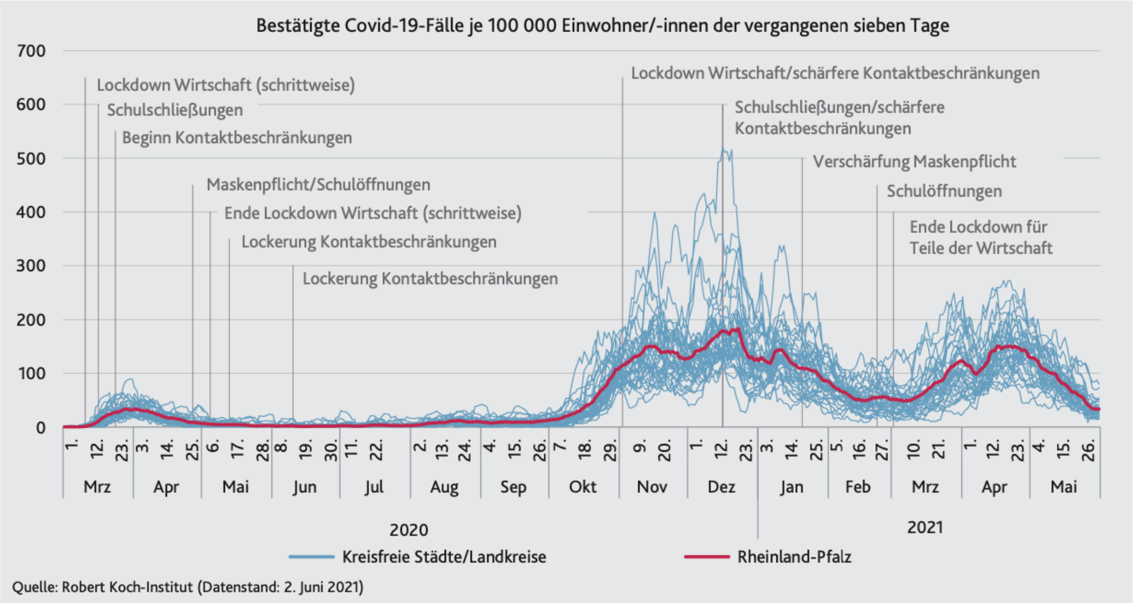
Nonetheless, the Metropolis-Hastings algorithm offers to calculate posteriors with different likelihoods and priors, which do not root from the same family. Many papers as Karim Barigou et al. (2022), or Jackie S.T. Wong et al. (2018) used methods like the negative Binomial or the Poisson model, with different priors, where the algorithm was essential. Therefore, the modeling influences the computational capabilities, and limits accuracy since the posterior distributions can not be calculated mathematically.

6 Interaction between political measures and Estimation

As seen in the Estimation in Chapter 4.2, there is a correlation between the geographic location of a county and the Bayesian estimate. The table below depicts the ten counties with the highest deviations from the ratio:

	County	State
1	Zweibrücken	Rheinland-Pfalz
2	Pirmasens	Rheinland-Pfalz
3	Suhl	Thüringen
4	Landau in der Pfalz	Rheinland-Pfalz
5	Lüchow-Dannenberg	Niedersachsen
6	Amberg	Bayern
7	Frankenthal (Pfalz)	Rheinland-Pfalz
8	Schwabach	Bayern
9	Coburg	Bayern
10	Neustadt an der Weinstraße	Rheinland-Pfalz

Most of these counties are in close vicinity to each other. ”Rheinland-Pfalz” makes up 23% of the largest 100 differences in the county estimation (Chapter 4.2) and is used as a proxy, to show the course and outcome of political measures. The next section will deal with examining the state of "Rheinland-Pfalz" and presents a timeline with cases per 100.000 inhabitants and major political measurements up until the third wave in 2021.



Quelle: Robert Koch-Institut (Datenstand: 2. Juni 2021)

Figure 6: Robert-Koch-Institut (2021), Rheinland-Pfalz Timeline

The function in the appendix allows to set a time-frame and and estimates the death probability in that window. This allows to compare and analyses between the regions and describe political measures taken in that window.

6.1 Period Division and Events

1. Period The first period depicts the beginning of the pandemic. The WHO declared the global pandemic on 11.03.2020, where several cases were already being monitored and recorded The government was obligated to implement a lockdown and closure of schools to switch to remote lecturing. (01.03.2020 - 06.05.2020)

2. Period The second period describes a down-phase in the summer, where case numbers and deaths were flattening. Lockdown was lifted and most of the hygiene precautions were rescinded. (07.05.2020 - 26.09.2020)

3. Period The third period shows the second wave of the pandemic, where the case numbers and deaths were increasing tremendously. During the next two waves, the state government was forced to incorporate strict hygiene laws, mask obligation, minimum distance, schook closure and many other measures (27.09.2020 - 27.02.2021).

4. Period The fourth period smoothly picks up from the third, as case numbers were high compared to the start of the pandemic. Same measures as before were taken, as there was no medication or vaccination stocked at that time. As the third wave was beginnig to decrease, older generations were a top priority to be vaccinated. (28.02.2021 - 26.05.2021).

6.2 Analysis of Events based on Estimations

The following tables show the point estimates for the probability, calculated with the MH Algorithm, and will present the credible intervals. The data selected represents the characteristics of: state = "Rheinland-Pfalz", Age = 80+.

Focussing on this age group for two reasons: First, the data quality is better, as senior citizens are more precautious and would instead alert the authorities of Covid Infections, unlike younger generations. Additionally, medical care is necessary at this age, and therefore, the reports of this age group are qualitatively better. Secondly, the measurements are predominantly used to protect senior citizens. Vaccines were available only for seniors and disease carriers with a precondition at that time. Therefore, those measurements have no lagged effect, as the age group is immediately affected.

Period	80 % Credibility Interval
1. Wave	{0.042, 0.151}
1. Cooldown	{0.055, 0.197}
2. Wave	{0.034, 0.119}
3. Wave	{0.055, 0.197}

This table depicts the credibility intervals for the respective period. One can see, that the intervals have almost no meaning, as the bounds barely differ in significance. Therefore, one hast to consider the development of estimates and compare them to the data ratios.

The table down below provides precisely these values. The estimates are all highly similar, and the ratios tend to showcase the same pattern. The following graph will

	estimates_MH	ratios	difference
1	0.0928	0.1170	0.0242
2	0.0934	0.1200	0.0265
3	0.0946	0.1185	0.0240
4	0.0937	0.1143	0.0206

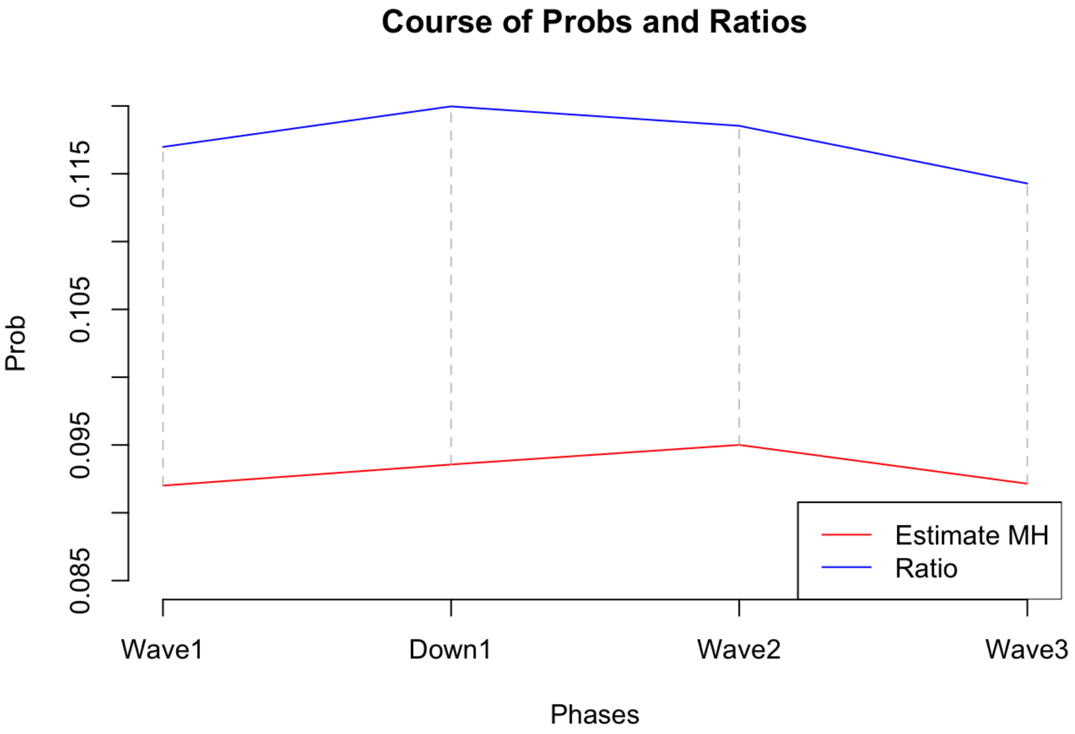


Figure 7: Course of death probabilities

show the differences in estimation over time. The development for the Bayesian estimate is concave, which means that the model assumes a peak in death ratios, even after implementing strict measures and rules. The data suggests that the government was dealing with a lack of information and resources against the deadly virus. However, Figure 7 also shows a decrease in the death rate, which suggests that especially the vaccines and more understanding of the virus led to an improved value of death - rates. The RKI explained the central part of this development in that specific age group to the progress and acceptance of the vaccine (RKI (2022a)). However, as margins are small, the value of expression these statements has to be questioned.

7 Conclusion and Comments

As shown, the Bayesian Model estimation technique helps deal with significant data discrepancies and allows the implementation of prior gathered information. However, it is also clear that the MH-Algorithm was too dependent on said prior, which underlines the criticism of the approach. Generally, the Bayesian approach is a viable solution to estimating death rates. Nonetheless, it brings up difficulties. Dealing with data representing ratios close to zero might be challenging to illustrate. As seen in the prior selection, without the usage of the beta prior, the estimation would be arbitrary:

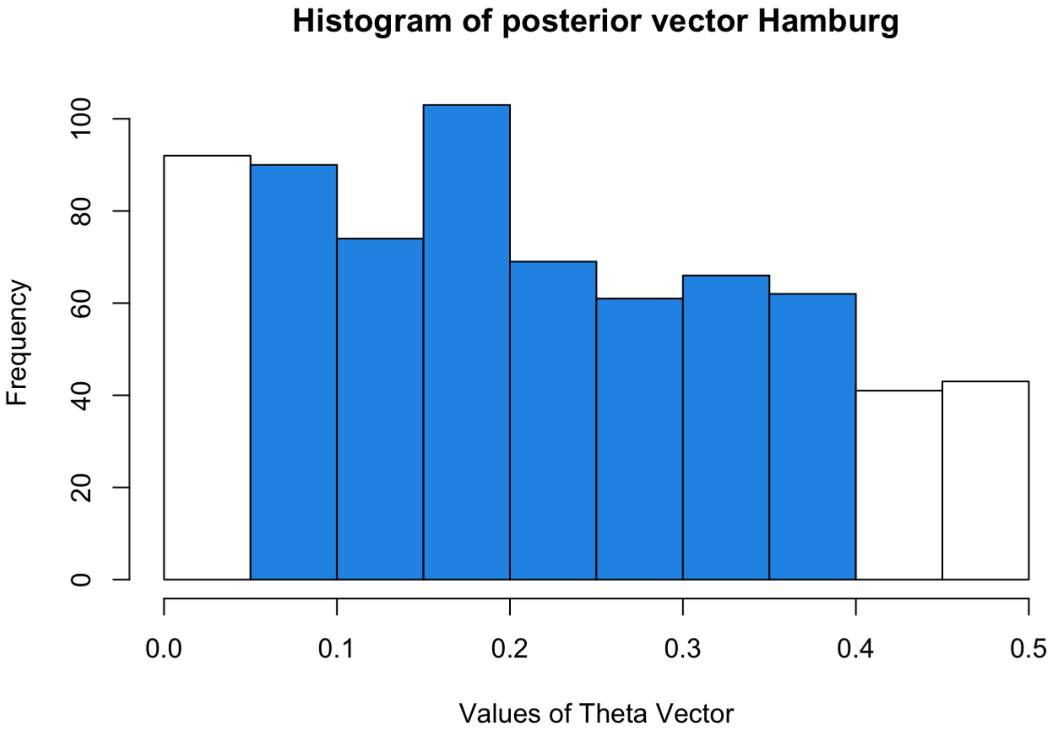


Figure 8: Posterior vector with a Uniform(0,0.5) Prior distribution

Here, the estimation represents the data from ”Hamburg”. The evidence only marginally influences the posterior, and the posterior still has no meaningful intuition. Therefore, the only viable solution would always be selecting the beta prior. Moreover, as mentioned in section 5.6, the best pick of the Bayesian Estimation technique would be to estimate the death rates without a more extensive computational effort. The modeling already indulges an abundance of cases. Therefore, the evidence is plenty so that the Bayesian model can learn from the given data.

Another factor that might obscure the estimation quality is the setting of the algorithm. If using the MH Algorithm, it is highly possible to misspecify priors. As shown in Chapter 5.4, the posterior distribution is highly dependent on the prior (also depending on the number of simulations), which means that the technique is time-consuming and biased computationally.

Nonetheless, the algorithm provides a good approximation, as it is only a mixture between the initial values of the prior and the evaluation through the likelihood of the evidence. With partitioning and asymmetric data, the Bayesian Model has opportunities to go deeper in the analysis, as one can compare different data sizes(Chapter 5.2). Also, Chapter 7. depicts a deeper viewing of the estimation. During these periods, the number of cases and deaths varied enormously. The Bayes estimate provided sample independent estimations, which could be evaluated against other periods. Therefore, conducting the Bayesian Model Estimation with sequential data is reasonable. One can identify certain developments commensurate to the evidence, measures, and other influences. Additionally, if a rolling data-set is implemented, the estimate will automatically converge to the evidence as seen in the prior and evidence importance (Chapter 5.1). Hence, the technique is justified in that setting.

A Appendix

A.1 Point Estimation Bayes

$$\epsilon(\hat{\theta}) := E[p((\hat{\theta} - \theta)^2 | D)] \tag{1}$$

Using optimization, we derive the latter equation and set it to zero:

$$\frac{d\epsilon}{d\hat{\theta}} = \frac{d}{d\hat{\theta}} \int_{\Theta} (\hat{\theta} - \theta)^2 p(\theta|D) d\theta = 0$$

$$2 \times \int_{\Theta} (\hat{\theta} - \theta) p(\theta|D) d\theta = 0$$

$$\hat{\theta} \int_{\Theta} p(\theta|D) d\theta = \int_{\Theta} \theta p(\theta|D) d\theta$$

The LHS term $\int_{\Theta} p(\theta|D) d\theta$ corresponds to the scaling condition, from which we then know that it has to be 1:

$$\hat{\theta} = \int_{\Theta} \theta p(\theta|D) d\theta = E[p(\theta|D)]$$

A.2 Calculate Posterior

Starting with Bayes' Theorem:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int_{\Theta} p(D|\theta)p(\theta) d\theta}$$

Set $p(D|\theta) = L(D|\theta)$ as the likelihood, and $\int_{\Theta} p(D|\theta)p(\theta)d\theta = m(D)$ as the marginal distribution of the evidence (data). The marginal distribution is equal for every characteristic of the binomial distribution as the Bernoulli experiments are i.i.d. against each other. Set $\pi(\theta|D)$ as the posterior:

$$\pi(\theta|D) = \frac{L(D|\theta)\pi(\theta)}{m(D)}$$

Plugging in the set distribution in Bayesian model:

$$\pi(p|D) = \binom{I}{D} p^D (1-p)^{I-D} \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} p^{\alpha_0-1}(1-p)^{\beta_0-1}\frac{1}{m}$$

Now using the Kernel method and only consider the fundamental and crucial part of the distribution (Gelman (2014)):

$$\pi(p|D) = a \times p^{D+\alpha_0-1}(1-p)^{I-D+\beta_0-1}$$

The Kernel method only extracts the decisive information and treats the rest of the posterior, which is analytically intractable due to the marginal distribution, as a constant. As known, the posterior is often defined as a proportional function, thus one can leave out the constant a, and proceed to analyze the Kernel of the posterior. It is apparent that the posterior follows a beta distribution depending on the parameters α_0 and β_0 predefined in the prior beta distribution:

$$\pi(p|D) \propto p^{D+\alpha_0-1}(1-p)^{I-D+\beta_0-1}$$

$$p|D \sim \text{Beta}(D+\alpha_0, I-D+\beta_0)$$

.

A.3 Weighted Probability Estimator

Derivation of the weighted Bayes estimator is as follows:

As introduced in A.1, the best point predictor is the expected value of the posterior distribution estimated, which in this case is derived from the beta distribution with:

$$E_{\text{Beta}} = \frac{\alpha}{\alpha+\beta}$$

with:

$$p|d \sim \text{Beta}(D+\alpha_0, I-D+\beta_0)$$

. Incorporating posterior values gives:

$$E[p_{\text{death}}|D] = \frac{D+\alpha_0}{(D+\alpha_0)+(I-D+\beta_0)}$$

$$= \frac{D+\alpha_0}{\alpha_0+\beta_0+I}$$

$$= \frac{D}{I}(\frac{I}{\alpha_0+\beta_0+I}) + (\frac{\alpha_0+\beta_0}{\alpha_0+\beta_0+I})(\frac{\alpha_0}{\alpha_0+\beta_0})$$

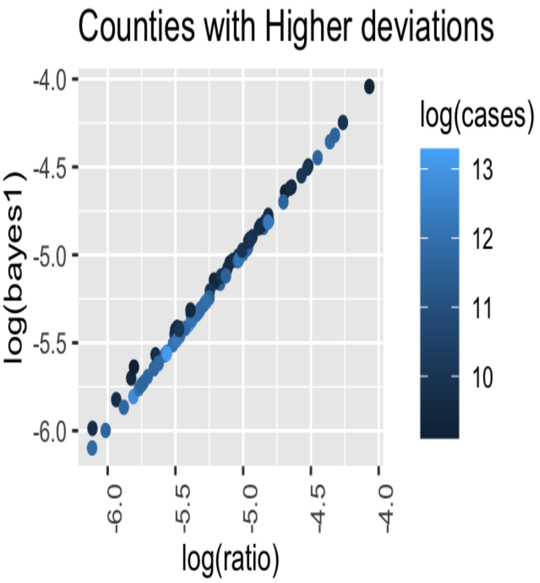
With $D = \sum d_i$, summation of all deaths:

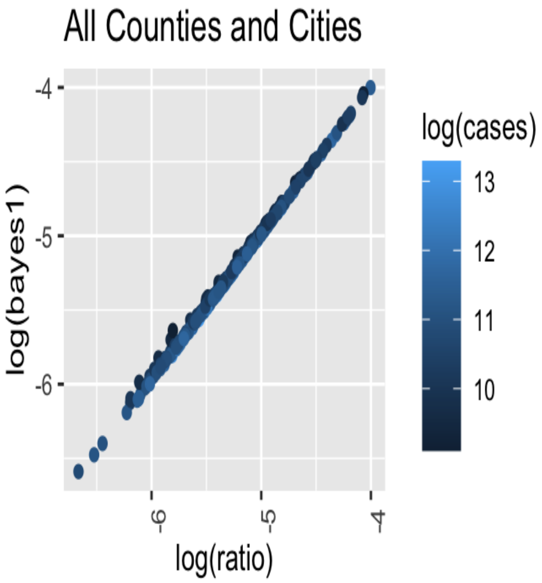
$$(\frac{I}{\alpha_0+\beta_0+I})\bar{D} + (\frac{\alpha_0+\beta_0}{\alpha_0+\beta_0+I})(\frac{\alpha_0}{\alpha_0+\beta_0})$$

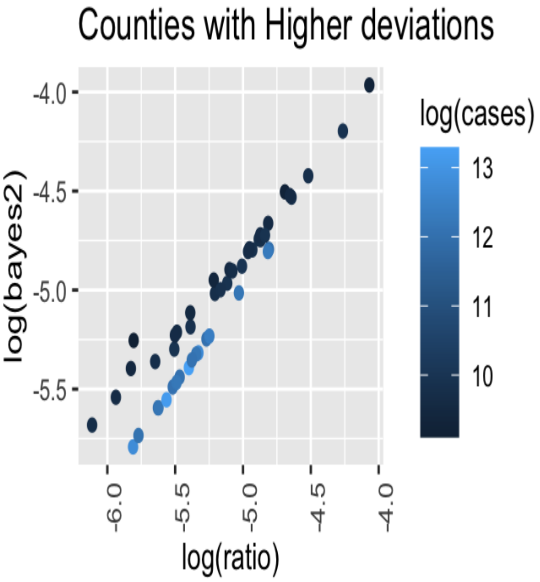
In this formula, one can see the trade-off between the importance of the data and the prior. As the evidence, thus I, rises, the importance of the Left-hand-side term will increase and will neglect the prior.

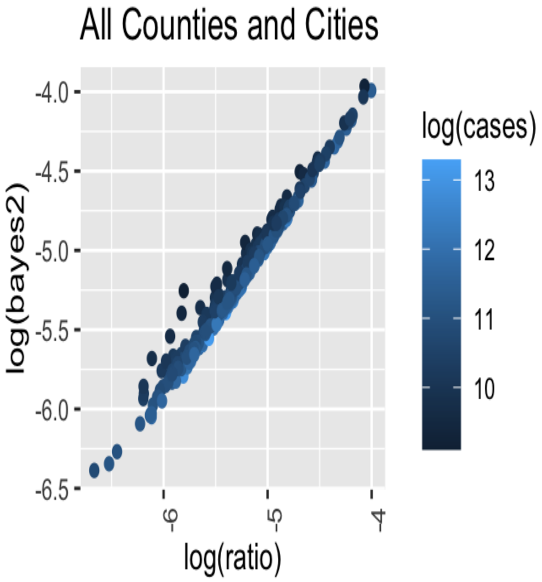
B Output results

In this section the output results for each of the age-groups, gender, region and time variation will be shown, with data provided by RKI (2022b)









B.0.1 Age Group Estimation

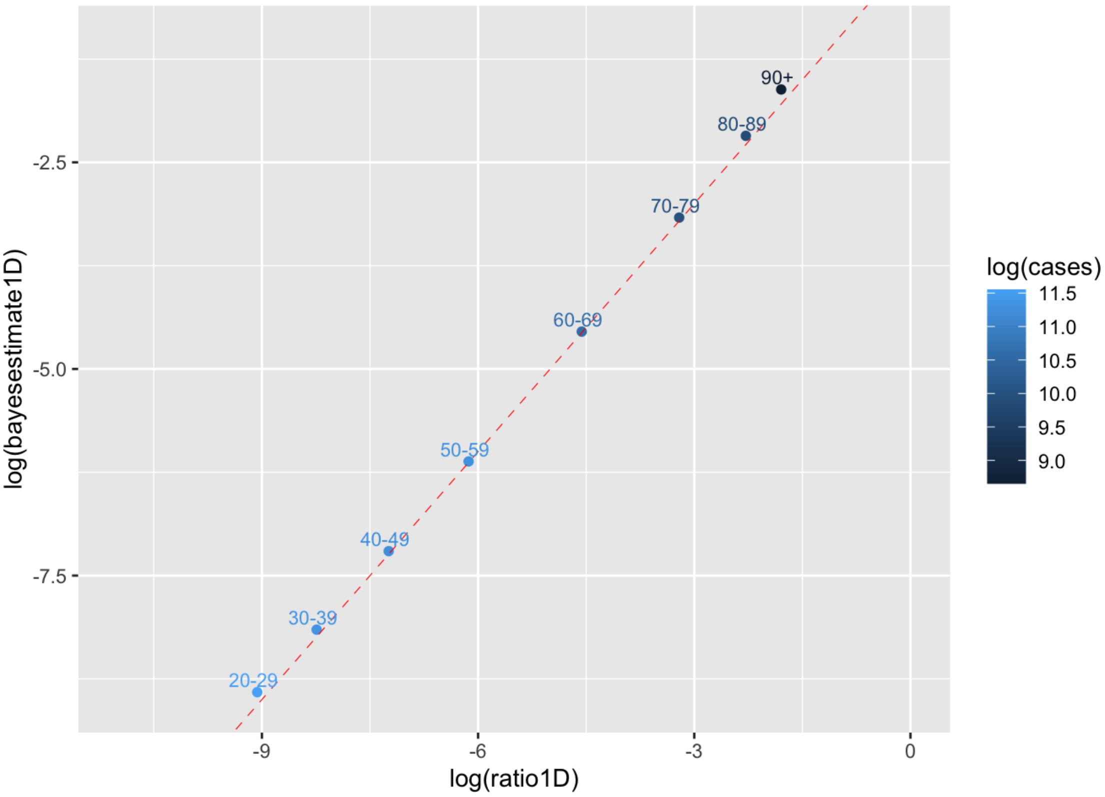


Figure 9: Age-Groups death rates in First Down phase

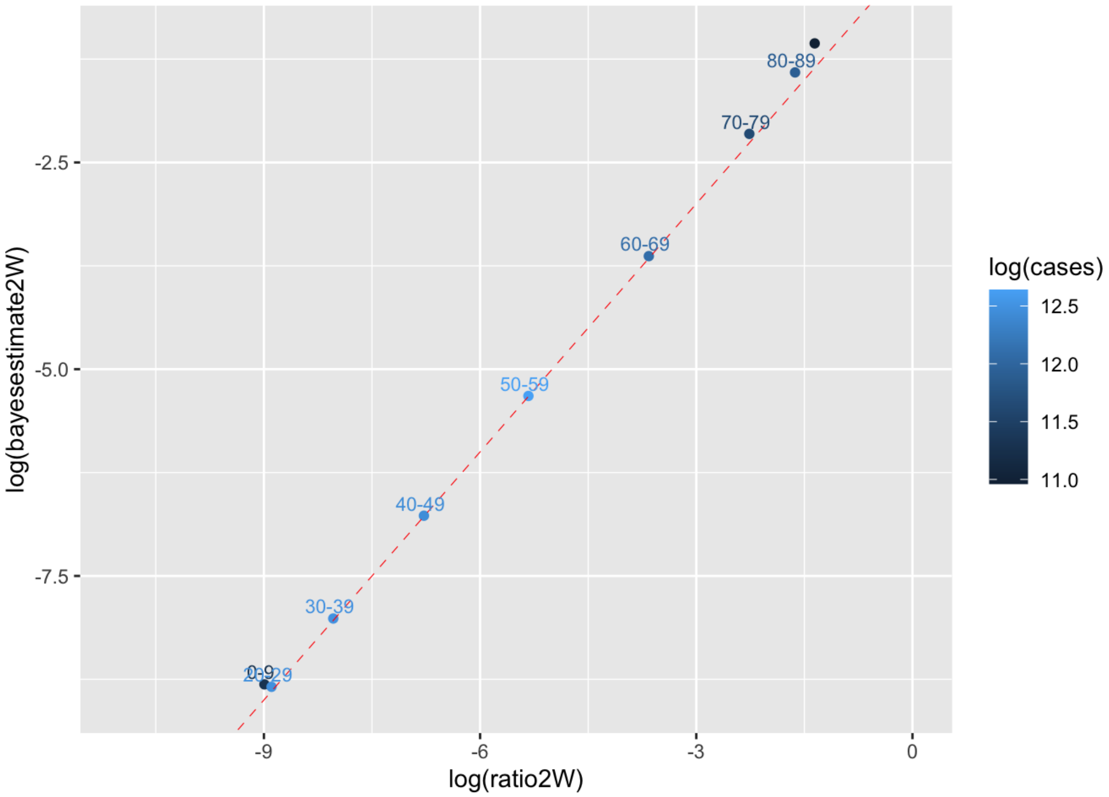
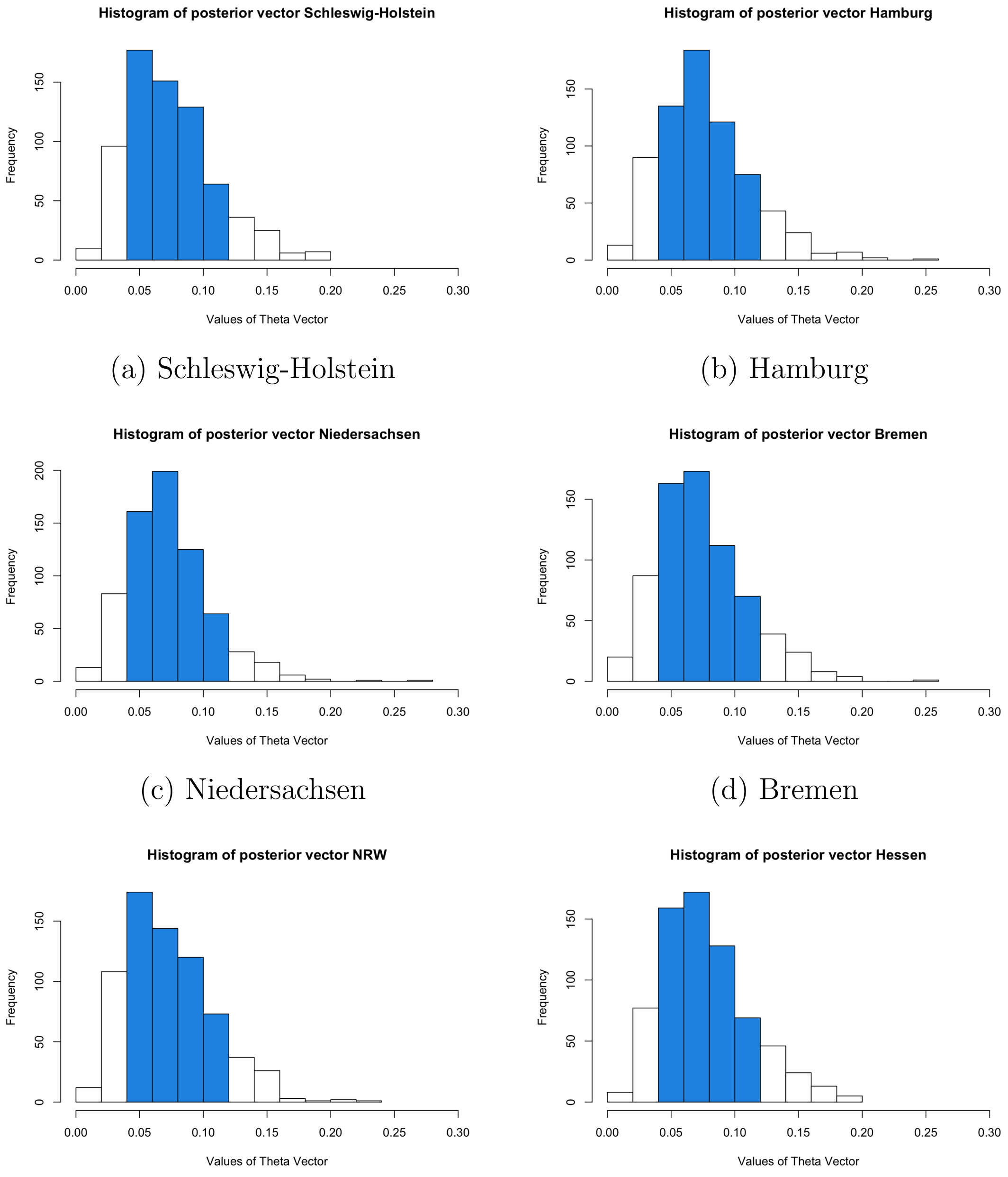


Figure 10: Age-Groups death rates in second Wave

B.0.2 Credible Intervals

”



(a) Schleswig-Holstein

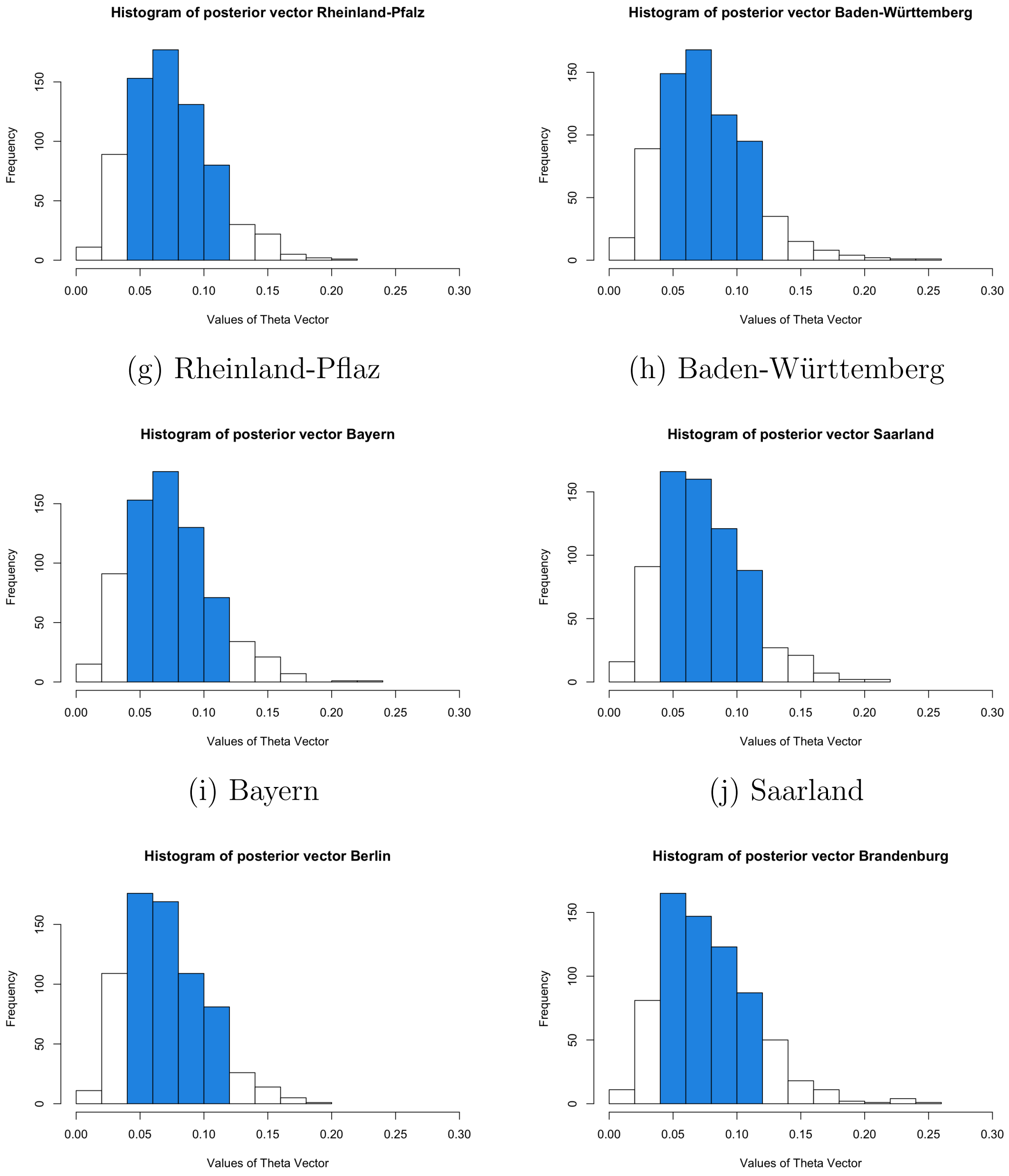
(b) Hamburg

(c) Niedersachsen

(d) Bremen

(e) NRW

(f) Hessen



(g) Rheinland-Pflaz

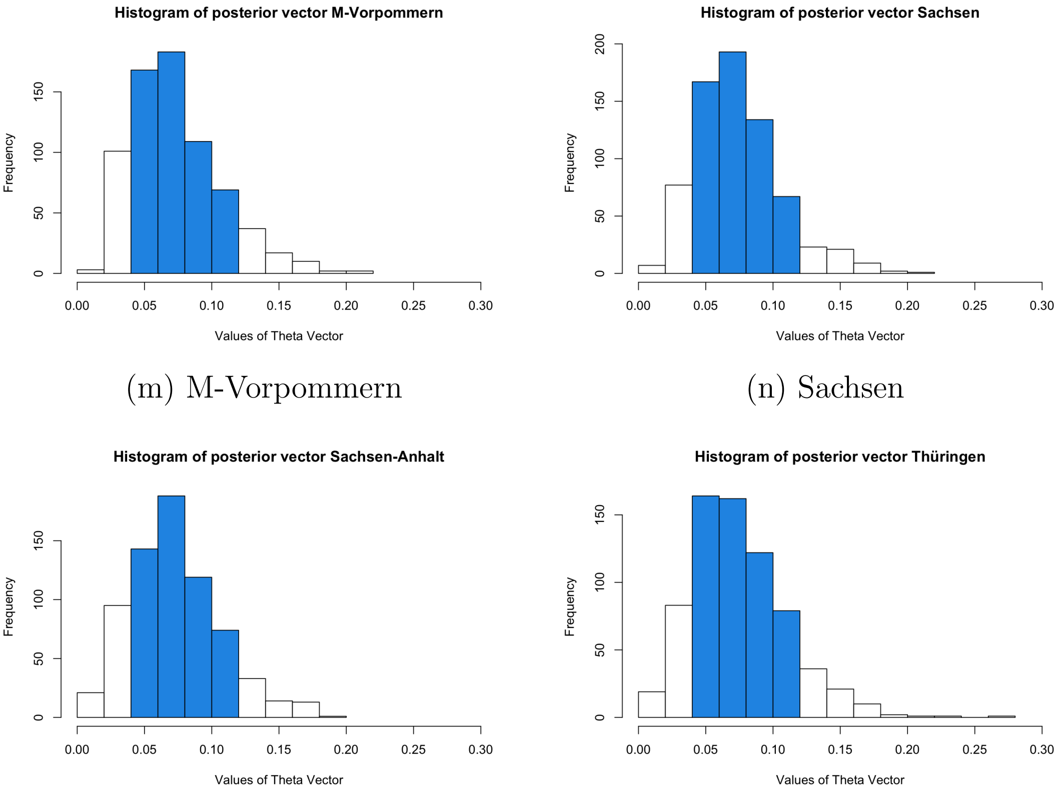
(h) Baden-Württemberg

(i) Bayern

(j) Saarland

(k) Berlin

(l) Brandenburg



(m) M-Vorpommern

(n) Sachsen

(o) Sachsen-Anhalt

(p) Thürigen

Figure 11: Credibility Intervals

C Code

Death-Rate Estimation

Shreyas Gupta

2022-07-17

```
library(readxl)
Altersverteilung <- as.data.frame(read_excel("Altersverteilung.xlsx",
    sheet = "Fallzahlen"))

Todesfaelle <- as.matrix(read_excel("COVID-19_Todesfaelle.xlsx",
                                    sheet = "COVID_Todesfälle_KW_AG10" , ))

for(k in 1:119){
  for(l in 1:12){
    if(Todesfaelle[k,l] == "<4"){
       Todesfaelle[k,l] = 2 #set the mean of the uncertain
                            #data as RKI only documented inequalities
    }
    Todesfaelle[k,l] = as.numeric(Todesfaelle[k,l])
  }
}
```

Merge and order both data frames in order to receive homogeneous data set of death ratios for each age group

```
row.names(Altersverteilung) <- Altersverteilung[,1]
Altersverteilung <- as.matrix(Altersverteilung[-1,][,-1])
Altersverteilung <- Altersverteilung[,-(dim(Altersverteilung)[2]:
                                          (dim(Altersverteilung)[2]-1))]

X <- matrix(0, nrow = 10, ncol = 119)
rownames(X) <- rev(c("0-9","10-19","20-29","30-39",
                     "40-49","50-59","60-69","70-79","80-89","90+"))
colnames(X) <- colnames(Altersverteilung)[1:119]
X[1,] <- Altersverteilung[1,]
```

Defining Y as Death Count and X as Infections: Combining the age group to 10 year sequences: Adhering to the death-dataset.

```
k <- 1
for (i in seq(2,19,2)){

  k = k+1

  X[k,] <- Altersverteilung[i,]+ Altersverteilung[(i+1),]

}
X<- X[rev(1:10),]
```

X is now the new Age distribution for Infections over the period of the pandemic.

```
Y <- Todesfaelle[,-(1:2)]
Y <- apply(Y, 2, as.numeric)
Y <- t(Y)


rownames(Y) <- rownames(X)
colnames(Y) <- colnames(X)
```

Now, both, *Todesfaelle* and *Fallzahl*, are now the same mode and the same dimension.

Bayesian Estimation

Calculating the Ratios per Age-Group over all time at first. Afterwards regardig the different time-periods and measures taken for each age group.

```
X_sum <- apply(X, 1, sum)
Y_sum <- apply(Y, 1, sum)

d_ratios <- Y_sum/X_sum
```

Now modeling the deaths as a Binomial Model regarding Infections:

Setting a Prior –> Beta Prior as Conjugate for Binomial Naming the Likelihood

```
plot(dbeta(x = seq(0,1, le=100),shape1 = 2, shape2 = 40))

alpha_0 <- 2
beta_0 <- 40

bayes_estimate <- (Y_sum + alpha_0)/(X_sum - Y_sum + beta_0)

comparison_frame <- data.frame(bayesestimate = bayes_estimate,
                               ratio = d_ratios, cases = X_sum)
```

```
require(ggplot2)
```

```
ggplot(data = comparison_frame, aes(x = log(ratio), y = log(bayes_estimate),
                                    colour = log(cases)))+
  geom_point()+
  geom_abline(lty = 2, col= "red", lwd= .2)+
  geom_text(
    label=rownames(comparison_frame),
    check_overlap = F,
    size = 3,
    nudge_x = -0.05, nudge_y = 0.15,
  )
```

Now performing same estimation with different data window.

Covid Wave 1 2020 Week 10 - Week 19

```
X_1Wave <- X[,1:10]
Y_1Wave <- Y[,1:10]

X_sum1W <- apply(X_1Wave, 1, sum)
Y_sum1W <- apply(Y_1Wave, 1, sum)
```

```
d_ratios_1W <- Y_sum1W/X_sum1W


bayes_estimate_1W <- (Y_sum1W + alpha_0)/(X_sum1W - Y_sum1W + beta_0)
comparison_frame_1W <- data.frame(bayesestimate1W = bayes_estimate_1W,
                                  ratio1W = d_ratios_1W, cases = X_sum1W)
```

Plotting differences

```
ggplot(data = comparison_frame_1W, aes(x = log(ratio1W),
                                       y = log(bayesestimate1W),colour = log(cases)))+
  geom_point()+
  geom_abline(lty = 2, col= "red", lwd= .2)+
  geom_text(
    label=rownames(comparison_frame_1W),
    check_overlap = F,
    size = 3,
    nudge_x = -0.05, nudge_y = 0.15,
  )+
  xlim(-11,0)+
  ylim(-9,-1)



summary(comparison_frame_1W)
```

Now downphase: from 2020 Week 20 - Week 45

```
X_1Down <- X[,11:36]
Y_1Down <- Y[,11:36]

X_sum1D <- apply(X_1Down, 1, sum)
Y_sum1D <- apply(Y_1Down, 1, sum)

d_ratios_1D <- Y_sum1D/X_sum1D


bayes_estimate_1D <-(Y_sum1D + alpha_0)/(X_sum1D - Y_sum1D + beta_0)
comparison_frame_1D <- data.frame(bayesestimate1D = bayes_estimate_1D,
                                  ratio1D = d_ratios_1D, cases = X_sum1D)
```

```
ggplot(data = comparison_frame_1D, aes(x = log(ratio1D),
                                       y = log(bayesestimate1D),
                                       colour = log(cases)))+
  geom_point()+
  geom_abline(lty = 2, col= "red", lwd= .2)+
  geom_text(
    label=rownames(comparison_frame_1D),
    check_overlap = F,
    size = 3,
    nudge_x = -0.05, nudge_y = 0.15,
  )+
  xlim(-11,0)+
  ylim(-9,-1)
```

2 Wave 2020 Week 46 - 2021 Week 9

```
X_2Wave <- X[,37:53]
Y_2Wave <- Y[,37:53]

X_sum2W <- apply(X_2Wave, 1, sum)
Y_sum2W <- apply(Y_2Wave, 1, sum)

d_ratios_2W <- Y_sum2W/X_sum2W


bayes_estimate_2W <- (Y_sum2W + alpha_0)/(X_sum2W - Y_sum2W + beta_0)
comparison_frame_2W <- data.frame(bayesestimate2W = bayes_estimate_2W,
                                  ratio2W = d_ratios_2W, cases = X_sum2W)
```

Plot

```
ggplot(data = comparison_frame_2W, aes(x = log(ratio2W),
                                       y = log(bayesestimate2W),
                                       colour = log(cases)))+
  geom_point()+
  geom_abline(lty = 2, col= "red", lwd= .2)+
  geom_text(
    label=rownames(comparison_frame_2W),
    check_overlap = F,
    size = 3,
    nudge_x = -0.05, nudge_y = 0.15,
  )+
  xlim(-11,0)+
  ylim(-9,-1)
```

(MCMCM scheme: different Priors)

Using MCMC Metropolis Hastings Algorithm to implement uniform(beta) uninformative prior

Using the data defined in the first wave and evaluating posterior for age group 90+ :

```
X_1Wave <- X[,1:10]; Y_1Wave <- Y[,1:10]
X_sum1W <- apply(X_1Wave, 1, sum); Y_sum1W <- apply(Y_1Wave, 1, sum)


proposal_frame <- data.frame(Infections = X_sum1W, Deaths = Y_sum1W)
```

Age Group data:

```
group1_dframe <- c(rep(1,proposal_frame[10,2]),rep(0,proposal_frame[10,1]))
group2_dframe <- c(rep(1,proposal_frame[9,2]),rep(0,proposal_frame[9,1]))
group3_dframe <- c(rep(1,proposal_frame[8,2]),rep(0,proposal_frame[8,1]))
group4_dframe <- c(rep(1,proposal_frame[7,2]),rep(0,proposal_frame[7,1]))
group5_dframe <- c(rep(1,proposal_frame[6,2]),rep(0,proposal_frame[6,1]))
group6_dframe <- c(rep(1,proposal_frame[5,2]),rep(0,proposal_frame[5,1]))
group7_dframe <- c(rep(1,proposal_frame[4,2]),rep(0,proposal_frame[4,1]))
group8_dframe <- c(rep(1,proposal_frame[3,2]),rep(0,proposal_frame[3,1]))
group9_dframe <- c(rep(1,proposal_frame[2,2]),rep(0,proposal_frame[2,1]))
group10_dframe <- c(rep(1,proposal_frame[1,2]),rep(0,proposal_frame[1,1]))
```

Defining a function for all age groups

```r
wrapper_MH_posterior <- function(n.sims = 1000, alpha_0 = 2, beta_0 = 20, data_frame){

  theta_0 <- rbeta(n = 1, shape1 = alpha_0,shape2 = beta_0)
  theta_vec <- rep(0, n.sims +1) #create posterior vector
  theta_vec[1] <- theta_0 #initialize posterior vector

  for (t in 1:n.sims){

    theta_star <- rbeta(n = 1, shape1 = alpha_0,shape2 = beta_0)#proposal from prior
    theta_t <- theta_vec[t] #previous accepted value for posterior

    #Create MH-Ratio using log-likelihoods

    MH_Ratio <- sum(dbinom(data_frame, size = 1, log = T, prob = theta_star))/
      sum(dbinom(data_frame, size = 1, log = T, prob = theta_t))

    #sample with MH ratio

    mh_prob <- c(min(1,MH_Ratio), 1-min(1,MH_Ratio))

    theta_vec[t+1] <- sample(x = c(theta_star, theta_t),size = 1,prob = mh_prob)
  }

  hist(theta_vec, probability = T)
  #lines(dbeta(x = seq(0,1, le=1000), shape1 = alpha_0, shape2 = beta_0))

  plot(theta_vec, type = "l", main = "data_frame")
  #return(list(posterior = theta_vec,
  #quantiles = quantile(theta_vec, probs = c(0.1,0.9)),
  #mean_posterior = mean(theta_vec)))

  return(mean(theta_vec))
}

#example

wrapper_MH_posterior(data_frame = group2_dframe, alpha_0 = 2 ,
                     beta_0 = 20, n.sims = 10000)
```

Scatter Plot All Counties

Shreyas Gupta

2022-07-29

Setup for data and estimation

```r
data <- read.csv(file = "RKI_Corona_Landkreise.csv", header = TRUE,sep = ",", dec = ".")

data.est = data.frame(data$GEN, data$deaths, data$cases)
data.est[,4] = data.est[,2]/data.est[,3]
colnames(data.est) = c("county", "deaths", "cases", "ratio")

##setting prior parameters such that estimates for only prior lie around either
## 0.2-0.9 --> experience actual shift in estimate

alpha_small <- 5
beta_small <- 10

alpha_big <- 20
beta_big <- 2

alpha_small_bayes <- data.est$deaths + alpha_small
beta_small_bayes <- data.est$cases + alpha_small + beta_small

alpha_big_bayes <- data.est$deaths + alpha_big
beta_big_bayes <- data.est$cases + alpha_big + beta_big

data.est[,5] <- alpha_big_bayes
data.est[,6] <- beta_big_bayes

data.est[,7] <- alpha_big_bayes
data.est[,8] <- beta_big_bayes

bayes_estimate1 = (data.est$deaths + alpha_small)/
  (data.est$cases+ alpha_small + beta_small)
bayes_estimate2 = (data.est$deaths + alpha_big)/
  (data.est$cases+ alpha_big + beta_big)

data.est[,9] = bayes_estimate1
data.est[,10] = bayes_estimate2

colnames(data.est) = c("county", "deaths", "cases", "ratio",
                       "alpha_small_bayes","beta_small_bayes","alpha_big_bayes"
                       ,"beta_big_bayes", "bayes1", "bayes2")

diff1 = log(abs(data.est$ratio - data.est$bayes1))
diff2 = log(abs(data.est$ratio - data.est$bayes2))
```

Understanding the reasoning behind selection of higher deviating states

```
hist(diff1)
hist(diff2)
```

Selecting all diffs in 1 from: < -10 or > -8.5 and diffs in 2: < -9 or > -7 in log values

Plotting the output

```
require(gridExtra)
require(ggplot2)

gg1 <- ggplot(data = data.est[diff1< -10|diff1> -8.5,])+
  geom_point(mapping = aes(log(ratio), log(bayes1), colour = log(cases)))+
  theme(axis.text.x = element_text(angle = 90, hjust = 2))+
  labs(title = "Counties with Higher deviations", fill = "N log cases")

gg2 <- ggplot(data = data.est)+
  geom_point(mapping = aes(log(ratio), log(bayes1), colour = log(cases)))+
  theme(axis.text.x = element_text(angle = 90, hjust = 2))+
  labs(title = "All Counties and Cities", fill = "N log cases")

gg3 <- ggplot(data = data.est[diff2< -9|diff2> -7,])+
  geom_point(mapping = aes(log(ratio), log(bayes2), colour = log(cases)))+
  theme(axis.text.x = element_text(angle = 90, hjust = 2))+
  labs(title = "Counties with Higher deviations", fill = "N log cases")

gg4 <- ggplot(data = data.est)+
  geom_point(mapping = aes(log(ratio), log(bayes2), colour = log(cases)))+
  theme(axis.text.x = element_text(angle = 90, hjust = 2))+
  labs(title = "All Counties and Cities", fill = "N log cases")

grid.arrange(gg1, gg2, gg3, gg4 ,ncol=2, nrow = 2)
```

Counties with the highest differences:

```
data$GEN[order(diff1, decreasing = T)[1:5]]
counties_top <- data.frame(data$GEN[order(diff2, decreasing = T)[1:15]],
                           data$BL[order(diff2, decreasing = T)[1:15]])
colnames(counties_top) <- c("County", "State")
counties_top

data$cases[order(diff1, decreasing = T)[1:5]]
mean(data$cases)


log_order <- log(order(diff1, decreasing = T)[1:5])
d1 <- data.est$bayes1[order(diff1, decreasing = T)[1:5]]
d2 <- data.est$bayes2[order(diff2, decreasing = T)[1:5]]
d3 <- data.est$ratio[order(diff1, decreasing = T)[1:5]]
d4 <- data.est$ratio[order(diff2, decreasing = T)[1:5]]
d5 <- data.est$alpha_small_bayes[order(diff1, decreasing = T)[1:5]]
d6 <- data.est$beta_small_bayes[order(diff1, decreasing = T)[1:5]]
d7 <- data.est$alpha_big_bayes[order(diff2, decreasing = T)[1:5]]
d8 <- data.est$beta_big_bayes[order(diff2, decreasing = T)[1:5]]
```

```
df <- data.frame(d1,d3,d5,d6,d2,d4,d7,d8, log_order)
colnames(df) <- c("bayes_1", "ratio","alpha1","beta1",
                  "bayes_2", "ratio","alpha2","beta2", "log_difference")
df
```

Presenting the posterior quantiles for the counties with the highest deviation

```
##small beta prior

for (i in 1:5){
  print(quantile(dbeta(seq(0,0.02, le = 10000), shape1 = df[i,3],
                       shape2 = df[i,4]), probs = c(0.25,0.75)))
}

##big beta prior

for (i in 1:5){
  print(quantile(dbeta(seq(0,0.02, le = 100), shape1 = df[i,7], shape2 = df[i,8]),
           probs = c(0.35,0.65)))
}
```

```
# with high prior
top_counties2 <- data$GEN[order(diff2, decreasing = T)[1:5]]

cases_top_deviate <- data.est$cases[order(diff2, decreasing = T)[1:5]]
death_top_deviate <- data.est$deaths[order(diff2, decreasing = T)[1:5]]
bayes_top_deviate <- data.est$bayes2[order(diff2, decreasing = T)[1:5]]
ratio_top_deviate <- data.est$deaths[order(diff2, decreasing = T)[1:5]]/
  data.est$cases[order(diff2, decreasing = T)[1:5]]

fraction_data_importance <- ((cases_top_deviate/
                            (alpha_big + beta_big + cases_top_deviate))*
                       (death_top_deviate/cases_top_deviate))/bayes_top_deviate

fraction_prior_importance <- (((alpha_big + beta_big)/
                              (alpha_big + beta_big + cases_top_deviate))*
                              (alpha_big/(alpha_big + beta_big)))/bayes_top_deviate




fraction_data_importance
fraction_prior_importance


##create prior-evidence exchange for estimation

df_interaction <- data.frame(row.names = top_counties2, fraction_data_importance,
                             fraction_prior_importance, cases_top_deviate,
                             death_top_deviate)
colnames(df_interaction) <- c("Evidence-Importance", "Prior Importance",
                              "Cases", "Deaths")
df_interaction
```

```
df_paper <- df[,5:9]
rownames(df_paper) <- top_counties2
colnames(df_paper) <- c("Bayes Estimate", "Death Ratio",
                        "Alpha Bayes", "Beta Bayes", "Log-Differences")


top_counties_10 <- data.frame(data$GEN[order(diff2, decreasing = T)[1:10]],
                              data$BL[order(diff2, decreasing = T)[1:10]])
colnames(top_counties_10) <- c("County",  "State")

#require(xtable)
#xtable(df_interaction)
#xtable(df_paper, digits = 4)
#xtable(top_counties_10)


importance <- data.frame(data$GEN[order(diff2, decreasing = T)[1:100]],
                         data$BL[order(diff2, decreasing = T)[1:100]])
colnames(importance) <- c("County", "State")
sum(importance$State=="Rheinland-Pfalz")/100
#23% of the data entries with the highest bayesian differences
#are from Rheinland-Pfalz therefore it seems sensible to lead with that state
```

Different-Likelihood-Poisson

Shreyas Gupta

2022-07-29

```r
Data <- read.csv(file = "Data.csv", header = TRUE, sep = ",", dec = ".")


n.states <- c("Schleswig-Holstein", "Hamburg","Niedersachsen","Bremen",
              "NRW","Hessen","Rheinland-Pfalz","Baden-Württemberg","Bayern","Saarland",
              "Berlin","Brandenburg","M-Vorpommern","Sachsen","Sachsen-Anhalt","Thüringen")

n.age <- c("A00-A04", "A05-A14", "A15-A34", "A35-A59", "A60-A79", "A80+")
gender <- c("M", "W", "unbekannt")

Data$Refdatum = as.Date(Data$Ref, format = "%Y-%m-%d")
```

```r
estimation <- function(data = Data, Bundeslaender = TRUE, Bundesland_Klasse,
                       Landkreise = FALSE,Landkreis_Klasse,Geschlechts_Klasse,
                       Alter_Klasse, alpha_0, beta_0, MH_algorithm = TRUE,
                       n.sims = 1000, time.stamp, Zeitpunkt){

  #####set prior distribution

  expected_prior <- alpha_0/(alpha_0 + beta_0)

  ##prior is set as defined distribution which will be drawn from random

  if(Bundeslaender){
    frame <- data[(data$Bundesland == Bundesland_Klasse),]
  }
  if(Landkreise){
    frame <- data[(data$IdLandkreis == Landkreis_Klasse),]
  }
  frame <- frame[(frame$Geschlecht == Geschlechts_Klasse),]

  frame <- frame[(frame$Altersgruppe == Alter_Klasse),]

  frame <- frame[,c(5,10,11)]

  time_frame <- as.Date(frame[,1])
  if(time.stamp){
    zeitpunkt_ende <- rev(time_frame)[1]
    zeitpunkt_neu = as.Date(Zeitpunkt)
    pos <- rev(which(time_frame %in% as.Date(zeitpunkt_ende,
                                                 format = "%Y-%m-%d")))[1]
    if(zeitpunkt_ende < zeitpunkt_neu){
      frame <- frame[1:pos ,]
  }
  else{
    while(sum(frame$Refdatum == Zeitpunkt, na.rm = T) == 0){
      zeitpunkt_neu <- zeitpunkt_neu +1
      zeitpunkt_neu <- as.Date(zeitpunkt_neu)
   }
  end_date <- rev(which(time_frame %in% zeitpunkt_neu))[1]
  frame <- frame[1:end_date ,]
  }
}
else{
  zeitpunkt_ende <- rev(time_frame)[1]
  pos <- rev(which(time_frame %in% as.Date(zeitpunkt_ende,
                                               format = "%Y-%m-%d")))[1]
  frame <- frame[1:pos ,]
}

specific_data <- apply(frame[,2:3], 2, sum)
ratio <- specific_data[2]/specific_data[1]

##build the data respective data

data_distribution <- c(rep(0, specific_data[1]), rep(1, specific_data[2]))

####MCMC
theta_vec1 <- NULL

if(MH_algorithm){

  theta_0 <- rbeta(n = 1, shape1 = alpha_0, shape2 = beta_0)
  theta_vec1 <- rep(0, n.sims+1) #create posterior vector
  theta_vec1[1] <- theta_0 #initialize posterior vector

  for (t in 1:n.sims){

    theta_star <- rbeta(n = 1, shape1 = alpha_0, shape2 = beta_0) #proposal from prior
    theta_t <- theta_vec1[t] #previous accepted value for posterior

    #Create MH-Ratio using log-likelihoods and taking the inverse of each likelihood
    #to penalize higher negative values as outcomes with a smaller likelihood

    MH_Ratio <- sum(dpois(x = data_distribution, lambda = specific_data[1]*theta_t, log = T))/
sum(dpois(x = data_distribution, lambda = specific_data[1]*theta_star, log = T))

    #sample with MH ratio

    mh_prob <- c(min(1,MH_Ratio), 1-min(1,MH_Ratio))

    theta_vec1[t+1] <- sample(x = c(theta_star, theta_t),size = 1,prob = mh_prob)
  }
  mean_pois <- mean(theta_vec1)
  quantiles1 <- quantile(theta_vec1, probs = c(0.1,0.9))
  #par(mfrow = c(2,2))
  h1 <- hist(theta_vec1[(0.3*n.sims):n.sims], plot = F, breaks = 9)
  cols <- c(0, 4, 0)
  k1<- cols[findInterval(h1$mids, quantile(theta_vec1, probs = c(0.1,0.9)),
                        rightmost.closed=T, all.inside=F)+1]
}

theta_vec2 <- NULL

if(MH_algorithm){

  theta_0 <- rbeta(n = 1, shape1 = alpha_0, shape2 = beta_0)
  theta_vec2 <- rep(0, n.sims+1) #create posterior vector
  theta_vec2[1] <- theta_0 #initialize posterior vector

  for (t in 1:n.sims){

    theta_star <- rbeta(n = 1, shape1 = alpha_0, shape2 = beta_0) #proposal from prior
    theta_t <- theta_vec2[t] #previous accepted value for posterior

    #Create MH-Ratio using log-likelihoods and taking the inverse of each likelihood
    #to penalize higher negative values as outcomes with a smaller likelihood

    MH_Ratio <- sum(dbinom(x = data_distribution, prob = theta_t, log = T, size = 1))/
sum(dbinom(x = data_distribution, prob = theta_star, log = T, size = 1))

    #sample with MH ratio

    mh_prob <- c(min(1,MH_Ratio), 1-min(1,MH_Ratio))

    theta_vec2[t+1] <- sample(x = c(theta_star, theta_t),size = 1,prob = mh_prob)
  }
  mean_binom <- mean(theta_vec2)
  quantiles2 <- quantile(theta_vec2, probs = c(0.1,0.9))
  #par(mfrow = c(2,2))

  h2 <- hist(theta_vec2[(0.3*n.sims):n.sims], plot = F, breaks = 9)
  cols <- c(0, 4, 0)
  k2<- cols[findInterval(h2$mids, quantile(theta_vec2, probs = c(0.1,0.9)),
                        rightmost.closed=T, all.inside=F)+1]

}

par(mfrow = c(2,2))
plot(h1, col = k1, main = paste("Poisson: Posterior vector", Bundesland_Klasse),
       xlab = "Values of Theta Vector",  xlim = c(0,0.2))
plot(h2, col = k2, main = paste("Binomial: Posterior vector", Bundesland_Klasse),
       xlab = "Values of Theta Vector", xlim = c(0,0.2))
plot(theta_vec1[(0.3*n.sims):n.sims], type = "l", ylab = "Mixing",
       main = "Mixing of posterior values")
plot(theta_vec2[(0.3*n.sims):n.sims], type = "l", ylab = "Mixing",
       main = "Mixing of posterior values")

return(list(mean = c(mean_pois, mean_binom), prior_exp = expected_prior, ratio = ratio))
```

```
}
```

```
estimation(Bundesland_Klasse = "NRW", Alter_Klasse = "A35-A59",alpha_0 = 4
            ,beta_0 = 50, time.stamp = TRUE, Zeitpunkt = "2021-01-21",
           Geschlechts_Klasse = "W")
```

Comparison MH - Bootstrap

```
Data <- read.csv(file = "Data.csv", header = TRUE, sep = ",", dec = ".")


n.states <- c("Schleswig-Holstein", "Hamburg","Niedersachsen","Bremen",
              "NRW","Hessen","Rheinland-Pfalz","Baden-Württemberg","Bayern","Saarland",
              "Berlin","Brandenburg","M-Vorpommern","Sachsen","Sachsen-Anhalt","Thüringen")

n.age <- c("A00-A04", "A05-A14", "A15-A34", "A35-A59", "A60-A79", "A80+")
gender <- c("M", "W", "unbekannt")

Data$Refdatum = as.Date(Data$Ref, format = "%Y-%m-%d")
```

```
estimation <- function(data = Data, Bundeslaender = TRUE, Bundesland_Klasse,
                       Landkreise = FALSE,Landkreis_Klasse,
                       Geschlechts_Klasse, Alter_Klasse, alpha_0, beta_0,
                       MH_algorithm = TRUE, Bayes_Estimate = TRUE,
                       n.sims = 1000, time.stamp, Zeitpunkt, Bootstrap = TRUE,
                       bootstrap.sim = 1000){

  #####set prior distribution

  expected_prior <- alpha_0/(alpha_0 + beta_0)

  ##prior is set as defined distribution which will be drawn from random

  if(Bundeslaender){
    frame <- data[(data$Bundesland == Bundesland_Klasse),]
  }
  if(Landkreise){
    frame <- data[(data$IdLandkreis == Landkreis_Klasse),]
  }
  frame <- frame[(frame$Geschlecht == Geschlechts_Klasse),]

  frame <- frame[(frame$Altersgruppe == Alter_Klasse),]

  frame <- frame[,c(5,10,11)]

  time_frame <- as.Date(frame[,1])
  if(time.stamp){
    zeitpunkt_ende <- rev(time_frame)[1]
    zeitpunkt_neu = as.Date(Zeitpunkt)
    pos <- rev(which(time_frame %in% as.Date(zeitpunkt_ende,
                                                 format = "%Y-%m-%d")))[1]
    if(zeitpunkt_ende < zeitpunkt_neu){
      frame <- frame[1:pos ,]
    }
    else{
```

```
    while(sum(frame$Refdatum == Zeitpunkt, na.rm = T) == 0){
      zeitpunkt_neu <- zeitpunkt_neu +1
      zeitpunkt_neu <- as.Date(zeitpunkt_neu)
   }
  end_date <- rev(which(time_frame %in% zeitpunkt_neu))[1]
  frame <- frame[1:end_date ,]
  }
}
else{
  zeitpunkt_ende <- rev(time_frame)[1]
  pos <- rev(which(time_frame %in% as.Date(zeitpunkt_ende,
                                               format = "%Y-%m-%d")))[1]
  frame <- frame[1:pos ,]
}

specific_data <- apply(frame[,2:3], 2, sum)
ratio <- specific_data[2]/specific_data[1]

##build the data respective binomial distribution

data_distribution <- c(rep(0, specific_data[1]), rep(1, specific_data[2]))

##### Bayes estimation -- no MH

if(Bayes_Estimate){
  alpha_bayes <- specific_data[2] + alpha_0
  beta_bayes <- specific_data[1] - specific_data[2] + beta_0
  bayes_estimate <- alpha_bayes/(alpha_bayes + beta_bayes)
}

####MCMC MH
theta_vec <- NULL

if(MH_algorithm){

  theta_0 <- rbeta(n = 1, shape1 = alpha_0, shape2 = beta_0)
  theta_vec <- rep(0, n.sims+1) #create posterior vector
  theta_vec[1] <- theta_0 #initialize posterior vector

  for (t in 1:n.sims){

    theta_star <- rbeta(n = 1, shape1 = alpha_0, shape2 = beta_0) #proposal from prior
    theta_t <- theta_vec[t] #previous accepted value for posterior

    #Create MH-Ratio using log-likelihoods

    MH_Ratio <- sum(log(dbinom(data_distribution, size = 1, log = F, prob = theta_t)))/
      sum(log(dbinom(data_distribution, size = 1, log = F, prob = theta_star)))


    #sample with MH ratio

    mh_prob <- c(min(1,MH_Ratio), 1-min(1,MH_Ratio))
    theta_vec[t+1] <- sample(x = c(theta_star, theta_t),size = 1,prob = mh_prob)
  }
  mh_estimate <- mean(theta_vec)
  quantiles <- quantile(theta_vec, probs = c(0.1,0.9))
  h <- hist(theta_vec[(0.3*n.sims):n.sims], plot = F)
  cols <- c(0, 4, 0)
  k<- cols[findInterval(h$mids, quantile(theta_vec, probs = c(0.1,0.9)),
                        rightmost.closed=T, all.inside=F)+1]
  if(Bootstrap == F){
  par(mfrow = c(2,2))
  plot(h, col = k, main = paste("Histogram of posterior vector", Bundesland_Klasse),
       xlab = "Values of Theta Vector")


  plot(dbeta(seq(0,0.2,le=1000), shape1 = alpha_bayes, shape2 = beta_bayes),
       type = "l", ylab = "Probability", main = "Posterior Distribution")
  plot(x = 1:4, y = c(ratio, bayes_estimate, mh_estimate, expected_prior),
       col = 1:4, pch = 16, ylim = c(0,expected_prior + 0.01), xlim = c(0,5),
       ylab = "Probability", main = "Comparison between estimates")
  text(x = 1:4,
       y = c(ratio,bayes_estimate, mh_estimate,
             expected_prior)+0.006,c("ratio","bayes_estimate","mh_estimate",
                                     "expected_prior"), cex = .7)
  plot(theta_vec[(0.3*n.sims):n.sims], type = "l", ylab = "Mixing",
       main = "Mixing of posterior values")


return(list(Bayes_Estimate = bayes_estimate, MH_Estimate = mean(theta_vec),
            ratio = ratio, infection_data = specific_data,
            expected_prior = expected_prior))
  }
}
if(Bootstrap){

  B <- bootstrap.sim
  estimates <- NULL
  n.bootstrap <- length(data_distribution)

  for(b in 1:B){
    new_sample_data <- sample(data_distribution, size = n.bootstrap,
                              replace = TRUE)
    estimates[b] <- mean(new_sample_data)
  }

  h2 <- hist(estimates, plot = F)


par(mfrow = c(1,2))
  plot(h, col = k, main = paste("Histogram of posterior vector", Bundesland_Klasse),
      xlab = "Values of Theta Vector")
  plot(h2, main = "Bootstrap distribution" )

return(list(MH_Estimate = mean(theta_vec), quantiles = quantiles, ratio = ratio))
```

```
  }
}
```

Showing credible intervals and comparisons

```
estimation(Bundeslaender = T, Bundesland_Klasse = "Rheinland-Pfalz",
           Geschlechts_Klasse = "M", Alter_Klasse = "A35-A59", alpha_0 = 4,
           beta_0 = 20, Zeitpunkt = "2020-10-27", Bootstrap = F, time.stamp = T,
           MH_algorithm = T, Bayes_Estimate = T)
```

```
estimation(Bundeslaender = T, Bundesland_Klasse = "Rheinland-Pfalz",
           Geschlechts_Klasse = "M", Alter_Klasse = "A35-A59", alpha_0 = 4,
           beta_0 = 20, Zeitpunkt = "2020-10-27", Bootstrap = T, time.stamp = T)
```

Political Evaluation CHhapter 6

```
Data <- read.csv(file = "Data.csv", header = TRUE, sep = ",", dec = ".")


n.states <- c("Schleswig-Holstein", "Hamburg","Niedersachsen","Bremen",
              "NRW","Hessen","Rheinland-Pfalz","Baden-Württemberg","Bayern","Saarland",
              "Berlin","Brandenburg","M-Vorpommern","Sachsen","Sachsen-Anhalt","Thüringen")

n.age <- c("A00-A04", "A05-A14", "A15-A34", "A35-A59", "A60-A79", "A80+")
gender <- c("M", "W", "unbekannt")

Data$Refdatum = as.Date(Data$Ref, format = "%Y-%m-%d")
```

```
estimation <- function(data = Data, Bundeslaender = TRUE, Bundesland_Klasse,
                       Landkreise = FALSE,Landkreis_Klasse,
                       Geschlechts_Klasse, Alter_Klasse, alpha_0, beta_0,
                       MH_algorithm = TRUE, Bayes_Estimate = TRUE,
                       n.sims = 1000, time.stamp, Zeitpunkt, Zeitpunkt_Anfang){

  #####set prior distribution

  expected_prior <- alpha_0/(alpha_0 + beta_0)

  ##prior is set as defined distribution which will be drawn from random

  if(Bundeslaender){
    frame <- data[(data$Bundesland == Bundesland_Klasse),]
  }
  if(Landkreise){
    frame <- data[(data$IdLandkreis == Landkreis_Klasse),]
  }
  frame <- frame[(frame$Geschlecht == Geschlechts_Klasse),]

  frame <- frame[(frame$Altersgruppe == Alter_Klasse),]

  frame <- frame[,c(5,10,11)]

  time_frame <- as.Date(frame[,1])
  if(time.stamp){
    zeitpunkt_ende <- rev(time_frame)[1]
    zeitpunkt_anfang_daten <- time_frame[1]
    zeitpunkt_neu = as.Date(Zeitpunkt)
    zeitpunkt_anfang <- as.Date(Zeitpunkt_Anfang)

    if(sum(time_frame == zeitpunkt_neu, na.rm = T) == 0){
      while(sum(time_frame == zeitpunkt_neu, na.rm = T) == 0){
          zeitpunkt_neu <- zeitpunkt_neu +1
          zeitpunkt_neu <- as.Date(zeitpunkt_neu)
    }
  }
  if(zeitpunkt_anfang < zeitpunkt_anfang_daten){
    while(zeitpunkt_anfang < zeitpunkt_anfang_daten){
        zeitpunkt_anfang <- zeitpunkt_anfang +1
        zeitpunkt_anfang <- as.Date(zeitpunkt_anfang)
    }
  }
  if(sum(time_frame == zeitpunkt_anfang, na.rm = T) == 0){
    while(sum(time_frame == zeitpunkt_anfang, na.rm = T) == 0){
        zeitpunkt_anfang <- zeitpunkt_anfang +1
        zeitpunkt_anfang <- as.Date(zeitpunkt_anfang)
    }
  }
  start_date <- rev(which(time_frame %in% zeitpunkt_anfang))[1]
  end_date <- rev(which(time_frame %in% zeitpunkt_neu))[1]
  end_date_data <- rev(which(time_frame %in% zeitpunkt_ende))[1]


  if(zeitpunkt_ende < zeitpunkt_neu){
    frame <- frame[start_date:end_date_data ,]
  }
  else{
  frame <- frame[start_date:end_date ,]
  }
}
else{
  frame <- frame[1:end_date_data ,]
}

specific_data <- apply(frame[,2:3], 2, sum)
ratio <- specific_data[2]/specific_data[1]

##build the data respective binomial distribution

data_distribution <- c(rep(0, specific_data[1]), rep(1, specific_data[2]))

##### Bayes estimation -- no MH

if(Bayes_Estimate){
  alpha_bayes <- specific_data[2] + alpha_0
  beta_bayes <- specific_data[1] + beta_0
  bayes_estimate <- alpha_bayes/(alpha_bayes + beta_bayes)
}

####MCMC
theta_vec <- NULL

if(MH_algorithm){

  theta_0 <- rbeta(n = 1, shape1 = alpha_0, shape2 = beta_0)
  theta_vec <- rep(0, n.sims+1) #create posterior vector
  theta_vec[1] <- theta_0 #initialize posterior vector
```

```
    for (t in 1:n.sims){

      theta_star <- rbeta(n = 1, shape1 = alpha_0, shape2 = beta_0) #proposal from prior
      theta_t <- theta_vec[t] #previous accepted value for posterior

      #Create MH-Ratio using log-likelihoods and taking the inverse of each likelihood
      #to penalize higher negative values as outcomes with a smaller likelihood

      MH_Ratio <- sum(dbinom(data_distribution, size = 1, log = T, prob = theta_t))/
        sum(dbinom(data_distribution, size = 1, log = T, prob = theta_star))

      #sample with MH ratio

      mh_prob <- c(min(1,MH_Ratio), 1-min(1,MH_Ratio))

      theta_vec[t+1] <- sample(x = c(theta_star, theta_t),size = 1,prob = mh_prob)
    }
    mh_estimate <- mean(theta_vec)
    quantiles <- quantile(theta_vec, probs = c(0.1,0.9))
    #par(mfrow = c(2,2))

    h <- hist(theta_vec[(0.3*n.sims):n.sims], plot = F, breaks = 10)
    cols <- c(0, 4, 0)
    k<- cols[findInterval(h$mids, quantile(theta_vec, probs = c(0.1,0.9)),
                          rightmost.closed=T, all.inside=F)+1]
    plot(h, col = k, main = paste("Histogram of posterior vector", Bundesland_Klasse),
         xlab = "Values of Theta Vector", xlim = c(0,0.3))
  }

  return(list(quantiles = quantiles, estimate = mean(theta_vec), ratio = ratio))
}
```

Function description

The function will find the nearest exact dates, if the dataset does not allow the selection of that time point.Uses MH Algorithm. Returns quantiles and estimate.

Age Group = A80+ State = Rheinland-Pfalz

Estimation during the second wave from 01.03.2020 - 06.05.2021

Pandemic was declared 11.03.2020

```
W1 <- estimation(Bundesland_Klasse = "Rheinland-Pfalz", Geschlechts_Klasse = "M",
           Alter_Klasse = "A80+", alpha_0 = 4, beta_0 = 40, time.stamp = TRUE,
           Zeitpunkt_Anfang = "2020-03-01",Zeitpunkt = "2020-05-06", n.sims = 1000)
```

Estimation during the first down-phase from 07.05.2020 - 26.09.2020

```
D1 <- estimation(Bundesland_Klasse = "Rheinland-Pfalz", Geschlechts_Klasse = "M",
           Alter_Klasse = "A80+", alpha_0 = 4, beta_0 = 40, time.stamp = TRUE,
           Zeitpunkt_Anfang = "2020-05-07",Zeitpunkt = "2020-09-26", n.sims = 1000)
```

Estimation during the second wave from 27.09.2020 - 27.02.2021

```
W2 <- estimation(Bundesland_Klasse = "Rheinland-Pfalz", Geschlechts_Klasse = "M",
           Alter_Klasse = "A80+", alpha_0 = 4, beta_0 = 40, time.stamp = TRUE,
           Zeitpunkt_Anfang = "2020-09-27",Zeitpunkt = "2021-02-27", n.sims = 1000)
```

Estimation during the third wave from 28.02.2021 - 26.05.2021

```
W3 <- estimation(Bundesland_Klasse = "Rheinland-Pfalz", Geschlechts_Klasse = "M",
           Alter_Klasse = "A80+", alpha_0 = 4, beta_0 = 40, time.stamp = TRUE,
           Zeitpunkt_Anfang = "2021-02-28",Zeitpunkt = "2021-05-26", n.sims = 1000)
```

```
estimates_MH <- c(W1$estimate, D1$estimate, W2$estimate, W3$estimate)
ratios <- c(W1$ratio, D1$ratio, W2$ratio, W3$ratio)
differene <- ratios - estimates_MH

comp_frame <- data.frame(estimates_MH, ratios, differene)

#xtable(comp_frame, digits = 4)

plot(1:4, estimates_MH, type = "l", col = "red", ylim = c(0.085, 0.12), axes = F,
     xlab = "Phases", ylab = "Prob", main = "Course of Probs and Ratios")
legend(x = "bottomright", legend = c("Estimate MH", "Ratio"), lty = c(1,1), col = c("red", "blue"))
axis(1, at = 1:4, labels = c("Wave1","Down1","Wave2", "Wave3"))
axis(2)
lines(1:4, ratios, col = "blue")
segments(x0 = 1, y0 = estimates_MH[1], x1 = 1, y1 = ratios[1], lty = 2, lwd = .8, col = "gray")
segments(x0 = 2, y0 = estimates_MH[2], x1 = 2, y1 = ratios[2], lty = 2, lwd = .8, col = "gray")
segments(x0 = 3, y0 = estimates_MH[3], x1 = 3, y1 = ratios[3], lty = 2, lwd = .8, col = "gray")
segments(x0 = 4, y0 = estimates_MH[4], x1 = 4, y1 = ratios[4], lty = 2, lwd = .8, col = "gray")
```

Function

Shreyas Gupta

2022-08-02

```r
Data <- read.csv(file = "Data.csv", header = TRUE, sep = ",", dec = ".")


n.states <- c("Schleswig-Holstein", "Hamburg","Niedersachsen","Bremen",
              "NRW","Hessen","Rheinland-Pfalz","Baden-Württemberg","Bayern","Saarland",
              "Berlin","Brandenburg","M-Vorpommern","Sachsen","Sachsen-Anhalt","Thüringen")

n.age <- c("A00-A04", "A05-A14", "A15-A34", "A35-A59", "A60-A79", "A80+")
gender <- c("M", "W", "unbekannt")

Data$Refdatum = as.Date(Data$Ref, format = "%Y-%m-%d")
```

```r
estimation <- function(data = Data, Bundeslaender = TRUE, Bundesland_Klasse,
                       Landkreise = FALSE,Landkreis_Klasse,
                       Geschlechts_Klasse, Alter_Klasse, alpha_0, beta_0,
                       MH_algorithm = TRUE, Bayes_Estimate = TRUE,
                       n.sims = 1000, time.stamp, Zeitpunkt, Zeitpunkt_Anfang,
                       beta = T, uniform = F, margins){

  #####set prior distribution

  expected_prior <- alpha_0/(alpha_0 + beta_0)

  ##prior is set as defined distribution which will be drawn from random

  if(Bundeslaender){
    frame <- data[(data$Bundesland == Bundesland_Klasse),]
  }
  if(Landkreise){
    frame <- data[(data$IdLandkreis == Landkreis_Klasse),]
  }
  frame <- frame[(frame$Geschlecht == Geschlechts_Klasse),]

  frame <- frame[(frame$Altersgruppe == Alter_Klasse),]

  frame <- frame[,c(5,10,11)]

  time_frame <- as.Date(frame[,1])
  if(time.stamp){
    zeitpunkt_ende <- rev(time_frame)[1]
    zeitpunkt_anfang_daten <- time_frame[1]
    zeitpunkt_neu = as.Date(Zeitpunkt)
    zeitpunkt_anfang <- as.Date(Zeitpunkt_Anfang)
  if(sum(time_frame == zeitpunkt_neu, na.rm = T) == 0){
    while(sum(time_frame == zeitpunkt_neu, na.rm = T) == 0){
      zeitpunkt_neu <- zeitpunkt_neu +1
      zeitpunkt_neu <- as.Date(zeitpunkt_neu)
    }
  }
  if(zeitpunkt_anfang < zeitpunkt_anfang_daten){
    while(zeitpunkt_anfang < zeitpunkt_anfang_daten){
      zeitpunkt_anfang <- zeitpunkt_anfang +1
      zeitpunkt_anfang <- as.Date(zeitpunkt_anfang)
    }
  }
  if(sum(time_frame == zeitpunkt_anfang, na.rm = T) == 0){
    while(sum(time_frame == zeitpunkt_anfang, na.rm = T) == 0){
      zeitpunkt_anfang <- zeitpunkt_anfang +1
      zeitpunkt_anfang <- as.Date(zeitpunkt_anfang)
    }
  }
  start_date <- rev(which(time_frame %in% zeitpunkt_anfang))[1]
  end_date <- rev(which(time_frame %in% zeitpunkt_neu))[1]
  end_date_data <- rev(which(time_frame %in% zeitpunkt_ende))[1]


  if(zeitpunkt_ende < zeitpunkt_neu){
    frame <- frame[start_date:end_date_data ,]
  }
  else{
    frame <- frame[start_date:end_date ,]
  }
}


specific_data <- apply(frame[,2:3], 2, sum)
ratio <- specific_data[2]/specific_data[1]

##build the data respective binomial distribution

data_distribution <- c(rep(0, specific_data[1]), rep(1, specific_data[2]))

##### Bayes estimation -- no MH

if(Bayes_Estimate){
  alpha_bayes <- specific_data[2] + alpha_0
  beta_bayes <- specific_data[1] + beta_0
  bayes_estimate <- alpha_bayes/(alpha_bayes + beta_bayes)
}

####MCMC
theta_vec <- NULL

if(MH_algorithm){

  if(beta){
      theta_0 <- rbeta(n = 1, shape1 = alpha_0, shape2 = beta_0)
    }
    if(uniform){
      theta_0 <- runif(0,0.5,n = 1)
    }
    theta_vec <- rep(0, n.sims+1) #create posterior vector
    theta_vec[1] <- theta_0 #initialize posterior vector

    for (t in 1:n.sims){

      if(beta){
        theta_star <- rbeta(n = 1, shape1 = alpha_0, shape2 = beta_0)
      }
      if(uniform){
        theta_star <- runif(0,0.5,n = 1)
      }

      #proposal from prior
      theta_t <- theta_vec[t] #previous accepted value for posterior

      #Create MH-Ratio using log-likelihoods and taking the inverse of each likelihood
      #to penalize higher negative values as outcomes with a smaller likelihood

      MH_Ratio <- sum(dbinom(data_distribution, size = 1, log = T, prob = theta_t))/
        sum(dbinom(data_distribution, size = 1, log = T, prob = theta_star))

      #sample with MH ratio

      mh_prob <- c(min(1,MH_Ratio), 1-min(1,MH_Ratio))

      theta_vec[t+1] <- sample(x = c(theta_star, theta_t),size = 1,prob = mh_prob)
    }
    mh_estimate <- mean(theta_vec)
    quantiles <- quantile(theta_vec, probs = c(0.1,0.9))
    #par(mfrow = c(2,2))

    h <- hist(theta_vec[(0.3*n.sims):n.sims], plot = F, breaks = 10)
    cols <- c(0, 4, 0)
    k<- cols[findInterval(h$mids, quantile(theta_vec, probs = c(0.1,0.9)),
                          rightmost.closed=T, all.inside=F)+1]
    plot(h, col = k, main = paste("Histogram of posterior vector", Bundesland_Klasse),
         xlab = "Values of Theta Vector", xlim = c(0,margins))
  }

  return(list(quantiles = quantiles, estimate = mean(theta_vec), ratio = ratio))
}
```

```
estimation(Bundeslaender = T, Bundesland_Klasse = "Hamburg",
           Alter_Klasse = "A60-A79", alpha_0 = 4, beta_0 = 40,
           Zeitpunkt = "2020-05-06", Geschlechts_Klasse = "M",
           time.stamp = T, n.sims = 1000, Zeitpunkt_Anfang = "2021-01-21",
           beta = F , uniform = T, margins = 0.5)
```

Bibliography

(2022a). Epidemiologisches bulletin. Accessed: 2022-08-01.

(2022b). Übersicht covid -19 deutschland. Accessed: 2022-08-01.

Cairns, A., D. Blake, A. R. Kessler, and M. Kessler (May 2020). *The impact of COVID-19 on future higher-age mortality* (rev2: 10 July 2020 ed.), Volume PI-20, 07 of *Discussion paper / Pensions Institute*. London, United Kingdom: The Pensions Institute Cass Business School City University of London.

de JONG, P. and L. TICKLE (2006). Extending lee–carter mortality forecasting. *Mathematical Population Studies 13*(1), 1–18.

Delignette-Muller, M. L. and C. Dutang (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software 64*(4), 1–34.

Dr. Ing. Gerd Ehret (2017/11). Parameterschaetzung und messunsicherheit mit der bayes-statistik.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap.* Number 57 in Monographs on Statistics and Applied Probability. Boca Raton, Florida, USA: Chapman & Hall/CRC.

Gelman, A. (2014). *Bayesian data analysis* (Third edition ed.). Texts in statistical science. Boca Raton: CRC Press.

Guzmán-Rincón, L. M., E. M. Hill, L. Dyson, M. J. Tildesley, and M. J. Keeling (2022). Bayesian estimation of real-time epidemic growth rates using gaussian processes: local dynamics of sars-cov-2 in england. *medRxiv*.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*(1), 97–109.

Jackie S.T. Wong, Jonathan J. Forster, and Peter W.F. Smith (2018). Bayesian mortality forecasting with overdispersion. *Insurance: Mathematics and Economics 83*, 206–221.

Karim Barigou, Pierre-Olivier Goffard, Stéphane Loisel, and Yahia Salhi (2022). Bayesian model averaging for mortality forecasting using leave-future-out validation. *International Journal of Forecasting*.

Marshall, R. J. (1991). Mapping disease and mortality rates using empirical bayes estimators. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 40*(2), 283–294.

Matt Keeling, P. R. (2008). *Modeling Infectious Diseases in Humans and Animals.* Princeton University Press.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*(6), 1087–1092.

Robert-Koch-Institut, R. (2021). Bestätigte covid.19 fälle je 100.000 einwohnerinnen der vergangen sieben tage.

Ronald Lee and Lawrence R. Carter (1992). Modeling and forecasting u. s. mortality. *Journal of the American Statistical Association 87*, 659–671.

Schmertmann, C. and M. R. Gonzaga (2018, February). Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. SocArXiv syzwx, Center for Open Science.

Affidavit

„I hereby declare that the paper “ Estimation of Covid-19 Death Probabilities in German Regions and Evaluating future Results using Bayesian Statistics” is my own work and that I have not called upon the help of a third party. In addition, I affirm that neither I nor anybody else has submitted this paper or parts of it to obtain credits elsewhere before. I have clearly marked and acknowledged all quotations or references that have been taken from the works of other. All secondary literature and other sources are marked and listed in the bibliography. The same applies to all charts, diagrams and illustrations as well as to all Internet sources. Moreover, I consent to my paper being electronically stores and sent anonymously in order to be checked for plagiarism. I am aware that the paper cannot be evaluated and may be graded “failed” (“nicht ausreichend”) if the declaration is not made.”

Mannheim, den 03.08.2022

Shreyas Prakash Gupta