Kansrekening voor Informatiekunde (I00051)

Bernd Souvignier

voorjaar 2005

Inhoud

Les 1 Combinatoriek

Als we het over de *kans* hebben dat iets gebeurt, hebben we daar wel intuïtief een idee over, wat we hiermee bedoelen. Bijvoorbeeld zeggen we, dat bij het werpen van een munt de kans $\frac{1}{2}$ is, dat de zijde met cijfer (munt) boven te liggen komt, evenzo als de kans voor de koningin (kop) $\frac{1}{2}$ is. Op een soortgelijke manier behandelen we het werpen van een dobbelsteen: de kans voor elke van de getallen $1, 2, 3, 4, 5, 6$ is $\frac{1}{6}$, maar we kunnen ook iets over de kans zeggen, dat we een even getal werpen, die is namelijk de som van de kansen voor 2, 4 en 6, dus $\frac{1}{2}$.

Het algemeen principe dat hier achter zit, is dat er een aantal mogelijke uitkomsten is, en we een deel hiervan als *gunstige* uitkomsten aanzien. De relatieve frequentie van gunstige uitkomsten interpreteren we dan als kans voor een gunstige uitkomst.

Principe van de relatieve frequentie: *De kans op een gunstige uitkomst berekenen we als het aantal gunstige uitkomsten gedeeld door het totaal aantal mogelijke uitkomsten.*

Het Simpson paradox

Soms kan zelfs het bepalen van kansen met behulp van relatieve frequenties tot verrassingen leiden. Stel we hebben een fruithandelaar die sinaasappels van minstens $100g$ per stuk wil verkopen. Hij heeft twee leveranciers, A en B, van sinaasappels.

In een eerste levering krijgt hij van A 110 sinaasappels waarvan er 50 te licht zijn en van B 70 sinaasappels waarvan 30 te licht zijn. Op dit moment zou hij ervan uit gaan dat B de betere leverancier is, omdat $\frac{5}{11} > \frac{3}{7}$ is.

Een week later krijgt hij van A een levering van 90 sinaasappels waarvan 60 te licht zijn en van B 140 sinaasappels waarvan 90 te licht zijn. Ook in dit geval is B de betere leverancier, want $\frac{6}{9} > \frac{9}{14}$.

Maar als we nu de twee leveringen bij elkaar nemen, waren bij A 110 van 200 sinaasappels te licht, terwijl bij B 120 van 210 sinaasappels te licht waren. Er geldt $\frac{11}{20} < \frac{12}{21}$, dus is over de twee weken gezien A de betere leverancier!

Het probleem is, dat we uit de twee leveringen apart kunnen concluderen dat $\frac{5}{11} + \frac{6}{9} > \frac{3}{7} + \frac{9}{14}$. Maar als we de leveringen gezamenlijk vergelijken, moeten we $\frac{5+6}{11+9}$ met $\frac{3+9}{7+14}$ vergelijken, en dat is niet de som van de breuken.

1.1 Tellen van uitkomsten

Om goed over kansen en kansverdelingen te kunnen praten, moeten we kijken, hoe we bij iets ingewikkeldere problemen dan het werpen van een dobbelsteen gunstige uitkomsten kunnen tellen. De kunst van het tellen van uitkomsten heet *combinatoriek*.

Bij het dobbelen met drie dobbelstenen kunnen we ons afvragen of de kans groter is dat de som van de ogen 11 of 12 is. Hiervoor moeten we 11 of 12 schrijven als sommen van drie getallen uit de verzameling $\{1, 2, 3, 4, 5, 6\}$. We

hebben

$$11 = 6+4+1 = 6+3+2 = 5+5+1 = 5+4+2 = 5+3+3 = 4+4+3$$

$$12 = 6+5+1 = 6+4+2 = 6+3+3 = 5+5+2 = 5+4+3 = 4+4+4$$

dus zijn er in elk geval 6 mogelijkheden en de kans lijkt even groot te zijn. Maar als we dit in een experiment na gaan (bijvoorbeeld met een computersimulatie), zien we dat de kans voor de som 11 ongeveer $P(11) = 0.125$ is en de kans voor de som 12 ongeveer $P(12) = 0.116$, dus kleiner dan die voor de som 11. Wat is hier mis gegaan?

Bij het tellen van de mogelijkheden hebben we alleen maar afstijgende sommen opgeschreven, maar als we even aannemen dat de drie dobbelstenen rood, blauw en groen zijn, is het duidelijk dat er verschillende manieren zijn, hoe we $6+4+1$ kunnen krijgen. De 6 kan namelijk op elke van de drie dobbelstenen verschijnen en in elk van deze drie gevallen hebben we nog twee mogelijkheden om 4 en 1 op de andere twee dobbelstenen te verdelen. We moeten dus de som $6+4+1$ zes keer tellen, omdat er zes verschillende manieren zijn hoe we deze som kunnen krijgen. Bij een som met twee verschillende getallen (zo als $5+5+1$) hebben we drie mogelijkheden en bij drie dezelfde getallen alleen maar eentje. Als we de mogelijkheden voor de som 11 zo bepalen vinden we $3\cdot 6+3\cdot 3 = 27$ mogelijkheden en voor de som 12 krijgen we $3\cdot 6+2\cdot 3+1 = 25$. Omdat er $6^3 = 216$ mogelijke uitkomsten met drie dobbelstenen zijn, is de kans voor de som 11 dus $\frac{27}{216} = \frac{1}{8}$ en die voor som 12 is $\frac{25}{216}$, en dit is wat we ook experimenteel zouden vinden.

Het belangrijke punt bij dit voorbeeld is, dat we de dobbelstenen kunnen onderscheiden en dat we daarom op de volgorde van de resultaten moeten letten. Het is afhankelijk van het experiment of we inderdaad op de volgorde willen letten of niet. Bijvoorbeeld zijn we bij een qualiteitscontrole alleen maar geïnteresseerd hoeveel slechte stukken we in een steekproef hebben, maar niet of de eerste of de laatste in de steekproef slecht is.

1.2 Geordende grepen

We gaan eerst na hoe we het aantal uitkomsten berekenen als de volgorde een rol speelt, dus als we het resultaat van de eerste greep en het resultaat van de tweede greep willen onderscheiden. Dit is bijvoorbeeld het geval voor het dobbelen met meerdere dobbelstenen, maar ook voor het toewijzen van nummers aan de spelers van een voetbalploeg.

Hier is een voorbeeld: Stel een exclusieve restaurant biedt een keuze van 4 voorgerechten, 3 hoofdgerechten en 3 desserts. Je mag elke combinatie van de drie gangen kiezen, hoeveel mogelijke menu's kun je dan bestellen? Het is duidelijk dat je $4\cdot 3\cdot 3$ mogelijkheden hebt. Algemeen geldt:

Principe van de vermenigvuldiging van uitkomsten: *Het aantal uitkomsten voor een geordende greep is* $\prod_{i=1}^{r} n_i$ *als we* r *keer trekken en er voor de* i*-de greep* n_i *mogelijkheden zijn.*

Van dit principe zijn er twee heel belangrijke speciale gevallen, het trekken *met* en het trekken *zonder* terugleggen.

Trekken met terugleggen

Uit een verzameling van n objecten kiezen we r keer een element, waarbij we het getrokken element weer terugleggen. Dan hebben we voor elke keuze n mogelijkheden en het aantal uitkomsten is dus

$$\underbrace{n \cdot n \cdot \ldots \cdot n}_{r} = n^r.$$

Dit is het aantal rijen $(a_1, \ldots, a_r)$ met $a_i \in \{1, \ldots, n\}$.

Trekken zonder terugleggen

Uit een verzameling van n objecten kiezen we r keer een element, maar een getrokken element wordt niet terug gelegd, dus is er na elke greep een element minder in de verzameling. Voor de eerste greep hebben we dus n mogelijkheden, voor de tweede $n-1$, voor de derde $n-2$ enzovoorts. Het aantal uitkomsten is dus

$$n \cdot (n-1) \cdot \ldots \cdot (n-r+1) = \frac{n!}{(n-r)!}.$$

Dit is het aantal rijen $(a_1, \ldots, a_r)$ met $a_i \in \{1, \ldots, n\}$ waarbij alle a_i verschillend zijn. In het bijzonder geldt:

Permutaties van n elementen: *Het aantal manieren hoe we de getallen* $\{1, \ldots, n\}$ *kunnen ordenen is gelijk aan* $n!$.

1.3 Ongeordende grepen

Bij veel toepassingen speelt de volgorde geen rol, bijvoorbeeld als we alleen maar geïnteresseerd zijn hoeveel objecten met een bepaalde eigenschap in een steekproef zitten. Als de volgorde geen rol speelt, kunnen we de elementen in de rij van getrokken elementen omordenen en zo ervoor zorgen dat ze in een zekere volgorde zitten. Op die manier zijn de uitkomsten van een ongeordende greep alleen maar de rijen $(a_1, \ldots, a_r)$ met $a_i \leq a_{i+1}$.

Merk op: Hier ligt een bron van mogelijke verwarring : Bij een *ongeordende greep* mogen we de elementen *omordenen* en krijgen dan een *geordende* rij.

Ook voor de ongeordende grepen zijn er weer twee mogelijkheden: We kunnen met of zonder terugleggen trekken. Omdat het geval zonder terugleggen eenvoudiger is, gaan we dit eerst bekijken.

Trekken zonder terugleggen

Het misschien meest bekende voorbeeld van een ongeordende greep zonder terugleggen is het trekken van de lottogetallen. Hierbij worden de ballen met de nummers weliswaar achter elkaar getrokken en we kunnen de ballen ook onderscheiden, maar op het eind worden de nummers in opstijgende volgorde gesorteerd, daarom speelt het geen rol in welke volgorde de nummers getrokken werden en de greep is dus ongeordend.

We hebben gezien, dat er $\frac{n!}{(n-r)!}$ mogelijke uitkomsten van een geordende greep zonder terugleggen zijn. Maar van zo'n greep zijn er precies $r!$ permutaties en alleen maar één van deze permutaties heeft de eigenschap dat de elementen opstijgend geordend zijn. Dus is het aantal uitkomsten voor ongeordende grepen zonder terugleggen

$$\frac{1}{r!} \cdot \frac{n!}{(n-r)!} = \frac{n!}{r!(n-r)!} =: \binom{n}{r}.$$

We noemen $\binom{n}{r}$ een *binomiaalcoëfficiënt* en spreken dit 'n over r'. De binomiaalcoëfficiënt $\binom{n}{r}$ geeft aan op hoeveel manieren we een deelverzameling van r elementen uit een verzameling van n elementen kunnen kiezen. Dit is hetzelfde als het aantal rijen $(a_1, \ldots, a_r)$ met $a_i \in \{1, \ldots, n\}$ en $a_i < a_{i+1}$. Merk op dat de binomiaalcoëfficiënt $\binom{n}{r} = 0$ voor $r > n$, omdat we geen r elementen uit $n < r$ kunnen kiezen.

In het geval van de lottogetallen is $n = 49$ en $r = 6$ (we negeren even extra- en supergetallen), dus is het aantal mogelijke uitkomsten van de lotto $\binom{n}{r} = 13983816$, dus bijna 14 miljoen.

Een andere samenhang waar we de binomiaalcoëfficiënt tegen komen (en waar ook de naam vandaan komt), is bij veeltermen: De (algemene) binomische formule is

$$(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r} = a^n + \binom{n}{1} a^{n-1}b + \binom{n}{2} a^{n-2}b^2 + \ldots + \binom{n}{n-1} ab^{n-1} + b^n$$

dus bijvoorbeeld $(a+b)^4 = b^4 + 4ab^3 + 6a^2b^2 + 4a^3b + a^4$. Het is geen toeval dat de binomiaalcoëfficiënt hier naar voren komt: Als we het product $(a+b)^n$ uitschrijven als $(a+b) \cdot (a+b) \cdot \ldots \cdot (a+b)$ en dan uitvoerig vermenigvuldigen krijgen we een term $a^r b^{n-r}$ als we in r van de factoren a kiezen en in de $n-r$ andere factoren b. Maar het aantal manieren om de r factoren met a uit de n factoren te kiezen is $\binom{n}{r}$, daarom wordt dit de coëfficiënt van $a^r b^{n-r}$.

We kunnen makkelijk een paar belangrijke eigenschappen van de binomiaalcoëfficiënten afleiden:

(i) $\binom{n}{r} = \binom{n}{n-r}$

Dit volgt meteen uit de definitie, omdat we alleen maar de factoren in de noemer omruilen. Maar we kunnen het ook anders inzien: Als we r uit de n elementen van een verzameling hebben gekozen, dan hebben we $n-r$ elementen niet gekozen, dus hoort bij elke deelverzameling van r elementen een eenduidige deelverzameling van $n-r$ elementen, dus is het aantal deelverzamelingen met r elementen gelijk aan het aantal deelverzamelingen met $n-r$ elementen. We noemen dit ook de *symmetrie* van de binomiaalcoëfficiënten.

(ii) $\sum_{r=0}^{n} \binom{n}{r} = 2^n$

Dit volgt uit de binomische formule als we $a = b = 1$ invullen. Maar we

kunnen dit ook uit het aftellen van deelverzamelingen zien: Een verzameling Ω van n elementen heeft $\binom{n}{r}$ deelverzamelingen met r elementen, dus is de som over de binomiaalcoëfficiënten het aantal van alle deelverzamelingen van Ω. Maar elk element $a \in \Omega$ is of in een deelverzameling $A \subseteq \Omega$ bevat of is er niet in bevat. Dit geeft 2 mogelijkheden voor elk element en dus 2^n mogelijkheden om de uitkomsten $a \in A$ of $a \notin A$ op de n elementen van Ω te verdelen en dus zijn er 2^n deelverzamelingen van Ω.

(iii) $\binom{n}{r-1} + \binom{n}{r} = \binom{n+1}{r}$
Hiervoor tellen we de $\binom{n+1}{r}$ deelverzamelingen $A \subseteq \{1, \ldots, n+1\}$ met r elementen op de volgende manier: Of het element $n+1$ ligt in een deelverzameling A, dan bevat A nog $r-1$ elementen uit de resterende n elementen en er zijn dus $\binom{n}{r-1}$ mogelijkheden voor A. Of het element $n+1$ zit niet in de deelverzameling A, dan zijn de r elementen van A uit de resterende n elementen gekozen en hiervoor zijn er $\binom{n}{r}$ mogelijkheden.

Een handige manier om de binomiaalcoëfficiënten op te schrijven (en uit te rekenen) is de *driehoek van Pascal* die in Figuur 1 afgebeeld is. In de driehoek van Pascal heeft de eerste rij één element, de tweede heeft twee elementen enz., de n-de rij heeft dus n elementen. Als r-de element in de n-de rij schrijven we de binomiaalcoëfficiënt $\binom{n-1}{r-1}$. Merk op dat $\binom{0}{0} = 1$ omdat $0! = 1$ is. De formule $\binom{n}{r-1} + \binom{n}{r} = \binom{n+1}{r}$ zegt nu dat we een element op een zekere plek in de driehoek van Pascal krijgen door de twee direct links en rechts boven dit element staande binomiaalcoëfficiënten op te tellen zo als in Figuur 1 voor het element $\binom{6}{2}$ aangetoond.

$\binom{0}{0}$	1
$\binom{1}{0}\ \binom{1}{1}$	1 1
$\binom{2}{0}\ \binom{2}{1}\ \binom{2}{2}$	1 2 1
$\binom{3}{0}\ \binom{3}{1}\ \binom{3}{2}\ \binom{3}{3}$	1 3 3 1
$\binom{4}{0}\ \binom{4}{1}\ \binom{4}{2}\ \binom{4}{3}\ \binom{4}{4}$	1 4 6 4 1
$\binom{5}{0}\ \binom{5}{1}\ \binom{5}{2}\ \binom{5}{3}\ \binom{5}{4}\ \binom{5}{5}$	1 5 10 10 5 1
$\binom{6}{0}\ \binom{6}{1}\ \binom{6}{2}\ \binom{6}{3}\ \binom{6}{4}\ \binom{6}{5}\ \binom{6}{6}$	1 6 15 20 15 6 1

Figuur 1: Driehoek van Pascal

Trekken met terugleggen

Als we na een greep het getrokken element weer terugleggen maar niet op de volgorde letten, willen we het aantal rijen $(a_1, \ldots, a_r)$ bepalen met $a_i \in$

$\{1, \ldots, n\}$ en $a_i \leq a_{i+1}$. Merk op dat we het aantal van dit soort rijen niet zo makkelijk uit het aantal van geordende rijen kunnen bepalen, omdat het aantal permutaties van een rij met herhalingen ervan afhangt hoeveel elementen hetzelfde zijn.

Maar hier komen we met een trucje en het resultaat voor het trekken zonder terugleggen verder: Stel we hebben een rij $(a_1, \ldots, a_r)$ met $a_i \leq a_{i+1}$, dan kunnen we hieruit een rij zonder herhalingen maken door $(i-1)$ bij het element a_i op te tellen. Dit geeft de rij $(b_1, \ldots, b_r)$ waarbij $b_i = a_i + i - 1 < a_{i+1} + i = b_{i+1}$. Voor de elementen b_i geldt $1 \leq b_i \leq n + r - 1$, dus hoort deze rij bij een ongeordende greep zonder terugleggen uit $n + r - 1$ elementen.

Omgekeerd kunnen we uit elke rij $(b_1, \ldots, b_r)$ met $b_i < b_{i+1}$ door aftrekken van $(i-1)$ van het element b_i een rij $(a_1, \ldots, a_r)$ maken met $a_i \leq a_{i+1}$. We zien dus dat er even veel rijen $(a_1, \ldots, a_r)$ zijn met $1 \leq a_i \leq n$ en $a_i \leq a_{i+1}$ als er rijen $(b_1, \ldots, b_r)$ zijn met $1 \leq b_i \leq n + r - 1$ en $b_i < b_{i+1}$. Maar we hebben gezien dat het aantal van het laatste soort rijen gelijk is aan

$$\binom{n+r-1}{r}$$

dus is dit ook het aantal van ongeordende r-grepen met terugleggen.

We hebben nu vier soorten van grepen gezien, namelijk geordende en ongeordende grepen die we telkens met of zonder terugleggen kunnen bekijken. Dit kunnen we overzichtelijk in een 2×2-schema beschrijven:

	geordend	ongeordend
met terugleggen	I	III
zonder terugleggen	II	IV

Deze vier gevallen kunnen we als volgt karakteriseren:

I: Noteer de uitslag van elke greep en leg terug $\Rightarrow n^r$ uitkomsten.

II: Noteer de uitslag van elke greep en leg niet terug $\Rightarrow \frac{n!}{(n-r)!} = \binom{n}{r} r!$ uitkomsten.

III: Noteer voor elke $a \in \Omega$ alleen maar het aantal grepen die a opleveren en leg terug $\Rightarrow \binom{n+r-1}{r}$ uitkomsten.

IV: Noteer voor elke $a \in \Omega$ alleen maar het aantal grepen die a opleveren en leg niet terug $\Rightarrow \binom{n}{r}$ uitkomsten.

Het Verjaardagsparadox

We willen de kans berekenen, dat er in en groep van r mensen twee mensen op dezelfde dag jarig zijn. Als verzameling nemen we de verzameling van verjaardagen, dus $|\Omega| = 365$ (we nemen aan dat niemand op 29 februari jarig is). Voor het aantal mogelijke uitkomsten zijn we in geval I, omdat we de mensen kunnen onderscheiden, dus het aantal is 365^r. Nu gebruiken we een klein trucje: We bepalen de kans van het complement van de gewenste uitkomst, dus

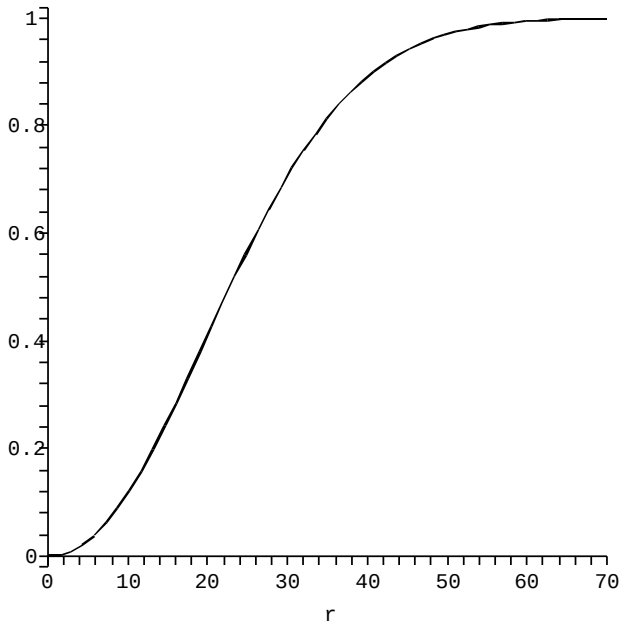
we bepalen de kans dat alle r mensen verschillende verjaardagen hebben. Dan zijn we voor de gunstige uitkomsten in geval II, want een verjaardag van één persoon mag niet meer het verjaardag van een andere persoon zijn. Er zijn dus $\binom{365}{r}r!$ gunstige uitkomsten (d.w.z. alle verjaardagen zijn verschillend). Bij elkaar genomen is de kans dat twee mensen op dezelfde dag jarig zijn dus

$$p = 1 - \frac{\binom{365}{r}r!}{365^r}.$$

Hier zijn een paar waarden van p voor verschillende grootten r van de groep:

$$\begin{aligned} r = 2 &\Rightarrow p = 0.003, & r = 5 &\Rightarrow p = 0.027, \\ r = 10 &\Rightarrow p = 0.117, & r = 15 &\Rightarrow p = 0.253, \\ r = 20 &\Rightarrow p = 0.411, & r = 23 &\Rightarrow p = 0.507, \\ r = 25 &\Rightarrow p = 0.569, & r = 30 &\Rightarrow p = 0.706, \\ r = 50 &\Rightarrow p = 0.970, & r = 70 &\Rightarrow p = 0.999. \end{aligned}$$

In Figuur 2 zie je de functie, die de kans op twee mensen met dezelfde verjaardag afhankelijk van de grootte r van de groep aangeeft. Omdat veel mensen het verrassend vinden dat de kans al voor $r = 23$ groter dan 0.5 is, noemt men dit ook het *verjaardagsparadox*. Er laat zich aantonen dat in het algemeen voor $r \approx \sqrt{n}$ geldt dat r grepen uit n objecten met kans $\frac{1}{2}$ twee dezelfde resultaten opleveren.



Figuur 2: Kans op dezelfde verjaardag bij r mensen

Belangrijke begrippen in deze les

- relatieve frequentie
- permutaties van n elementen
- geordende en ongerordende grepen
- grepen met en zonder terugleggen
- binomiaalcoëfficiënt
- verjaardagsparadox

Opgaven

1. We dobbelen met twee dobbelstenen. Bepaal de kansen voor de volgende uitkomsten:

   (i) De som van de twee getallen is 5.
   (ii) Beide dobbelstenen tonen een oneven getal.
   (iii) De eerste dobbelsteen toont een kleiner getal dan de tweede.
   (iv) De som van de twee getallen is even.
   (v) De som van de twee getallen is minstens 4.
   (vi) De som van de twee getallen is of even of minstens 4 (of allebei).

   De absolute waarde van het verschil van de twee getallen ligt tussen 0 en 5. Geef de kansverdeling $P(k)$ aan dat bij een worp met twee dobbelstenen de absolute waarde van het verschil precies k is.

2. Bij het *Poker* spel krijg je 5 kaarten uit een kaartspel met 52 kaarten. Verschillende combinaties van kaarten hebben een bijzondere waarde:

   (i) tweeling: twee kaarten van dezelfde soort (bijvoorbeeld twee boeren),
   (ii) dubbele tweeling: twee verschillende tweelingen (bijvoorbeeld twee vrouwen en twee azen),
   (iii) drieling: drie kaarten van dezelfde soort,
   (iv) vierling: vier kaarten van dezelfde soort,
   (v) full house: een tweeling en een drieling,
   (vi) straight: vijf kaarten in de goede volgorde (bijvoorbeeld 9, 10, boer, vrouw, heer),
   (vii) straight flush: een straight van dezelfde kleur.

   Bepaal voor elke van deze combinaties de kans en breng de combinaties hierdoor in een volgorde van opstijgende waarde.

3. Bij het skaat spel krijg je 10 kaarten uit een kaartspel met 32 kaarten. Om een solo te spelen wil je van één kleur veel kaarten hebben en een andere kleur helemaal niet. We vereenvoudigen het probleem door de boeren als gewone kaarten te tellen. Hoe groot is de kans dat je een blad krijgt waarbij je van elke kleur minstens 2 kaarten hebt (zo iets is meestal een slecht blad)?

   Als je een solo speelt, krijg je er nog 2 kaarten bij en leg je aansluitend weer 2 van je 12 kaarten weg. Hoe groot is de kans, dat je nu alleen nog met 2 kleuren zit (wat meestal een goed blad is)?

4. Een groep van 18 personen verdeeld zich in een restaurant over drie tafels van 4, 6 en 8 plaatsen. Hoeveel verschillende arrangements zijn er, als de plaatsing aan een tafel geen rol speelt?

5. Je spreekt met een vriend af om op de volgende dag in de rij te staan om kaarten voor Bruce Springsteen (of AC/DC of David Helfgott) te kopen. Op een gegeven moment staan jullie allebei in de rij, maar hebben elkaar niet gezien. Hoe groot is de kans, dat in een rij van n mensen precies r mensen tussen jullie staan? Hoe groot is de kans dat jullie elkaar kunnen zien als er 1000 mensen in de rij staan en je aanneemt dat je je vriend onder de 100 mensen naast je kunt herkennen?

Les 2 Kansverdelingen

We hebben in het begin gesteld dat we de kans voor een zekere gunstige uitkomst berekenen als het aantal gunstige uitkomsten gedeelt door het totaal aantal mogelijke uitkomsten. Maar vaak is het handig, dat we verschillende uitkomsten samenvatten en dit als een nieuwe soort uitkomst bekijken. Bijvoorbeeld kunnen we bij het werpen van twee dobbelstenen de som van de twee geworpen getallen als uitkomst nemen. Als we met $P(s)$ de kans op de som s noteren, zien we (door de mogelijke gevallen na te gaan) makkelijk in, dat $P(2) = 1/36$,P(3) $= 2/36$,P(4) $= 3/36$,P(5) $= 4/36$,P(6) $= 5/36$,P(7) $= 6/36$, $P(8) = 5/36$, $P(9) = 4/36$, $P(10) = 3/36$, $P(11) = 2/36$, $P(12) = 1/36$. Hieruit laat zich bijvoorbeeld snel aflezen, dat de kans op het dobbelen van een som die een priemgetal is, gelijk is aan $(1 + 2 + 4 + 6 + 2)/36 = 5/12$.

Om ook voor dit soort algemenere situaties over kansen te kunnen praten, hebben we een algemener begrip dan de relatieve frequenties nodig, namelijk het begrip van een *kansverdeling*, waarvan de relatieve frequenties een belangrijk speciaal geval zijn.

Het algemeen principe van een kansverdeling is nog altijd redelijk voor de hand liggend, we eisen alleen maar eigenschappen die heel natuurlijk zijn:

Zij Ω de verzameling van mogelijke uitkomsten. We willen nu graag aan elke deelverzameling $A \subseteq \Omega$ een kans $P(A)$ toewijzen. Hiervoor hebben we een functie

$$P : \mathcal{P}(\Omega) := \{A \subseteq \Omega\} \to \mathbb{R}$$

nodig, die op de *machtsverzameling* van Ω, d.w.z. de verzameling van alle deelverzamelingen van Ω, gedefinieerd is. We noemen zo'n functie $P : \mathcal{P}(\Omega) \to \mathbb{R}$ een *kansverdeling* als P aan de volgende eisen voldoet:

(i) $P(A) \geq 0$ voor alle $A \subseteq \Omega$,

(ii) $P(\Omega) = 1$,

(iii) $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$.

De eerste eigenschap zegt alleen maar, dat kansen niet negatief mogen zijn, en de tweede eigenschap beweert, dat alle mogelijke uitkomsten inderdaad in Ω liggen. De derde eigenschap is een soort van additiviteit, die zegt dat we de kansen voor uitkomsten die niet overlappen (dus niets met elkaar te maken hebben) gewoon mogen optellen. We hadden in principe ook nog kunnen eisen, dat $P(A) \leq 1$ is voor alle $A \subseteq \Omega$, maar dit kunnen we inderdaad uit (i)-(iii) al afleiden en willen graag zo zuinig als mogelijk met onze eisen zijn.

2.1 Discrete kansverdelingen

We hebben tot nu toe alleen maar naar voorbeelden gekeken, waarbij de verzameling Ω van mogelijke uitkomsten eindig is. In deze situatie spreken we van *discrete* kansverdelingen, in tegenstelling tot *continue* kansverdelingen die we in de volgende paragraaf gaan behandelen.

Een belangrijk voorbeeld van een discrete kansverdeling hebben we al gezien, namelijk de *gelijkverdeling* die vaak ook *Laplace-verdeling* heet: Elke mogelijke uitkomst $w \in \Omega$ moet dezelfde kans hebben (vandaar de naam), dan is $P(w) = \frac{1}{|\Omega|}$ voor elke $w \in \Omega$. Hieruit volgt met eigenschap (iii) dat $P(A) = \frac{|A|}{|\Omega|}$ en dit is precies de relatieve frequentie.

We gaan nu een aantal voorbeelden bekijken waarin we het tellen van uitkomsten toepassen en daarbij verschillende belangrijke discrete kansverdelingen tegen komen.

Voorbeeld 1: Bij de lotto 6 uit 49 worden uit een vaas met 49 ballen 6 ballen getrokken en vervolgens in opstijgende volgorde gebracht. Omdat de volgorde hier geen rol speelt en zonder terugleggen getrokken wordt, zijn we in het geval IV (volgens de lijst uit de vorige les). Het aantal mogelijke uitkomsten is dus $\binom{49}{6}$. We willen nu de kans bepalen dat we bij ons 6 kruisjes k goede getallen hebben waarbij $0 \leq k \leq 6$. De k goede getallen kunnen we op $\binom{6}{k}$ manieren uit de 6 juiste getallen kiezen. Maar ook voor de verkeerd aangekruisde getallen moeten we nog iets zeggen, want we willen *precies* k goede getallen hebben, dus mogen we niet per ongeluk nog een verder goed getal krijgen. We moeten dus onze $6-k$ resterende getallen uit de $49-6=43$ verkeerde getallen kiezen en hiervoor zijn er $\binom{43}{6-k}$ mogelijkheden. Het aantal manieren hoe we precies k goede getallen kunnen kiezen is dus $\binom{6}{k} \cdot \binom{43}{6-k}$ en de kans op k goede getallen is dus

$$\frac{\binom{6}{k} \cdot \binom{43}{6-k}}{\binom{49}{6}}.$$

De waarden voor deze kansen zijn:

$k=0:$	43.6%	(1 in 2.3)
$k=1:$	41.3%	(1 in 2.4)
$k=2:$	13.2%	(1 in 7.6)
$k=3:$	1.8%	(1 in 57)
$k=4:$	0.1%	(1 in 1032)
$k=5:$	0.002%	(1 in 54201)
$k=6:$	0.000007%	(1 in 13983816)

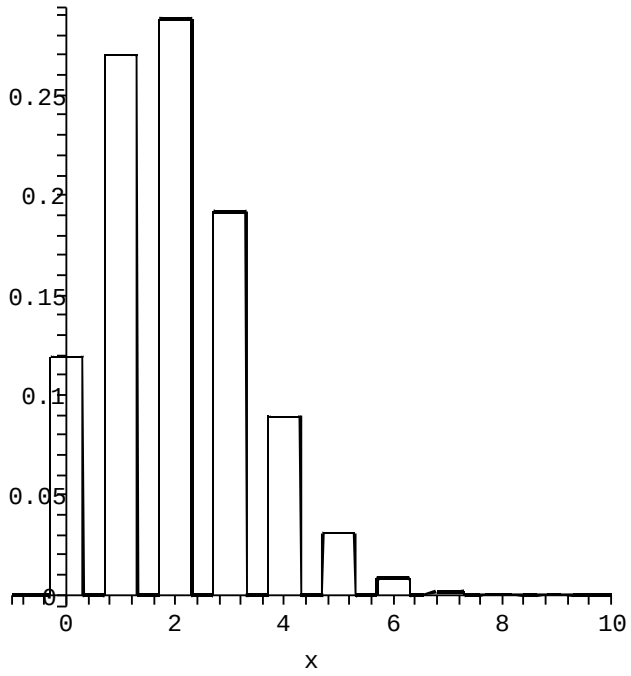
Voorbeeld 2: Bij een qualiteitstoets kiezen we uit een levering van n stukken een steekproef van m stukken die we testen en niet terugleggen. Dit is bijvoorbeeld het geval als de test het object beschadigt, zo als bij het testen van lucifers. We nemen aan dat de levering s slechte stukken bevat en willen de kans berekenen, dat we in onze steekproef k slechte stukken vinden. Omdat we alleen maar in het aantal slechte stukken geïnteresseerd zijn, maar niet of de eerste of laatste slecht zijn, zijn we weer in het geval IV. We kunnen de kans nu net als in het voorbeeld van de lotto berekenen: Er zijn $\binom{s}{k}$ mogelijkheden om k slechte uit de s slechte stukken de vissen, dan zijn er $\binom{n-s}{m-k}$ mogelijkheden om nog $m-k$ goede stukken te kiezen en het totale aantal van mogelijke grepen

is $\binom{n}{m}$. De kans, om k slechte te vinden is dus

$$h(n, m, s; k) := \frac{\binom{s}{k} \cdot \binom{n-s}{m-k}}{\binom{n}{m}}.$$

Omdat dit zo'n belangrijk geval is, heeft deze kansverdeling een eigen naam, ze heet de *hypergeometrische verdeling*.

Ook de kansverdeling die we in Voorbeeld 1 hebben bekeken, is een hypergeometrische kansverdeling, namelijk $h(49, 6, 6; k)$. Figuur 3 laat een histogram voor de hypergeometrische verdeling $h(1000, 100, 20; k)$ zien: Bij een levering van 1000 stukken, waarvan 2% slecht is maken we een steekproef van 100 stuk en kijken, met welke kans we k slechte stukken vinden. Zo als men dat misschien zou verwachten, is de kans bij $k = 2$ maximaal.



Figuur 3: Hypergeometrische verdeling $h(1000, 100, 20; k)$

De praktijk van een qualiteitstoets ziet er natuurlijk eigenlijk iets anders eruit: We weten niet hoeveel slechte stukken er in de levering zitten, maar de leverancier beweert dat het er minder dan s_0 zijn. Wij kennen wel de waarden n, m en k en schatten nu de waarde $\hat{s}$ van s zo dat $h(n, m, \hat{s}; k)$ maximaal wordt. Als onze schatting $\hat{s}$ groter dan s_0 is, zullen we de levering waarschijnlijk niet accepteren.

Een andere toepassing van dit soort schatting vinden we in de ecologie. Als we het aantal vissen in een vijver willen bepalen, kunnen we een aantal s van

vissen markeren en op de volgende dag het aantal k van gemarkeerde vissen in een greep van m vissen bepalen. We schatten dan het aantal $\hat{n}$ van vissen in de vijver zo dat $h(\hat{n}, m, s; k)$ maximaal wordt.

Stel we markeren 1000 vissen en vangen op de volgende dag ook 1000 vissen, waaronder we 100 gemarkeerde vissen vinden. We weten nu dat er minstens nog 900 gemarkeerde vissen in de vijver zitten, dus is $n \geq 1900$. Maar $h(1900, 1000, 1000; 100) \approx 5 \cdot 10^{-430}$, dus deze kans is heel erg klein. Evenzo is de kans op een miljoen vissen heel klein, namelijk $h(10^6, 1000, 1000; 100) \approx 2 \cdot 10^{-163}$. We vinden de maximale waarde voor $\hat{n} = 10000$ en nemen daarom aan dat er ongeveer 10000 vissen in de vijver zijn. Zo'n soort schatting noemen we een *maximum likelihood* schatting, omdat we de parameter n zo kiezen dat de kans $h(n, m, s; k)$ *maximaal* wordt.

Voorbeeld 3: Als we een qualiteitstoets uitvoeren waarbij de stukken niet beschadigt worden en we misschien ook iets heel kostbaars testen (bijvoorbeeld het gewicht van een staaf goud) zullen we getoetste stukken waarschijnlijk weer terugleggen. Dan zijn we niet meer in het geval IV maar moeten de kans op een andere manier bepalen. We letten nu wel op de volgorde en zijn dus in het geval I. Er zijn s^k manieren om k slechte uit de s slechte stukken te kiezen en er zijn $(n-s)^{m-k}$ manieren om $m-k$ goede uit de $n-s$ goede stukken te kiezen. Maar omdat de goede niet van de slechte stukken gescheiden zijn moeten we ook nog tellen hoe we de k slechte stukken op de m grepen kunnen verdelen. Hiervoor zijn er $\binom{m}{k}$ mogelijkheden. Als we de relatieve frequentie van slechte stukken $p := \frac{s}{n}$ noemen vinden we dus voor de kans om k slechte stukken te kiezen:

$$b(n, m, s; k) := \frac{\binom{m}{k} s^k (n-s)^{m-k}}{n^m} = \binom{m}{k} p^k (1-p)^{m-k} =: b(m, p; k).$$

Ook deze kansverdeling is heel fundamenteel een heet de *binomiale verdeling*.

Intuïtief zullen we zeggen, dat het voor het geval dat n veel groter is dan m bijna geen verschil maakt of we met of zonder terugleggen trekken, want de kans dat we een element twee keer pakken is heel klein. Er laat zich inderdaad zuiver aantonen, dat voor $n \gg m$ de hypergeometrische verdeling meer en meer op de binomiale verdeling lijkt en in de limiet geldt

$$\lim_{n \to \infty} h(n, m, np; k) = b(m, p; k).$$

Deze samenhang tussen hypergeometrische en binomiale verdeling wordt meestal de *binomiale benadering* van de hypergeometrische verdeling genoemd. Merk op dat de binomiale verdeling (behalve van de grootte m van de greep) alleen maar van één parameter afhangt, namelijke het relatieve aantal $p = \frac{s}{n}$ van slechte stukken, terwijl de hypergeometrische verdeling van het totaal aantal n van stukken en het aantal s van slechte stukken afhangt. Dit maakt het natuurlijk veel handiger om met de binomiale verdeling te werken, vooral als je bedenkt dat deze functies vaak in de vorm van tabellen aangegeven worden.

Er laat zich geen algemene regel aangeven, wanneer de binomiale benadering goed genoeg is. Soms leest men iets van $n > 2000$ en $\frac{m}{n} < 0.1$, maar in sommige

gevallen heeft de benadering dan al een behoorlijke afwijking. Voor $n = 2000$, $m = 100$, $s = 20$ en $k = 2$ hebben we bijvoorbeeld $h(2000, 100, 20; 2) = 18.95\%$ en de binomiale benadering geeft in dit geval $b(100, \frac{20}{2000}; 2) = 18.49\%$ wat al een tamelijke afwijking is. Als we aan de andere kant naar de kans op 2 goede getallen in de lotto kijken, hebben we $h(49, 6, 6; 2) = 13.24\%$. De binomiale benadering hiervan is $b(6, \frac{6}{49}; 2) = 13.34\%$ en dit is een redelijke benadering terwijl we hier niet aan het criterium voldoen.

De Poisson-verdeling

Vaak willen we bij experimenten de kans weten, dat er bij m pogingen k keer een bepaalde uitkomst plaats vindt. We hebben gezien dat we dit met de binomiale verdeling kunnen beschrijven: Als de kans voor een gunstige uitkomst p is, dan is $b(m, p; k) := \binom{m}{k} p^k (1-p)^{m-k}$ de kans op k gunstige uitkomsten bij m pogingen.

Voor heel zeldzame gebeurtenissen zullen we verwachten dat er veel pogingen nodig zijn tot dat een gunstige uitkomst optreed en als de kans p maar nog half zo groot is, zullen we verwachten twee keer zo vaak te moeten proberen. Om voor gebeurtenissen waar p tegen 0 loopt nog een gunstige uitkomst te kunnen verwachten, moeten we dus m zo laten groeien dat $m \cdot p = \lambda$ constant blijft. De constante λ geeft aan hoeveel gunstige uitkomsten we bij m pogingen eigenlijk verwachten.

De vraag is nu wat er met de binomiale verdeling $b(m, p; k)$ gebeurt als we de limiet $p \to 0$, $m \to \infty$ bekijken met $p \cdot m = \lambda$. We hebben

$$\begin{aligned}
\binom{m}{k} p^k (1-p)^{m-k} &= \frac{m!}{k!(m-k)!} \frac{\lambda^k}{m^k} (1 - \frac{\lambda}{m})^{m-k} \\
&= \frac{\lambda^k}{k!} (1 - \frac{\lambda}{m})^m (\frac{m}{m} \cdot \frac{m-1}{m} \cdot \ldots \cdot \frac{m-k+1}{m}) (1 - \frac{\lambda}{m})^{-k} \\
&\to \frac{\lambda^k}{k!} e^{-\lambda},
\end{aligned}$$

omdat $(1 - \frac{\lambda}{m})^m \to e^{-\lambda}$ voor $m \to \infty$ en $\frac{m-k+1}{m} \to 1$ en $(1 - \frac{\lambda}{m}) \to 1$ voor $m \to \infty$.

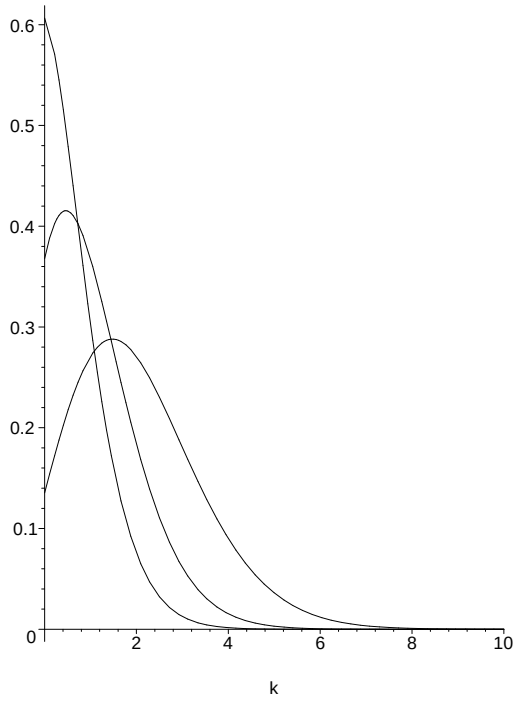
Voor zeldzame gebeurtenissen gaat de binomiale verdeling dus in de limiet tegen de *Poisson-verdeling*

$$P(k) = po_\lambda(k) := \frac{\lambda^k}{k!} e^{-\lambda}.$$

Merk op dat bij de binomiale verdeling het aantal gunstige uitkomsten natuurlijk door het aantal pogingen begrensd is. In de Poisson-verdeling is de enige parameter het aantal verwachte successen λ en we kunnen dus met deze verdeling de kans voor elk aantal gunstige uitkomsten berekenen.

Hoe goed de Poisson-verdeling de binomiale verdeling benadert hangt natuurlijk van de parameters af. Als een vuistregel geldt, dat men de Poisson-benadering mag gebruiken als $p < 0.1$ en $\lambda \leq 5$ of $\lambda \leq 10$, maar hierbij speelt natuurlijk ook weer de benodigde nauwkeurigheid een rol.

De afhankelijkheid van de Poisson-verdeling van de parameter λ kunnen we in Figuur 4 zien, waar de Poisson-verdelingen voor de parameters $\lambda = 0.5, 1, 2$ als continue functies van k getekend zijn. De kansen worden alleen maar op de punten $k \in \mathbb{N}$ afgelezen.



Figuur 4: Poisson-verdelingen voor parameters $\lambda = 0.5, 1, 2$

Omdat $lim_{k\to 0}\frac{\lambda^k}{k!} = 1$ is, heeft de Poisson-verdeling in 0 de waarde $e^{-\lambda}$ en we zien dat voor kleinere waarden van λ de grafiek bij een hogere waarde voor $k = 0$ begint maar dan sneller naar 0 toe gaat. Dit klopt ook met onze intuïtie, want als de kans voor een zeldzaam gebeurtenis minder groot is, verwachten we met een hogere waarschijnlijkheid dat het helemaal niet gebeurt. In het plaatje hoort dus de grafiek die bij $e^{-0.5} \approx 0.61$ begint bij de parameter $\lambda = 0.5$, de grafiek die bij $e^{-1} \approx 0.37$ begint hoort bij de parameter $\lambda = 1$, en de grafiek die bij $e^{-2} \approx 0.14$ begint hoort bij de parameter $\lambda = 2$.

Het maximum van de continue Poisson-verdeling laat zich alleen maar door een ingewikkelde functie (de Ψ-functie) beschrijven, voor $\lambda = 1$ ligt het ongeveer bij 0.46 en voor $\lambda = 2$ bij 1.48. Voor kleine waarden van λ is de grafiek van de Poisson-verdeling dalend, een maximum bestaat alleen maar voor waarden $\lambda \gtrsim 0.562$.

De maximale waarde van de Poisson-verdeling voor $k \in \mathbb{N}$ laat zich wel berekenen. We hebben $\frac{po_\lambda(k+1)}{po_\lambda(k)} = \frac{\lambda^{k+1}}{(k+1)!} \cdot \frac{k!}{\lambda^k} = \frac{\lambda}{k+1}$. Dit toont aan dat de waarden van po_λ voor $k \leq \lambda$ groeien en dan weer dalen. De maximale waarde is bereikt voor het grootste gehele getal $\leq \lambda$. Als λ zelf een geheel getal is, zijn de waarden voor $k = \lambda - 1$ en $k = \lambda$ hetzelfde.

De Poisson-verdeling is altijd van belang als het erom gaat zeldzame gebeurtenissen te beschrijven. Voorbeelden hiervoor zijn:

- Gevallen met een heel hoge schade voor verzekeringsmaatschappijen.
- Het uitzenden van α-deeltjes door een radioactief preparaat.
- Het aantal drukfouten op een bladzijde.

We kijken naar een voorbeeld: We dobbelen met vier dobbelstenen, dan is de kans om vier 6en te hebben gelijk aan $\frac{1}{6^4}$. Als we nu 1000 keer dobbelen is de parameter $\lambda = m \cdot p = \frac{1000}{1296} \approx 0.77$. De kans om bij de 1000 werpen geen enkele keer vier zessen te hebben is dus $e^{-\lambda} \approx 0.46$, de kans dat het een keer gebeurd is $\lambda e^{-\lambda} \approx 0.36$, de kans op twee keer zo'n werp is $\frac{\lambda^2}{2} e^{-\lambda} \approx 0.14$. De kans op drie of meer keer vier zessen is ongeveer 4.3%.

Merk op dat we altijd het aantal m van grepen kennen en de parameter λ kunnen uitrekenen als we de kans p van gunstige uitkomsten kennen. Vaak komen we in de praktijk het omgedraaide probleem tegen: We kennen het aantal k van gunstige uitkomsten bij een aantal m van pogingen. Hieruit willen we nu de kans p op een gunstige uitkomst schatten. Hiervoor kiezen we de parameter λ zo dat de bijhorende Poisson-verdeling een maximale waarde in k heeft. Dit is weer een *maximum likelihood* schatting.

2.2 Continue kansverdelingen

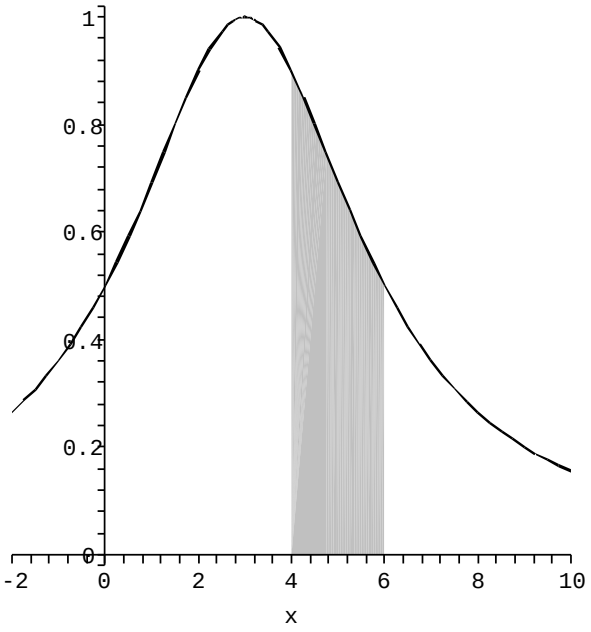
We hebben tot nu toe alleen maar naar eindige uitkomstenruimten Ω gekeken, d.w.z. naar uitkomstenruimten met $|\Omega| = n < \infty$. Met analoge technieken laten zich ook kansverdelingen op oneindige maar aftelbare ruimten Ω definiëren, d.w.z. op ruimten Ω die in bijectie zijn met de natuurlijke getallen $\mathbb{N}$. Zo'n bijectie geeft gewoon nummers aan de elementen en we krijgen $\Omega = \{\omega_1, \omega_2, \ldots\} = \{\omega_i \mid i \in \mathbb{N}\}$. Door ω_i door het gewone getal i te vervangen kunnen we elke aftelbare ruimte Ω tot de natuurlijke getallen $\mathbb{N}$ terugbrengen en we hoeven dus bij aftelbaar oneindige uitkomstenruimten alleen maar aan de natuurlijke getallen te denken.

De normering $P(\Omega) = 1$ van de kansverdeling komt in dit geval neer op een uitspraak over een oneindige reeks, namelijk $\sum_{i=0}^{\infty} P(i) = 1$. Ook kansverdelingen voor aftelbare uitkomstenruimten noemen we nog *discrete kansverdelingen* omdat we de punten van de natuurlijke getallen als gescheiden punten op de reële lijn beschouwen.

Vaak hebben experimenten echter helemaal geen discrete uitkomsten. Als we bijvoorbeeld naar de wachttijd kijken die we als klant in een rij doorbrengen voordat we geholpen worden, kan de uitkomst een willekeurige tijd t zijn (met misschien een zekere bovengrens). Net zo kunnen we bij een test van het invloed van doping-middelen op de prestatie van kogelstoters willekeurige waarden tussen $10m$ en $25m$ verwachten. In dit voorbeeld leert onze ervaring al een mogelijke oplossing, hoe we naar discrete uitkomsten terug komen. De prestaties worden namelijk alleen maar tot op centimeters nauwkeurig aangegeven en we vatten dus alle waarden in een zeker interval tot een enkele uitkomst samen.

Maar we kunnen ook kansverdelingen met continue uitkomsten beschrijven. Het idee hiervoor is als volgt: We beschrijven de kans dat de uitkomst x van

een experiment in het interval $[a, b]$ valt als oppervlakte onder de grafiek van een geschikte functie $f(x)$ op het interval $[a, b]$.



Figuur 5: Kans op een uitkomst in een interval als oppervlakte onder de grafiek van een functie.

De oppervlakte onder een grafiek noteren we als *integraal*, we krijgen dan voor de kans $P(a \leq x \leq b)$ dat x in het interval $[a, b]$ ligt:

$$P(a \leq x \leq b) = \int_a^b f(t)\, dt.$$

Als de kans groot is, moet de gemiddelde waarde van $f(x)$ op het interval dus ook groot zijn, als de kans klein is, heeft ook de functie $f(x)$ kleine waarden. Om op deze manier echt een kansverdeling te krijgen, moet de functie $f(x)$ aan de volgende eisen voldoen:

(i) $f(x) \geq 0$ voor alle $x \in \mathbb{R}$,

(ii) $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

De eerste eis zorgt ervoor dat we steeds niet-negatieve kansen krijgen en de tweede eis zegt dat de totale oppervlakte onder de grafiek 1 is en geeft dus de normering van de kansverdeling weer. We noemen een functie $f(x) : \mathbb{R} \to \mathbb{R}$ die aan deze eisen voldoet een *dichtheidsfunctie*.

Merk op dat we met de definitie van de kans als oppervlakte op een interval automatisch aan de eis voldoen dat $P(A \cup B) = P(A) + P(B)$ als $A \cap B = \emptyset$ (eis (iii) uit de oorspronkelijke definitie van een kansverdeling) want voor niet

overlappende deelintervallen $[a, b]$ en $[c, d]$ worden de oppervlakten gewoon bij elkaar opgeteld.

In nauw verband met de dichtheidsfunctie $f(x)$ staat de *verdelingsfunctie* $F(a)$, die voor elke waarde van a de kans $P(x \leq a)$ dat de uitkomst hoogstens a is aangeeft. Omdat dit betekent dat $-\infty < x \leq a$, krijgen we deze kans als oppervlakte onder de grafiek van $f(x)$ tussen $-\infty$ en a, dus (weer als integraal geschreven) als

$$F(a) := \int_{-\infty}^{a} f(x)\, dx.$$

De verdelingsfunctie heeft de eigenschappen:

(i) $\lim_{a\to-\infty} F(a) = 0$, $\lim_{a\to\infty} F(a) = 1$.

(ii) $F(a)$ is stijgend, dus $a_2 \geq a_1 \Rightarrow F(a_2) \geq F(a_1)$.

(iii) $P(a \leq x \leq b) = F(b) - F(a)$.

(iv) $F'(a) = f(a)$, dus de afgeleide van $F(a)$ geeft de dichtheidsfunctie.

We gaan nu een aantal belangrijke voorbeelden van continue kansverdelingen bekijken:

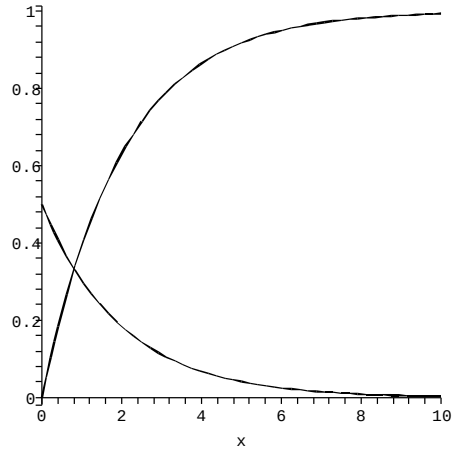
Voorbeeld 1: De uniforme verdeling (homogene verdeling, rechthoekverdeling).
Dit is het continue analoog van de discrete gelijkverdeling. Op een bepaald interval $[a, b]$ (of een vereniging van intervallen) heeft elke punt dezelfde kans en buiten het interval is de kans 0. De normering $\int_{-\infty}^{\infty} f(x)\, dx = 1$ geeft dan de waarde voor $f(x)$ op het interval $[a, b]$. De dichtheidsfunctie $f(x)$ en verdelingsfunctie $F(x)$ van de uniforme verdeling zijn

$$f(x) = \begin{cases} 0 & \text{als } x < a \\ \frac{1}{b-a} & \text{als } a \leq x \leq b \\ 0 & \text{als } x > b \end{cases} \quad \text{en} \quad F(x) = \begin{cases} 0 & \text{als } x < a \\ \frac{x-a}{b-a} & \text{als } a \leq x \leq b \\ 1 & \text{als } x > b \end{cases}$$

Voorbeeld 2: De exponentiële verdeling.
Bij het bepalen van de levensduur van dingen als radioactieve preparaten of borden in de kast gaan we ervan uit dat het aantal verdwijnende objecten evenredig is met het aantal objecten die er nog zijn. Dit soort processen voldoet aan een differentiaalvergelijking $f'(x) = \lambda f(x)$ die de oplossing $e^{-\lambda x}$ heeft. De dichtheidsfunctie en verdelingsfunctie die de levensduur van dit soort objecten beschrijft, zijn:

$$f(x) = \begin{cases} 0 & \text{als } x < c \\ \lambda e^{-\lambda(x-c)} & \text{als } x \geq c \end{cases} \quad \text{en} \quad F(x) = \begin{cases} 0 & \text{als } x < c \\ 1 - e^{-\lambda(x-c)} & \text{als } x \geq c \end{cases}$$

Merk op dat de constante factor λ bij de exponentiële functie weer door de normering bepaald is, want $\int_c^{\infty} e^{-\lambda(x-c)}\, dx = \frac{-1}{\lambda} e^{-\lambda(x-c)}\big|_c^{\infty} = \frac{1}{\lambda}$.



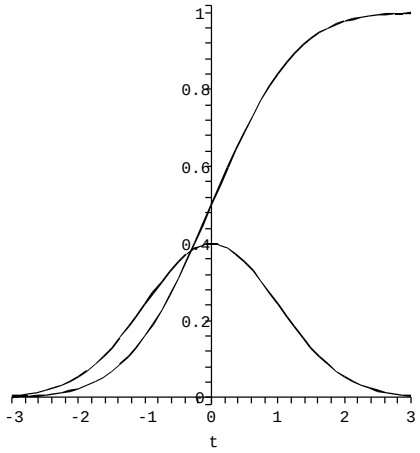
Figuur 6: Dichtheidsfunctie en verdelingsfunctie voor de exponentiële verdeling met $\lambda = 0.5$ en $c = 0$

Voorbeeld 3: De normaalverdeling (Gauss verdeling).
De belangrijkste continue verdeling is de normaalverdeling die centraal in de statistiek staat. De dichtheidsfunctie heeft de vorm van een klok en is gegeven door

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

In dit geval kunnen we de verdelingsfunctie alleen maar door de integraal van $f(x)$ beschrijven.

De normaalverdeling met $\mu = 0$ en $\sigma = 1$ noemen we *standaardnormaalverdeling*.



Figuur 7: Dichtheidsfunctie en verdelingsfunctie voor de standaardnormaalverdeling

Belangrijke begrippen in deze les

- kansverdeling
- gelijkverdeling (Laplace-verdeling)
- hypergeometrische verdeling
- binomiale verdeling
- Poisson-verdeling
- continue kansverdeling
- dichtheidsfunctie, verdelingsfunctie

Opgaven

6. Bij een hockeytoernooi zijn er 18 teams aangemeld. In de eerste ronde worden de teams in twee groepen van 9 teams geloot. Onder de deelnemers zijn 5 teams uit de hoogste klasse. Hoe groot is de kans dat deze 5 teams in dezelfde groep terecht komen? Hoe groot is de kans dat er in een groep 2 en in de andere 3 teams uit de hoogste klasse terecht komen.
7. In een kast liggen n paren schoenen (dus $2n$ schoenen) willekeurig door elkaar. Je grijpt blindelings $k \leq n$ schoenen. Hoe groot is de kans dat je er minstens één passend paar uit vist? Hoe groot is de kans dat je precies één paar uit vist?
8. Een test bestaat uit 10 ja-nee vragen. Iemand die van toeten nog blazen weet, besluit de vragen op goed geluk te beantwoorden (dit betekent dat hij voor elke vraag een kans van $\frac{1}{2}$ op een goed antwoord heeft). Met 6 goede antwoorden ben je in de test geslaagd. Wat is de kans voor onze kandidaat om de test te halen?
9. Volgens een statistiek vinden in Nederland per jaar 3 op de 100.000 mensen een portemonnee met meer dan 1000 €. Wat is de kans dat in en stad als Nijmegen (met 150.000 inwoners) dit geluk (a) 3, (b) 5, (c) 10, (d) hooguit 2 mensen overkomt.
10. De kans dat een eerstejaars student in een bepaald vak afstudeert is 40%. Wat zijn de kansen dat uit een groep van 5 eerstejaars:
    (i) niemand afstudeert,
    (ii) precies 1 afstudeert,
    (iii) minstens 3 afstuderen?

Les 3 Verwachtingswaarde en spreiding

3.1 Stochasten

In een paar voorbeelden hebben we al gezien dat we bij een experiment vaak niet zo zeer in een enkele uitkomst geïneteresseerd zijn, maar bijvoorbeeld wel in het aantal uitkomsten van een zeker soort. Zo willen we bij een steekproef weten, hoeveel stukken defect zijn, maar niet of nu het eerste of laatste stuk defect is.

Vaak zijn de uitkomsten waarin we geïneteresseerd zijn veel eenvoudiger dan de uitkomstenruimte zelf, bijvoorbeeld kijken we naar het aantal k van defecte stukken in plaats van alle combinaties van m testresultaten, waarvan k negatief zijn. We kunnen dus zeggen, dat we verschillende uitkomsten die een zekere eigenschap gemeenschappelijk hebben in een cluster samenvatten, Zo'n eigenschap laat zich door een functie

$$X : \Omega \to \mathbb{R}, \quad \omega \mapsto X(\omega)$$

beschrijven, die aan elk element ω van de uitkomstenruimte een waarde $X(\omega)$ toekent. Zo'n functie X noemen we een *random variable* (in het Engels), een *stochastische variabele*, een *kansvariabele* of kort een *stochast*.

In het voorbeeld van de kwaliteitsproef is de stochast dus de functie die aan een rij van testresultaten het aantal negatieve (of positieve) resultaten toekent.

Een ander voorbeeld is het dobbelen met twee dobbelstenen: Als we alleen maar in de som van de geworpen getallen geïneteresseerd zijn, nemen we als stochast de functie $X(\omega_1, \omega_2) := \omega_1 + \omega_2$.

Het belangrijke aan de stochasten is, dat we makkelijk een kansverdeling hiervoor kunnen definiëren: De kans $P(X = x)$ dat de stochast de waarde x aanneemt definiëren we door

$$P(X = x) := \sum_{X(\omega)=x} P(\omega)$$

dus we tellen gewoon de kansen voor alle elementen van Ω op, waar de stochast de waarde x oplevert.

Onbewust hebben we al eerder stochasten op deze manier gebruikt, bijvoorbeeld voor het uitrekenen van de kans dat we met twee dobbelstenen een som van 5 werpen.

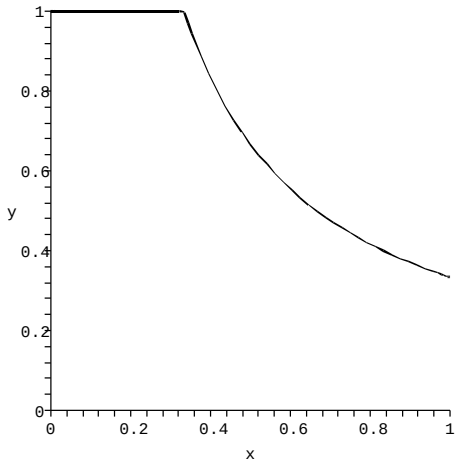
Voor continue kansverdelingen gaat de som over de uitkomsten met $X(\omega) = x$ over in een integraal. Omdat de kans op een enkele uitkomst steeds 0 is, wordt hier de kans bepaald, dat de stochast X een waarde onder een gegeven grens aanneemt. Voor een continue kansverdeling met dichtheidsfunctie $f(x)$ krijgen we:

$$P(X \leq x) = \int_{X(t)\leq x} f(t)\, dt.$$

Meestal zijn continue stochasten door hun eigen dichtheidsfunctie aangegeven, er geldt dan

$$P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt.$$

Voorbeeld: Stel we hebben een randomgenerator die toevalsgetallen tussen 0 en 1 volgens de uniforme verdeling voortbrengt. We vragen ons af, wat de kans is dat het product van twee opeenvolgende van die toevalsgetallen kleiner is dan een grens $0 \leq a \leq 1$. De stochast die bij dit probleem hoort is $X(x, y) := x \cdot y$ en omdat we het met de uniforme verdeling te maken hebben, moeten we alleen maar de oppervlakte van het gebied $G = \{(x, y) \in \mathbb{R}^2 \mid x \cdot y \leq a\}$ bepalen. Als $x \leq a$ kan y elke waarde tussen 0 en 1 hebben, maar voor $x \geq a$ hebben we $y \leq \frac{a}{x}$ nodig. De volgende schets laat dit (voor $a = \frac{1}{3}$) zien:



Met behulp van een eenvoudige integratie kunnen we de kansverdeling van deze stochast ook expliciet bepalen, er geldt:

$$P(X \leq a) = \int_0^a dx + \int_a^1 \frac{a}{x}\, dx = a + a(\log(1) - log(a)) = a(1 - \log(a)).$$

Voor $a = 0.5$ is deze kans bijvoorbeeld $P(X \leq 0.5) \approx 0.85$ en pas voor $a < 0.187$ is $P(X \leq a) < 0.5$.

3.2 Verwachtingswaarde

Als we in het casino roulette gaan spelen, zijn we er niet in geïnteresseerd of we in het eerste of laatste spel winnen of verliezen en ook niet hoe vaak we winnen of verliezen. Eigenlijk willen we alleen maar weten of we kunnen verwachten dat we aan het eind van de dag met een winst naar huis komen. Als we N keer spelen en bij elke keer 10€ op rood zetten, dan is bij elk spel de kans dat we 10€ winnen gelijk aan $\frac{18}{37}$, want er zijn 18 rode en 18 zwarte getallen en de groene 0. De kans dat we de 10€ verliezen is dus $\frac{19}{37}$. Als we heel vaak spelen, kunnen we verwachten dat we $\frac{18 \cdot N}{37}$ keer winnen en $\frac{19 \cdot N}{37}$ keer verliezen. Dit betekend dat we een verlies van $N \cdot \frac{1}{37} \cdot 10$€ kunnen verwachten.

Uit het perspectief van het casino is dit natuurlijk heel wenselijk. Omdat alle winsten alleen maar op de getallen 1 t/m 36 zijn gebaseerd (dus als je op de 3 getallen $4, 5, 6$ zet maak je een winst van 12 keer je inzet) heeft de groene 0 het effect dat het casino gemiddeld een zevenendertigste van alle inzetten wint.

In het voorbeeld van het roulette spel hebben we een stochast gebruikt die het bedrag van de winst of verlies aangeeft. Waar we in geïnteresseerd zijn is de gemiddelde winst die we per spel zullen maken. Dit is het gemiddelde van de mogelijke waarden van de stochast, waarbij elke waarde met zijn kans gewogen wordt. Wat we zo krijgen is de winst die we per spel gemiddeld verwachten, en daarom noemen we dit ook de *verwachtingswaarde*.

Algemeen definiëren we voor een stochast X de verwachtingswaarde $E(X)$ (de E staat voor het Engelse *expectation*) door

$$E(X) := \sum_{x \in X} x \cdot P(X = x) = \sum_{x \in X} x \cdot (\sum_{X(\omega)=x} P(\omega)) = \sum_{\omega \in \Omega} X(\omega) P(\omega).$$

Voor een stochast X met continue kansverdeling is de verwachtingswaarde met behulp van zijn dichtheidsfunctie $f(x)$ analoog gedefinieerd door de integraal

$$E(X) := \int_{-\infty}^{\infty} x \cdot f(x)\, dx.$$

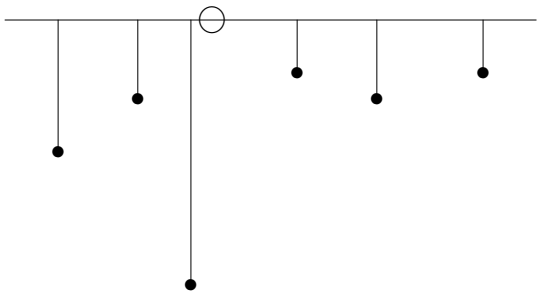
Merk op dat we van een continu verdeelde stochast door samenvatten van de waarden in een deelinterval naar een discreet verdeelde stochast kunnen komen:

Er geldt $P(X \in [x, x+\delta]) = \int_x^{x+\delta} f(t)\ dt$ en voor kleine δ kunnen we aannemen dat $f(t)$ op het interval $[x, x+\delta]$ bijna constant is, dit geeft

$$P(X \in [x, x+\delta]) \approx \delta \cdot f(x).$$

Als we nu de reële lijn in stukjes $[i \cdot \delta, (i+1) \cdot \delta]$ van lengte δ onderverdelen en de uitkomsten $x \in [i \cdot \delta, (i+1) \cdot \delta]$ tot de uitkomst $x = i \cdot \delta$ samenvatten, hebben we alleen maar nog de discrete verzameling $\{i \cdot \delta \mid i \in \mathbb{Z}\}$ van uitkomsten. Voor deze *gediscretiseerde* stochast is de verwachtingswaarde $\sum_{i \in \mathbb{Z}, x = i \cdot \delta} x \cdot P(X \in [x, x+\delta]) \approx \sum_{i \in \mathbb{Z}, x = i \cdot \delta} x \cdot \delta \cdot f(x)$ en dit is juist de discrete benadering van de integraal $\int_{-\infty}^{\infty} x \cdot f(x)\, dx = E(X)$.

We kunnen de verwachtingswaarde aanschouwelijk zien als het evenwichtspunt van een balk (oneindig lang, zonder gewicht), waar we in het punt x een gewicht van massa $P(x)$ aan hangen. Het evenwichtspunt is dan juist het punt $E(X)$. In het volgende plaatje zijn de gewichten gerepresenteerd door de lengten van de verticale ribben.



Een aantal belangrijke elementaire eigenschappen van de verwachtingswaarde kunnen we meteen uit de definitie aflezen. Als X en Y stochasten zijn, dan geldt:

(i) $E(X+Y) = E(X) + E(Y)$, dus de som van de verwachtingswaarden van twee stochasten is de verwachtingswaarde van de som van de stochasten.

(ii) $E(\alpha X) = \alpha E(X)$.

(iii) $X(\omega) \geq Y(\omega)$ voor alle $\omega \in \Omega \Rightarrow E(X) \geq E(Y)$.

Als we in (i) voor Y de constante stochast $Y(\omega) = c$ nemen, volgt hieruit dat een verschuiving van de stochast om c ook de verwachtingswaarde om c verschuift (omdat de constante stochast verwachtingswaarde c heeft). We kunnen dus een stochast door aftrekken van zijn verwachtingswaarde altijd zo verschuiven dat hij verwachtingswaarde 0 heeft:

$$X_0 := X - E(X) \Rightarrow E(X_0) = E(X - E(X)) = E(X) - E(X) = 0.$$

We gaan nu de verwachtingswaarden van de belangrijkste kansverdelingen berekenen.

Binomiale verdeling

We hebben $P(X = k) = b(m, p; k) = \binom{m}{k} p^k (1-p)^{m-k}$, dus:

$$\begin{aligned} E(X) &= \sum_{k=0}^{m} k \binom{m}{k} p^k (1-p)^{m-k} = \sum_{k=0}^{m} k \, \frac{m!}{k!(m-k)!} \, p^k (1-p)^{m-k} \\ &= m \cdot p \cdot \sum_{k=1}^{m} \frac{(m-1)!}{(k-1)!(m-k)!} \, p^{k-1} (1-p)^{m-k} \\ &= m \cdot p \cdot \sum_{k=0}^{m-1} \binom{m-1}{k} p^k (1-p)^{m-1-k} \\ &= m \cdot p \cdot \sum_{k=0}^{m-1} b(m-1, p; k) = m \cdot p. \end{aligned}$$

In de laatste stap hebben we hierbij gebruik van het feit gemaakt, dat de som over de kansen $b(m-1, p; k)$ voor alle waarden van k de totale kans 1 oplevert. De verwachtingswaarde van de binomiale verdeling is dus $m \cdot p$ en dit is precies het verwachte aantal van gunstige uitkomsten als we m pogingen doen bij een kans van p voor een gunstige uitkomst.

Hypergeometrische verdeling

We hebben $P(X = k) = h(n, m, s; k) = \frac{\binom{s}{k} \cdot \binom{n-s}{m-k}}{\binom{n}{m}}$, en er geldt: $k \cdot \binom{s}{k} = k \cdot \frac{s!}{k!(s-k)!} = s \cdot \frac{(s-1)!}{(k-1)!(s-k)!} = s \cdot \binom{s-1}{k-1}$ en $\binom{n}{m} = \frac{n!}{m!(n-m)!} = \frac{n}{m} \cdot \frac{(n-1)!}{(m-1)!(n-m)!} =$

$\frac{n}{m} \cdot \binom{n-1}{m-1}$. Hieruit volgt:

$$E(X) = \sum_{k=0}^{m} k \frac{\binom{s}{k} \cdot \binom{n-s}{m-k}}{\binom{n}{m}} = \sum_{k=1}^{m} \frac{s\binom{s-1}{k-1} \cdot \binom{n-s}{m-k}}{\frac{n}{m}\binom{n-1}{m-1}} = m\frac{s}{n} \sum_{k=1}^{m} \frac{\binom{s-1}{k-1} \cdot \binom{n-s}{m-k}}{\binom{n-1}{m-1}}$$
$$= m\frac{s}{n} \sum_{k=0}^{m-1} \frac{\binom{s-1}{k} \cdot \binom{n-s}{m-1-k}}{\binom{n-1}{m-1}} = m\frac{s}{n} \sum_{k=0}^{m-1} h(n-1, m-1, s; k) = m\frac{s}{n}.$$

In de stap naar de laatste regel hebben hierbij k door $k+1$ verplaatst, de som die voor k van 1 tot m loopt, loopt voor $k+1$ van 0 tot $m-1$. In de laatste stap loopt de som over de kansen $h(n-1, m-1, s; k)$ voor alle waarden van k, dus is deze som gelijk aan 1. Het resultaat hadden we ook intuïtief kunnen afleiden, want de kans om bij een greep een van de s slechte stukken uit de totale n stukken te pakken is $\frac{s}{n}$, en als we m keer grijpen zouden we gemiddeld $m\frac{s}{n}$ slechte stukken verwachten.

Poisson-verdeling

We hebben $P(X = k) = po_\lambda(k) = \frac{\lambda^k}{k!}e^{-\lambda}$ en maken gebruik van de gelijkheid $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^\lambda$:

$$E(X) = \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = \lambda \cdot e^{-\lambda} \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda \cdot e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda \cdot e^{-\lambda} \cdot e^\lambda = \lambda.$$

Ook hier vinden we het verwachte resultaat, omdat de Poisson-verdeling de limiet van de binomiale verdeling is als $p \to 0$ gaat en $m \cdot p = \lambda$ constant is.

Uniforme verdeling

We hebben $P(X = x) = \frac{1}{b-a}$ als $a \leq x \leq b$ en 0 anders, dus

$$E(X) = \int_a^b x\frac{1}{b-a}\, dx = \frac{1}{2(b-a)}(b^2 - a^2) = \frac{1}{2}(a+b).$$

De verwachtingswaarde is dus het middelpunt van het interval waarop de dichtheidsfunctie niet 0 is.

Exponentiële verdeling

We nemen aan dat we de dichtheidsfunctie zo hebben verschoven dat de beginwaarde $c = 0$ is. Dan is $f(x) = \lambda e^{-\lambda x}$ als $x \geq 0$ en $f(x) = 0$ anders. Dit geeft

$$E(X) = \int_0^\infty x\lambda e^{-\lambda x}\, dx = -x\lambda e^{-\lambda x}\big|_0^\infty + \int_0^\infty e^{-\lambda x}\, dx = -\frac{1}{\lambda}e^{-\lambda x}\big|_0^\infty = \frac{1}{\lambda}$$

(merk op dat we hierbij gebruiken dat $\lim_{x\to\infty} xe^{-x} = 0$ is). Ook hier is het resultaat voor de verwachtingswaarde plausibel, want als λ groter wordt, gaat de functie $f(x)$ sneller naar nul en moeten we dus een kleinere verwachtingswaarde krijgen.

Normaalverdeling

In dit geval kunnen we de verwachtingswaarde zonder enig rekenwerk bepalen. Als we de dichtheidsfunctie $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ zo verschuiven dat $\mu = 0$ is, is de functie symmetrisch ten opzichte van de y-as en dan is $E(X) = 0$. De verwachtingswaarde voor de algemene normaalverdeling is dus μ en dit is ook geen verrassing omdat de dichtheidsfunctie juist zo gemaakt is.

3.3 Spreiding

Als we de verwachtingswaarde van een stochast kennen, weten we wat we op lange termijn gemiddeld kunnen verwachten. Maar vaak willen we toch iets meer weten, bijvoorbeeld hoe ver de daadwerkelijke uitkomsten van de verwachtingswaarde verwijderd zijn. Als we namelijk een stochast X zo verschuiven dat de verwachtingswaarde 0 is, dan heeft ook de stochast αX verwachtingswaarde 0, maar voor $\alpha > 1$ zijn de enkele uitkomsten verder van de verwachtingswaarde verwijderd.

In het model van de balk met gewichten kunnen we het verschil tussen de stochasten X en αX duidelijk zien. Als de gewichten dicht bij het evenwichtspunt zijn, kunnen we de balk makkelijk om dit punt draaien. Als we nu bijvoorbeeld naar de stochast $10 \cdot X$ kijken, worden de afstanden van het evenwichtspunt met 10 vermenigvuldigd. Nu hebben we meer kracht nodig om de balk te draaien. Dit ligt eraan dat het traagheidsmoment van de balk groter geworden is, dit is namelijk gegeven als als de som over $m \cdot r^2$ waarbij m de massa in een punt is die afstand r van het draaipunt heeft. Als we het traagheidsmoment naar de stochast vertalen wordt dit

$$Var(X) := \sum_{x \in X} (x - E(X))^2 \cdot P(X = x) = E((X - E(X))^2)$$

en dit noemen we de *variantie* of *spreiding* van X. De variantie is dus de verwachtingswaarde van de kwadratische afstand van de stochast van zijn verwachtingswaarde en is dus een maat ervoor hoe dicht de waarden van een stochast bij de verwachtingswaarde liggen.

Vaak wordt in plaats van de variantie de wortel uit de variantie als maat voor de afwijkingen gebruikt, omdat deze lineair met de stochast verandert (d.w.z. als we X met een factor α vermenigvuldigen, wordt ook de wortel uit de variantie met α vermenigvuldigt). We definiëren dus

$$\sigma_X := \sqrt{Var(X)}$$

en noemen dit de *standaardafwijking* van X.

Voorbeeld: Bij het werpen van een dobbelsteen is de verwachtingswaarde $E(X) = \sum_{k=1}^{6} k \cdot \frac{1}{6} = \frac{7}{2}$. De variantie is dan $Var(X) = \sum_{k=1}^{6} (k - \frac{7}{2})^2 \cdot \frac{1}{6} = \frac{35}{12}$ en de standaardafwijking $\sigma_X = \sqrt{\frac{35}{12}} \approx 1.7$.

We hebben boven opgemerkt dat de variantie van een stochast aangeeft hoe sterk de uitkomsten van de verwachtingswaarde afwijken. Deze samenhang

tussen verwachtingswaarde en spreiding kunnen we heel expliciet aangeven, namelijk in de *Ongelijkheid van Chebyshev*. Hierbij maken we een afschatting voor de kans dat een uitkomst een grotere afstand dan $a > 0$ van de verwachtingswaarde $E(X)$ heeft.

Volgens de definitie berekenen we de variantie door $Var(X) = \sum_{x \in X}(x - E(X))^2 \cdot P(X = x)$. Als we de som beperken tot de waarden van x met $|x - E(X)| \geq a$, krijgen we

$$Var(X) \geq \sum_{|x-E(X)|\geq a} (x - E(X))^2 \cdot P(X = x) \geq \sum_{|x-E(X)|\geq a} a^2 \cdot P(X = x)$$

en dit is juist $a^2 \cdot P(|X - E(X)| \geq a)$. We hebben dus bewezen:

Ongelijkheid van Chebyshev: Voor een stochast X met verwachtingswaarde $E(X)$ en variantie $Var(X)$ geldt voor elke $a > 0$ de ongelijkheid

$$P(|X - E(X)| \geq a) \leq \frac{1}{a^2} Var(X).$$

Als voorbeeld kunnen we met de ongelijkheid van Chebyshev eens afschatten, wat de kans op het dobbelen van een zes is. We hebben boven gezien dat de verwachtingswaarde bij het dobbelen $\frac{7}{2}$ en de variantie $\frac{35}{12}$ is. De afstand tussen een 6 en de verwachtingswaarde $\frac{7}{2}$ is $\frac{5}{2}$ en volgens de ongelijkheid van Chebyshev geldt $P(|X - E(X)| \geq \frac{5}{2}) \leq \frac{4}{25} \cdot \frac{35}{12} = \frac{7}{15} \approx 0.467$. Omdat deze kans ook het dobbelen van een 1 insluit, moeten we nog door twee delen en schatten de kans op een 6 dus met 23.3% (naar boven) af. Natuurlijk weten we dat de kans in feite $\frac{1}{6} = 16.7\%$ is en dit laat zien dat de afschatting niet eens zo slecht is.

In de statistiek wordt vaak als vuistregel de zo genoemde 2σ-regel gebruikt: Voor een stochast X met standaardafwijking σ_X liggen meestal 95% van de gebeurtenissen in het interval $(E(X) - 2\sigma_X, E(X) + 2\sigma_X)$. De ongelijkheid van Chebyshev geeft aan dat dit interval minstens 75% van de gebeurtenissen bevat, maar voor de meeste kansverdelingen (in het bijzonder voor de normaalverdeling) geldt de sterkere uitspraak van de 2σ-regel.

Naast de ongelijkheid van Chebyshev kunnen we een aantal verdere belangrijke eigenschappen voor de variantie van een stochast X meteen uit de definities afleiden:

(i) $Var(X) = 0$ dan en slechts dan als $X = c$ constant is.

(ii) $Var(\alpha X) = \alpha^2 Var(X)$ en $\sigma_{\alpha X} = \alpha \cdot \sigma_X$.

(iii) $Var(X+c) = Var(X)$, dus zo als we dit zouden verwachten is de variantie onafhankelijk van een verschuiving van de stochast.

(iv) $Var(X) = E(X^2) - E(X)^2$, want:

$$\begin{aligned} Var(X) &= \sum_{x \in X} (x - E(X))^2 \cdot P(X = x) \\ &= (\sum_{x \in X} x^2 \cdot P(X = x)) - 2E(X)(\sum_{x \in X} x \cdot P(X = x)) + E(X)^2 \\ &= E(X^2) - 2E(X) \cdot E(X) + E(X)^2 = E(X^2) - E(X)^2. \end{aligned}$$

Dit is in veel gevallen een handige formule om de variantie van een stochast uit te rekenen.

Vaak is het nuttig een stochast zo te normeren dat hij verwachtingswaarde 0 en variantie 1 heeft. Dit kunnen we met behulp van (ii) en (iii) makkelijk bereiken, want voor $X_0 := \frac{X-E(X)}{\sigma_X}$ geldt $E(X_0) = \frac{1}{\sigma_X}(E(X) - E(X)) = 0$ en $Var(X_0) = Var(\frac{X}{\sigma_X}) = \frac{1}{\sigma_X^2}Var(X) = 1$.

We gaan nu ook de varianties van de meest belangrijke kansverdelingen berekenen.

Binomiale verdeling

Dit pakken we met de formule $Var(X) = E(X^2) - E(X)^2$ aan:

$$\begin{aligned} E(X^2) &= \sum_{k=0}^{m} k^2 \binom{m}{k} p^k (1-p)^{m-k} \\ &= m \cdot p \cdot \sum_{k=1}^{m} k \frac{(m-1)!}{(k-1)!(m-k)!} p^{k-1}(1-p)^{m-k} \\ &= m \cdot p \cdot \sum_{k=0}^{m-1} (k+1) \binom{m-1}{k} p^k (1-p)^{m-1-k}. \end{aligned}$$

De som $\sum_{k=0}^{m-1}(k+1)\binom{m-1}{k}p^k(1-p)^{m-1-k}$ is de verwachtingswaarde van de verschoven stochast $X+1$ voor de parameter $m-1$, dus is de waarde hiervan $(m-1)p+1$. We hebben dus $E(X^2) = mp((m-1)p+1) = mp(mp+(1-p))$ en dus

$$Var(X) = E(X^2) - E(X)^2 = mp(mp+(1-p)) - (mp)^2 = mp(1-p).$$

Hypergeometrische verdeling

Dit is een beetje omslachtig om uit te werken, dus geven voor de volledigheid alleen maar het resultaat aan. Voor een stochast X met $P(X=k) = h(n,m,s;k)$ geldt

$$Var(X) = m\frac{s}{n}(1-\frac{s}{n})\frac{n-m}{n-1}.$$

Als n veel groter is dan m geldt $\frac{n-m}{n-1} \approx 1$ en met $p = \frac{s}{n}$ wordt de variantie van de hypergeometrische verdeling dan benadert door de variantie van de binomiale verdeling.

Poisson-verdeling

We gebruiken weer de formule $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^\lambda$. Er geldt:

$$E(X^2) = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} k \frac{\lambda^k}{(k-1)!} e^{-\lambda} = \sum_{k=1}^{\infty} ((k-1)+1) \frac{\lambda^k}{(k-1)!} e^{-\lambda}$$
$$= (\sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} e^{-\lambda}) + (\sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda})$$
$$= \lambda^2 e^{-\lambda} (\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}) + \lambda e^{-\lambda} (\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}) = \lambda^2 + \lambda.$$

We hebben dus

$$Var(X) = E(X^2) - E(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Dit hadden we ook uit de variantie voor de binomiale verdeling kunnen gokken, want de Poisson-verdeling is de limiet voor $p \to 0$ met $mp = \lambda$ en bij deze limiet gaat $mp(1-p)$ naar $mp = \lambda$.

Uniforme verdeling

Er geldt

$$E(X^2) = \int_a^b x^2 \frac{1}{b-a}\, dx = \frac{1}{3(b-a)}(b^3 - a^3) = \frac{1}{3}(a^2 + ab + b^2)$$

dus hebben we

$$Var(X) = E(X^2) - E(X)^2 = \frac{1}{3}(a^2 + ab + b^2) - \frac{1}{4}(a^2 + 2ab + b^2) = \frac{1}{12}(a-b)^2.$$

Exponentiële verdeling

Er geldt

$$E(X^2) = \int_0^{\infty} x^2 \lambda e^{-\lambda x}\, dx = -x^2 \lambda e^{-\lambda x}\big|_0^{\infty} + 2\int_0^{\infty} x e^{-\lambda x}\, dx = 2\int_0^{\infty} x e^{-\lambda x}\, dx$$

en dit is gelijk aan $\frac{2}{\lambda} E(X) = \frac{2}{\lambda^2}$. We hebben dus

$$Var(X) = E(X^2) - E(X)^2 = \frac{2}{\lambda^2} - (\frac{1}{\lambda})^2 = \frac{1}{\lambda^2}.$$

Normaalverdeling

Dit is iets ingewikkelder te berekenen maar de parameters in de normaalverdeling zijn zo gekozen dat σ^2 de variantie aangeeft en dus σ de standaardafwijking.

3.4 Covariantie en correlatie

Het is iets moeilijker om iets over de variantie van de som van twee stochasten te zeggen dan dit bij de verwachtingswaarde het geval was. We hebben

$$\begin{aligned} Var(X+Y) &= E((X+Y)^2) - (E(X+Y))^2 \\ &= E(X^2 + 2X \cdot Y + Y^2) - (E(X) + E(Y))^2 \\ &= E(X^2) + 2E(X \cdot Y) + E(Y^2) - E(X)^2 - 2E(X)E(Y) - E(Y)^2 \\ &= E(X^2) - E(X)^2 + E(Y^2) - E(Y)^2 + 2E(X \cdot Y) - 2E(X)E(Y) \\ &= Var(X) + Var(Y) + 2(E(X \cdot Y) - E(X) \cdot E(Y)). \end{aligned}$$

We noemen $E(X \cdot Y) - E(X) \cdot E(Y)$ de *covariantie* van X en Y en noteren dit met $Cov(X, Y)$. Er geldt dus

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

en dit betekent dat de covariantie aangeeft hoe sterk de variantie van de som van twee stochasten afwijkt van de som van de varianties.

De covariantie laat zich ook beschrijven als de verwachtingswaarde van het product van $(X - E(X))$ en $(Y - E(Y)$, want:

$$\begin{aligned} E((X - E(X)) \cdot (Y - E(Y))) &= E(X \cdot Y - E(X)Y - E(Y)X - E(X)E(Y)) \\ &= E(X \cdot Y) - E(E(X)Y) - E(E(Y)X) + E(E(X)E(Y)) \\ &= E(X \cdot Y) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) \\ &= E(X \cdot Y) - E(X)E(Y) = Cov(X, Y), \end{aligned}$$

dus hebben we

$$Cov(X, Y) = E((X - E(X))(Y - E(Y)).$$

We zullen in de volgende les uitgebreid bediscussiëren wat het betekent dat twee stochasten *onafhankelijk* zijn, maar intuïtief zou men al zeggen, dat de uitkomst van de ene stochast de uitkomst van de andere niet mag beïnvloeden. We zullen twee stochasten X en Y onafhankelijk noemen, als de kans $P(X = x, Y = y)$ op de gecombineerde uitkomst $X = x$ en $Y = y$ gelijk is aan het product $P(X = x) \cdot P(Y = y)$ van de kansen op de aparte uitkomsten en als dit voor alle paren (x, y) geldt.

Stel nu dat X en Y onafhankelijke stochasten zijn, dan geldt:

$$\begin{aligned} E(X \cdot Y) &= \sum_{(x,y) \in X \times Y} x \cdot y \cdot P(X = x, Y = y) \\ &= \sum_{(x,y) \in X \times Y} x \cdot y \cdot P(X = x) \cdot P(Y = y) \\ &= (\sum_{x \in X} x \cdot P(X = x))(\sum_{y \in Y} y \cdot P(Y = y)) = E(X) \cdot E(Y). \end{aligned}$$

We hebben dus gezien:

Voor onafhankelijke stochasten X en Y geldt $E(X \cdot Y) = E(X) \cdot E(Y)$, dus $Cov(X, Y) = 0$ en dus $Var(X + Y) = Var(X) + Var(Y)$.

Waarschuwing: De omkering hiervan geldt niet. Twee stochasten kunnen covariantie 0 hebben zonder onafhankelijk te zijn.

We hebben gezien dat de covariantie $Cov(X, Y)$ in zekere zin en maat voor de afhankelijkheid van X en Y is. Er laat zich aantonen dat $|Cov(X, Y)| \leq \sigma_X \sigma_Y$ is, dus de covariantie van twee stochasten is begrensd door het product van de standaardafwijkingen van de stochasten. Met behulp van de standaardafwijkingen kunnen we dus de covariantie op waarden tussen -1 en 1 normeren. We noemen

$$\rho_{X,Y} := \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

de *correlatiecoëfficiënt* van X en Y. De waarde van de correlatiecoëfficiënt ligt tussen -1 en 1 de waarde $\rho_{X,Y} = -1$ treedt alleen maar op voor $Y = -\alpha X + \beta$ met $\alpha > 0$, de waarde $\rho_{X,Y} = 1$ alleen maar voor $Y = \alpha X + \beta$ met $\alpha > 0$. Precies gezegd geeft de correlatiecoëfficient dus aan, in hoeverre de stochasten X en Y *lineair* van elkaar afhangen, d.w.z. hoe goed zich Y door $\alpha X + \beta$ laat benaderen. Voor $\rho_{X,Y} > 0$ spreekt men van *positieve afhankelijkheid* voor $\rho_{X,Y} < 0$ van *negatieve afhankelijkheid*.

Belangrijke begrippen in deze les

- stochasten
- verwachtingswaarde
- variantie, standaardafwijking
- covariantie, correlatiecoëfficiënt

Opgaven

11. Bij een spel met een dobbelsteen win je n€ als je n dobbelt en n even is en je verliest n€ als n oneven is. Wat is de verwachtingswaarde van je winst/verlies.
12. Bij het skaat spel krijg je 10 kaarten uit een kaartspel met 32 kaarten (8 soorten, 4 kleuren). Wat is de verwachtingswaarde voor het aantal boeren dat je krijgt?
13. In een loterij heb je 70% nieten en 30% winnende lotjes. Iemand beslist zo lang lotjes te kopen tot dat hij een winnende lot krijgt, maar hooguit vijf keer. Wat kan hij voor een uitgave verwachten, als een lot 2€ kost?
14. Je koopt een nieuwe speelautomaat voor je kroeg. In de automaat draaien twee onafhankelijke wielen die in tien even grote segmenten zijn opgedeeld en volgens een gelijkverdeling in een van de segmenten stoppen. De segmenten hebben de nummers 1 t/m 10. Een speler heeft alleen maar de volgende winstmogelijkheden (bij alle andere uitkomsten verliest hij zijn inzet):
    - Als beide wielen 10 tonen wint hij 5€.

- Als beide wielen hetzelfde getal maar niet 10 tonen wint hij 2€.
- Als precies een van de wielen 10 toont wint hij 1€.

Je wilt natuurlijk winst met je automaat maken. Wat is de minimale inzet die je per spel moet vragen om een winst te kunnen verwachten?

15. Twee tennissters A en B spelen vaker tegen elkaar en gemiddeld wint A 60% van de sets. De speelsters ontmoeten elkaar op een toernooi in een best-of-five match (dus wie het eerst drie sets wint heeft gewonnen).

    (i) Wat zijn de kansen dat A in 3, 4 of 5 sets wint? Hoe zit het met B? Wat is de kans dat B überhaupt wint?

    (ii) Bereken de verwachtingswaarde voor het aantal sets die het match duurt.

    (iii) Bereken apart de verwachtingswaarden voor het aantal sets in het geval dat A wint en dat B wint.

    (iv) Bereken de spreiding en de standaardafwijking voor het aantal sets die het match duurt: onafhankelijk van wie er wint, als A wint en als B wint.

Les 4 Voorwaardelijke kansen, de Bayes regel en onafhankelijkheid

Sommige vragen uit de kanstheorie hebben een antwoord dat niet met de intuïtie van iedereen klopt. Een voorbeeld hiervoor is het *Monty-Hall probleem* ook bekend als *Geitenprobleem*:

> Bij een TV-show valt er voor de kandidaat een auto te winnen. Het enige wat de kandidaat moet doen is uit drie deuren de goede deur te kiezen waar de auto achter staat. Achter de andere twee deuren zijn er geiten. Nadat de kandidaat een deur heeft gekozen, wordt deze niet meteen geopend, maar de showmaster (die weet waar de auto staat) opent een van de niet gekozen deuren en een geit blaat tegen het publiek (en de kandidaat). De vraag is nu: Is het voor de kandidaat verstandig is om bij zijn keuze te blijven, of is het gunstiger om te wisselen of maakt het niets uit.

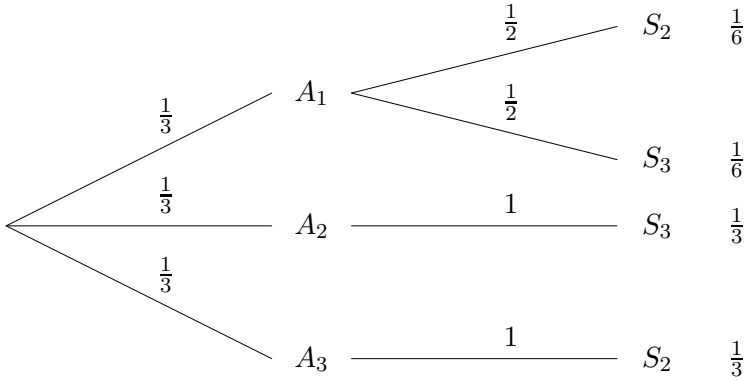
Intuïtief zullen veel mensen denken, dat na het openen van een van de deuren met een geit daarachter de kans 50 : 50 is, dat de auto achter de door de kandidaat gekozen deur staat. Dus zou het niets uitmaken of de kandidaat wisselt of niet. In de VS heeft een journaliste, Marilyn vos Savant, de oplossing voor dit probleem in haar column in de tijdschrift *Parade* gepubliceerd. Deze vrouw heeft een van de hoogste IQ's ter wereld en haar antwoord was dat de kans op de auto groeit als de kandidaat wisselt. Haar column resulteerde in een lawine van boosaardige en verontwaardigde brieven, waaronder veel van wiskundigen, die het antwoord van vos Savant bespottelijk maakten. Als reactie op dit gebeuren werd in Duitsland door de journalist Gero von Randow in de weekkrant *Die Zeit* een artikel gepubliceerd, waarin hij het geitenprobleem en een oplossing met dezelfde conclusie als die van vos Savant voorstelde. Ook hier was de reactie opmerkelijk: Over weken kwamen er brieven binnen, waarin professoren, gepromoveerde en dergelijk 'geleerden' uitlegden waarom de oplossing van vos Savant en von Randow onzin is. Ook hier waren er behoorlijk veel wiskundigen bij.

Hoe zit het nu met de oplossing van het geitenprobleem? De reden waarom veel mensen voor de 50 : 50 oplossing kiezen is dat ze ervan uit gaan, dat de situatie na het openen van een van de deuren door de showmaster onafhankelijk is van wat er eerder is gebeurd. Dit is echter niet het geval! Als de kandidaat een deur met een geit daarachter heeft gekozen, heeft de showmaster geen keuze welke deur hij gaat openen, terwijl hij in het geval dat de kandidaat de deur met de auto heft gekozen twee mogelijkheden heeft.

We kunnen dit als volgt analyseren: Stel de kandidaat heeft deur 1 gekozen. De auto kan nu achter deur 1, 2 of 3 staan, deze gevallen noemen we A_1, A_2 en A_3 en we gaan ervan uit dat elk van deze gevallen een kans van $\frac{1}{3}$ heeft. In het geval A_1 kan de showmaster deur 2 of deur 3 openen. Deze gevallen noemen we S_2 en S_3 en omdat er geen verschil tussen de deuren (en de geiten) is, kunnen we aannemen dat S_2 en S_3 dezelfde kans $\frac{1}{2}$ hebben. De kans dat de auto achter deur 1 staat en de showmaster deur 2 opent is dus $\frac{1}{6}$, hetzelfde

geldt voor het openen van deur 3. Maar in het geval A_2 heeft de showmaster geen keuze, hij moet deur 3 openen, dus is de kans voor dit geval $\frac{1}{3}$. Evenzo moet de showmaster in het geval A_3 deur 2 openen, dus is ook hier de kans $\frac{1}{3}$.

Deze situatie kunnen we door het volgende boomdiagram beschrijven:



In het geval dat de showmaster deur 2 heeft geopend is de kans dus twee keer zo groot dat de auto achter deur 3 staat dan dat hij achter deur 1 staat. Hetzelfde geldt voor het geval dat de showmaster deur 3 heeft geopend. In elk geval is het dus verstandig dat de kandidaat van keuze verandert, want hierdoor wordt zijn kans op de auto twee keer zo groot.

We zullen later nog eens op het geitenprobleem terug komen en het antwoord uit de regel van Bayes afleiden. Maar eerst gaan we algemeen naar het probleem kijken dat de kans voor een uitkomst kan veranderen als aanvullende informatie over gerelateerde gebeurtenissen bekend wordt.

4.1 Voorwaardelijke kansen

Het idee dat de kans voor een uitkomst kan veranderen als we aanvullende informatie hebben, is zo natuurlijk dat we er meestal niet over nadenken. Bijvoorbeeld kan de kans op vorst op 30 april over de afgelopen 150 jaar eenvoudig afgelezen worden uit de tabellen van de weerkundige dienst. Als er bijvoorbeeld 10 keer in de afgelopen 150 jaren vorst op 30 april was, kunnen we aannemen dat de kans op vorst op 30 april 2005 ongeveer 6.67% is. Als aanvullende informatie kunnen we gebruiken dat er ook 10 keer vorst op 29 april is geweest en dat er in 5 jaren vorst op 29 en 30 april gevallen is. Zo ver maakt dit nog geen verschil voor de kans op vorst op 30 april 2005. Maar als er inderdaad vorst op 29 april 2005 valt, kunnen we zeggen dat de kans op vorst op 30 april 2005 opeens 50% is, want in 5 van de 10 jaren met vorst op 29 april was er ook vorst op 30 april.

De kans dat er vorst op 30 april valt, gegeven het feit dat er vorst op 29 april is, noemen we een *voorwaardelijke kans*.

Abstract gaan we dit zo beschrijven: Stel we willen de kans van $A \subseteq \Omega$ bepalen onder de voorwaarde dat $B \subseteq \Omega$ plaats vindt. Deze kans definiëren we als de kans dat A *en* B gebeuren, gegeven het feit dat B gebeurt. Als de kansen door relatieve frequenties gegeven zijn, dus $P(A) = \frac{|A|}{|\Omega|}$, hebben

we $\frac{|A\cap B|}{|B|} = \frac{\frac{|A\cap B|}{|\Omega|}}{\frac{|B|}{|\Omega|}} = \frac{P(A\cap B)}{P(B)}$ en het laatste nemen we als definitie voor de voorwaardelijke kans:

Voor een kansverdeling P of Ω en $B \subseteq \Omega$ met $P(B) > 0$ noemen we

$$P(A \mid B) := \frac{P(A,B)}{P(B)} := \frac{P(A \cap B)}{P(B)}$$

de *voorwaardelijke kans op A gegeven B*.

Notatie: De kans voor het gemeenschappelijke optreden van de gebeurtenissen A en B wordt meestal met $P(A,B)$ in plaats van $P(A \cap B)$ genoteerd.

Om te rechtvaardigen, dat we $P(A \mid B)$ een *kans* noemen, moeten we even aantonen dat $P(\cdot \mid B)$ voor $P(B) > 0$ een kansverdeling is, waarbij we natuurlijk gebruiken dat $P(\cdot)$ al een kansverdeling is. (Voor $P(B) = 0$ is het onzin een kans onder de voorwaarde B te bekijken, want B gebeurt nooit.)

(i) $P(A \mid B) = \frac{P(A\cap B)}{P(B)} \geq 0$.

(ii) $P(\Omega \mid B) = \frac{P(\Omega\cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$.

(iii) Voor $A_1, A_2 \subseteq \Omega$ met $A_1 \cap A_2 = \emptyset$ geldt $(A_1 \cup A_2) \cap B = (A_1 \cap B) \cup (A_2 \cap B)$. Verder is $(A_1 \cap B) \cap (A_2 \cap B) = \emptyset$ omdat $A_1 \cap B$ een deelverzameling van A_1 en $A_2 \cap B$ een deelverzameling van A_2 is. Daarom geldt:
$P(A_1 \cup A_2 \mid B) = \frac{P((A_1\cup A_2)\cap B)}{P(B)} = \frac{P((A_1\cap B)\cup(A_2\cap B))}{P(B)} = \frac{P(A_1\cap B)+P(A_2\cap B)}{P(B)} = \frac{P(A_1\cap B)}{P(B)} + \frac{P(A_2\cap B)}{P(B)} = P(A_1 \mid B) + P(A_2 \mid B)$.

Voorbeeld: Hier is een typisch voorbeeld van een vraag die met voorwaardelijke kansen te maken heeft:

Aan 1000 werknemers wordt gevraagd of ze een hoog of een laag salaris hebben. Van de werknemers geven 210 vrouwen aan een hoog salaris te hebben en 360 geven aan een laag salaris te hebben. Van de mannen blijken 210 een hoog en 220 een laag salaris te hebben. Deze gegevens vinden we in het volgende schema terug:

	hoog salaris	laag salaris	som
vrouw	0.21	0.36	0.57
man	0.21	0.22	0.43
totaal	0.42	0.58	1.00

De vraag is nu of vrouwen en mannen dezelfde kans op een hoog salaris hebben. De kans voor een vrouw om een hoog salaris te hebben is de voorwaardelijke kans $P(hoog \mid vrouw) = \frac{P(hoog \text{ en } vrouw)}{P(vrouw)} = \frac{0.21}{0.57} \approx 0.37$. Voor mannen is de kans $P(hoog \mid man) = \frac{P(hoog \text{ en } man)}{P(man)} = \frac{0.21}{0.43} \approx 0.49$ dus hebben mannen in dit voorbeeld een behoorlijk grotere kans op een hoog salaris dan vrouwen.

We kunnen voorwaardelijke kansen niet alleen maar voor twee deelverzamelingen van Ω maar ook algemeen voor n deelverzamelingen definiëren. Het

idee hierbij is hetzelfde, we kijken naar de kans van het gemeenschappelijke optreden van de voorwaarden met een gebeurtenis, gedeeld door de kans voor de voorwaarden en krijgen dus:

$$P(A_n \mid A_1 \cap \ldots \cap A_{n-1}) = P(A_n \mid A_1, \ldots, A_{n-1}) = \frac{P(A_1, \ldots, A_n)}{P(A_1, \ldots, A_{n-1})}.$$

We hebben dus bijvoorbeeld $P(A_3 \mid A_1, A_2) = \frac{P(A_1,A_2,A_3)}{P(A_1,A_2)}$.

Omgekeerd kunnen we de kans voor het gemeenschappelijke optreden van gebeurtenissen (iteratief) door voorwaardelijke kansen uitdrukken en krijgen zo de zogeheten *kettingregel*:
$P(A_1, A_2) = P(A_2 \mid A_1) \cdot P(A_1)$,
$P(A_1, A_2, A_3) = P(A_3|A_1, A_2) \cdot P(A_1, A_2) = P(A_3|A_1, A_2) \cdot P(A_2|A_1) \cdot P(A_1)$,
en in het algemeen
$P(A_1, ..., A_n) = P(A_n|A_1, ..., A_{n-1}) \cdot P(A_{n-1}|A_1, ..., A_{n-2}) \cdot \ldots \cdot P(A_2|A_1) \cdot P(A_1)$.

4.2 Regel van Bayes

Omdat de doorsnede $A \cap B$ symmetrisch in A en B is, vinden we uit de definitie voor de voorwaardelijke kans dat

$$P(A \mid B) \cdot P(B) = P(A \cap B) = P(B \mid A) \cdot P(A)$$

en dit geeft de eenvoudigste vorm van de *regel van Bayes*, namelijk

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A)}.$$

De nut van deze regel ligt in het omdraaien van de rollen van voorwaarde en uitkomst. Denk hierbij bijvoorbeeld aan een test op een ziekte. Als de uitslag van de test gegeven is, zijn we geïnteresseerd in de kans dat we de ziekte hebben of niet. Maar bekend is alleen maar de nauwkeurigheid van de test die zegt met welke kans de test bij een gezonde mens het verkeerde resultaat geeft en andersom.

De regel van Bayes wordt vaak op een iets slimmere manier toegepast. Hiervoor wordt de deelverzameling $B \subseteq \Omega$ in verschillende gevallen onderverdeeld die elkaar uitsluiten, dus we schrijven $B = \cup_{i=1}^{n} B_i$ met $B_i \cap B_j = \emptyset$ als $i \neq j$. Een belangrijk speciaal geval hiervoor is $B = B_1 \cup B_2$ met $B_2 = B \setminus B_1 = B_1^c$. We noemen B_2 het *complement* van B_1 in B.

Er geldt:

$$P(A \cap B) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i)$$

en dus

$$P(A \mid B) = \frac{1}{P(B)} \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i).$$

In het bijzonder kunnen we in het geval dat $A \subseteq B$ de *totale kans* $P(A)$ berekenen als $P(A) = P(A \cap B) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i)$ en het belangrijkste geval hiervoor is $B = \Omega$, d.w.z. we delen alle mogelijke uitkomsten in een aantal klassen van uitkomsten op.

We kunnen nu de regel van Bayes algemeen formuleren:

Regel van Bayes: Zij $B \subseteq \Omega$ met $B = \cup_{i=1}^{n} B_i$ en $B_i \cap B_j = \emptyset$ als $i \neq j$. Verder zij $A \subseteq B$. Dan geldt

$$P(B_j \mid A) = \frac{P(A \mid B_j) \cdot P(B_j)}{P(A)} = \frac{P(A \mid B_j) \cdot P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i)}$$

Om de abstracte concepten duidelijk te maken, passen we de regel van Bayes op een aantal voorbeelden toe.

Voorbeeld 1: De uitkomst van een HIV-test noemen we A als de test positief was en A^c als de test negatief was. Het geïnfecteerd zijn noemen we I en het niet geïnfecteerd zijn I^c. Over de kwaliteit van de test is bekend, dat hij voor geïnfecteerden in 99% van de gevallen een positief resultaat oplevert en voor niet geïnfecteerden in 99.9% van de gevallen een negatief resultaat. We hebben dus $P(A \mid I) = 0.99$, $P(A^c \mid I) = 0.01$ en $P(A^c \mid I^c) = 0.999$, $P(A \mid I^c) = 0.001$. Verder nemen we aan dat 1 uit 10000 mensen HIV-geïnfecteerd is, dus $P(I) = 0.0001$ en $P(I^c) = 0.9999$. De vraag is nu, hoe groet bij een positieve HIV-test de kans is, inderdaad geïnfecteerd te zijn, dus hoe groot de voorwaardelijke kans $P(I \mid A)$ is. Met de regel van Bayes hebben we

$$P(I \mid A) = \frac{P(A \mid I) \cdot P(I)}{P(A)} = \frac{P(A \mid I) \cdot P(I)}{P(A \mid I) \cdot P(I) + P(A \mid I^c) \cdot P(I^c)}$$

$$= \frac{0.99 \cdot 0.0001}{0.99 \cdot 0.0001 + 0.001 \cdot 0.9999} \approx 9.0\%.$$

Deze verrassend lage kans is opmerkelijk maar toch goed te begrijpen. Als we 10000 mensen testen, dan is er gemiddeld 1 HIV-geïnfecteerde mens bij en die krijgt waarschijnlijk ook een positieve test-uitslag. Maar bij de 9999 niet-geïnfecteerden zal de test in 0.1% van de gevallen een (verkeerd) positief resultaat opleveren, dus komen er nog 10 positieve resultaten bij. Als we dus naar de 11 positieve resultaten kijken, is dit alleen maar in één geval veroorzaakt door een geïnfecteerde, maar in 10 gevallen door een test-fout.

Merk op dat er in dit soort vragen vaak verkeerd geargumenteerd wordt. Dit vind je zelfs in wetenschappelijke publicaties, bijvoorbeeld in de medicijn of in de rechtsgeleerdheid terug. Denk hier bijvoorbeeld aan een misdadiger waarbij de schuld door een DNA-analyse wordt bewezen. Het probleem is, dat zelfs bij een test met een hoge nauwkeurigheid het aantal verkeerde uitslagen vaak hoger is dan het aantal van de gezochte zeldzame uitkomsten.

Voorbeeld 2: Een student moet bij een tentamen een multiple-choice vraag met n mogelijkheden oplossen. Als hij voorbereid is, zal zijn antwoord juist zijn, als niet zal hij willekeurig een antwoord gokken en dus een kans van $\frac{1}{n}$ op een juiste antwoord hebben. De kans dat de student voorbereid is, zij p. Voor de

docent is het nu interessant om de kans te bepalen, dat de student inderdaad voorbereid was, als hij een juiste antwoord heeft gegeven. Als we een juiste antwoord met J en een voorbereide student met V betekenen hebben we dus:

$$P(V \mid J) = \frac{P(J \mid V) \cdot P(V)}{P(J \mid V) \cdot P(V) + P(J \mid V^c) \cdot P(V^c)}$$

$$= \frac{1 \cdot p}{1 \cdot p + \frac{1}{n}(1-p)} = \frac{np}{np + (1-p)}.$$

Het is duidelijk dat dit voor grote waarden van n dicht bij 1 ligt, want dan is $(1-p)$ tegen np te verwaarlozen. Maar voor $n = 4$ en $p = 0.5$ hebben we bijvoorbeeld $P(V \mid J) = \frac{4}{5} = 80\%$ en voor $n = 4$ en $p = 0.2$ geldt al $P(V \mid J) = \frac{1}{2} = 50\%$. Als de docent dus weet dat gewoon maar een vijfde van de studenten voorbereid is, weet hij ook dat de helft van de goede antwoorden goede gokken zijn.

Voorbeeld 3: In de automatische spraakherkenning gaat het erom, gegeven een akoestisch signaal X het woord w te vinden dat hier het beste bij past, d.w.z. waarvoor de voorwaardelijke kans $P(w \mid X)$ maximaal is. Hiervoor gebruiken we ook de regel van Bayes en schrijven

$$P(w \mid X) = \frac{P(X \mid w) \cdot P(w)}{P(X)}.$$

Omdat we alleen maar aan het woord met de hoogste kans geïnteresseerd zijn, kunnen we de noemer gewoon vergeten, omdat die voor elk woord hetzelfde is. In de teller geeft $P(X \mid w)$ de kans, dat een zeker woord w tot het signaal X lijdt. Deze kans wordt tijdens het *training* van een systeem bepaald, waarbij een aantal mensen het woord spreekt en uit de zo verkregen signalen een kansverdeling geschat wordt. De kans $P(w)$ is de totale kans dat een woord gesproken wordt. Dit noemen we de a-priori kans voor het woord, en deze kansen worden als relatieve frequenties op heel grote tekst-corpora (bijvoorbeeld 10 jaar NRC Handelsblad) bepaald.

Hetzelfde principe geldt trouwens voor de meeste soorten van patroonherkenning (beeld-herkenning, handschrift-herkenning).

Voorbeeld 4: We komen nog eens terug op het Monty-Hall probleem. Stel de kandidaat heeft deur 1 gekozen, dan nemen we aan dat de showmaster deur 2 heeft geopend (S_2), het geval S_3 geeft een analoog resultaat. We zijn nu geïnteresseerd in de kansen $P(A_1 \mid S_2)$ en $P(A_3 \mid S_2)$, dus de voorwaardelijke kansen dat de auto achter deur 1 of deur 3 staat, gegeven het feit dat de showmaster deur 2 heeft geopend. Er geldt

$$P(A_1 \mid S_2) = \frac{P(S_2 \mid A_1) \cdot P(A_1)}{P(S_2 \mid A_1) \cdot P(A_1) + P(S_2 \mid A_2) \cdot P(A_2) + P(S_2 \mid A_3) \cdot P(A_3)}$$

$$= \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 0 + 1 \cdot \frac{1}{3}} = \frac{1}{3}.$$

Evenzo berekenen we de kans $P(A_3 \mid S_2)$ als

$$P(A_3 \mid S_2) = \frac{P(S_2 \mid A_3) \cdot P(A_3)}{P(S_2 \mid A_1) \cdot P(A_1) + P(S_2 \mid A_2) \cdot P(A_2) + P(S_2 \mid A_3) \cdot P(A_3)}$$

$$= \frac{1 \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 0 + 1 \cdot \frac{1}{3}} = \frac{2}{3}.$$

We zien dus weer dat het voor de kandidaat verstandig is om naar deur 3 te wisselen, omdat de kans dat de auto daar achter zit twee keer zo groot is.

4.3 Onafhankelijkheid

Nu dat we goed naar voorwaardelijke kansen hebben gekeken kunnen we ook zeggen wat het betekent dat twee uitkomsten onafhankelijk zijn. Intuïtief zullen we zeggen, dat twee uitkomsten A en B onafhankelijk zijn, als de kans voor A niet ervan afhangt of B optreed of niet. Met de voorwaardelijke kans kunnen we dit zo formuleren:

Twee uitkomsten A en B heten onafhankelijk als $P(A) = P(A \mid B)$.
Equivalent hiermee is dat $P(A \cap B) = P(A) \cdot P(B)$.

De equivalentie van de twee formuleringen volgt uit de definitie van de voorwaardelijke kans, want $P(A \cap B) = P(A \mid B) \cdot P(B)$ geeft $P(A) = P(A \mid B) \Leftrightarrow P(A \cap B) = P(A \mid B) \cdot P(B) = P(A) \cdot P(B)$. Omdat ook $P(A \cap B) = P(B \mid A) \cdot P(A)$ geldt, volgt hieruit ook dat $P(A) = P(A \mid B) \Leftrightarrow P(B) = P(B \mid A)$, dus het maakt niets uit welke voorwaardelijke kans we bekijken.

Een eenvoudig voorbeeld zijn de soorten en kleuren in een kaartspel. De kans om uit een kaartspel met 52 kaarten een aas te trekken is $\frac{1}{13}$, de kans om een kaart van kleur klaver te trekken is $\frac{1}{4}$. De doorsnede van de uitkomsten *aas* en *klaver* is alleen maar de kaart *klaver aas* en de kans om deze kaart te trekken is $\frac{1}{52} = \frac{1}{13} \cdot \frac{1}{4}$. Omdat we ook elke andere soort of kleur hadden kunnen kiezen, toont dit aan, dat de soorten en de kleuren onafhankelijk zijn.

In een ander voorbeeld kijken we naar een familie met twee kinderen. We vragen ons af of de uitkomsten

A : er is een meisje en een jongen $\qquad$ B : er is hoogstens een meisje

onafhankelijk zijn. Als we m voor een meisje en j voor een jongen schrijven, zijn de mogelijkheden voor de twee kinderen (m, m), (m, j), (j, m) en (j, j). We zien makkelijk dat $P(A) = \frac{1}{2}$ en $P(B) = \frac{3}{4}$, maar $P(A \cap B) = \frac{1}{4} \neq \frac{1}{2} \cdot \frac{3}{4} = \frac{3}{8}$. Dus zijn de uitkomsten A en B niet onafhankelijk.

Als we de familie nu van twee naar drie kinderen uitbreiden maar dezelfde uitkomsten bekijken, is de situatie veranderd. De mogelijkheden voor de drie kinderen zijn nu (m, m, m), (m, j, m), (j, m, m), (j, j, m), (m, m, j), (m, j, j), (j, m, j) en (j, j, j). In dit geval is $P(A) = \frac{3}{4}$, $P(B) = \frac{1}{2}$ en $P(A \cap B) = \frac{3}{8} = P(A) \cdot P(B)$, dus zijn de uitkomsten nu inderdaad onafhankelijk.

Aan de hand van dit voorbeeld zien we, dat soms uitkomsten kanstheoretisch onafhankelijk zijn, die we in het echte leven niet onafhankelijk zouden noemen.

De onafhankelijkheid van uitkomsten A en B heeft ook nuttige consequenties voor de complementen A^c en B^c. Er geldt namelijk dat met (A, B) ook de paren (A, B^c), (A^c, B) en (A^c, B^c) onafhankelijk zijn. Dit kunnen we met behulp van een paar eenvoudige manipulaties van de betrokken verzamelingen uit $P(A \cap B) = P(A) \cdot P(B)$ concluderen:
$P(A \cap B^c) = P(A \cup B) - P(B) = P(A) + P(B) - P(A \cap B) - P(B) = P(A) - P(A \cap B) = P(A) - P(A) \cdot P(B) = P(A)(1 - P(B)) = P(A) \cdot P(B^c)$. Dit werkt evenzo voor $P(A^c \cap B)$.
$P(A^c \cap B^c) = P((A \cup B)^c) = 1 - P(A \cup B) = 1 - P(A) - P(B) + P(A \cap B) = 1 - P(A) - P(B) + P(A) \cdot P(B) = (1 - P(A))(1 - P(B)) = P(A^c) \cdot P(B^c)$.

We kunnen het begrip van onafhankelijkheid ook naar stochasten uitbreiden: Voor twee stochasten X, Y zij $A_x := \{\omega \in \Omega \mid X(\omega) = x\}$ en $B_y := \{\omega \in \Omega \mid Y(\omega) = y\}$. We noemen de uitkomsten A_x en B_y onafhankelijk als $P(A_x \cap B_y) = P(A_x) \cdot P(B_y)$. Maar in de taal van stochasten heet dit dat

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$$

en we noemen twee stochasten X, Y onafhankelijk als dit voor alle paren (x, y) geldt.

Tot nu toe hebben we het alleen maar over de onafhankelijkheid van *twee* uitkomsten gehad. Als we meerdere uitkomsten bekijken, zijn er verschillende mogelijkheden om hun onafhankelijkheid te definiëren:

(1) We noemen de n uitkomsten $A_1, \ldots, A_n$ *paarsgewijs onafhankelijk* als $P(A_i \cap A_j) = P(A_i) \cdot P(A_j)$ voor alle $i \neq j$.

(2) We noemen n uitkomsten $A_1, \ldots, A_n$ *onafhankelijk* als $P(A_{i_1} \cap \ldots \cap A_{i_k}) = P(A_{i_1}) \cdot \ldots \cdot P(A_{i_k})$ voor elke deelverzameling $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$.

Als we de begrippen op deze manier definiëren is het duidelijk dat onafhankelijke uitkomsten ook paarsgewijs onafhankelijk zijn. Het omgekeerde geldt niet, wat aan het volgende tegenvoorbeeld duidelijk wordt:

We dobbelen met twee dobbelstenen en bekijken de kansen van de volgende uitkomsten:

A_1 : de eerste dobbelsteen toont een oneven getal,

A_2 : de tweede dobbelsteen toont een oneven getal,

A_3 : de som van de getallen is even.

We hebben $P(A_1) = P(A_2) = P(A_3) = \frac{1}{2}$ en $P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) = \frac{1}{4}$, dus zijn de uitkomsten paarsgewijs onafhankelijk. Maar $P(A_1 \cap A_2 \cap A_3) = P(A_1 \cap A_2)$ omdat de som van twee oneven getallen even

is, dus is $P(A_1 \cap A_2 \cap A_3) \neq P(A_1) \cdot P(A_2) \cdot P(A_3) = \frac{1}{8}$ en dus zijn de drie uitkomsten niet onafhankelijk.

We zouden bij de definitie van onafhankelijkheid voor meerdere uitkomsten ook kunnen hopen dat het voldoende is om $P(A_1 \cap \ldots \cap A_n) = P(A_1) \cdot \ldots \cdot P(A_n)$ te eisen, maar het volgende tegenvoorbeeld laat zien dat hieruit niet eens volgt dat de A_i paarsgewijs onafhankelijk zijn: We werpen een munt drie keer en kijken naar de volgende uitkomsten:

A_1 : de eerste worp toont kop,

A_2 : er valt vaker kop dan munt,

A_3 : de laatste twee worpen leveren hetzelfde resultaat.

Door naar de mogelijke uitkomsten te kijken zien we dat $P(A_1) = P(A_2) = P(A_3) = \frac{1}{2}$ en dat $P(A_1 \cap A_2 \cap A_3) = \frac{1}{8}$. Aan de andere kant hebben we $P(A_1 \cap A_2) = \frac{3}{8}$, dus zijn A_1 en A_2 niet (paarsgewijs) onafhankelijk. De andere paren zijn wel onafhankelijk, want $P(A_1 \cap A_3) = P(A_2 \cap A_3) = \frac{1}{4}$.

4.4 Bernoulli-model

Een belangrijke toepassing van de onafhankelijkheid van uitkomsten is de herhaalde uitvoering van een experiment. We nemen aan dat we in de uitkomstenruimte Ω een deelverzameling $A \subseteq \Omega$ van gunstige uitkomsten hebben. Bij de eenmalige uitvoering van het experiment is de kans op een gunstige uitkomst gegeven door $p = \frac{|A|}{|\Omega|}$. De kans voor een ongunstige uitkomst is dan $1-p$. Als we het experiment twee keer uitvoeren is de kans dat we twee gunstige uitkomsten hebben de kans van de doorsnede van een gunstige uitkomst bij de eerste keer en een gunstige uitkomst bij de tweede keer. Omdat we ervan uitgaan dat het eerste en het tweede experiment onafhankelijk zijn, kunnen we de kans voor de doorsnede als product van de enkele kansen berekenen, dus als $p \cdot p = p^2$.

Merk op dat de eis dat herhalingen van een experiment onafhankelijk zijn een voorwaarde voor de opzet van het experiment is. Als je bijvoorbeeld de kans wilt bepalen waarmee een vaccinatie tot de uitbraak van een ziekte lijdt mag je bij het herhalen van het experiment geen mensen nemen die al bij de vorige keer gevaccineerd zijn, omdat deze een hoger aantal antilichamen hebben en dus een kleinere kans lopen dat de ziekte uitbreekt.

Als we ervan uitgaan dat het herhalen van een experiment onafhankelijke uitkomsten heeft, dan is de kans dat we bij m herhalingen k keer een gunstige uitkomst hebben gegeven door de binomiale verdeling:

$$b(m, p; k) = \binom{m}{k} p^k (1-p)^{m-k}.$$

De kans dat de eerste k uitkomsten gunstig zijn is namelijk p^k en de kans dat de laatste $m-k$ uitkomsten ongunstig zijn is $(1-p)^{m-k}$. Nu kunnen we de gunstige uitkomsten nog op $\binom{m}{k}$ manieren over de m experimenten verdelen.

De beschrijving van uitkomsten door onafhankelijke herhaling van een experiment noemen we het *Bernoulli-model*.

Belangrijke begrippen in deze les

- voorwaardelijke kans
- regel van Bayes
- onafhankelijkheid, paarsgewijs onafhankelijk
- Bernoulli-model

Opgaven

16. Een socioloog wil de kans bepalen dat mensen een keer een winkeldiefstal hebben gepleegd. Omdat mensen op een rechtstreekse vraag waarschijnlijk niet eerlijk zouden antwoorden heeft hij de volgende opzet verzonnen: Elke persoon krijgt 10 kaarten waarvan op 4 de vraag staat:
Heb je ooit een winkeldiefstal gepleegd?
en op de andere 6 de vraag
Heb je nog nooit een winkeldiefstal gepleegd?
De mensen worden nu gevraagd om toevallig één van de tien kaarten te trekken, het (waarheidsgetrouwe) antwoord op een briefje te schrijven en alleen maar dit briefje aan de onderzoeker te geven. Zo hoeft niemand om zijn anonimiteit te vrezen.

    Bij 1000 testpersonen krijgt de onderzoeker 516 keer het antwoord *ja* en 484 keer het antwoord *nee*. Hoe kan hij nu de gezochte kans berekenen en wat is deze kans?

17. Er wordt met twee dobbelstenen gedobbeld. Gegeven de informatie dat de twee dobbelstenen verschillende getallen tonen (bijvoorbeeld in een spel waar je bij gelijke getallen nog een keer dobbelt), wat is de kans dat de som oneven is?

18. In een zak zitten drie munten, waarvan twee eerlijk zijn maar de derde heeft twee kop-zijden. Er wordt blindelings een munt getrokken, vervolgens wordt deze munt twee keer geworpen, waarbij twee keer kop valt. Bepaal de kans, dat de getrokken munt een eerlijke munt is.

    Hoe zit het met het geval dat in de zaak een miljoen in plaats van drie munten zitten, waarvan weer één oneerlijk is. Nu werp je twintig keer in plaats van twee keer en krijgt twintig keer het resultaat kop. Hoe groot is nu de kans dat de getrokken munt een eerlijke munt is.

19. In sommige studies is er na het eerste semester een advies aan de studenten die weliswaar niet bindend is. Neem aan dat in een (zware) studie gemiddeld 40% van de studenten vroegtijdig afhaken. Het blijkt dat van de afhakende studenten 90% een negatief studieadvies kregen, terwijl slechts 1% van de studenten die afstuderen een negatief advies hadden. Wat is de kans dat een student met negatief studieadvies wel in dit vak zou afstuderen?

20. Bij een rechtbank zal een leugendetector geraadpleegd worden. Het is bekend dat voor een schuldige verdachte de detector in 90% van de gevallen het juiste resultaat (schuldig) geeft en voor een onschuldige verdachte in 99% van de gevallen het resultaat onschuldig. Uit de statistieken van de belastingdienst is bekend dat 5% van de burgers in hun belastingaangifte ernstig bedriegen. Bij een verdachte geeft de leugendetector aan dat de man/vrouw schuldig is. Wat is de kans, dat de verdachte toch onschuldig is?

Les 5 Schatten en simuleren

5.1 Maximum likelihood schatting

Tot nu toe hebben we meestal naar voorbeelden gekeken waar we van een kansverdeling zijn uitgegaan en dan voorspellingen hebben gemaakt. In de praktijk komen we echter vaak een iets andere situatie tegen. We weten dat er iets volgens een zekere kansverdeling zal gebeuren, maar deze hangt van een parameter af die we niet kennen. Bijvoorbeeld kunnen we aannemen dat de kans p waarmee een machine defecte stukken produceert constant is, maar dat we de waarde van p niet kennen. Als we nu in een steekproef defecte stukken tellen, kunnen we het aantal defecte stukken door de binomiale (of hypergeometrische) verdeling beschrijven. Wat we nu nodig hebben is een *schatting* voor de kans p, gegeven de aantallen van defecte stukken in een paar steekproeven. Neem aan dat we altijd een steekproef van m stukken nemen, dan vinden we in de verschillende steekproeven $k_1, k_2, \ldots, k_n$ defecte stukken. We kunnen nu op verschillende manieren een waarde voor p schatten, bijvoorbeeld:

- simplistisch: We schatten $p = \frac{k_1}{m}$, dus we nemen aan dat de eerste steekproef typisch was en negeren de anderen (dit kunnen we natuurlijk ook met k_3 of k_n in plaats van k_1 doen).
- optimistisch: We schatten $p = \frac{k_{min}}{m}$, waarbij k_{min} de minimale waarde van de k_i is.
- pessimistisch: We schatten $p = \frac{k_{max}}{m}$, waarbij k_{max} de maximale waarde van de k_i is.
- pragmatisch: We schatten $p = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} k_i}{m}$, dus we nemen het gemiddelde van de relatieve frequenties in de enkele steekproeven.

Een algemene methode om parameters van kansverdelingen te schatten is gebaseerd op het volgende argument: Voor elke keuze van een parameter (of meerdere parameters) heb je een kansverdeling, die aan een waargenomen resultaat een zekere kans geeft. In het voorbeeld is dit $P_p(X = k) = b(m, p; k) = \binom{m}{k} p^k (1-p)^{m-k}$. Bij onafhankelijke herhaling kunnen we de kans voor een rij observaties als product van de kansen voor de aparte observaties berekenen, in het voorbeeld hebben we dus

$$P_p(k_1, \ldots, k_n) = \prod_{i=1}^{n} \binom{m}{k_i} p^{k_i} (1-p)^{m-k_i}.$$

De kans voor de observaties is nu een functie van de parameter p en we noemen deze functie de *aannemelijkheidsfunctie* of *likelihood* functie. We maken nu een schatting voor p door te zeggen, dat we p zo kiezen dat de aannemelijkheidsfunctie aan maximale waarde heeft, dus dat de kans voor onze observatie maximaal wordt. Deze methode van schatting noemt men de *meest aannemelijke* of *maximum likelihood* schatting van de parameter.

Om een maximum likelihood schatting uit te werken, moeten we in principe de functie $P_p(k_1, \ldots, k_n)$ naar p afleiden en de nulpunten van de afgeleide bepalen. Omdat de kans een product van de enkele kansen is, zal het afleiden een hele hoop termen opleveren, want we moeten altijd de productregel toepassen. Hier is het volgende trucje vaak erg handig: In plaats van het maximum van $P_p(k_1, \ldots, k_n)$ te berekenen, bepalen we het maximum van $\log(P_p(k_1, \ldots, k_n))$. Dit zit namelijk op dezelfde plek, omdat de logaritme een monotone functie is. De (negatieve) logaritme van de kans noemt men soms ook de *score* van een uitkomst.

We gaan nu de maximum likelihood schatting voor een aantal kansverdelingen uitwerken:

Binomiale verdeling: In n steekproeven van grootte $m_1, \ldots, m_n$ vinden we $k_1, \ldots, k_n$ gunstige uitkomsten. We hebben

$$P_p(k_1, \ldots, k_n) = \prod_{i=1}^{n} \binom{m_i}{k_i} p^{k_i}(1-p)^{m_i - k_i}.$$

We definiëren nu

$$L(p) = \log(P_p(k_1, \ldots, k_n)) = \sum_{i=1}^{n} \left(\log(\binom{m_i}{k_i}) + \log(p^{k_i}) + \log((1-p)^{m_i-k_i}) \right)$$

$$= \sum_{i=1}^{n} \log(\binom{m_i}{k_i}) + \sum_{i=1}^{n} k_i \log(p) + \sum_{i=1}^{n} (m_i - k_i) \log(1-p).$$

De afgeleide (met betrekking tot p) hiervan is

$$L'(p) = \frac{1}{p}(\sum_{i=1}^{n} k_i) - \frac{1}{1-p}(\sum_{i=1}^{n}(m_i - k_i)).$$

We hebben

$$L'(p) = 0 \Leftrightarrow (1-p)(\sum_{i=1}^{n} k_i) = p(\sum_{i=1}^{n}(m_i - k_i)) \Leftrightarrow \sum_{i=1}^{n} k_i = p(\sum_{i=1}^{n} m_i)$$

$$\Leftrightarrow p = \frac{\sum_{i=1}^{n} k_i}{\sum_{i=1}^{n} m_i}.$$

Dit betekent dat we de parameter als de relatieve frequentie van gunstige uitkomsten in alle steekproeven bij elkaar kiezen. Dit komt op de pragmatische keuze neer, maar we hebben nu een betere onderbouwing voor onze keuze. Het is namelijk de parameter die de observaties het beste verklaart.

Poisson-verdeling: Een zeldzaam gebeurtenis zien we $k_1, \ldots, k_n$ keer gebeuren. We hebben

$$P_\lambda(k_1, \ldots, k_n) = \prod_{i=1}^{n} \frac{\lambda^{k_i}}{k_i!} e^{-\lambda}.$$

We definiëren nu

$$L(\lambda) = \log(P_\lambda(k_1, \ldots, k_n)) = \sum_{i=1}^{n} (\log(\lambda^{k_i}) - \log(k_i!) - \lambda)$$

$$= \sum_{i=1}^{n} k_i \log(\lambda) - \sum_{i=1}^{n} \log(k_i!) - n\lambda.$$

De afgeleide hiervan is

$$L'(\lambda) = \frac{1}{\lambda}(\sum_{i=1}^{n} k_i) - n$$

en we hebben

$$L'(\lambda) = 0 \Leftrightarrow \lambda = \frac{1}{n}(\sum_{i=1}^{n} k_i).$$

De schatting voor de verwachtingswaarde λ van de Poisson-verdeling is dus het rekenkundig gemiddelde van de aantallen geobserveerde zeldzame gebeurtenissen. Ook dit klopt met onze intuïtie, dat we na een aantal pogingen aannemen, dat we vervolgens ook weer gebeurtenissen met ongeveer hetzelfde gemiddelde zullen krijgen.

Normaalverdeling: De normaalverdeling wordt in de opgaven behandeld.

Exponentiële verdeling: Voor een gebeurtenis dat volgens een exponentiële verdeling optreedt maken we de observaties $x_1, \ldots, x_n$. Er geldt

$$f_\lambda(x_1, \ldots, x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda(\sum_{i=1}^{n} x_i)}.$$

Merk op dat we het hier met een dichtheidsfunctie voor de kansverdeling te maken hebben. Maar we kunnen aannemen, dat we steeds een klein interval rond een geobserveerde waarde bekijken, dan is de kans voor een observatie in het interval $[x, x+\delta]$ gegeven door $P_\lambda(X \in [x, x+\delta]) = \lambda e^{-\lambda x}\delta$. Maar δ heeft als constante factor geen invloed op het maximum van de functie, dus kunnen we meteen naar de dichtheidsfunctie kijken.

We definiëren nu

$$L(\lambda) = \log(f_\lambda(x_1, \ldots, x_n)) = \log(\lambda^n) - \lambda(\sum_{i=1}^{n} x_i) = n\log(\lambda) - \lambda(\sum_{i=1}^{n} x_i).$$

De afgeleide hiervan is

$$L'(\lambda) = \frac{n}{\lambda} - (\sum_{i=1}^{n} x_i)$$

en we hebben

$$L'(\lambda) = 0 \Leftrightarrow \frac{1}{\lambda} = \frac{1}{n}(\sum_{i=1}^{n} x_i).$$

De schatting voor de verwachtingswaarde $\frac{1}{\lambda}$ van de exponentiële verdeling is dus weer het rekenkundig gemiddelde van de observaties.

Hypergeometrische verdeling: In de eerste les hebben we al het voorbeeld bekeken dat we het aantal vissen in een vijver willen bepalen. Het idee hiervoor is, dat we s vissen markeren en dan kijken hoeveel gemarkeerde vissen we in een (latere) steekproef van m vissen vinden. De kans dat we er k gemarkeerde vissen in vinden is gegeven door de hypergeometrische verdeling

$$h(n, m, s; k) = \frac{\binom{s}{k}\binom{n-s}{m-k}}{\binom{n}{m}}$$

waarbij n het onbekende aantal vissen in de vijver is. In dit voorbeeld gaan we niet de logaritme gebruiken, maar bepalen we het maximum van $h(n, m, s; k)$ als een functie van n op een andere manier. We kijken naar de quotiënt

$$q(n) := \frac{h(n, m, s; k)}{h(n-1, m, s; k)}.$$

Als $q(n) \geq 1$ is $h(n, m, s; k)$ stijgend, als $q(n) \leq 1$ is $h(n, m, s; k)$ dalend. We hebben

$$q(n) = \frac{\binom{s}{k}\binom{n-s}{m-k}}{\binom{n}{m}} \cdot \frac{\binom{n-1}{m}}{\binom{s}{k}\binom{n-1-s}{m-k}} = \frac{\frac{(n-s)!}{(m-k)!(n-s-m+k)!} \cdot \frac{(n-1)!}{m!(n-1-m)!}}{\frac{n!}{m!(n-m)!} \cdot \frac{(n-1-s)!}{(m-k)!(n-1-s-m+k)!}}$$

$$= \frac{(n-s)!(n-1)!(n-m)!(n-1-s-m+k)!}{(n-s-m+k)!(n-1-m)!n!(n-1-s)!} = \frac{(n-s)(n-m)}{(n-s-m+k)n}$$

$$= \frac{n^2 - sn - nm + sm}{n^2 - sn - mn + kn}.$$

We zien dus dat $q(n) \geq 1$ als $sm \geq kn$ en $q(n) \leq 1$ als $sm \leq kn$. Het maximum wordt dus bereikt voor $n = \frac{sm}{k}$, d.w.z. voor $\frac{k}{m} = \frac{s}{n}$. Dit betekent dat we de grootte van de populatie zo schatten dat het relatieve aantal gemarkeerde vissen in onze vangst hetzelfde is als het relatieve aantal in de hele vijver.

Merk op: In de voorbeelden die we hier hebben behandeld, kunnen we de maximum likelihood schatting expliciet uitrekenen en krijgen meestal een resultaat dat we ook intuïtief hadden verwacht. Voor ingewikkeldere kansverdelingen (bijvoorbeeld met veel parameters) is het vaak niet mogelijk de nulpunten van de partiële afgeleiden expliciet te bepalen. Hier worden dan iteratieve benaderingsmethoden toegepast, bijvoorbeeld het *EM-algoritme* (hierbij staat *EM* voor *expectation maximization*).

Er zijn ook andere schatters dan de maximum likelihood schatter, bijvoorbeeld de *momentenschatters*. Het k-de moment van een stochast X is de verwachtingswaarde $E(X^k)$ van de k-de macht van de stochast. Bij een momentenschatter wordt geprobeerd de parameters van een kansverdeling zo te bepalen dat de momenten van de kansverdeling gelijk zijn aan de momenten die in een steekproef waargenomen zijn.

We zullen ons hier niet verder in verdiepen omdat het probleem van het schatten van parameters van een kansverdeling meer in de statistiek thuis hoort.

5.2 Simulatie

Soms heb je bij experimenten na een aantal observaties een idee erover wat er gebeurt en maak je er een model om de resultaten te beschrijven. De kwaliteit van een model ligt in het vermogen om toekomstige resultaten te kunnen voorspellen en dit is ook de manier hoe een model getoetst wordt. Vaak zijn experimenten zo ingewikkeld of kostbaar dat je bij een aanpassing van het beschrijvende model niet meteen weer veel experimenten kunt of wilt doen. Dan is het handig om het nieuwe model met een simulatie te testen, waarbij je zekere parameters volgens een kansverdeling kiest.

Een andere motivatie voor het simuleren van kansverdeling is dat sommige effecten pas naar heel veel herhalingen van een experiment naar voren komen. Voor een computer is het veel makkelijker om iets 10000 keer te herhalen dan dit in de realiteit te doen, bijvoorbeeld een munt 10000 keer te werpen.

We gaan daarom in deze paragraaf bekijken hoe we voor een aantal kansverdelingen een stochast met gegeven verdelingsfunctie kunnen simuleren.

Randomgenerator

Het startpunt voor alle soorten simulaties is een *toevalsgenerator* of *randomgenerator*. Dit is een procedure (een soort orakel) die een rij getallen $U_1, U_2, \ldots$ tussen 0 en 1 produceert die aan de volgende eisen voldoet:

(1) De kansverdeling op de i-de plek in de rij is de uniforme verdeling op het interval $[0, 1)$, d.w.z. er geldt $P(U_i \leq u) = u$ voor elke i.

(2) De stochasten $U_1, U_2, \ldots$ zijn onafhankelijk, d.w.z. voor elke keuze van indices $i_1 < i_2 < \ldots < i_k$ hebben we $P(U_{i_1} \leq u_1, U_{i_2} \leq u_2, \ldots, U_{i_k} \leq u_k) = P(U_{i_1} \leq u_1) \cdot P(U_{i_2} \leq u_2) \cdot \ldots \cdot P(U_{i_k} \leq u_k) = u_1 \cdot u_2 \cdot \ldots \cdot u_k$.

Als de rij $U_1, U_2, \ldots$ van getallen aan deze eisen voldoet, noemt men de U_i *toevalsgetallen*. Helaas kan een praktische implementatie van een toevalsgenerator nooit perfect aan deze eisen voldoen, men spreekt daarom strikt genomen beter ervan dat een randomgenerator *pseudo-toevalsgetallen* en geen 'echte' toevalsgetallen produceert.

Een veel gebruikte type van randomgeneratoren zijn de *lineaire congruentie modellen*: Kies een getal $m \in \mathbb{N}$, constanten $a, c \in \mathbb{Z}$ en een *zaad* (Engels: seed) I_0. Vervolgens bereken je iteratief

$$I_{n+1} := (aI_n + c) \bmod m$$

waarbij $x \bmod m$ de rest bij het delen van x door m is. De waarden van de getallen I_n liggen tussen 0 en $m - 1$, hieruit krijgt men toevalsgetallen U_n in het interval $[0, 1)$ door $U_n := \frac{I_n}{m}$ te definiëren.

Omdat I_n alleen maar de waarden $0, 1, \ldots m-1$ kan hebben, is deze randomgenerator altijd periodiek met een periode van lengte hoogstens m. Maar behalve voor speciale (slechte) waarden van m, a, c en I_0 wordt deze lengte van de periode ook bereikt en levert deze methode een redelijk goede randomgenerator. Vaak wordt voor m een macht van 2 zoals 2^{32} gekozen, omdat dit op

een computer met 32-bit of 64-bit getallen de *modulo* operatie heel eenvoudig maakt. In dit geval laat zich aantonen, dat een lineaire congruentie model met $a \equiv 1 \bmod 4$ en een oneven c altijd een periode van maximale lengte oplevert.

Voordat een randomgenerator voor simulaties wordt gebruikt, is het verstandig om te toetsen of de pseudo-toevalsgetallen die hij oplevert inderdaad redelijk goed gelijkverdeeld en onafhankelijk zijn. Hiervoor zijn er een aantal tests, die op methoden uit de statistiek gebaseerd zijn.

Om een eerste indruk te krijgen, kan men de punten (U_{2i-1}, U_{2i}) in het 2-dimensionale vlak plotten en kijken of dit er redelijk toevallig uitziet. Als er hier al een soort structuur of patroon opvalt, is er zeker iets mis met de randomgenerator.

In een iets systematischere test deelt men het interval $[0, 1)$ in d (even grote) deelintervallen, telt hoe veel van $U_1, U_2, \ldots, U_n$ in elk van die deelintervallen ligt en toetst deze verdeling met een χ^2-test tegen de gelijkverdeling. (De χ^2-test is een standaardtest uit de statistiek die toetst of de gevonden verdeling te veel of te weinig van de gelijkverdeling afwijkt, want het is ook heel onwaarschijnlijk, dat in elk deelinterval precies d/n getallen terecht komen.) Een soortgelijke test kan men in plaats van de enkele toevalsgetallen ook op paren of in het algemeen op k-dimensionale vectoren $(U_1, \ldots, U_k)$, $(U_{k+1}, \ldots, U_{2k}), \ldots, (U_{(n-1)k+1}, \ldots, U_{nk})$ toepassen, die gelijkverdeeld in de k-dimensionale kubus $[0, 1)^k$ moeten zijn.

Met andere tests wordt de onafhankelijkheid getoetst. Bijvoorbeeld wordt in de *gap test* een deelinterval $[a, b]$ van $[0, 1)$ gekozen en vervolgens gekeken, hoe lang de stukken van de rij (U_i) zijn die niet in $[a, b]$ liggen. Als we $p := |b - a|$ definiëren, dan is de kans op een stuk van lengte k tussen twee getallen die wel in $[a, b]$ liggen, gelijk aan $p(1 - p)^k$ (dit noemt men een geometrische verdeling met parameter p). De gevonden verdeling van lengtes van stukken kunnen we nu ook weer tegen de verwachte geometrische verdeling toetsen (bijvoorbeeld met een χ^2-toets).

We gaan er vanaf nu van uit dat we een (wel getoetste) randomgenerator ter beschikking hebben, die elke keer dat we hem gebruiken een toevalsgetal $U_i \in [0, 1)$ oplevert zo dat deze getallen gelijk verdeeld en onafhankelijk zijn.

Er zijn een aantal algemene principes, hoe we een gewenste kansverdeling met behulp van een randomgenerator kunnen simuleren. De meest belangrijke zijn de methode van de *inverse verdelingsfunctie* en de *wegwerp* (rejection) methode. In principe zijn deze methoden op discrete en continue kansverdelingen toepasbaar, omdat we ook voor discrete kansverdelingen vaak eenvoudig een verdelingsfunctie $F(x)$ en dichtheidsfunctie $f(x)$ kunnen aangeven. Maar voor zekere discrete kansverdelingen zullen we later nog andere (meer directe) methoden aangeven.

Simulatie met behulp van de inverse verdelingsfunctie

Voor een algemene (continue) kansverdeling met dichtheidsfunctie $f(x)$ en verdelingsfunctie $F(x) = \int_{-\infty}^{x} f(t)\, dt$ passen we de inverse F^{-1} van de verdelings-

functie op de uniforme verdeling toe: Zij U een stochast met uniforme verdeling op $[0,1)$, dan definiëren we een nieuwe stochast X door $X := F^{-1}(U)$. Voor de kans $P(X \leq a)$ geldt nu $P(X \leq a) = P(F(X) \leq F(a)) = P(U \leq F(a)) = F(a)$ omdat U uniform verdeeld is. De stochast X heeft dus de verdelingsfunctie $F(x)$.

Voorbeeld 1: We willen een algemene rechthoekverdeling op het interval $[a,b]$ simuleren. De verdelingsfunctie voor deze verdeling is $F(x) = \frac{1}{b-a}(x-a)$ en uit $y = \frac{1}{b-a}(x-a) \Leftrightarrow (b-a)y = (x-a) \Leftrightarrow x = a + (b-a)y$ volgt $F^{-1}(y) = a + (b-a)y$.

We krijgen dus een toevalsrij (V_i) met waarden in het interval $[a,b]$ door $V_i := a + (b-a)U_i$. Dit hadden we natuurlijk ook zonder de inverse van de verdelingsfunctie kunnen bedenken.

Voorbeeld 2: Ook voor de exponentiële verdeling krijgen op deze manier een simulatie. Na een mogelijke verschuiving op de x-as heeft de exponentiële verdeling de dichtheidsfunctie $f(x) = \lambda e^{-\lambda x}$ en de verdelingsfunctie $F(x) = 1 - e^{-\lambda x}$. Omdat $y = 1 - e^{-\lambda x} \Leftrightarrow -\lambda x = \log(1-y) \Leftrightarrow x = -\frac{1}{\lambda}\log(1-y)$, hebben we $F^{-1}(y) = -\frac{1}{\lambda}\log(1-y)$. Voor een uniform verdeelde stochast U is dus de stochast $X := -\frac{1}{\lambda}\log(1-U)$ exponentieel verdeeld met parameter λ. Maar omdat met U ook $1-U$ gelijkverdeeld op $[0,1)$ is, kunnen we net zo goed ook $X := -\frac{1}{\lambda}\log(U)$ definiëren.

Simulatie met behulp van de wegwerp methode

Soms is de inverse F^{-1} van de verdelingsfunctie $F(x)$ van een kansverdeling niet makkelijk te bepalen of zelfs onmogelijk expliciet op te schrijven. Het meest prominente voorbeeld hiervoor is de normaalverdeling.

Maar we kunnen een kansverdeling met dichtheidsfunctie $f(x)$ op een eindig interval $[a,b]$ als volgt simuleren: Stel de dichtheidsfunctie is op het interval $[a,b]$ door een waarde c begrensd, d.w.z. $f(x) \leq c$ voor alle $x \in [a,b]$. Dan produceren we een rij toevalsgetallen (X_i) volgens een gelijkverdeling op $[a,b]$ en een rij (Y_i) volgens een gelijkverdeling op $[0,c]$. We accepteren nu alleen maar de X_i voor de indices i waarvoor geldt dat $Y_i \leq f(X_i)$ en werpen de andere X_i weg. Het is niet moeilijk om in te zien dat de geaccepteerde toevalsgetallen X_i de dichtheidsfunctie $f(x)$ hebben, want een waarde $X_i = x$ wordt juist met kans $\frac{f(x)}{c}$ geaccepteerd.

Simulatie van speciale verdelingen

Voor een aantal belangrijke kansverdelingen geven we nu aan hoe we met behulp van een randomgenerator die toevalsgetallen U_i op het interval $[0,1)$ produceert een stochast X met deze kansverdeling kunnen simuleren.

Discrete gelijkverdeling: Voor een eindige uitkomstenruimte Ω met $|\Omega| = n$ kunnen we aannemen dat $\Omega = \{0, \ldots, n-1\}$. We krijgen een gelijkverdeling op Ω door $X := \lfloor n \cdot U_i \rfloor$, waarbij $\lfloor x \rfloor$ het grootste gehele getal is dat $\leq x$ is.

Binomiale verdeling: We kunnen algemeen een uitkomst met kans p simuleren door $X := \lfloor p+U_i \rfloor$, want $p+U_i$ is een gelijkverdeling op het verschoven interval $[p, 1+p]$ en we hebben een waarde ≥ 1 met kans p.

Voor de binomiale verdeling $b(m, p; k)$ herhalen we m keer een simulatie met kans p en krijgen: $X := \sum_{i=1}^{m} \lfloor p + U_i \rfloor$.

Hypergeometrische verdeling: Om de hypergeometrische verdeling met parameters n, m en s te simuleren volgen we in principe de procedure van een echte proef. We noemen s_i het aantal slechte stukken die voor de i-de greep nog in de verzameling zitten en $p_i = \frac{s_i}{n}$ de kans dat we in de i-de greep een slecht stuk kiezen. Onze stochast X is het aantal slechte stukken die we grijpen. We beginnen dus met $X := 0$, $s_1 := s$ en $p_1 := \frac{s_1}{n} = \frac{s}{n}$ en voeren de volgende procedure voor $i = 1, 2, \ldots, m$ uit:

> Laat $A_i := \lfloor p_i + U_i \rfloor$ dan geeft $A_i = 1$ aan dat een slecht stuk werd getrokken, en $A_i = 0$ dat geen slecht stuk werd getrokken. We zetten nu $X := X + A_i$, $s_{i+1} := s_i - A_i$ en $p_{i+1} := \frac{s_{i+1}}{n}$.

Poisson-verdeling: Als m groot is kunnen we met behulp van de simulatie van de binomiale verdeling ook de Poisson-verdeling met parameter $\lambda = m \cdot p$ simuleren. Maar hiervoor maken we beter gebruik van *Poisson-processen* die we in de volgende les uitgebreid gaan behandelen. Een Poisson-proces beschrijft gewoon de tijdstippen van gebeurtenissen die volgens een Poisson-verdeling optreden. Het cruciale punt is dat de tussentijden tussen twee gebeurtenissen van een Poisson-proces exponentieel verdeeld zijn en de parameter van deze exponentiele verdeling noemen we de *intensiteit* van het Poisson-proces. We zullen zien dat voor een Poisson-proces met intensiteit 1 het aantal waarnemingen in het tijdsinterval $[0, \lambda]$ een Poisson-verdeling met parameter λ heeft. We moeten dus het tijdsinterval $[0, \lambda]$ met exponentieel verdeelde tussentijden overdekken en tellen hoeveel tussentijden er nodig zijn. Hiervoor nemen we onafhankelijke stochasten $Y_1, Y_2, \ldots$ die exponentieel verdeeld zijn met parameter 1. Als we nu een stochast X definiëren door de eigenschap

$$\sum_{i=1}^{X} Y_i \leq \lambda < \sum_{i=1}^{X+1} Y_i$$

dan heeft X een Poisson-verdeling met parameter λ. Maar de Y_i kunnen we zo als boven gezien met behulp van een randomgenerator U_i simuleren door $Y_i := -\log(U_i)$ (de parameter van de exponentiële verdeling is 1), dus is $-\sum_{i=1}^{X} \log(U_i) \leq \lambda < -\sum_{i=1}^{X+1} \log(U_i)$ en dit is equivalent met

$$\prod_{i=1}^{X} U_i \geq e^{-\lambda} > \prod_{i=1}^{X+1} U_i.$$

We vermenigvuldigen dus exponentieel verdeelde toevalsgetallen tussen 0 en 1 tot dat het product kleiner is dan $e^{-\lambda}$, het aantal X van benodigde getallen is dan een stochast met Poisson-verdeling met parameter λ.

Normaalverdeling: Voor de normaalverdeling bestaat er behalve de wegwerpmethode nog een andere mogelijkheid om tot een efficiënte simulatie te komen. Deze methode berust op de

Centrale limietstelling: Als $X_1, X_2, \ldots$ onafhankelijke stochasten zijn met verwachtingswaarde $E(X_i)$ en variantie $Var(X_i)$, dan is de limiet

$$X := \lim_{n\to\infty} \frac{\sum_{i=1}^{n}(X_i - E(X_i))}{\sqrt{\sum_{i=1}^{n} Var(X_i)}}$$

onder zwakke verdere voorwaarden aan de X_i een stochast met standaardnormaalverdeling. In het bijzonder wordt aan de voorwaarden voldaan als alle X_i dezelfde standaardafwijking σ hebben, in dit geval convergeert $\sum_{i=1}^{n} \frac{X_i - E(X_i)}{\sqrt{n}\cdot\sigma}$ tegen de standaardnormaalverdeling.

Voor de door de randomgenerator (U_i) gesimuleerde uniforme verdeling op $[0,1)$ hebben we $E(X) = \frac{1}{2}$ en $Var(X) = \int_0^1 (x - \frac{1}{2})^2 dx = \frac{1}{12}$. Als we nu n waarden van de rij (U_i) optellen hebben we $S_n = \sum_{i=1} U_i$ en er geldt $E(S_n) = \frac{n}{2}$ en $Var(S_n) = \frac{n}{12}$. Als benadering van de standaardnormaalverdeling krijgen we dus

$$X := \sqrt{\frac{12}{n}}((\sum_{i=1}^{n} U_i) - \frac{n}{2}).$$

Deze benadering is al voor $n = 10$ heel goed en voor de meeste toepassingen voldoende.

Voorbeeld: We kijken tot slot naar een simulatie van het Monty-Hall probleem, om mensen die de theoretische argumenten niet accepteren door een experiment te kunnen overtuigen. De simulatie volgt de stappen in de show:

(1) Kies een deur A waar de auto achter staat: $A := \lfloor 3 \cdot U_i \rfloor$ (we noemen de deuren 0, 1 en 2).

(2) De kandidaat kiest een deur K: $K := \lfloor 3 \cdot U_i \rfloor$.

(3) De moderator opent een deur M. Hier zijn twee gevallen mogelijk:

   (i) $A = K$: in dit geval heeft de moderator de keuze tussen $A + 1$ en $A + 2$ (als we nummers van de deuren modulo 3 nemen) we nemen dus $M := A + \lfloor 2 \cdot U_i \rfloor + 1 \bmod 3$.

   (ii) $A \neq K$: in dit geval heeft de moderator geen keuze, hij moet de deur M openen met $M \neq A$ en $M \neq K$.

(4) Hier zijn er twee versies:

   (A) De kandidaat blijft bij zijn keuze, dus $K' = K$.

   (B) De kandidaat wisselt van keuze, dus K' zo dat $K' \neq K$ en $K' \neq M$.

(5) Als $K' = A$ krijgt de kandidaat de auto, anders alleen maar de geit.

Dit kunnen we voor de versies A en B in stap (4) op een computer heel makkelijk 10000 keer doorspelen. Na drie herhalingen voor beide versies krijgen we bijvoorbeeld 3319, 3400 en 3327 successen voor versie A en 6583, 6655 en 6675 successen voor versie B.

Het blijkt dus ook uit het experiment dat het verstandig voor de kandidaat is om van keuze te wisselen.

Belangrijke begrippen in deze les

- maximum likelihood schatting
- simulatie
- randomgenerator, toevalsgetallen
- methode van de inverse verdelingsfunctie
- wegwerp methode
- centrale limietstelling

Opgaven

21. Voor een gebeurtenis dat volgens een normaalverdeling met dichtheidsfunctie

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

optreedt, zijn de observaties $x_1, \ldots, x_n$ gemaakt.

(i) Bepaal de maximum-likelihood schatting voor de verwachtingswaarde μ als de variantie σ^2 bekend is.

(ii) Bepaal de maximum-likelihood schatting voor de variantie σ^2 als de verwachtingswaarde μ bekend is.

(Opmerking: Als de verwachtingswaarde μ en de variantie σ^2 onbekend zijn, zijn de waarden uit (i) en (ii) de nulpunten van de partiële afgeleiden van de likelihood-functie en geven dus noodzakelijke voorwaarden voor een maximum van de likelihood-functie. Er laat zich aantonen dat men zo inderdaad een maximum vindt, dus laten zich μ en σ simultaan schatten.)

22. Laten U_1 en U_2 twee uniform verdeelde stochasten op $[0, 1)$ zijn. Laat zien dat $\sqrt{U_1}$ en $\max(U_1, U_2)$ dezelfde verdeling hebben. (Dit geeft een zuinige manier om het maximum van twee uniforme kansverdeling te simuleren.)

23. Een symmetrische *driehoeksverdeling* op het interval $[-1, 1]$ heeft de dichtheidsfunctie $f(x) = 1 - |x| = \begin{cases} 1+x & \text{als } x < 0 \\ 1-x & \text{als } x \geq 0 \end{cases}$.

Laten U en V twee stochasten zijn die uniform verdeeld zijn op het interval $[0, 1)$.

(i) Laat zien dat de stochast $X_1 := U + V - 1$ de boven aangegeven driehoeksverdeling als dichtheidsfunctie heeft.

(ii) Ga na dat de stochast $X_2 := U - V$ dezelfde kansverdeling als X_1 heeft. (We hebben dus twee manieren om de driehoeksverdeling met behulp van een randomgenerator te simuleren.)

(iii) Laat zien dat X_1 en X_2 covariantie 0 hebben, d.w.z. dat $E(X_1 \cdot X_2) = E(X_1) \cdot E(X_2)$ is.

(iv) Toon aan dat X_1 en X_2 *niet* onafhankelijk zijn.

24. Bedenk en beschrijf een efficiënte simulatie voor het trekken van de lottogetallen.

Les 6 Poisson processen

Als gebeurtenissen op willekeurige tijdstippen plaats vinden, kunnen we dit opvatten als een soort proces die de gebeurtenissen op een toevallige manier voortbrengt. Voorbeelden van dit soort processen zijn:

- klanten die een winkel binnen lopen,
- inkomende aanvragen bij een telefooncentrale,
- uitvallen van servers van een groot internet-bedrijf,
- emissie van een radioactief preparaat.

Een groot aantal van processen die we als *toevallig* beschouwen, laat zich door een paar heel eenvoudige regels karakteriseren. Als $N(t_1, t_2)$ het aantal gebeurtenissen van een proces in het tijdsinterval $[t_1, t_2]$ aangeeft, zou men bij een proces bijvoorbeeld het volgende kunnen eisen:

(1) De kansverdeling van $N(t, t+h)$ is onafhankelijk van t, d.w.z. de kans voor de gebeurtenissen is invariant onder een verschuiving in de tijd.

(2) Voor $t_1 < t_2 < t_3 < t_4$ zijn $N(t_1, t_2)$ en $N(t_3, t_4)$ onafhankelijk, d.w.z. de gebeurtenissen op niet overlappende intervallen zijn onafhankelijk.

(3) $P(N(t, t+h) = 1) = \lambda h + o(h)$ en $P(N(t, t+h) = 0) = 1 - \lambda h + o(h)$. Hierbij geeft $o(h)$ een term aan die voor $h \to 0$ sneller naar 0 gaat dan h, dus waarvoor geldt dat $\lim_{h\to 0} \frac{o(h)}{h} = 0$. Hieruit volgt dat voor kleine tijdsintervallen de kans op één gebeurtenis in het interval evenredig aan de lengte van het tijdsinterval is, want $\lim_{h\to 0} \frac{\lambda h + o(h)}{h} = \lambda$.

De eisen (1) en (2) betekenen dat de kansverdeling $N(t_1, t_2)$ alleen maar van de lengte $|t_2 - t_1|$ van het tijdsinterval afhangt en onafhankelijk van eerdere gebeurtenissen is. Omdat de kansverdeling $N(t_1, t_2)$ niet van het verleden of de *geschiedenis* van het proces afhangt, noemt men zo'n proces ook *geheugenloos*.

De parameter λ heet de *intensiteit* van het proces. Uit eis (3) kunnen we concluderen dat de kans op twee of meer gebeurtenissen in een tijdsinterval gegeven is door $P(N(t, t+h) \geq 2) = o(h)$ en voor $h \to 0$ gaat deze kans naar 0, want $\lim_{h\to 0} \frac{o(h)}{h} = 0$. Dit betekent in het bijzonder dat er nooit twee gebeurtenissen op hetzelfde tijdstip plaats vinden.

Een proces die aan deze eisen voldoet, noemen we een *Poisson-proces*, de naamgeving zal straks toegelicht worden.

Nemen we als model voor dit soort processen klanten die een winkel binnen lopen, dan kunnen we ons ook voorstellen dat de klanten in een rij staan voor dat ze geholpen worden. Daarom spreekt men hier ook van *wachtrijen*. We kunnen nu verschillende typen van vragen stellen, bijvoorbeeld:

- Wat is de kans dat er binnen 5 minuten twee klanten de winkel binnen lopen?

- Wat is de kans dat er meer dan 5 minuten niemand binnen komt?

De eerste vraag gaat over het aantal mensen die in een rij staan en we zullen zien dat we dit met behulp van de Poisson-verdeling kunnen bepalen. De tweede vraag is over de tussentijden tussen twee gebeurtenissen, en het aardige is dat we uit onze aannamen over onafhankelijkheid kunnen afleiden dat de tussentijden tussen de n-de en $(n+1)$-de gebeurtenis alle onafhankelijk van elkaar zijn en door een *exponentiële verdeling* beschreven worden. In het bijzonder vinden we hier een koppeling tussen de discrete verdeling van de gebeurtenissen en de continue verdeling van de tussentijden.

6.1 Tussentijden bij een Poisson-proces

Om de Poisson-processen te beschrijven kijken we eerst naar de kans dat er tot een tijdstip t helemaal geen gebeurtenis waargenomen wordt. We schrijven $P_0(t) := P(N(0,t) = 0)$ voor deze kans. Omdat de tijdsintervallen $[0,t]$ en $[t,t+h]$ niet overlappen, geldt volgens eis (2) dat

$$P_0(t+h) = P_0(t) \cdot P(N(t,t+h) = 0) = P_0(t)(1-\lambda h).$$

Hierbij hebben we de $o(h)$ termen weggelaten, omdat we later de limiet $h \to 0$ bekijken, waarbij deze termen sowieso wegvallen. Hieruit volgt

$$\frac{P_0(t+h) - P_0(t)}{h} = -\lambda P_0(t)$$

en door hiervan de limiet $h \to 0$ te nemen, krijgen we aan de linkerkant de afgeleide van $P_0(t)$, dus

$$P_0'(t) = -\lambda P_0(t).$$

Zo'n soort vergelijking die een functie en hun afgeleide bevat, heet een *differentiaalvergelijking*. In het algemeen is het enigszins moeilijk om de oplossingen van differentiaalvergelijking te vinden, maar in ons geval is het zelfs mogelijk, alle functies $f(x)$ die aan de differentiaalvergelijking $f'(x) = -\lambda f(x)$ voldoen expliciet aan te geven:

Men gaat (door afleiden) snel na dat $f(x) = e^{-\lambda x}$ inderdaad een oplossing is. Als we nu nog een tweede functie $g(x)$ hebben, waarvoor ook geldt dat $g'(x) = -\lambda g(x)$, dan kunnen we de quotiënt $\frac{g(x)}{f(x)}$ van de twee functies bekijken. Voor de afgeleide van deze quotiënt geldt (met behulp van de quotiëntenregel):

$$\left(\frac{g(x)}{f(x)}\right)' = \frac{g'(x)f(x) - g(x)f'(x)}{g(x)^2} = \frac{-\lambda g(x)f(x) - g(x)(-\lambda f(x))}{g(x)^2} = \frac{0}{g(x)^2} = 0.$$

Maar als een functie afgeleide 0 heeft, is de functie zelf constant, d.w.z. $\frac{g(x)}{f(x)} = c$ en dus $g(x) = c \cdot f(x)$. De oplossingen van de differentiaalvergelijking zijn dus juist de veelvouden van $f(x)$.

We hebben op deze manier gezien dat $P_0(t)$ noodzakelijk van de vorm $P_0(t) = c \cdot e^{-\lambda t}$ is. Als we $t = 0$ invullen, kijken we naar de kans dat er geen gebeurtenis in het interval $[0,0]$ plaats vindt. Maar deze kans is 1 omdat het om een enkele punt gaat, dus is $P_0(0) = 1$ en dus $c = 1$.

De kans dat tot een tijdstip t geen enkel gebeurtenis waargenomen wordt is dus $e^{-\lambda t}$. Als we het tijdstip van de eerste waarneming T noemen, betekent dit dat $P(T > t) = e^{-\lambda t}$ en dus

$$P(T < t) = 1 - e^{-\lambda t}.$$

Dit betekent dat de kansverdeling voor het tijdsinterval tot de eerste waarneming een *exponentiële verdeling met parameter* λ is.

Herinnering: De exponentiële verdeling met parameter λ heeft dichtheidsfunctie $f(x) = \lambda e^{-\lambda x}$ en verdelingsfunctie $F(x) = 1 - e^{-\lambda x}$. Dit betekent dat voor een exponentieel verdeelde stochast X de kans op een uitkomst van hoogstens x gegeven is door $P(X \leq x) = 1 - e^{-\lambda x}$. De verwachtingswaarde van zo'n exponentieel verdeelde stochast X is $E(X) = \frac{1}{\lambda}$.

Onze veronderstelling over de onafhankelijkheid tegenover verschuivingen in de tijd zegt nu, dat we het tijdstip $t = 0$ ook op het tijdstip T van de eerste waarneming kunnen leggen. Dan kunnen we meteen concluderen dat de kansverdeling voor het tijdsinterval tussen de eerste en de tweede waarneming van een gebeurtenis ook een exponentiële verdeling met parameter λ is. Met hetzelfde argument vinden we, dat de kansverdeling voor het tijdsinterval tussen de n-de en de $(n + 1)$-de waarneming dezelfde exponentiële verdeling met parameter λ is.

Dat de tussentijden tussen de verschillende gebeurtenissen onafhankelijk van elkaar zijn volgt uit de onafhankelijkheid voor niet overlappende intervallen

Merk op: Omdat de tussentijden bij een Poisson-proces met intensiteit λ exponentieel verdeeld met parameter λ zijn en dus verwachtingswaarde $\frac{1}{\lambda}$ hebben, kunnen we uit de kennis van de gemiddelde tussentijden de intensiteit van het proces bepalen. Als namelijk de tussentijden gemiddeld τ zijn, heeft het Poisson-proces de intensiteit $\frac{1}{\tau}$.

6.2 Aantallen gebeurtenissen bij een Poisson-proces

We gaan nu nog aantonen dat bij een Poisson-proces het aantal gebeurtenissen in een gegeven tijdsinterval door een *Poisson-verdeling* beschreven wordt.

We hebben al gezien dat $P(N(0,t) = 0) = e^{-\lambda t}$. Als afkorting schrijven we nu $P_1(t) := P(N(0,t) = 1)$ voor de kans dat er precies één gebeurtenis tot het tijdstip t plaats vindt. Als er in het tijdsinterval $[0, t + h]$ één waarneming is, is die of in het interval $[0, t]$ of in het interval $[t, t + h]$. Omdat deze twee intervallen niet overlappen, geldt volgens onze onafhankelijkheidseisen voor Poisson-processen dat

$$\begin{aligned} P_1(t+h) &= P_0(t) \cdot P(N(t,t+h) = 1) + P_1(t) \cdot P(N(t,t+h) = 0) \\ &= P_0(t) \cdot \lambda h + P_1(t) \cdot (1 - \lambda h). \end{aligned}$$

Hieruit volgt

$$\frac{P_1(t+h) - P_1(t)}{h} = \lambda P_0(t) - \lambda P_1(t)$$

en door hiervan de limiet $h \to 0$ te nemen en $P_0(t)$ van boven in te vullen, krijgen we de differentiaalvergelijking

$$P_1'(t) = \lambda e^{-\lambda t} - \lambda P_1(t).$$

Ook hier is het niet zo moeilijk om aan te tonen, dat $P_1(t)$ van de vorm $P_1(t) = \lambda t e^{-\lambda t} + ce^{-\lambda t}$ moet zijn, en uit $P_1(0) = 0$ volgt $c = 0$. Er geldt dus

$$P(N(0,t) = 1) = (\lambda t)e^{-\lambda t}.$$

We kunnen nu op een soortgelijke manier doorgaan om aan te tonen dat

$$P(N(0,t) = k) = \frac{(\lambda t)^k}{k!} \cdot e^{-\lambda t}.$$

Stel we hebben dit voor $k-1$ al gezien, dan schrijven we net als boven $P_k(t) := P(N(0,t) = k)$ voor de kans op precies k gebeurtenissen tot het tijdstip t. Omdat we nooit twee gebeurtenissen in een klein interval h hebben, geldt:

$$\begin{aligned} P_k(t+h) &= P_{k-1}(t) \cdot P(N(t,t+h) = 1) + P_k(t) \cdot P(N(t,t+h) = 0) \\ &= P_{k-1}(t) \cdot \lambda h + P_k(t) \cdot (1 - \lambda h). \end{aligned}$$

Hieruit volgt

$$\frac{P_k(t+h) - P_k(t)}{h} = \lambda P_{k-1}(t) - \lambda P_k(t)$$

en door hiervan de limiet $h \to 0$ te nemen en $P_{k-1}(t)$ van boven in te vullen, krijgen we

$$P_k'(t) = \lambda \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t} - \lambda P_k(t).$$

Hieruit volgt $\frac{d}{dt}(e^{\lambda t} P_k(t)) = (P_k'(t) + \lambda P_k(t))e^{\lambda t} = \lambda \frac{(\lambda t)^{k-1}}{(k-1)!}$ en door integreren krijgen we $e^{\lambda t} P_k(t) = \frac{(\lambda t)^k}{k!} + c$. Er geldt dus $P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} + ce^{-\lambda t}$ en uit $P_k(0) = 0$ volgt weer $c = 0$.

Conclusie: Alles bij elkaar genomen hebben we aangetoond dat voor een Poisson-proces met intensiteit λ het aantal gebeurtenissen in het interval $[0,t]$ een Poisson-verdeling met parameter λt heeft, dus dat

$$P(N(0,t) = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

Merk op: Een Poisson-verdeling met parameter λt heeft verwachtingswaarde λt. De intensiteit λ van een Poisson-proces is dus het gemiddelde aantal gebeurtenissen in het eenheidstijdsinterval $[0,1]$.

Voorbeeld 1: Klanten komen een winkel binnen volgens een Poisson-proces met intensiteit 3 (per uur). Elke klant blijft twintig minuten in de winkel. We willen de kans berekenen dat twee klanten elkaar ontmoeten. Hiervoor kijken we naar het tijdsinterval tussen twee klanten en we hebben de kans nodig, dat zo'n interval hoogstens 20 minuten is. Maar we weten dat de tussentijden

exponentieel met parameter $\lambda = 3$ verdeelt zijn, dus is de kans op een interval $T < \frac{1}{3}uur$ gegeven door $P(T < \frac{1}{3}) = 1 - e^{-3 \cdot \frac{1}{3}} = 1 - e^{-1} \approx 0.632$.

Voorbeeld 2: Een interactief systeem kan maximaal 15 transacties per seconde verwerken. In een spitsuur zijn er gemiddeld 10 transacties per seconde die volgens een Poisson-proces binnen komen. Wat is de kans dat het systeem tijdens een spitsuur overbelast raakt?

We hebben een intensiteit van $\lambda = 10$ (transacties per seconde) en willen de kans op meer dan 15 transacties in een tijdsinterval van een seconde bepalen. Dit is het complement van de kans op hoogstens 15 transacties en deze is

$$P(X \leq 15) = (1 + \frac{10}{1} + \frac{10^2}{2!} + \ldots + \frac{10^{15}}{15!})e^{-10} \approx 0.9513.$$

De gezochte kans is dus $1 - P(X \leq 15) \approx 4.87\%$.

We gaan nu eens omgekeerd uit van een proces die niet noodzakelijk een Poisson-proces is, maar waarvan we weten, dat de kansverdeling $P(N(0,t) = k)$ een Poisson-verdeling met parameter λt is. In dit geval is het eenvoudig om aan te tonen dat het tijdsinterval T tot de eerste waarneming exponentieel met parameter λ verdeeld is; in feite hebben we het argument boven al toegepast.

De kans dat T groter dan t is, is namelijk gelijk aan de kans dat in het interval $[0,t]$ geen waarneming ligt, en die is bij een Poisson-verdeling $e^{-\lambda t}$. We hebben dus $P(T < t) = 1 - e^{-\lambda t}$ en de verdeling van het tijdsinterval tot de eerste waarneming van een gebeurtenis is inderdaad een exponentiële verdeling met parameter λ.

We kunnen nog een verder aspect bekijken, dat laat zien dat een Poisson-proces inderdaad toevallige gebeurtenissen beschrijft. Hiervoor nemen we aan dat we weten dat er een gebeurtenis in het interval $[0,t]$ valt. We gaan nu de kansverdeling voor het tijdstip T bepalen, waarop de gebeurtenis plaats vindt. Voor de voorwaardelijke kans $P(T \leq x \mid N(0,t) = 1)$ dat T hoogstens x is, geldt $P(T \leq x \mid N(0,t) = 1) = \dfrac{P(T \leq x, N(x,t) = 0)}{P(N(0,t) = 1)}$ en de teller hiervan kunnen we opsplitsen, namelijk $P(T \leq x, N(x,t) = 0) = P(N(0,x) = 1, N(x,t) = 0) = P(N(0,x) = 1) \cdot P(N(x,t) = 0)$, omdat de twee tijdsintervallen niet overlappen. Hieruit volgt

$$\begin{aligned} P(T \leq x \mid N(0,t) = 1) &= \frac{P(N(0,x) = 1) \cdot P(N(x,t) = 0)}{P(N(0,t) = 1)} \\ &= \frac{\lambda x e^{-\lambda x} \cdot e^{-\lambda(t-x)}}{\lambda t e^{-\lambda t}} = \frac{x}{t}. \end{aligned}$$

en dit betekent dat het tijdstip T uniform op het interval $[0,t]$ verdeeld is. Als we weten dat er een gebeurtenis in het interval $[0,t]$ valt, hebben we dus geen verdere informatie over het tijdstip van de gebeurtenis, elk punt in het interval is even goed.

Samenvattend kunnen we zeggen, dat een Poisson-proces met intensiteit λ gekarakteriseerd is door een van de volgende eigenschappen:

(1) Het aantal gebeurtenissen in een tijdsinterval van lengte t is gegeven door een Poisson-verdeling met parameter λt.

(2) De tussentijden tussen de gebeurtenissen zijn onafhankelijk van elkaar en alle verdeelt volgens een exponentiële verdeling met parameter λ.

Belangrijke begrippen in deze les

- Poisson-proces
- intensiteit
- wachtrijen
- exponentiële verdeling

Opgaven

25. In een zeker gebied treden aardbevingen bij benadering op volgens een Poisson-proces met een gemiddelde van 2 aardbevingen per maand.

    (i) Bereken de kans dat er de komende twee maanden minstens drie aardbevingen optreden.

    (ii) Wat is de kans dat de eerstvolgende aardbeving minstens drie maanden op zich laat wachten?

26. Op een computer systeem komen aanvragen volgens een Poisson-proces binnen, gemiddeld 60 per uur. Bepaal de kansen voor de volgende tussentijden tussen twee op elkaar volgende aanvragen:

    (i) meer dan 4 minuten,

    (ii) minder dan 8 minuten,

    (iii) tussen 2 en 6 minuten.

27. Op een kantoor komen gemiddeld 12 gesprekken per uur binnen. Het aantal gesprekken dat per 10 minuten binnenkomt kan beschouwd worden als een stochast met Poisson-verdeling. Bereken de kans dat er

    (i) meer dan 3,

    (ii) hoogstens 4,

    (iii) meer dan 1 maar hoogstens 4

    klanten geen gehoor krijgen als de telefoniste gedurende 10 minuten afwezig is.

28. Een telefooncentrale kan per minuut maximaal 20 telefoongesprekken aan. Het aantal gesprekken per uur is een Poisson-verdeelde stochast met verwachtingswaarde 600. Bereken de kans dat de telefooncentrale gedurende een bepaalde minuut overbelast zal raken.

Les 7 Betrouwbaarheid en levensduur

7.1 Betrouwbaarheid van systemen

Als een systeem of netwerk uit verschillende componenten bestaat, kan men zich de vraag stellen hoe groot de kans is dat het systeem faalt. Dit hangt zeker van de kwaliteit van de componenten af, maar ook hoe deze componenten van elkaar afhangen en samengeschakeld zijn.

Voor een enkele component C noemen we de kans dat C (normaal) werkt de *betrouwbaarheid* van C. Als A de gebeurtenis is dat C werkt, dan is de betrouwbaarheid R (voor *reliability*) van C dus gegeven door $R = P(A)$.

Rijschakeling

De eenvoudigste manier om verschillende componenten te combineren, is een *rijschakeling*. Hierbij wordt een opdracht achter elkaar van de componenten $C_1, C_2, \ldots, C_n$ verwerkt, waarbij C_{i+1} pas begint als C_i klaar is.

Het hele systeem werkt alleen maar als alle componenten werken, als we met A_i het normaal werken van de component C_i noteren en met A het goed functioneren van het hele systeem, hebben we $P(A) = P(A_1, A_2, \ldots, A_n)$. Als de component C_i de betrouwbaarheid $R_i = P(A_i)$ heeft, dan is de betrouwbaarheid $R = P(A)$ van het hele systeem dus gegeven door

$$R = R_1 \cdot R_2 \cdot \ldots \cdot R_n = \prod_{i=1}^{n} R_i.$$

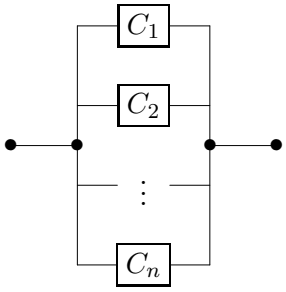
Hierbij veronderstellen we dat het falen van de enkele componenten onafhankelijke gebeurtenissen zijn.

Het is duidelijk dat zelfs bij zeer betrouwbare componenten de betrouwbaarheid van de rijschakeling met een groeiend aantal van componenten snel afneemt. Bijvoorbeeld hebben we voor $n = 10$ en $R_1 = \ldots = R_{10} = 0.99$ voor de rij van 10 componenten slechts nog een betrouwbaarheid van $R = 0.99^{10} \approx 0.904$.

Omdat systemen vaak uit heel veel componenten opgebouwd zijn en deze geen willekeurig hoge betrouwbaarheid kunnen hebben, wordt er vaak *redundantie* in een systeem ingebouwd. Dit betekent dat een systeem componenten bevat die overbodig zijn als alles goed werkt, maar die ervoor zorgen dat het hele systeem ook nog blijft werken als er een of meer componenten falen.

Parallelschakeling

De eenvoudigste vorm van redundantie is een *parallelschakeling*. Hierbij hoeft voor het correcte functioneren van het systeem slechts een van een aantal componenten benodigd.



Het hele systeem werkt als een van de componenten $C_1, \ldots, C_n$ werkt, dus is de kans $P(A^c)$ dat het hele systeem niet werkt gelijk aan de kans dat *geen* van de componenten werkt en dus gelijk aan het product van de kansen $P(A_i^c)$ dat de enkele componenten niet werken, dus $P(A^c) = (1-R_1)\cdot(1-R_2)\cdot\ldots\cdot(1-R_n)$. Voor de betrouwbaarheid $R = 1 - P(A^c)$ van het hele systeem geldt dus

$$R = 1 - (1 - R_1) \cdot (1 - R_2) \cdot \ldots \cdot (1 - R_n) = 1 - \prod_{i=1}^{n} (1 - R_i).$$

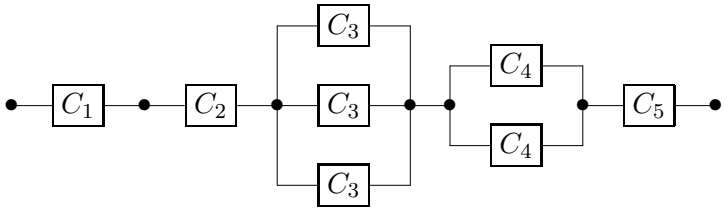
Ook hier gaan we weer ervan uit dat de componenten onafhankelijk van elkaar uitvallen.

Door een parallelschakeling kunnen we de betrouwbaarheid van een systeem snel verhogen, voor $n = 2$ en $R_1 = R_2 = 0.9$ hebben we bijvoorbeeld $R = 1 - (1 - 0.1)^2 = 0.99$. Maar we hebben hierbij de kosten verdubbeld om de betrouwbaarheid van 90% op 99% te verhogen.

Rij-parallel-schakeling

Als rij- en parallelschakelingen in een systeem gecombineerd worden, spreekt men van een *rij-parallel-schakeling*. Zo'n systeem krijgt men als men een rij- of parallelschakeling van een aantal componenten als een enkele component beschouwd en met dit soort componenten weer rij- en parallelschakelingen construeert. Door dit proces te herhalen, kan men redelijk ingewikkelde systemen realiseren. Tegelijkertijd levert dit proces ook het pad van de analyse van zo'n systeem, want door de stappen na te gaan kan ook de betrouwbaarheid van het systeem bepaald worden. We kijken hiervoor naar twee voorbeelden.

Voorbeeld 1: Een eenvoudig geval van een rij-parallel-schakeling is een rij van parallelschakelingen. In het volgende voorbeeld gaan we ervan uit dat de componenten in een parallelschakeling alle hetzelfde zijn.



De betrouwbaarheid R van het volledige systeem krijgen we als

$$R = R_1 \cdot R_2 \cdot (1 - (1 - R_3)^3) \cdot (1 - (1 - R_4)^2) \cdot R_5.$$

Als de componenten C_i bijvoorbeeld de betrouwbaarheden $R_1 = 0.95$, $R_2 = 0.99$, $R_3 = 0.7$, $R_4 = 0.75$ en $R_5 = 0.9$ hebben, geeft dit

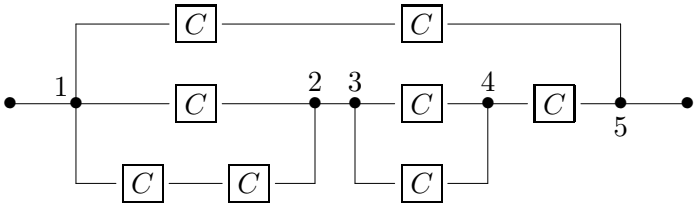
$$R = 0.95 \cdot 0.99 \cdot (1 - 0.3^3) \cdot (1 - 0.25^2) \cdot 0.9 \approx 0.772.$$

Algemeen heeft een rij van s parallelschakelingen, waarbij de i-de parallelschakeling n_i componenten met betrouwbaarheid R_i bevat de betrouwbaarheid

$$R = (1 - (1 - R_1)^{n_1}) \cdot \ldots \cdot (1 - (1 - R_s)^{n_s}) = \prod_{i=1}^{s} (1 - (1 - R_i)^{n_i}).$$

Merk op dat voor een component zonder parallel alternatief geldt dat $n_i = 1$, in dit geval is $1-(1-R_i)^{n_i} = 1-(1-R_i) = R_i$ en we moeten de betrouwbaarheid gewoon met R_i vermenigvuldigen.

Voorbeeld 2: We kijken naar een communicatienetwerk waarbij verschillende mogelijke paden tussen twee punten bestaan. De verbindingen op de paden bevatten zekere componenten waarvan het functioneren van de verbindingen afhangt, bijvoorbeeld versterkers voor GSM-signalen. Stel we hebben het volgende netwerk, waarbij de componenten C alle dezelfde betrouwbaarheid R hebben.



We zien dat het hele netwerk een parallelschakeling van het rechtstreekse pad $1-5$ en de verbinding $1-2-3-4-5$ is, waarbij $1-2-3-4-5$ een rij van eenvoudige parallelschakelingen is. De enkele betrouwbaarheden voor de deelverbindingen vinden we nu als volgt:

$$\begin{aligned}
R_{1-5} &= R^2 \\
R_{1-2} &= 1 - (1 - R)(1 - R^2) = 1 - (1 - R - R^2 + R^3) = R + R^2 - R^3 \\
R_{3-4} &= 1 - (1 - R)^2 = 1 - (1 - 2R + R^2) = 2R - R^2 \\
R_{1-2-3-4-5} &= R_{1-2} \cdot R_{3-4} \cdot R = (R + R^2 - R^3)(2R - R^2) \cdot R \\
&= 2R^3 - R^4 + 2R^4 - R^5 - 2R^5 + R^6 = 2R^3 + R^4 - 3R^5 + R^6.
\end{aligned}$$

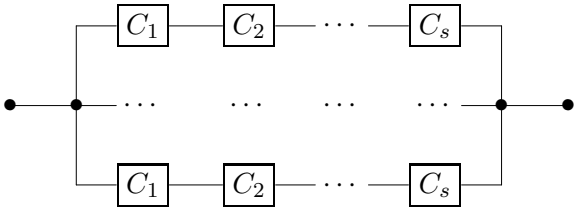
Als betrouwbaarheid R_{net} van het hele netwerk vinden we hieruit

$$\begin{aligned}
R_{net} &= 1 - (1 - R_{1-5})(1 - R_{1-2-3-4-5}) \\
&= 1 - (1 - R^2)(1 - 2R^3 - R^4 + 3R^5 - R^6) \\
&= 1 - 1 + 2R^3 + R^4 - 3R^5 + R^6 + R^2 - 2R^5 - R^6 + 3R^7 - R^8 \\
&= R^2 + 2R^3 + R^4 - 5R^5 + 3R^7 - R^8.
\end{aligned}$$

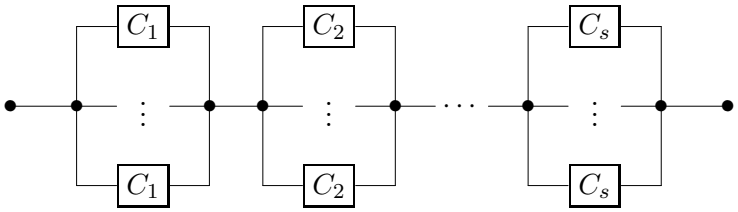
Als we $R = 0.9$ aannemen, krijgen we $R_{1-5} = 0.81$, $R_{1-2-3-4-5} \approx 0.874$ en $R_{net} \approx 0.976$. Voor $R = 0.99$ hebben we $R_{1-5} = 0.9801$, $R_{1-2-3-4-5} \approx 0.9897$ en $R_{net} \approx 0.9998$.

Als een systeem uit s stappen bestaat die achter elkaar uitgevoerd moeten worden, zijn er twee voor de hand liggende manieren om zo'n systeem met een *redundantiefactor* van m, d.w.z. met m componenten voor ieder stap, op te bouwen:

(1) Als parallelschakeling van m rijen die de s verschillende componenten bevatten.



(2) Als rijschakeling van s parallelschakelingen van telkens m componenten.



Als we de betrouwbaarheid van de component C_i met R_i noteren, vinden we voor de betrouwbaarheden van de versies (1) en (2):

$$\begin{aligned} R_{(1)} &= 1-(1-R_1\cdot R_2\cdot\ldots\cdot R_s)^m = 1-(1-\prod_{i=1}^{s} R_i)^m, \\ R_{(2)} &= (1-(1-R_1)^m)\cdot(1-(1-R_2)^m)\cdot\ldots\cdot(1-(1-R_s)^m) \\ &= \prod_{i=1}^{s}(1-(1-R_i)^m). \end{aligned}$$

Als alle componenten C_i dezelfde betrouwbaarheid R hebben, krijgen we hieruit de eenvoudigere uitdrukkingen:

$$R_{(1)} = 1-(1-R^s)^m \quad \text{en} \quad R_{(2)} = (1-(1-R)^m)^s.$$

Er laat zich algemeen aantonen dat versie (2) steeds een hogere betrouwbaarheid geeft, maar dit gaan we hier niet bewijzen. We zullen wel voor een paar kleine waarden van s en m nagaan, dat het zo is.

(i) $s=2$, $m=2$: Er geldt

$$R_{(1)} = 1-(1-R_1R_2)^2 = 1-(1-2R_1R_2+R_1^2R_2^2) = 2R_1R_2-R_1^2R_2^2.$$

Aan de andere kant is

$$\begin{aligned} R_{(2)} &= (1-(1-R_1)^2)(1-(1-R_2)^2) = (2R_1-R_1^2)(2R_2-R_2^2) \\ &= 4R_1R_2-2R_1^2R_2-2R_1R_2^2+R_1^2R_2^2 \\ &= 2R_1R_2-R_1^2R_2^2+2R_1R_2(1-R_1-R_2+R_1R_2) \\ &= R_{(1)}+2R_1R_2(1-R_1)(1-R_2). \end{aligned}$$

Omdat $R_i>0$ en $1-R_i>0$ is dus inderdaad $R_{(2)}>R_{(1)}$.

(ii) $s = 3$, $m = 2$, $R_1 = R_2 = R_3 = R$: Er geldt

$$R_{(1)} = 1 - (1 - R^3)^2 = 1 - (1 - 2R^3 + R^6) = 2R^3 - R^6.$$

Aan de andere kant is

$$\begin{aligned} R_{(2)} &= (1 - (1 - R)^2)^3 = (1 - (1 - 2R + R^2))^3 = (2R - R^2)^3 \\ &= 8R^3 - 12R^4 + 6R^5 - R^6 = 2R^3 - R^6 + 6R^3(1 - 2R + R^2) \\ &= R_{(1)} + 6R^3(1 - R)^2. \end{aligned}$$

Uit $R > 0$ volgt dat $R_{(2)} > R_{(1)}$.

(iii) $s = 2$, $m = 3$, $R_1 = R_2 = R$: Er geldt

$$R_{(1)} = 1 - (1 - R^2)^3 = 1 - (1 - 3R^2 + 3R^4 - R^6 = 3R^2 - 3R^4 + R^6.$$

Aan de andere kant is

$$\begin{aligned} R_{(2)} &= (1 - (1 - R)^3)^2 = (1 - (1 - 3R + 3R^2 - R^3))^2 \\ &= (3R - 3R^2 + R^3)^2 = 9R^2 + 9R^4 + R^6 - 18R^3 + 6R^4 - 6R^5 \\ &= 3R^2 - 3R^4 + R^6 + 6R^2(1 - 3R + 3R^2 - R^3) \\ &= R_{(1)} + 6R^2(1 - R)^3. \end{aligned}$$

Ook hier volgt uit $1 - R > 0$ dat $R_{(2)} > R_{(1)}$.

Andere typen van redundantie

Naast rij-parallel-schakelingen zijn er natuurlijk nog andere manieren om redundantie in een netwerk in te bouwen. Een mogelijkheid zijn zogeheten *m-uit-n* blokken, die uit n componenten bestaan en goed werken als minstens m van de n componenten normaal functioneren.

Als in een m-uit-n blok alle componenten betrouwbaarheid R hebben, heeft de hele blok de betrouwbaarheid

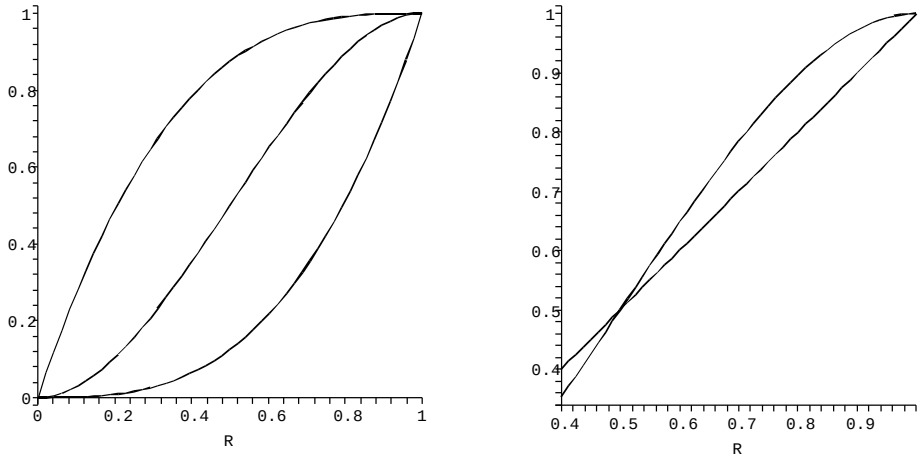
$$R_{m|n} = \sum_{k=m}^{n} \binom{n}{k} R^k (1 - R)^{n-k}.$$

Het zou geen verrassing zijn dat hier de binomiale verdeling een rol speelt, want we moeten gewoon $k \geq m$ van de n componenten kiezen de goed werken, en de kans voor het goed werken van k componenten is R^k terwijl de kans dat de resterende $n - k$ componenten niet werken $(1 - R)^{n-k}$ is.

Als speciale gevallen vinden we voor $m = 1$ de parallelschakeling en voor $m = n$ de rijschakeling terug. In het linkerplaatje van Figuur 8 zijn de betrouwbaarheden van een parallelschakeling, een 2-uit-3 blok en een rij schakeling met telkens 3 componenten als functies van de betrouwbaarheid R van een van de componenten te zien.

Een belangrijk voorbeeld van een m-uit-n blok is een 2-uit-3 blok, die de naam *triple modular redundancy system*, afgekort TMR, heeft. Zo'n systeem heeft de betrouwbaarheid

$$R_{TMR} = R_{2|3} = R^3 + 3R^2(1 - R) = R^3 + 3R^2 - 3R^3 = 3R^2 - 2R^3.$$

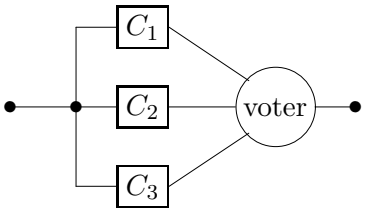


Figuur 8: Links: Betrouwbaarheden van m-uit-3 blokken voor $m = 1, 2, 3$. Rechts: Vergelijk van TMR met een van zijn componenten.

Om de betrouwbaarheid van een TMR met een van zijn componenten te vergelijken moeten we kijken wanneer $3R^2 - 2R^3 > R$ is. Voor $R > 0$ is dit het geval als $3R - 2R^2 > 1$, dus als $2R^2 - 3R + 1 < 0$. De nulpunten van $2R^2 - 3R + 1$ kunnen we (bijna) gokken, er geldt $2R^2 - 3R + 1 = (2R - 1)(R - 1)$, dus zijn de punten waar $R_{TMR} = R$ geldt de waarden $R = 0$, $R = \frac{1}{2}$ en $R = 1$. Voor $\frac{1}{2} < R < 1$ geldt dat $R_{TMR} > R$ en voor $0 < R < \frac{1}{2}$ dat $R_{TMR} < R$.

Het rechterplaatje van Figuur 8 laat het verschil tussen de betrouwbaarheid van TMR en een van zijn componenten voor $R > 0.4$ zien.

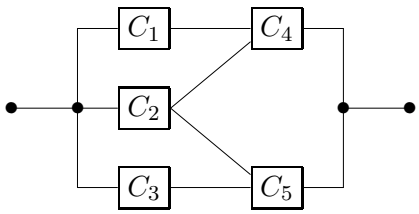
In de praktijk wordt een m-uit-n blok vaak door een *voter* gerealiseerd. De voter neemt gewoon een meerderheidsbeslissing over de resultaten van de enkele componenten.



Een probleem hierbij is dat bij componenten die alleen maar een keuze uit twee mogelijkheden hebben (ja-nee beslissing), een meerderheid van foutieve componenten tot een foutieve beslissing van de voter leidt. Daarom wordt meestal nog een detectie component achter de voter gezet, die deze gevallen moet herkennen. Maar omdat deze component ook niet perfect werkt, leidt dit tot een lagere betrouwbaarheid van de TMR.

Het voordeel van rij-parallel-schakelingen was dat we de betrouwbaarheid op de eenvoudige gevallen van zuivere rij- of parallelschakelingen terug konden brengen. Maar er is natuurlijk geen reden waarom systemen niet ook anders opgebouwd kunnen zijn. Om in dit soort systemen de betrouwbaarheid te

berekenen, is het vaak handig om naar voorwaardelijke kansen te kijken. Het volgende systeem is een voorbeeld hiervan:



Dit is geen rij-parallel-schakeling omdat we een vertakking vanuit C_2 hebben.

Een mogelijkheid om de betrouwbaarheid van dit systeem te bepalen is, alle combinaties van werken/falen voor de vijf componenten door te lopen. Dit geeft 2^5 combinaties, en voor elke combinatie kunnen we de kans van het optreden uit de betrouwbaarheden R_i van de componenten berekenen: Bijvoorbeeld heeft de combinatie waarbij C_1, C_2 en C_5 werken, maar C_3 en C_4 falen, de kans $R_1 \cdot R_2 \cdot (1 - R_3) \cdot (1 - R_4) \cdot R_5$. De betrouwbaarheid van het systeem vinden we, door de kansen van degene combinaties bij elkaar op te tellen, waarvoor er een pad langs werkende componenten door het systeem bestaat.

Maar dit is een heel moeizame methode en bij een groot aantal van componenten ondoenlijk.

Met behulp van voorwaardelijke kansen kunnen we in dit voorbeeld de betrouwbaarheid veel sneller bepalen. We bekijken gewoon de gevallen apart of de component C_2 werk of niet. We noteren met A_2 de gebeurtenis dat C_2 werkt en met A_2^c het geval dat C_2 niet werkt. Verder geven we met de gebeurtenis A aan dat het hele systeem werkt. Omdat we in een van de gevallen A_2 of A_2^c zijn, geldt $A = (A \cap A_2) \cup (A \cap A_2^c)$ en dus $P(A) = P(A, A_2) + P(A, A_2^c)$. Maar we weten dat $P(A, B) = P(A \mid B) \cdot P(B)$, daarom kunnen we de betrouwbaarheid $R = P(A)$ van het systeem schrijven als

$$R = P(A \mid A_2) \cdot P(A_2) + P(A \mid A_2^c) \cdot P(A_2^c) = P(A \mid A_2) \cdot R_2 + P(A \mid A_2^c) \cdot (1 - R_2).$$

De voorwaardelijke kansen $P(A \mid A_2)$ en $P(A \mid A_2^c)$ kunnen we nu makkelijk bepalen. Als C_2 werkt, maakt het niets uit wat er met C_1 of C_3 aan de hand is, achter C_2 hebben we een gewone parallelschakeling en er geldt

$$P(A \mid A_2) = 1 - (1 - R_4)(1 - R_5) = 1 - (1 - R_4 - R_5 + R_4 R_5) = R_4 + R_5 - R_4 R_5.$$

Aan de andere kant, als C_2 niet werkt, hebben we een parallelschakeling van de twee rijen $C_1 - C_4$ en $C_3 - C_5$, en er geldt

$$P(A \mid A_2^c) = 1 - (1 - R_1 R_4)(1 - R_3 R_5) = R_1 R_4 + R_3 R_5 - R_1 R_3 R_4 R_5.$$

Als we dit in de formule $R = P(A \mid A_2) \cdot R_2 + P(A \mid A_2^c) \cdot (1 - R_2)$ invullen, krijgen we

$$R = (R_4 + R_5 - R_4 R_5) \cdot R_2 + (R_1 R_4 + R_3 R_5 - R_1 R_3 R_4 R_5) \cdot (1 - R_2).$$

Codering van bits

Een belangrijke toepassing van het verhogen van de betrouwbaarheid met behulp van redundantie is het versturen van digitale informatie. Als we een bit versturen, wordt slechts met een zekere kans p ook hetzelfde bit ontvangen, met kans $1-p$ wordt het bit door een storing omgeschakeld. Het versturen van een bit heeft dus de betrouwbaarheid p.

Om een hoge betrouwbaarheid te bereiken, kan men een bit n keer herhalen en dan een voter gebruiken. Het is slim om n oneven te kiezen, dus $n = 2m-1$, dan is bij minstens m correct ontvangen bits de meerderheid correct en we hebben dus een m-uit-$(2m-1)$ blok. De betrouwbaarheid van deze manier van versturen is

$$R = \sum_{k=m}^{2m-1} \binom{2m-1}{k} p^k (1-p)^{2m-1-k}.$$

Het versturen van een boodschap door middel van herhaling noemt men een *repetition code*. Het is duidelijk dat dit geen efficiënte manier is, zelfs voor een 2-uit-3 blok die niet zo'n grote verbetering in de betrouwbaarheid geeft moeten we 3 keer zo veel bits versturen dan de boodschap eigenlijk bevat.

Een efficiëntere manier om digitale boodschappen met zekere redundantie te versturen zijn *foutverbeterende codes*. Het idee is dat een boodschap op een slimme manier aangevuld wordt door verdere bits, bijvoorbeeld door *parity-check bits*.

Een parity-check bit werkt als volgt: Als een boodschap van 7 bits verstuurd wordt, wordt de achtste bit op 0 gezet als de eerste 7 bits een even aantal 1en bevat, en op 1 als de eerste 7 bits een oneven aantal 1en bevat. De achtste bit is dus de (binaire) som van de eerste 7 bits. Als nu een boodschap ontvangen wordt, kan men weer checken, of het laatste bit de som van de andere bits is. Als dit niet het geval is, is er minstens een bit veranderd, maar het is niet mogelijk te zeggen welke bit veranderd is. Als de som wel klopt, is de boodschap of correct ontvangen of er zijn minstens twee bits veranderd. Maar de kans hierop is al behoorlijk kleiner dan die op een enkel veranderde bit.

Omdat we niet kunnen zeggen welke bit veranderd is als de som niet klopt, heet een code met alleen maar een parity-check bit een *foutherkennende code*.

Maar hoe werkt een foutverbeterende code? We zullen dit aan een voorbeeld bekijken, namelijk de $(7,4)$-Hamming code. Bij deze code wordt een boodschap van 4 bits aangevuld met 3 verdere bits die van de eerste 4 bits afhangen. Op deze manier zijn maar $2^4 = 16$ van de mogelijke $2^7 = 128$ woorden van 7 bits geldige codewoorden. De Hamming code wordt aangegeven door een matrix die de codewoorden voor de boodschappen 1000, 0100, 0010 en 0001 bevat. Deze matrix is

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

We kunnen elke boodschap van 4 bits als som van de vier aangegeven bood-

schappen krijgen, het bijhorende codewoord is dan de som van de overeenkomstige rijen van de matrix, bijvoorbeeld hoort bij 1010 het codewoord 1010101.

Men gaat nu na dat verschillende codewoorden steeds op minstens drie plekken verschillen. Het aantal verschillende bits van twee woorden noemt men ook de *Hamming-afstand* of kort *afstand* van de woorden. Maar als we nu een ontvangen boodschap waarbij één bit is veranderd, kunnen we de verstuurde boodschap eenduidig reconstrueren, want de ontvangen boodschap heeft slechts van een geldig codewoord afstand 1 en van alle andere geldige codewoorden minstens afstand 2. Op deze manier kunnen we dus één fout verbeteren.

Merk op dat de $(7,4)$-Hamming code heel zuinig is: Als we de woorden die afstand 1 van een geldig codewoord hebben de *buren* van dit codewoord noemen, heeft ieder van de 16 geldige codewoorden 7 buren. Maar omdat verschillende codewoorden minstens afstand 3 hebben, mogen twee codewoorden geen gemeenschappelijke buren hebben. Maar als we nu de geldige codewoorden en hun buren tellen, zijn dit er $16+16\cdot 7 = 128$, d.w.z. we hebben alle 7-bit woorden gebruikt.

Een foutverbeterende code zo als de $(7,4)$-Hamming code wordt als volgt toegepast: Een boodschap van 4 bits heeft een kans van p^4 om correct ontvangen te worden als we geen redundantie inbouwen.

De $(7,4)$-Hamming code is een 6-uit-7 blok, want we kunnen bij 6 correct ontvangen bits het verstuurde codewoord reconstrueren. De betrouwbaarheid van het versturen met behulp van de $(7,4)$-Hamming code is dus

$$R = p^7 + 7p^6(1-p) = p^7 + 7p^6 - 7p^7 = 7p^6 - 6p^7 = p^4(7p^2 - 6p^3).$$

Dit is een verbetering tegenover het gewone versturen als $7p^2 - 6p^3 > 1$. Men gaat snel na dat dit het geval is als $p > \frac{1}{2}$.

De volgende tabel voor drie waarden van p laat zien dat de Hamming-code veel beter is dan het gewone versturen en bijna zo goed als een repetition code met drievoudige herhaling, waarvoor de betrouwbaarheid bij het ontvangen van 4 bits gegeven is door $(3p^2 - 2p^3)^4$. Merk op dat bij de drievoudige herhaling 12 bits verstuurd worden, terwijl dit er bij de Hamming-code slechts 7 bits zijn, deze is dus veel zuiniger.

	direct	Hamming	repetition
0.9	0.66	0.85	0.89
0.99	0.9606	0.9980	0.9988
0.999	0.996006	0.999979	0.999988
p	p^4	$7p^6 - 6p^7$	$(3p^2 - 2p^3)^4$

7.2 Levensduur

Vaak is het bij systemen interessant de (verwachte) levensduur te bepalen. Dit is de tijd tot het falen van het systeem. Omdat we de kans dat een systeem normaal werkt met de betrouwbaarheid van het systeem aangeven, kunnen we de levensduur bepalen als we de betrouwbaarheid van de tijd laten afhangen.

Als we met T de (niet bekende) levensduur van een systeem noteren, is de betrouwbaarheid $R(t)$ op het tijdstip t de kans dat T groter is dan t. Als we T

als een stochast zien, hebben we dus

$$R(t) = P(T \geq t) = 1 - P(T \leq t).$$

Maar een stochast X met continue kansverdeling wordt meestal door een dichtheidsfunctie $f(x)$ en een verdelingsfunctie $F(x)$ beschreven, waarbij $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du$ geldt. Als nu $f(t)$ de dichtheidsfunctie voor het falen op tijdstip t is, hebben we $P(T \leq t) = \int_0^t f(\tau)\, d\tau$ en dus

$$R(t) = 1 - \int_0^t f(\tau)\, d\tau.$$

We kunnen de dichtheidsfunctie $f(t)$ ook nog iets anders interpreteren: Volgens de hoofdstelling van de calculus geldt $(\int f(x)\, dx)' = f(x)$, dus hebben we

$$R'(t) = -f(t),$$

de dichtheidsfunctie $f(t)$ geeft dus de snelheid van verandering van de betrouwbaarheid aan. Merk op dat de betrouwbaarheid met de tijd nooit toeneemt, daarom is de afgeleide $R'(t)$ steeds ≤ 0.

Met een trucje uit de calculus kunnen we de verwachtingswaarde $E(T)$ van de levensduur van een systeem rechtstreeks uit de functie $R(t)$ voor de betrouwbaarheid berekenen.

De productregel voor de afgeleide zegt dat $(f(x) \cdot g(x))' = f'(x)g(x) + f(x)g'(x)$. Als we hier op de twee zijden de integraal van nemen en gebruiken dat $\int (f(x) \cdot g(x))'\, dx = f(x) \cdot g(x)$, volgt hieruit $f(x) \cdot g(x) = \int f'(x)g(x)\, dx + \int f(x)g'(x)\, dx$. Door een van de integralen naar de andere kant te brengen, krijgen we de regel voor de *partiële integratie*:

$$\int f(x)g'(x)\, dx = f(x) \cdot g(x) - \int f'(x)g(x)\, dx.$$

Voor bepaalde integralen krijgen we

$$\int_a^b f(x)g'(x)\, dx = (f(b)g(b) - f(a)g(a)) - \int_a^b f'(x)g(x)\, dx.$$

De verwachtingswaarde $E(T)$ is gedefinieerd als

$$E(T) = \int_0^\infty t \cdot f(t)\, dt.$$

Als we hier $f(t)$ door $-R'(t)$ vervangen en de regel van de partiële integratie toepassen, krijgen we

$$E(T) = -\int_0^\infty t \cdot R'(t)\, dt = -t \cdot R(t)\, |_0^\infty + \int_0^\infty R(t)\, dt = \int_0^\infty R(t)\, dt$$

want voor $t = 0$ is $t \cdot R(t) = 0$ en voor $t \to \infty$ moet $R(t)$ sneller naar 0 gaan dan $\frac{1}{t}$, omdat de integraal $\int_0^\infty R(t)\, dt$ bestaat.

We krijgen dus de verwachtingswaarde van de levensduur door de integraal over de betrouwbaarheid te berekenen.

Tot nu toe hebben we nog geen concrete dichtheidsfunctie $f(t)$ verondersteld. We zullen hier alleen maar één belangrijk speciaal geval voor de verdeling van de levensduur behandelen, namelijk exponentieel verdeelde levensduur. Dat dit een belangrijk geval is hebben we in de vorige les bij de Poisson-processen gezien, waar de tussentijden steeds exponentieel verdeeld waren. Als we het falen van componenten als Poisson-proces zien, zijn we dus precies in dit geval. Dit veronderstelt natuurlijk dat het falen van een component onafhankelijk van de geschiedenis van de component is, in het bijzonder vindt geen ouderdomsverzwakking plaats.

We zeggen dat de levensduur T van een systeem *exponentieel verdeeld* met parameter λ is, als de dichtheidsfunctie $f(t)$ voor het falen gegeven is door $f(t) = \lambda e^{-\lambda t}$. In dit geval geldt $P(T \leq t) = 1 - e^{-\lambda t}$ en dus

$$R(t) = 1 - (1 - e^{-\lambda t}) = e^{-\lambda t}.$$

De verwachte levensduur van zo'n systeem is

$$E(T) = \int_0^\infty e^{-\lambda t}\, dt = -\frac{1}{\lambda} e^{-\lambda t} \,|_0^\infty = \frac{1}{\lambda}.$$

(We wisten natuurlijk al lang dat een exponentiële verdeling met parameter λ verwachtingswaarde $\frac{1}{\lambda}$ heeft.)

We zullen nu de levensduur voor bepaalde combinaties van componenten met exponentieel verdeelde levensduur bepalen, te weten voor een rijschakeling, een parallelschakeling en een TMR blok.

Rijschakeling

We gaan ervan uit dat we n componenten in een rij hebben waarvan de i-de een exponentieel verdeelde levensduur met parameter λ_i heeft. Voor de betrouwbaarheid van een rijschakeling hadden we gezien dat $R = \prod_{i=1}^n R_i$, dus hebben we nu

$$R(t) = \prod_{i=1}^n R_i(t) = \prod_{i=1}^n e^{-\lambda_i t} = e^{-\lambda_1 t} \cdot \ldots \cdot e^{-\lambda_n t} = e^{-\lambda_1 t - \ldots - \lambda_n t} = e^{-(\sum_{i=1}^n \lambda_i)t}.$$

De levensduur van het systeem is dus exponentieel verdeeld met parameter $\lambda_1 + \ldots \lambda_n = \sum_{i=1}^n \lambda_i$ en de verwachte levensduur is

$$E(T) = \frac{1}{\sum_{i=1}^n \lambda_i}.$$

In het bijzonder is de verwachte levensduur van het systeem korter dan die van elke van zijn componenten. Als alle componenten dezelfde parameter λ hebben, is de verwachte levensduur van het systeem slechts een $n$de van die van de componenten.

Parallelschakeling

Bij een parallelschakeling is de levensduur van het systeem het maximum van de levensduren van zijn componenten.

De verwachte levensduur van een systeem met twee parallelle componenten met exponentieel verdeelde levensduren met parameters λ_1 en λ_2 vinden we als volgt: De betrouwbaarheid van een gewoon systeem met twee parallelle componenten was $R = 1 - (1 - R_1)(1 - R_2)$, dus hebben we

$$\begin{aligned} R(t) &= 1 - (1 - R_1(t))(1 - R_2(t)) = 1 - (1 - e^{-\lambda_1 t})(1 - e^{-\lambda_2 t}) \\ &= e^{-\lambda_1 t} + e^{-\lambda_2 t} - e^{-(\lambda_1+\lambda_2)t}. \end{aligned}$$

We hebben al gezien dat $\int_0^\infty e^{-\lambda t}\ dt = \frac{1}{\lambda}$, daarom geldt voor de verwachte levensduur:

$$E(T) = \int_0^\infty R(t)\ dt = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} - \frac{1}{\lambda_1 + \lambda_2}.$$

De verwachte levensduur is dus korter dan de som van de verwachte levensduren van de componenten. In het geval $\lambda_1 = \lambda_2 = \lambda$ is de verwachte levensduur van het systeem $\frac{3}{2} \cdot \frac{1}{\lambda}$, dus om 50% tegenover de componenten verhoogd.

Bij een parallelschakeling van n componenten met exponentieel verdeelde levensduren met dezelfde parameter λ laat zich aantonen dat het systeem de verwachte levensduur

$$E(T) = \frac{1}{\lambda}(1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n})$$

heeft. Voor grote waarden van n geldt $1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} \approx \log(n) + 0.577$, dit betekent dat de levensduur van het systeem voor groeiende n slechts zo als $\log(n)$ toeneemt. Omdat de logaritme heel langzaam groeit is een parallelschakeling dus alleen maar voor 2 of 3 componenten een efficiënte manier om de levensduur te verhogen.

TMR blok

We hadden gezien dat een TMR blok (of 2-uit-3 blok) met betrouwbaarheid R van de componenten betrouwbaarheid $R_{TMR} = 3R^2 - 2R^3$ heeft. Voor de betrouwbaarheid van een TMR blok waarbij de componenten exponentieel verdeelde levensduur met parameter λ hebben, geldt dus

$$R(t) = 3e^{-2\lambda t} - 2e^{-3\lambda t}.$$

Voor de verwachte levensduur $E(T)$ krijgen we

$$\begin{aligned} E(T) &= \int_0^\infty R(t)\ dt = \int_0^\infty 3e^{-2\lambda t}\ dt - \int_0^\infty 2e^{-3\lambda t}\ dt \\ &= -\frac{3}{2\lambda}e^{-2\lambda t}\ |_0^\infty + \frac{2}{3\lambda}e^{-3\lambda t}\ |_0^\infty = \frac{3}{2\lambda} - \frac{2}{3\lambda} = \frac{5}{6} \cdot \frac{1}{\lambda}. \end{aligned}$$

De verwachte levensduur is dus *korter* dan bij de enkele componenten! Dit lijkt paradox, want we hadden gezien dat voor $R > \frac{1}{2}$ de TMR blok een grotere betrouwbaarheid heeft dan R. Maar dit is juist de sleutel voor de verklaring van

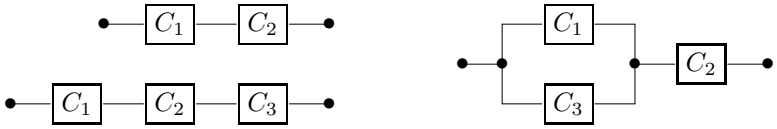
de paradox. De TMR blok is betrouwbarer dan een van zijn componenten als deze betrouwbaarheid $> \frac{1}{2}$ hebben. Maar de betrouwbaarheid van de componenten is gegeven door $e^{-\lambda t}$ en $e^{-\lambda t} > \frac{1}{2}$ is equivalent met $t < \log(2) \cdot \frac{1}{\lambda} \approx 0.7 \cdot \frac{1}{\lambda}$. De TMR blok is dus alleen maar voor kortere tijdsduren betrouwbarer dan een van de componenten, de grotere betrouwbaarheid van de enkele component bij grotere tijden zorgt ervoor dat de verwachtingswaarde voor de levensduur groter is dan bij de TMR blok.

Belangrijke begrippen in deze les

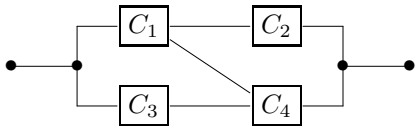
- betrouwbaarheid
- redundantie
- rijschakeling, parallelschakeling, rij-parallel-schakeling
- m-uit-n blok, voter
- foutverbeterende code
- (verwachte) levensduur

Opgaven

29. Laten C_1, C_2 en C_3 drie componenten met bijhorende betrouwbaarheden R_1, R_2 en R_3 zijn. Bereken de betrouwbaarheden van de volgende combinaties van componenten voor algemene R_i en voor de speciale waarden $R_1 = 0.8$, $R_2 = 0.75$, $R_3 = 0.98$.



30. Bereken de betrouwbaarheid van het volgende systeem, afhankelijk van de betrouwbaarheden R_1, R_2, R_3 en R_4 van de enkele componenten.



Stel je hebt twee componenten met een betrouwbaarheid van 90% beschikbaar en twee componenten met een betrouwbaarheid van slechts 80%. Op welke posities moet je de betere componenten plaatsen en welke betrouwbaarheid kun je zo maximaal bereiken?

31. Om een zeker programma uit te voeren, moeten op een multiprocessor-computer met een kans van 95% minstens 2 processoren beschikbaar zijn. Als designer van de multiprocessor-computer kun je verschillende typen van processoren inbouwen: De goedkoopste processor met een betrouwbaarheid van 60% kost 1000€ en elke verhoging van de betrouwbaarheid om 10% kost 800€ extra (dus betaal je 3400€ voor een processor met een betrouwbaarheid van 90%.

Wat is de goedkopste manier om aan de eisen van het programma te voldoen als je een parallelschakeling van *gelijksoortige* processoren bouwt? Is er een goedkopere manier als je verschillende typen van processoren mag gebruiken?

32. De $(23, 12)$-Golay code is een lineaire code met 2^{12} legale codewoorden van lengte 23 (d.w.z. boodschappen van 12 bits worden door 11 check-bits aangevuld). De legale codewoorden in de Golay code hebben een Hamming-afstand van minstens 7, daarom kan de Golay code 3 fouten verbeteren. We kunnen daarom bij een ontvangen boodschap van 23 bits de oorspronkelijke boodschap eenduidig reconstrueren, als er hoogstens 3 bits zijn veranderd. Stel dat één bit met kans p correct wordt ontvangen. Bereken de betrouwbaarheid van het versturen van een boodschap van 12 bits met behulp van de $(23, 12)$-Golay code.

    Je kunt een boodschap van 12 bits natuurlijk ook als 3 blokken van 4 bits met behulp van de $(7, 4)$-Hamming code versturen. Vergelijk voor $p = 0.9$ en $p = 0.99$ de betrouwbaarheden van het versturen met behulp van de $(23, 12)$-Golay code en met behulp van de $(7, 4)$-Hamming code (in drie blokken). Vergelijk de betrouwbaarheden ook met de betrouwbaarheid bij het versturen met drievoudige herhaling, dus met een 2-uit-3 blok voor elke bit.

33. In een systeem dat twee stappen bevat is elke component verdubbeld. Bepaal voor de volgende twee designs van het systeem de verwachte levensduur als alle componenten een exponentieel verdeelde levensduur met parameter λ hebben. Welke design geeft een grotere levensduur?

    Bepaal de verwachte levensduur van de systemen ook voor het geval dat de componenten C_1 een exponentieel verdeelde levensduur met parameter λ_1 hebben en de componenten C_2 een exponentieel verdeelde levensduur met parameter λ_2.

    Laat zien dat het design in het rechterplaatje voor alle waarden van λ_1 en λ_2 een hogere verwachte levensduur geeft. (Hiervoor is het handig, de verwachte levensduren van de twee designs van elkaar af te trekken en aan te nemen dat $\lambda_2 = 1$ is.)

Les 8 Proces analyse

Veel processen laten zich door netwerken beschrijven, waarin knopen acties aangeven en opdrachten langs verbindingen tussen de knopen verwerkt worden. Typische elementen van dit soort netwerken zijn:

- **rijen:** een aantal acties worden achter elkaar uitgevoerd,
- **en-splitsingen:** een aantal alternatieven worden parallel toegepast,
- **of-splitsingen:** van een aantal alternatieven wordt er één toegepast,
- **iteratie:** een actie wordt meerdere keren achter elkaar toegepast.

We zullen in deze les kijken hoe we netwerken kunnen analyseren, om bijvoorbeeld de volgende vragen te kunnen beantwoorden:

(1) Wat is de zwakste schakel in het netwerk, dus welke component van het netwerk moeten we verbeteren om vooruitgang in de algemene prestatie van het netwerk te boeken?

(2) Hoe kunnen we bepalen wat de gemiddelde tijd is, waarmee een opdracht verwerkt wordt en wat is de tijd waarin 90% van de opdrachten door het netwerk gesluisd wordt?

Om dit soort vraagstukken op te lossen, hebben we natuurlijk zekere informatie over de componenten van het netwerk nodig.

Vaak wordt er voor elke actie in het netwerk een verwerkingstijd aangegeven die het uitvoeren van deze actie in beslag neemt. Dit kan de gemiddelde tijdsduur voor de taak zijn, maar soms ook de tijd waarin een zekere percentage (bijvoorbeeld 95%) van de gevallen voltooid wordt. We zullen zien, dat dit soort informatie voldoende is, om kritieke componenten van een netwerk te identificeren.

Voor vragen over de verdeling van de verwerkingstijden van opdrachten door het netwerk hebben we natuurlijk meer informatie nodig, namelijk kansverdelingen voor de tijdsduur die elke actie in beslag neemt. Deze verdelingen zijn of discrete of continue kansverdelingen.

(a) **Discrete kansverdeling:** Als we aannemen dat de tijd voor een actie alleen maar veelvouden van een 'eenheidsinterval' t_0 (bijvoorbeeld tijden afgerond op hele seconden) aanneemt, kunnen we een kansverdeling door $P(T = k \cdot t_0)$ voor $k = 0, 1, 2, \ldots$ aangeven. Hiervoor zullen we kort $P(k)$ schrijven.

Natuurlijk is het vaak ook interessant, naar de kans te kijken dat de tijdsduur T van een actie onder een zekere bovengrens ligt. Deze kans is gelijk aan de som over de kansen voor tijden tot en met de grens, we hebben dus $P(T \leq k \cdot t_0) = \sum_{j=0}^{k} P(j)$. Deze kans gaan we met $F(k)$ afkorten en noemen dit de *discrete verdelingsfunctie*.

Omgekeerd kunnen we uit de discrete verdelingsfunctie $F(k)$ ook de kansverdeling $P(k)$ weer makkelijk achterhalen, er geldt namelijk $P(k) = F(k) - F(k-1)$.

(b) **Continue kansverdeling:** Als we willekeurige positieve tijden toelaten, hebben we een continue kansverdeling nodig. Deze wordt door een dichtheidsfunctie $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ met $\int_0^\infty f(t)\ dt = 1$ aangegeven. De kans $P(T \leq t)$ dat een actie hoogstens de tijdsduur t in beslag neemt is dan gegeven door de verdelingsfunctie $F(t) := \int_0^t f(\tau)\ d\tau$, waarbij de integraal de oppervlakte onder de grafiek van $f(t)$ op het interval $[0, t]$ aangeeft.

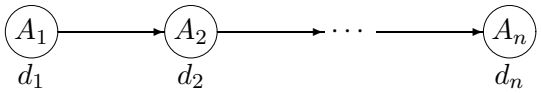
8.1 Elementen van netwerken

We zullen nu naar de typische componenten van netwerken kijken en hun eigenschappen met betrekking tot verwerkingstijden en tijdsduurverdelingen beschrijven.

Actie

Een actie A wordt door een knoop gerepresenteerd. De verwerkingstijd d (voor *duration*) wordt aangegeven door een waarde die aan de knoop hangt.

Rij



De verwerkingstijd van een rij is natuurlijk de som

$$d = d_1 + d_2 + \ldots + d_n = \sum_{i=1}^{n} d_i$$

van de verwerkingstijden van de enkele componenten.

Stel nu dat we discrete kansverdelingen voor de tijdsduur van de enkele acties hebben. We hoeven alleen maar naar het geval van twee acties te kijken, voor langere rijen kunnen we het resultaat hiervoor herhaaldelijk toepassen. De kansverdeling voor de actie A_1 noteren we met $P_1(k) := P_1(T = k)$ en de verdeling voor A_2 met $P_2(k) := P_2(T = k)$.

De achter elkaar geschakelde acties A_1 en A_2 nemen t tijdseenheden in beslag als A_1 een tijd van k en A_2 een tijd van $t - k$ duurt. Om de kans hiervoor te berekenen, moeten we over alle mogelijke waarden van k tussen 0 en t lopen en krijgen zo

$$P(T = t) = \sum_{k=0}^{t} P_1(k) \cdot P_2(t - k).$$

Een op deze manier *gevouwen* som van producten speelt in verschillende gebieden van de wiskunde een belangrijke rol (bijvoorbeeld ook bij Fourier transformaties) en heet een (discreet) *convolutieproduct*. Het convolutieproduct wordt meestal met een sterretje genoteerd, dus

$$P(t) = (P_1 * P_2)(t).$$

Let op dat je zo'n convolutieproduct niet met het gewone product $P_1(t) \cdot P_2(t)$ verwisselt.

> Er is een situatie waar we een convolutieproduct toepassen zonder hier verder over na te denken. Als we namelijk twee veeltermen $f(x) = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$ en $g(x) = b_4x^4 + b_3x^3 + b_2x^2 + b_1x + b_0$ met elkaar vermenigvuldigen, dan is het duidelijk dat we de coëfficiënt van x^5 in het product $f(x) \cdot g(x)$ vinden als $a_1b_4 + a_2b_3 + a_3b_2 + a_4b_1$. We tellen dus de producten van coëfficiënten bij elkaar op die bij machten van x met som 5 horen. Maar dit is juist een convolutieproduct, namelijk $\sum_{k=0}^{5} a_k b_{5-k}$ (waarbij we a_5 en b_5 als 0 definiëren).

Met behulp van het convolutieproduct kunnen we ook de discrete verdelingsfunctie F voor de kans op een verwerkingstijd van hoogstens t uit de verdelingsfuncties F_1 en F_2 voor A_1 en A_2 berekenen. We hebben namelijk

$$\begin{aligned} F(t) = P(T \leq t) &= \sum_{s=0}^{t} P(T = s) = \sum_{s=0}^{t} \left(\sum_{k=0}^{s} P_1(k) \cdot P_2(s-k) \right) \\ &= \sum_{s=0}^{t} \left(\sum_{k=0}^{s} P_1(k) \right) \cdot P_2(t-s) = \sum_{s=0}^{t} F_1(s) \cdot P_2(t-s) = (F_1 * P_2)(t). \end{aligned}$$

Net zo goed kunnen we F ook door $F(t) = (P_1 * F_2)(t)$ berekenen.

Op een soortgelijke manier kunnen we de verdeling van de tijdsduren van een rij acties ook voor continue kansverdelingen bepalen. Hierbij nemen de dichtheidsfuncties $f_1(t)$ en $f_2(t)$ de plaats van de kansverdelingen P_1 en P_2 in en wordt de som een integraal. De dichtheidsfunctie $f(t)$ voor de kans dat A_1 en A_2 samen t seconden in beslag nemen is dan

$$f(t) = \int_0^t f_1(\tau) \cdot f_2(t-\tau)\, d\tau$$

een dit noemen we het convolutieproduct van de functies f_1 en f_2, weer genoteerd door $f(t) = (f_1 * f_2)(t)$. De verdelingsfunctie voor de gecombineerde verwerkingstijd van de acties A_1 en A_2 is dan

$$F(t) = P(T \leq t) = \int_0^t f(\tau)\, d\tau = \int_0^t \left(\int_0^\tau f_1(x) \cdot f_2(\tau - x)\, dx \right).$$

Voorbeeld: We kijken naar een rij van twee acties A_1 en A_2 met exponentieel verdeelde tijdsduren met parameters λ_1 en λ_2 (dit betekent dat de verwachtingswaarden voor de verwerkingstijden λ_1^{-1} en λ_2^{-1} zijn). De dichtheidsfunctie voor een exponentiële verdeling met parameter λ is $\lambda e^{-\lambda t}$, dus hebben we voor de dichtheidsfunctie van de acties:

$$f(t) = \int_0^t \lambda_1 e^{-\lambda_1 \tau} \cdot \lambda_2 e^{-\lambda_2 (t-\tau)}\, d\tau = \lambda_1 \lambda_2 e^{-\lambda_2 t} \int_0^t e^{(\lambda_2 - \lambda_1)\tau}\, d\tau.$$

Voor het speciaal geval $\lambda_1 = \lambda_2 = \lambda$ geeft dit de dichtheidsfunctie

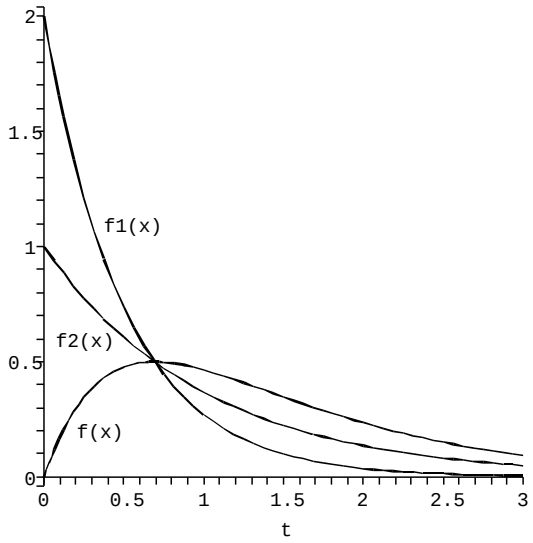
$$f(t) = \lambda^2 t e^{-\lambda t}$$

en in dit geval is de verdelingsfunctie voor de rij van acties

$$F(t) = \int_0^t f(\tau)\, d\tau = \int_0^t \lambda^2 \tau e^{-\lambda\tau}\, d\tau = 1 - (1+\lambda t)e^{-\lambda t}.$$

Voor het algemeen geval $\lambda_1 \neq \lambda_2$ hebben we

$$f(t) = \lambda_1\lambda_2\, e^{-\lambda_2 t}\frac{1}{\lambda_2 - \lambda_1}(e^{(\lambda_2-\lambda_1)t} - 1) = \frac{\lambda_1\lambda_2}{\lambda_2 - \lambda_1}(e^{-\lambda_1 t} - e^{-\lambda_2 t}).$$



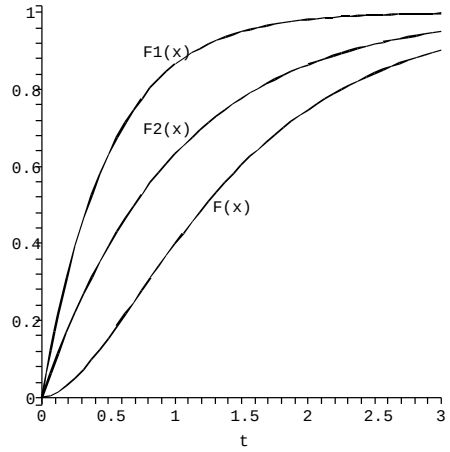
Figuur 9: Dichtheidsfuncties voor twee exponentieel verdeelde acties met parameters $\lambda_1 = 2$ en $\lambda_2 = 1$ en de rij van deze acties.

In Figuur 9 zijn de dichtheidsfuncties voor twee exponentiële verdelingen $f_1(t)$ en $f_2(t)$ met parameters $\lambda_1 = 2$ en $\lambda_2 = 1$ en hun convolutieproduct $f(t) = (f_1 * f_2)(t)$ te zien. De functies zijn makkelijk te identificeren omdat een exponentiële verdeling met parameter λ in het punt $t = 0$ de waarde λ heeft.

Om de verdelingsfunctie voor de gecombineerde actie te berekenen, moeten we de dichtheidsfunctie $f(t)$ nog integreren, we hebben

$$F(t) = \int_0^t \frac{\lambda_1\lambda_2}{\lambda_2 - \lambda_1}(e^{-\lambda_1 t} - e^{-\lambda_2 t}) = 1 - (\frac{\lambda_2}{\lambda_2 - \lambda_1}e^{-\lambda_1 t} - \frac{\lambda_1}{\lambda_2 - \lambda_1}e^{-\lambda_2 t}).$$

Voor de verdeling $f(t)$ van de rij van acties laat zich aantonen dat de verwachtingswaarde $\int_0^\infty f(t)\, dt = \frac{\lambda_1+\lambda_2}{\lambda_1\lambda_2}$ is, dit is juist de som $\frac{1}{\lambda_1} + \frac{1}{\lambda_2}$ van de verwachtingswaarden van de twee acties. Voor het geval $\lambda_1 = 2$ en $\lambda_2 = 1$ is de verwachtingswaarde dus $\frac{3}{2}$.

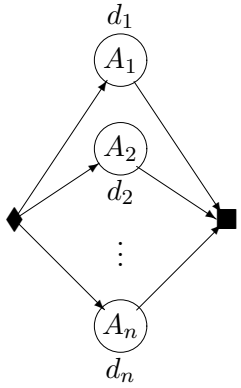


Figuur 10: Verdelingsfuncties voor twee exponentieel verdeelde acties met parameters $\lambda_1 = 2$ en $\lambda_2 = 1$ en de rij van deze acties.

In Figuur 10 zijn de verdelingsfuncties $F_1(t)$, $F_2(t)$ en $F(t)$ voor de dichtheidsfuncties $f_1(t)$, $f_2(t)$ en $f(t)$ van de acties van boven te zien. Het is duidelijk dat de verdelingsfunctie voor de verdeling met de kleinste verwachtingswaarde het hoogste ligt, daarom zijn de functies van boven naar beneden $F_1(T)$, $F_2(T)$, $F(T)$.

En-splitsing

Een en-splitsing ligt tussen een ingevulde ruit ♦ en een ingevulde vierkant ■.



Als we een en-splitsing met de alternatieve acties $A_1, A_2, \ldots, A_n$ hebben, die alle parallel uitgevoerd moeten worden, is de verwerkingstijd d natuurlijk het maximum van de verwerkingstijden van de enkele componenten, dus we hebben

$$d = \max\{d_1, d_2, \ldots, d_n\}.$$

Om de kansverdeling voor de tijdsduur van de splitsing te vinden, gebruiken we de verdelingsfuncties F_i van de enkele acties A_i. Hierbij maakt het niets uit of we met discrete of continue kansverdelingen te maken hebben. De en-splitsing

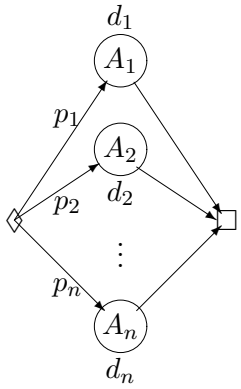
verwerkt een opdracht in een tijd $T \leq t$ als elke actie A_i de opdracht in een tijd $T \leq t$ verwerkt. Omdat we aannemen dat de tijden voor de verschillende acties onafhankelijk van elkaar zijn, is de kans dat alle acties A_i hoogstens een duur van t hebben juist het product van de kansen voor de enkele acties, daarom hebben we voor de verdelingsfunctie $F(t)$ van de en-splitsing:

$$F(t) = F_1(t) \cdot F_2(t) \cdot \ldots \cdot F_n(t) = \prod_{i=1}^{n} F_i(t).$$

Als de kansverdelingen discreet zijn, krijgen we uit de verdelingsfunctie de kans $P(k) = P(T = k \cdot t_0)$ makkelijk door $P(k) = F(k) - F(k-1)$. Voor continue kansverdelingen moeten we de verdelingsfunctie afleiden en krijgen de dichtheidsfunctie als $f(t) = F'(t)$.

Of-splitsing

Een of-splitsing ligt tussen een niet ingevulde ruit $\Diamond$ en een niet ingevulde vierkant $\Box$.



Een of-splitsing bestaat uit alternatieve acties, waarvan slechts één uitgevoerd hoeft te worden. Als aanvullende informatie moeten we hierbij nog aangeven, met welke kans p_i de enkele acties ingegaan worden.

Voor een *worst-case* analyse zouden we ook hier het maximum van de verwerkingstijden van de enkele acties moeten nemen. Maar als we willen bepalen, wat voor een tijd we typisch kunnen verwachten, is het voor de hand liggend de verwachtingswaarde van de verwerkingstijden te berekenen, dus we krijgen

$$d = p_1 d_1 + p_2 d_2 + \ldots + p_n d_n = \sum_{i=1}^{n} p_i d_i.$$

Net als bij de en-splitsing gebruiken we de verdelingsfuncties van de enkele acties om de verdelingsfunctie voor de hele splitsing te vinden. De kans, dat de splitsing een opdracht in een tijd $T \leq t$ verwerkt is de gewogen som van de kansen dat een enkele actie de opdracht in een tijd $T \leq t$ verwerkt. De gewichten zijn hierbij de kansen p_i waarmee de actie A_i ingegaan wordt. We hebben dus (weer voor discrete en continue kansverdelingen):

$$F(t) = p_1 F_1(t) + p_2 F_2(t) + \ldots + p_n F_n(t) = \sum_{i=1}^{n} p_i F_i(t).$$

Omdat we de verdelingsfunctie als gewogen som krijgen, kunnen we er ook meteen de kansverdeling of dichtheidsfunctie van afleiden. In het discrete geval geldt

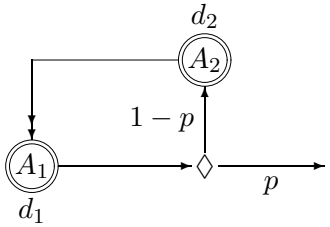
$$P(k) = P(T = k) = p_1 P_1(k) + p_2 P_2(k) + \ldots + p_n P_n(k) = \sum_{i=1}^{n} p_i P_i(k)$$

en in het continue geval

$$f(t) = P(T \leq t)' = p_1 f_1(t) + p_2 f_2(t) + \ldots + p_n f_n(t) = \sum_{i=1}^{n} p_i f_i(t).$$

Als we aannemen dat de bij actie A_i aangegeven verwerkingstijd d_i de verwachtingswaarde van de kansverdeling voor deze actie is, volgt hieruit dat de gemiddelde kans $d = \sum_{i=1}^{n} p_i d_i$ juist de verwachtingswaarde voor de verwerkingstijd door de hele of-splitsing is.

Iteratie



In een iteratie hebben we een actie A_1, waarna een beslissing wordt genomen of we door gaan (dit gebeurt met kans p) of via een actie A_2 terug naar A_1 gaan (dit gebeurt met kans $1-p$). Beide acties A_1 en A_2 kunnen in principe leeg zijn en een duur van 0 hebben.

In principe kunnen we een iteratie zien als een speciale of-splitsing met oneindig veel alternatieven, want we kunnen de lus willekeurig vaak doorlopen voor dat we er uit komen. De alternatieven en hun verwerkingstijden zijn dus:

0 iteraties: actie A_1, tijd d_1, kans p

1 iteratie: acties A_1, A_2, A_1, tijd $2d_1 + d_2$, kans $(1-p)p$

2 iteraties: acties A_1, A_2, A_1, A_2, A_1, tijd $3d_1 + 2d_2$, kans $(1-p)^2 p$

$\vdots$

k iteraties: acties $A_1, A_2, A_1, \ldots, A_2, A_1$, tijd $(k+1)d_1 + kd_2$, kans $(1-p)^k p$

Als we k keer itereren voeren we dus $k+1$ keer de actie A_1 en k keer de actie A_2 uit en hebben hiervoor de tijd $(k+1)d_1 + kd_2$ nodig en dit gebeurt met kans $(1-p)^k p$.

Zo als bij de of-splitsing bepalen we de verwerkingstijd als verwachtingswaarde voor de verwerkingstijden van de enkele alternatieven, maar deze keer is dit een som over oneindig veel alternatieven, omdat we de lus willekeurig vaak kunnen doorlopen. Om dit uit te werken, hebben we als hulpmiddel de *meetkundige reeks* nodig.

De meetkundige reeks is de oneindige som $1 + x + x^2 + \ldots = \sum_{k=0}^{\infty} x^k$ en er geldt dat

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$$

als $|x| < 1$ is. (Als $|x| > 1$ is, bestaat de som niet, omdat de waarden van x^k steeds groter worden.) Deze gelijkheid ziet men in door na te gaan dat $(1-x)(\sum_{k=0}^{n} x^k) = 1 - x^{n+1}$ en dus $\sum_{k=0}^{n} x^k = \frac{1-x^{n+1}}{1-x}$. Als $|x| < 1$ gaat $x^n \to 0$ voor $n \to \infty$ en dus gaat $\sum_{k=0}^{\infty} x^k \to \frac{1}{1-x}$.

Als een eerste toepassing van de meetkundige reeks kunnen we nagaan dat de kansen voor het doorlopen van 0, 1, 2, enz. iteraties inderdaad som 1 hebben. De kans voor k iteraties is $(1-p)^k p$ en we hebben

$$\sum_{k=0}^{\infty}(1-p)^k p = (\sum_{k=0}^{\infty}(1-p)^k) \cdot p = \frac{1}{1-(1-p)} \cdot p = \frac{1}{p} \cdot p = 1.$$

Als we de relatie $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$ op beide zijden naar x afleiden (waarbij we de som aan de linkerkant termsgewijs afleiden), krijgen we $\sum_{k=0}^{\infty} k x^{k-1} = (\sum_{k=0}^{\infty} x^k)' = (\frac{1}{1-x})' = \frac{1}{(1-x)^2}$. Als we hier voor x weer $1-p$ invullen, hebben we

$$\sum_{k=0}^{\infty} k(1-p)^{k-1} = \frac{1}{(1-(1-p))^2} = \frac{1}{p^2}.$$

Voor de verwachtingswaarde van de verwerkingstijd van de iteratie volgt hiermee:

$$\begin{aligned}
d &= \sum_{k=0}^{\infty}(1-p)^k p((k+1)d_1 + kd_2) \\
&= pd_1 \cdot \sum_{k=0}^{\infty}(1-p)^k(k+1) + pd_2 \cdot \sum_{k=0}^{\infty}(1-p)^k k \\
&= pd_1 \cdot \sum_{k=0}^{\infty}(1-p)^k(k+1) + pd_2(1-p) \cdot \sum_{k=0}^{\infty}(1-p)^{k-1} k \\
&= pd_1 \cdot \frac{1}{p^2} + pd_2(1-p) \cdot \frac{1}{p^2} = \frac{d_1}{p} + \frac{d_2(1-p)}{p} = \frac{d_1+d_2}{p} - d_2
\end{aligned}$$

Door de interpretatie van de iteratie als of-splitsing kunnen we in principe ook de resultaten van boven over de kansverdelingen voor de tijdsduur toepassen. We hebben het echter hierbij met een oneindige som te maken, waarin steeds langere convolutieproducten voorkomen. In de praktijk wordt daarom een maximum voor het aantal doorlopen van de iteratie gekozen en de som daar afgebroken. Dit is een redelijke aanpak, omdat $(1-p)^k$ met groeiende k snel klein wordt en de verdere termen daarom bijna geen rol meer spelen.

Merk op: Bij het afbreken van een iteratie moeten we opletten, dat we de kansmassa van de *vergeten* opties niet kwijt raken. Als we bijvoorbeeld na

één rondje van de iteratie afbreken, hebben we een of-splitsing met de optie A_1 met kans p en de optie A_1, A_2, A_1 met kans $(1-p)p$. Maar deze twee kansen bij elkaar opgeteld geven niet 1, maar slechts $1-(1-p)^2$, we zijn namelijk de iteraties met 2 of meer lusjes vergeten. De eenvoudigste (en ook redelijkste) manier om dit op te lossen, is, de kansen voor de opties met 2 of meer lusjes bij de optie met 1 iteratie op te tellen. Dit kunnen we zo interpreteren, dat we na 1 iteratie gedwongen zijn om de iteratie te verlaten.

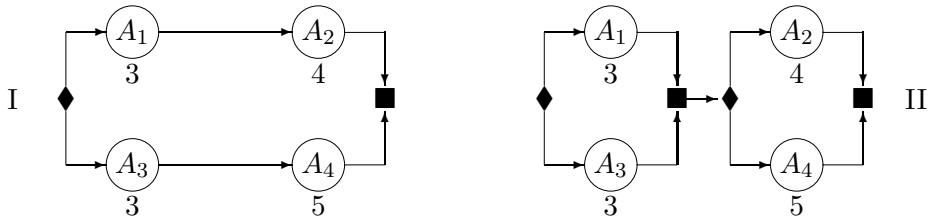
Als we met deze aanpak bij $k = 2$ afbreken, krijgen we (in het continue geval) de dichtheidsfunctie:

$$f(t) = pf_1(t) + p(1-p)(f_1 * f_2 * f_1)(t) + (1-p)^2(f_1 * f_2 * f_1 * f_2 * f_1)(t).$$

Merk op dat we bij $f_1 * f_2 * f_1 * f_2 * f_1$ eigenlijk de kans $p(1-p)^2$ zouden hebben, maar deze is nu tot $(1-p)^2$ verhoogd.

Twee voorbeelden

We gaan de behandelde principes nu eens op de twee volgende eenvoudige netwerken toepassen:



Als we de verwerkingstijden voor de netwerken bepalen, krijgen we $d = \max(3+4, 3+5) = 8$ voor netwerk I en $d = \max(3,3) + \max(4,5) = 3+5 = 8$ voor netwerk II.

Nu veronderstellen we dat de aangegeven verwerkingstijden verwachtingswaarden zijn en dat de kansverdeling voor de actie A_i met verwachtingswaarde d_i gegeven is door

$$P_i(d_i - 1) = 0.25, \quad P_i(d_i) = 0.5, \quad P_i(d_i + 1) = 0.25.$$

Dit betekent dat de helft van de acties in de aangegeven tijd uitgevoerd worden, dat een kwart om een tijdseenheid sneller en een kwart om een tijdseenheid langzamer uitgevoerd wordt. Voor deze kansverdelingen gaan we nu de verdelingen van de verwerkingstijden voor de twee netwerken bepalen. We beginnen met netwerk I.

Voor de kansverdeling $P_{12}(t)$ van de rij $A_1 - A_2$ van acties krijgen we met behulp van het convolutieproduct:

$$\begin{aligned}
P_{12}(5) &= 0.25 \cdot 0.25 = 0.0625 \\
P_{12}(6) &= 0.25 \cdot 0.5 + 0.5 \cdot 0.25 = 0.25 \\
P_{12}(7) &= 0.25 \cdot 0.25 + 0.5 \cdot 0.5 + 0.25 \cdot 0.25 = 0.375 \\
P_{12}(8) &= 0.5 \cdot 0.25 + 0.25 \cdot 0.5 = 0.25 \\
P_{12}(9) &= 0.25 \cdot 0.25 = 0.0625.
\end{aligned}$$

De verdelingsfunctie $F_{12}(t)$ voor deze rij van acties is dus

$$\begin{aligned}
F_{12}(4) &= 0 \\
F_{12}(5) &= F_{12}(4) + P_{12}(5) = 0 + 0.0625 = 0.0625 \\
F_{12}(6) &= F_{12}(5) + P_{12}(6) = 0.0625 + 0.25 = 0.3125 \\
F_{12}(7) &= F_{12}(6) + P_{12}(7) = 0.3125 + 0.375 = 0.6875 \\
F_{12}(8) &= F_{12}(7) + P_{12}(8) = 0.6875 + 0.25 = 0.9375 \\
F_{12}(9) &= F_{12}(8) + P_{12}(9) = 0.9375 + 0.0625 = 1.
\end{aligned}$$

Op dezelfde manier krijgen we voor de rij $A_3 - A_4$ van acties de kansverdeling $P_{34}(t)$ en verdelingsfunctie $F_{34}(t)$ met

$$\begin{aligned}
P_{34}(5) &= 0, & F_{34}(5) &= 0 \\
P_{34}(6) &= 0.0625, & F_{34}(6) &= 0.0625 \\
P_{34}(7) &= 0.25, & F_{34}(7) &= 0.3125 \\
P_{34}(8) &= 0.375, & F_{34}(8) &= 0.6875 \\
P_{34}(9) &= 0.25, & F_{34}(9) &= 0.9375 \\
P_{34}(10) &= 0.0625, & F_{34}(10) &= 1.
\end{aligned}$$

Omdat we het met een en-splitsing te maken hebben, krijgen we de verdelingsfunctie $F(t)$ voor het hele netwerk als product van de verdelingsfuncties $F_{12}(t)$ en $F_{34}(t)$, dus er geldt $F(t) = F_{12}(t) \cdot F_{34}(t)$. Dit geeft de verdelingsfunctie

$$\begin{aligned}
F(5) &= 0.0625 \cdot 0 = 0 \\
F(6) &= 0.3125 \cdot 0.0625 = 0.0195 \\
F(7) &= 0.6875 \cdot 0.3125 = 0.2148 \\
F(8) &= 0.9375 \cdot 0.6875 = 0.6445 \\
F(9) &= 1 \cdot 0.9375 = 0.9375 \\
F(10) &= 1 \cdot 1 = 1
\end{aligned}$$

Door $P(t) = F(t) - F(t-1)$ krijgen we hieruit de kansverdeling $P(t)$ voor het netwerk, dus

$$P(6) = 0.0195,\ P(7) = 0.1953,\ P(8) = 0.4297,\ P(9) = 0.2930,\ P(10) = 0.0625.$$

We kunnen nu ook de verwachtingswaarde van de verwerkingstijd door netwerk I bepalen, dit is gewoon de verwachtingswaarde van de kansverdeling P, dus

$$\sum_{t=6}^{10} t \cdot P(t) \approx 8.18.$$

Het is geen verrassing dat dit hoger ligt dan de verwachtingswaarde voor beide deelpaden van de en-splitsing, want de rij $A_3 - A_4$ heeft inderdaad verwachtingswaarde 8, maar de gevallen waar de rij $A_1 - A_2$ een langere verwerkingstijd dan 8 hebben veroorzaken een hogere verwachtingswaarde voor het netwerk.

Het netwerk II kunnen we op een soortgelijke manier analyseren. Voor de verdelingsfuncties $F_i(t)$ van de enkele acties A_i geldt

$$F_i(d_i - 2) = 0, \quad F_i(d_i - 1) = 0.25, \quad F_i(d_i) = 0.75, \quad F_i(d_i + 1) = 1.$$

Hieruit krijgen we voor de en-splitsing met acties A_1 en A_3 de verdelingsfunctie $F_{13}(t) = F_1(t) \cdot F_3(t)$ als

$$F_{13}(2) = 0.25 \cdot 0.25 = 0.0625, \quad F_{13}(3) = 0.75 \cdot 0.75 = 0.5625, \quad F_{13}(4) = 1$$

en dus de kansverdeling $P_{13}(t)$ met

$$P_{13}(2) = 0.0625,\ P_{13}(3) = 0.5625 - 0.0625 = 0.5,\ P_{13}(4) = 1 - 0.5625 = 0.4375.$$

Voor de en-splitsing met acties A_2 en A_4 krijgen we analoog de verdelingsfunctie $F_{24}(t) = F_2(t) \cdot F_4(t)$ als

$$\begin{aligned} &F_{24}(3) = 0.25 \cdot 0 = 0, \quad F_{24}(4) = 0.75 \cdot 0.25 = 0.1875, \\ &F_{24}(5) = 1 \cdot 0.75 = 0.75, \quad F_{24}(6) = 1 \end{aligned}$$

en de kansverdeling $P_{24}(t)$ met

$$P_{24}(4) = 0.1875,\ P_{24}(5) = 0.75 - 0.1875 = 0.5625,\ P_{24}(6) = 1 - 0.75 = 0.25.$$

De kansverdeling $P(t)$ voor de verwerkingstijd door het netwerk II krijgen we nu met behulp van het convolutieproduct van $P_{13}(t)$ en $P_{24}(t)$, dit geeft:

$$\begin{aligned} P(6) &= P_{13}(2) \cdot P_{24}(4) = 0.0625 \cdot 0.1875 = 0.0117 \\ P(7) &= P_{13}(2) \cdot P_{24}(5) + P_{13}(3) \cdot P_{24}(4) \\ &= 0.0625 \cdot 0.5625 + 0.5 \cdot 0.1875 = 0.1289 \\ P(8) &= P_{13}(2) \cdot P_{24}(6) + P_{13}(3) \cdot P_{24}(5) + P_{13}(4) \cdot P_{24}(4) \\ &= 0.0625 \cdot 0.25 + 0.5 \cdot 0.5625 + 0.4375 \cdot 0.1875 = 0.3789 \\ P(9) &= P_{13}(3) \cdot P_{24}(6) + P_{13}(4) \cdot P_{24}(5) \\ &= 0.5 \cdot 0.25 + 0.4375 \cdot 0.5625 = 0.3711 \\ P(10) &= P_{13}(4) \cdot P_{24}(6) = 0.4375 \cdot 0.25 = 0.1094 \end{aligned}$$

In dit geval vinden we als verwachtingswaarde voor de verwerkingstijd door netwerk II

$$\sum_{t=6}^{10} t \cdot P(t) \approx 8.44.$$

Het is ook hier geen verrassing dat de verwerking door netwerk II langer duurt, want we moeten in ieder geval met de acties A_2 en A_4 wachten tot dat A_1 en A_3 beide klaar zijn.

8.2 Kritieke pad analyse

Als we in een netwerk de zwakste schakel willen vinden, moeten we het pad door het netwerk vinden dat de grootste verwerkingstijd heeft. Dit noemen we een *kritiek pad*. We hebben gezien hoe we voor de verschillende elementen van een netwerk de verwerkingstijden moeten combineren, en dat moeten we bij een netwerk stapsgewijs toepassen. Voor een rij van acties is er natuurlijk geen keuze, maar bij een splitsing hoeven we in principe alleen maar naar het alternatief met de maximale verwerkingstijd te kijken.

Bij een en-splitsing is dit helder, maar bij een of-splitsing leidt dit ertoe dat we alleen maar het ergste geval bekijken, dit noemen we een *worst-case* analyse.

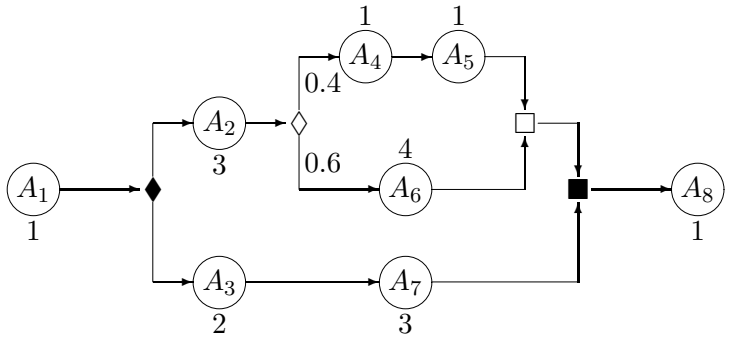
Het kan natuurlijk zo zijn dat de ergste optie van een of-splitsing slechts met een kleine kans ingegaan wordt, daarom kijkt men bij een *gewone analyse* naar alle alternatieven van een of-splitsing. Als verwerkingstijd van de of-splitsing neemt men dan de verwachtingswaarde van de verwerkingstijden van de alternatieven, maar alle acties van de of-splitsing komen in het kritieke pad terecht.

Bij een iteratie krijgen we in een worst-case analyse een oneindige verwerkingstijd, omdat we de lus willekeurig vaak kunnen doorlopen. Daarom behandelen we een iteratie altijd zo als een of-splitsing in de gewone analyse met de verwachtingswaarde van de verwerkingstijd als looptijd.

We krijgen zo de volgende strategie voor het vinden van een kritiek pad:

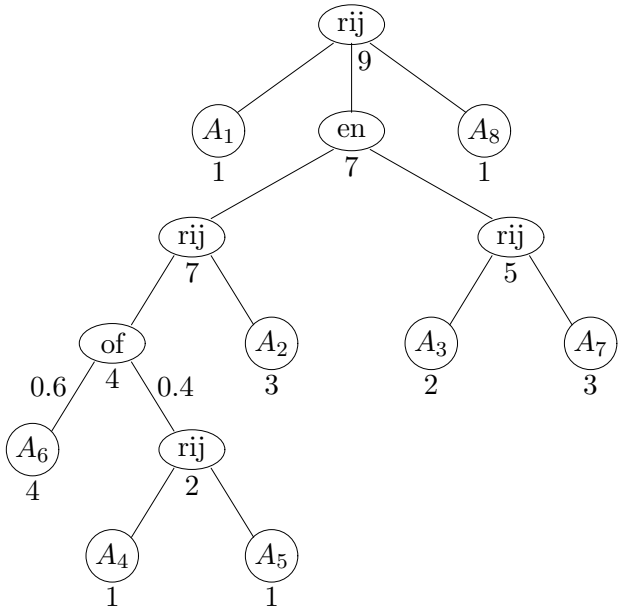
- En-splitsing: alleen maar alternatief met maximale verwerkingstijd in het kritieke pad.
- Of-splitsing:
  - worst-case analyse: alleen maar alternatief met maximale verwerkingstijd in het kritieke pad.
  - gewone analyse: alle alternatieven in het kritieke pad; verwerkingstijd is de verwachtingswaarde van de verwerkingstijden van de alternatieven.
- Iteratie: alle acties van de iteratie in het kritieke pad; verwerkingstijd is de verwachtingswaarde van de looptijd van de iteratie.

We kijken naar het volgende voorbeeld van een netwerk:



Een handige manier om zo'n netwerk te representeren is een *boom* met rijen, en-splitsingen en of-splitsingen als knopen en de acties als bladeren. De takken

vanuit een knoop zijn bij een rij de componenten van de rij (dit kunnen acties, maar ook splitsingen zijn) en bij splitsingen de alternatieven van de splitsing. In zo'n boom laten zich de verwerkingstijden makkelijk vanuit de bladeren tot de wortel bepalen. Voor het gegeven netwerk is dit in Figuur 11 te zien.



Figuur 11: Boom structuur voor een netwerk

De verwerkingstijden zijn van beneden naar boven bepaald, bij een rij is de verwerkingstijd natuurlijk de som van de tijden voor de takken, bij een en- of of-splitsing het maximum van de alternatieven (worst-case analyse).

Het is duidelijk dat de acties A_1 en A_8 in ieder pad door het netwerk bevat zijn. Het lagere pad in de en-splitsing met de acties A_3 en A_7 heeft een verwerkingstijd van $2 + 3 = 5$. Voor een pad door A_2 moeten we nog naar de of-splitsing met A_4, A_5 en A_6 kijken. Als het om een worst-case analyse gaat, hebben we hier alleen maar het alternatief met de maximale tijd nodig, dat is in dit geval A_6 met een verwerkingstijd van 4. Voor een gewone analyse krijgen we $0.6 \cdot 4 + 0.4 \cdot (1 + 1) = 3.2$ als verwerkingstijd voor de of-splitsing. In elk geval is de tijd voor dit alternatief van de en-splitsing groter dan voor het alternatief met A_3 en A_7, namelijk $3 + 4 = 7$ in de worst-case analyse en $3 + 3.2 = 6.2$ voor de gewone analyse.

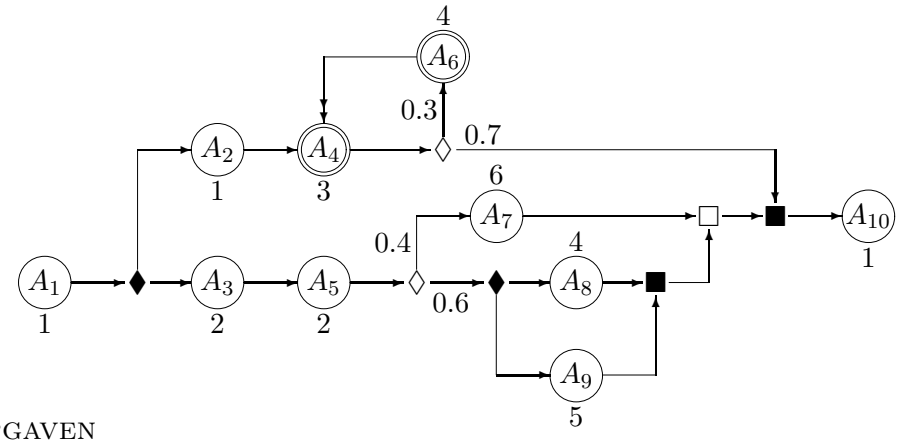
In een worst-case analyse is er dus een eenduidig kritiek pad, namelijk het pad door A_1, A_2, A_6 en A_8 en de verwerkingstijd hiervoor is $1 + 3 + 4 + 1 = 9$. Bij de gewone analyse moeten we ook de andere alternatieven in de of-splitsing in het kritieke pad meenemen, omdat een verandering in deze componenten ook de verwerkingstijd van het kritieke pad gaat veranderen. In dit geval zouden we dus ook A_4 en A_5 in het kritieke pad hebben.

Als we in dit voorbeeld één component zouden mogen verbeteren, zou A_6 de meest voor de hand liggende keuze zijn. Als we dit namelijk op een verwerkingstijd van $d_6 = 2$ kunnen reduceren, hebben we een leuk evenwicht in het

netwerk bereikt. De twee alternatieven in de of-splitsing hebben dan namelijk dezelfde verwerkingstijd en ook de verwerkingstijden voor de twee alternatieven van de en-splitsing hebben dan dezelfde tijden.

Belangrijke begrippen in deze les

- netwerk van acties
- en-splitsing, of-splitsing, iteratie
- convolutieproduct
- kritiek pad
- worst-case analyse, gewone analyse



Opgaven

34. Bepaal welke acties er in het kritieke pad van het bovenstaande netwerk liggen:

    (i) bij een worst-case analyse, d.w.z. bij een of-splitsing komt alleen maar het slechtste alternatief in het kritieke pad terecht.

    (ii) bij een gewone analyse, d.w.z. alle alternatieven van een of-splitsing komen in het kritieke pad terecht.

35. Je mag 3000€ in het netwerk investeren. Met 1000€ kan je één actie om 1 tijdseenheid versnellen. Je mag het geld over een, twee of drie acties verdelen, maar elke actie zal steeds minstens een verwerkingstijd van 1 hebben (dus aan een actie met een verwerkingstijd van 2 kan je niet meer dan 1000€ besteden).

    (i) Wat is bij een worst-case analyse de beste keuze voor je investering, d.w.z. hoe kan je de verwerkingstijd van het kritieke pad het meeste reduceren?

    (ii) Hoe zit het bij een gewone analyse?

36. De aangegeven verwerkingstijden zijn gemiddelde tijden. In werkelijkheid hebben de enkele acties de volgende kansverdelingen voor hun verwerkingstijden: Voor een actie met verwerkingstijd d is

$$P(t = d) = 0.5, \quad P(t = d - 1) = 0.25 \quad \text{en} \quad P(t = d + 1) = 0.25$$

    (d.w.z. de aangegeven verwerkingstijden gelden voor de helft van de opdrachten, 25% zijn een tijdseenheid sneller en 25% een tijdseenheid langzamer).

(i) Bepaal de kansverdeling voor de verwerkingstijd van de en-splitsing met acties A_8 en A_9.

(ii) Bepaal de kansverdeling voor de verwerkingstijd van de of-splitsing met acties A_7, A_8 en A_9.

(iii) Bepaal de kansverdeling voor de verwerkingstijd van de iteratie met acties A_4 en A_6. Hierbij mag je de iteratie na 1 iteratie afbreken.

(iv) Bereken dezelfde kansverdeling als in deel (iii), maar breek pas na 2 iteraties af. Vergelijk de kansverdeling met degene uit deel (iii).

37. Net als in de vorige opgave heeft een actie met aangegeven verwerkingstijd d de kansverdeling

$$P(t = d) = 0.5, \quad P(t = d - 1) = 0.25 \quad \text{en} \quad P(t = d + 1) = 0.25$$

voor de verwerkingstijden. De iteratie mag weer na 1 iteratie afgebroken worden.

(i) Bepaal de kansverdeling voor de verwerkingstijd van het netwerk.

(ii) In welke tijd kan je 90% van de opdrachten afhandelen?

(iii) Bereken ook de verwachtingswaarde van de verwerkingstijd van het netwerk.