Estimating Parameters of the SIR Epidemiology Model using Simulated Survey Data and Multinomial Maximum Likelihood

A Thesis
Presented in Partial Fulfillment of the Requirements for the
Degree of Master of Science
with a
Major in Statistical Science
in the
College of Graduate Studies
University of Idaho
by
Dylan Hull-Nye

Major Professor: Brian Dennis, Ph.D.
Committee Members: Chris Williams, Ph.D.; Chris Remien, Ph.D.
Department Administrator: Chris Williams, Ph. D.

August 2020

Authorization to Submit Thesis

This thesis of Dylan Hull-Nye, submitted for the degree of Master of Science with a major in Statistical Science and titled "Estimating Parameters of the SIR Epidemiology Model using Simulated Survey Data and Multinomial Maximum Likelihood," has been reviewed in final form. Permission, as indicated by the signatures and dates given below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: ______________________________Date____________
Brian Dennis, Ph.D.

Committee
Members: ______________________________Date____________
Chris Williams, Ph.D.

______________________________Date____________
Chris Remien, Ph.D.

Department
Administrator: ______________________________Date____________
Chris Williams, Ph.D.

Abstract

Brian Dennis and William Kemp (et al.) in "Stochastic Model of Insect Phenology Estimation and Testing" use a Multinomial Maximum likelihood approach to estimate model parameters via ratios between variables. The model in this paper attempts to estimate deterministic epidemiological population parameters for the Kermack-McKendrick SIR model using simulated sample survey data. Ratios between populations in the SIR model are used as probabilities in the Multinomial distribution and resulting estimates of population parameters are analyzed via Bootstrap confidence intervals, visual analysis, and mean squared error (MSE) estimates. The results show that certain surveys schemes perform better than others based on survey sample size, number of surveys, and survey spacing. All simulations generally produce unbiased estimates with some producing smaller variances than others. Applications include the estimating of infectious disease dynamics using small survey sampling in rural or small town universities and populations.

Acknowledgements

I would like to thank my advisor, Brian Dennis, for introducing me to the investigation from the beginning and inspiring me to explore the ideas that lead to this thesis. I also want to thank him for his courses which taught me the statistical rigor and coding skills needed to carry out this project. I would also like to thank the Committee members, Chris Remien and Chris Williams, for their support and patience.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Much has been achieved on the topic of parameter estimation in epidemiology, especially in light of pandemics such as influenza, the 2019 measles outbreak, and the novel COVID-19. Many researchers use least squares curve fitting and other maximum likelihood approaches based on hospital incidence or positive test case data. One example is the stochastic model for the 2009 H1N1 pandemic in rural Washington [2]. Although these approaches are well respected, incidence data from clinics or hospitals does not randomly sample the population overall. Random survey samples capture population dynamics more realistically, and in small university towns such as Pullman, WA or Moscow, ID, random survey samples are achievable. In this paper, I describe how I used random survey samples of a small city or town to estimate model parameters.

The Kermack-McKendrick Susceptible-Infected-Recovered (SIR) deterministic model was the focus of all my investigations. It was used by researchers in modeling the 2009 H1N1 pandemic in Pullman, WA [1]. I chose this model because it is the foundation of most other epidemiology models. My investigations, however, can certainly be extended to other models with more complicated terms. One natural extension would be SIR models capturing the effect of vaccination or quarantine[5]. All variables in the model are continuous and follow a system of three ordinary differential equations. The model was not solved using any analytical methods but only with numerical packages in the software R. I wrote all of the code in R myself but found some basic ideas in a paper that modeled measles in Niger [3]. Though I did not borrow their code literally, I gained ideas from their use of certain packages in R for optimization and numerical solutions of ODEs.

For the random sampling of the population, I investigated how well I can estimate parameters from data generated by hypothetical surveys. Though I did not carry out the surveys myself, if researchers were to do so, they would only need to ask two very simple questions from a respondent about a particular infectious disease: (1) Do you currently have the disease? and (2) Have you had the disease recently and recovered? These questions could be asked as part of a larger health survey by the university or asked in a short survey given in various locations on campus. Surveys bring their own challenges, such as whether people actually know they have contracted a particular disease. This is not always obvious in the case of influenza, but more likely to be known in a disease such as measles or rare diseases where people have unusual symptoms and tend to visit a hospital or clinic for testing. Random population clinical testing, if it were efficient, could also produce suitable data. There would likely be other typical survey problems such as nonresponse as well. In my investigations I assumed reasonably accurate and reliable survey data. In Brian Dennis' paper estimating the proportion of insects at various stages of development, he takes random samples from a multinomial distribution for the probabilities and then estimates parameters in his model using maximum likelihood [4]. I applied this approach to the numerically solved deterministic SIR model and built a multinomial log-likelihood function, based on the one from Samaniego [6]. Overall my goal was to examine how effective multinomial-generated surveys can estimate population parameters of the SIR model using numerical maximum likelihood optimization.

CHAPTER 2

Model

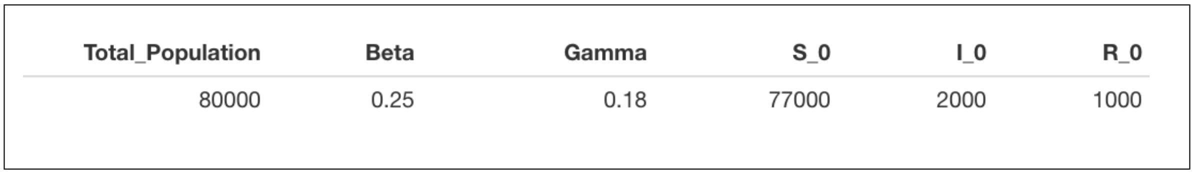
2.1 Susceptible-Infected-Recovered (SIR)

The Kermack-McKendrick SIR model is given by the following three ordinary differential equations:

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta SI}{N}\\ \frac{dI}{dt} &= \frac{\beta SI}{N} - \gamma I\\ \frac{dR}{dt} &= \gamma I\end{aligned}$$

The three populations are Susceptible in time $S(t)$, Infected in time $I(t)$, and Recovered in time $R(t)$. I chose not to scale the equations so all numbers represent actual people at time t in the population. The parameter N is the population size, set constant for all simulations, and time is measured in days. The parameter β is scaled by N so that it is in the range of 10^-1 rather than 10^-5. The parameter β measures the rate of disease transmission and the parameter γ measures the rate of recovery from the disease. If one is not familiar with the terms in this model, notice that the term involving β is removed from the Susceptible population $S(t)$, representing the proportion of those who interact with those infected and therefore join the infected population. That same term is added as recruitment to the infected population $I(t)$, and finally the term with γ leaves the infected population (for those healed) and enters the recovered population.

All simulations in my investigation were based on one initial population with specific parameter values for β and γ as well as the three initial conditions $S(0), I(0), R(0)$. The specific values are given in Table 2.1, and I chose them as general parameters representing various diseases found in the literature. In the original population, the infected population reaches a peak in about 30 days and the susceptible population levels off over time at about half the original size. Though each disease will have its own underlying disease parameters, for the sake of simulation and efficiency, I chose to focus on one in this paper. The total population size is representative of a small city or large town, such as the combined Moscow-Pullman area, using a value of $N = 80,000$.



Total_Population	Beta	Gamma	S_0	I_0	R_0
80000	0.25	0.18	77000	2000	1000

Table 2.1: The initial conditions and parameters were set to reflect a realistic disease start. The multinomial maximum likelihood function estimated back these population parameters and initial conditions.

2.2 Multinomial Surveys and Likelihood Function

Figure 2.1 shows plotted trajectories for the three populations early in time (first 40 days), with vertical lines at different spacing. The parameters of the multinomial distribution are sample sizes and probabilities of success. The probabilities in the context of an SIR model were calculated as ratios in the number of each population over the total sum of all three populations at each time point. For example, take time $t = 8$ days as a time point, as shown by the first vertical line in plot A of Figure 2.1. The probabilities associated with this time point are the ratios: $p_{s_1} = \frac{S(8)}{S(8)+I(8)+R(8)}$, $p_{i_1} = \frac{I(8)}{S(8)+I(8)+R(8)}$, and $p_{r_1} = \frac{R(8)}{S(8)+I(8)+R(8)}$. The corresponding ratios from the data sample were calculated with sample size n and susceptible $(\frac{s_1}{n})$, infected $(\frac{i_1}{n})$, and recovered $(\frac{r_1}{n})$ at $t = 8$. The subscript refers to the number sample, with the first sample shown here.

Sampling schemes were chosen so as to cover various different sampling scenarios. Each scenario or scheme can be identified by three parameters: n, the sample size, q, the number of survey samples, and x, the number of days between survey samples (i.e. spacing of samples). Although in my investigation I chose to keep the sample size consistent at every time step, my code actually allows for differing sample sizes at each time point. For each sampling scheme I calculated the population proportions at each spacing based on the numerically solved ODE trajectories. The population proportions became the underlying multinomial probabilities for all parameter estimations. It was assumed that the underlying population disease dynamics follow these proportions. Though I assume knowledge of population parameters for the simulations, in reality one doesn't actually know them and would use simulations to produce estimates. The purpose of the simulations was to see how well the multinomial maximum likelihood with smaller sample sizes estimates back the population proportions.

The q-by-3 data matrix was generated from q samples of size n: each row is a sample at a specific time point and each column is the resulting number of Susceptible, Infected, and Recovered randomly drawn at that time. The simulated data matrix depended both on the overall chosen sampling scheme and on randomness derived from the R function ***rmultinom***. A total of 1000 random draws were made with the same underlying population probabilities, sample size, and sampling scheme. The multinomial log likelihood function used each data matrix and estimated proportions that generated the best fit SIR equations to the data. For each random sample it tested and searched for parameters and initial conditions leading to maximum likelihood probabilities. Output data took the form of 1000 maximum likelihood estimates for each SIR parameter or initial condition. The likelihood function is as follows:

$$L(\mathbf{p}) = \prod_{k=1}^{q} \frac{n_k!}{s_k! i_k! r_k!} (p_{s_k})^{s_k} (p_{i_k})^{i_k} (p_{r_k})^{r_k}$$

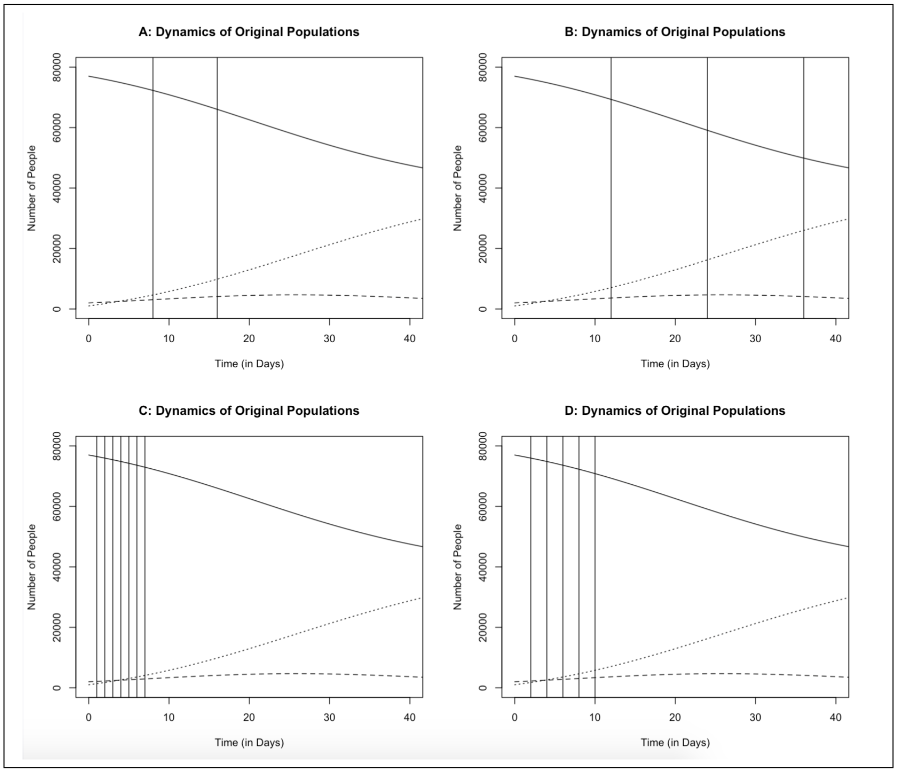


Figure 2.1: The four chosen examples of the S-I-R plotted trajectories from the ODE model. Vertical lines represent four sampling schemes. In all cases, the S-I-R model was the model described in Section 2.1, with A being q = 2 sample surveys at x = 8 days apart, B being q = 3 sample surveys at x = 12 days apart, C being q = 7 sample surveys at x = 1 day apart, and D being q = 5 sample surveys at x = 2 days apart. All four give an example of the sampling scheme for any sample size. The solid line is the Susceptible, the dotted line is the Recovered, and the dashed lined is the Infected populations.

2.3 Computational Methods in R

All resulting plots will show five estimated quantities: β, γ, $S(0)$, $I(0)$, $R(0)$. However, in my actual code, in order to increase computational efficiency, I forced an assumption, which may or may not be true, but can be changed if further simulations were carried out. I set the initial condition $R(0) = 0.5I(0)$, which as a consequence also fixed $S(0) = N - R(0) - I(0)$ since population size N was constant. The program essentially estimated only three parameters: β, γ, and $I(0)$, with $S(0)$ and $R(0)$ derived from the estimates of $I(0)$.

I used the ***deSolve*** package to numerically solve the SIR equations, and I implemented the ***mle2*** package. The main optimization program in ***mle2*** is ***optim***. My choice of algorithm was Nelder-Mead, which was also used in the insect phenology paper [4]. As an unbounded search algorithm, Nelder-Mead only required starting values for each simulation, which I set to lower but roughly nearby values: $\beta = 0.1$, $\gamma = 0.1$, and $I(0) = 100$. The only problem I found with it was computational expense, since each set of simulations for one sampling scheme took on average 20 minutes. In turn, I ran all simulations on cloud computing in parallel through multiple sessions of *R Studio Cloud*.

Sampling schemes were set to try to cover most situations throughout the epidemic. Values for survey size and spacing were set to: q = 2 with x = 3, 8, 15 and q = 3 with x = 8, 12, 18 then q = 5 with x = 2, 8 and q = 7 with x = 1, 3, 8. Two sample sizes were set for each of the 11 schemes: n = 500 and n = 3000. Given the initial results of all 22 simulations, I decided to run further simulations: q = 2, 3, 4, 5, 6, 7, 8 while x = 1 for sample sizes n = 500 and n = 3000. The reasons for doing further simulations for specific numbers will be explained in the Results section.

CHAPTER 3

Results

3.1 Histograms and QQ Plots

One hypothesis before running the simulations was that the best scheme would be the one with several samples spaced widely, so as to cover the trajectories. The best should visually follow a symmetric, Normal-appearing distribution centered around the population estimate, which is known and pre-set as in Table 2.1. Numerically the best should also have a smaller variance, also shown by thinner histograms. In the simulations I found that all estimates were centered adequately in their mean around each population parameter. Both sample sizes $n = 500$ and $n = 3000$ did comparably well in this respect, appearing to be unbiased estimators. However, the histograms and the data tables show that some performed more precisely than others and surprisingly different than my original hypothesis.

The histograms appearing the least Normal and with widest variance in the $n = 500$ sample size were $\{q = 2, x = 3\}$, $\{ q = 2, x = 8\}$ and $\{q = 2, x = 15\}$. These had the minimum number of samples but spaced within roughly a week and two weeks apart. At $q = 3$ samples $x = 8$ days apart, the estimates for β and γ improved but the estimates for the initial conditions were poor. The split between good estimates of the model parameters and mediocre estimates for the initial conditions continued for all other simulations in sample size 500, except at $\{q = 7, x = 1\}$ and $\{q = 7, x = 3\}$, which were slightly better than all others. The worst case can be seen in Figure 3.4. A full set of histograms and all other plots and tables are available in the Appendix link for Supplementary Material.

For sample size $n = 3000$, which is six times the sample size of the first set of simulations, all schemes did generally well for both parameters and initial conditions. As expected an increase in sample size produced better estimates. The scheme $\{q = 2, x = 3\}$ again performed the poorest just as in the $n = 500$ case, but the rest of the cases perform adequately. The most exceptional were $\{q = 7, x = 1\}$ and $\{q = 7, x = 3\}$, but even better than the

$n = 500$ case. The scheme $\{q = 5, x = 2\}$ for some reason was the third most successful. In other words if a researcher wanted to estimate back these population parameters it would be best to give a survey every day for a week. Estimates early on in short intervals were all better than over time and spread out. Even ones that cover time over the entire course of the disease outbreak were not as good. For histograms of the best results refer to Figure 3.1, Figure 3.2, and Figure 3.3.

Given the peculiar nature of the better estimators, I decided to run a few more simulations with the one-day-apart scenarios. I ran further simulations for both sample sizes again $n = 500$ and $n = 3000$ but for $q = 2, 3, 4, 5, 6, 7, 8$ while holding $x = 1$ for all of them. It turned out that all of them except $\{q = 2, x = 1\}$ were decent, and as the number of surveys increased, the accuracy and variance improved as well. The increase in sample size also improved estimation again. The last one $\{q = 8, x = 1\}$ was the best.

As for assessing Normality, I examined QQ Plots for all parameters for all 1000 runs for each scheme, producing 110 QQ Plots in total, plus QQ Plots for the further simulations. The better schemes mentioned had the best QQ Plots, appearing most Normal, and the poorer ones had heavy tails with less conformity to Normality. One of the QQ plots is shown for the best scheme overall in Figure 3.5 and the poorest scheme overall in Figure 3.6.

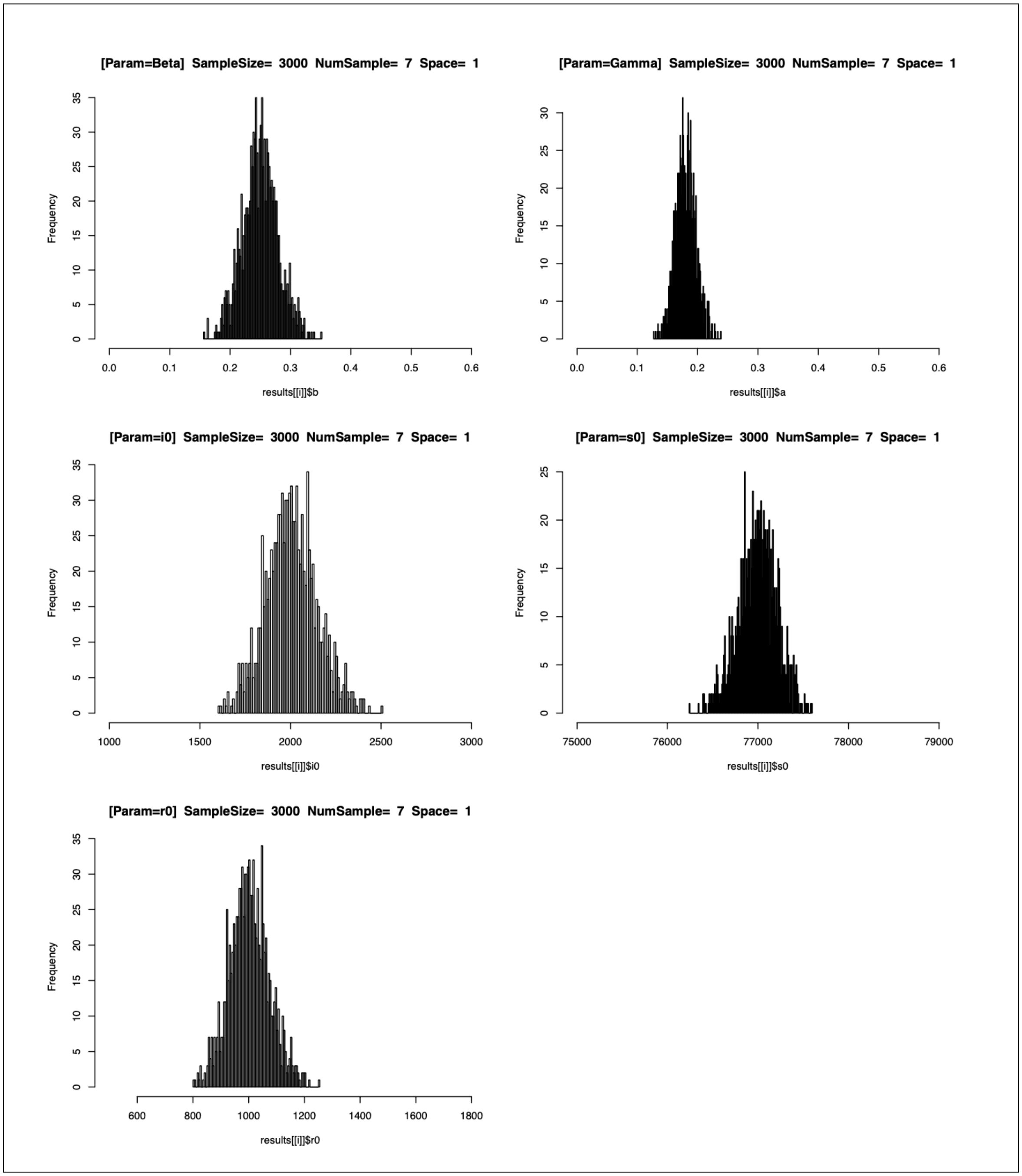


Figure 3.1: Histograms of the five estimated parameters $\beta, \gamma, S(0), I(0), R(0)$ resulting from 1000 simulations of Multinomial Maximum Likelihood with sample size n = 3000 people. Each histogram can be identified by its parameter in square brackets. This was the best estimator over all simulations, following the setup q = 7 surveys, x = 1 day apart.

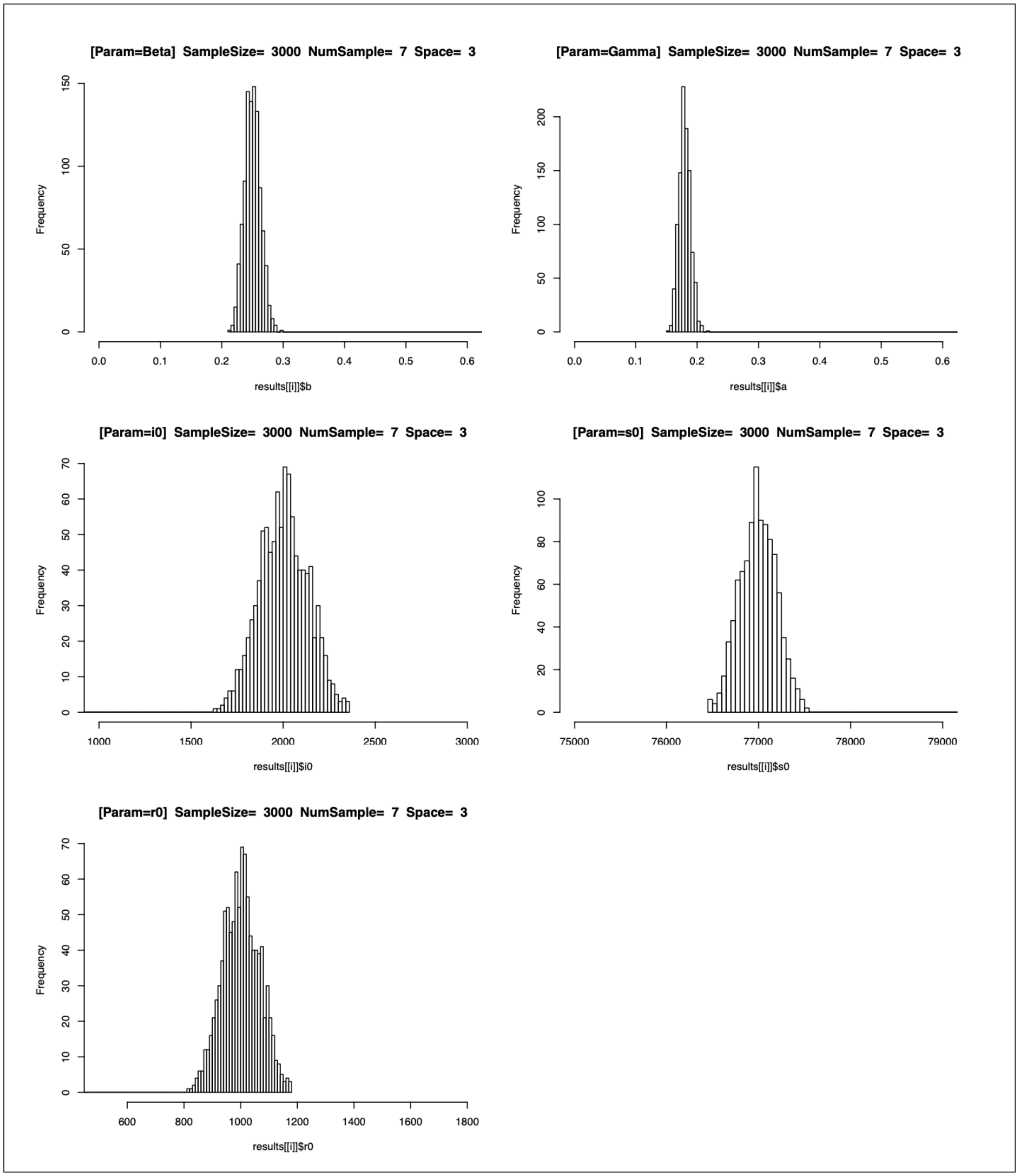


Figure 3.2: Histograms of the five estimated parameters $\beta, \gamma, S(0), I(0), R(0)$ resulting from 1000 simulations of Multinomial Maximum Likelihood with sample size n = 3000 people. Each histogram can be identified by its parameter in square brackets. This was the best estimator over all simulations, following the setup q = 7 surveys, x = 3 days apart.

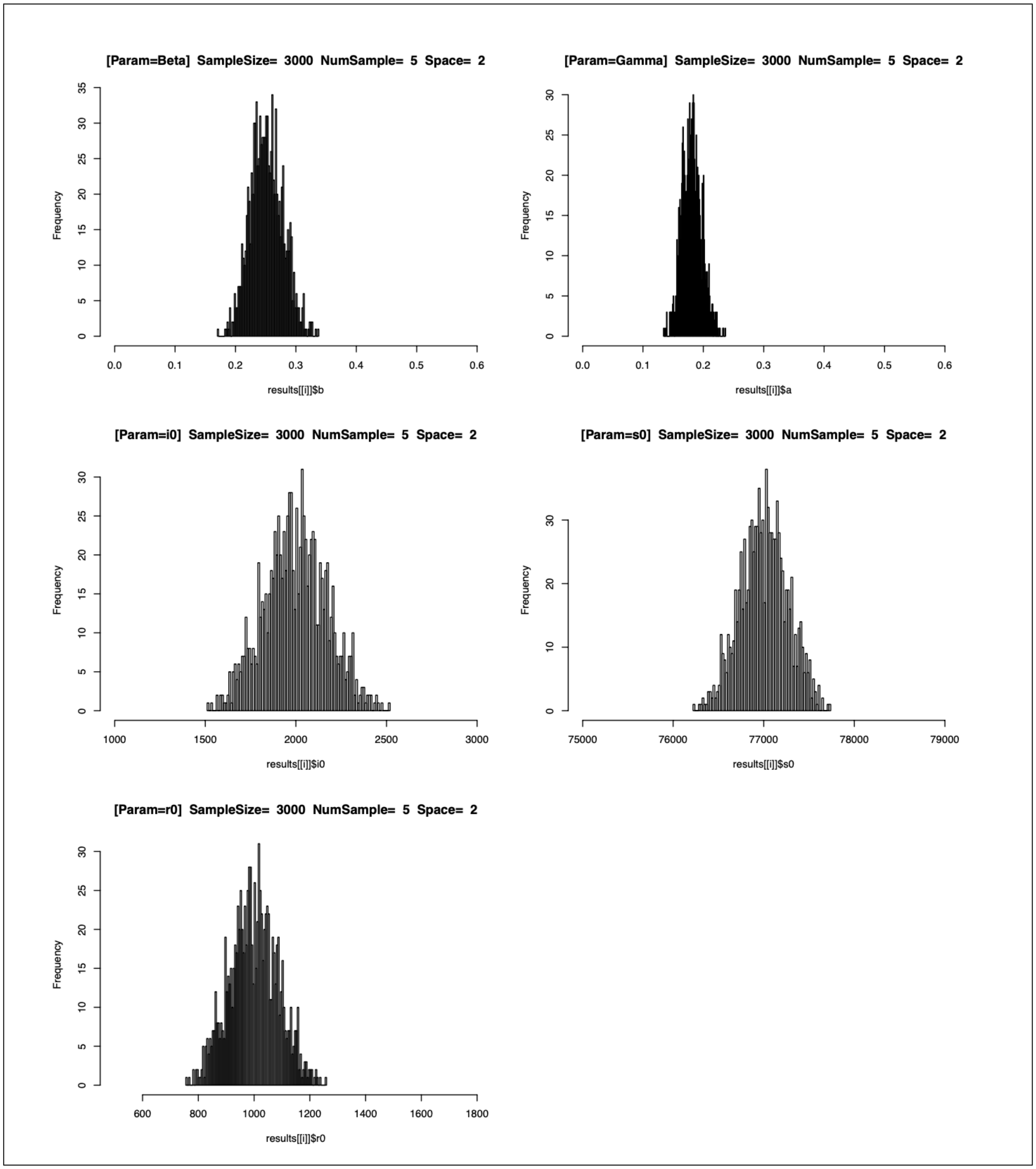


Figure 3.3: Histograms of the five estimated parameters $\beta, \gamma, S(0), I(0), R(0)$ resulting from 1000 simulations of Multinomial Maximum Likelihood with sample size n = 3000 people. Each histogram can be identified by its parameter in square brackets. This was the best estimator over all simulations, following the setup q = 5 surveys, x = 2 days apart.

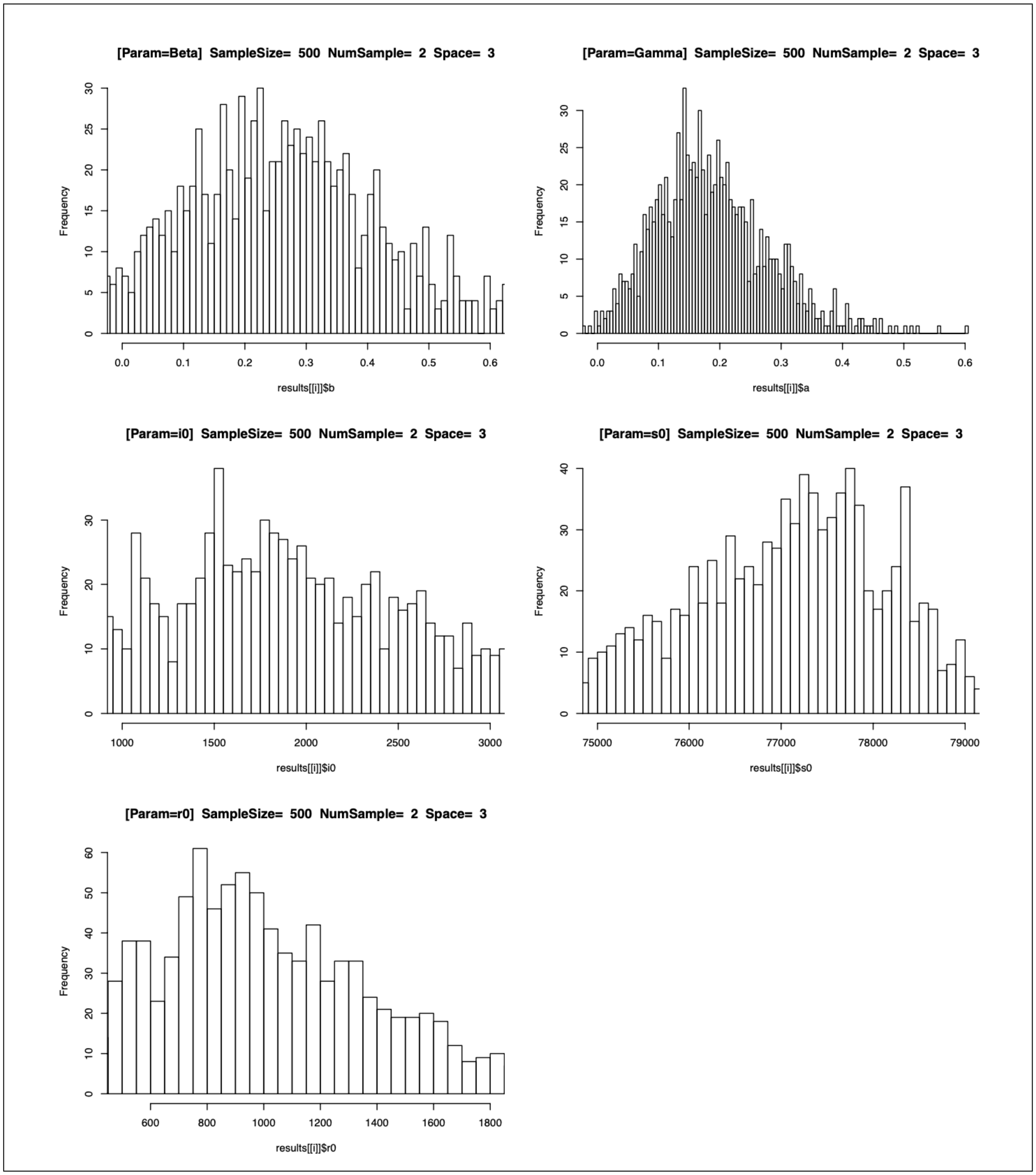


Figure 3.4: Histograms of the five estimated parameters $\beta, \gamma, S(0), I(0), R(0)$ resulting from 1000 simulations of Multinomial Maximum Likelihood with sample size n = 500 people. Each histogram can be identified by its parameter in square brackets. This was the best estimator over all simulations, following the setup q = 2 surveys, x = 3 days apart.

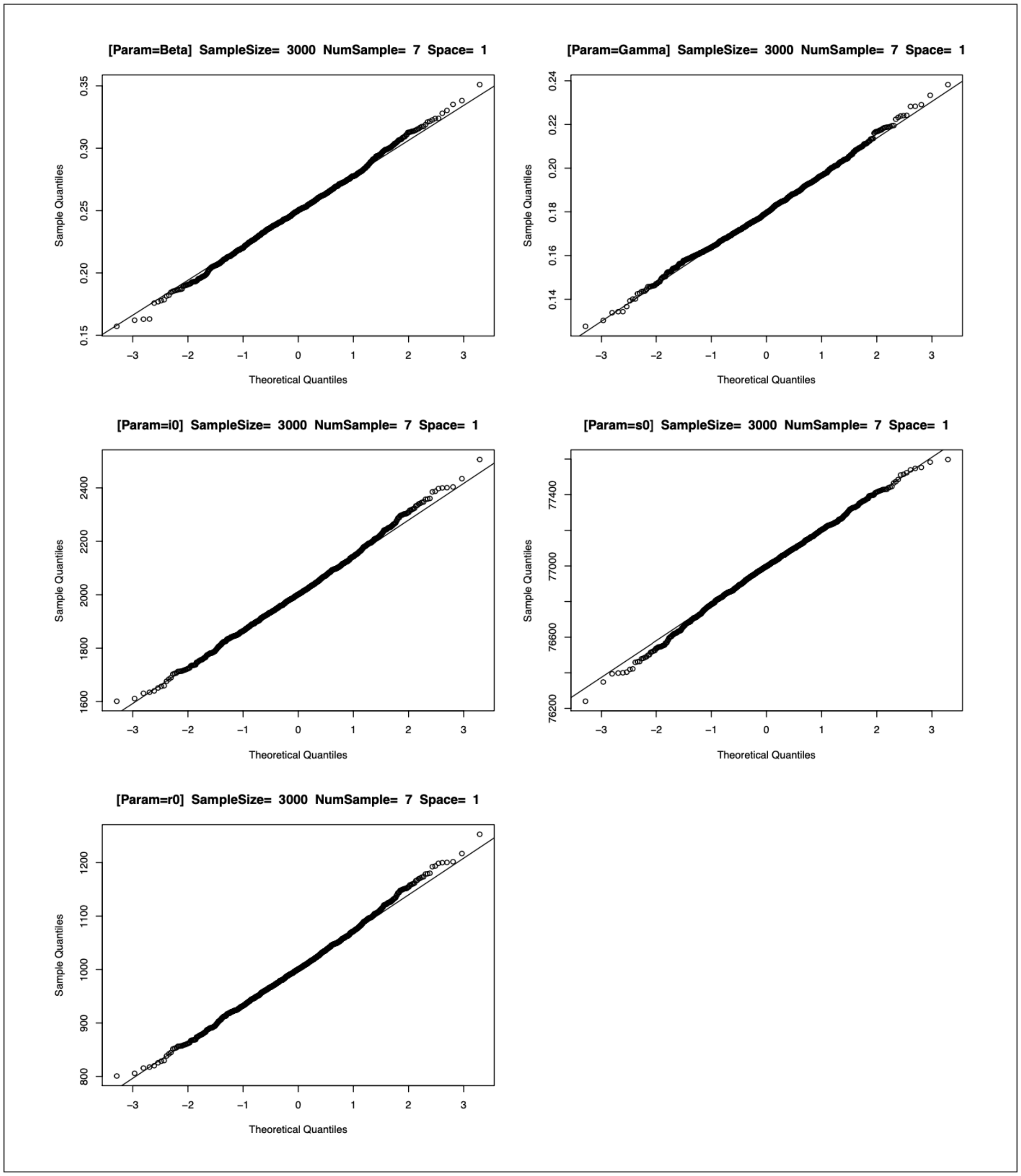


Figure 3.5: QQ Plots of the five estimated parameters $\beta, \gamma, S(0), I(0), R(0)$ resulting from 1000 simulations of Multinomial Maximum Likelihood with sample size n = 3000 people. Each QQ Plot can be identified by its parameter in square brackets. Shown is the best estimator over all simulations, following q = 7 surveys, x = 1 days apart. Normality appears more than adequate.

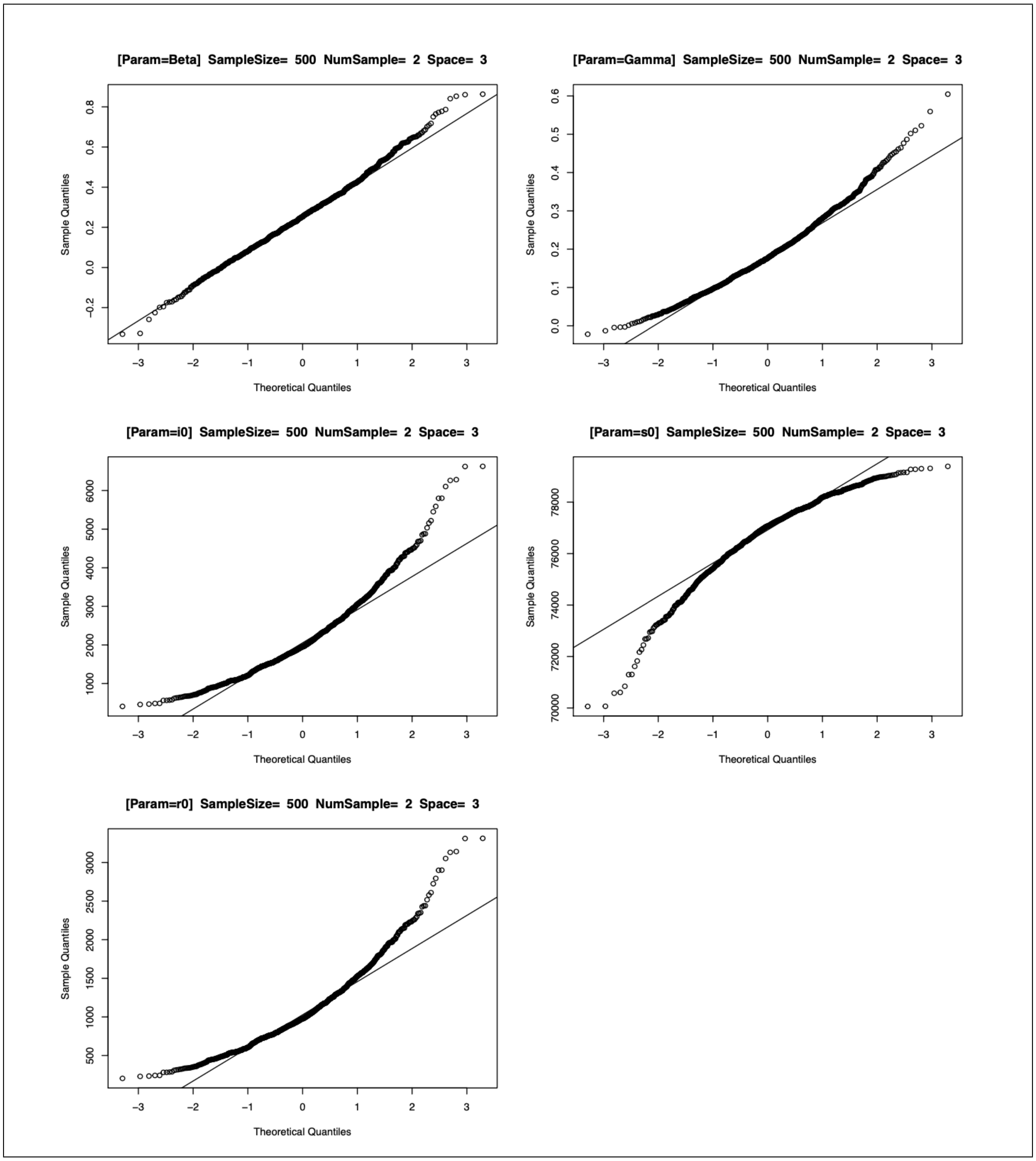


Figure 3.6: QQ Plots of the five estimated parameters $\beta, \gamma, S(0), I(0), R(0)$ resulting from 1000 simulations of Multinomial Maximum Likelihood with sample size n = 500 people. Each QQ Plot can be identified by its parameter in square brackets. Shown is the best estimator over all simulations, following q = 2 surveys, x = 3 days apart. Normality appears generally not acceptable, especially with deviations in the tails.

3.2 Selected Confidence Intervals Plots

In Figures 3.7- 3.10 plots of confidence intervals are shown for the parameter β over all simulations at both sample sizes. One can see that apart from the first one or two schemes, the schemes within a sample size were roughly the same width. The widths did not change much in the original simulations but did change slightly in the further simulations. Almost all schemes produced good numerical estimates. The trend shown was similar for other parameters and it lead me to perform a more robust analysis using Mean Squared Errors.

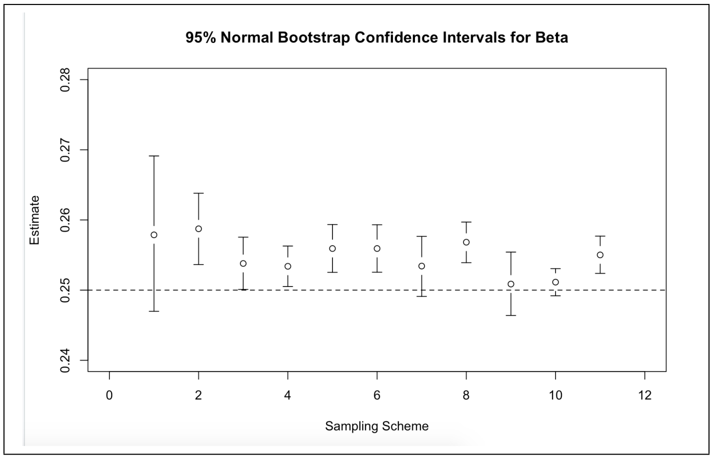


Figure 3.7: Shown are bootstrap Confidence Intervals for β estimates for all initial simulations of sample size n = 500. Sampling Schemes are in the same order as Table 3.5. The true population parameter is $\beta = 0.25$, seen as a dashed line.

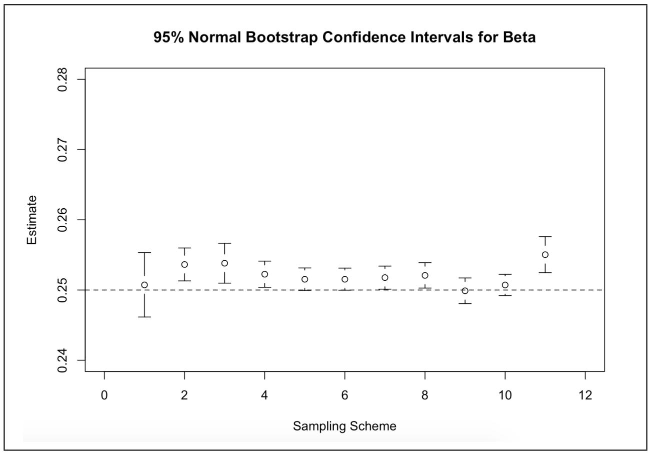


Figure 3.8: Shown are bootstrap Confidence Intervals for β estimates for all initial simulations of sample size n = 3000. Sampling Schemes are in the same order as Table 3.6. The true population parameter is $\beta = 0.25$, seen as a dashed line.

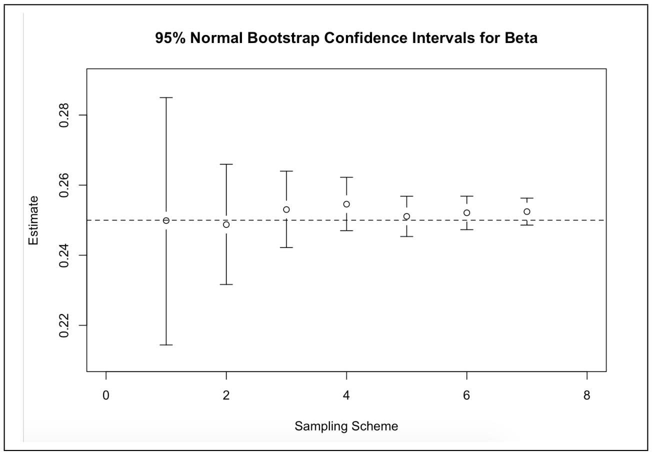


Figure 3.9: Shown are bootstrap Confidence Intervals for β estimates for all further simulations of sample size n = 500. Sampling Schemes are in the same order as Table 3.7. The true population parameter is $\beta = 0.25$, seen as a dashed line.

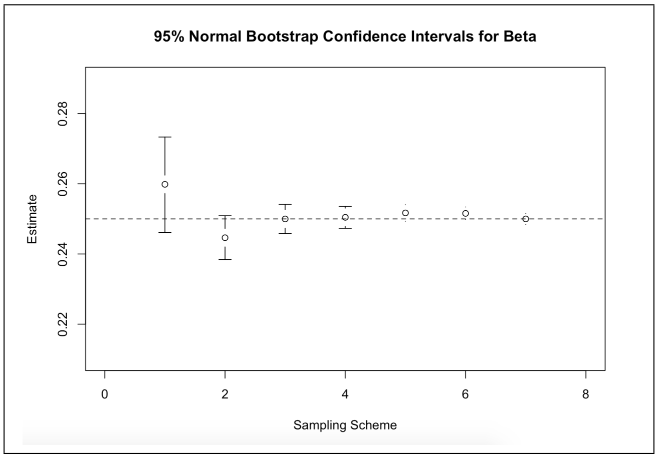
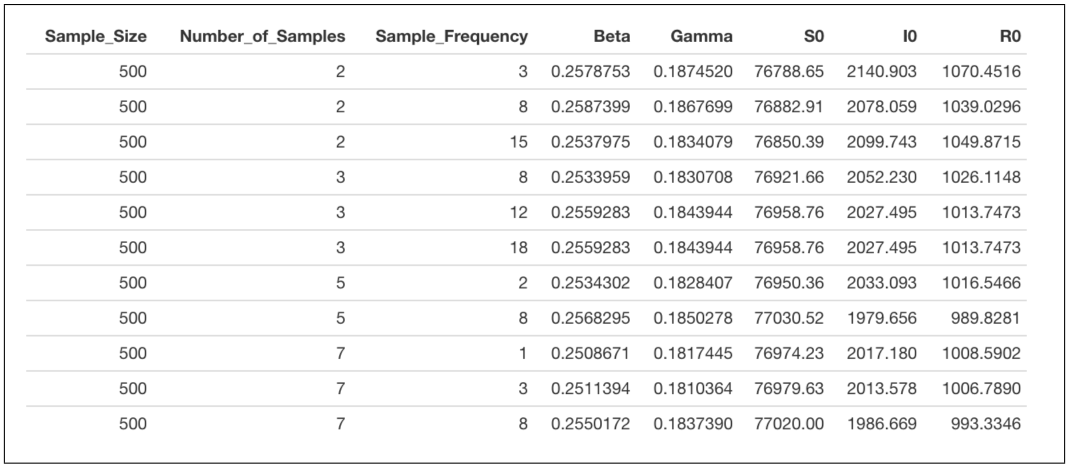


Figure 3.10: Shown are bootstrap Confidence Intervals for β estimates for all further simulations of sample size n = 3000. Sampling Schemes are in the same order as Table 3.8. The true population parameter is $\beta = 0.25$, seen as a dashed line.

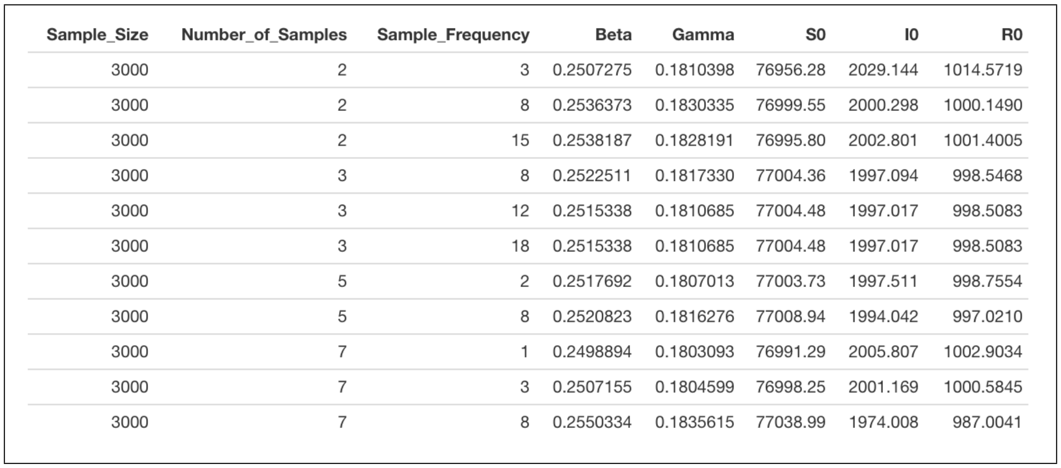
3.3 Mean Maximum-Likelihood-Estimate (MLE) Tables

The following tables give the mean of the 1000 maximum likelihood estimates for each parameter and initial condition. The numerical results confirmed the visual analysis since they were mostly centered at the population parameters and initial conditions.



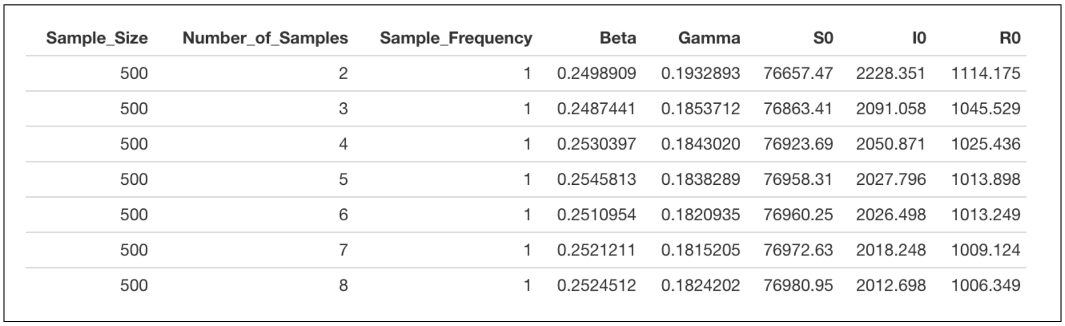
Sample_Size	Number_of_Samples	Sample_Frequency	Beta	Gamma	S0	I0	R0
500	2	3	0.2578753	0.1874520	76788.65	2140.903	1070.4516
500	2	8	0.2587399	0.1867699	76882.91	2078.059	1039.0296
500	2	15	0.2537975	0.1834079	76850.39	2099.743	1049.8715
500	3	8	0.2533959	0.1830708	76921.66	2052.230	1026.1148
500	3	12	0.2559283	0.1843944	76958.76	2027.495	1013.7473
500	3	18	0.2559283	0.1843944	76958.76	2027.495	1013.7473
500	5	2	0.2534302	0.1828407	76950.36	2033.093	1016.5466
500	5	8	0.2568295	0.1850278	77030.52	1979.656	989.8281
500	7	1	0.2508671	0.1817445	76974.23	2017.180	1008.5902
500	7	3	0.2511394	0.1810364	76979.63	2013.578	1006.7890
500	7	8	0.2550172	0.1837390	77020.00	1986.669	993.3346

Table 3.1: Shown are means of all 1000 Maximum Likelihood estimates for all initial simulations of sample size n = 500.



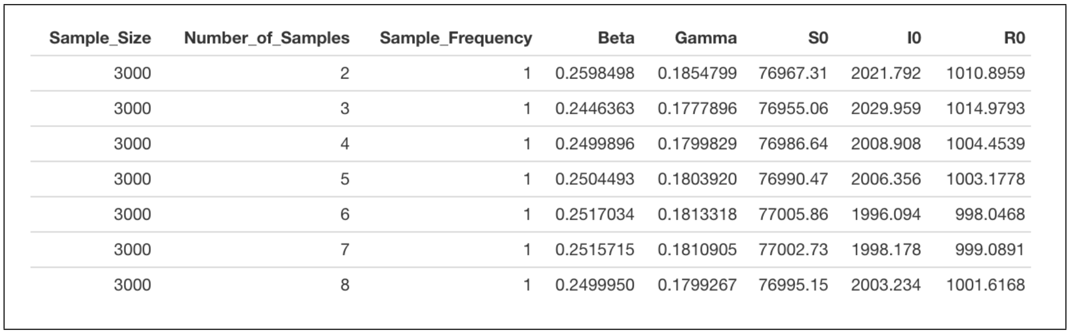
Sample_Size	Number_of_Samples	Sample_Frequency	Beta	Gamma	S0	I0	R0
3000	2	3	0.2507275	0.1810398	76956.28	2029.144	1014.5719
3000	2	8	0.2536373	0.1830335	76999.55	2000.298	1000.1490
3000	2	15	0.2538187	0.1828191	76995.80	2002.801	1001.4005
3000	3	8	0.2522511	0.1817330	77004.36	1997.094	998.5468
3000	3	12	0.2515338	0.1810685	77004.48	1997.017	998.5083
3000	3	18	0.2515338	0.1810685	77004.48	1997.017	998.5083
3000	5	2	0.2517692	0.1807013	77003.73	1997.511	998.7554
3000	5	8	0.2520823	0.1816276	77008.94	1994.042	997.0210
3000	7	1	0.2498894	0.1803093	76991.29	2005.807	1002.9034
3000	7	3	0.2507155	0.1804599	76998.25	2001.169	1000.5845
3000	7	8	0.2550334	0.1835615	77038.99	1974.008	987.0041

Table 3.2: Shown are means of all 1000 Maximum Likelihood estimates for all initial simulations of sample size n = 3000.



Sample_Size	Number_of_Samples	Sample_Frequency	Beta	Gamma	S0	I0	R0
500	2	1	0.2498909	0.1932893	76657.47	2228.351	1114.175
500	3	1	0.2487441	0.1853712	76863.41	2091.058	1045.529
500	4	1	0.2530397	0.1843020	76923.69	2050.871	1025.436
500	5	1	0.2545813	0.1838289	76958.31	2027.796	1013.898
500	6	1	0.2510954	0.1820935	76960.25	2026.498	1013.249
500	7	1	0.2521211	0.1815205	76972.63	2018.248	1009.124
500	8	1	0.2524512	0.1824202	76980.95	2012.698	1006.349

Table 3.3: Shown are means of all 1000 Maximum Likelihood estimates for all further simulations of sample size n = 500.

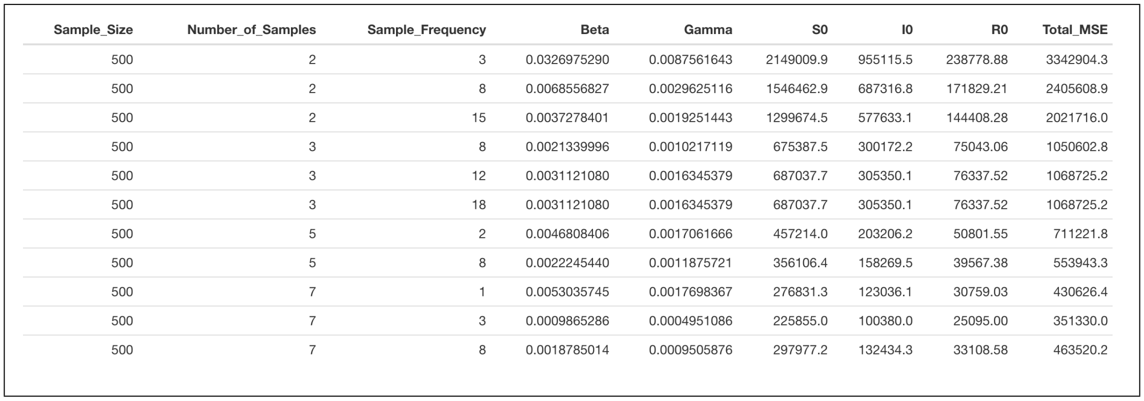


Sample_Size	Number_of_Samples	Sample_Frequency	Beta	Gamma	S0	I0	R0
3000	2	1	0.2598498	0.1854799	76967.31	2021.792	1010.8959
3000	3	1	0.2446363	0.1777896	76955.06	2029.959	1014.9793
3000	4	1	0.2499896	0.1799829	76986.64	2008.908	1004.4539
3000	5	1	0.2504493	0.1803920	76990.47	2006.356	1003.1778
3000	6	1	0.2517034	0.1813318	77005.86	1996.094	998.0468
3000	7	1	0.2515715	0.1810905	77002.73	1998.178	999.0891
3000	8	1	0.2499950	0.1799267	76995.15	2003.234	1001.6168

Table 3.4: Shown are means of all 1000 Maximum Likelihood estimates for all further simulations of sample size n = 3000.

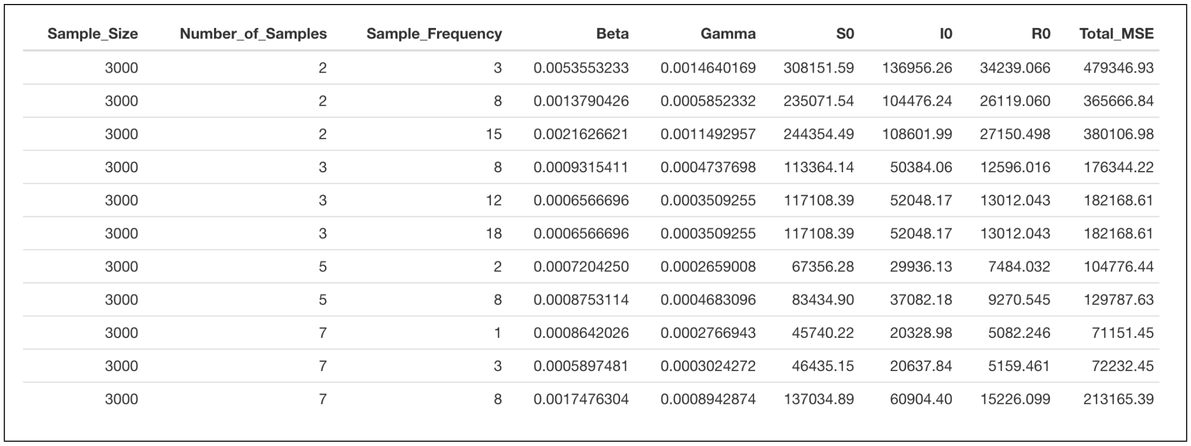
3.4 Mean-Squared-Error (MSE) Tables

I relied on the Mean Squared Error as a numerical assessment of the parameters. For an estimated parameter, $\hat{\theta}$, the MSE was calculated as $MSE(\hat{\theta}) = var(\hat{\theta}) + (mean(\hat{\theta}) - \theta)^2$. Here θ was the true parameter value from Table 2.1, and the variance and mean of $\hat{\theta}$ were calculated on the data sets of each 1000 estimates for each parameter. The results are shown in Tables 3.5 - 3.8 for the first simulations and the further simulations. The reason I used MSE was that it captured both the variability and the biasedness in all estimates. Lower MSE values I deemed better estimators. The highest MSE value of β turned out to be in the $\{n = 500, q = 2, x = 1\}$, then $\{n = 500, q = 3, x = 1\}$, then $\{n = 500, q = 2, x = 3\}$. This confirms comparable results to the visual analysis. One of the difficulties with assessing the MSE values was that there were five different parameters with MSE values. In turn, I created a column for total sum of MSE over all parameters. The lowest MSE values over all simulations were found to be (in order of lowest): $\{n = 3000, q = 7, x = 1\}$, $\{n = 3000, q = 7, x = 3\}$, then $\{n = 3000, q = 5, x = 2\}$. I also confirmed that the ranks were roughly matched up by calculating the rank of each column. The conclusions therefore stayed the same. The $\{n = 3000, q = 6, x = 1\}$ and $\{n = 3000, q = 8, x = 1\}$ simulations also performed comparably well.



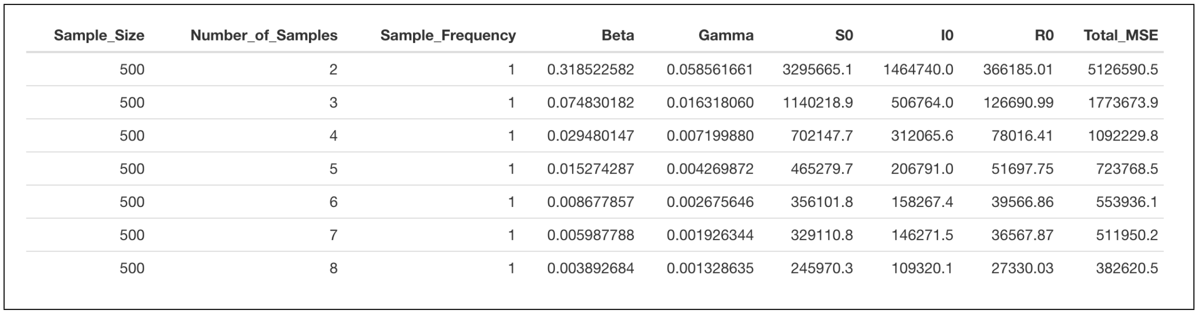
Sample_Size	Number_of_Samples	Sample_Frequency	Beta	Gamma	S0	I0	R0	Total_MSE
500	2	3	0.0326975290	0.0087561643	2149009.9	955115.5	238778.88	3342904.3
500	2	8	0.0068556827	0.0029625116	1546462.9	687316.8	171829.21	2405608.9
500	2	15	0.0037278401	0.0019251443	1299674.5	577633.1	144408.28	2021716.0
500	3	8	0.0021339996	0.0010217119	675387.5	300172.2	75043.06	1050602.8
500	3	12	0.0031121080	0.0016345379	687037.7	305350.1	76337.52	1068725.2
500	3	18	0.0031121080	0.0016345379	687037.7	305350.1	76337.52	1068725.2
500	5	2	0.0046808406	0.0017061666	457214.0	203206.2	50801.55	711221.8
500	5	8	0.0022245440	0.0011875721	356106.4	158269.5	39567.38	553943.3
500	7	1	0.0053035745	0.0017698367	276831.3	123036.1	30759.03	430626.4
500	7	3	0.0009865286	0.0004951086	225855.0	100380.0	25095.00	351330.0
500	7	8	0.0018785014	0.0009505876	297977.2	132434.3	33108.58	463520.2

Table 3.5: Shown are Mean Squared Error (MSE) values of all 1000 Maximum Likelihood estimates for all initial simulations of sample size n = 500.



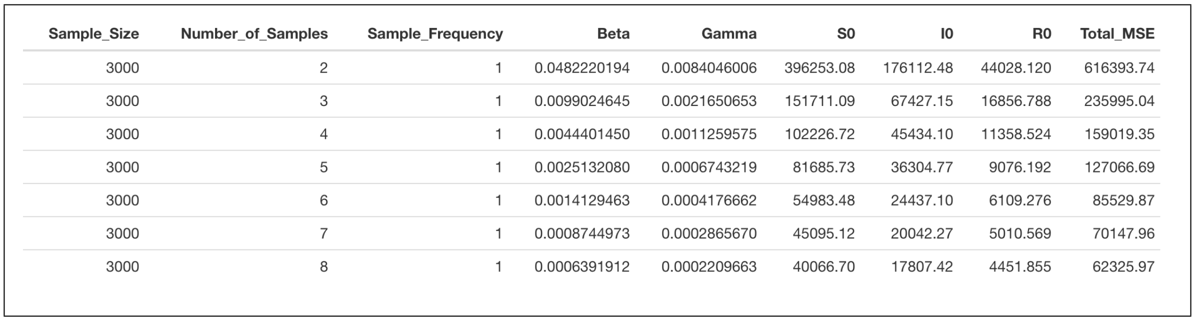
Sample_Size	Number_of_Samples	Sample_Frequency	Beta	Gamma	S0	I0	R0	Total_MSE
3000	2	3	0.0053553233	0.0014640169	308151.59	136956.26	34239.066	479346.93
3000	2	8	0.0013790426	0.0005852332	235071.54	104476.24	26119.060	365666.84
3000	2	15	0.0021626621	0.0011492957	244354.49	108601.99	27150.498	380106.98
3000	3	8	0.0009315411	0.0004737698	113364.14	50384.06	12596.016	176344.22
3000	3	12	0.0006566696	0.0003509255	117108.39	52048.17	13012.043	182168.61
3000	3	18	0.0006566696	0.0003509255	117108.39	52048.17	13012.043	182168.61
3000	5	2	0.0007204250	0.0002659008	67356.28	29936.13	7484.032	104776.44
3000	5	8	0.0008753114	0.0004683096	83434.90	37082.18	9270.545	129787.63
3000	7	1	0.0008642026	0.0002766943	45740.22	20328.98	5082.246	71151.45
3000	7	3	0.0005897481	0.0003024272	46435.15	20637.84	5159.461	72232.45
3000	7	8	0.0017476304	0.0008942874	137034.89	60904.40	15226.099	213165.39

Table 3.6: Shown are Mean Squared Error (MSE) values of all 1000 Maximum Likelihood estimates for all initial simulations of sample size n = 3000.



Sample_Size	Number_of_Samples	Sample_Frequency	Beta	Gamma	S0	I0	R0	Total_MSE
500	2	1	0.318522582	0.058561661	3295665.1	1464740.0	366185.01	5126590.5
500	3	1	0.074830182	0.016318060	1140218.9	506764.0	126690.99	1773673.9
500	4	1	0.029480147	0.007199880	702147.7	312065.6	78016.41	1092229.8
500	5	1	0.015274287	0.004269872	465279.7	206791.0	51697.75	723768.5
500	6	1	0.008677857	0.002675646	356101.8	158267.4	39566.86	553936.1
500	7	1	0.005987788	0.001926344	329110.8	146271.5	36567.87	511950.2
500	8	1	0.003892684	0.001328635	245970.3	109320.1	27330.03	382620.5

Table 3.7: Shown are Mean Squared Error (MSE) values of all 1000 Maximum Likelihood estimates for all further simulations of sample size n = 500.



Sample_Size	Number_of_Samples	Sample_Frequency	Beta	Gamma	S0	I0	R0	Total_MSE
3000	2	1	0.0482220194	0.0084046006	396253.08	176112.48	44028.120	616393.74
3000	3	1	0.0099024645	0.0021650653	151711.09	67427.15	16856.788	235995.04
3000	4	1	0.0044401450	0.0011259575	102226.72	45434.10	11358.524	159019.35
3000	5	1	0.0025132080	0.0006743219	81685.73	36304.77	9076.192	127066.69
3000	6	1	0.0014129463	0.0004176662	54983.48	24437.10	6109.276	85529.87
3000	7	1	0.0008744973	0.0002865670	45095.12	20042.27	5010.569	70147.96
3000	8	1	0.0006391912	0.0002209663	40066.70	17807.42	4451.855	62325.97

Table 3.8: Shown are Mean Squared Error (MSE) values of all 1000 Maximum Likelihood estimates for all further simulations of sample size n = 3000.

CHAPTER 4

Example: 2009 H1N1 Pandemic in Pullman Washington

4.1 Least Squares Fitting to H1N1 Data

In order to verify results I applied simulations to the 2009 H1N1 pandemic in Pullman, WA. Taking clinical data from the campus hospital on reported cases, I ran my own curve-fitting to obtain parameters for the SIR ODE equations. The data is from [2] and is shown as unfilled circles in Figure 4.1 along with the least-squares fit in red for $I(t)$. The resulting parameters in Table 4.1 were set according to my own fit, along with initial conditions from [2]. I ran 1000 simulations to estimate back these specific population parameters by using the best scenario: $\{q = 7, x = 1\}$. This example shows how a real random survey scheme could be carried out in a small university town such as Pullman. To reflect Pullman's population of 18,223 at the time I chose survey sample sizes of $n = 500$ and $n = 1500$.

4.2 Example H1N1 Results

The resulting histograms generally appeared centered at the population parameters and also mostly symmetric, with small variances (refer to Figures 4.2 and 4.3). The larger sample size again proved better at capturing the dynamics. The QQ plots looked adequate, though not very Normal due to heavy tails. The confidence intervals and MSE values decreased in width and size for the larger sample size $n = 1500$. The total normalized MSE value for the $n = 500$ sample was of similar value to the normalized MSE value of the best estimator in the main simulations above $\{n = 3000, q = 7, x = 1\}$, which means a satisfactory performance with real data.

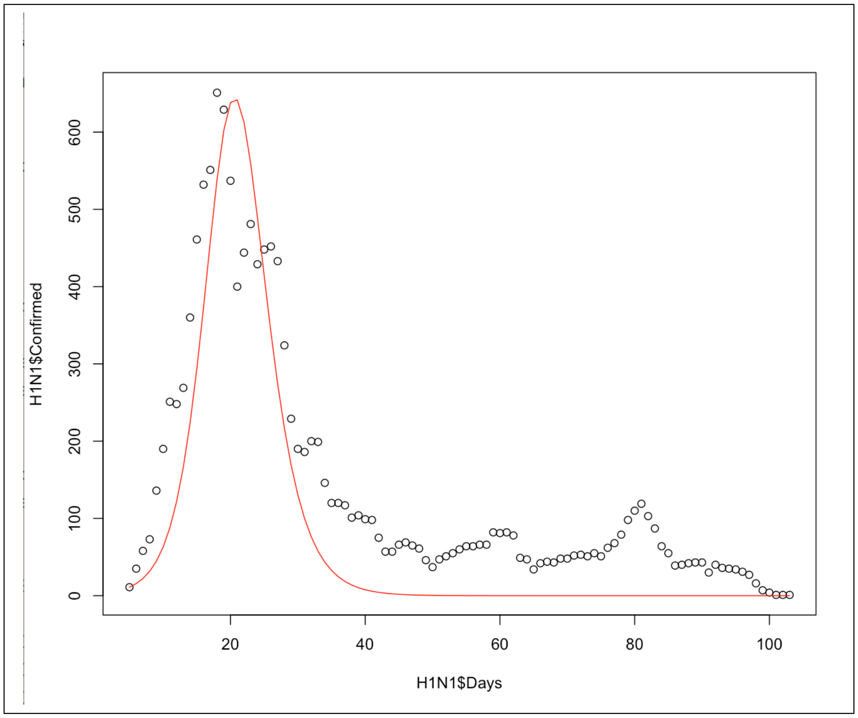
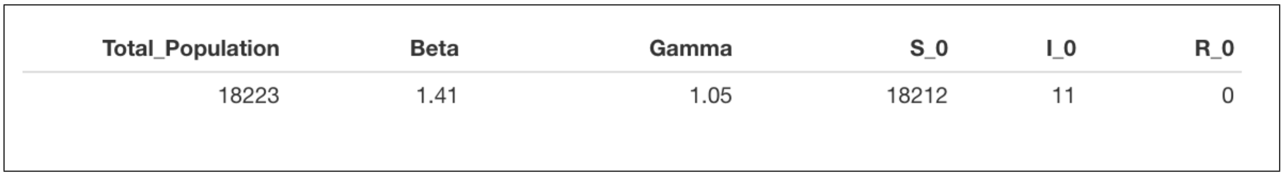
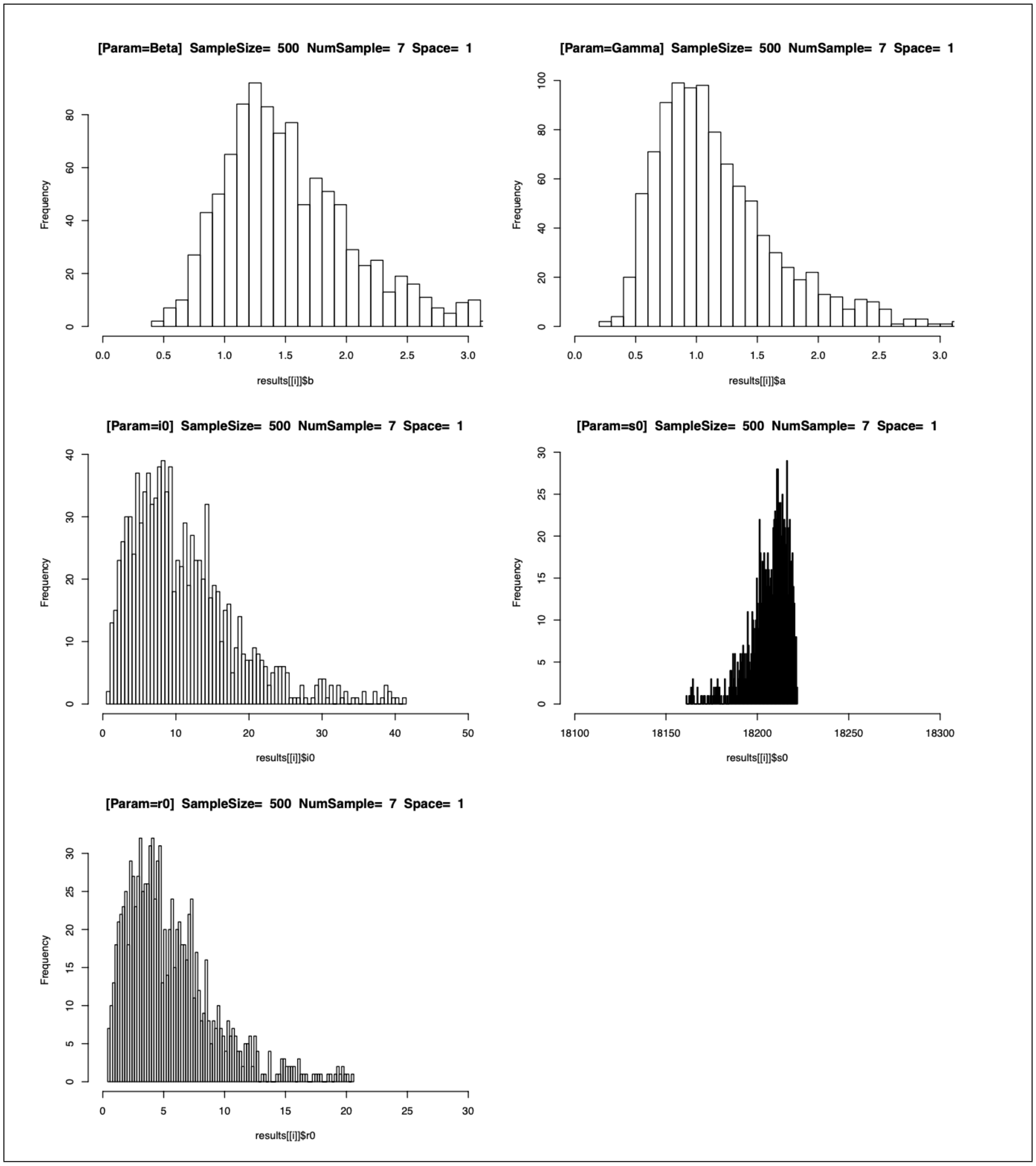


Figure 4.1: Shown are the Number of Cases over time (Days) for H1N1 in Pullman, WA plotted with the least squares fit of the SIR ODE model, showing in red the estimated $I(t)$ trajectory.



Total_Population	Beta	Gamma	S_0	I_0	R_0
18223	1.41	1.05	18212	11	0

Table 4.1: Shown is the set up for the initial population parameters and initial conditions for H1N1. Population dynamics are estimated back with Multinomial Maximum Likelihood. The best sampling scheme resulted in q = 7, x = 1 at both sample sizes, n = 500 and n = 1500.



[Param=Beta] SampleSize= 500 NumSample= 7 Space= 1

[Param=Gamma] SampleSize= 500 NumSample= 7 Space= 1

[Param=i0] SampleSize= 500 NumSample= 7 Space= 1

[Param=s0] SampleSize= 500 NumSample= 7 Space= 1

[Param=r0] SampleSize= 500 NumSample= 7 Space= 1

Figure 4.2: Shown are histograms of the five estimated parameters $\beta, \gamma, S(0), I(0), R(0)$ resulting from 1000 simulations of Multinomial Maximum Likelihood estimates with sample size n = 500 people. Each histogram can be identified by its parameter in square brackets. The setup used: q = 7 surveys, x = 1 day apart.

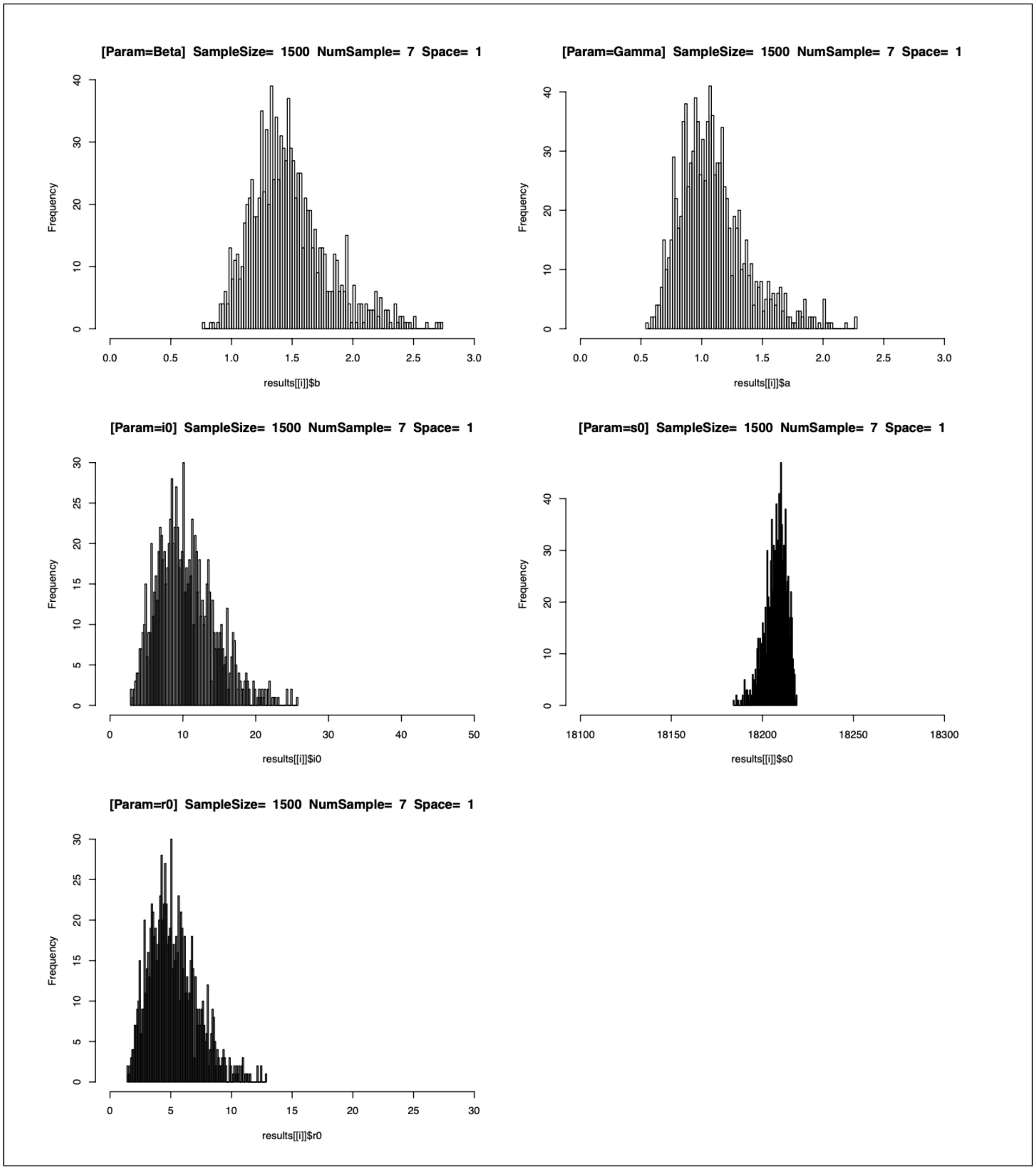


Figure 4.3: Shown are histograms of the five estimated parameters $\beta, \gamma, S(0), I(0), R(0)$ resulting from 1000 simulations of Multinomial Maximum Likelihood estimates with sample size n = 1500 people. Each histogram can be identified by its parameter in square brackets. The setup used: q = 7 surveys, x = 1 day apart.

CHAPTER 5

Discussion

My original hypothesis was that more samples further apart (larger q and x) would best capture the dynamics, as well as a larger sample size n. Larger sample size generally produced better estimates; however, larger q and x did not necessarily improve estimates. I had initially thought more samples spread apart would capture more changes in the ratios between the S, I, and R curves in the trajectory plots. Surprisingly, the better estimators were short spurts of at least 5 samples, 1 day apart, with the best being $\{n = 3000, q = 7, x = 1\}$, the second best being $\{n = 3000, q = 7, x = 3\}$, and the third best $\{n = 3000, q = 5, x = 2\}$. This leads to the conclusion that giving several surveys close together, early in the disease progression, will produce the best estimates of disease dynamics. In other words, counter-intuitively, knowing how a disease takes off early with precision is the best way to predict its eventual course. It is important to remember that though these shorter spurts were the best, most sampling schemes overall, except a few, were generally still good estimators. This leaves a good degree of flexibility for researchers conducting surveys with limited resources and time. In some cases, it is unknown early on as to when the disease has actually taken off, so surveys are more likely to be conducted a little later than the very beginning. However, the above simulations estimated not only the transmission parameters but also the initial conditions, so potentially, the results imply success for surveys conducted at different starting points.

Overall, the multinomial model estimated the original parameters independent of the original population size. This is key because the two sample sizes could be changed and run with further simulations, even while original population size N and initial conditions aren't the same. One major implication for researchers is that they can run the surveys on a small scale such as a small town or university town, capture the dynamics of the disease well, and then, presuming the disease strain is the same elsewhere, set up a model with these parameters for other towns or cities. Not only would this save resources, but examining a smaller closed population such as a university would likely capture the dynamics better.

Conducting one week of surveys every day, one could examine the same disease in another area promptly, such as a city with a larger population. This might prove helpful in outbreaks such as influenza and measles where transmission takes place at different times in different locations. If one conducts the surveys early on in one location, one can use the results to predict outcomes on the populations effected later in time. In all cases I would recommend running a selection of simulations with hypothetical surveys and see how well they estimate back before conducting an actual survey.

One of the major disadvantages in the simulations was time inefficiency in the algorithm. The optimization technique relied entirely on the high precision Nelder-Mead algorithm, but the code took 20 minutes per sampling scheme, for a total of 440 minutes for the main results. Some of the code could be improved by writing parallel computations and possibly trying other optimization techniques, including Monte Carlo and Simulated Annealing algorithms. Another improvement would be running the code on high computing machines or cloud computing platforms. With greater efficiency or computing power, the code could certainly be turned into an R package or applet for epidemiologists or other researchers in the field.

This project can and should be extended to alternate original population parameters, different sample sizes, and further sampling schemes. In other words, the simulations should be extended to other sets of q and x (number of samples and spacing of samples), as well as to various possibilities for n and for $N, \beta, \gamma, I(0), R(0), S(0)$. To truly examine how well the multinomial model performs, it would be better to run all the above simulations with various examples for starting population parameters and then compare results. In addition, the investigation can certainly be extended to other epidemiological models, such as SIR models with vaccine, SEIR models, SIS models, and stochastic models. It remains unknown how well these can be estimated. Complexity increases when there are more parameters to be estimated and the survey becomes difficult to design. For example, how does one create a survey for the exposed class in an SEIR model? With vaccine models, whether one has been vaccinated would need to be included as a question. Due to computational limitations

I chose to focus on the most basic model, but eventually I hope to extend the investigation to these other models and sampling scenarios.

References

[1] N. K. Vaidya, M. Morgan, T. Jones, L. Miller, S. Lapin, E. J. Schwartz, ”Modelling the epidemic spread of an H1N1 influenza outbreak in a rural university town”, Cambridge University Press 2014. Epidemiol. Infect.m(2015), 143, 1610-1620

[2] L. Miller, T. Jones, M. Morgan, S. Lapin, E. J. Schwartz, ”Individual-based computational model used to explain 2009 pandemic H1N1 in rural campus community”, Journal of Biological Systems. World Scientific Publishing Company, Vol. 21, No. 4 (2013) 1340005.

[3] John M. Drake and Pejman Rohani with contributions from Ben Bolker, Matt Ferrari, Aaron King and Dave Smith, ”Estimating model parameters by maximum likelihood Measles in Niamey, Niger” 2006 Sep;100(9):867-73. E Pub 2006 Mar 15.

[4] Brian Dennis, William P. Kemp, Roy C. Beckwith ”Stochastic Model of Insect Phenology Estimation and Testing” Environmental Entomology. Vol 15 no. 3 (1986). pp. 540-546

[5] Zhilan Feng, Sherry Towers, and Yiding Yang, ”Modeling the Effects of Vaccination and Treatment on Pandemic Influenza”, The AAPS Journal, Vol. 13, No. 3, September 2011.

[6] Francisco J. Samaniego, STOCHASTIC MODELING AND MATHEMATICAL STATISTICS. CRC Press. Taylor and Francis Group. 2014. pp. 191-192

Appendix

All further simulated histograms, QQ plots, and summary tables can be accessed online via Supplemental Material at the following link:

https://github.com/hull1893/masters-thesis-supplemental