# Kleine kansen en grote afwijkingen

**In dit artikel wordt een onwaarschijnlijke (of zeldzame) gebeurtenis in een wachtrijmodel bestudeerd en gesimuleerd. De zeldzame gebeurtenis treedt op omdat het model zich gedurende een periode niet gemiddeld gedraagt, maar ver daarvan afwijkt. Een theorie uit de kansrekening laat zien hoe de onwaarschijnlijke gebeurtenis hoogstwaarschijnlijk optreedt. Daarvan gebruikmakend kan de simulatie aanzienlijk versneld worden.**

**Dr. A.A.N. Ridder** studeerde wiskunde aan de UvA, promoveerde in 1987 aan de Universiteit Leiden en is sinds 1992 UHD besliskunde aan de VU. Zijn onderzoek ligt op het gebied van wachttijdtheorie en simulatie.

## Inleiding

Kun je je voorstellen dat een verzekeringsmaatschappij failliet gaat? Ja, want de inkomsten zijn min of meer constant, gegeven door het aantal verzekerden en hun premies. Maar de uitgaven kunnen een grillig patroon volgen want ze hangen af van de (goedgekeurde) claims van die verzekerden. Als toevallig heel veel mensen hele hoge claims terecht indienen, dan loopt het misschien fout af voor de maatschappij. Natuurlijk, de kans dat dit gebeurt is klein.

Wat is de kans dat een vliegtuig neerstort op de Rivierenbuurt in Amsterdam vóór 2010? Als je wekelijks de lotto invult, verwacht je dan binnen een jaar de jackpot te winnen? Zal er in de 21-ste eeuw een aardbeving plaatsvinden in Nederland van een kracht 7 op de schaal van Richter? Dit zijn allemaal voorbeelden van onwaarschijnlijke situaties, je verwacht niet dat ze zullen gebeuren. Maar als ze gebeuren hebben ze enorme economische of maatschappelijke consequenties (het lottovoorbeeld slechts in de privésfeer). Bovendien, als genoeg mensen wekelijks de lotto invullen, is er met hele grote kans wel een winnaar tussen. Ieder jaar verongelukt er wel een vliegtuig ergens in de wereld. Het onwaarschijnlijke is dat de gespecificeerde situatie gebeurt.

Neem nu eens aan dat het systeem waarin de onwaarschijnlijke situatie zou kunnen gebeuren, beschreven wordt door een kansmodel. Misschien kan de kans dan berekend worden. Zo niet, dan is simulatie een methode om de kans te schatten. In dit artikel zal ik dit toelichten aan de hand van een specifiek kansmodel (namelijk een wachtrijmodel) en een specifieke situatie (namelijk een lange wachtrij in een korte tijd). Daarbij zal blijken dat het het simuleren niet probleemloos gaat en dat het vereist is de simulatie te versnellen door een variantiereductietechniek.

## De M/M/1 Wachtrij

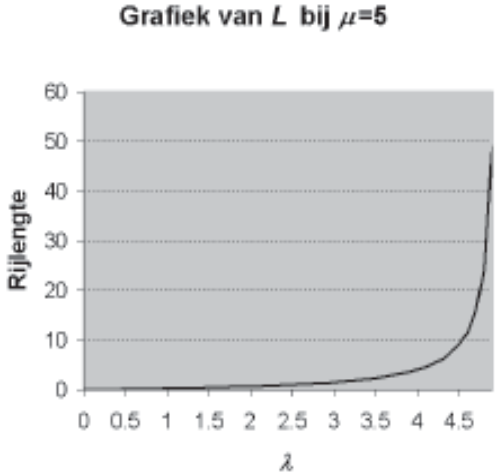
Iedere econometrist en besliskundige leert tijdens zijn/haar studie het bekende M/M/1 wachtrijsysteem. Wachttijdtheorie is de studie van abstracte modellen om het fenomeen wachten te beschrijven in wiskundige relaties. Grofweg gezegd houdt een wachtrijmodel in dat klanten arriveren bij een servicedienst die hen één voor één afhandelt. Klanten wachten als ze nog niet aan de beurt zijn en vormen zo de wachtrij. Wachtrijmodellen worden toegepast bij telefonische informatiediensten (call centers) om personeel te plannen, en bij computernetwerken (internet) om de capaciteit te bepalen.

De genoemde M/M/1 wachtrij is het meest eenvoudige model: de tijdsduren tussen de aankomsten van klanten zijn onafhankelijke gelijkverdeelde stochastische variabelen, en ook de gevraagde services zijn onafhankelijke gelijkverdeelde stochastische variabelen (en onafhankelijk van de aankomstvariabelen). Bovendien zijn de twee verdelingen voor tussenaankomsttijd en service exponentieel. De bijhorende parameters worden meestal $\lambda$

en $\mu$ genoemd. Die parameters kun je interpreteren als: $\lambda$ is het verwachte aantal aankomsten per tijdseenheid, en $\mu$ is het verwachte aantal klanten dat de servicedienst kan afhandelen per tijdseenheid. Je voelt dus wel aan dat geëist wordt dat $\lambda<\mu$ is, want anders kan de serviceverlener de stroom klanten niet aan en er zal een oneindig grote wachtrij ontstaan. Echter, als $\lambda$ wel kleiner dan $\mu$ is maar er vlakbij ligt, wordt de wachtrij (gemiddeld over een lange tijd) zeer groot. Ongetwijfeld ken je de beroemde formule voor de gemiddelde wachtrij in de M/M/1 waarin dat is af te lezen:

$$L = \frac{\lambda}{\mu - \lambda}$$

Hierbij heb ik voor het gemak ook de klant in service meegerekend als zijnde in de wachtrij. Zo'n formule wordt veel duidelijker door het tekenen van de grafiek, zie onderstaand figuur.




Vaak maakt men bezwaar tegen dergelijke wiskundige exercities omdat die veronderstellen dat de kansverdelingen van de aankomsten en de gevraagde diensten gelijk blijven gedurende een oneindige lange periode. In de werkelijkheid zitten daar fluctuaties in en mag gerust $\lambda$ groter dan $\mu$ zijn over een korte periode. Als gemiddeld $\lambda<\mu$ is, dan corrigeert het systeem zich na enige tijd wel weer naar een stabiele toestand. Zo mogen we in de eerste dagen (weken?) van 2002 lange wachtrijen verwachten bij de kassa's van supermarkten omdat de bedieningstijden langer zullen zijn ten gevolge van de nieuwe Euromunten en wegens de mogelijkheid tot omwisselen. Maar in februari zal de situatie weer zijn zoals de huidige.

## Een overschrijdingsprobleem

Ik wil nu het volgende probleem bespreken in het M/M/1 model. Veronderstel dat er op tijdstip 0 geen wachtrij is en veronderstel dat inderdaad $\lambda<\mu$, hoe groot is dan de kans dat de wachtrij een lengte $B$ bereikt binnen $T$ tijdseenheden? Hierbij is $B$ relatief groot. (Dit is typisch een probleem in het internet, de wachtrij stelt dan een buffer voor in een centrale waar vele verbindingen van internetsessies samenkomen en vervolgens gesplitst worden naar de bestemmingen. Natuurlijk is de buffer eindig, en soms is die vol. We hebben dan te maken met zogeheten bufferoverschrijding.) Dit betekent in ons M/M/1 model dat gedurende een periode van $T$ tijdseenheden toevallig veel aankomsten zijn, veel meer dan de verwachte $\lambda T$. Ook kunnen de services toevallig langer duren dan de verwachte $1/\mu$. Let op: ik zeg niet dat de kansverdelingen voor aankomst en service in deze periode anders zijn (waardoor $\lambda>\mu$). Omdat we te maken hebben met stochastische variabelen, kunnen de uitkomsten soms ver afwijken van wat we verwachten. Zo kun je ook voorstellen dat je ooit tien keer achter elkaar 6 gooit met een gewone dobbelsteen.

De genoemde kans kan in het M/M/1 model exact berekend worden. Een analytische oplossing is helaas slechts mogelijk in een paar wachtrijmodellen. In alle andere gevallen moet je iets anders bedenken. Een gebruikelijke methode is de kans te schatten door middel van het simuleren van het model door een computerprogramma. Dat houdt in dat we de computer het model laten naspelen door onafhankelijke realiseringen van hoogstens $T$ tijdseenheden te genereren. De fractie van het aantal realiseringen waarin $B$ bereikt wordt, is dan de geschatte kans. Een realisatie van het M/M/1 model over een eindige tijd noem ik een pad. Ik heb een simulatieprogramma in MATLAB geschreven waarmee paden gegenereerd kunnen worden. Laat ik als tijdseenheid de verwachte serviceduur nemen: $\mu = 1$. Dan is bij $\lambda = 0.5$, $T = 25$, $B = 19$, de geschatte overschrijdingskans gelijk aan $1.2 \times 10^{-6}$. Dit is een kleine kans behorend bij een gebeurtenis die zelden optreedt: ongeveer één op de miljoen keer wordt de lengte van de wachtrij binnen 25

tijdseenheden 19 klanten groot (bij de genoemde $\lambda$ en $\mu$ en startend vanaf 0).

Wat betekent dat nu voor het uitvoeren van de simulaties? Veronderstel dat we $N$ paden genereren en dat $k$ daarvan resulteren in een succes ($B$ bereikt binnen $T$ tijdseenheden), dan is de fractie $f = k/N$ de schatting. Zoals gebruikelijk stellen we een betrouwbaarheidsinterval (bti) om $f$ op dat aangeeft hoe nauwkeurig (betrouwbaar) de schatting is: $(f\text{-}w,f+w)$. Nu is eenvoudig na te gaan (door ook de variantie van de schatter te schatten) dat voor een 95%-bti:

$$w = 1{,}96\sqrt{f(1-f)/(N-1)}$$

Het is natuurlijk belangrijk om relatief een smal bti te hebben, zeg relatief 10% naar beide kanten om de schatting: $w/f = 0{,}1$.

Substitueren we nu $f = 1.2 \times 10^{-6}$ dan krijgen we dat de steekproefomvang $N$ ongeveer 320 miljoen moet zijn. Dat wil zeggen, we bootsen 320 miljoen keer het M/M/1 model na over $T$ tijdseenheden, steeds te beginnen met een leeg systeem, pas dan zijn we zeker van de gestelde (95%,10%) nauwkeurigheid van de schatting. Op mijn PC (een Pentium III-450 PC met 63MB RAM onder Windows 95) gaat dat te lang duren. Uit een proef van 10000 paden bleek dat een pad gemiddeld 2.4 msec duurt. (In MATLAB: Programma's in C werken veel sneller. Deze simulatie gaat in C zelfs 20 keer zo snel!) Dat komt uit op een simulatieduur van 8 dagen en 21 uur tot de (95%,10%) nauwkeurigheid bereikt is. Daarbij moet je bedenken dat je bij een overschrijdingsprobleem in het computernetwerk eigenlijk geïnteresserd bent in het omgekeerde: hoe groot moet $B$ zijn zodat de kans hoogstens $10^{-6}$ is bij de gegeven $\lambda$, $\mu$, $T$. Dan simuleer je allerlei scenario's (verschillende $B$) totdat dit bereikt is. In situaties waar sprake is van gevoelige data wil men meestal een nog veel kleinere kans bereiken, van de orde $10^{-9}$.

Behalve de tijdsfactor is er bij dergelijke simulaties van zeldzame gebeurtenissen nog een probleem. Dat heeft te maken met de random generator. Elke gerealiseerde tussenaankomsttijd en servicetijd in de simulaties is het resultaat van een aanroep van de random generator. In een goede random generator mogen de afzonderlijke gegenereerde getallen beschouwd worden als zijnde realiseringen van onafhankelijke stochastische variabelen. Maar, die rij is begrensd door de (zogeheten) cycluslengte van de random generator. In een PC is de lengte $2^{31}$, dat is ongeveer $2 \times 10^9$. In het getalvoorbeeld van zojuist zijn er per pad ongeveer 25 aankomsten en vertrekken, dus 25 aanroepen van de random generator. Bij $N$ = 320 miljoen paden kom je dan uit op $8 \times 10^9$. Je zou dus tot de conclusie kunnen komen dat een simulatiestudie niet de gewezen weg is om kansen op zeldzame gebeurtenissen te schatten.

## Importance Sampling

Gelukkig bestaan er allerlei variantiereductietechnieken om de simulatie te versnellen. Importance sampling is zo'n methode. Importance sampling is gebaseerd op het veranderen van kansverdelingen. De achterliggende gedachte is: een overschrijding vindt plaats omdat er een afwijkend pad van het systeem optreedt. Doe nu alsof het afwijkende pad een gemiddeld pad is van een ander M/M/1 model. In deze nieuwe M/M/1 zijn de parameters $\lambda^*$ en $\mu^*$ en daarom schrijven we (M/M/1)*. Over $T$ tijdseenheden verwachten we $\lambda^*T$ aankomsten en $\mu^*T$ vertrekken. Opdat een overschrijding gebeurt, eisen we

$$(1)\quad \lambda^*T - \mu^*T \geq B$$

Hoe krijgen we nu een correcte schatting van de overschrijdingskans in het oorspronkelijke M/M/1 model als we de (M/M/1)* wachtrij simuleren? Veronderstel dat $a_1, a_2, \ldots$ en $s_1, s_2, \ldots$ de gerealiseerde waarden zijn van tussenaankomsttijden en servicetijden in een pad van het (M/M/1)* model dat overschrijding tot gevolg heeft. Dit pad telt dan niet mee voor 1 (zoals oorspronkelijk), maar voor zijn likelihood ratio $L$. De likelihood ratio van een enkele gerealiseerde tussenaankomsttijd $a$ is

$$L(a) = \frac{\lambda e^{-\lambda a}}{\lambda^* e^{-\lambda^* a}}$$

Zo wordt de likelihood ratio van het pad het product (wegens onafhankelijkheid) van al de afzonderlijke likelihood ratios van de gerealiseerde tussenaankomsttijden en servicetijden van het pad.

De schatting wordt nu :

$$\frac{\text{som van likelihood ratios van succespaden}}{\text{aantal gegenereerde paden}}$$

Dat de betreffende schatter zuiver is, volgt uit de volgende observatie. Laat $X$ exponentieel $\lambda$ verdeeld zijn, en $X^*$ exponentieel $\lambda^*$, dan is

$$E[X] = \int_0^\infty x\lambda e^{-\lambda x} dx$$

$$= \int_0^\infty L(x) x\lambda^* e^{-\lambda^* x} dx = E[LX^*].$$

Zoals gebruikelijk kunnen we weer de variantie van de schatter schatten en daarmee een 95%-bti opstellen. De vraag is natuurlijk: is de variantie inderdaad kleiner dan de oorspronkelijke (bij dezelfde steekproefomvang $N$)? Zo ja, dan kunnen we met een kleinere steekproef volstaan om toch weer een (95%,10%) nauwkeurigheid te garanderen. Om de vraag te beantwoorden keer ik terug naar vergelijking (1). Om daaraan te voldoen zijn er oneindig veel mogelijkheden voor de twee nieuwe parameters. Je kunt zelf eens experimenteren, en dan blijkt dat heel vaak geen variantie-reductie zal optreden. Een goede keuze wordt verkregen door de theorie van grote afwijkingen (Large Deviations) toe te passen.

## Large Deviations

De theorie van grote afwijkingen is een onderdeel van de kansrekening waarmee zeldzame gebeurtenissen worden bestudeerd. Ze kan verklaren hoe een zeldzame gebeurtenis hoogstwaarschijnlijk optreedt. In het overschrijdingsprobleem van de M/M/1 zijn er zeer veel (oneindig veel) paden die leiden tot overschrijding binnen $T$ tijdseenheden. Maar alle zijn onwaarschijnlijk zoals blijkt uit de simulatiestudie. Een stelling uit de large deviations theorie zegt dat van al deze overschrijdingspaden degene met een constante drift $B/T$ de grootste kans heeft. Bovendien is het aantal aankomsten langs dit pad $\lambda xT$ en het aantal vertekken $\mu T/x$, waarbij $x$ een factor is zodat

(2) $\lambda xT - \mu T/x = B$

Vergelijking (2) is een standaard vierkantvergelijking in $x$ en gemakkelijk op te lossen. De aankomsten worden versneld met de factor $x$, en de services worden vertraagd met dezelfde factor. Duidelijk is dat als we $\lambda^* = \lambda x$ en $\mu^* = \mu/x$ kiezen, de nieuwe parameters voldoen aan (1) voor de importance sampling simulaties. Nogmaals de large deviations stelling toepassen geeft dat deze keuze van parameters de beste keuze is, dat wil zeggen dat de grootste variantiereductie wordt verkregen. Uitwerken in het getalvoorbeeld van boven geeft dat $\lambda^* = 1.1827$ en $\mu^* = 0.4227$. Ik heb ook de importance sampling simulaties met deze parameters uitgevoerd in MATLAB. Bij steekproefomvang $N = 5000$ wordt de (95%,10%) nauwkeurigheid al bereikt. Wegens het grotere rekenwerk (likelihood) duurt de generatie van een pad langer, ongeveer 10 msec, maar ruim binnen één minuut is de schatting (95%,10%) nauwkeurig. Trouwens, de snelheid van C ten opzichte van MATLAB is nu nog frappanter: 30 keer zo snel.

---

NB: een waarschuwing mocht je zelf gaan experimenteren. De genoemde nieuwe parameters zijn alleen optimaal als $B/T \geq \lambda - \mu$. Is $B$ te klein of $T$ te groot, dan moet je een hybride simulatie toepassen: simuleer het oorspronkelijke M/M/1 model tot het tijdstip

$$t^* = T - \frac{B}{\mu - \lambda}$$

Vanaf $t^*$ simuleer je (M/M/1)* met $\lambda^* = \mu$ en $\mu^* = \lambda$. Ik zal hier verder niet op ingaan.

*Verdere informatie:*

De codes van MATLAB programma's zijn te vinden op mijn internetpagina:
http://www.econ.vu.nl/medewerkers/aridder/divers.html#aenorm

Een grafische illustratie van paden en overschrijdingen is de java applet:
http://www.econ.vu.nl/medewerkers/aridder/applets/simRW/simRW.html

Een boek waarin large deviations en importance sampling simulaties aan de orde komen is:
J.A. Bucklew. *Large Deviation Techniques in Decision, Simulation, and Estimation*, Wiley, New York, 1990.