Een nieuw onderwerp dat is voorgesteld voor het wiskundeprogramma van de profielen N&T en N&G in de vernieuwde Tweede Fase is wachttijden. **Ronald Meester** legt uit om welke praktische vragen het draait in de wachttijdtheorie en licht een tipje van de wiskundige sluier op.

Hoeveel wachtenden zijn er voor u?

Wat is wachttijdtheorie?

Hoeveel loketten moeten er open zijn op een postkantoor opdat het aantal wachtende klanten niet al te groot wordt? Hoeveel mensen zijn er nodig bij de 06-8008 informatielijn? (Bij deze 06-8008 informatielijn ben ik trouwens altijd verrast door het volgende feit: het aantal wachtenden vóór mij neemt af totdat er nog één wachtende voor mij is; daarna kom ik direct aan de beurt, terwijl er toch ook nog een periode moet zijn waarin er géén wachtende voor mij is...). Wat is de kans dat ik in de supermarkt langer dan vijf minuten moet wachten bij de kassa? Als ik op een willekeurig moment bij een bushalte aankom, hoe lang moet ik dan gemiddeld op de eerste bus wachten?
De tak van de kansrekening die zich met dit soort vragen bezighoudt, wordt wachttijdtheorie genoemd. Bovenstaande vragen zijn in principe geen wiskundige vragen. Door een beetje te experimenteren, kun je er wel achter komen hoeveel loketten er nodig zijn op bepaalde tijden in een postkantoor en door een aantal keren op een willekeurig moment naar de bushalte te gaan, kun je wel een aardig idee krijgen van de gemiddelde wachttijd. Vaak zal zo'n experiment echter niet mogelijk zijn. Als je bijvoorbeeld een productielijn van een nieuwe fabriek wilt ontwerpen, dan kun je gewoonlijk op geen enkele manier experimenteren. Voor dit soort situaties ben je aangewezen op een wiskundig model van het systeem. Dit model moet de reële situatie natuurlijk zo nauwkeurig mogelijk beschrijven.
Hoe ziet zo'n wiskundig model er uit? Ik zal nu eerst een informele beschrijving van zo'n model geven en daarna zal ik alles wat preciezer doen. We beperken ons tot een systeem met één bediende die binnenkomende klanten in volgorde van binnenkomst bedient. Na bediening verlaten de klanten het systeem.

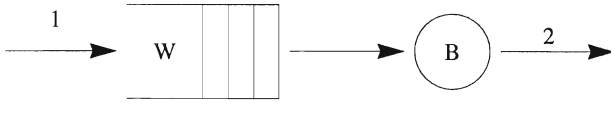

fig. 1 Binnenkomende klanten (1) sluiten aan in de wachtrij (W). Vervolgens wachten ze op hun beurt, worden bediend door de bediende (B) en verlaten het pand (2).

Verder nemen we aan dat er gemiddeld λ binnenkomende klanten per uur zijn en dat de bediende per uur gemiddeld μ klanten kan bedienen. (Voor de eenheid van tijd mag ook best iets anders genomen worden, bijvoorbeeld minuten in plaats van uren.) Het zal geen verbazing wekken dat het quotiënt $\rho = \frac{\lambda}{\mu}$ van belang is. Als $\rho > 1$ is, komen er gemiddeld meer klanten het systeem binnen dan er bediend kunnen worden en zal de rij wachtenden waarschijnlijk wel lang worden. Als $\rho < 1$ is, heeft de bediende gemiddeld voldoende capaciteit om de binnenkomende klanten te bedienen. Het geval $\rho = 1$ is heel speciaal en laten we hier verder buiten beschouwing.
Het getal ρ is zo belangrijk dat er een speciale term voor bestaat, ρ wordt de *verkeersintensiteit* genoemd. Als bijvoorbeeld $\lambda = \frac{4}{5}$ en $\mu = 1$, dan is de verkeersintensiteit gelijk aan $\frac{4}{5}$. Dit eenvoudige model kan op een computer gesimuleerd worden.

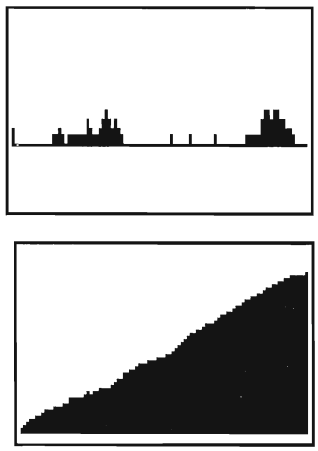

fig. 2

In figuur 2 is het aantal wachtenden als functie van de tijd afgebeeld. In het bovenste plaatje is $\lambda = \frac{4}{5}$ en $\mu = 1$, dus $\rho = \frac{4}{5}$. In het onderste plaatje is $\lambda = 3$ en $\mu = 1$, dus $\rho = 3$. In het eerste plaatje is de lengte van de wachtrij redelijk stabiel. Er zijn wel uitschieters, maar die zijn altijd tijdelijk. Het aantal wachtende klanten in het tweede plaatje stijgt echter snel. Dit komt omdat $\rho < 1$ in het eerste plaatje en $\rho > 1$ in het tweede.

Drie vragen

Bij dit model is een aantal voor de hand liggende vragen te stellen die alleen zinvol te beantwoorden zijn als $\rho < 1$, wat we vanaf nu dan ook aannemen. In deze paragraaf geven we een overzicht van die vragen met hun antwoorden. Waar die antwoorden vandaan komen, is onderwerp van de paragrafen daarna.

1. *Wat is de kans om bij aankomst bij het systeem precies k klanten vóór je te treffen?*

Deze kans zal gelijk blijken te zijn aan $\rho^k(1-\rho)$. Als $\lambda = \frac{4}{5}$ en $\mu = 1$ zoals in ons eerste voorbeeld, dan is bijvoorbeeld de kans dat er géén wachtenden voor je zijn gelijk aan $\rho^0(1-\rho) = \frac{1}{5}$. Deze kansverdeling wordt de *geometrische verdeling* genoemd. We kunnen nu ook heel eenvoudig de kans uitrekenen dat er minstens m klanten voor je zijn. Deze kans is nu immers gelijk aan

$$\sum_{k=m}^{\infty} (1-\rho)\rho^k = (1-\rho)\sum_{k=0}^{\infty} \rho^m \rho^k$$

$$= (1-\rho)\rho^m \sum_{k=0}^{\infty} \rho^k = (1-\rho)\rho^m \frac{1}{1-\rho}$$

$$= \rho^m.$$

We zien dat als de verkeersintensiteit afneemt, het gemiddelde aantal wachtende klanten voor je ook afneemt, wat natuurlijk heel logisch is.

2. *Wat is de gemiddelde tijd die een klant moet wachten voor hij/zij aan de beurt is?*

Het antwoord op deze vraag is

$$\frac{\rho}{\mu - \lambda}.$$

In ons voorbeeld waar $\lambda = \frac{4}{5}$ en $\mu = 1$ geeft dit $\frac{4}{25}$. In figuur 3 is de gemiddelde wachttijd als functie van μ in het geval dat $\lambda = 1$ afgebeeld.

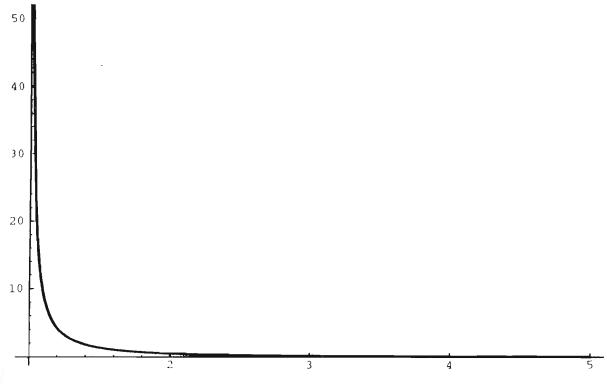

fig. 3 De gemiddelde wachttijd als $\lambda = 1$ als functie van μ

3. *Hoe lang zal de bediende gemiddeld onafgebroken bezig blijven met klanten vanaf het moment dat een klant binnenkomt in een leeg systeem?*

Dit gemiddelde is:

$$\frac{1}{\mu - \lambda}.$$

Voor ons voorbeeld met $\lambda = \frac{4}{5}$ en $\mu = 1$ geeft dit 5. We zien ook hier weer dat als λ en μ heel dicht bij elkaar liggen, de bediende weinig vrije tijd zal hebben...

Eindeloos wachten...

Voor we ons in het wiskundige diepe storten, is het aardig om eens te kijken wat de antwoorden op de drie vragen voor consequenties hebben.

Als een medisch specialist gemiddeld tien minuten doet over een behandeling van een patiënt, dan is $\mu = 6$ en zal deze specialist geneigd zijn om per uur zes afspraken te maken (en λ wordt dan dus gelijk aan 6). Maar wat gebeurt er als λ heel dicht bij μ ligt? Uit het antwoord op de tweede vraag volgt dat de limiet van de gemiddelde wachttijd voor $\lambda \uparrow \mu$ gelijk is aan ∞! Je ziet dus dat de zaken vreselijk uit de hand kunnen lopen: zelfs als λ een heel klein beetje kleiner is dan μ, dan nog is de gemiddelde wachttijd heel erg groot. Dit verklaart waarom we vaak zo lang moeten wachten in een polikliniek. De specialisten zijn erg bang om zonder patiënten (en dus ook zonder inkomsten) te zitten en daarom maken ze dus in feite te veel afspraken.

Ditzelfde effect kan optreden bij kaartjesautomaten op een station en bij informatielijnen. Stel dat uit onderzoek blijkt dat er 40 mensen per uur kaartjes kopen via een automaat en dat het apparaat er 45 aan kan per uur (dus $\lambda = 40$ en $\mu = 45$). Denk dan niet dat alles uitstekend geregeld is! De capaciteit overstijgt weliswaar de vraag, maar de gemiddelde wachttijd van een klant is nu

$$\frac{40/45}{45-40} = 0{,}178.$$

Dit lijkt niet lang, maar bedenk dat we in *uren* rekenen, dus dit betekent een gemiddelde wachttijd die langer is dan tien minuten...

Een wiskundig wachtrijmodel

In de paragrafen die volgen, beschrijven we een formeel wiskundig model voor zo'n systeem. Het lijkt misschien dat dat niet meer zo nodig is na de informele beschrijving van hiervoor, maar het is bijvoorbeeld niet zo duidelijk hoe het feit dat er gemiddeld λ klanten per uur aankomen wiskundig precies geformuleerd kan worden. Tenslotte volgt een schets van hoe de wiskundige analyse van zo'n model in zijn werk gaat.

Zoals we hiervoor al zagen, wordt het model beschreven door het antwoord op twee vragen:

1. Hoe komen klanten het systeem binnen?
2. Hoe groot is de bedieningstijd van de klanten?

We beginnen met de eerste vraag. Wat zijn nu redelijke

aannames om het aankomstproces van klanten te modelleren? Als eerste nemen we aan dat het aantal binnenkomende klanten in niet-overlappende tijdsintervallen onafhankelijk is van elkaar: hoeveel klanten er tussen 10.00 en 11.00 binnenkomen zal niets zeggen over hoeveel klanten er bijvoorbeeld tussen 12.00 en 12.30 binnenkomen.
Voor de tweede aanname bekijken we een heel klein tijdsintervalletje ter lengte δ. We nemen aan dat de kans dat er precies één klant binnenkomt in dat intervalletje, voor kleine δ, ongeveer evenredig is met de lengte van het interval. Dit betekent dat er een positief getal λ bestaat, zodanig dat deze kans ongeveer gelijk is aan $\lambda\delta$. (Straks zal blijken dat deze λ overeenkomt met de λ uit de eerste paragraaf.) Tenslotte nemen we aan dat de kans dat er géén klant binnenkomt in een tijdsintervalletje ter lengte δ, voor kleine δ, ongeveer gelijk is aan $1 - \lambda\delta$. Aangezien de kans op één klant in het intervalletje plus de kans op geen klanten in het intervalletje samen ongeveer 1 zijn, moet de kans op twee of meer aankomsten in een klein interval ongeveer 0 zijn.

> Voor diegenen voor wie de term 'ongeveer' in de alinea hierboven onbevredigend is, kunnen we de aanname als volgt preciseren. Als $F(y)$ de kans is dat er precies één klant binnenkomt in een interval ter lengte y, dan is natuurlijk $F(0) = 0$. De aanname hierboven betekent dat $F(y)$ differentieerbaar is in $y = 0$ met afgeleide $F'(0) = \lambda$. Als $G(y)$ de kans is op géén klant in een interval ter lengte y, dan is $G(0) = 1$. De laatste aanname hierboven betekent dat $G(y)$ in $y = 0$ differentieerbaar is met afgeleide $G'(0) = -\lambda$.

Het getal λ wordt de *intensiteit* van het aankomstproces genoemd. De drie aannames die we nu hebben gemaakt, zijn heel natuurlijk. Het aardige is dat deze aannames het aankomstproces volledig vastleggen: als we de tijdstippen waarop klanten aankomen, aangeven met T_1, T_2, T_3, ... (dus de i-de klant komt binnen op tijstip T_i, waarbij we voor het gemak zeggen dat $T_0 = 0$) dan noemen we de verschillen $X_i = T_i - T_{i-1}$ de *tussenaankomsttijden* van de klanten, voor $i = 1, 2, \ldots$

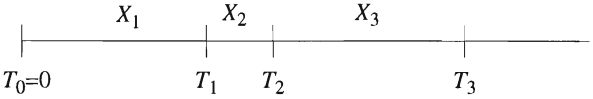

fig. 4 De aankomsttijdstippen T_i van de klanten en de bijbehorende tussenaankomsttijden X_i

Onder de bovenstaande aannames *blijkt* het zo te zijn dat alle tussenaankomsttijden onafhankelijk zijn van elkaar én dat

$$P(X_i \leq y) = 1 - e^{-\lambda y},\ y \geq 0.$$

Dit is dus geen keuze die we maken, maar een feit dat volgt uit de aannames. Op het moment dat er een klant binnenkomt, is dus de kans dat de volgende klant binnen y uur arriveert gelijk aan $1 - e^{-\lambda y}$. Deze kansverdeling wordt de *exponentiële verdeling* genoemd.

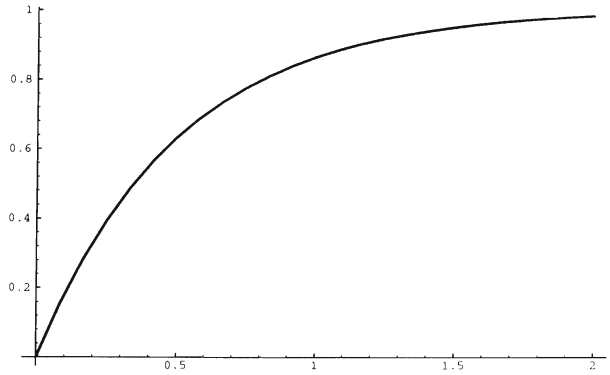

fig. 5 De kans $P(X_i \leq y) = 1 - e^{-2y}$ als functie van y. Dit wordt de exponentiële verdeling genoemd

Je kunt nu met wat elementaire kansrekening de gemiddelde tussenaankomsttijd uitrekenen. Deze blijkt gelijk te zijn aan $\frac{1}{\lambda}$. Dit betekent dus dat er gemiddeld λ klanten per uur aankomen! Dit is precies wat we in het begin aannamen in onze informele discussie. Als bijvoorbeeld $\lambda = 2$, betekent dit dat er gemiddeld elk half uur een klant binnenkomt.

Vervolgens bespreken we de tweede vraag, de bedieningstijden van klanten. In analogie met het aankomstproces nemen we aan dat alle bedieningstijden onafhankelijk zijn van elkaar. De bedieningstijden zijn natuurlijk stochastisch, dus we moeten een keuze maken voor de verdeling van deze bedieningstijden. Misschien denkt u direct aan de normale verdeling; in de praktijk blijken immers heel veel dingen ongeveer een normale verdeling te hebben. In dit geval is dat echter niet zo. Uit experimenten blijkt dat opnieuw de exponentiële verdeling een goede benadering is. We nemen daarom aan dat de kans dat een klant minder dan y uur bedieningstijd nodig heeft gelijk is aan

$$1 - e^{-\mu y},\ y \geq 0$$

voor een bepaalde positieve constante μ. Net als hierboven is de gemiddelde bedieningstijd nu gelijk aan $\frac{1}{\mu}$ en dit betekent weer dat de bedieningscapaciteit van de bediende μ klanten per uur is. Dit is opnieuw in overeenstemming met de aanname in de eerste paragraaf. Als bijvoorbeeld $\mu = 8$, betekent dit alles dat de bediende gemiddeld $7\frac{1}{2}$ minuten per klant nodig heeft.

Geheugenloosheid van het model

Een essentieel begrip dat voor het vervolg nodig is, is het begrip *geheugenloosheid*. Dit slaat op een fenomeen dat het beste aan de hand van een voorbeeld is uit te leggen. Stel dat we twee dezelfde, maar onafhankelijke, wachtrijsystemen naast elkaar observeren. Het aankomstproces en het bedieningsproces van de twee systemen hebben dus dezelfde λ en dezelfde μ, maar verder hebben ze niets

met elkaar te maken. Stel nu dat op een gegeven moment in systeem 1 klant A bediend wordt en in systeem 2 klant B. Op een bepaald moment is één van klanten klaar, zeg klant A, en vervolgens gaat de bediende in systeem 1 een nieuwe klant, zeg klant C, bedienen. Wie heeft er nu de grootste kans om als eerste klaar te zijn, klant B of klant C? In eerste instantie ligt het voor de hand om te zeggen klant B, want die wordt al een tijdje bediend, terwijl klant C net begint. Het blijkt echter dat als de bedieningstijden van de klanten dezelfde exponentiële verdeling hebben, de klanten B en C allebei kans $\frac{1}{2}$ hebben om als eerste klaar te zijn!

Hetzelfde verrassende effect treedt op bij het aankomstproces. Als we eerst wachten totdat bij één van de twee systemen een klant binnenkomt, is de kans dat de *volgende* klant bij *hetzelfde* systeem binnenkomt weer gelijk aan $\frac{1}{2}$. Het feit dat we in het andere systeem al een tijdje aan het wachten waren, doet er dus niet toe.

Dit verschijnsel kan als volgt wiskundig bewezen worden. We noemen de (stochastische) bedieningstijden van de klanten B en C even Y_B en Y_C. We willen de kans uitrekenen dat B eerder klaar is dan C, gegeven de gebeurtenis dat B al een bepaalde tijd bediening, zeg t minuten, achter de rug heeft. Daartoe gaan we eerst de verdeling van Y_B uitrekenen, gegeven de gebeurtenis $\{Y_B > t\}$:

$$P(Y_B > t+s \mid Y_B > t) = \frac{P(Y_B > t+s \cap Y_B > t)}{P(Y_B > t)}$$

$$= \frac{P(Y_B > t+s)}{P(Y_B > t)} = \frac{e^{-\mu(t+2)}}{e^{-\mu t}}$$

$$= e^{-\mu s}.$$

Dit is weer een exponentiële verdeling met parameter μ. Dus $P(Y_C > s) = P(Y_B > t + s \mid Y_B > t)$ en aangezien de bedieningstijden van B en C onafhankelijk zijn, volgt hieruit eenvoudig dat B en C ook gelijke kans hebben om als eerste klaar te zijn.

In het algemeen is geheugenloosheid als volgt samen te vatten. Als je op een willekeurig moment t het wachtrijsysteem begint te observeren, mag je net doen alsof de laatst binnengekomen klant juist op tijdstip t is aangekomen. Ook kun je net doen alsof de klant die op dat moment bediend wordt, juist op tijdstip t met de bediening is aangevangen. Deze belangrijke eigenschap zal hierna heel erg van pas komen bij de wiskundige analyse van het model.

Om die te kunnen uitvoeren, zijn eerst nog enkele andere nieuwe begrippen nodig en wel de begrippen *Markovprocessen*, *'rates'* en *evenwichtsverdelingen*. Nadat deze begrippen besproken zijn, komen we toe aan waar het eigenlijk om gaat: de wiskundige analyse van de vragen uit het begin van dit artikel.

Markovprocessen

Laten we het aantal klanten in het systeem op tijdstip t (inclusief de mogelijke klant die op dat moment bediend wordt) aanduiden met $Q(t)$. Dan is $(Q(t); t \geq 0)$ een voorbeeld van een stochastisch proces. Uit de geheugenloosheid van de vorige paragraaf volgt dat op elk tijdstip t_0 de verdeling van de toekomst $(Q(t); t > t_0)$ van het proces alleen maar afhangt van het heden $Q(t_0)$ en niet van het verleden $(Q(t); 0 \leq t < t_0)$. Immers, we hebben hiervoor gezien dat het bijvoorbeeld voor de verdeling van de resterende bedieningstijd van een klant op tijdstip t_0 niet uitmaakt hoe lang de bediening al aan de gang is.

Deze eigenschap wordt de *Markoveigenschap* genoemd en een proces met deze eigenschap heet een *Markovproces*.

Een Markovproces is in feite hetzelfde als een Markovketen, het enige verschil is dat een Markovproces een continue tijdsparameter heeft en een Markovketen een discrete. Bij een Markovketen hoort een overgangsmatrix waarin vastgelegd is wat de kans is om vanuit toestand i naar toestand j te springen. Deze kans wordt vaak aangegeven met $p_{i,j}$. In een continu Markovproces zijn ook dergelijke overgangskansen te definiëren. Als je op een bepaald tijdstip t_0 in toestand i bent, dan is de kans dat je op tijdstip $t_0 + t$ in toestand j bent $p_{i,j}(t)$. De Markoveigenschap vertelt ons dat deze kans inderdaad alleen maar afhangt van de positie op tijdstip t_0; wat daarvoor gebeurde is irrelevant.

'Rates'

Een ander belangrijk begrip is het begrip *rate*. (We gebruiken de Engelse term bij gebrek aan een goed Nederlands woord; *snelheid* zou nog het dichtst in de buurt komen.)

We hebben al gezien dat de kans dat er in een heel klein tijdsintervalletje $[t, t + \delta]$ een klant binnenkomt, gelijk is aan $F(\delta) = 1 - e^{-\lambda\delta}$ en dat de afgeleide van F in het punt 0 gelijk is aan

$$F'(0) = \lim_{\delta \downarrow 0} \frac{F(\delta) - F(0)}{\delta} = \lambda.$$

We noemen dit de *rate* (of snelheid) waarmee het systeem, als het in toestand k is, van k naar $k + 1$ klanten 'wil' en noteren dit als $r(k, k + 1) = \lambda$. Deze rate kan beschouwd worden als een soort 'kans per tijdseenheid' dat het systeem van toestand k naar toestand $k + 1$ gaat.

Op dezelfde manier is te zien dat de rate waarmee het systeem van $k + 1$ naar k klanten wil gelijk is aan μ. Daarvoor moeten we kijken naar de gebeurtenis dat een bedieningstijd afloopt in het tijdsinterval $[t, t + \delta]$. We schrijven $r(k + 1, k) = \mu$.

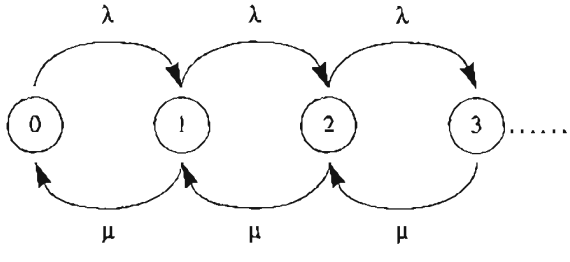

fig. 6 Een schematisch plaatje van de rates tussen de verschillende toestanden

Voorbeeld van een wachtrij

"Wachten in een rij is een van de grootste kwellingen van het moderne bestaan. Het geeft het gevoel overgeleverd te zijn aan een tergend noodlot. Weinig mensen zijn in staat hun goede humeur te bewaren als ze, staande in een rij voor een loket, merken dat de rij naast hen sneller gaat. Vooral als iemand die later is aangekomen daardoor eerder wordt geholpen."
(*De Volkskrant* 25 april 1992)

Het model

Wiskundig gezien is wachten voor een loket een interessant verschijnsel. Voor het wachten voor een loket heeft men het volgende model opgesteld:

$$A = \frac{Ak}{b}$$

A: bezettingsgraad.
k: aantal klanten per uur.
b: aantal klanten dat per uur bediend wordt.

Met behulp van de volgende formule kun je het gemiddeld aantal klanten bij het loket uitrekenen:

$$N = \frac{A}{1 - A}$$

N: gemiddeld aantal klanten bij het loket.
A: bezettingsgraad.

Als je weet hoeveel tijd een klant gemiddeld nodig heeft om geholpen te worden, kun je uitrekenen hoe lang je gemiddeld moet wachten in de rij.

Dit kan met de volgende formule:

$$W = t \times N$$

W: gemiddelde wachttijd in minuten.
N: gemiddeld aantal klanten bij het loket.
t: gemiddelde bedieningstijd per klant in minuten.

Opgave 1
Er komen gemiddeld 10 klanten in een half uur aan en er kunnen gemiddeld 10 klanten in een kwartier geholpen worden. Laat zien dat de bezettingsgraad A gelijk is aan 0,5.

Opgave 2
Op een postkantoor is één loket. Er arriveren gemiddeld 4 klanten per kwartier. Een klant helpen kost gemiddeld 3 minuten.

a. Bereken de bezettingsgraad A van het loket.
b. Bereken hoeveel klanten er gemiddeld in het postkantoor zijn.
c. Bereken hoeveel minuten een klant gemiddeld moet wachten.

Opgave 3
Als je een model bestudeert, kan het soms erg verhelderend zijn om naar extreme omstandigheden te kijken.

a. Hoeveel minuten moet een klant wachten als er gemiddeld 9 klanten per half uur aankomen en een bediening gemiddeld 3 minuten duurt?
b. Teken een grafiek met op de horzontale as de bezettingsgraad A en op de verticale as het gemiddeld aantal klanten in de wachtrij.
c. Wat gaat er mis als er 5 klanten per kwartier aankomen en de gemiddelde bedieningstijd 3 minuten is? Wat gebeurt er dan eigenlijk?

Opgave 4
Een bepaald bedrijf heeft een afdeling, waar tamelijk gecompliceerde machines staan opgesteld, die nogal wat onderhoud vergen. Deze machines gaan volgens een willekeurig patroon kapot, met een gemiddelde van 3 machines per uur. Elk uur machinestilstand kost het bedrijf *f* 50,- in de vorm van produktieverlies.
Het bedrijf kan voor het onderhoud van deze machines beschikken over een geschoolde onderhoudsmonteur, die gemiddeld 5 machines per uur kan repareren, of over een leerlingmonteur, die gemiddeld 4 machines per uur kan repareren. De geschoolde monteur kost 40 gulden per uur en de leerlingmonteur kost 20 gulden per uur.
Welke monteur verdient de voorkeur?

Bron: Ir.F. Huisman. *Inleiding tot de Operationele Research*, 3e druk, Wolters-Noordhoff, Groningen.

Deze opgaven over wachttijden zijn te vinden op de website van Willem van Ravenstein, docent wiskunde.
Het adres is http://www1.tip.nl/users/t434883

Evenwichtsverdelingen

Een vector $\pi = (\pi_0, \pi_1, \ldots)$ heet een *kansvector* als $\pi_n \geq 0$ voor alle n, en $\sum_{n=0}^{\infty} \pi_n = 1$.

Een kansvector heet een *evenwichtsverdeling* voor een Markovproces dan en slechts dan als voor elke toestand k geldt

$$\pi_k \sum_{i \neq k} r(k, i) = \sum_{i \neq k} \pi_i r(i, k). \qquad (1)$$

Deze vergelijking ziet er wat ingewikkeld uit, maar bij nadere beschouwing valt het nogal mee. Het getal π_k interpreteren we als de kans dat het systeem zich in toestand k bevindt, dat wil zeggen als de kans dat er k klanten aanwezig zijn. We hebben al gezien dat $r(k, i)$ te beschouwen is als de kans per tijdseenheid om vanuit toestand k naar toestand i te gaan. De som van $r(k, i)$ voor $i \neq k$ is dus de totale kans per tijdseenheid om uit toestand k naar een andere toestand te gaan. Het linkerlid stelt dan de kans per tijdseenheid voor om vanuit toestand k naar een andere toestand te gaan. Op dezelfde manier is het rechterlid te interpreteren als de kans per tijdseenheid om vanuit een andere toestand juist náár toestand k te gaan. Vergelijking (1) geeft dus aan dat de kans per tijdseenheid om uit toestand k te verdwijnen net zo groot is als de kans per tijdseenheid om naar toestand k te gaan. Dit betekent dat de kans om in toestand k te zijn niet verandert in de loop van de tijd; het systeem is in evenwicht.

Bovenstaande vergelijkingen (het zijn er vele, voor elke k één) worden de *evenwichtsvergelijkingen* genoemd. Je ziet direct dat $\pi_k = 0$ voor alle k een oplossing is van de evenwichtsvergelijkingen, maar deze doet natuurlijk niet mee, omdat de nulvector geen kansvector is.

Wat heb je hier aan?

Waarom zijn evenwichtsverdelingen nu zo belangrijk? Het antwoord is even simpel als prachtig: het blijkt zo te zijn dat als er een evenwichtsverdeling π_k bestaat, er geldt dat

$$\lim_{t \to \infty} P(Q(t) = k) = \pi_k.$$

We zeggen dat het systeem naar zijn eigen evenwichtstoestand convergeert.

Dit betekent dat als het systeem al een tijdje aan de gang is, de kans dat er k klanten in het systeem aanwezig zijn ongeveer gelijk is aan π_k. Reden genoeg dus om in ons model de evenwichtsvergelijkingen op te stellen en op te lossen. Het antwoord voor ons model blijkt te zijn:

$$\pi_k = (1 - \rho)\rho^k, \; k = 0, 1, 2, \ldots$$

Voor wie geïnteresseerd is in de afleiding stellen we de evenwichtsvergelijkingen voor ons wachtrijmodel op. Voor $k = 0$ geldt dat je alleen naar $k = 1$ kunt en je kunt ook alleen maar via 1 naar 0. De eerste vergelijking wordt dus

$$\lambda\pi_0 = \mu\pi_1. \qquad (2)$$

Voor andere toestanden $k \neq 0$ geldt dat je naar $k - 1$ en $k + 1$ kunt met rates μ en λ respectievelijk. Je kunt naar k komen via opnieuw $k - 1$ en $k + 1$ met rates λ en μ. De vergelijkingen worden dan

$$\pi_k(\lambda + \mu) = \lambda \pi_{k-1} + \mu \pi_{k+1}. \qquad (3)$$

Dit stelsel vergelijkingen kunnen we als volgt oplossen. Herschrijf (2) als $\pi_1 = \rho\pi_0$. Vergelijking (3) is te herschrijven als

$$\pi_{k+1} = (1 + \rho)\pi_k - \rho \pi_{k-1}. \qquad (4)$$

Als we $k = 1$ kiezen, dan volgt $\pi_2 = (1 + \rho)\pi_1 - \rho\pi_0 = (1 + \rho)\rho\pi_0 - \rho\pi_0 = \rho^2\pi_0$. Het lijkt er dus op dat $\pi_k = \rho^k\pi_0$. Voor $k = 1, 2$ hebben we dit net bewezen. Stel nu dat we

$$\pi_k = \rho^k\pi_0 \qquad (5)$$

bewezen hebben voor $k = n - 1$ en $k = n$. Dan vinden we met behulp van (3) dat

$$\begin{aligned} \pi_{n+1} &= (1 + \rho)\pi_n - \rho\pi_{n-1} \\ &= (1 + \rho)\rho^n\pi_0 - \rho\rho^{n-1}\pi_0 \\ &= \rho^{n+1}\pi_0. \end{aligned}$$

We hebben nu dus laten zien dat als (5) voor $k = n - 1$ en $k = n$ waar is, dat het dan ook waar is voor $k = n+1$. Aangezien (5) waar is voor $k = 1$ en $k = 2$, volgt nu dat het ook waar is voor $k = 3$. Hieruit volgt dan weer dat het ook waar is voor $k = 4$, enzovoort. Het is dus voor alle k waar. (Dit heet een bewijs met *volledige inductie*.)

Aangezien we nu een oplossing van de vorm $\pi_k = \rho^k\pi_0$ hebben, kunnen we voor π_0 een willekeurig getal invullen en dit geeft ons, zo lijkt het, vele oplossingen van de evenwichtsvergelijkingen. Maar we vergeten één ding, namelijk dat $\sum_{n=0}^{\infty} \pi_k$ gelijk moet zijn aan 1.

Deze eis vertelt ons dat $\pi_0 \sum_{n=0}^{\infty} \rho^k = \frac{1}{1 - \rho}$.

Aangezien $\rho < 1$ (dit is de enige plaats waar we dit gebruiken!) geldt dat $\sum_{k=0}^{\infty} \rho^k = \frac{1}{1 - \rho}$,

(dit is opnieuw de bekende meetkundige reeks) en we vinden dan dat $\pi_k = (1 - \rho)\rho^k, k = 0, 1, 2, \ldots$

Vanwege het feit dat het systeem naar zijn eigen evenwicht convergeert, mag je aannemen dat wanneer je naar het postkantoor gaat, het systeem in evenwicht is.

Nogmaals de drie vragen

Nu zijn de vragen uit het begin te beantwoorden:

1. *Wat is de kans om bij aankomst in het systeem precies k klanten voor je te treffen?*

Deze kans is de kans op k klanten in het systeem en dus gelijk aan $(1 - \rho)\rho^k$.

2. *Wat is de gemiddelde tijd die een klant moet wachten voor hij/zij aan de beurt is?*

Als we deze stochastische wachttijd even aangeven met W, is dit gemiddelde (aangeduid met EW; de E staat voor de Engelse term 'expectation') uit te rekenen door te conditioneren op het aantal klanten voor je. Bedenk dat als je k klanten voor je treft, de gemiddelde wachttijd gelijk is aan $\frac{k}{\mu}$, namelijk k maal de gemiddelde bedieningstijd van één klant.

$$EW = \sum_{k=0}^{\infty} P(k \text{ klanten voor je})\, E(W|k \text{ klanten voor je})$$

$$= \sum_{k=0}^{\infty} (1-\rho)\rho^k \frac{k}{\mu}$$

$$= \frac{\rho(1-\rho)}{\mu} \sum_{k=0}^{\infty} \rho^{k-1}$$

$$= \frac{\rho(1-\rho)}{\mu} \frac{d}{d\rho}\left(\sum_{k=0}^{\infty} \rho^k\right)$$

$$= \frac{\rho}{\mu(1-\rho)} = \frac{\rho}{\mu-\lambda}.$$

3. *Hoe lang zal de bediende gemiddeld onafgebroken bezig blijven met klanten vanaf het moment dat een klant binnenkomt in een leeg systeem?*

We zullen het antwoord op deze vraag alleen aannemelijk maken, het argument kan echter ook exact uitgewerkt worden. De evenwichtsverdeling kan geïnterpreteerd worden als de fractie van de tijd waarin het systeem zich in een bepaalde toestand bevindt. In het bijzonder zal tot en met tijdstip t (denk aan t als heel groot) het systeem ongeveer een tijdsduur $\pi_0 t = (1-\rho)t$ zonder klanten hebben doorgebracht. De gemiddelde lengte van zo'n periode is de verwachte tussenaankomsttijd en dus gelijk aan $\frac{1}{\lambda}$. Dat betekent dus dat er op tijdstip t ongeveer $\lambda(1-\rho)t$ van zulke 'lege' perioden geweest zijn. Als gevolg hiervan zijn er ook ongeveer zoveel perioden geweest waarin het systeem onafgebroken klanten bevatte. De totale tijdsduur doorgebracht in niet-lege toestand is ongeveer $(1-\pi_0)t = \rho t$. Dat betekent dat de lengte van één onafgebroken interval met klanten in het systeem gemiddeld ongeveer gelijk is aan

$$\frac{\rho t}{\lambda(1-\rho)t} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu-\lambda}.$$

Deze laatste uitdrukking is het gevraagde gemiddelde.

De wachttijdparadox

Deze paragraaf gaat over een heel merkwaardig fenomeen dat *wachttijdparadox* wordt genoemd. We hebben al eerder opgemerkt dat de gemiddelde tussenaankomst tussen twee klanten gelijk is aan $\frac{1}{\lambda}$. Neem nu een willekeurig tijdstip $t_0 > 0$. We zijn geïnteresseerd in de gemiddelde tussenaankomsttijd tussen de laatste klant vóór t_0 en de eerste klant ná t_0. Ik geef nu twee redeneringen die tot verschillende antwoorden leiden. Misschien is het aardig, voor u daarna verder leest, een keuze te maken uit de twee argumenten:

1. De tussenaankomst tussen de laatste klant vóór t_0 en de eerste klant ná t_0 is een gewone tussenaankomsttijd. Het gemiddelde van zulke tussenaankomsttijden is gelijk aan $\frac{1}{\lambda}$, dus het gemiddelde van deze tussenaankomsttijd is ook $\frac{1}{\lambda}$.
2. Als t_0 groot is, zal t_0 gemiddeld ongeveer halverwege een tussenaankomsttijd liggen. Omdat de wachttijd tot de volgende klant vanaf tijdstip t_0 gewoon weer exponentieel verdeeld is, is de gemiddelde wachttijd vanaf t_0 tot de *volgende* klant ongeveer $\frac{1}{\lambda}$. Dit is ongeveer de helft van het totale interval waar het om gaat, dus we concluderen dat de lengte van het hele interval ongeveer $\frac{1}{\lambda}$ is.

De eerste redenering is onjuist en dit fenomeen wordt ook wel de *wachttijdparadox* genoemd. De reden dat de eerste redenering fout is, is vrij subtiel. Het idee is dat grotere tussenaankomsttijden een grotere 'kans' hebben om t_0 in hun interval te 'vangen'. We kunnen dit duidelijker illustreren met een heel extreem voorbeeld. Stel ik heb een vreemd aankomstproces waarbij de tussenaankomsttijden gelijk zijn aan 1000 of 1, allebei met kans $\frac{1}{2}$. De gemiddelde tussenaankomsttijd is dan natuurlijk $500\frac{1}{2}$. Maar als je nu een tijdstip t_0 kiest, dan moet het wel heel toevallig zijn als t_0 in een tussenaankomstinterval ter lengte 1 ligt. Gemiddeld ligt $N(t_0)$ dan ook heel dicht bij 1000 (zie figuur 7).

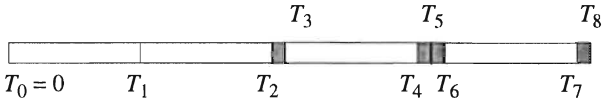

fig. 7 Een situatie met zeer grote en zeer kleine tussenaankomsttijden. Slechts een klein gedeelte van de tijdsbalk wordt overdekt door de kleine tussenaankomstintervalletjes.

Met fors rekenwerk is het mogelijk een exacte uitdrukking te vinden voor het gevraagde gemiddelde: dit blijkt gelijk te zijn aan

$$\frac{1}{\lambda}\left(2 - e^{-\lambda t_0}\right).$$

De limiet $t_0 \to \infty$ van deze uitdrukking is gelijk aan $\frac{2}{\lambda}$. Als $\lambda = 2$ dan is de verwachting van een 'gewone' tussenaankomsttijd gelijk aan $\frac{1}{2}$, maar voor t_0 heel groot is de verwachte tussenaankomsttijd tussen de laatste klant vóór t_0 en de eerste klant ná t_0 ongeveer gelijk aan 1...

Realistischer modellen

Misschien heeft u zich bij het lezen van dit artikel afgevraagd hoe realistisch het model eigenlijk is. Laten we eens een paar problemen noemen:

- De instroom van klanten in een postkantoor is niet altijd even groot. Tussen 12.00 en 13.00 zal het waarschijnlijk veel drukker zijn dan tussen 10.00 en 11.00 uur.

Als we deze stochastische wachttijd even aangeven met W, is dit gemiddelde (aangeduid met EW; de E staat voor de Engelse term 'expectation') uit te rekenen door te conditioneren op het aantal klanten voor je. Bedenk dat als je k klanten voor je treft, de gemiddelde wachttijd gelijk is aan $\frac{k}{\mu}$, namelijk k maal de gemiddelde bedieningstijd van één klant.

$$EW = \sum_{k=0}^{\infty} P(k \text{ klanten voor je})\, E(W|k \text{ klanten voor je})$$

$$= \sum_{k=0}^{\infty} (1-\rho)\rho^k \frac{k}{\mu}$$

$$= \frac{\rho(1-\rho)}{\mu} \sum_{k=0}^{\infty} \rho^{k-1}$$

$$= \frac{\rho(1-\rho)}{\mu} \frac{d}{d\rho}\left(\sum_{k=0}^{\infty} \rho^k\right)$$

$$= \frac{\rho}{\mu(1-\rho)} = \frac{\rho}{\mu-\lambda}.$$

3. *Hoe lang zal de bediende gemiddeld onafgebroken bezig blijven met klanten vanaf het moment dat een klant binnenkomt in een leeg systeem?*

We zullen het antwoord op deze vraag alleen aannemelijk maken, het argument kan echter ook exact uitgewerkt worden. De evenwichtsverdeling kan geïnterpreteerd worden als de fractie van de tijd waarin het systeem zich in een bepaalde toestand bevindt. In het bijzonder zal tot en met tijdstip t (denk aan t als heel groot) het systeem ongeveer een tijdsduur $\pi_0 t = (1-\rho)t$ zonder klanten hebben doorgebracht. De gemiddelde lengte van zo'n periode is de verwachte tussenaankomsttijd en dus gelijk aan $\frac{1}{\lambda}$. Dat betekent dus dat er op tijdstip t ongeveer $\lambda(1-\rho)t$ van zulke 'lege' perioden geweest zijn. Als gevolg hiervan zijn er ook ongeveer zoveel perioden geweest waarin het systeem onafgebroken klanten bevatte. De totale tijdsduur doorgebracht in niet-lege toestand is ongeveer $(1-\pi_0)t = \rho t$. Dat betekent dat de lengte van één onafgebroken interval met klanten in het systeem gemiddeld ongeveer gelijk is aan

$$\frac{\rho t}{\lambda(1-\rho)t} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu-\lambda}.$$

Deze laatste uitdrukking is het gevraagde gemiddelde.

De wachttijdparadox

Deze paragraaf gaat over een heel merkwaardig fenomeen dat *wachttijdparadox* wordt genoemd. We hebben al eerder opgemerkt dat de gemiddelde tussenaankomst tussen twee klanten gelijk is aan $\frac{1}{\lambda}$. Neem nu een willekeurig tijdstip $t_0 > 0$. We zijn geïnteresseerd in de gemiddelde tussenaankomsttijd tussen de laatste klant vóór t_0 en de eerste klant ná t_0. Ik geef nu twee redeneringen die tot verschillende antwoorden leiden. Misschien is het aardig, voor u daarna verder leest, een keuze te maken uit de twee argumenten:

1. De tussenaankomst tussen de laatste klant vóór t_0 en de eerste klant ná t_0 is een gewone tussenaankomsttijd. Het gemiddelde van zulke tussenaankomsttijden is gelijk aan $\frac{1}{\lambda}$, dus het gemiddelde van deze tussenaankomsttijd is ook $\frac{1}{\lambda}$.
2. Als t_0 groot is, zal t_0 gemiddeld ongeveer halverwege een tussenaankomsttijd liggen. Omdat de wachttijd tot de volgende klant vanaf tijdstip t_0 gewoon weer exponentieel verdeeld is, is de gemiddelde wachttijd vanaf t_0 tot de *volgende* klant ongeveer $\frac{1}{\lambda}$. Dit is ongeveer de helft van het totale interval waar het om gaat, dus we concluderen dat de lengte van het hele interval ongeveer $\frac{1}{\lambda}$ is.

De eerste redenering is onjuist en dit fenomeen wordt ook wel de *wachttijdparadox* genoemd. De reden dat de eerste redenering fout is, is vrij subtiel. Het idee is dat grotere tussenaankomsttijden een grotere 'kans' hebben om t_0 in hun interval te 'vangen'. We kunnen dit duidelijker illustreren met een heel extreem voorbeeld. Stel ik heb een vreemd aankomstproces waarbij de tussenaankomsttijden gelijk zijn aan 1000 of 1, allebei met kans $\frac{1}{2}$. De gemiddelde tussenaankomsttijd is dan natuurlijk $500\frac{1}{2}$. Maar als je nu een tijdstip t_0 kiest, dan moet het wel heel toevallig zijn als t_0 in een tussenaankomstinterval ter lengte 1 ligt. Gemiddeld ligt $N(t_0)$ dan ook heel dicht bij 1000 (zie figuur 7).

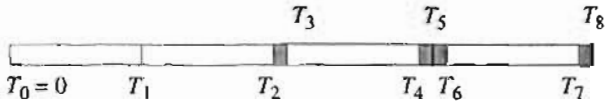

fig. 7 Een situatie met zeer grote en zeer kleine tussenaankomsttijden. Slechts een klein gedeelte van de tijdsbalk wordt overdekt door de kleine tussenaankomstintervalletjes.

Met fors rekenwerk is het mogelijk een exacte uitdrukking te vinden voor het gevraagde gemiddelde: dit blijkt gelijk te zijn aan

$$\frac{1}{\lambda}\left(2 - e^{-\lambda t_0}\right).$$

De limiet $t_0 \to \infty$ van deze uitdrukking is gelijk aan $\frac{2}{\lambda}$. Als $\lambda = 2$ dan is de verwachting van een 'gewone' tussenaankomsttijd gelijk aan $\frac{1}{2}$, maar voor t_0 heel groot is de verwachte tussenaankomsttijd tussen de laatste klant vóór t_0 en de eerste klant ná t_0 ongeveer gelijk aan 1...

Realistischer modellen

Misschien heeft u zich bij het lezen van dit artikel afgevraagd hoe realistisch het model eigenlijk is. Laten we eens een paar problemen noemen:

- De instroom van klanten in een postkantoor is niet altijd even groot. Tussen 12.00 en 13.00 zal het waarschijnlijk veel drukker zijn dan tussen 10.00 en 11.00 uur.

- Als het drukker is, zal de bediende waarschijnlijk harder gaan werken, waardoor de bedieningstijden niet meer allemaal dezelfde verdeling hebben.
- Als het erg druk is, zullen aankomende klanten wellicht besluiten dat ze liever eerst boodschappen gaan doen in de hoop dat het later rustiger zal zijn.
- Als een klant heel veel bedieningstijd vraagt, zal de bediende misschien wel harder werken. De bedieningstijden krijgen dan een andere verdeling.
- Meestal zullen er meerdere bedienden zijn. De wachtrij kan individueel, per bediende, zijn of gemeenschappelijk, zoals tegenwoordig in postkantoren het geval is.

Het is niet al te moeilijk deze lijst met een groot aantal andere problemen uit te breiden. Ons model houdt met al deze zaken geen rekening en het is dus maar de vraag of ons model enige praktische waarde heeft.
In veel gevallen is het mogelijk het wiskundig model zo uit te breiden dat wel rekening wordt gehouden met deze zaken. Dit leidt echter vaak tot wiskundige complicaties. Deze afweging moet vaak gemaakt worden: als het model realistischer wordt, wordt in het algemeen de wiskundige analyse moeilijker. Ik sluit dit artikel af met twee voorbeelden van dergelijke uitbreidingen.

Andere bedieningstijden

Als we de bedieningstijden niet meer exponentieel verdeeld nemen, dan is het stochastisch proces $(Q(t);\ t \geq 0)$ geen Markovproces meer en dat betekent dat de evenwichtsvergelijkingen geen betekenis meer hebben. In zulke gevallen is geavanceerdere wiskunde nodig om toch nog iets exacts te kunnen zeggen over het model. In dit geval blijkt het opnieuw zo te zijn dat het systeem naar een evenwicht convergeert zolang de gemiddelde tussenaankomsttijd groter is dan de gemiddelde bedieningstijd.

Klanten die weggaan bij grote drukte

Stel dat een aankomende klant eerst telt hoeveel klanten er voor hem/haar in de rij staan. Als dat er n zijn, dan verlaat hij de rij met kans $\frac{1}{n+1}$ en komt niet meer terug. Natuurlijk is ook dit model niet heel realistisch, maar in ieder geval betekent het dat er bij grotere drukte minder klanten binnenkomen. In dit model kunnen we weer de evenwichtsverdeling uitrekenen. Deze blijkt gelijk te zijn aan

$$\pi_k = \frac{\rho^k e^{-\rho}}{k!}$$

voor *alle* waarden van ρ, dus ook voor $\rho \geq 1$. Er bestaat dus een evenwichtsverdeling voor alle waarden van ρ. Dit komt doordat het systeem 'zichzelf in de hand houdt' in die zin dat er automatisch minder klanten binnenkomen als het erg druk is. Op deze manier zorgt het systeem er zelf voor dat de lengte van de wachtrij gemiddeld niet al te groot wordt.

De evenwichtsvergelijkingen in dit nieuwe model worden, analoog aan (2) en (4), gegeven door

$$\pi_1 = \rho\,\pi_0$$

en

$$\pi_{k+1} = \left(1 + \frac{\rho}{k+1}\pi_k\right) - \frac{\rho}{k-1}\pi_{k-1}$$

voor $k = 1, 2, \ldots$
Deze laatste vergelijking voor $k = 1$ levert $\pi_2 = \frac{1}{2}\rho^2\pi_0$, en voor $k = 2$ wordt het $\pi_3 = \frac{1}{6}\rho^3\pi_0$.

Met volledige inductie is nu te bewijzen dat $\pi_n = \frac{1}{n!}\rho^n\pi_0$.

Omdat alles bij elkaar op moet tellen tot 1, vinden we dan dat

$$\pi_0 \sum_{n=0}^{\infty} \frac{1}{n!}\rho^n = \pi_0 e^{\rho}$$

gelijk moet zijn aan 1. Dit betekent dat $\pi_0 = e^{-\rho}$ en de evenwichtsverdeling is gevonden.

Ik wil Corrie Quant (en haar vader) en Hansje Huson bedanken voor nuttig commentaar op eerdere versies van dit artikel.

Ronald Meester is verbonden aan de Vakroep Wiskunde van de Faculteit Wiskunde & Informatica van de Universiteit Utrecht.