Schatting van SARS-CoV-2-transmissieparameters in huishoudens

Bachelorscriptie Wiskunde en toepassingen

Frida Bomhof - studentnummer 2509393
Begeleider: Martin Bootsma

14 juni 2024

Abstract

We hebben verspreiding van SARS-CoV-2 binnen huishoudens onderzocht, door transmissieparameters in huishoudens te schatten met data van het Universitair Medisch Centrum Utrecht. Een transmissieparameter geeft aan hoe groot de kans is dat één persoon iemand anders in het huishouden besmet. Er hebben 307 huishoudens meegedaan aan de studie, bij 59 van deze huishoudens was minstens één besmetting gedurende de tijd dat het huishouden gevolgd is. De periode waarin huishoudens gevolgd zijn ligt tussen september 2020 en juli 2021. We gebruiken *final size data*. We bekijken één model zonder leeftijdsonderscheid en één model met twee leeftijdscategorieën. Bij het eerste model schatten we de transmissieparameter eerst met behulp van een waarschijnlijkheidsfunctie en daarna met behulp van het *Monte Carlo Markov Chain*-algoritme (MCMC-algoritme). Bij het tweede model, gebruiken we alleen het MCMC-algoritme om vier transmissieparameters te schatten. Uit het eerste model volgt dat de kans dat een persoon een ander besmet 0.19 is, beide methoden geven dezelfde schatting. Uit het tweede model volgt dat de kans dat een kind een kind besmet het grootst is, de kans dat een volwassene een kind besmet is het kleinst. Bij het besmetten van kinderen, zijn kinderen significant besmettelijker. In vervolgonderzoek zou er gekeken kunnen worden naar een model waarbij de contactintensiteit van elk paar van individuen in een huishouden kleiner wordt als het huishouden groter wordt, in ons model blijft deze intensiteit gelijk. Ook kan er bijvoorbeeld gekeken worden naar het gebruik van de gerapporteerde besmettingstijden of data uit een andere tijdsperiode.

Inhoudsopgave

1 Inleiding

SARS-CoV-2 is het virus dat de oorzaak is van COVID-19. Dit virus had veel invloed op de volksgezondheid en ook op de maatschappij door de maatregelen om verspreiding van het virus tegen te gaan. Het is daarom belangrijk om meer te weten te komen over de verspreiding van dit virus. In deze scriptie onderzoeken we het deelonderwerp verspreiding binnen huishoudens. Verspreiding binnen huishoudens is relatief goed te modelleren omdat er binnen huishoudens weinig mogelijkheden voor verspreiding zijn in vergelijking met verspreiding buiten huishoudens. Een transmissieparameter geeft aan hoe groot de kans is dat één persoon iemand anders in het huishouden besmet. Door transmissieparameters binnen huishoudens te schatten kunnen we dus iets zeggen over de besmettelijkheid van individuen in huishoudens.

Het is belangrijk om goede schattingen voor de transmissieparameters te hebben, omdat dit bijvoorbeeld informatie kan geven over effectieve maatregelen tegen de verspreiding van SARS-CoV-2. Ook de verhouding tussen besmettelijkheid van kinderen en volwassenen kan hier informatie over geven. Door verschillende manieren te gebruiken om de parameters te schatten, hebben we meer zekerheid over de betrouwbaarheid van de schattingen. Ook kan er vergeleken worden welke manier het beste werkt voor vervolgonderzoek. Door te kijken of mogelijke correlatie tussen verschillende transmissieparameters klopt met de verwachtingen, kan ook meer zekerheid over schattingen verkregen worden.

We beginnen deze scriptie met een methode waarin we eerst uitleg geven over de verkregen data (paragraaf 2.1) en een model zonder onderscheid in leeftijdscategorieën (paragraaf 2.2). We geven hier een uitleg over het model zelf en het bepalen van uitdrukkingen voor kansen (handmatig en algoritmisch). Daarna laten we zien hoe we de transmissieparameter voor dit model kunnen schatten, hiervoor beginnen we met het bepalen van het maximum van de waarschijnlijkheidsfunctie en vervolgens gebruiken we het *Monte Carlo Markov Chain*-algoritme (MCMC-algoritme). In paragraaf 2.3 bekijken we een model met twee leeftijdscategorieën en meerdere transmissieparameters. Hier gebruiken we alleen het MCMC-algoritme. Verder kijken we bij dit model naar correlatie tussen transmissieparameters en verhoudingen tussen de besmettelijkheid van kinderen en volwassenen. In hoofdstuk 3 zijn de resultaten te vinden. We eindigen met een conclusie (hoofdstuk 4) en discussie (hoofdstuk 5).

2 Methode

2.1 Data

We gebruiken in deze scriptie data van een studie waarin huishoudens met minstens één kind jonger dan 18 jaar zijn geïncludeerd. Deze huishoudens komen uit drie verschillende Nederlandse studies, namelijk: 'Respiratory Syncytial Virus Consortium in Europe', 'Microbiome Utrecht Infant Study' en 'Generation R cohort'. Bij de eerste waren de kinderen in de bijbehorende huishoudens het jongst en bij de laatste het oudst. Uiteindelijk hebben 307 huishoudens met 1209 gezinsleden meegedaan. Hiervan waren 582 personen jonger dan 18, de rest was volwassen [2].

Voor het verzamelen van de data door het Universitair Medisch Centrum Utrecht is gebruik gemaakt van een *prospective household-based* studie. Dit betekent dat alle individuen in de gekozen huishoudens gedurende een bepaalde periode gevolgd worden om bij te houden of er infecties in deze huishoudens plaatsvinden. Er zijn verschillende manieren gebruikt om uitbraken (van SARS-CoV-2 of een ander virus) in de huishoudens te detecteren. Als eerste moesten de personen in de huishoudens elke dag in een app aangeven of ze last hadden van koorts of symptomen met betrekking tot de ademhalingswegen. Bovendien werd - ook als de personen geen symptomen hebben - er één keer per 4 tot 6 weken getest of ze besmet zijn met een virus met betrekking tot de ademhalingswegen. Een zogenaamde 'uitbraak studie' begint zodra aan één van de volgende criteria voldaan is: iemand wordt bij bovenvermelde test positief getest, iemand wordt extern positief getest, iemand krijgt koorts of iemand krijgt symptomen met betrekking tot de ademhalingswegen. Een uitbraak studie houdt in dat het hele huishouden vaker en op verschillende manieren getest wordt [6]. Om zeker te weten dat de uitbraak in het huishouden afgelopen is, worden alle personen in het huishouden nog één keer getest tien dagen na het einde van de uitbraak. Aan de criteria voor het beginnen van een uitbraak studie kan dus worden voldaan door een besmetting met SARS-CoV-2 of door een besmetting met een ander virus. In het geval van besmetting met SARS-CoV-2, wordt na deze uitbraak het huishouden niet meer gevolgd. Dat betekent dat als er na het einde van zo'n uitbraak in een huishouden, een nieuwe besmetting met SARS-CoV-2 in ditzelfde huishouden plaatsvindt, deze niet meer wordt geregistreerd [2].

Tijdens de studie waren 205 volwassenen ongevaccineerd, 74 volwassenen zijn tenminste één keer gevaccineerd tijdens de studie (voordat de uitgebreide studie begon). 287 volwassenen zijn ook tenminste één keer gevaccineerd tijdens de studie, maar pas tijdens de uitgebreide studie. Verder is het van 58 volwassenen niet bekend of ze gevaccineerd zijn en er zijn geen kinderen gevaccineerd [2]. Tijdens het grootste deel van de studie, waren de varianten van het virus die het meeste voorkwamen het wildtype virus en de alfavariant, vanaf eind juni 2021 werd dit de deltavariant.

We gebruiken alleen de data over de huishoudens waarin minstens 1 besmetting is geweest gedurende de tijd dat het huishouden gevolgd is, dit is het geval bij 59 van de 307 huishoudens. In deze 59 huishoudens zijn in totaal 119 van de 237 personen besmet.

We maken gebruik van *final size data*, deze geeft weer hoeveel personen in een huishouden aan het eind van de periode waarin het huishouden gevolgd wordt besmet zijn. De startdatum voor het volgen van een huishouden verschilt per huishouden. Tussen september 2020 en januari 2021 startte steeds het volgen van ongeveer evenveel huishoudens, vanaf januari 2021 nam het aantal af. 161 dagen is de maximale volglengte van de huishoudens; de meeste huishoudens zijn ook voor 161 dagen gevolgd. Er is ook een meer uitgebreide studie gedaan waarbij een deel van de huishoudens nog langer gevolgd zijn, tot maximaal juli 2021.

Voor alle personen in de data is de leeftijdscategorie (kind jonger dan 12 jaar, jongere vanaf 12 tot 18 jaar of volwassene) gegeven en is bekend of ze besmet zijn geraakt en of ze een indexgeval

zijn. Indexgevallen zijn de personen die besmet zijn bij de start van de uitbraak. Als er meerdere indexgevallen in een huishouden zijn, nemen we aan dat deze tegelijk besmet zijn geraakt. Als alle personen in een huishouden indexgeval zijn, verwijderen we dit huishouden uit de data [6].

We definiëren de volgende 3 stochastische variabelen:

$X :=$ totaal aantal besmette personen in een huishouden aan het eind van de gevolgde periode,

$N :=$ aantal niet-geïnfecteerde personen in een huishouden bij start van de uitbraak,

$I :=$ aantal indexgevallen in een huishouden.

Verder gebruiken we voor deze stochastische variabelen het subscript k om aan te geven dat we alleen naar het aantal kinderen en jongeren (leeftijdscategorie tot en met 17 jaar) kijken. Het subscript v wordt gebruikt voor het aantal volwassenen vanaf 18 jaar. De data is te zien in Tabel 4 en Tabel 5. Een samenvatting van de data is te zien in Tabel 1.

	X	X_k	X_v	N	N_k	N_v	I	I_k	I_v
Gemiddelde	2.02	0.86	1.15	2.71	1.54	1.17	1.31	0.49	0.81
Standaardafwijking	1.30	0.95	0.73	1.04	0.93	0.69	0.64	0.62	0.62

Tabel 1: Samenvatting van data.

2.2 Geen leeftijdsonderscheid

We beschrijven eerst de methode om de transmissieparameter te schatten voor het model zonder leeftijdsonderscheid.

2.2.1 Model

We beginnen met een geïdealiseerde situatie waarin twee individuen in een huishouden contact met elkaar hebben. Deze individuen hebben verder geen contacten.

Definieer

$p(t) :=$ kans dat een op tijd 0 besmette persoon een andere persoon niet besmet heeft op tijd t,

$c :=$ het aantal contacten per tijdseenheid dat de huisgenoten met elkaar hebben,

$g(t) :=$ kans op besmetting van een vatbaar persoon bij contact met een besmette huisgenoot als de besmette huisgenoot een tijdsduur t geleden besmet is geraakt.

Dan is $cg(t)$ is de infectiedruk die een vatbaar persoon in het huishouden ervaart van het besmette individu. Er geldt

$$\frac{dp(t)}{dt} = -cg(t)p(t). \tag{1}$$

We kunnen $cg(t)$ als volgt herschrijven:

$$cg(t) = c \int_0^t g(\tau)\, d\tau \frac{g(t)}{\int_0^t g(\tau)\, d\tau}.$$

Hierbij nemen we aan dat $g(t)$ een integreerbare functie is. Dit is in de praktijk altijd het geval. De besmettelijk periode is vaak kort en $g(t)$ zal zeker gelijk zijn aan nul als t groter is dan de maximale

leeftijd die mensen bereiken.

Definieer

$$R := c \int_0^t g(\tau)\, d\tau,$$
$$f(t) := \frac{g(t)}{\int_0^t g(\tau)\, d\tau}.$$

R is dan een maat voor de besmettelijkheid van een individu en $f(t)$ is de generatietijdverdeling. De generatietijd is de hoeveelheid tijd tussen het moment dat het indexgeval geïnfecteerd is en het moment dat het indexgeval andere personen infecteert [5]. Door in (1) $cg(t)$ te vervangen door $Rf(t)$ krijgen we:

$$\frac{dp(t)}{dt} = -Rf(t)p(t).$$

Verder geldt

$$p(0) = 1,$$

omdat op tijd $t = 0$ er nog geen besmetting plaatsgevonden kan hebben.

Oplossen van de differentiaalvergelijking geeft:

$$p(t) = e^{-R\int_0^t f(\tau)\, d\tau}. \tag{2}$$

Omdat we naar *final size data* kijken, willen we geen uitdrukking die afhangt van t. Daarom doen we de aanname dat als een willekeurige persoon iemand besmet, dit altijd snel gebeurt na de besmetting van de willekeurige persoon. Dus de infectiedruk wordt heel klein als t groter wordt. Daarom kunnen we in (2) de limiet voor t naar ∞ nemen. Per definitie van de generatietijd volgt dat $\int_0^\infty f(\tau)\, d\tau = 1$. We krijgen nu dat de kans p dat een persoon een ander persoon niet besmet gelijk is aan

$$p := \lim_{t\to\infty} p(t) = e^{-R\int_0^\infty f(\tau)\, d\tau} = e^{-R}.$$

Het bovenstaande is geldig voor een huishoudgrootte gelijk aan 2. Voor grotere huishoudens kunnen we dezelfde manier gebruiken. Het verschil is dat een persoon dan niet wordt geïnfecteerd, als deze persoon aan de infectiedruk van alle andere personen in het huishouden ontsnapt. We gebruiken hierbij een *density-dependent* model, dit betekent dat de contactintensiteit van elk paar van individuen in een huishouden niet afhangt van de grootte van het huishouden. Dus het totale aantal contacten van een individu is groter bij een groter huishouden [1].

2.2.2 Handmatig kansen bepalen

Om de parameters te schatten hebben we een uitdrukking voor de waarschijnlijkheidsfunctie nodig, voor de waarschijnlijkheidsfunctie hebben we weer uitdrukkingen voor $\mathbb{P}(X = x \mid N = n, I = i)$ nodig voor verschillende combinaties van x, n en i. We bekijken nu hoe we deze kansen handmatig kunnen bepalen.

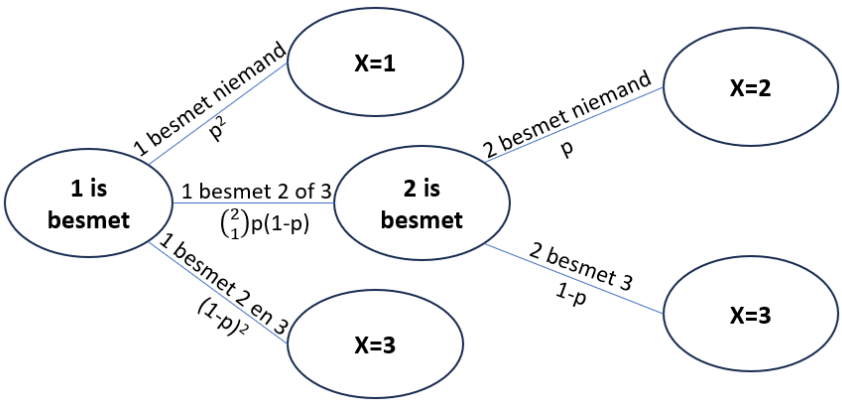
We beginnen met de gevallen waar $I = 1$. We nummeren de personen in een huishouden als $1, ..., n+i$. Zonder verlies van algemeenheid nemen we aan dat de indexgevallen de laagste nummers

hebben.

Voor $N = 1$ krijgen we per definitie van p dat

$$\mathbb{P}(X = 1 \mid N = 1, I = 1) = p \quad \text{en}$$
$$\mathbb{P}(X = 2 \mid N = 1, I = 1) = 1 - p.$$

Voor $N = 2$ maken we de kansboom te zien in Figuur 1.



Figuur 1: Kansboom voor $N = 2$. De verschillende mogelijkheden staan boven de takken en de bijbehorende kansen eronder, de uitkomsten staan in de ovalen.

In Figuur 1 is op het begin alleen persoon 1 besmet. Er zijn dan drie mogelijkheden gegeven in de kansboom: '1 besmet niemand', '1 besmet 2 of 3' en '1 besmet 2 en 3'. De bijbehorende kansen staan onder deze takken. Voor de eerste mogelijkheid zijn er geen nieuwe personen besmet die weer anderen zouden kunnen besmetten. Dus we hebben geen nieuwe takken vanaf de uitkomst '$X = 1$'. Bij de derde mogelijkheid is iedereen besmet dus er komen ook geen nieuwe takken meer vanaf '$X = 3$'. Voor de uitkomsten in de ovalen nemen we zonder verlies van algemeenheid aan dat steeds (ook in de latere kansbomen) de persoon met het laagste nummer besmet is geraakt. Daarom hebben we na '1 besmet 2 of 3', de uitkomst '2 is besmet'. Vanaf '2 is besmet' hebben we wel nieuwe takken: '2 besmet niemand' (uitkomst '$X = 2$') of '2 besmet 3' (uitkomst '$X = 3$'). In het algemeen geldt dat we vanaf een uitkomst nieuwe takken krijgen, als in deze uitkomst minstens één persoon is besmet en er zijn andere personen nog niet besmet. Als we geen nieuwe takken krijgen, schrijven we in de ovaal het totaal aantal besmettingen X.

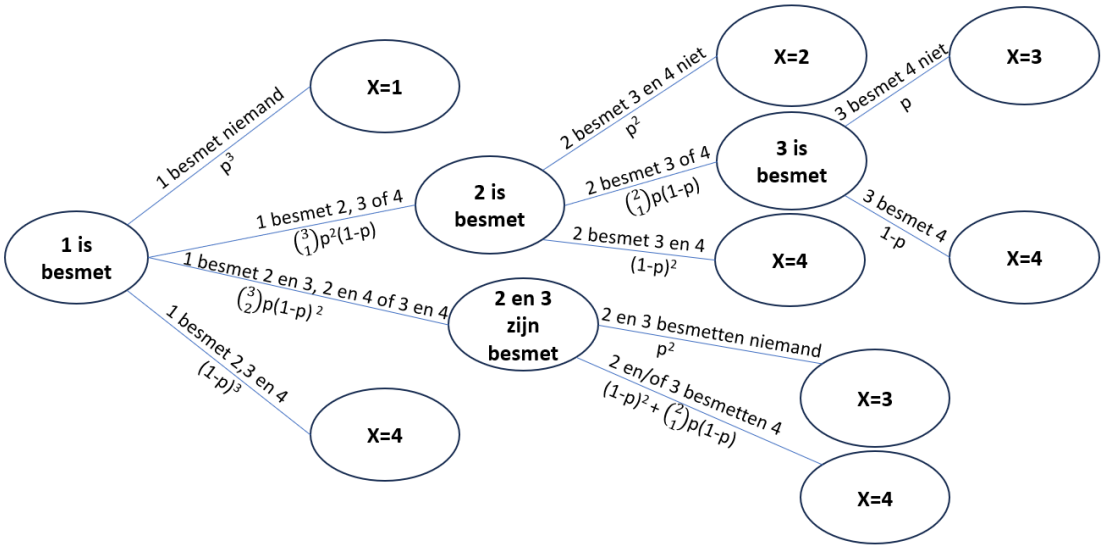
Er zijn dus twee routes die uitkomen in $X = 3$. We kunnen de route volgen die begint met '1 besmet 2 of 3' en daarna verder gaat met '2 besmet 3'. De kans hierop is $\binom{2}{1}p(1-p) \cdot (1-p)$. De andere mogelijkheid is de route '1 besmet 2 en 3'. De kans hierop is $(1-p)^2$. Hieruit volgt

$$\mathbb{P}(X = 3 \mid N = 2, I = 1) = 2p(1-p)^2 + (1-p)^2.$$

Op dezelfde manier krijgen we

$$\mathbb{P}(X = 1 \mid N = 2, I = 1) = p^2 \quad \text{en}$$
$$\mathbb{P}(X = 2 \mid N = 2, I = 1) = 2p^2(1-p).$$

Voor $N = 3$ krijgen we op dezelfde manier als voor $N = 2$ de kansboom in Figuur 2.



Figuur 2: Kansboom voor $N = 3$. De verschillende mogelijkheden staan boven de takken en de bijbehorende kansen eronder, de uitkomsten staan in de ovalen.

We zien in de kansboom voor $N = 3$ dat vanaf de ovaal '2 is besmet' we dezelfde kansboom als bij $N = 2$ hebben. Het enige verschil is dat de nummering van de personen één is opgeschoven, doordat we in deze ovaal beginnen met '2 is besmet' en bij $N = 2$ begonnen we met '1 is besmet'. Bij het maken van kansbomen moet rekening worden gehouden met de mogelijkheid dat 2 personen tegelijk dezelfde persoon kunnen besmetten. Dit is te zien bij de ovaal '2 en 3 zijn besmet', voor de onderste tak vanaf deze ovaal is het ook mogelijk dat personen 2 en 3 allebei persoon 4 besmetten.

We kunnen nu met Figuur 2 de kans $\mathbb{P}(X = x \mid N = 3, I = 1)$ bepalen voor $x = 1$, $x = 2$, $x = 3$ en $x = 4$. Niet alle situaties komen ook daadwerkelijk voor in de data, de kansen voor de situaties die wel voorkomen zijn te vinden in de bijlage in Tabel 6. Hier zijn ook voor $N = 4$ en $N = 5$ de benodigde kansen te vinden, die ook bepaald zijn met behulp van een kansboom. Verder is voor elke situatie in de tabel te vinden, bij hoeveel huishoudens precies deze situatie voorkomt. Dit getal definiëren we als m_r, m_r wordt later precies gedefinieerd.

In de data hebben we ook situaties waar $I = i$ met $i > 1$. Voor deze situaties kunnen we de kansbomen voor $I = 1$ en $N = 3$, $N = 4$, $N = 5$ en $N = 6$ gebruiken. Als voorbeeld bekijken we $I = 2$ en $N = 1$. In dit geval kunnen we Figuur 2 gebruiken. De nieuwe kansboom die we nu nodig hebben is de kansboom die we krijgen als we in Figuur 2 alleen de ovaal '2 en 3 zijn besmet' en de takken en ovalen die vanaf daar naar rechts gaan bekijken. Dan hebben we namelijk een situatie waarbij twee personen aan het begin besmet zijn (personen 2 en 3) en één persoon aan het begin niet besmet is (persoon 4). Doordat persoon 1 nu niet besmet is, krijgen we als uitkomst $X = 2$ of $X = 3$ in plaats van $X = 3$ of $X = 4$. Uit deze nieuwe kansboom volgt dan dat

$$\mathbb{P}(X = 2 \mid N = 1, I = 2) = p^2 \quad \text{en}$$
$$\mathbb{P}(X = 3 \mid N = 1, I = 2) = 2p(1-p) + (1-p)^2.$$

Voor andere situaties waar $I = i$ met $i > 1$ kunnen we op dezelfde manier een nieuwe kansboom halen uit de gemaakte kansbomen met $I = 1$. Om de bijbehorende kansen te bepalen moeten in de meeste gevallen deze kansen opnieuw bepaald worden door de verschillende routes voor een bepaalde waarde van X te bekijken en voor elke route de kansen bij de bijbehorende takken te

vermenigvuldigen. Alle op deze manier bepaalde kansen die we nodig hebben zijn te vinden in de bijlage, in Tabel 6.

2.2.3 Kansen algoritmisch bepalen

We hebben de formules voor deze kansen ook algoritmisch bepaald. We gebruiken Python om de kansen iteratief te bepalen. Alle Python codes zijn te vinden op Github, de gebruikte versie van Python is Python 3.8. Hiervoor definiëren we eerst de stochastische variabele

$$H = N + I,$$

deze stochastische variabele geeft de huishoudgrootte weer. De kansen worden voor elke combinatie van huishoudgrootte, aantal indexgevallen en aantal besmettingen opgeslagen in een driedimensionale matrix. We gebruiken nu dus de stochastische variabelen H, I en X, in plaats van N, I en X. We initialiseren de driedimensionale matrix met nullen. We definiëren eerst alle kansen behorend bij huishoudgrootte 1 en alle kansen behorend bij geen indexgevallen in deze matrix. Deze kansen zijn te zien in Tabel 2.

x	h	i	$\mathbb{P}(X = x \mid H = h, I = i)$
0	0,1,2,...	0	1
1,2,...	0,1,...	0	0
1	1	1	1
0	1	1	0

Tabel 2: Kansen nodig voor iteratief kansen bepalen.

De rest van de kansen kan nu iteratief bepaald worden. Hiervoor gebruiken we geneste *for-loops* over x, h en i. De structuur is in Algoritme 1 te zien.

Algoritme 1 Structuur geneste *for-loops* zonder leeftijdsonderscheid

```
for h = 2 to 6 do
  for i = 1 to h do
    for x = 1 to h do
      // Bepaal kans
    end for
  end for
end for
```

Met *to* bedoelen we tot en met. h loopt dus tot en met 6, 6 is de maximale huishoudgrootte. In de binnenste *for-loop* bepalen we de bijbehorende kans, hiervoor definiëren we

$$J := \text{aantal besmettingen in de eerste generatie.}$$

Dus dit geeft het aantal vatbare personen die door indexgevallen besmet zijn geraakt. Voor elke mogelijke waarde j die J kan aannemen bepalen we de kans en deze kansen tellen we op. Er geldt

$$\mathbb{P}(X = x \mid H = h, I = i) = \sum_{j=0}^{h-i} \mathbb{P}(X = x - i \mid H = h - i, I = j)(1 - p^i)^j p^{i^{h-i-j}} \binom{h-i}{j}.$$

p^i geeft hierbij de kans dat i indexgevallen een vatbare persoon niet besmetten. Er zijn in totaal $h-i$ vatbare personen. Dus $J \sim \text{Bin}(h-i, p^i)$. Hieruit volgt dat de kans op j besmettingen door i indexgevallen gelijk is aan $(1-p^i)^j p^{i^{h-i-j}} \binom{h-i}{j}$. De besmettingen door indexgevallen zijn de besmettingen in de eerste generatie. We moeten nu kijken naar besmettingen in de volgende generaties. Voor de tweede generatie bekijken we de personen die besmet zijn geraakt door de j personen. Dan krijgen we $\mathbb{P}(X = x - i \mid H = h - i, I = j)$. We nemen hier $I = j$, omdat er nu j personen kunnen besmetten. Verder hoeven we nu niet meer naar de indexgevallen te kijken, dus we nemen voor X de waarde $x-i$ en voor H de waarde $h-i$. De kans $\mathbb{P}(X = x-i \mid H = h-i, I = j)$ wordt iteratief bepaald.

2.2.4 Waarschijnlijkheidsfunctie

Er komen 22 verschillende combinaties van x, n en i voor in de data. Deze nummeren we als x_r, n_r en i_r voor r van 1 tot en met 22. We definiëren m_r als het aantal keer dat een combinatie van x_r, n_r en i_r voorkomt in de data. De waarschijnlijkheidsfunctie voor p is dan gegeven door [4]:

$$L(p) = \prod_{r=1}^{22} (\mathbb{P}(X = x_r \mid N = n_r, I = i_r, p))^{m_r}.$$

De log waarschijnlijkheidsfunctie is gegeven door [4]:

$$l(p) := \log(L(p)) = \sum_{r=1}^{22} m_r \log(\mathbb{P}(X = x_r \mid N = n_r, I = i_r, p)).$$

Om een schatting voor p te bepalen, maken we gebruik van *sympy* in Python. Hiermee stellen we de afgeleide van de log waarschijnlijkheidsfunctie gelijk aan 0. Dit geeft een schatting voor p, deze noemen we $\hat{p}$. Door een plot te maken van de waarschijnlijkheidsfunctie, kan er gecontroleerd worden dat in $\hat{p}$ inderdaad een maximum is.

We gebruiken een *Wald interval estimator*, gegeven door

$$\left[\hat{p} - z\left(\frac{\alpha}{2}\right) \frac{1}{\sqrt{nI(\hat{p})}}, \hat{p} + z\left(\frac{\alpha}{2}\right) \frac{1}{\sqrt{nI(\hat{p})}} \right],$$

als $100(1-\alpha)\%$ betrouwbaarheidsinterval [4]. Hierbij geldt dat

$$I(\hat{p}) = \frac{1}{59} \sum_{i=1}^{59} \left(\frac{\partial}{\partial \hat{p}} \log \mathbb{P}(X = x_r \mid N = n_r, I = i_r, \hat{p}) \right)^2.$$

Verder geldt dat $z\left(\frac{\alpha}{2}\right) = 1.96$ voor een 95%-betrouwbaarheidsinterval. Dit bepalen we ook in Python, we krijgen een betrouwbaarheidsinterval die symmetrisch is rond $\hat{p}$. Als laatste kunnen we door te gebruiken dat $p = e^{-R}$ een schatting en betrouwbaarheidsinterval voor R bepalen.

2.2.5 Monte Carlo Markov Chain

In plaats van de afgeleide van de waarschijnlijkheidsfunctie gelijk te stellen aan 0, gebruiken we nu het MCMC-algoritme om een schatting voor p te bepalen. Het MCMC-algoritme is een voorbeeld van een Bayesiaanse methode. Bij dit algoritme beginnen we met een bepaalde waarde voor p.

We passen deze waarde van p iets aan, deze nieuwe waarde accepteren we of wijzen we af. Bij acceptatie gaan we verder met de nieuwe waarde en passen deze aan, bij afwijzing gaan we verder met het aanpassen van de oude waarde. Zo krijgen we een reeks van waarden van p.

Om te bepalen of we de nieuw voorgestelde waarde voor p accepteren of afwijzen, hebben we de posterior-verdeling nodig. Er geldt

$$f_{P|X,N,I}(p|x,n,i) \propto L(p) \cdot f_P(p),$$

waarbij $f_{P|X,N,I}(p|x,n,i)$ de kansdichtheidsfunctie van de posterior-verdeling is en $f_P(p)$ de kansdichtheidsfunctie van de prior-verdeling. De functie voor de prior-verdeling geeft weer wat we verwachten voor p voordat we de data voor X, N en I hebben. De functie voor de posterior-verdeling geeft de informatie over p weer nadat we de data hebben [4]. In ons geval nemen we voor de prior-verdeling een uniforme verdeling van 0 tot 1. We definiëren $g(p)$ als de kansdichtheidsfunctie voor de posterior-verdeling:

$$g(p) := f_{P|X,N,I}(p|x,n,i)$$

Het MCMC-algoritme die wij gebruiken bestaat uit de volgende stappen [7]:

1. We kiezen de startwaarde $p_{\text{oud}} = 0.5$.
2. Voor de nieuwe waarde nemen we $p_{\text{nieuw}} = p_{\text{oud}} + a$ waarbij a normaal verdeeld is met gemiddelde 0 en variantie 0.09. Dus de kans dat p groter wordt is gelijk aan de kans dat p kleiner wordt. Het is mogelijk dat p_{nieuw} buiten $[0,1]$ komt te liggen. Als $p_{\text{nieuw}} > 1$, dan nemen we $2 - p_{\text{nieuw}}$ in plaats van p_{nieuw}. Als $p_{\text{nieuw}} < 0$, dan nemen we $-p_{\text{nieuw}}$ in plaats van p_{nieuw}. Dit doen we net zolang totdat $p_{\text{nieuw}} \in [0,1]$.
3. Als $g(p_{\text{nieuw}}) \geq g(p_{\text{oud}})$, dan accepteren we p_{nieuw}. Als $g(p_{\text{nieuw}}) < g(p_{\text{oud}})$, dan bepalen we $\frac{g(p_{\text{nieuw}})}{g(p_{\text{oud}})}$, dit geeft de kans waarmee de nieuwe waarde van p geaccepteerd wordt.
4. Als p_{nieuw} geaccepteerd is, nemen we $p_{\text{oud}} = p_{\text{nieuw}}$. Als p_{nieuw} afgewezen is, dan houden we $p_{\text{oud}} = p_{\text{oud}}$.
5. Dit is de eerste iteratie, we slaan de waarde van p_{oud} op. We gaan daarna terug naar stap 2 en stoppen na 100000 iteraties.

Doordat we een variantie van 0.09 voor a hebben gekozen, is de kans op afwijzing ongeveer twee derde. Door de variantie groter te maken, wordt er vaker afgewezen. Als we de variantie juist kleiner maken, wordt er minder vaak afgewezen maar verandert p langzaam.

We maken een *traceplot* van de opeenvolgende waarden van p. We gebruiken nu een belangrijke eigenschap behorend bij het MCMC-algoritme. Zodra convergentie van de reeks waarden van p is bereikt, geeft deze reeks een willekeurige steekproef van de posterior verdeling $g(p)$ [3]. We bepalen een schatting voor p met behulp van een histogram, we willen hiervoor dus alleen de reeks waarden van p gebruiken behorend bij iteraties nadat er convergentie heeft plaatsgevonden. In de *traceplot* is te zien dat dit altijd het geval is voordat we iteratie 500 hebben bereikt. We maken ook twee plots die de waarde die de waarschijnlijkheidsfunctie aanneemt in de iteraties geven. Voor één plot bekijken we alleen de eerste 2500 iteraties en voor de andere plot bekijken we alle iteraties. In de eerste plot kunnen we zien dat vanaf iteratie 500 de waarschijnlijkheid gemiddeld ook niet meer stijgt, met de tweede plot kunnen we controleren dat dit inderdaad nog steeds geldt vanaf iteratie 2500. Dit betekent dat we de waarden van p vanaf iteratie 500 aan het histogram toevoegen. De periode tot iteratie 500 wordt de *burn-in period* genoemd [3].

Van het histogram bepalen we het 2,5de; 50ste en 97,5de percentiel, waarbij het 50ste percentiel

de schatting voor p geeft. We gebruiken dus de mediaan als schatting voor p. Het 2,5de en 97,5ste percentiel geven samen het 95%-geloofwaardigheidsinterval. Dit doen we allemaal in Python. Als laatste kunnen we weer met $p = e^{-R}$ een schatting en geloofwaardigheidsinterval voor R bepalen.

2.3 Twee leeftijdscategorieën

We beschrijven nu hoe de transmissieparameters geschat kunnen worden als we verschillende parameters definiëren voor transmissie tussen verschillende leeftijdscategorieën. Er is weinig data over jongeren beschikbaar, daarom voegen we de leeftijdscategorieën voor kinderen en jongeren samen. Dit zorgt er ook voor dat de uitdrukkingen voor de kansen niet te complex worden en dat we minder parameters hoeven te schatten. Als we erg veel schattingen moeten maken, is het gewenst om voor betrouwbare schattingen ook een dataset met veel huishoudens te hebben. We maken dus onderscheid tussen twee leeftijdscategorieën: kinderen/jongeren (in de rest van deze scriptie hebben we het over kinderen) en volwassenen.

2.3.1 Model

Het model is hetzelfde als omschreven in paragraaf 2.2.1, maar we gebruiken nu p_1, p_2, p_3 en p_4 in plaats van p. We gebruiken R_1, R_2, R_3 en R_4 in plaats van R. Deze parameters zijn als volgt gedefinieerd:

$p_1(t) :=$ kans dat een op tijd 0 besmet kind een ander kind niet besmet heeft op tijd t,
$p_2(t) :=$ kans dat een op tijd 0 besmet kind een volwassene niet besmet heeft op tijd t,
$p_3(t) :=$ kans dat een op tijd 0 besmette volwassene een kind niet besmet heeft op tijd t,
$p_4(t) :=$ kans dat een op tijd 0 besmette volwassene een andere volwassene niet besmet heeft op tijd t.

Verder geldt dat

$$p_i = e^{-R_i} \text{ voor } i = 1, 2, 3, 4$$

De stochastische variabelen X_k, X_v, N_k, N_v, I_k en I_v zijn gedefinieerd in paragraaf 2.1. Verder definiëren we

$$r_1 := \frac{1 - p_4}{1 - p_2},$$
$$r_2 := \frac{1 - p_3}{1 - p_1}.$$

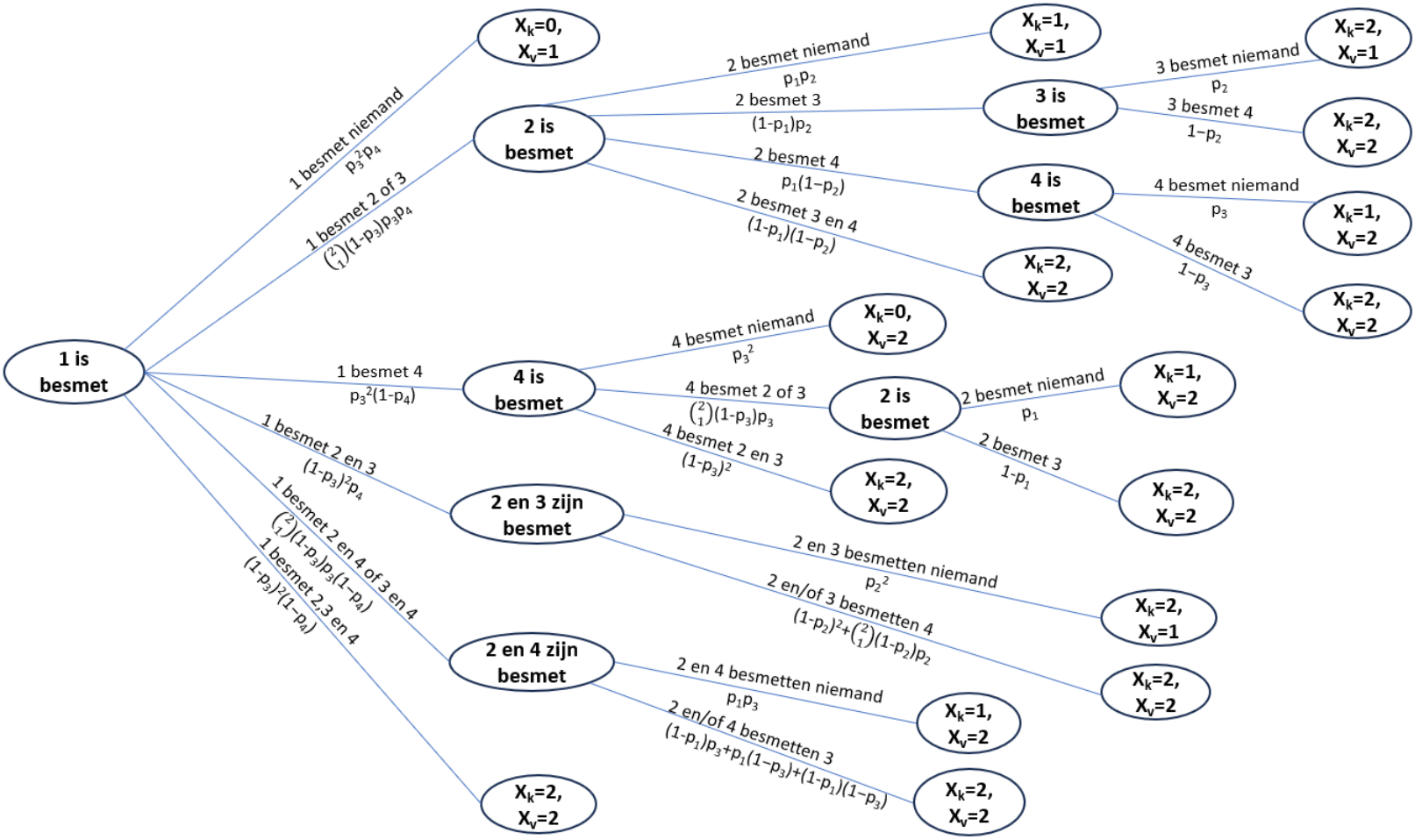
Dus r_1 geeft de verhouding tussen de besmettelijkheid van volwassenen en kinderen, bij het besmetten van volwassenen. r_2 geeft dezelfde verhouding, maar dan bij het besmetten van kinderen. Door ook een schatting voor r_1 en r_2 te bepalen, kunnen we bekijken of er een verschil is tussen de besmettelijkheid van volwassenen en kinderen.

2.3.2 Handmatig kansen bepalen

Om de waarschijnlijkheidsfunctie te bepalen hebben we de kansen $\mathbb{P}(X_k = x_k, X_v = x_v \mid N_K = n_k, N_v = n_v, I_k = i_k, I_v = i_v)$ nodig voor verschillende combinaties van x_k, x_v, n_k, n_v, i_k en i_v. Deze kansen bepalen we eerst handmatig met behulp van kansbomen op een vergelijkbare manier

als in paragraaf 2.2.2. We nummeren de personen in een huishouden weer als $1, ..., n+i$, waarbij $n = n_k + n_v$ en $i = i_k + i_v$. We nemen aan dat de indexgevallen de laagste nummers hebben. Verder nemen we aan dat als er meerdere indexgevallen zijn, kinderen de laagste nummers hebben. Als er meerdere personen geen indexgeval zijn, nemen we voor deze personen ook aan dat kinderen de laagste nummers hebben.

We bepalen als voorbeeld $\mathbb{P}(X_k = x_k, X_v = x_v \mid N_K = 2, N_v = 1, I_k = 0, I_v = 1)$. Hiervoor maken we een kansboom waarbij $N_K = 2$, $N_v = 1$,$I_k = 0$ en $I_v = 1$. We nemen dus aan dat persoon 1 een volwassene en indexgeval is, personen 2 en 3 zijn geen indexgevallen en kinderen en persoon 4 is geen indexgeval en volwassene. De kansboom is te zien in Figuur 3.



Figuur 3: Kansboom voor $N_K = 2$, $N_v = 1$,$I_k = 0$ en $I_v = 1$. De verschillende mogelijkheden staan boven de takken en de bijbehorende kansen eronder, de uitkomsten staan in de ovalen.

De kansboom is vergelijkbaar met Figuur 2. Het verschil is dat we nu meer takken krijgen, doordat we een andere uitdrukking voor de kans krijgen als een volwassene in plaats van een kind geïnfecteerd wordt of andersom. Boven een tak staan de mogelijkheden voor de personen die besmet raken. Als we voor elke mogelijkheid de som van de nummers van de personen nemen, kiezen we de mogelijkheid met de laagste som om mee verder te werken (deze mogelijkheid komt dus als uitkomst in een ovaal). De voorwaarden voor wanneer er rechts van een ovaal geen nieuwe takken meer komen zijn hetzelfde als bij de eerdere kansbomen. De kans op een combinatie van X_k en X_v

wordt weer bepaald door alle routes te bekijken voor hoe we in een ovaal met deze waarde van X_k en X_v kunnen uitkomen. Voor elke route bepalen we de kans, door de kansen bij de bijbehorende takken te vermenigvuldigen, deze kansen tellen we op.

Er ontstaat geen probleem bij het bepalen van kansen waarbij $I = I_k + I_v > 1$. Als we bijvoorbeeld $N_K = 1$, $N_v = 0$,$I_k = 1$ en $I_v = 1$ hebben, krijgen we als kansboom een deel van de kansboom uit Figuur 3, namelijk het deel bestaande uit de ovaal '2 en 4 zijn besmet' en de takken en ovalen die vanaf daar naar rechts gaan. Voor andere waarden van $N_K = 1$, $N_v = 0$,$I_k = 1$ en $I_v = 1$ met $I = I_k + I_v > 1$ kunnen we dus ook kansbomen maken.

Voor huishoudens met huishoudgrootte 5 of 6, werkt deze manier niet goed. De kansbomen voor deze huishoudgrootte werden al erg groot zonder onderscheid in leeftijd, maar met onderscheid in leeftijd hebben we nog meer mogelijkheden en worden de bomen nog een stuk groter. Daarom hebben we de kansen voor huishoudgrootte 5 en 6 niet bepaald op deze manier. 44 van de 59 huishoudens hebben een grootte kleiner dan 5, de kansen voor deze huishoudens zijn te vinden in Tabel 7. We hebben algoritmisch wel alle kansen kunnen bepalen, dit wordt uitgelegd in de volgende paragraaf.

2.3.3 Kansen algoritmisch bepalen

We bepalen nu de kansen algoritmisch. De methode die we hiervoor gebruiken lijkt op de methode uit paragraaf 2.2.3. Definieer hiervoor de volgende stochastische variabelen die respectievelijk het aantal kinderen en het aantal volwassenen in een huishouden weergeven:

$$H_k = N_k + I_k,$$

$$H_v = N_v + I_v.$$

Voor elke combinatie van H_k, H_v, I_k, I_v, X_k en X_v slaan we de bijbehorende kans op in een zesdimensionale matrix. We gebruiken nu dus de stochastische variabelen H_k en H_v in plaats van N_k en N_v. We initialiseren de matrix met nullen.

Vervolgens definiëren we de kansen uit Tabel 3 in Python.

h_k	h_v	i_k	i_v	x_k	x_v	$\mathbb{P}(X_k = x_k, X_v = x_v \mid H_k = h_k, H_v = h_v, I_k = i_k, I_v = i_v)$
1	0	1	0	1	0	1
1	0	1	0	0	0	0
0	1	0	1	0	1	1
0	1	0	1	0	0	0
0,1,...	0,1,...	0	0	0	0	1
0,1,...	0,1,...	0	0	1,2,...	1,2,...	0

Tabel 3: Kansen nodig voor iteratief kansen bepalen met twee leeftijdscategorieën

De eerste vier regels geven de kansen wanneer de totale huishoudgrootte 1 is. De een-na-laatste regel in deze tabel geeft de kansen als er geen indexgevallen en geen besmettingen zijn. De laatste regel geeft de kansen als er geen indexgevallen zijn, maar wel besmettingen.

De structuur om de overige kansen iteratief te bepalen is te zien in Algoritme 2.

Algoritme 2 Structuur geneste *for-loops* met twee leeftijdscategorieën

```
for h_k = 0 to 6 do
  for h_v = 0 to 6 do
    if h_k + h_v > 1 then
      for i_k = 0 to h_k do
        for i_v = 0 to h_v do
          if i_k + i_v > 0 then
            for x_k = i_k to h_k do
              for x_v = i_v to h_v do
                // Bepaal kans
              end for
            end for
          end if
        end for
      end for
    end if
  end for
end for
```

De variabelen h_k en h_v lopen allebei tot en met 6, 6 is de maximale huishoudgrootte. De eerste *if-statement* gebruiken we omdat de kansen voor huishoudgrootte 1 al van tevoren gedefinieerd zijn. Hetzelfde geldt voor de tweede *if-statement*: de kansen zonder indexgevallen zijn al van tevoren gedefinieerd. In de binnenste *for-loop* bepalen we de kans, hiervoor definiëren we

$$\begin{aligned}
J_k &:= \text{aantal besmettingen van kinderen in de eerste generatie,}\\
J_v &:= \text{aantal besmettingen van volwassenen in de eerste generatie,}\\
p_k &:= p_1^{i_k} p_3^{i_v},\\
p_v &:= p_2^{i_k} p_4^{i_v}.
\end{aligned}$$

Als er i_k kinderen indexgeval zijn en i_v volwassenen, dan geeft p_k de kans dat een kind niet besmet wordt door deze indexgevallen. p_v geeft de kans dat een volwassene niet wordt besmet door de indexgevallen.

Voor elke mogelijke combinatie j_k en j_v die J_k en J_v kunnen aannemen bepalen we de kans, deze kansen tellen we op. We krijgen:

$$\begin{aligned}
&\mathbb{P}(X_k = x_k, X_v = x_v \mid H_k = h_k, H_v = h_v, I_k = i_k, I_v = i_v) =\\
&\sum_{j_k=0}^{h_k-i_k} \sum_{j_v=0}^{h_v-i_v} (1-p_k)^{j_k} {p_k}^{h_k-i_k-j_k} \binom{h_k-i_k}{j_k} (1-p_v)^{j_v} {p_v}^{h_v-i_v-j_v} \binom{h_v-i_v}{j_v}.\\
&\mathbb{P}(X_k = x_k - i_k, X_v = x_v - i_v \mid H_k = h_k - i_k, H_v = h_v - i_v, I_k = j_k, I_v = j_v)
\end{aligned}$$

Hierbij geeft de term $(1-p_k)^{j_k} {p_k}^{h_k-i_k-j_k} \binom{h_k-i_k}{j_k}$ de kans dat van de $h_k - i_k$ vatbare kinderen, j_k kinderen wel besmet raken en $h_k - i_k - j_k$ kinderen niet. De term $(1-p_v)^{j_v} {p_v}^{h_v-i_v-j_v} \binom{h_v-i_v}{j_v}$ geeft dezelfde kans, maar voor volwassenen in plaats van kinderen. Samen geven deze termen de kans op j_k besmetten van kinderen en j_v besmettingen van volwassenen in de eerste generatie. Als laatste geeft de term $\mathbb{P}(X_k = x_k - i_k, X_v = x_v - i_v \mid H_k = h_k - i_k, H_v = h_v - i_v, I_k = j_k, I_v = j_v)$ de kans voor de tweede generatie, deze kans wordt weer iteratief bepaald.

2.3.4 Waarschijnlijkheidsfunctie

We definiëren $\tilde{m}_r$ als het aantal keer dat een combinatie van x_k, x_v, n_k, n_v, i_k en i_v voorkomt in de data. We maken één keer schattingen van de transmissieparameters waarbij we alleen de huishoudens tot en met grootte 4 meenemen en één keer schattingen waarbij we alle huishoudens meenemen. Op deze manier kunnen we later bekijken of de aanname van een *density-dependent* model nog steeds realistisch is voor grote huishoudens. Er zijn 22 verschillende combinaties voor huishoudgroottes tot en met 4. Als we alle huishoudgroottes bekijken, zijn er 35 verschillende combinaties. Als we alleen de huishoudgroottes tot en met vier bekijken, krijgen we dus op dezelfde manier als in paragraaf 2.2.4 de volgende waarschijnlijkheidsfunctie:

$$L_1(p) = \prod_{r=1}^{22} (\mathbb{P}(X_k = x_{k,r}, X_v = x_{v,r} \mid N_k = n_{k,r}, N_v = n_{v,r}, I_k = i_{k,r}, I_v = i_{v,r}))^{\tilde{m}_r}.$$

waarbij $p^T = (p_1, p_2, p_3, p_4)$. Als we alle huishoudgroottes bekijken, krijgen we de volgende waarschijnlijkheidsfunctie:

$$L_2(p) = \prod_{r=1}^{35} (\mathbb{P}(X_k = x_{k,r}, X_v = x_{v,r} \mid N_k = n_{k,r}, N_v = n_{v,r}, I_k = i_{k,r}, I_v = i_{v,r}))^{\tilde{m}_r}.$$

We gebruiken in het vervolg de algemene notatie $L(p)$ in plaats van $L_1(p)$ of $L_2(p)$. We bepalen bij dit model geen schatting met behulp van de afgeleiden van de waarschijnlijkheidsfunctie, doordat p vierdimensionaal is werkt dit niet goed.

2.3.5 Monte Carlo Markov Chain

We moeten een paar aanpassingen doen in vergelijking met het MCMC-algoritme uitgelegd in paragraaf 2.2.5. omdat p vierdimensionaal is. $f_P(p)$ is weer de kansdichtheidsfunctie van de prior-verdeling. Maar voor de prior-verdeling nemen we nu het product van vier uniforme verdelingen: namelijk de uniforme verdeling van 0 tot 1 voor p_1, voor p_2, voor p_3 en voor p_4. Net als in paragraaf 2.2.5 hebben we dat $g(p) := f_{P|X,N,I}(p|x,n,i)$ en $f_{P|X,N,I}(p|x,n,i) \propto L(p) \cdot f_P(p)$ waarbij $P = (P_1, P_2, P_3, P_4)^T$; $X = X_k, X_V$; $N = N_K, N_v$ en $I = I_K, I_v$.

We voeren de volgende stappen uit voor het MCMC-algoritme:

1. We kiezen de startwaarde $p_{\text{oud}} = (0.5, 0.5, 0.5, 0.5)^T$.

2. Kies een van de getallen 1, 2, 3 of 4. De kans op elk van deze getallen is 0,25. Geef de variabele y de waarde van het gekozen getal.

3. Voor de nieuwe waarde nemen we $(p_{\text{nieuw}})_z = (p_{\text{oud}})_z$ voor $z \neq y$ en $(p_{\text{nieuw}})_y = (p_{\text{oud}})_y + a$ waarbij a normaal verdeeld is met gemiddelde 0 en variantie 0.48. Als $(p_{\text{nieuw}})_y > 1$, dan nemen we $2 - (p_{\text{nieuw}})_y$ in plaats van $(p_{\text{nieuw}})_y$. Als $(p_{\text{nieuw}})_y < 0$, dan nemen we $-(p_{\text{nieuw}})_y$ in plaats van $(p_{\text{nieuw}})_y$. Dit doen we net zolang totdat $(p_{\text{nieuw}})_y \in [0, 1]$.

4. Vanaf het bepalen of p_{nieuw} geaccepteerd wordt, werkt het algoritme hetzelfde als in paragraaf 2.2.5.

Doordat we een variantie van 0.48 voor a hebben gekozen, is de kans op afwijzing ongeveer twee derde.

Bij het MCMC-algoritme met meer parameters, duurt het wat langer voordat convergentie is

bereikt. We veranderen per iteratie maar één van de parameters p_1, p_2, p_3 en p_4. Daarom voeren we in totaal meer iteraties uit, namelijk 1000000 iteraties. Om te bepalen welke periode we kunnen kiezen voor de *burn-in period*, maken we als eerste vier *traceplots* voor p_1, p_2, p_3 en p_4. Verder maken we twee plots van de waarschijnlijkheidsfunctie: één voor de eerste 10000 iteraties en één met alle iteraties. Uit deze *traceplots* en plots van de waarschijnlijkheidsfunctie volgt dat er altijd convergentie heeft plaatsgevonden voor iteratie 2000, de waarschijnlijkheidsfunctie stijgt gemiddeld niet meer vanaf iteratie 2000. De periode tot iteratie 2000 kiezen we dus als *burn-in period*.

We maken zes histogrammen voor p_1, p_2, p_3, p_4, r_1 en r_2. Bij deze histogrammen bepalen we het 2,5de; 50ste en 97,5de percentiel. Deze geven de schattingen en bijbehorende 95%-geloofwaardigheidsintervallen voor p_1, p_2, p_3, p_4, r_1 en r_2. Verder maken we twee spreidingsdiagrammen: één met p_1 op de horizontale as en p_3 op de verticale as en één met p_2 op de horizontale as en p_4 op de verticale as. Voor deze twee grafieken worden ook de bijbehorende correlatiecoëfficiënten in Python berekend. Door te gebruiken dat $p_i = e^{-R_i}$ kunnen we schattingen en geloofwaardigheidsintervallen voor R_i bepalen.

3 Resultaten

3.1 Geen leeftijdsonderscheid

3.1.1 Waarschijnlijkheidsfunctie met handmatig bepaalde kansen

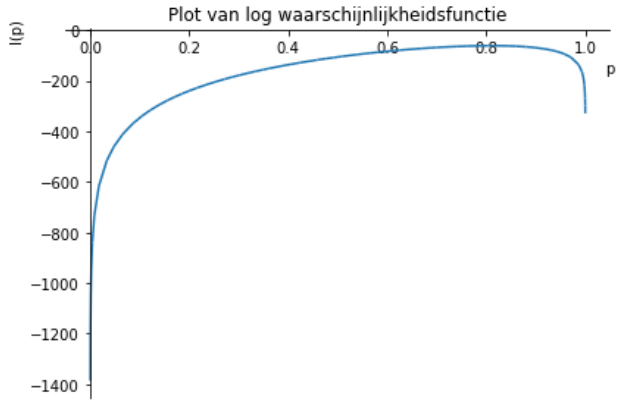
We hebben als eerste p geschat door de afgeleide van de log waarschijnlijkheidsfunctie gelijk te stellen aan 0. We krijgen als schatting $\hat{p} = 0.81$. Het bijbehorende 95%-betrouwbaarheidsinterval is gegeven door $[p_{\text{onder}}, p_{\text{boven}}] = [0.77, 0.85]$. Dus de schatting voor R is

$$\hat{R} = -\log(\hat{p}) = 0.21.$$

Verder hebben we het volgende 95%-betrouwbaarheidsinterval voor R:

$$[R_{\text{onder}}, R_{\text{boven}}] = [-\log(p_{\text{boven}}), -\log(p_{\text{onder}})] = [0.16, 0.26].$$

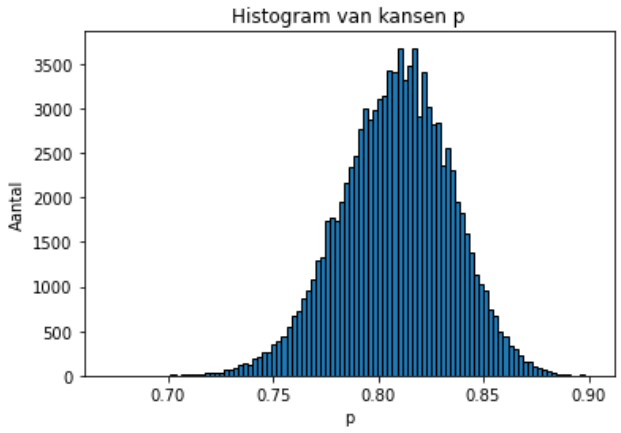
Een plot van de log waarschijnlijkheidsfunctie is te zien in Figuur 4, hier is te zien dat het maximum inderdaad rond $\hat{p} = 0.81$ ligt.



Figuur 4: Plot van log waarschijnlijkheidsfunctie.

3.1.2 Monte Carlo Markov Chain met algoritmisch bepaalde kansen

We hebben vervolgens $\hat{p}$ geschat met behulp van het MCMC-algoritme. Het histogram is te zien in Figuur 5.



Figuur 5: Histogram van kansen p.

De schatting voor p is $\hat{p} = 0.81$. Het bijbehorende 95%-geloofwaardigheidsinterval is gegeven door $[p_{\text{onder}}, p_{\text{boven}}] = [0.75, 0.86]$. Hieruit volgt de schatting voor R: $\hat{R} = 0.21$. Daarnaast krijgen we hieruit het volgende 95%-geloofwaardigheidsinterval voor R: $[R_{\text{onder}}, R_{\text{boven}}] = [0.15, 0.28]$.

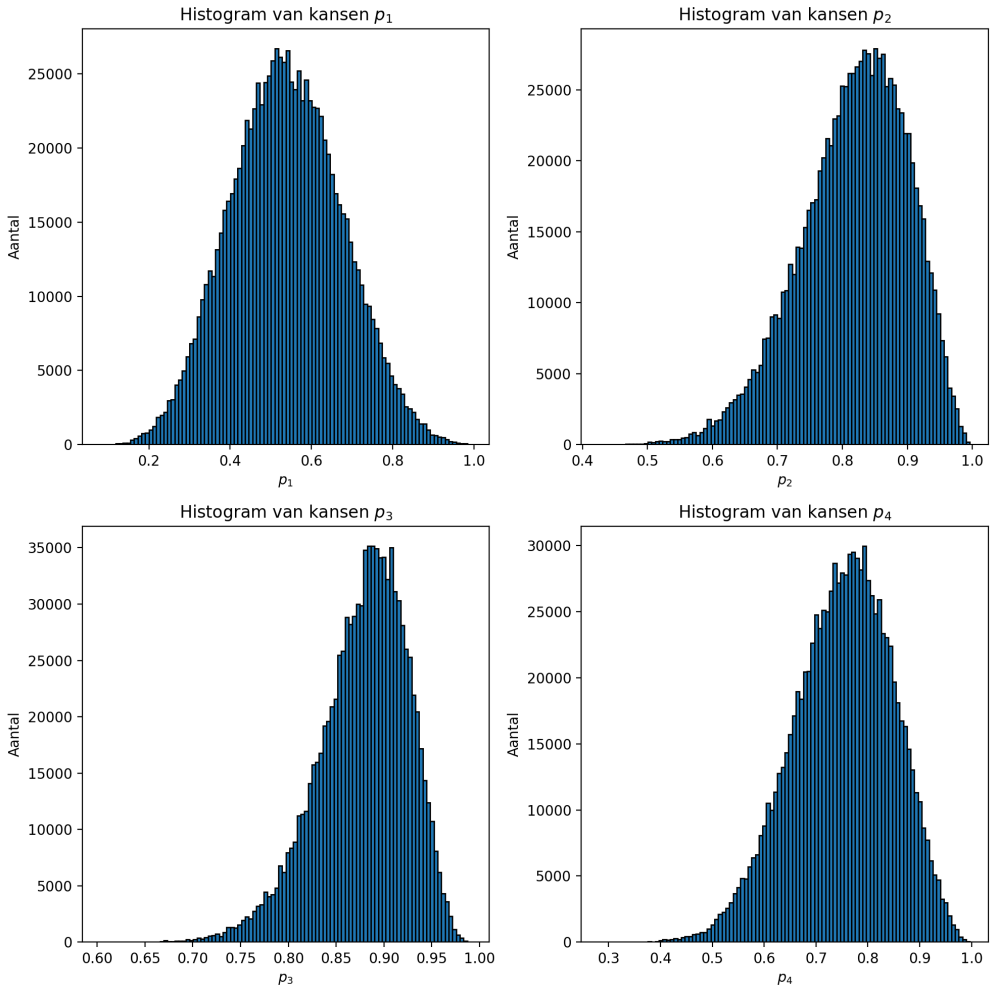
3.1.3 Vergelijking van de twee methoden

De schattingen voor p en R zijn bij beide methoden hetzelfde. Het geloofwaardigheidsinterval is iets groter bij het gebruik van het MCMC-algoritme en de looptijd van de MCMC-code is langer.

3.2 Twee leeftijdscategorieën

3.2.1 Monte Carlo Markov Chain met huishoudgrootte tot en met vier

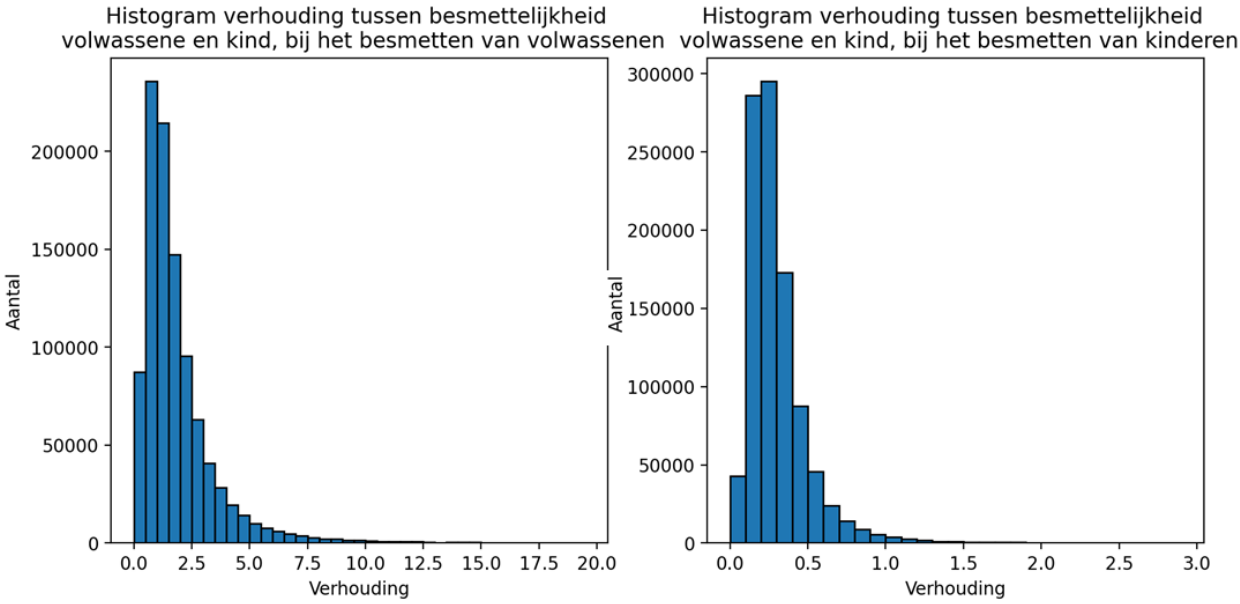
We hebben eerst p_1, p_2, p_3 en p_4 geschat met behulp van het MCMC-algoritme, waarbij we alleen huishoudgroottes tot en met vier hebben meegenomen. In dit geval is in de plots te zien dat de maximale waarde van de waarschijnlijkheidsfunctie die tijdens de iteraties wordt aangenomen rond de $1.4 \cdot 10^{-18}$ ligt. Bij ongeveer 66 procent van de iteraties wordt de nieuwe waarde niet geaccepteerd. De histogrammen zijn te zien in Figuur 6.



Figuur 6: Histogrammen van kansen p_1, p_2, p_3 en p_4.

We krijgen de schattingen $\hat{p_1} = 0.54$, $\hat{p_2} = 0.83$, $\hat{p_3} = 0.88$ en $\hat{p_4} = 0.76$. De bijbehorende 95%-geloofwaardigheidsintervallen zijn gegeven door $[p_{1,\text{onder}}, p_{1,\text{boven}}] = [0.28, 0.81]$, $[p_{2,\text{onder}}, p_{2,\text{boven}}] = [0.64, 0.95]$, $[p_{3,\text{onder}}, p_{3,\text{boven}}] = [0.77, 0.95]$ en $[p_{4,\text{onder}}, p_{4,\text{boven}}] = [0.55, 0.92]$. Dus de schattingen voor R_1, R_2, R_3 en R_4 zijn $\hat{R_1} = 0.62$, $\hat{R_2} = 0.19$, $\hat{R_3} = 0.12$ en $\hat{R_4} = 0.28$. De geloofwaardigheidsintervallen voor R_1, R_2, R_3 en R_4 zijn gegeven door $[R_{1,\text{onder}}, R_{1,\text{boven}}] = [0.22, 1.27]$, $[R_{2,\text{onder}}, R_{2,\text{boven}}] = [0.05, 0.45]$, $[R_{3,\text{onder}}, R_{3,\text{boven}}] = [0.05, 0.26]$ en $[R_{4,\text{onder}}, R_{4,\text{boven}}] = [0.08, 0.60]$.

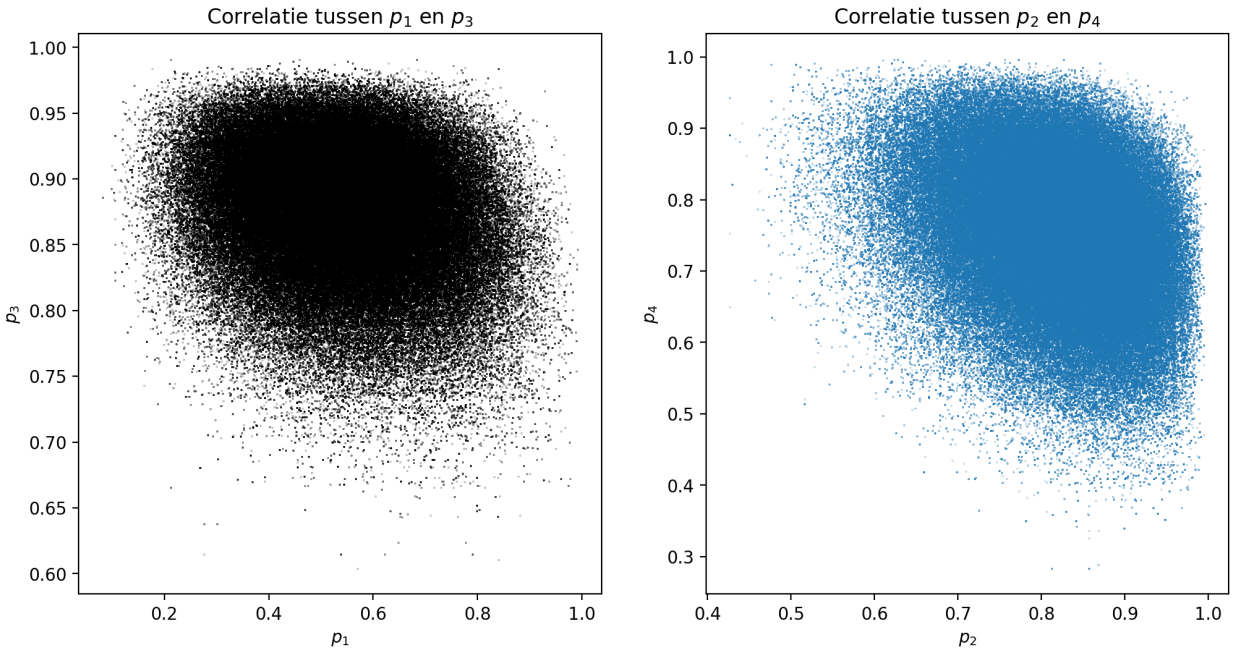
De schatting voor r_1 is gegeven door $\hat{r_1} = 1.39$ met 95%-geloofwaardigheidsinterval $[r_{1,\text{onder}}, r_{1,\text{boven}}] = [0.30, 6.74]$. Voor r_2 krijgen we $\hat{r_2} = 0.25$ met 95%-geloofwaardigheidsinterval $[r_{2,\text{onder}}, r_{2,\text{boven}}] = [0.09, 0.86]$. In Figuur 10 zijn de bijbehorende histogrammen voor r_1 en r_2 te zien.



Figuur 7: Linkerfiguur: histogram van r_1, rechterfiguur: histogram van r_2.

De grootste waarde van r_1 die voorkomt tijdens de iteraties is 105.92, voor r_2 is deze waarde 21.51. Als p_4 erg klein is en p_2 juist erg groot, kan r_1 erg groot worden. Hetzelfde geldt voor r_2 als p_3 klein is en p_1 groot. Dit komt niet vaak voor, in de histogrammen zijn de aantallen voor grote waarden van r_1 en r_2 zo klein in vergelijking met de andere waarden, dat de staven bij grote waarden met het blote oog niet te zien zijn. Daarom laten we het histogram voor r_1 lopen tot en met 20 en voor r_2 tot en met 3.

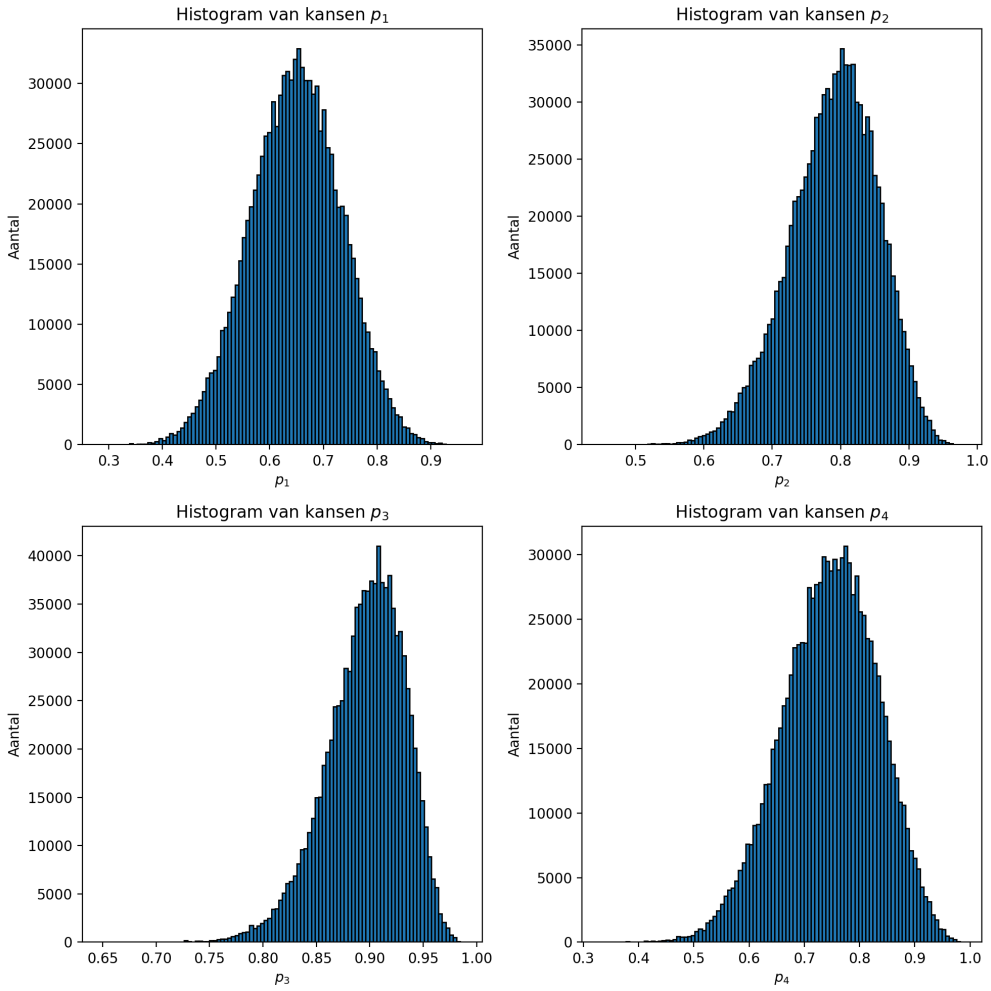
De correlatiecoëfficiënt voor de correlatie tussen p_1 en p_3 is -0.20, er is dus een negatieve correlatie. Ook tussen p_2 en p_4 is de correlatie negatief, in dit geval is de correlatiecoëfficiënt -0.32. In Figuur 8 is links een spreidingsdiagram van p_1 en p_3 te zien en rechts van p_2 en p_4, hier is ook te zien dat de correlatie in beide gevallen negatief is.



Figuur 8: Correlatie tussen p_1 en p_3 en correlatie tussen p_2 en p_4

3.2.2 Monte Carlo Markov Chain met alle huishoudens

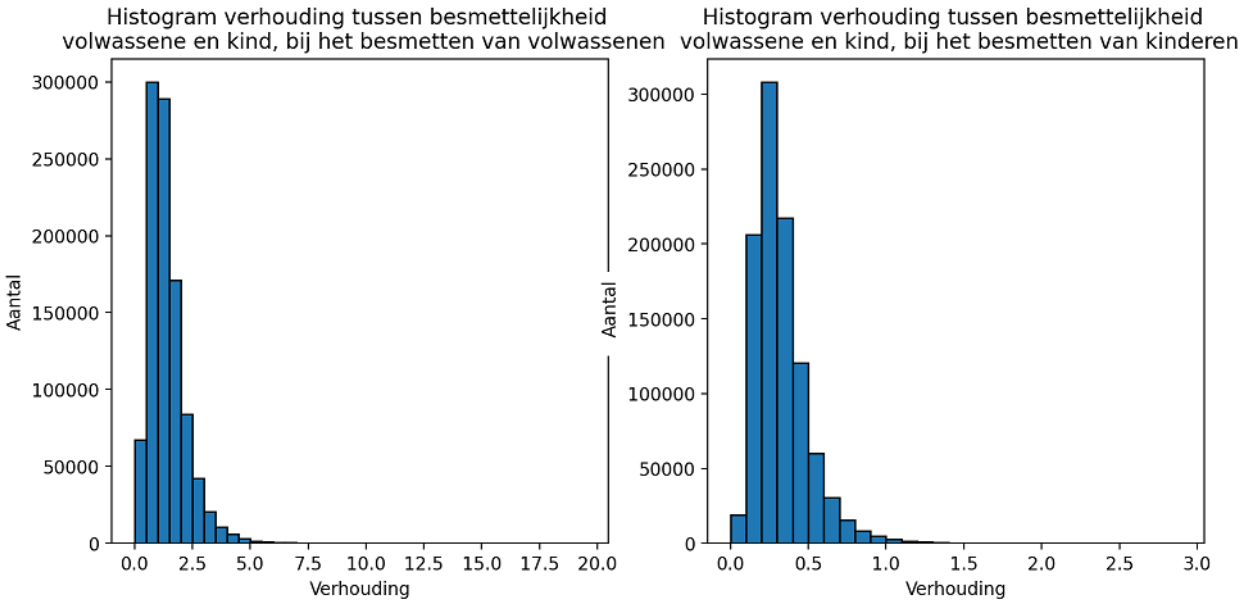
Vervolgens hebben we de parameters p_1, p_2, p_3 en p_4 geschat, waarbij we alle huishoudens hebben meegenomen. Hier ligt de maximale waarde van de waarschijnlijkheidsfunctie die tijdens de iteraties wordt aangenomen rond de $2.0 \cdot 10^{-29}$. Bij ongeveer 73 procent van de iteraties wordt de waarde niet geaccepteerd. De histogrammen zijn te zien in Figuur 9.



Figuur 9: Histogrammen van kansen p_1, p_2, p_3 en p_4.

De bijbehorende schattingen zijn $\hat{p_1} = 0.65$, $\hat{p_2} = 0.79$, $\hat{p_3} = 0.90$ en $\hat{p_4} = 0.75$. Verder zijn de 95%-geloofwaardigheidsintervallen gegeven door $[p_{1,\text{onder}}, p_{1,\text{boven}}] = [0.48, 0.81]$, $[p_{2,\text{onder}}, p_{2,\text{boven}}] = [0.65, 0.90]$, $[p_{3,\text{onder}}, p_{3,\text{boven}}] = [0.81, 0.96]$ en $[p_{4,\text{onder}}, p_{4,\text{boven}}] = [0.57, 0.90]$. Hieruit volgen de schattingen voor R_1, R_2, R_3 en R_4: $\hat{R_1} = 0.43$, $\hat{R_2} = 0.23$, $\hat{R_3} = 0.11$ en $\hat{R_4} = 0.29$. De geloofwaardigheidsintervallen voor R_1, R_2, R_3 en R_4 zijn: $[R_{1,\text{onder}}, R_{1,\text{boven}}] = [0.21, 0.73]$, $[R_{2,\text{onder}}, R_{2,\text{boven}}] = [0.10, 0.43]$, $[R_{3,\text{onder}}, R_{3,\text{boven}}] = [0.05, 0.21]$ en $[R_{4,\text{onder}}, R_{4,\text{boven}}] = [0.10, 0.57]$.

De schatting voor r_1 is gegeven door $\hat{r_1} = 1.21$ met 95%-geloofwaardigheidsinterval $[r_{1,\text{onder}}, r_{1,\text{boven}}] = [0.36, 3.47]$. Voor r_2 krijgen we $\hat{r_2} = 0.29$ met 95%-geloofwaardigheidsinterval $[r_{2,\text{onder}}, r_{2,\text{boven}}] = [0.11, 0.77]$. In Figuur 10 zijn de bijbehorende histogrammen voor r_1 en r_2 te zien.



Figuur 10: Linkerfiguur: histogram van r_1, rechterfiguur: histogram van r_2.

De grootste waarde van r_1 die voorkomt tijdens de iteraties is 18.62, voor r_2 is deze waarde 5.46, r_1 en r_2 kunnen dus in uitzonderlijke gevallen weer groot worden. We hebben weer het histogram voor r_1 laten lopen tot en met 20 en voor r_2 tot en met 3, de staven voor grotere waarden van r_1 en r_2 zijn weer zo klein dat ze niet zichtbaar zijn.

De correlatiecoëfficiënt voor de correlatie tussen p_1 en p_3 is -0.23, tussen p_2 en p_4 is deze -0.32. De correlatie is dus nog steeds in beide gevallen negatief. De spreidingsdiagrammen zien er vergelijkbaar uit als in Figuur 8.

3.2.3 Vergelijking van de twee methoden

Het eerste verschil tussen de methoden is dat bij de tweede methode (het meenemen van alle huishoudens) de maximale waarde van de waarschijnlijkheidsfunctie die tijdens de iteraties wordt aangenomen, een stuk kleiner is. Dit is logisch omdat we bij deze methode voor het product van kansen in de waarschijnlijkheidsfunctie meer kansen hebben, doordat we hier ook de huishoudens met grootte groter dan 4 hebben bekeken. In Figuur 9 is de standaardafwijking in het algemeen kleiner en de piek hoger dan in Figuur 6. Dit is ook te zien aan de geloofwaardigheidsintervallen, deze zijn smaller. De schattingen voor p_1, p_2, p_3 en p_4 (en dus voor R_1, R_2, R_3 en R_4) liggen bij het gebruik van beide methoden dicht bij elkaar. Voor p_1 is het verschil tussen de schattingen wat groter. Doordat de standaardafwijking smaller is bij het meenemen van alle huishoudens en we wel dezelfde standaardafwijking gebruiken in het algoritme om een nieuwe waarde te krijgen voor de volgende iteratie, worden bij het meenemen van alle huishoudens meer iteraties niet geaccepteerd.

De schattingen voor r_1 en r_2 zijn bij het gebruik van beide methoden niet heel nauwkeurig, wel zien we weer dat deze schattingen alsnog een stuk nauwkeuriger zijn bij de tweede methode en dat er geen groot verschil is tussen de schattingen van de eerste en tweede methode. De maximale waarden voor deze twee parameters zijn bij de tweede methode een stuk kleiner, wat overeenkomt met meer nauwkeurige schattingen.

Het verschil in de correlatiecoëfficiënten tussen beide methoden is erg klein. Bij beide methoden zijn allebei de correlatiecoëfficiënten negatief, waarbij de correlatiecoëfficiënt voor de correlatie

tussen p_2 en p_4 het meest negatief is.

4 Conclusie

4.1 Vergelijking manieren om kansen te bepalen

We hebben de kansen die nodig zijn voor de waarschijnlijkheidsfunctie op twee manieren bepaald: handmatig en algoritmisch. In het algemeen gaat de voorkeur uit naar de kansen algoritmisch bepalen. Het eerste voordeel hiervan is dat als we nieuwe data krijgen over nieuwe huishoudens of over dezelfde huishoudens in een andere tijdsperiode, we meteen het programma ook voor deze nieuwe data kunnen uitvoeren. We hoeven dan niet eerst allerlei kansen zelf te bepalen. Een ander voordeel is dat we bij twee leeftijdscategorieën de kansen voor de huishoudens met huishoudgrootte groter dan vier alleen algoritmisch en niet handmatig konden bepalen. Bij geen leeftijdsonderscheid was het in ons geval nog mogelijk om alle kansen ook handmatig te bepalen, maar als er bijvoorbeeld nieuwe data verkregen wordt met grotere huishoudens, wordt dit ook te veel werk. Voor het inzicht was het wel nuttig om de kansen ook handmatig te bepalen.

4.2 Geen leeftijdsonderscheid

Het programma in Python is bij het gebruik van de afgeleide van de waarschijnlijkheidsfunctie makkelijker en werkt sneller. Het voordeel van het gebruiken van het MCMC-algoritme is hier bijvoorbeeld dat dit algoritme hierdoor makkelijk gegeneraliseerd kon worden naar het MCMC-algoritme met meerdere parameters. De schatting voor de kans dat een persoon een ander persoon besmet is 0.19.

4.3 Twee leeftijdscategorieën

Bij het meenemen van alle huishoudens krijgen we als schatting voor de kans dat een kind een kind besmet 0.35. Voor de kans dat een kind een volwassene besmet krijgen we 0.21, voor de kans dat een volwassene een kind besmet 0.10 en voor de kans dat een volwassene een volwassene besmet 0.25. De vier verschillende situaties waarbij een kind of volwassene een ander kind of volwassene niet besmet, komen niet allemaal even vaak voor. Toch krijgen we als we het gemiddelde van deze vier kansen bepalen 0.23, wat in de buurt zit van de kans 0.19 die we vonden in de vorige paragraaf. De kans dat een kind een kind besmet is het grootst, de kans dat een volwassene een kind besmet is het kleinst. Er waren geen kinderen gevaccineerd tijdens de studie, dit zou een reden kunnen zijn dat kinderen de grootste kans hebben om elkaar te besmetten. Een andere reden zou kunnen zijn dat ze meer contact met elkaar hebben of dat ze minder letten op hygiëne.

Als er bijvoorbeeld één kind en één volwassene indexgevallen zijn en er is één andere persoon besmet in het huishouden, is het onzeker door wie de andere persoon besmet is geraakt. Deze onzekerheid zorgt voor spreiding in de schattingen. Huishoudens waar niet meerdere mogelijkheden zijn, zorgen voor het patroon in de waarden die de parameters aannemen tijdens de iteraties. Er waren geen huishoudens in de data met alleen kinderen of alleen volwassenen. Hierdoor is de spreiding het grootst voor de kans dat een kind een kind niet besmet en de kans dat een volwassene een volwassene niet besmet.

De schattingen voor p_1, p_2, p_3 en p_4, r_1 en r_2 veranderen niet veel als we ook de grotere huishoudens bekijken, behalve voor de kans dat een kind een kind niet besmet. Dit kan deels verklaard worden doordat we bij het bekijken van alleen de kleinere huishoudens een stuk minder data hebben over het besmet van kinderen door kinderen. De grotere huishoudens hebben vaak drie of vier kinderen en de kleinere huishoudens vaak maar één of twee. Door grotere huishoudens aan de data toe te voegen krijgen we dus in vergelijking het meeste informatie over het besmetten van kinderen door kinderen. De schatting voor de kans dat een kind een kind besmet wordt kleiner

als we alle huishoudens meenemen. De aanname van een *density-dependent* model, betekent dat we aannemen dat een paar kinderen in een groot huishouden evenveel contact met elkaar hebben als in een kleiner huishouden. In werkelijkheid zou het kunnen dat een paar kinderen minder contact met elkaar hebben in een groter huishouden en er dus een kleinere kans is dat ze elkaar besmetten. Voor het contact tussen kinderen en ouders of tussen ouders onderling, lijkt het realistischer dat de contactintensiteit gelijk blijft bij grotere huishoudens, dit correspondeert ook met het resultaat dat de schattingen voor de andere transmissieparameters minder veranderen bij het meenemen van wel of geen grote huishoudens. Een *density-dependent* model kan dus minder realistisch zijn bij grote huishoudens.

De correlatie tussen de kans dat een kind een kind niet besmet en de kans dat een volwassene een kind niet besmet is negatief. De verklaring hiervoor is dat als er een grote kans is dat een willekeurig kind besmet wordt door een kind, dan is de kans klein dat dit willekeurige kind door een volwassene is besmet. De correlatie tussen de kans dat een kind een volwassene niet besmet en de kans dat een volwassene een volwassene niet besmet is ook negatief. De verklaring is hetzelfde, maar dan kijken we door wie een volwassene besmet is.

Voor de verhouding tussen de besmettelijkheid van volwassenen en kinderen, bij het besmetten van kinderen, hebben we gezien dat de waarde 1 niet in het 95%-geloofwaardigheidsinterval ligt, de bovengrens voor het geloofwaardigheidsinterval is 0.77 bij het meenemen van alle huishoudens. Bij het besmetten van kinderen is de verhouding tussen de besmettelijkheid van volwassenen en de besmettelijkheid van kinderen dus klein. Kinderen zijn in dit geval significant besmettelijker. Bij het besmetten van volwassenen kunnen we geen conclusie trekken over de besmettelijkheid van kinderen in vergelijking met de besmettelijkheid van volwassenen, hier ligt de waarde 1 wel in het 95%-geloofwaardigheidsinterval.

5 Discussie

Er zijn een aantal opmerkingen over het huidige onderzoek:

- We hebben aangenomen dat als persoon 1 iemand anders besmet, dit altijd snel gebeurt na de besmetting van persoon 1. Deze aanname was noodzakelijk, omdat we *final size data* wilden gebruiken. Dit hoeft echter niet altijd realistisch te zijn.
- Een *density-dependent* model hoeft ook niet altijd realistisch te zijn, de contactintensiteit van elk paar van individuen zou kleiner kunnen zijn bij een groter huishouden.
- We hebben aangenomen dat als er meerdere indexgevallen in een huishouden zijn, deze tegelijk aan het begin van de uitbraak besmet zijn geraakt. Dit hoeft niet altijd te kloppen, maar doordat we *final size data* gebruiken moesten we dit aannemen.
- We hebben bij het model zonder leeftijdsonderscheid en het bepalen van een schatting met de afgeleide van de waarschijnlijkheidsfunctie een *Wald interval estimator* gebruikt, deze is vooral geschikt als er veel huishoudens zouden zijn omdat het een asymptotische benadering is. Hierdoor kan het betrouwbaarheidsinterval wat minder nauwkeurig zijn.
- We hebben bij het model met leeftijdsonderscheid en het bepalen van een schatting met het MCMC-algoritme gebruik gemaakt van een uniforme prior-verdeling. Hier is verder niet uitgebreid naar gekeken.

Verder zijn er een aantal mogelijke ideeën voor vervolgonderzoek, waarbij een deel voortkomt uit de bovenstaande opmerkingen:

- We hebben nu de leeftijdscategorieën voor kinderen en jongeren bij elkaar gevoegd. Een idee voor vervolgonderzoek is om deze niet samen te voegen. Dan krijgen we negen verschillende parameters voor de kansen in plaats van vier, de besmettelijkheid van kinderen en jongeren kan dan bijvoorbeeld vergeleken worden.
- In vervolgonderzoek zou er gekeken kunnen worden naar het gebruik van een *frequency-dependent* model in plaats van *density-dependent* model, dit betekent dat de contactintensiteit van elk paar van individuen in een huishouden kleiner wordt als het huishouden groter wordt. We hebben gezien dat de aanname van een *density-dependent* model misschien niet zo realistisch is voor grote huishoudens. Een *frequency-dependent* model lijkt alsnog iets minder goed te corresponderen met de data, maar dit is niet zeker [6]. Resultaten van beide modellen kunnen met elkaar vergeleken worden.
- Er is ook een dataset met data uit een latere tijdsperiode. Door dezelfde analyses met deze dataset uit te voeren, kunnen resultaten tussen de twee tijdsperioden vergeleken worden. De immuniteit van de populatie en de variant van het virus verschillen tussen deze twee tijdsperioden. Bij deze latere tijdsperiode is een deel van de personen bij de startdatum gevaccineerd. Door in plaats van alleen onderscheid in leeftijdscategorieën, ook onderscheid in wel of niet gevaccineerd te maken, kan het effect van vaccinatie geschat worden.
- Het kan interessant zijn om een algemene theoretische afleiding te geven voor de transmissieparameters van de *final size data*.
- We hebben nu alleen *final size data* gebruikt. De gerapporteerde besmettingstijden zijn niet openbaar beschikbaar, maar deze kunnen wel interessant zijn om te bestuderen om zo een

volledigere beeld van de verspreiding te krijgen. Dit kan bijvoorbeeld extra informatie over de kansen, de infectietijd en het bestaan van superverspreiders geven. Verder hoeven we dan niet aan te nemen dat als persoon 1 iemand anders besmet, dit altijd snel gebeurt na de besmetting van persoon 1. We hadden ook aangenomen dat als er meerdere indexgevallen zijn, deze tegelijk besmet zijn geraakt aan het begin van de uitbraak. Deze aanname is ook niet meer nodig als we gerapporteerde besmettingstijden gebruiken.

- Als laatste kan in vervolgonderzoek gekeken worden naar het gebruik van de programmeertaal Stan. Hier is het MCMC-algoritme al automatisch in geprogrammeerd, waardoor het efficiënter werkt. Dit kan bijvoorbeeld handig zijn als we ook meer parameters willen introduceren, zoals hierboven genoemd. Het introduceren van meer parameters zorgt voor een langere tijd om het programma uit te voeren, maar door dit in Stan te programmeren, kan een te lange tijd waarschijnlijk worden voorkomen. Ook zou het gebruiken van Stan voor meerdere parameters betere schattingen kunnen geven, omdat Stan ook gebruik maakt van de gradiënt van de waarschijnlijkheidsfunctie.

Referenties

[1] Benny Borremans e.a. "The shape of the contact–density function matters when modelling parasite transmission in fluctuating populations". In: *Royal Society Open Science* 4.11 (nov 2017), p. 171308. DOI: 10.1098/rsos.171308. URL: https://doi.org/10.1098/rsos.171308.

[2] Marieke L A De Hoog e.a. "Longitudinal Household Assessment of respiratory illness in children and parents during the COVID-19 pandemic". In: *JAMA network open* 5.10 (okt 2022), e2237522. DOI: 10.1001/jamanetworkopen.2022.37522. URL: https://doi.org/10.1001/jamanetworkopen.2022.37522.

[3] *Markov Chain Monte Carlo*. Mrt 2023. URL: https://www.publichealth.columbia.edu/research/population-health-methods/markov-chain-monte-carlo.

[4] John A. Rice. *Mathematical statistics and data analysis*. Brooks/Cole, jan 2007.

[5] Åke Svensson. "A note on generation times in epidemic models". In: *Mathematical Biosciences* 208.1 (jul 2007), p. 300–311. DOI: 10.1016/j.mbs.2006.10.010. URL: https://doi.org/10.1016/j.mbs.2006.10.010.

[6] Michiel Van Boven e.a. "Estimation of introduction and transmission rates of SARS-COV-2 in a prospective household study". In: *medRxiv (Cold Spring Harbor Laboratory)* (jun 2023). DOI: 10.1101/2023.06.02.23290879. URL: https://doi.org/10.1101/2023.06.02.23290879.

[7] Don Van Ravenzwaaij, Pete Cassey en Scott Brown. "A simple introduction to Markov chain Monte–Carlo Sampling". In: *Psychonomic Bulletin & Review* 25.1 (mrt 2016), p. 143–154. DOI: 10.3758/s13423-016-1015-8. URL: https://doi.org/10.3758/s13423-016-1015-8.

6 Bijlage

6.1 Data

Tabel 4: *Final size data.* j geeft het aantal infecties in een huishouden van personen die geen indexgeval zijn, a geeft het aantal indexgevallen, n geeft het aantal onbesmette personen bij de start van de uitbraak. 1 staat voor de leeftijdscategorie kind, 2 voor jongvolwassene en 3 voor volwassene.

Nummer van het huishouden	j_1	j_2	j_3	a_1	a_2	a_3	n_1	n_2	n_3
1	0	0	0	0	0	1	2	2	1
3	0	0	0	0	0	1	3	1	1
5	0	0	0	0	1	1	0	0	1
7	0	0	0	0	1	0	0	1	2
13	0	0	0	0	0	1	0	1	1
24	0	0	0	0	1	0	1	0	2
27	0	2	0	0	0	2	1	2	0
37	0	0	1	0	0	1	0	3	1
49	0	1	0	0	1	2	0	2	0
58	0	0	0	0	0	1	1	1	1
61	0	0	0	0	0	2	0	2	0
62	0	0	0	0	0	1	0	1	0
63	0	0	0	0	1	0	1	1	1
80	0	0	2	0	1	0	0	0	2
85	0	0	0	0	1	0	1	1	2
86	0	0	0	1	0	0	0	3	2
92	0	0	0	0	1	1	0	2	1
105	2	0	0	1	0	0	2	0	2
116	0	0	0	0	0	2	1	0	0
117	0	0	0	0	0	1	1	0	2
123	2	0	1	0	0	1	2	0	1
125	0	0	1	0	0	1	2	0	1
127	0	0	0	1	0	0	0	0	2
136	1	0	0	1	0	0	1	0	2
138	0	0	0	1	0	0	1	0	2
139	0	0	0	0	0	1	2	0	1
145	0	0	1	2	0	1	0	0	1
152	0	0	1	0	0	1	2	0	1
157	0	0	0	0	0	1	2	0	1
159	0	0	0	0	0	1	2	0	1
163	2	0	1	0	0	1	2	0	1
171	0	0	1	1	0	1	0	0	1
173	0	0	0	0	0	1	1	0	1
174	2	0	1	0	0	1	2	0	1
176	1	0	0	0	0	2	1	0	0
184	0	0	0	3	0	1	0	0	1
191	0	0	0	1	0	0	0	0	2
196	0	0	0	1	0	0	1	0	2

Tabel 5: Vervolg van Tabel 4.

Nummer van het huishouden	j_1	j_2	j_3	a_1	a_2	a_3	n_1	n_2	n_3
200	0	0	1	1	0	1	1	0	1
203	2	0	0	0	0	2	2	0	0
210	2	0	2	1	0	0	2	0	2
217	1	0	0	1	0	2	1	0	0
224	1	0	2	1	0	0	2	0	2
228	0	0	0	0	0	1	2	0	1
231	0	0	0	0	0	1	1	0	1
245	0	0	0	1	0	0	0	0	3
253	0	0	1	1	0	0	2	0	2
254	0	0	0	0	0	1	2	0	1
257	0	0	0	0	0	1	1	0	1
273	0	0	0	0	0	1	1	0	1
280	1	0	2	1	0	0	1	0	2
286	0	0	0	0	0	1	3	0	1
287	0	0	0	0	0	1	2	0	1
289	0	0	0	0	0	1	1	0	0
293	0	0	0	0	0	1	2	0	1
298	0	0	0	0	0	1	2	0	1
299	2	0	2	1	0	0	2	0	2
300	0	0	0	0	0	1	2	0	1
304	0	0	0	0	0	1	1	0	1

6.2 Kansen waarschijnlijkheidsfunctie

Tabel 6: Kansen nodig voor waarschijnlijkheidsfunctie.

r	m_r	x_r	n_r	i_r	$\mathbb{P}(X = x_r \mid N = n_r, I = i_r)$
1	2	1	1	1	p
2	8	1	2	1	p^2
3	1	3	2	1	$2p(1-p)^2 + (1-p)^2$
4	17	1	3	1	p^3
5	3	2	3	1	$3(1-p)p^4$
6	4	4	3	1	$(1-p)^3 + 6(1-p)^3p^3 + 3(1-p)^3p^2 + 3(1-p)^4p + 6(1-p)^3p^2$
7	2	1	4	1	p^4
8	2	2	4	1	$4(1-p)p^6$
9	1	3	4	1	$12(1-p)^2p^7 + 6(1-p)^2p^6$
10	1	4	4	1	$24(1-p)^3p^7 + 12(1-p)^3p^6 + 24(1-p)^3p^6 + 12(1-p)^4p^5 + 4(1-p)^3p^4$
11	2	5	4	1	$(1-p)^4 + 16(1-p)^4p^3 + 24(1-p)^4p^6 + 36(1-p)^4p^5 +$ $12(1-p)^5p^4 + 6(1-p)^6p^2 + 24(1-p)^5p^3 + 24(1-p)^4p^4 + 12(1-p)^5p^2 +$ $4(1-p)^6p + 24(1-p)^4p^5 + 12(1-p)^5p^4$
12	3	1	5	1	p^5
13	2	2	1	2	p^2
14	2	3	1	2	$2p(1-p) + (1-p)^2$
15	1	2	2	2	p^4
16	2	3	2	2	$4(1-p)p^4 + 2(1-p)^2p^3$
17	1	4	2	2	$(1-p)^4 + 4(1-p)^3p + 4(1-p)^2p^2 + 4(1-p)^2p^3 + 2(1-p)^3p^2$
18	1	2	3	2	p^6
19	1	4	3	2	$12(1-p)^2p^6 + 12(1-p)^3p^5 + 3(1-p)^4p^4 + 12(1-p)^2p^7 + 6(1-p)^3p^6$
20	2	4	1	3	$3(1-p)p^2 + 3(1-p)^2p + (1-p)^3$
21	1	4	2	3	$2(1-p)^3p^4 + 6(1-p)^2p^5 + 6(1-p)p^6$
22	1	4	1	4	p^4

6.3 Kansen waarschijnlijkheidsfunctie met onderscheid in leeftijd

Tabel 7: Kansen nodig voor waarschijnlijkheidsfunctie met onderscheid in leeftijd. De rechterkolom geeft de kans $\mathbb{P}(X_k = x_{k,r}, X_v = x_{v,r} \mid N_k = n_{k,r}, N_v = n_{v,r}, I_k = i_{k,r}, I_v = i_{v,r})$.

r	$\tilde{m}_r$	$x_{k,r}$	$x_{v,r}$	$n_{k,r}$	$n_{v,r}$	$i_{k,r}$	$i_{v,r}$	Kans op situatie
1	2	0	1	1	0	0	1	p_3
2	6	0	1	1	1	0	1	p_3p_4
3	1	0	1	1	2	0	1	$p_3p_4^2$
4	10	0	1	2	1	0	1	$p_3^2p_4$
5	2	0	2	2	1	0	1	$p_3^4(1-p_4)$
6	3	2	2	2	1	0	1	$(1-p_3)^2(1-p_4) + 2p_3^2(1-p_1)(1-p_3)(1-p_4)+$ $2p_1(1-p_3)^2p_3(1-p_4) + 2(1-p_1)(1-p_3)^2p_3(1-p_4)+$ $p_4(1-p_3)^2(1-p_2)^2 + 2p_2(1-p_2)(1-p_3)^2p_4+$ $p_3^2(1-p_3)^2(1-p_4) + 2p_3^3(1-p_4)(1-p_3)(1-p_1)$ $+2(1-p_3)^2p_3p_4p_1(1-p_2)+$ $2(1-p_3)p_3p_4(1-p_1)p_2(1-p_2)+$ $2(1-p_3)p_3p_4(1-p_1)(1-p_2)$
7	1	0	2	1	0	0	2	p_3^2
8	1	1	2	1	0	0	2	$(1-p_3)^2 + 2p_3(1-p_3)$
9	1	0	2	2	0	0	2	p_3^4
10	2	2	2	2	0	0	2	$(1-p_3)^4 + 4(1-p_3)^3p_3 + 4p_3^2(1-p_3)^2+$ $4p_3^3(1-p_3)(1-p_1) + 2p_3^2(1-p_3)^2(1-p_1)$
11	2	1	0	0	2	1	0	p_2^2
12	1	1	2	0	2	1	0	$(1-p_2)^2 + 2p_2(1-p_2)(1-p_4)$
13	1	1	0	0	3	1	0	p_2^3
14	4	1	0	1	2	1	0	$p_1p_2^2$
15	1	2	0	1	2	1	0	$(1-p_1)p_2^4$
16	1	2	2	1	2	1	0	$2(1-p_2)p_2p_1(1-p_3)(1-p_4)+$ $2(1-p_2)p_2p_1(1-p_4)p_3(1-p_3)+$ $2(1-p_2)p_2p_1p_4(1-p_3)(1-p_2)+$ $2(1-p_1)p_2^3(1-p_2)(1-p_4) + (1-p_1)p_2^2(1-p_2)^2+$ $(1-p_2)^2p_1(1-p_3)^2 + 2(1-p_2)^2p_1(1-p_3)p_3+$ $2(1-p_2)(1-p_1)p_2^2(1-p_4)+$ $2(1-p_2)^2(1-p_1)p_2(1-p_4)+$ $2(1-p_2)^2(1-p_1)p_2p_4$
17	1	1	0	2	1	1	0	$p_1^2p_2$
18	1	1	1	0	1	1	1	p_2p_4
19	1	1	2	0	1	1	1	$(1-p_3)p_1 + p_3(1-p_1) + (1-p_3)(1-p_1)$
20	1	1	2	1	1	1	1	$(1-p_2)(1-p_4)p_1p_3^2 + (1-p_4)p_1p_3^2p_2 + (1-p_2)p_1p_3^3p_4$
21	1	2	2	1	0	1	2	$(1-p_1)p_3^2 + (1-p_1)(1-p_3)^2 + 2p_1p_3(1-p_3)+$ $2(1-p_1)p_3(1-p_3) + (1-p_3)^2p_1$
22	1	2	2	0	1	2	1	$(1-p_2)^2(1-p_4) + (1-p_2)^2p_4 + 2(1-p_2)p_2(1-p_4)+$ $p_2^2(1-p_4) + 2p_2(1-p_2)p_4$