Voortgezette Kansrekening

voor het 2e jaar wiskunde

docent: Hans Maassen

September 2007

Onderwijsinstituut voor Wiskunde, Natuur- en Sterrenkunde
Radboud Universiteit Nijmegen

Inhoudsopgave

Hoofdstuk 8

Rekenen met voorwaardelijke kansen

8.1 Een nuttige eigenschap. We gaan het begrip voorwaardelijke kans, dat we aan het begin van hoofdstuk 3 hebben ingevoerd, aan een nader onderzoek onderwerpen. De fundamentele eigenschap van de voorwaardelijke kans is:

(a) $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B|A)$. Zie paragraaf 3.2.

We breiden deze eigenschap nu uit:

(b) $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A) \cdot \mathbb{P}(B|A) \cdot \mathbb{P}(C|A \cap B)$

$$\text{Bewijs: } \begin{aligned} \mathbb{P}(A \cap B \cap C) &= \mathbb{P}((A \cap B) \cap C) = \\ &= \mathbb{P}(A \cap B) \cdot \mathbb{P}(C|A \cap B) = \\ &= \mathbb{P}(A) \cdot \mathbb{P}(B|A) \cdot \mathbb{P}(C|A \cap B). \ \square \end{aligned}$$

Met behulp van volledige inductie bewijs je eenvoudig dat

(c) $\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_3|A_1 \cap A_2) \cdots \mathbb{P}(A_n|A_1 \cap \cdots \cap A_{n-1})$.

Toepassing. Men neemt blindelings 3 kaarten uit het kaartspel. De kans dat het drie ♠ kaarten zijn, is $\frac{13}{52} \cdot \frac{12}{51} \cdot \frac{11}{50}$. Hierin stelt bijv. de factor $\frac{11}{50}$ de voorwaardelijke kans voor dat de 3e kaart ♠ is als de beide voorgaande ook ♠ kaarten waren.
In hoofdstuk 1 hebben we de kans op 3 ♠ uitgerekend volgens $\frac{\binom{13}{3}}{\binom{52}{3}}$. Dit leidt tot dezelfde uitkomst. In paragraaf 8.3 gaan we hier verder op in.

8.2 Als $\{A_1, A_2, \ldots\}$ een partitie (van Ω) is, dan is

$$\boxed{\mathbb{P}(B) = \sum_i \mathbb{P}(A_i \cap B).}$$

en dus is ook

$$\boxed{\mathbb{P}(B) = \sum_i \mathbb{P}(A_i)\mathbb{P}(B|A_i).}$$

Vergeet in de tweede formule de factor $\mathbb{P}(A_i)$ niet! Vooral deze tweede formule is van groot praktisch nut om vraagstukken op te lossen.

Voorbeeld. 0,01 % van de bevolking van een land lijdt aan TBC. Een test geeft in 90 % van de gevallen waarin inderdaad sprake is van TBC, een positieve uitslag. Heeft een persoon geen TBC, dan is er toch nog 1 % kans dat de test positief uitvalt. Wat is de kans op een positieve uitslag?

Voorbeeld. Op de hele wereld zijn er maar twee fabrieken die gorgels maken. 20 % van de gorgels die van fabriek I komen en 5 % van die van fabriek II zijn defect. Fabriek I produceert elke week twee keer zoveel gorgels als fabriek II.

Wat is de kans dat een gorgel, willekeurig gekozen, niet defect is?

Dat hangt natuurlijk af van de fabriek, waarin hij is gemaakt. Als B de gebeurtenis [niet defect] is, en A de gebeurtenis [gemaakt in fabriek I], dan is

$$\begin{aligned}
\mathbb{P}(B) &= \mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(A^c)\mathbb{P}(B|A^c) \\
&= \tfrac{2}{3} \cdot \tfrac{4}{5} + \tfrac{1}{3} \cdot \tfrac{19}{20} \\
&= \tfrac{51}{60}.
\end{aligned}$$

8.3 Kaarten uit een kaartspel.

Uit een spel kaarten worden 2 kaarten getrokken, zonder teruglegging. We willen de kans op 2 ♡ berekenen. Dat kan op twee manieren:

(a) We gebruiken de kansdefinitie van Laplace ofwel de hypergeometrische verdeling:

$$\begin{aligned}
\mathbb{P}[\heartsuit\heartsuit] &= \frac{\text{Aantal paren } \heartsuit}{\text{Totaal aantal paren kaarten}} \\
&= \frac{\binom{13}{2}}{\binom{52}{2}} \\
&= \tfrac{13\cdot 12}{2} \cdot \tfrac{2}{52\cdot 51} \\
&= \tfrac{13}{52} \cdot \tfrac{12}{51}.
\end{aligned}$$

(b) We gebruiken voorwaardelijke kansen:

$$\begin{aligned}\mathbb{P}[\heartsuit\heartsuit] &= \mathbb{P}[\text{1e kaart } \heartsuit] \cdot \mathbb{P}(\text{2e kaart } \heartsuit | \text{1e kaart } \heartsuit) \\ &= \tfrac{13}{52} \cdot \tfrac{12}{51}.\end{aligned}$$

Methode (a) gaat er van uit dat elk paar kaarten dezelfde kans heeft om getrokken te worden. Methode (b) heeft als uitgangspunt, dat alle nog niet getrokken kaarten dezelfde kans hebben om als eerstvolgende getrokken te worden.
De berekening hierboven laat zien, dat (in dit geval) methode (a) en (b) hetzelfde resultaat hebben!
Dit geldt meer algemeen:
Als we uit een vaas met N knikkers, waarvan er r rood en w wit zijn ($r+w=N$), n knikkers nemen ($n \leq N$), en de kans op k rode ($k \leq n$ en $k \leq r$) willen berekenen, dan is die kans gelijk aan

$$\begin{aligned}&\frac{\binom{r}{k} \cdot \binom{w}{n-k}}{\binom{N}{n}} \\ = \; & \frac{\overbrace{r \cdot (r-1) \cdots (r-k+1)}^{k \text{ factoren}}}{k!} \times \frac{\overbrace{w \cdot (w-1) \cdots (w-n+k+1)}^{n-k \text{ factoren}}}{(n-k)!} \times \frac{n!}{\underbrace{N \cdot (N-1) \cdots (N-n+1)}_{n \text{ factoren}}} \\ = \; & \binom{n}{k} \times \frac{r}{N} \cdot \frac{r-1}{N-1} \cdots \frac{r-k+1}{N-k+1} \times \frac{w}{N-k} \cdots \frac{w-n+k+1}{N-n+1}.\end{aligned}$$

De aanpak met behulp van voorwaardelijke kansen levert nog een extra gegeven op. Kijk maar in het voorbeeld: [1e kaart $\heartsuit$] en [2e kaart $\heartsuit$] zijn afhankelijk, want $\mathbb{P}$[2e kaart $\heartsuit$|1e kaart $\heartsuit$] $\neq$ $\mathbb{P}$[2e kaart $\heartsuit$| 1e kaart geen $\heartsuit$]. Het resultaat van de eerste kaart beïnvloedt het resultaat van de tweede kaart.

8.4 De formule van Bayes

Stel er doen zich een aantal mogelijkheden voor, beschreven door de gebeurtenissen $A_1, A_2, \ldots$ Dus: $\{A_1, A_2, \ldots\}$ is een partitie van Ω. Neem aan: we kennen de kansen $\mathbb{P}(A_i)$ op elk van deze mogelijkheden; we noemen $\mathbb{P}(A_i)$ de "a priori"-kans op A_i.
Neem aan dat we ook de voorwaardelijke kansen $\mathbb{P}(B|A_i)$ op een zekere gebeurtenis B kennen.
Nu wordt ons verteld, dat B inderdaad optreedt. Door dit gegeven veranderen de kansen op de A_i's in de zogenaamde "a posteriori" kansen $\mathbb{P}(A_i|B)$, die te berekenen zijn met behulp van de **formule van Bayes**:

$$\boxed{\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A_i)\cdot\mathbb{P}(B|A_i)}{\sum_j \mathbb{P}(A_j)\cdot\mathbb{P}(B|A_j)}}$$

De formule van Bayes vertelt je dus, hoe kansen veranderen door het verkrijgen van extra informatie.

Voorbeeld. We keren terug naar het TBC-voorbeeld uit paragraaf 8.2. Stel je voor: Je ondergaat de desbetreffende test. Vóórdat de uitslag bekend is, is de kans op TBC 0,01 %. Nu blijkt de uitslag positief te zijn. Bereken met behulp van de formule van Bayes de (a posteriori) kans, dat je inderdaad drager van TBC bent.

Toepassing. Drie collegezalen zitten boordevol studenten. Student S bevindt zich in één van de drie (a priori: kans $\frac{1}{3}$ voor elke zaal). De voorwaardelijke kans dat we S zien bij het vluchtig doorkijken van zaal i , gegeven dat S in zaal i is, noemen we α_i, voor $i = 1,\ 2,\ 3$ ($\alpha_i < 1$ is toegestaan). Tijdens een vluchtig bezoek aan zaal 1 treffen we S niet aan. Wat is nù de kans dat S zich toch in zaal 1 bevindt?

Oplossing. Zij A_i de mogelijkheid [S is in zaal i], $i = 1,\ 2,\ 3$, en B het feitelijk gegeven [we zien S niet tijdens een vluchtig bezoek aan zaal 1]. Gevraagd wordt: $\mathbb{P}(A_1|B)$. De formule van Bayes geeft (controleer!)

$$\mathbb{P}(A_1|B) = \frac{\mathbb{P}(A_1 \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A_1)\cdot\mathbb{P}(B|A_1)}{\sum_{i=1}^{3} \mathbb{P}(A_i)\cdot\mathbb{P}(B|A_i)} = \frac{1-\alpha_1}{3-\alpha_1}.$$

8.5 De eerste zes.

Opgave 9 van hoofdstuk 2 werd door Huygens als volgt opgelost. Laat q de kans zijn dat Jan wint. Dit kan direct bij de eerste worp gebeuren, of later. We definiëren:

$$\begin{aligned} A_i &= [i^{\mathrm{e}} \text{ worp levert een zes op}],\ i = 1,2,3; \\ B &= [\text{Jan wint}], \end{aligned}$$

en vinden:

$$\begin{aligned} q &= \mathbb{P}(B) = \mathbb{P}(A_1 \cap B) + \mathbb{P}(A_1^c \cap B) = \mathbb{P}(A_1) + \mathbb{P}(A_1^c \cap A_2^c \cap B) = \\ &= \mathbb{P}(A_1) + \mathbb{P}(A_1^c \cap A_2^c)\cdot\mathbb{P}(B|A_1^c \cap A_2^c) = \\ &= \mathbb{P}(A_1) + \mathbb{P}(A_1^c)\cdot\mathbb{P}(A_2^c)\cdot\mathbb{P}(B). \end{aligned}$$

Toelichting: de voorwaardelijke kans $\mathbb{P}(B|A_1^c \cap A_2^c)$ is gewoon $\mathbb{P}(B)$, want onder de gegeven voorwaarde (1^{e} en 2^{e} worp geen zes) zijn we weer terug in de begintoestand. Dus $q = \frac{1}{6} + \frac{5}{6}\cdot\frac{5}{6}\cdot q$. Hieruit volgt: $q = \frac{6}{11}$.

Opmerking. Op dezelfde manier kunnen we aantonen dat de kans dat ooit een zes optreedt in een rij worpen met een dobbelsteen, 1 is. Stel deze kans nl. q. Dan geldt: $q = \frac{1}{6} + \frac{5}{6}q$. Hieruit volgt $q = 1$.

8.6 Het machine-voorbeeld.

In paragraaf 1.9 rekenden we de kans uit dat het gevonden slijtage-resultaat optreedt onder de hypothese dat de kwaliteit van beide merken hetzelfde is (dus dat kwaliteit en merk onafhankelijk zijn). Ook deze kans kunnen we opvatten als een voorwaardelijke kans. Immers zij X het aantal A-machines en Y het aantal B-machines dat na twee jaar versleten is.

Dan is $X \sim \text{Bin}(4, p_1)$ en $Y \sim \text{Bin}(6, p_2)$ voor zekere succeskansen p_1 en p_2. X en Y zijn onafhankelijk: het al of niet kapotgaan van de ene machine staat los van dat van de andere. Het gaat om de voorwaardelijke kans

$$\mathbb{P}(X \leq 1 | X + Y = 5).$$

Nemen we nu aan dat de gestelde hypothese geldt, dus dat $p_2 = p_1$, dan is $X + Y \sim \text{Bin}(10, p_1)$ (opgave 11 van hoofdstuk 3). Nu volgt:

$$\begin{aligned}
\mathbb{P}(X = 1 | X + Y = 5) &= \frac{\mathbb{P}[X = 1, Y = 4]}{\mathbb{P}[X + Y = 5]} = \\
&= \frac{\mathbb{P}[X = 1] \cdot \mathbb{P}[Y = 4]}{\mathbb{P}[X + Y = 5]} = \\
&= \frac{\binom{4}{1} p_1^1 (1 - p_1)^3 \cdot \binom{6}{4} p_1^4 (1 - p_1)^2}{\binom{10}{5} p_1^5 (1 - p_1)^5} = \\
&= \frac{\binom{4}{1}\binom{6}{4}}{\binom{10}{5}} = \frac{5}{21}
\end{aligned}$$

en dat is precies de (in hoofdstuk 1 geponeerde) hypergeometrische kans. Net zo bereken je $\mathbb{P}(X = 0 | X + Y = 5)$.

Algemener geldt: als $X \sim \text{Bin}(n, p)$, $Y \sim \text{Bin}(m, p)$, en X en Y zijn onafhankelijk, dan is

$$\mathbb{P}[X = k | X + Y = j] = \frac{\binom{n}{k} \cdot \binom{m}{j-k}}{\binom{n+m}{j}}.$$

Opgaven

1. Het weer in een weekeinde. Zij A = [mooi weer op zaterdag], B = [mooi weer op zondag]. Stel dat $\mathbb{P}(A) = \mathbb{P}(B) = 0,4$ en $\mathbb{P}(A \cap B) = 0,3$. Bepaal $\mathbb{P}(B|A)$, $\mathbb{P}(B^c|A)$, $\mathbb{P}(B|A^c)$ en $\mathbb{P}(B^c|A^c)$.

2. In opgave 1 blijkt $\mathbb{P}(B|A) > \mathbb{P}(B)$ (er is een positieve afhankelijkheid: als het eenmaal mooi weer is, dan is het waarschijnlijker dat het de volgende dag ook nog zo is). Laat zien dat in het algemeen uit $\mathbb{P}(B|A) > \mathbb{P}(B)$ volgt dat ook $\mathbb{P}(B^c|A^c) > \mathbb{P}(B^c)$. (neem aan: $0 < \mathbb{P}(A) < 1$).

3. In een zak zitten drie kaarten. Eén ervan is aan beide kanten wit. De tweede is aan beide kanten rood en de derde heeft een witte en een rode kant. Men pakt blindelings een kaart en legt die op tafel zó dat we alleen de bovenkant zien. Die blijkt wit te zijn. Hoe groot is de kans dat de onderkant ook wit is? Aanwijzing: het antwoord $\frac{1}{2}$ is niet goed.

4. Nog een keer het probleem van de eerste zes. Jan en Piet gooien om de beurt een dobbelsteen. Jan begint. Winnaar is degene die als eerste een zes gooit. We zeggen dat de beslissing in de eerste ronde valt als bij de eerste of tweede worp een zes verschijnt. De beslissing valt in de tweede ronde als de eerste zes pas bij de derde of vierde worp optreedt, enzovoort.
   Zij A_k = [beslissing in k^{e} ronde], B = [Jan wint].
   Bepaal eerst $\mathbb{P}(B|A_k)$ en bereken dan $\mathbb{P}(B)$ m.b.v. de tweede formule van paragraaf 8.2. Merk op dat het in dit geval niet nodig is de afzonderlijke kansen $\mathbb{P}(A_k)$ te kennen. Het is genoeg dat we weten dat $\sum_k \mathbb{P}(A_k) = 1$.

5. We doen een rij onafhankelijke proeven waarbij steeds de kans op succes p is. Stel p_n de kans op een oneven aantal successen in de eerste n proeven; $q_n = 1 - p_n$. Bewijs:

   (a) $p_{n+1} = q \cdot p_n + p \cdot q_n$;

   (b) $p_n = \frac{1}{2} - \frac{1}{2}(q-p)^n$.

6. In een rij onafhankelijke spelen kan men bij ieder spel een gulden winnen of een gulden verliezen, beide met kans $\frac{1}{2}$. Laat, voor $n \geq 1$, p_n de kans zijn dat we op den duur al ons geld verliezen als we beginnen met n gulden op zak. Bewijs:

   (a) $p_n = \frac{1}{2}p_{n+1} + \frac{1}{2}p_{n-1}$ $(n \geq 2)$; $p_1 = \frac{1}{2}p_2 + \frac{1}{2}$.

(b) $p_n = 1$ voor iedere n.

7. Hoe moeten we 10 witte en 10 zwarte knikkers over 2 vazen verdelen als we willen dat, wanneer we op goed geluk een vaas pakken (kans $\frac{1}{2}$ voor beide vazen) en uit die vaas blindelings een knikker, de kans dat die knikker wit is, zo groot mogelijk is?

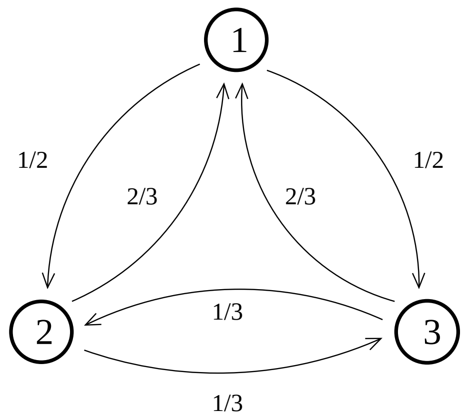
8. Doos 1 bevat 1 witte en 9 zwarte knikkers, doos 2 bevat 3 witte en 3 zwarte knikkers, doos 3 bevat 4 witte en 0 zwarte knikkers.
We kiezen op goed geluk één van de drie dozen, en uit die doos nemen we (zonder terugleggen) 2 knikkers. Hoe groot zijn de voorwaardelijke kansen dat we doos i gekozen hebben als die twee knikkers beide wit bleken te zijn?

9. De scheidsrechter gooit een dobbelsteen, resultaat Z ogen. De spelers A en B krijgen dan ieder Z munten en gooien die op. A krijgt daarbij X keer kop en B krijgt Y keer kop. A wint als $X > Y$, B wint als $Y > X$.

   (a) Bepaal de kans op gelijkspel.

   (b) Bepaal $\mathbb{P}[X = 6]$ en $\mathbb{P}[X = 6 \text{ en } Y = 6]$. Zijn X en Y onafhankelijk?

10. Een vlo springt heen en weer over de hoekpunten van een driehoek, die we toestanden 1, 2 en 3 zullen noemen. Uit toestand 1 springt hij met kans $\frac{1}{2}$ naar 2 en met kans $\frac{1}{2}$ naar 3; uit 2 met kans $\frac{2}{3}$ naar 1 en met kans $\frac{1}{3}$ naar 3; uit 3 met kans $\frac{2}{3}$ naar 1 en met kans $\frac{1}{3}$ naar 2. Dit zijn de zogenaamde overgangskansen. Laat X_n de toestand na n sprongen zijn. Als de begintoestand $X_0 = 1$, bepaal



dan de voorwaardelijke kansen dat de vlo na 2 sprongen in toestand i is (i=1, 2, 3) (Zoiets noemt men een Markovketen; zie ook hoofdstuk 13).

11. Als het in Nijmegen een bepaalde dag regent, dan is het de volgende dag met kans $\frac{1}{4}$ de gehele dag droog, terwijl het met kans $\frac{3}{4}$ ook de volgende dag regent. Als het een bepaalde dag droog is, dan is het ook de volgende dag droog met kans $\frac{3}{5}$, terwijl het met kans $\frac{2}{5}$ de volgende dag regent. Verder nemen we aan dat het geen enkele dag met zekerheid droog is; evenmin regent het met zekerheid een bepaalde dag.

    (a) Stel dat het vandaag regent. Wat is onder dit gegeven de voorwaardelijke kans dat het overmorgen droog is?

    (b) Zijn de gebeurtenissen [donderdag 16 april 2020 regent het] en [vrijdag 17 april 2020 regent het] onafhankelijk?

    (c) Stel nu dat de kans, dat het een bepaalde dag regent, voor alle dagen hetzelfde is. Bereken die kans.

12. In een stad rijden 70 % gele taxi's en 30 % groene taxi's. Op een donkere avond veroorzaakt een taxi een aanrijding en rijdt door. Een getuige verklaart dat een groene taxi de aanrijding veroorzaakte. In een proces, dat tegen het groene taxibedrijf wordt aangespannen, gelast de rechter na te gaan hoe goed de getuige bij matig licht een groene van een gele taxi kan onderscheiden. Een experiment wijst uit dat de getuige in 80 % van de gevallen de juiste kleur noemt. Bereken de voorwaardelijke kans dat inderdaad een groene taxi de aanrijding heeft veroorzaakt (aangenomen mag worden dat taxi's en chauffeurs van beide bedrijven gemiddeld even betrouwbaar zijn).

13. (Knock-out-systeem) Gegeven 2^n spelers, allen even sterk. Ieder van hen speelt tegen een van de anderen. Vervolgens spelen de 2^{n-1} winnaars twee-aan-twee tegen elkaar, enzovoorts, tot uiteindelijk 2 spelers overblijven en tegen elkaar spelen. Via loting wordt telkens vastgesteld, welk tweetal spelers tegen elkaar speelt. In elke partij hebben beide spelers kans $\frac{1}{2}$ op winst. A en B zijn twee van de deelnemers.

    (a) Bepaal de kans dat A in precies i partijen meespeelt ($i = 1, 2, \ldots, n$).

    (b) Bepaal de kans p_n dat A en B elkaar ooit treffen. Aanwijzing: leid eerst een verband af tussen p_n en p_{n-1}. Bereken dan zonodig p_1, p_2, $p_3, \ldots$

14. Je doet mee aan een TV-quiz en bent er, wonder boven wonder, in geslaagd om de finale te bereiken. De presentator houdt je nu drie dozen voor. In één van deze dozen zit een fantastische hoofdprijs; in de twee andere dozen zit een zak

aardappelen. Helaas kun je niet zien in welke doos de hoofdprijs zit. Je mag nu kiezen welke van de drie dozen je wilt hebben. Nadat je je keuze gemaakt hebt, maakt de presentator (die weet waar de hoofdprijs zit) één van de dozen die je *niet* gekozen hebt open, en haalt er een zak aardappelen uit. Hij vraagt vervolgens, of je nog van keuze wilt veranderen. Wat biedt de meeste kans op de hoofdprijs, bij je aanvankelijke keuze blijven, of wisselen? Of maakt het niet uit?

Hoofdstuk 9

Kansverdelingen in $\mathbb{R}^2$

9.1 We laten het toeval een punt (X,Y) in het vlak aanwijzen; (X,Y) is dus een stochastische vector in $\mathbb{R}^2$. We zeggen dat we de kansverdeling van (X,Y) kennen als de kans dat (X,Y) in A valt bekend is voor genoeg deelverzamelingen A van $\mathbb{R}^2$, bijvoorbeeld voor alle rechthoeken $(a,b] \times (c,d] = \{(x,y) : a < x \leq b \text{ en } c < x \leq d\}$.

Voorbeelden:

(a) X en Y zijn de aantallen ogen van twee worpen met een dobbelsteen. Het punt (X,Y) ligt dan in het 6×6-rooster

$$\{(i,j) \in \mathbb{N} \times \mathbb{N} : 1 \leq i \leq 6 \text{ en } 1 \leq j \leq 6\}.$$

(b) Een worp met een dobbelsteen geeft X ogen. We gooien daarna X munten op en krijgen daarbij Y keer kop. Welke mogelijkheden zijn er voor het punt (X,Y)?

(c) Men werpt pijltjes op een schijf waarop een assenstelsel is aangebracht; de roos is $(0,0)$. De uitkomst van een worp is dan een punt (X,Y) in $\mathbb{R}^2$.

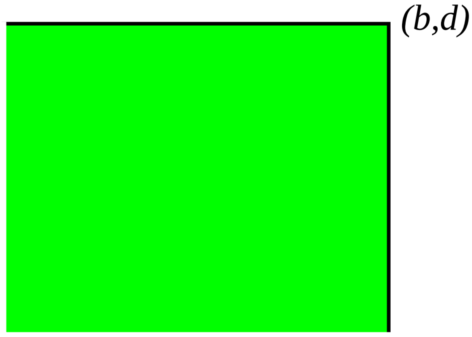
(d) X en Y zijn de aankomsttijden van de twee cafébezoekers uit opgave 10 van hoofdstuk 2. (X,Y) is een punt in het vierkant $[0,1] \times [0,1]$.

Verdelingsfuncties

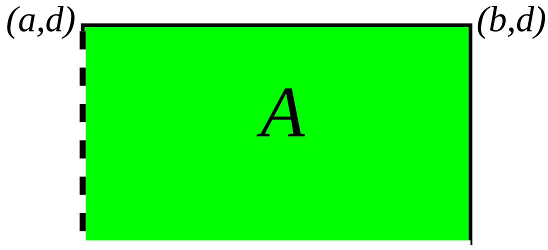
9.2 De kansverdeling van een stochastische vector (X,Y) over het vlak ($\mathbb{R}^2$) is weer te beschrijven door middel van de *(simultane) verdelingsfunctie* F *van* (X,Y), notatie ook wel $F_{X,Y}$, die gedefinieerd wordt door

$$\boxed{F(x,y) = \mathbb{P}[X \leq x \text{ en } Y \leq y].}$$

Bij ieder punt (b, d) in $\mathbb{R}^2$ geeft $F(b, d)$ de kans dat (X, Y) terechtkomt in de "zuid-west-sector" met hoekpunt (b, d). Men kan laten zien dat F de kansverdeling geheel bepaalt. Anders gezegd: Als we bovenstaande kansen eenmaal kennen, kunnen we voor ontzaglijk veel deelverzamelingen A van $\mathbb{R}^2$ de kans berekenen dat (X, Y) in A terechtkomt.

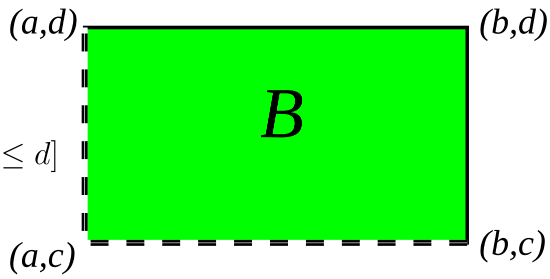


Om te beginnen voor het gebied A hiernaast, dat we immers kunnen beschouwen als het verschil van twee zuid-west-sectoren:



$$\begin{aligned} \mathbb{P}[(X, Y) \in A] &= \mathbb{P}[a < X \leq b \text{ en } Y \leq d] \\ &= F(b, d) - F(a, d) \end{aligned}$$

Maar dan kunnen we ook wel de kans berekenen dat (X, Y) in nevenstaande rechthoek B terechtkomt:



$$\begin{aligned} \mathbb{P}[(X, Y) \in B] &= \mathbb{P}[a < X \leq b \text{ en } c < Y \leq d] \\ &= (F(b, d) - F(a, d)) \\ &\quad - (F(b, c) - F(a, c)) \\ = F(b, d) &- F(a, d) - F(b, c) + F(a, c) \end{aligned}$$

Zo voortgaande kun je laten zien dat je $\mathbb{P}[(X, Y) \in C]$ kunt berekenen voor elk gebied C dat opgebouwd is uit rechthoeken (eindig of aftelbaar veel); bv. voor de verzameling $C = \{(x, y) : x + y \leq 3\}$. De praktische uitvoerbaarheid is echter niet zo groot; e zullen daarvoor aanstonds dan ook wat technieken leren.
Het gaat er nu maar even om, op te merken dat kennis van F, dus kennis van $\mathbb{P}[(X, Y)] \in A]$ voor alle zuid-west-sectoren A, voldoende is om $\mathbb{P}[(X, Y) \in C]$ voor allerlei gebieden C te berekenen.

9.3 Verband tussen simultane ($F_{X,Y}$) en marginale (F_X, F_Y) verdeling.

(a) Berekening van F_X en F_Y uit $F_{X,Y}$:

$$\boxed{\begin{aligned}\lim_{x\to\infty} F_{X,Y}(x,y) &= F_Y(y)\\ \lim_{y\to\infty} F_{X,Y}(x,y) &= F_X(x)\end{aligned}}$$

BEWIJS (voor de eerste bewering; de tweede gaat natuurlijk net zo):

$$\begin{aligned}\lim_{x\to\infty} F_{X,Y}(x,y) &= \lim_{n\to\infty} F_{X,Y}(n,y)\\ &= \lim_{n\to\infty} \mathbb{P}[X \le n \text{ en } Y \le y]\\ &\overset{2.5(c)}{=} \mathbb{P}\left(\bigcup_{n=1}^{\infty}[X \le n \text{ en } Y \le y]\right)\\ &= \mathbb{P}[X < \infty \text{ en } Y \le y]\\ &= \mathbb{P}[Y \le y]\\ &= F_Y(y) \quad \square\end{aligned}$$

(b) Als we F_X en F_Y kennen, kunnen we $F_{X,Y}$ nog niet uitrekenen; als we niet weten of X en Y onafhankelijk zijn, ligt $F_{X,Y}$ niet vast. Anders gezegd: bij gegeven F_X en F_Y zijn dan verschillende $F_{X,Y}$'en mogelijk.
Wel zien we onmiddellijk m.b.v. 6.10:
Als X en Y *onafhankelijk* zijn, kunnen we $F_{X,Y}$ berekenen uit

$$\boxed{F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y).}$$

Zelfs geldt equivalentie: X en Y zijn onafhankelijk $\Leftrightarrow$ $F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y)$ voor alle x, y.

Continue verdelingen in $\mathbb{R}^2$.

9.4 Definitie. We zeggen dat de stochastische vector (X,Y) *continu verdeeld is met dichtheid* $f = f_{X,Y}$ als voor iedere $(x,y) \in \mathbb{R}^2$:

$$\boxed{F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dvdu.}$$

f is hier een functie van $\mathbb{R}^2$ naar $[0, \infty)$.
Hierbij merken we op:

(a) Daar de verdelingsfunctie F de kansverdeling vastlegt, kan hieruit afgeleid worden dat voor ieder fatsoenlijk gebied A in $\mathbb{R}^2$ geldt:

$$\boxed{\mathbb{P}[(X, Y) \in A] = \iint_A f(u, v) dv du.}$$

Fatsoenlijke gebieden zijn bv. rechthoeken, halfvlakken, cirkelschijven. Door voor A de zuid-west-sector $A = (-\infty, x] \times (-\infty, y]$ te kiezen, krijg je precies de formule uit de definitie terug.

(b) Integralen zoals in (a) worden altijd berekend door herhaalde integratie: eerst over v, dan over u, of omgekeerd. De volgorde doet er daarbij niet toe, omdat de te integreren functie postief is. Dit op grond van de stelling van Fubini (zie Analyse).

(c) Natuurlijk is $\mathbb{P}[(X, Y) \in \mathbb{R}^2] = 1$, dus: $\iint_{\mathbb{R}^2} f = 1$. En net als in paragraaf 6.4 kiezen we f zó dat $f \geq 0$.

9.5 Verband tussen simultane ($f_{X,Y}$) en marginale (f_X, f_Y) dichtheid.

(a) Als (X, Y) continu verdeeld is met dichtheid $f_{X,Y}$, dan is X continu verdeeld met een dichtheid f_X gegeven door

$$\boxed{f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy.}$$

Evenzo is Y continu verdeeld met dichtheid

$$\boxed{f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx.}$$

Bewijs. Zij $A = (-\infty, a] \times \mathbb{R}$.

$$\begin{aligned} F_X(a) &= \mathbb{P}[X \leq a] \\ &= \mathbb{P}[(X, Y) \in A] \\ &= \iint_A f_{X,Y}(x, y) dy dx \\ &= \int_{-\infty}^{a} \left(\int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \right) dx, \end{aligned}$$

dus de afbeelding

$$x \mapsto \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

is een dichtheid van X. $\square$

(b) Als X continu verdeeld is met dichtheid f_X en Y continu verdeeld is met dichtheid f_Y, en X en Y zijn *onafhankelijk*, dan is (X,Y) continu verdeeld met dichtheid $f_{X,Y}$ gegeven door

$$\boxed{f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y).}$$

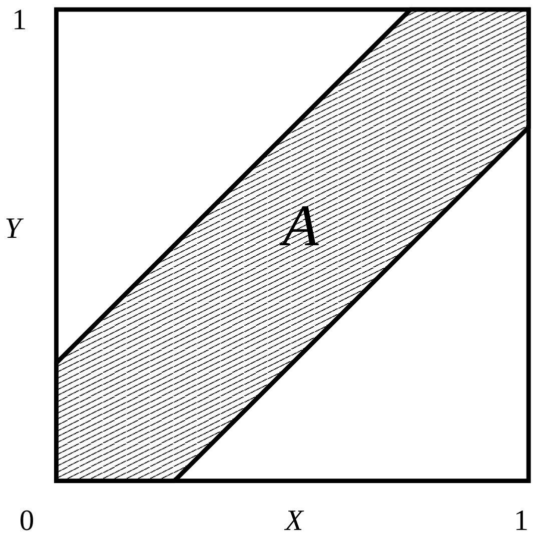
Bewijs.

$$\begin{aligned} F_{X,Y}(x,y) &\overset{9.3(b)}{=} F_X(x) \cdot F_Y(y) \\ &= \int_{-\infty}^{x} f_X(u)du \cdot \int_{-\infty}^{y} f_Y(v)dv \\ &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_X(u) \cdot f_Y(v)dvdu. \end{aligned}$$

(De laatste stap lijkt misschien moeilijk, maar het enige dat er gebeurt is,dat het getal $\int_{-\infty}^{y} f_Y(v)dv$ binnen de integraal $\int_{-\infty}^{x} f_X(u)du$ wordt gezet, m.a.w. we gebruiken dat $c \cdot \int_a^b f(x)dx = \int_a^b c \cdot f(x)dx$.)
We hebben nu dus $F_{X,Y}(x,y)$ geschreven in de vorm van paragraaf 9.4. De functie $(u,v) \mapsto f_X(u) \cdot f_Y(v)$ kunnen we dan dus nemen als dichtheid van (X,Y). $\square$
Opmerking. weer geldt zelfs equivalentie: X en Y zijn onafhankelijk $\Leftrightarrow$ er zijn dichtheden f_X, f_Y, $f_{X,Y}$ van X, Y, resp. (X,Y) met: $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$ voor alle x,y.

(c) Als in (b) X en Y niet onafhankelijk zijn, hoeft (X,Y) niet eens continu verdeeld te zijn met een dichtheid $f_{X,Y}$. Bijvoorbeeld als $X = Y$: dan ligt alle kansmassa op de lijn $y = x$, en de kansverdeling van (X,Y) kan niet beschreven worden d.m.v. een dichtheid.

9.6 Voorbeeld: de uniforme verdeling. Stel $X \sim \mathrm{Un}(0,1)$ en $Y \sim \mathrm{Un}(0,1)$, en X en Y onafhankelijk.
Bereken $\mathbb{P}\left[|X-Y| \le \frac{1}{4}\right]$. Misschien had je het al herkend: dit is precies de opgave over het café-bezoek (hoofdstuk 2, opgave 10), alleen nu met wat minder proza geformuleerd.



We tekenen hetzelfde plaatje als destijds, en noemen het "gunstige" gebied A: Gegeven is dat $f_X = f_Y = \mathbf{1}_{(0,1)}$; wegens 9.5(b) is dan

$$f_{X,Y}(x,y) = \mathbf{1}_{(0,1)}(x) \cdot \mathbf{1}_{(0,1)}(y).$$

M.b.v. 9.4 vinden we dan dat

$$\begin{aligned} \mathbb{P}\left[|X-Y| \leq \tfrac{1}{4}\right] &= \mathbb{P}[(X,Y) \in A] \\ &= \iint_A \mathbf{1}_{(0,1)}(x) \cdot \mathbf{1}_{(0,1)}(y) dy dx. \end{aligned}$$

Formeel moeten we deze integraal uitrekenen door onderscheid te maken naar $x \in [0, \frac{1}{4}]$, $x \in (\frac{1}{4}, \frac{3}{4}]$, $x \in (\frac{3}{4}, 1]$, en de bijbehorende y-waarden te bepalen. We vinden dan

$$\int_0^{\frac{1}{4}} \left(\tfrac{1}{4} + x\right) dx + \int_{\frac{1}{4}}^{\frac{3}{4}} \tfrac{1}{2} dx + \int_{\frac{3}{4}}^{1} \left(\tfrac{5}{4} - x\right) dx = \ldots = \tfrac{7}{16}.$$

Maar het kan ook veel gemakkelijker. Uit de definitie van integratie in $\mathbb{R}^2$ (zie Analyse) volgt onmiddelijk dat

$$\iint_A 1 dy dx = \text{oppervlakte}(A) = 1 - \left(\tfrac{3}{4}\right)^2 = \tfrac{7}{16}.$$

We formuleren bovenstaande algemener:
Zij G een gebied in $\mathbb{R}^2$ met oppervlakte$(G) = a < \infty$. Als de stochastische vector (X,Y) slechts waarden aanneemt in G, zonder voorkeur voor delen binnen G, m.a.w. als $f_{X,Y} = \frac{1}{a}\mathbf{1}_G$ (men noemt (X,Y) dan uniform verdeeld over G), dan is voor elke $A \subset G$:

$$\boxed{\mathbb{P}[(X,Y) \in A] = \frac{\text{oppervlakte}(A)}{\text{oppervlakte}(G)}.}$$

In het café-voorbeeld is $G = (0,1) \times (0,1)$.

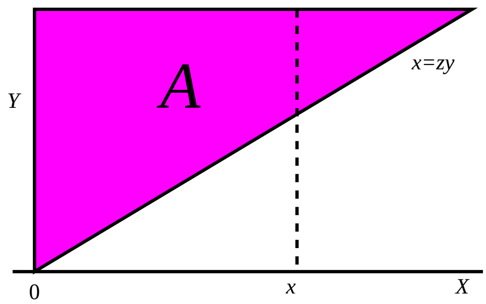
9.7 Voorbeeld met de exponentiële verdeling. Stel X en Y zijn onafhankelijke stochasten, beide exponentieel verdeeld met parameter 1. Bepaal verdelingsfunctie, dichtheidsfunctie en verwachting van $Z = \frac{X}{Y}$.

Oplossing. het is duidelijk dat $F_Z(z) = 0$ als $z \leq 0$. Zij verder $z > 0$. Omdat X en Y onafhankelijk zijn, en $f_X(x) = e^{-x}$ $(x > 0)$ en $f_Y(y) = e^{-y}$ $(y > 0)$, heeft de stochastische vector (X,Y) wegens 9.5(b) dichtheid

$$f_{X,Y}(x,y) = e^{-x}e^{-y} \qquad (x, y > 0).$$

Zo volgt, met

$$A = \{(x,y) : x, y > 0 \text{ en } x \le zy\}:$$



$$\begin{aligned}
F_Z(z) &= \mathbb{P}\left[\frac{X}{Y} \le z\right] = \mathbb{P}[(X,Y) \in A] \\
&= \iint_A e^{-x} e^{-y} dy dx \\
&= \int_0^\infty e^{-x} \left(\int_{\frac{x}{z}}^\infty e^{-y} dy \right) dx \\
&= \int_0^\infty e^{-x} e^{-\frac{x}{z}} dx = \int_0^\infty e^{-\left(1+\frac{1}{z}\right)x} dx = \frac{1}{1+\frac{1}{z}} = \frac{z}{z+1},
\end{aligned}$$

zodat

$$F_Z(z) = \begin{cases} 0 & (z < 0) \\ \frac{z}{z+1} & (z \ge 0) \end{cases} \quad \text{en } f_Z(z) = \begin{cases} 0 & (z < 0) \\ \frac{1}{(z+1)^2} & (z > 0) \end{cases}$$

Om $\mathbb{E}(Z)$ te bepalen, bereken we $\int_0^n z f_Z(z) dz$;
we vinden

$$\begin{aligned}
\int_0^n \frac{z}{(z+1)^2} dz &= \int_0^n \left(\frac{1}{z+1} - \frac{1}{(z+1)^2} \right) dz \\
&= \left[\log(z+1) + \frac{1}{z+1} \right]_{z=0}^{z=n} \\
&= \log(n+1) + \frac{1}{n+1} - 1,
\end{aligned}$$

zodat

$$\mathbb{E}(Z) = \lim_{n\to\infty} \int_0^n \frac{z}{(z+1)^2} dz = \infty.$$

9.8 Sommen van normaal verdeelde stochasten.

$$\boxed{\left.\begin{array}{r} U_1 \sim \mathrm{N}(\mu_1, \sigma_1^2) \\ U_2 \sim \mathrm{N}(\mu_2, \sigma_2^2) \\ U_1 \perp\!\!\!\perp U_2 \end{array}\right\} \Longrightarrow U_1 + U_2 \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)}$$

Bewijs: We mogen wel aannemen dat $\mu_1 = \mu_2 = 0$ (ga anders over op de stochasten $U_1 - \mu_1$ en $U_2 - \mu_2$).
De meest voor de hand liggende wijze is de convolutie te berekenen van de stochasten U_1 en U_2. Dat blijkt echter erg veel werk te zijn. Het volgende gaat veel mooier. Schrijf

$$S = U_1 + U_2$$

en

$$X = \frac{U_1}{\sigma_1} \qquad Y = \frac{U_2}{\sigma_2}.$$

Merk op dat X en Y standaard-normaal verdeeld zijn.
Wegens de onafhankelijkheid van X en Y is de simultane dichtheid $f_{X,Y}$ (zie 9.5(b)):

$$f_{X,Y}(x,y) = \varphi(x)\varphi(y) = \frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)}.$$

We zien dat $f_{X,Y}$ op elke cirkel in $\mathbb{R}^2$ met de oorsprong als middelpunt een constante functie is, m.a.w. $f_{X,Y}$ is invariant voor rotaties om de oorsprong.

Nog anders gezegd: ls het gebied A door een rotatie om de oorsprong overgaat in gebied B, dan is

$$\iint_A f_{X,Y}(x,y)dydx = \iint_B f_{X,Y}(x,y)dydx.$$

Zij nu $s \geq 0$. Dan is

$$\begin{aligned} \mathbb{P}[S \leq s] &= \mathbb{P}[\sigma_1 X + \sigma_2 Y \leq s] \\ &= \mathbb{P}[(X,Y) \in A] \\ &= \iint_A f_{X,Y}(x,y)dydx, \end{aligned}$$

waarin

$$A = \{(x,y) : \sigma_1 x + \sigma_2 y \leq s\}.$$

Door een rotatie over de hoek ϑ (zie plaatje) in de richting van de wijzers van de klok gaat A over in het halfvlak

$$B = \left\{(x,y) : \sqrt{\sigma_1^2 + \sigma_2^2}\, x \leq s\right\}$$

(ga dit na). We vinden zo:

$$\begin{aligned} \mathbb{P}[(X,Y) \in A] &= \iint_A f_{X,Y}(x,y)dydx \\ &= \iint_B f_{X,Y}(x,y)dydx \\ &= \mathbb{P}\left[(X,Y) \in B\right] \\ &= \mathbb{P}\left[X \leq \frac{s}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right] \\ &= \mathbb{P}\left[\sqrt{\sigma_1^2 + \sigma_2^2} X \leq s\right] \end{aligned}$$

Dus $S \sim \sqrt{\sigma_1^2+\sigma_2^2}X$, d.w.z. $S \sim \mathrm{N}(0, \sigma_1^2+\sigma_2^2)$. $\square$

Opmerking. Het is nogal bijzonder dat de som van twee onafhankelijke stochasten met eenzelfde verdeling, weer eenzelfde verdeling heeft. Zoals opgemerkt in paragraaf **??** krijg je door het optellen van onafhankelijke stochasten (convolutie van bijbehorende dichtheidsfuncties) verdelingen met steeds mooieren klokvormige dichtheden. De normale verdeling hééft kennelijk al deze "ideale" vorm; er valt dus niets meer te verfraaien.

Convoluties

9.9 De verdeling van $X+Y$. Vaak moeten in de stochastiek stochasten opgeteld worden: wat is de kansverdeling van $X+Y$ als de kansverdelingen van X en Y bekend zijn en X en Y onafhankelijk verondersteld worden? Het blijkt niet nodig dit telkens opnieuw uit te rekenen: in deze paragraaf leiden we een formule af die voor alle continu verdeelde stochasten gebruikt kan worden.

Concreet. Stel X en Y zijn onafhankelijke stochasten, beide continu verdeeld met dichtheden f en g. Wat is dan de verdeling van $Z = X+Y$?

Antwoord: Z is ook continu verdeeld en wel met een dichtheid h die gegeven wordt door

$$\boxed{h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)dx = \int_{-\infty}^{\infty} f(z-y)g(y)dy}$$

(beide formules leveren hetzelfde resultaat).

Men noemt h de *convolutie* van f en g, notatie: $h = f * g$.

(Vergelijk dit met het discrete geval: Als X en Y onafhankelijke geheelwaardige stochasten zijn, dan is

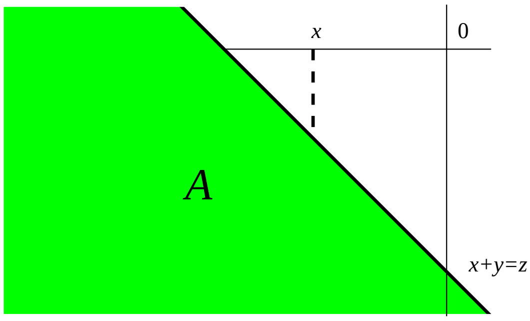
$$\mathbb{P}[X+Y=n] = \sum_{k=0}^{n} \mathbb{P}[X{=}k] \cdot \mathbb{P}[Y{=}n{-}k].)$$

Bewijs. De verdelingsfunctie H van $Z = X+Y$ wordt gegeven door

$$H(z) = \mathbb{P}[Z \le z] = \mathbb{P}[(X,Y) \in A]$$

met $A = \{(x,y) : x+y \le z\}$. Met 9.4(a) zien we:

$$\begin{aligned} H(z) &= \iint_A f_{X,Y}(x,y)dydx \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z-x} f(x)g(y)dy \right) dx. \end{aligned}$$



We nemen nu even x vast en beschouwen
de "binnenste" integraal:

$$\int_{-\infty}^{z-x} f(x)g(y)dy.$$

We vervangen (substitueren) de variabele y door de nieuwe variabele $t = x + y$, en vinden dan:

$$\int_{-\infty}^{z-x} f(x)g(y)dy = \int_{-\infty}^{z} f(x)g(t-x)dt.$$

Zo zien we:

$$\begin{aligned} H(z) &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z} f(x)g(t-x)dt \right) dx \\ &= \int_{-\infty}^{z} \left(\int_{-\infty}^{\infty} f(x)g(t-x)dx \right) dt \end{aligned}$$

en blijkt dat we de functie

$$t \mapsto \int_{-\infty}^{\infty} f(x)g(t-x)dx$$

kunnen nemen als dichtheid van Z. $\square$

9.10 Sommen van uniform verdeelde stochasten.

Bij numerieke berekeningen komen vaak problemen voor als beschreven in het begin van paragraaf 9.9 (vergelijk paragraaf 6.6, voorbeeld 3). Laat $X_1, \ldots, X_n$ de afrondingsfouten zijn in de afgeronde grootheden; $Y_n = X_1 + \ldots + X_n$ is dan de fout in de som van die grootheden.
We nemen aan dat $X_1, \ldots, X_n$ onafhankelijk zijn en allemaal uniform verdeeld over $(-\frac{1}{2}, \frac{1}{2})$.
Laat f_k de dichtheidsfunctie van Y_k zijn; dan is f_1 dus de dichtheid van zowel X_1, als $X_2, \ldots, X_n$.
Omdat $Y_{k+1} = Y_k + X_{k+1}$ is (wegens paragraaf 9.9) $f_{k+1} = f_k * f_1$, dus

$$f_{k+1}(x) = \int_{-\infty}^{\infty} f_k(t)f_1(x-t)dt = \int_{x-\frac{1}{2}}^{x+\frac{1}{2}} f_k(t)dt.$$

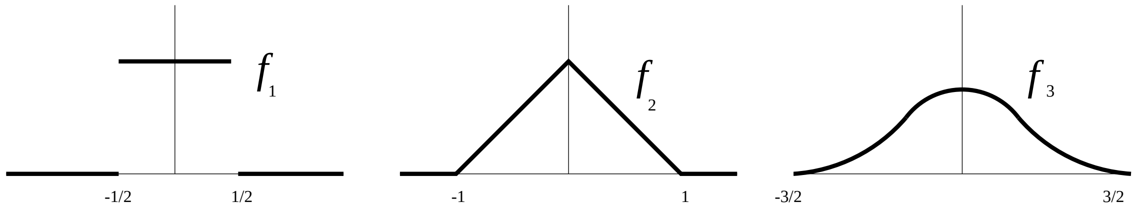
Door $k = 1$ te kiezen vinden we

$$f_2(x) = \int_{x-\frac{1}{2}}^{x+\frac{1}{2}} \mathbf{1}_{(-\frac{1}{2},\frac{1}{2})}(t)dt.$$

Voor vaste x moeten we dus t's hebben met

$$x - \tfrac{1}{2} < t < x + \tfrac{1}{2} \text{ en } -\tfrac{1}{2} < t < \tfrac{1}{2}.$$

Voor $|x| > 1$ lukt dit niet, dus dan is $f_2(x) = 0$.
Voor $-1 < x < 0$ betekent het bovenstaande: $-\frac{1}{2} < t < x + \frac{1}{2}$, zodat $f_2(x) = \int_{-\frac{1}{2}}^{x+\frac{1}{2}} 1dt = x + 1$.
Voor $0 < x < 1$ vind je $x - \frac{1}{2} < t < \frac{1}{2}$, dus $f_2(x) = \int_{x-\frac{1}{2}}^{\frac{1}{2}} 1dt = 1 - x$.
Daarmee is f_2 gevonden.
Zo doorgaande kun je f_3 uitrekenen m.b.v. f_2, f_4 m.b.v. f_3, enzovoorts; het rekenwerk wordt wel steeds omvangrijker. We besluiten met de grafieken van f_1, f_2 en f_3 en merken op dat deze gladder worden naarmate het rangnummer toeneemt:



9.11 Sommen van exponentieel verdeelde stochasten. Aangezien mijn theepotten nogal snel sneuvelen, besluit ik er maar 12 tegelijk te kopen (dat is nog voordeliger ook). Wat is de kans dat ik binnen drie jaar opnieuw naar de winkel moet?
Nummer de theepotten; zij X_n de levensduur van de n^{e} theepot (ingaande op het moment van ingebruikneming, dus als de $(n-1)^{\text{e}}$ kapotvalt). We veronderstellen $X_1, \ldots, X_{12}$ onafhankelijk, en alle exponentieel verdeeld met parameter λ; zij $Y_n = X_1 + \ldots + X_n$ de totale levensduur van de eerste n potten. Noem de dichtheid van Y_n: f_n, dan is weer $f_{n+1} = f_n * f_1$ zodat

$$\begin{aligned} f_{n+1}(t) &= \int_{-\infty}^{\infty} f_n(s) f_1(t-s) ds \\ &= \int_0^t f_n(s) \lambda e^{-\lambda(t-s)} ds \\ &= e^{-\lambda t} \int_0^t f_n(s) \lambda e^{\lambda s} ds. \end{aligned}$$

Dit geeft achtereenvolgens:

$$\begin{aligned} f_2(t) &= \lambda^2 t e^{-\lambda t} \\ f_3(t) &= \frac{\lambda^3}{2} t^2 e^{-\lambda t} \end{aligned}$$

en met behulp van volledige inductie:

$$f_n(t) = \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t}.$$

Dit is een van de zeldzame gevallen waarin convoluties gemakkelijk te berekenen zijn.
Sommen van exponentieel verdeelde stochasten hangen nauw samen met het Poissonproces: zie paragraaf 10.8.

Opgaven.

1. Beantwoord de vraag in de eerste alinea van paragraaf 9.9.

2. Laat (X, Y) continu verdeeld zijn met dichtheid

$$f(x,y) = \begin{cases} \frac{1}{\pi} & \text{als } x^2 + y^2 < 1; \\ 0 & \text{als } x^2 + y^2 > 1. \end{cases}$$

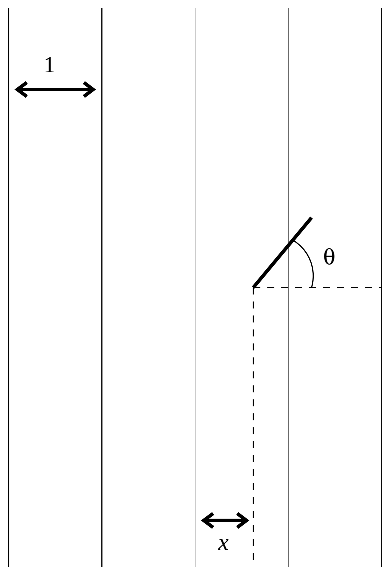
   (a) Bepaal de dichtheidsfuncties van X en Y.

   (b) Bepaal verdelings- en dichtheidsfunctie van $R = \sqrt{X^2 + Y^2}$.

   (c) Zijn X en Y onafhankelijk?

3. Zij $X \sim \text{Un}(0,1)$, $Y \sim \text{Exp}(1)$, $Z = X + Y$. Stel X en Y onafhankelijk. Bepaal dichtheids- en verdelingsfunctie van Z.

4. Een munt met diameter d ($d < 1$) wordt op een schaakbord geworpen, met vierkante vakjes van 1×1. Bepaal de kans dat de munt helemaal binnen één vakje terechtkomt.



5. Een vloer bestaat uit planken van 1 meter breed. Een breinaald met lengte ℓ ($\ell < 1$) wordt op de vloer geworpen. We willen de kans bepalen dat de breinaald geheel op één plank valt. We noteren zijn positie met (X, ϑ), waarin X de afstand is van het meest linkse punt van de breinaald tot de naad er links naast (dus $X \in (0,1)$) en ϑ de hoek van de naald met de positieve X-as (dus $\vartheta \in (-\frac{1}{2}\pi, \frac{1}{2}\pi)$).
We nemen aan: $X \sim \mathrm{Un}(0,1)$, $\vartheta \sim \mathrm{Un}(-\frac{1}{2}\pi, \frac{1}{2}\pi)$, X en ϑ onafhankelijk.

   (a) Leid af dat moet gelden: $X + \ell \cos\vartheta < 1$, wil de naald geheel op één plank vallen.

   (b) Teken een (X, ϑ)-assenstelsel ($0 < X < 1$, $-\frac{1}{2}\pi < \vartheta < \frac{1}{2}\pi$) en daarin het gebied dat voldoet aan (a).

   (c) Bereken nu de gevraagde kans.

6. De stochasten X en Y zijn onafhankelijk en uniform verdeeld over $(0,1)$. Verdeel $(0,1)$ in drie delen:

$$(0, \min\{X,Y\}), \quad (\min\{X,Y\}, \max\{X,Y\}), \quad (\max\{X,Y\}, 1).$$

   Probeer een driehoek te vormen met zijden, congruent aan deze intervallen. Wat is de kans dat dit lukt?

7. X en Y zijn onafhankelijke, exponentieel(1) verdeelde stochasten.

   (a) Bepaal de dichtheidsfunctie van (X, Y).

   (b) Zij $Z = |X - Y|$. Bewijs dat ook $Z \sim \mathrm{Exp}(1)$.

8. X en Y zijn onafhankelijke stochasten. $X \sim \mathrm{Un}(0,1)$, $Y \sim \mathrm{Un}(0,a)$, met $a > 1$.

   (a) Bereken verdelings- en dichtheidsfunctie van $Y - X$.

   (b) Bereken a als gegeven is dat $\mathbb{E}(Y - X) = \frac{19}{12}$.

9. Een mier (dikte nul) beweegt zich kriskras over een tafelblad met lengte $2n$ en breedte $2m$ ($n > m > 0$). De plaats waar de mier zich op zeker ogenblik bevindt, is uniform verdeeld over het blad. Zij X de afstand tot de dichtstbijzijnde rand.

   (a) Bereken de verdelingsfunctie van X.

   (b) Bereken de dichtheidsfunctie van X.

(c) Zij $n = 2m$. Bereken $\mathbb{E}(X)$ en $\mathbb{V}\mathrm{ar}(X)$.

10. De levensduur van grote vliegtuigmotoren is exponentieel (λ) verdeeld, die van kleine exponentieel (2λ). De levensduren van verschillende motoren zijn onafhankelijk.

    (a) In vliegtuigtype A bevinden zich een grote en een kleine motor die slechts samen het vliegtuig in werking houden. Zij T_A de levensduur van het vliegtuig (reparatie of vervanging is niet mogelijk). Bereken verdelingsfunctie en verwachting van T_A.

    (b) In type B bevinden zich een grote en een kleine motor die ook elk apart het vliegtuig in werking houden. Bij de start wordt de grote motor ingeschakeld; als deze uitvalt, start automatisch de kleine motor. Bereken dichtheid en verwachting van de levensduur T_B.

    (c) Type C is uitgerust met één grote en twee kleine motoren. Dit vliegtuig blijft in werking tot ofwel de grote motor uitvalt, ofwel beide kleine uitgevallen zijn. Alle drie de motoren werken vanaf de start. Bereken verdelingsfunctie en verwachting van de levensduur T_C.

Hoofdstuk 10

De Poisson-verdeling

10.1 Definitie en berekening.

Laat K het aantal klanten zijn dat in een tijdsinterval I van bijvoorbeeld een uur een winkel binnengaat. We willen de kansverdeling van K achterhalen.

Natuurlijk moeten we daartoe iets over het gedrag van de klanten aannemen. Eerst wat notatie: als $J \subset I$ een deelinterval is, dan verstaan we onder $K(J)$ het aantal klanten dat gedurende J binnenkomt. We veronderstellen:

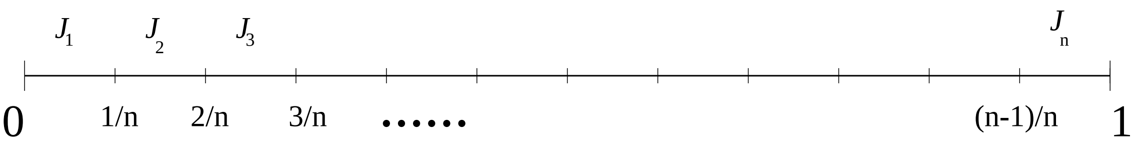
1. Als de deelintervallen J_1 en J_2 van I even lang zijn, dan hebben $K(J_1)$ en $K(J_2)$ dezelfde kansverdeling.

2. Als $J_1 \cap J_2 = \emptyset$, dan zijn $K(J_1)$ en $K(J_2)$ onafhankelijk.

3. Er komen nooit twee of meer klanten tegelijk binnen.

4. De verwachting $\lambda := \mathbb{E}(K)$ bestaat (en is eindig).

Het blijkt dat deze uitgangspunten de kansverdeling van K (en ook van $K(J)$ voor elk deelinterval J) helemaal vastleggen. Dat gaan we nu bewijzen.



Hiertoe verdelen we $I = (0, 1]$ in n gelijke stukken $J_1, \ldots, J_n$, dus $J_k = (\frac{k-1}{n}, \frac{k}{n}]$, $k = 1, \ldots, n$. Met $A_j^{(n)}$ duiden we de volgende gebeurtenis aan:

$$A_j^{(n)} := [\text{er komt minstens één klant binnen gedurende } J_j].$$

Wegens aanname 1 is de kans $\mathbb{P}\left(A_j^{(n)}\right)$ dezelfde voor alle waarden van j; we noemen deze kans: p_n.

Wegens aanname 2 zijn de gebeurtenissen $A_1^{(n)}, A_2^{(n)}, \ldots, A_n^{(n)}$ onafhankelijk. Het aantal K_n van de $A_j^{(n)}$'s die optreden (dus van de tijdsintervallen waarin er iemand binnenkomt) is dus binomiaal verdeeld met parameters n en p_n: voor alle $k \in \mathbb{N}$ is

$$\mathbb{P}[K_n = k] = \binom{n}{k} p_n^k (1 - p_n)^{n-k}. \tag{1}$$

Nu kan het aantal tijdsintervallen waarop er iemand binnenkomt natuurlijk niet groter zijn dan het totale aantal klanten dat binnenkomt: $K_n \leq K$. Maar ook volgt, uit aanname 3, dat, als er maar voldoende deelstreepjes in het interval I worden geplaatst, elke klant zijn eigen tijdsintervalletje krijgt toebedeeld:

$$\forall \omega \in \Omega : \lim_{n \to \infty} K_n(\omega) = K(\omega). \tag{2}$$

Het lijkt daarom een goed idee, in (1) de limiet $n \to \infty$ te nemen. Of nog liever: we stellen $n = 2^l$ en nemen $l \to \infty$. Dit is handig omdat (ga na!)

$$K_1 \leq K_2 \leq K_4 \leq K_8 \leq K_{16} \leq \ldots. \tag{3}$$

We duiden deze "limiet over de machten van 2" aan met $\lim^*_{n \to \infty}$. We zullen aantonen dat:

(a) $\quad \lim^*_{n \to \infty} \mathbb{P}[K_n = k] = \mathbb{P}[K = k]$ voor $k \in \mathbb{N}$);

(b) $\quad \lim^*_{n \to \infty} n p_n = \lambda.$

(a) bewijzen we als volgt: Voor willekeurige k geldt wegens (3) dat de gebeurtenissen $[K_n{>}k]$ met $n = 2^l$ een stijgende rij vormen: $[K_1{>}k] \subset [K_2{>}k] \subset [K_4{>}k] \subset \ldots$.
De vereniging van deze rij gebeurtenissen is (wegens (2)) gelijk aan de gebeurtenis $[K{>}k]$.
Volgens eigenschap 2.5(c) geldt dus

$$\lim^*_{n \to \infty} \mathbb{P}[K_n{>}k] = \mathbb{P}[K{>}k]. \tag{4}$$

En dus ook

$$\begin{aligned} \lim^*_{n \to \infty} \mathbb{P}[K_n{=}k] &= \lim^*_{n \to \infty} \left(\mathbb{P}[K_n{>}k{-}1] - \mathbb{P}[K_n{>}k] \right) \\ &= \mathbb{P}[K{>}k{-}1] - \mathbb{P}[K{>}k] \\ &= \mathbb{P}[K{=}k], \end{aligned}$$

waarmee (a) is aangetoond.

(b) gaat nu als volgt: Uit (3) volgt dat de verwachtingen $\mathbb{E}(K_n)$ met $n = 2^l$ een stijgende rij vormen. Deze rij is begrensd, want $K_n \leq K$, dus $\mathbb{E}(K_n) \leq \mathbb{E}(K)$ voor alle n.
De rij verwachtingen is dus convergent, en

$$\lim_{n\to\infty}^* \mathbb{E}(K_n) \le \mathbb{E}(K). \tag{5}$$

Wegens $\mathbb{E}(K_n) = \sum_{k=0}^{\infty} \mathbb{P}[K_n{>}k] \ge \sum_{k=0}^{N} \mathbb{P}[K_n{>}k]$ krijgen we met (4):

$$\lim_{n\to\infty}^* \mathbb{E}(K_n) \ge \lim_{n\to\infty}^* \sum_{k=0}^{N} \mathbb{P}[K_n{>}k] = \sum_{k=0}^{N} \mathbb{P}[K{>}k].$$

Daar dit voor iedere N geldt, krijgen we:

$$\lim_{n\to\infty}^* \mathbb{E}(K_n) \ge \sum_{k=0}^{\infty} \mathbb{P}[K{>}k] = \mathbb{E}(K).$$

Dit gecombineerd met (5) geeft (b).

Met behulp van (a) en (b) kunnen we nu de limiet $n \to \infty$ over de machten van 2 in (1) uitvoeren:

$$\begin{aligned}
\mathbb{P}[K = k] &= \lim_{n\to\infty}^* \mathbb{P}[K_n = k] \\
&= \lim_{n\to\infty}^* \binom{n}{k} p_n^k (1-p_n)^{n-k} \\
&= \frac{1}{k!} \lim_{n\to\infty}^* (np_n)\left((n{-}1)p_n\right)\cdots\left((n{-}k{+}1)p_n\right) \cdot \frac{(1-p_n)^n}{(1-p_n)^k} \\
&= \frac{\lambda^k}{k!} e^{-\lambda},
\end{aligned} \tag{6}$$

De verdeling van K wordt de *Poisson-verdeling met parameter* λ genoemd. Als een stochast X deze kansverdeling heeft, schrijven we $X \sim \text{Poisson}(\lambda)$.

10.2 Verwachting en variantie.

Dat $\lambda = \mathbb{E}(K)$ moet achteraf natuurlijk ook uit de kansverdeling van K berekend kunnen worden. Dat gaat zó:

$$\begin{aligned}
\mathbb{E}(K) &= \sum_{k=0}^{\infty} k\mathbb{P}[K = k] \\
&= \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} \\
&= e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \\
&= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda e^{-\lambda} e^{\lambda} \\
&= \lambda.
\end{aligned}$$

Berekening van $\mathbb{V}\mathrm{ar}(K)$ gaat het gemakkelijkst via dezelfde methode: we berekenen eerst

$$\begin{aligned}
\mathbb{E}(K(K-1)) &= \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} \\
&= e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^k}{(k-2)!} \\
&= \lambda^2 e^{-\lambda}e^{\lambda} \\
&= \lambda^2.
\end{aligned}$$

Vervolgens: $\mathbb{V}\mathrm{ar}(K) = \mathbb{E}(K^2) - \mathbb{E}(K)^2 = \mathbb{E}(K(K-1)) + \mathbb{E}(K) - \mathbb{E}(K)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$. Dus is ook

$$\boxed{\mathbb{V}\mathrm{ar}(K) = \lambda.}$$

10.3 Verband met de binomiale verdeling.

De laatste drie stappen uit berekening (6) zijn ook geldig voor binomiaal (n, p_n) verdeelde stochasten K_n mits

$$\lim_{n\to\infty} np_n = \lambda \tag{7}$$

(in dat geval hoeft de limiet ook niet meer over de machten van 2 genomen te worden).
Hier wordt triviaal aan voldaan door: $p_n = \frac{\lambda}{n}$.
Dus als de rij stochasten $X_1, X_2, X_3, \ldots$ voldoet aan: $X_n \sim \mathrm{Bin}(n, \frac{\lambda}{n})$, dan is

$$\lim_{n\to\infty} \mathbb{P}[X_n = k] = \frac{\lambda^k}{k!}e^{-\lambda}. \tag{8}$$

Men gebruikt dit resultaat vaak om voor kansen uit een binomiale verdeling met grote n en kleine p en benadering te vinden door niet de binomiale maar de "overeenkomstige" Poisson-kans te berekenen; dat scheelt een parameter en het is vaak wat minder lastig rekenwerk. Zie opgave 2. Het resultaat werd gevonden door Siméon Poisson, al in 1837, en was de reden dat zijn naam aan deze kansverdeling gekoppeld werd. Pas veel later zag men de zelfstandige betekenis van deze verdeling, die we in paragraaf 10.1 hebben gezien.
In (8) staat dat een binomiale verdeling met grote n en kleine p sterk lijkt op een Poisson-verdeling met parameter $\lambda = np$. Hun verwachtingen zijn gelijk, en hun varianties bijna gelijk (npq resp. $\lambda = np$; bedenk dat $q = 1 - p$ bijna 1 is). Omdat de Poisson-verdeling te voorschijn komt als limiet van een binomiale verdeling

met aanzwellende n (het aantal proeven) en krimpende succeskans p, noemt men de Poisson-verdeling wel de verdeling der "zeldzame gebeurtenissen". Deze benaming is eigenlijk misleidend. De succeskans p mag dan klein zijn, het verwachte aantal successen $np = \lambda$ hoeft helemaal niet klein te zijn. Een belangrijk *verschil* tussen de binomiale verdeling en de Poisson-verdeling is dat de uitkomstenruimte bij de binomiale verdeling *eindig* is ($\{0, 1, \dots, n\}$) en bij de Poisson-verdeling *oneindig* ($\mathbb{N}$).

10.4 Zeldzame gebeurtenissen en paardebloemen.

Er zijn in het dagelijks leven veel voorbeelden van stochasten die kunnen worden opgevat als het aantal successen in een lange reeks onafhankelijke pogingen met allemaal dezelfde, kleine succeskans. Zulke stochasten hebben in benadering een Poisson-verdeling. We noemen ze "zeldzame gebeurtenissen".
Voorbeelden:

(a) Het aantal drukfouten op een bladzijde.

(b) Het aantal telefoonnummers dat in Nijmegen op één dag verkeerd wordt gedraaid.

(c) Het aantal gebarsten straatklinkers in 10 meter straat.

Nog groter is het aantal voorbeelden waarbij, zoals in 10.1, de "pogingen" niet meer goed te onderscheiden zijn, maar elk stukje tijd (of lengte, of oppervlak, of...) als "poging" kan worden opgevat, waarbij de succeskans evenredig is met de duur (of lengte of oppervlakte of...). We noemen ze "paardebloemen". Voorbeelden:

(d) Het aantal paardebloemen in 1 are weiland.

(e) Het aantal klanten dat in één uur een winkel binnenkomt.

(f) Het aantal klanten dat in één uur een winkel binnenkomt en ook iets koopt.

(g) Het aantal borelingen op één dag in Nijmegen.

(h) Het aantal α-deeltjes dat in een bepaald tijdsinterval uitgezonden wordt door een radioactief preparaat.

10.5 Eigenschappenvan de Poisson-verdeling.

(a) Opteleigenschap.

$$\left.\begin{array}{r} K_1 \sim \text{Poisson}(\lambda_1) \\ K_2 \sim \text{Poisson}(\lambda_2) \\ K_1 \perp\!\!\!\perp K_2 \end{array}\right\} \Longrightarrow K_1 + K_2 \sim \text{Poisson}(\lambda_1 + \lambda_2).$$

Dus de som van onafhankelijke Poisson-verdeelde stochasten is weer Poisson-verdeeld.
Dit is heel logisch als je het formuleert in de termen van 10.1:
Wanneer $K_1 = K(J_1)$ en $K_2 = K(J_2)$, terwijl $J_1 \cap J_2 = \emptyset$ (zodat $K_1 \perp\!\!\!\perp K_2$) dan is natuurlijk $K_1 + K_2 = K(J_1 \cup J_2)$, en dit is Poisson met parameters $\mathbb{E}(K(J_1 \cup J_2)) = \mathbb{E}(K_1) + \mathbb{E}(K_2) = \lambda_1 + \lambda_2$.
We geven hier echter een direct bewijs:
Bewijs:

$$\begin{aligned}
\mathbb{P}[K_1 + K_2 = n] &= \sum_{k=0}^{n} \mathbb{P}[K_1 = k, K_2 = n-k] \\
&= \sum_{k=0}^{n} \mathbb{P}[K_1 = k]\mathbb{P}[K_2 = n-k] \\
&= \sum_{k=0}^{n} \frac{\lambda_1^k}{k!} e^{-\lambda_1} \cdot \frac{\lambda_2^{n-k}}{(n-k)!} e^{-\lambda_2} \\
&= e^{-(\lambda_1+\lambda_2)} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} \lambda_1^k \lambda_2^{n-k} \\
&\overset{\text{Bin. v. Newton}}{=} e^{-(\lambda_1+\lambda_2)} \cdot \frac{(\lambda_1+\lambda_2)^n}{n!}.
\end{aligned}$$

□

(b) Onafhankelijkheid van de individuen

$$\left.\begin{array}{r} K_1 \sim \text{Poisson}(\lambda_1) \\ K_2 \sim \text{Poisson}(\lambda_2) \\ K_1 \perp\!\!\!\perp K_2 \end{array}\right\} \implies \mathbb{P}(K_1 = k | K_1 + K_2 = n) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\text{met } p = \frac{\lambda_1}{\lambda_1 + \lambda_2}, \text{ alle } k, n.$$

In termen van het klantenprobleem: als gegeven is dat in het interval $J_1 \cup J_2$ in totaal n klanten binnenkomen, dan kunnen hun tijdstippen van binnenkomst als n onafhankelijke grootheden worden opgevat (elke klant maakt onafhankelijk van de andere klanten een keuze tussen J_1 en J_2); in het bijzonder zijn de n keuzes tussen J_1 en J_2 die de klanten maken, onafhankelijk van elkaar.
Het aantal k van de klanten die J_1 uitkiezen om binnen te komen is dus $\text{Bin}(n, \frac{\lambda_1}{\lambda_1+\lambda_2})$ verdeeld.
Deze eigenschap van de Poisson-verdeling is veel minder begrijpelijk op grond van de berekening in paragraaf 10.1. Ze kan desgewenst als alternatieve karakterisering van de Poisson-verdeling worden gebruikt. Het bewijs echter is wel eenvoudig:

Bewijs:

$$\begin{aligned}
\mathbb{P}(K_1 = k | K_1 + K_2 = n) &= \frac{\mathbb{P}[K_1 = k, K_1 + K_2 = n]}{\mathbb{P}[K_1 + K_2 = n]} \\
&= \frac{\mathbb{P}[K_1 = k, K_2 = n - k]}{\mathbb{P}[K_1 + K_2 = n]} \\
&= \frac{n!}{k!(n-k)!} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{n-k}.
\end{aligned}$$

□

(c) Splitsings-eigenschap.

$$\left. \begin{array}{r} K \sim \text{Poisson}(\lambda) \\ K = K_1 + K_2 \\ \mathbb{P}(K_1 = k | K = n) = \binom{n}{k} p^k (1-p)^{n-k} \end{array} \right\} \Longrightarrow \left\{ \begin{array}{l} K_1 \sim \text{Poisson}(\lambda p) \\ K_2 \sim \text{Poisson}(\lambda(1-p)) \\ K_1 \perp\!\!\!\perp K_2 \end{array} \right.$$

Als je door loting een Poisson-verdeeld aantal individuen in twee groepen verdeelt, dan zijn de aantallen van deze groepen onafhankelijk en weer Poisson-verdeeld.

Dit is een soort omkering van (b) met gebruikmaking van (a).

In het klantenmodel is deze eigenschap weer goed te verklaren (als we de onafhankelijkheid van de klanten, (b), aannemen): zij $K = K(I)$. Op grond van (b) kunnen we de loting per klant mooi uitvoeren door eenvoudig het interval I in twee stukken te hakken: $J_1 \cup J_2 = I$ en $J_1 \cap J_2 = \emptyset$. Natuurlijk zijn nu $K_1 = K(J_1)$ en $K_2 = K(J_2)$ Poisson-verdeeld en onafhankelijk.

Bewijs:

$$\begin{aligned}
\mathbb{P}[K_1 = k, K_2 = m] &= \mathbb{P}[K_1 + K_2 = k + m, K_1 = k] \\
&= \mathbb{P}[K_1 + K_2 = k + m] \cdot \mathbb{P}(K_1 = k | K_1 + K_2 = k + m) \\
&= \frac{\lambda^{k+m}}{(k+m)!} e^{-\lambda} \cdot \binom{k+m}{k} \cdot p^k \cdot (1-p)^m \\
&= \left(\frac{(\lambda p)^k}{k!} e^{-\lambda p} \right) \cdot \left(\frac{(\lambda(1-p))^m}{m!} e^{-\lambda(1-p)} \right).
\end{aligned}$$

Hieruit volgt:

$$\begin{aligned}
\mathbb{P}[K_1 = k] &= \sum_{m=0}^{\infty} \mathbb{P}[K_1 = k, K_2 = m] \\
&= \sum_{k=0}^{\infty} \left(\frac{(\lambda p)^k}{k!} e^{-\lambda p} \right) \cdot \left(\frac{(\lambda(1-p))^m}{m!} e^{-\lambda(1-p)} \right) \\
&= \frac{(\lambda p)^k}{k!} e^{-\lambda p},
\end{aligned}$$

dus $K_1 \sim \text{Poisson}(\lambda p)$. Net zo volgt dat $K_2 \sim \text{Poisson}(\lambda(1-p))$. Tevens hebben we gevonden dat

$$\mathbb{P}[K_1 = k, K_2 = m] = \mathbb{P}[K_1 = k] \cdot \mathbb{P}[K_2 = m]$$

voor elke k, m, dus dat K_1 en K_2 onafhankelijk zijn. □

Voorbeeld 1. Als het aantal klanten Poisson(λ) verdeeld is, en als iedere klant met kans p iets koopt, dan is het aantal kopers Poisson(λp) verdeeld, en het aantal niet-kopers Poisson($\lambda(1-p)$).

Voorbeeld 2. Als het aantal kinderen dat in een jaar geboren wordt Poisson(λ) verdeeld is, en bij iedere geboorte de kans op een jongen p is, dan is het aantal jongens K_1 dat geboren wordt Poisson(λp) verdeeld. Het aantal meisjes K_2 is Poisson($\lambda(1-p)$) verdeeld, en K_1 en K_2 zijn onafhankelijk.

10.6 Een toepassing.

Het aantal ongelukken K is Poisson(λ) verdeeld. Bij elk ongeluk bedraagt de schade 1, 2 of 3 eenheden, met kans p_1, p_2 resp. p_3 (en $p_1 + p_2 + p_3 = 1$). Zij S de totale schade. Bepaal $\mathbb{E}(S)$.

1e oplossing. $\mathbb{E}(S) = \sum_{k=0}^{\infty} \mathbb{E}(S|K = k)\mathbb{P}[K = k]$. Als $K = k$, kun je S schrijven als $S = Y_1 + \ldots + Y_k$, waarin Y_j de schade bij het j^{e} ongeluk voorstelt. Dan is $\mathbb{E}(Y_j) = p_1 + 2p_2 + 3p_3$ voor elke j, dus $\mathbb{E}(S|K = k) = k(p_1 + 2p_2 + 3p_3)$. We vinden dan

$$\begin{aligned}
\mathbb{E}(S) &= \sum_{k=0}^{\infty} k(p_1 + 2p_2 + 3p_3)e^{-\lambda}\frac{\lambda^k}{k!} \\
&= (p_1 + 2p_2 + 3p_3)e^{-\lambda}\lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda(p_1 + 2p_2 + 3p_3).
\end{aligned}$$

2e oplossing. Schrijf $K = K_1 + K_2 + K_3$, waarbij K_i het aantal ongelukken is dat schade i veroorzaakt. Dan is $S = K_1 + 2K_2 + 3K_3$. Wegens 10.5(c) en opgave 11 is elke K_i Poisson(λp_i) verdeeld, zodat $\mathbb{E}(K_i) = \lambda p_i$. Gevolg:

$$\mathbb{E}(S) = \mathbb{E}(K_1) + 2\mathbb{E}(K_2) + 3\mathbb{E}(K_3) = \lambda(p_1 + 2p_2 + 3p_3).$$

10.7 Het Poissonproces.

Over het klantenmodel uit paragraaf 10.1 valt nog meer interessants te zeggen.

Laat I eens het tijdsinterval $(0, t]$ voorstellen. Met een kleine wijziging in de oorspronkelijke opzet kunnen we best de tijd ná het tijdstip 1 dóór laten lopen. Zij

nu $K_t = K([0,t))$, het aantal klanten dat vóór tijdstip t binnenkomt. Natuurlijk is voor elke t de stochast K_t Poisson-verdeeld. Op grond van aannamen 1 en 4 leid je gemakkelijk af dat $\mathbb{E}(K_t) = \lambda t$ (Immers de functie $f : t \mapsto \mathbb{E}(K_t)$ is stijgend en additief, en voldoet aan $f(1) = \lambda$).
Dus $K_t \sim \text{Poisson}(\lambda t)$.
λ is het verwachte aantal klanten per tijdseenheid.
De familie van stochasten $(K_t)_{t\geq 0}$ wordt wel het *Poissonproces* met *intensiteit* λ genoemd.
Opmerking. Als we in het model hadden willen opnemen dat de intensiteit van het klantenbezoek gedurende de dag varieert (om half zes zal het gemiddeld drukker zijn dan om 10 uur), dan kunnen we de constante intensiteit λ vervangen door een intensiteitsfunctie $\lambda : t \mapsto \lambda(t)$. K_t heeft dan Poisson-verdeling met $\mathbb{E}(K_t) = \int_0^t \lambda(s)ds$. Dit komt erop neer dat we aanname 1 vervangen door:

1'. Als $\int_{J_1} \lambda(t)dt = \int_{J_2} \lambda(t)dt$, dan hebben $K(J_1)$ en $K(J_2)$ dezelfde kansverdeling.

10.8 Tussenpozen. De volgende eigenschap van het Poissonproces is van groot belang in toepassingen en simulaties.

> De tussenpozen $X_1 := T_1,\ X_2 := T_2 - T_1,\ X_3 := T_3 - T_2,\ \cdots$ in het Poissonproces zijn exponentieel (λ) verdeeld en onderling onafhankelijk.

Dat de aankomsttijd T_1 van de eerste klant exponentieel (λ) verdeeld is zul je zelf bewijzen in opgave 6. de exponentiele verdeling van de overige tussenpozen, en hun onafhankelijkheid kan men bewijzen op eenzelfde manier als in het geval van de geometrische verdeling (zie Hoofdstuk 3, opgave 9), maar we laten het bewijs hier achterwege.
De aankomsttijden kunnen dus worden opgevat als sommen van onafhankelijke exponentieel verdeelde stochasten, zoals we bekeken hebben in paragraaf 9.11. En dus is de verdeling van de aankomsttijd T_n van de n^{e} klant ($T_n = X_1 + \ldots + X_n$) precies dezelfde als we daar gevonden hebben . Hun dichtheidsfuncties zijn:

$$f_{T_n}(t) = \frac{\lambda^n t^{n-1}}{(n-1)!} e^{-\lambda t}.$$

Opmerking. Deze dichtheidsfuncties kan men ook anders afleiden. Als het aantal klanten K_t in $(0,t]$ Poisson (λt) verdeeld is, dan is nl.

$$F_{T_n}(t) \quad = \quad \mathbb{P}[T_n \leq t]$$

$$\begin{aligned} &= \mathbb{P}[K_t \geq n] \\ &= 1 - \mathbb{P}[K_t \leq n-1] \\ &= 1 - e^{-\lambda t}\left((1 + \lambda t + \frac{(\lambda t)^2}{2!} + \ldots + \frac{(\lambda t)^{n-1}}{(n-1)!}\right). \end{aligned}$$

Differentieer vervolgens: $f_{T_n}(t) = F'_{T_n}(t)$ en we vinden de gezochte dichtheid (opgave 7).

10.9 Het Poissonproces in $\mathbb{R}^n$.

In plaats van de tijdstippen van binnenkomst van klanten en tijdsinterval I hadden we in paragraaf 10.1 ook wel kunnen kiezen: paardebloemen in een weiland W. Verdeling in hokjes $H \subset W$ in plaats van deelintervallen leidt tot dezelfde uitkomst voor de verdeling van $K(W)$, en ook van $K(H)$ met $H \subset W$: een Poisson-verdeling met als parameter de gemiddelde dichtheid $\lambda \times$ het oppervlak van H.
Op overeenkomstige wijze kan men de verdeling van sterren over een stukje heelal beschrijven door een Poissonproces in $\mathbb{R}^3$. Dat geldt ook voor de verdeling van melkwegstelsels over een groter stuk heelal.

10.10 Simulatie van de Poissonverdeling. We kunnen de eigenschap van de onafhankelijke exponentieel verdeelde tussenpozen uit paragraaf 10.8 gebruiken om het Poissonproces, en daarmee de Poissonverdeling, te simuleren. Immers K_t, het aantal aankomsttijdstippen vóór tijdstip t, is Poisson-verdeeld met parameter λt. Kies nu even $\lambda = 1$ en noem t weer: λ. We simuleren de tijdstippen $T_1, T_2, \ldots$ door exponentieel(1) verdeelde stochasten bij elkaar te tellen: $T_1 = -\ln U_1$, $T_2 = T_1 + (-\ln U_2)$, ...
Stop zodra $T_{k+1} \geq \lambda$. Als $T_k < \lambda$ en $T_{k+1} \geq \lambda$, dan geven we X de waarde k.
In het bijzonder: $X = 0$ als $T_1 \geq \lambda$.
We moeten dus steeds controleren of T_k nog kleiner dan λ is.

$$\begin{aligned} T_k &= -\ln U_1 - \ln U_2 - \ldots - \ln U_k \\ &= -\ln(U_1 \cdots U_k) < \lambda \end{aligned}$$

en dit is precies het geval als

$$U_1 \cdots U_k > e^{-\lambda}.$$

We vermenigvuldigen dus randomgetallen $U_1, U_2, \ldots$ zolang het product groter blijft dan $e^{-\lambda}$. Op den duur duikt het product onder het niveau $e^{-\lambda}$. De laatste k waarvoor $U_1 \cdots U_k > e^{-\lambda}$ is de gesimuleerde van onze Poisson(λ)-stochast.

Opgaven

1. Zij X Poisson(λ) verdeeld. Voor welke $k \in \mathbb{N}$ is $\mathbb{P}[X = k]$ maximaal?

2. Stel dat iedere schroef die door een machine gemaakt wordt, met kans 0,01 een gebrek heeft (onafhankelijk van de andere schroeven). Benader de kans dat er in een doosje met 200 schroeven minder dan 5 defecte schroeven zitten (Aanwijzing: paragraaf 10.3).

3. Bepaal in paragraaf 10.6 de kans $\mathbb{P}[S{=}k]$ voor $k = 0, 1, 2, 3$. Heeft S een Poisson-verdeling?

4. Het aantal geboorten X in Nijmegen op één dag heeft een Poisson(λ) verdeling. Bij iedere geboorte kan het (onafhankelijk van andere geboorten) een eenling (kans $p_1 \in (0, 1)$) of een tweeling (kans p_2) zijn; $p_1 + p_2 = 1$. Dus als X_1 het aantal eenlingen en X_2 het aantal tweelingen is, dan is $X = X_1 + X_2$ het aantal geboorten en $Y = X_1 + 2X_2$ het aantal kinderen.

   (a) Bepaal $\mathbb{P}[Y = 4]$.

   (b) Bepaal $\mathbb{E}(Y)$ en $\mathbb{V}\mathrm{ar}(Y)$.

   (c) Heeft ook Y een Poisson-verdeling?

5. In een zeker gebied treden aardbevingen bij benadering op volgens een Poissonproces met een gemiddelde van 2 aardbevingen per week.

   (a) Bereken de kans dat er de komende twee weken minstens drie aardbevingen optreden.

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

6. (a) Klanten komen aan volgens een Poisson(λ) proces. Zij T de aankomsttijd van de 1e klant. Zij $t > 0$. Bepaal de kans $\mathbb{P}[T > t]$ (Aanwijzing: Druk de gebeurtenis $[T > t]$ uit in een gebeurtenis m.b.v. X_t, het aantal klanten in $(0, t]$).

   (b) Zij in paragraaf 10.9 R de afstand tot de oorsprong van de dichtstbijzijnde paardebloem. Bepaal $\mathbb{P}[R > t]$.

   (c) Zij in paragraaf 10.9 R de afstand van de dichtstbijzijnde ster tot de oorsprong. Bepaal $\mathbb{P}[R > t]$.

7. Bepaal f_{T_n} in paragraaf 10.8 door de aldaar genoemde differentiatie uit te voeren.

8. Hagelstenen vallen op een vlak. Voor ieder deel van het vlak met oppervlakte a is het aantal hagelstenen Poisson verdeeld met verwachting λa (voor elke $a > 0$). (X, Y) zijn de coördinaten van de steen die het dichtst bij $(0, 0)$ valt. Zij $R = \sqrt{X^2 + Y^2}$. Bepaal verdelingsfunctie en dichtheid van R. Is R exponentieel verdeeld?

9. Grassprietjes groeien in een tuin volgens een Poisson(λ) proces[1]. Er staat een paaltje (dikte nul) in de tuin.

   (a) Er staan 10.000 grassprietjes op hoogstens 1 m afstand van het paaltje. Hoe groot is de kans dat er hiervan precies één op hoogstens 1 cm van het paaltje staat?

   (b) Geef een Poissonbenadering voor de in (a) gevonden kans.

10. Tussen 11 en 12 uur komen klanten een winkel binnen volgens een Poissonproces met intensiteit 3. Elke klant blijft twintig minuten in de winkel. Hoe groot is de kans dat twee klanten elkaar ontmoeten?

11. (Een uitbreiding van de "Splitsings-eigenschap") Bewijs:

$$\left.\begin{array}{r} K \sim \text{Poisson}(\lambda) \\ K = K_1 + K_2 + K_3 \\ \mathbb{P}[K_i = k | K = n] = \binom{n}{k} p_i^k (1-p_i)^{n-k} \end{array}\right\} \Rightarrow \left\{\begin{array}{l} K_i \sim \text{Poisson}(\lambda p_i) \\ K_1, K_2, K_3 \text{ zijn onafhankelijk.} \end{array}\right.$$

[1] d.w.z. gemiddeld λ sprietjes per cm^2.

Hoofdstuk 11

Simulatie

11.1 De experimentele weg. Toevalsverschijnselen kunnen we analyseren met behulp van de theorie in de voorgaande hoofdstukken.

Voorbeeld. Hoeveel worpen met een dobbelsteen zijn nodig om iedere kant minstens één keer boven te krijgen? Langs theoretische weg (opgave 12 uit hoofdstuk 5) vind je de verwachting: 14,7 $(= 1 + \frac{6}{5} + \frac{6}{4} + \frac{6}{3} + \frac{6}{2} + 6)$, dus gemiddeld lukt het in ongeveer 15 worpen.

Hoe groot is de kans dat je 20 worpen nodig hebt? Als je geen zin hebt deze kans uit te rekenen (hoe zou je dat trouwens moeten doen?), dan kun je nog altijd proefondervindelijk een indruk krijgen van de grootte van deze kans, en wel op twee manieren:

(a) Doe de proef echt. Gooi 20 dobbelstenen en kijk of je “alles hebt”. Herhaal dit (100 keer bijvoorbeeld).

(b) Doe alsof (simuleer). Gebruik de computer. De randomgenerator produceert een getal U uniform verdeeld tussen 0 en 1. Met $[6U] + 1$ krijg je met gelijke kansen uitkomsten 1, 2, 3, 4, 5, en 6 net als bij een echte dobbelsteen. Je kunt de computer dit laten herhalen, bijvoorbeeld tot iedere uitkomst aan de beurt geweest is. Zo krijg je een gesimuleerde waarde van de stochast X: het benodigde aantal worpen. Laat de computer 1000 keer een waarde voor X produceren. Daarbij wordt bijgehouden hoe vaak ieder van de mogelijke waarden $6, 7, 8, \ldots$ optreedt. Het resultaat kan op het scherm gebracht worden in de vorm van een histogram en een tabel.

11.2 Randomgenerator. De ideale randomgenerator produceert een rij toevalsgetallen $U_1, U_2, U_3, \ldots$ die voldoet aan:

(a) Iedere U_i is uniform verdeeld over het interval $(0,1)$: $\mathbb{P}[U_i \leq u] = u$ voor u tussen 0 en 1.

(b) De stochasten U_1, U_2, $U_3, \ldots$ zijn onafhankelijk.

Aan deze eisen is in de praktijk nooit helemaal exact voldaan. Wegens het gebruik van een beperkt aantal (zeg k) decimalen is de verdeling van een randomgetal eigenlijk niet eens continu, zodat aan de eerste eis niet voldaan is. Dit is voor de meeste praktische doeleinden niet zo erg, als alle 10^k mogelijke uitkomsten maar dezelfde kans hebben, d.w.z. op den duur ongeveer even vaak voorkomen.
Ernstiger zijn meestal de tekortkomingen wat betreft de tweede eis. De onafhankelijkheid van de uniforme stochasten U_n en U_{n+1} impliceert dat de punten (U_n, U_{n+1}) uniform verdeeld zijn over het eenheidsvierkant. Eenvoudige contrôle: Breng een groot aantal van die punten op het scherm. Doe dit een paar keer. Ontstaan er systematisch bepaalde patronen dan is er iets mis met eis (b).

Er zijn allerlei methoden en principes voor het maken van randomgeneratoren. Zie hiervoor bijvoorbeeld: D.E. Knuth, *The Art of Computer Programming*, deel 2, hoofdstuk 3. Daar wordt ook beschreven hoe je kunt testen of de geproduceerde toevalsrijen wel voldoende toevallig, chaotisch, onvoorspelbaar zijn. Wij zullen in het volgende uitgaan van een ideale randomgenerator en enkele dingen bespreken die we ermee kunnen doen.
In de rest van dit hoofdstuk stelt U steeds een uniforme stochast voor, en $U_1, U_2, \ldots$ een rij van onafhankelijke uniforme stochasten zoals geproduceerd door een ideale randomgenerator.

11.3 Enkele discrete verdelingen

(a) Discreet uniform. Zo noemt men een stochast X met $\mathbb{P}[X{=}k] = \frac{1}{N}$ voor $k = 1, 2, \ldots, N$. Deze verdeling krijgen we met $X = [NU] + 1$.

(b) Alternatieve verdeling: $\mathbb{P}[X{=}1] = 1 - \mathbb{P}[X{=}0] = p$.
Recept: $X = [p + U]$.

(c) Binomiale verdeling (met parameters n en p).
Uit de interpretatie van X: het aantal successen in n onafhankelijke proeven met succeskans p, volgt dat we X kunnen simuleren met $X = X_1 + \ldots + X_n$, waarbij $X_i = [p + U_i]$ voor $i = 1, 2, \ldots, n$.

(d) Geometrische verdeling. $\mathbb{P}[X{=}k] = pq^{k-1}$ voor $k = 1, 2, \ldots$.

Interpretatie: X is het aantal proeven dat nodig is om één succes te behalen. Dit geeft ons de *eerste methode* om X te simuleren: Maak $X_1, X_2, \ldots$ met $X_i = [p + U_i]$ voor $i = 1, 2, \ldots$. Stop zodra je een 1 krijgt. Bijvoorbeeld: Als $X_1 = 1$ dan $X = 1$. Als $X_1 = X_2 = \ldots = X_{k-1} = 0$ en $X_k = 1$, dan $X = k$. Het aantal randomgetallen dat nodig is om één waarde van X te produceren is bij deze methode variabel (gemiddeld $\frac{1}{p}$ stuks).

Het is ook mogelijk om het met slechts één randomgetal klaar te spelen. *Tweede methode:* $X := 1 + \left[\dfrac{\ln U}{\ln q}\right]$ (ga na dat $\mathbb{P}[X{>}k] = q^k$ en dus $\mathbb{P}[X{=}k] = \mathbb{P}[X > k{-}1] - \mathbb{P}[X{>}k] = q^{k-1} - q^k = pq^{k-1}$).

11.4 De trekking van de lottogetallen. Uit de bol met ballen, genummerd $1, 2, \ldots, N$, wordt n keer een bal getrokken. Dit gaat zonder teruglegging. Zo krijgen we een rijtje: $X_1, X_2, \ldots, X_n$ waarbij X_i het resultaat van de i^{e} trekking voorstelt. Bij de Nederlandse lotto: $N = 41$, $n = 6$ (of 7 met het reservegetal erbij). In Duitsland: $N = 47$, $n = 6$ (of 7).

Je speelt mee door op een formuliertje de nummers te voorspellen die getrokken gaan worden. Is je voorspelling goed of bijna goed, dan heb je een prijs.

Na de trekking worden de getrokken nummers op volgorde van grootte gelegd. Het *geordende* resultaat geven we aan met $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$.

Bijvoorbeeld bij een trekking $(X_1, \ldots, X_6) = (23, 16, 34, 2, 15, 20)$ wordt de gerangschikte uitslag: $(X_{(1)}, \ldots, X_{(6)}) = (2, 15, 16, 20, 23, 34)$.

Hoe kunnen we zulke lotto-uitslagen met de computer nabootsen? Eerst een paar algemene opmerkingen:

1. We hebben hier te doen met een steekproef van omvang n uit de verzameling $\{1, \ldots, N\}$. Deze steekproef veronderstellen we *aselect* te zijn. Dat betekent: Alle $\binom{N}{n}$ mogelijkheden voor $(X_{(1)}, \ldots, X_{(n)})$ hebben dezelfde kans, en ook: alle $n!\binom{N}{n}$ mogelijkheden voor $(X_1, \ldots, X_n)$.

2. Ieder element van $\{1, \ldots, N\}$ heeft dezelfde kans om bij de i^{e} trekking te voorschijn te komen: $\mathbb{P}[X_i{=}k] = \frac{1}{N}$, voor $i = 1, \ldots, n$ en $k = 1, \ldots, N$. Dus: iedere X_i is discreet uniform.

3. Ieder element van de verzameling $\{1, \ldots, N\}$ heeft dezelfde kans om in de steekproef terecht te komen. Deze kans is $\frac{n}{N}$.

De beweringen 2 en 3 volgen uit de veronderstellingen in 1. Bewijs dit zelf.

Simulatie van $(X_1, \ldots, X_n)$**.**
EERSTE METHODE: (primitief). De X_i's zijn discreet uniform, dus neem $X_i = [NU_i] + 1$, $i = 1, \ldots, n$ (dus net als bij een steekproef met teruglegging). Controleer nu of alle n getallen verschillend zijn. Zo nodig uitkomsten die eerder zijn voorgekomen weggooien en extra trekkingen doen totdat er n verschillende resultaten zijn.
Je kunt dit op verschillende manieren programmeren. Je kunt bijvoorbeeld bij iedere trekking controleren of het resultaat al eerder is voorgekomen, zo ja, weg ermee en opnieuw trekken. Bij deze methode is het aantal trekkingen (en dus het aantal benodigde randomgetallen) variabel.
De volgende methode heeft altijd precies n randomgetallen nodig.
TWEEDE METHODE. Het idee is als volgt. Zet N bakjes op een rij. In ieder bakje een bal. $A(i)$ is de inhoud van bakje i (het nummer van de bal in het i^{e} bakje).
Beginsituatie: $A(i) = i$ voor $i = 1, \ldots, N$ (bal i in bakje i).
Eerste trekking: Kies een bakje, ieder bakje met dezelfde kans. Neem de bal uit dat bakje. Dat is de eerste getrokken bal. Die kan ook niet opnieuw getrokken worden, want het bakje waar hij in zat is nu leeg.
We doen nu de bal uit bakje N in dat lege bakje, zodat bakjes 1 t/m $N-1$ gevuld zijn en bakje N leeg (had je bakje N gekozen, dan hoef je niets te doen). We zijn nu klaar voor de
Tweede trekking: Kies een van de bakjes $1, 2, \ldots, N-1$ met gelijke kansen. De bal uit dat bakje is de tweede getrokken bal.
De bal uit bakje $N-1$ doen we in het bakje dat net geleegd is. Zo gaan we door.
Iets formeler: Beginsituatie: $A(i) = i$ voor $i = 1, \ldots, N$.
1^{e} trekking: Kies i_1 uniform uit $\{1, \ldots, N\}$:

```
i1 := [NU1] + 1                    (Bakje i1 gekozen)
X1 := A(i1)                        (Eerste trekking wordt: bal uit bakje i1)
A(i1) := A(N)                      (Nieuwe inhoud bakje i1 wordt gelijk aan inhoud bakje N)
```

2^{e} trekking: Kies i uniform uit $\{1, \ldots, N-1\}$:

```
i2 := [(N-1)U2] + 1
X2 := A(i2)
A(i2) := A(N-1)
```

Het is duidelijk hoe dit verder gaat.
Door een variabele v in dienst te nemen die bijhoudt hoeveel bakjes nog gevuld zijn,

kunnen we de procedure iets netter beschrijven:

Methode A

Beginsituatie: $v = N$; $A(i) = i$ voor $i = 1, \ldots, N$.
Voor $k = 1, \ldots, n$ doe je het volgende:

`i := [vU] + 1`	(keuze bakje i uniform uit $\{1, \ldots, v\}$)
`X`$_k$ `:= A(i)`	(k^e trekking: bal uit bakje i)
`A(i) := A(v)`	(Nieuwe vulling bakje i wordt bal uit meest rechtse bakje)
`v := v - 1`	(Aantal gevulde bakjes wordt één minder)

Het is niet moeilijk in te zien dat deze methode correct is, d.w.z. de steekproef is aselect zoals gedefinieerd op blz.121. Immers, wat gebeurt er precies?
Er worden achtereenvolgens bakjes $i_1, i_2, \ldots, i_n$ gekozen. Het aantal mogelijkheden voor $i_1, i_2, \ldots, i_n$ is

$$N(N-1)\cdots(N-n+1) = \frac{N!}{(N-n)!}.$$

Bij ieder rijtje $(i_1, \ldots, i_n)$ hoort precies één steekproef $(x_1, \ldots, x_n)$ (bijvoorbeeld bij $(i_1, \ldots, i_n) = (1, 1, \ldots, 1)$ krijg je $(x_1, \ldots, x_n) = (1, N, N-1, \ldots, N-n+2)$).
Verander je het rijtje $(i_1, \ldots, i_n)$ dan verandert ook $(x_1, \ldots, x_n)$.
Het aantal mogelijkheden voor $(x_1, \ldots, x_n)$ is $n!\binom{N}{n} = \frac{N!}{(N-n)!}$, evenveel als voor $(i_1, \ldots, i_n)$. Er is dus een 1-1 correspondentie.
De kans op een bepaald rijtje $(i_1, \ldots, i_n)$ is gelijk aan

$$\frac{1}{N} \cdot \frac{1}{N-1} \cdots \frac{1}{N-n+1}$$

(hier gebruik je de onafhankelijkheid van de gebruikte randomgetallen $U_1, \ldots, U_n$).
Dus iedere mogelijkheid heeft dezelfde kans.

Random permutatie. Neem je $n = N$, dan maakt de voorgaande procedure een random rangschikking van de getallen $1, \ldots, N$. Dit vindt allerlei toepassingen: van CD-spelers tot bridgedrives.

Simulatie van $(X_{(1)}, \ldots, X_{(n)})$.
De gerangschikte steekproef $(X_{(1)}, \ldots, X_{(n)})$ kunnen we krijgen door eerst $(X_1, \ldots, X_n)$ te maken zoals we hiervoor besproken hebben, en daar de getrokken nummers in de goede volgorde te zetten (sorting). Er is echter ook een procedure die ons rechtstreeks de gerangschikte steekproef geeft.

Het idee is als volgt: Bij ieder element van $\{1,\ldots,N\}$ kiest het toeval of dit element zal worden opgenomen in de steekproef of niet. Het resultaat van dit kiesproces wordt weergegeven als een rijtje $(Y_1,\ldots,Y_N)$ bestaande uit n enen en $N-n$ nullen, $Y_i = 1$ als element i gekozen wordt, en $Y_i = 0$ als dat niet zo is (voorbeeld: de steekproef $(X_{(1)},\ldots,X_{(4)}) = (3,4,7,9)$ uit $\{1,\ldots,10\}$ wordt weergegeven door $(Y_1,\ldots,Y_{10}) = (0,0,1,1,0,0,1,0,1,0)$). Volgens opmerking 3 op blz. 121 geldt voor iedere i: $\mathbb{P}[Y_i = 1] = \frac{n}{N}$. Maar de stochasten $Y_1,\ldots,Y_N$ zijn niet onafhankelijk. De afhankelijkheid is als volgt: Als van de elementen $1,\ldots,k-1$ er al s gekozen zijn dan is de kans dat k gekozen wordt gelijk aan $\frac{n-s}{N-k+1}$, immers van de overblijvende $N-k+1$ elementen $k,\ldots,N$ moeten er nog $n-s$ gekozen worden om de steekproef vol te maken. Dus

$$\mathbb{P}(Y_k = 1 | Y_1 + \ldots + Y_{k-1} = s) = \frac{n-s}{N-k+1}.$$

We kunnen het ook zo zeggen: De kans op een één (of een nul) bij de volgende keuze is evenredig met het aantal ontbrekende enen (nullen). Dit leidt automatisch tot het juiste aantal enen want zodra je n enen hebt wordt de kans op nog een één nul. Deze intuïtieve gedachtengang leidt tot de volgende

Methode B

Beginsituatie: $s = 0$ (s houdt bij hoeveel énen je al hebt, dus hoeveel elementen er al gekozen zijn).
Voor $k = 1,\ldots,N$ doe je het volgende:

$$\texttt{Y}_{\texttt{k}} \texttt{ := [U + } \frac{\texttt{n-s}}{\texttt{N-k+1}}\texttt{]}$$

```
s := s + Yk
ALS Yk = 1 DAN X(s) := k
```

Op deze manier maken we een rijtje $(Y_1,\ldots,Y_N)$ en tegelijk de bijbehorende gerangschikte steekproef $(X_{(1)},\ldots,X_{(n)})$.
De methode is correct: alle $\binom{N}{n}$ mogelijkheden hebben dezelfde kans. Neem maar zo'n rijtje $(y_1,\ldots,y_n)$ met n enen en $N-n$ nullen.
De kans op dit rijtje is een breuk met noemer $N(N-1)\cdots 2\cdot 1$. De teller bevat alle factoren n, $n-1$,..., 2, 1 en $N-n$, $N-n-1$,... (de ontbrekende aantallen enen en nullen). Dus ieder rijtje $(y_1,\ldots,y_N)$ heeft kans $\frac{n!(N-n)!}{N!}$.

11.5 Steekproef en populatie. Behalve steekproeven uit $\{1,\ldots,N\}$ zoals op zondagavond bij de trekking van de lottogetallen, komt de oplettende burger door

de week nog heel wat steekproeven tegen uit allerlei verzamelingen. Dikwijls staat men er niet bij stil dat de gegevens die we tegenkomen slechts een steekproef zijn uit een groter geheel, een topje van de ijsberg. Van de grote massa onder water hebben we een vaag vermoeden.
Door het nemen van steekproeven kan men iets te weten komen over een grotendeels onbekende verzameling die men wil onderzoeken.

De algemene situatie kan, abstract, zo voorgesteld worden:
Er is een populatie $\mathcal{P}$. De elementen van $\mathcal{P}$ kunnen getallen zijn of vectoren, of functies, of ...
Voor de eenvoud houden wij het bij getallen: $\mathcal{P}$ is een eindig rijtje van getallen $(x_1, \ldots, x_N)$. Het gaat bijvoorbeeld over de studenten in een bepaalde plaats. x_i is het geldbedrag dat student i per maand van zijn/haar ouders krijgt, of het aantal glazen dat i drinkt, of waar men zich ook maar voor interesseert.
Men kiest nu, op goed geluk, aselect, wat elementen uit $\mathcal{P}$. Deze vormen de steekproef.
Hoe kunnen we zo'n steekproef trekken?
Omdat de elementen van $\mathcal{P}$ genummerd zijn, komt dit neer op het trekken van een steekproef uit $\{1, \ldots, N\}$. Je kunt je voorstellen dat $\mathcal{P}$ een kaartsysteem is met kaarten genummerd van 1 tot en met N. Het rangnummer i staat onopvallend in de linkerbovenhoek van kaart i en het getal x_i waar het om gaat staat dik gedrukt midden op die kaart. Dus: trek maar n nummers uit $\{1, \ldots, N\}$. De gegevens op de bijbehorende kaartjes vormen dan de steekproef $(X_1, \ldots, X_n)$.
Voor het *simuleren* van een steekproef (aselect en zonder terugleggen) $(X_1, \ldots, X_n)$ *uit* $\{x_1, \ldots, x_N\}$ kunnen we methode A gebruiken. We hoeven alleen de eerste regel maar te veranderen. Die wordt nu:

Methode A′

Beginsituatie: $v = N$; $A(i) = x_i$ voor $i = 1, \ldots, N$.
De rest is hetzelfde als bij **Methode A**

11.6 De hypergeometrische verdeling. Uit een doos met N ballen, m witte en $N-m$ rode, neemt men een aselecte steekproef (zonder teruglegging) van n ballen. In de steekproef treft men X witte ballen aan. X heeft de bekende hypergeometrische verdeling.
Dit is een bijzonder geval van de algemene situatie hiervoor. De populatie $\mathcal{P}$ is hier

$(x_1,\ldots,x_N)$ met $x_i = 1$ als $1 \le i \le m$ en $x_i = 0$ als $m < i \le N$. Tellen we de n getallen $X_1,\ldots,X_n$ bij elkaar op, dan hebben we X.

Simulatie van X kan dus zoals hierboven beschreven, maar het gaat ook heel goed als volgt:

> Beginsituatie: $X = 0$ (X houdt bij hoeveel witte ballen er getrokken zijn)
> Voor $k = 1$ tot en met n doen we het volgende:
> $\mathtt{X_k\ :=\ [\frac{m-X}{N-k+1}\ +\ U]}$
> $\mathtt{X\ :=\ X\ +\ X_k}$

Er zijn bij deze methode n randomgetallen nodig. Soms kan het met minder. Bijvoorbeeld als $m < n$.
Voorbeeld: Het aantal azen in een hand van 13 kaarten heeft dezelfde verdeling als het aantal ♡ in een greep van 4 kaarten. In het eerste geval hebben we een hypergeometrische verdeling met $n = 13$, $m = 4$ ($N = 52$). In het tweede geval: $n = 4$, $m = 13$.
De verwisselbare rol van m en n blikt als we de kans op $X = k$ uitschrijven:

$$\mathbb{P}[X = k] = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} = \frac{n!m!(N-n)!(N-m)!}{N!k!(n-k)!(m-k)!(N-n-m+k)!}.$$

Ook de uitdrukkingen voor verwachting en variantie vertonen deze symmetrie:

$$\mathbb{E}(X) = \frac{nm}{N} \text{ en } \mathbb{V}\mathrm{ar}(X) = \frac{nm(N-m)(N-n)}{N^2(N-1)}.$$

Je kunt het ook zo bekijken: De populatie is ingedeeld naar kleur: wit (m stuks) en rood ($N - m$). Door het trekken van de steekproef komt er een tweede, toevallige, indeling bij: Een bal kan in de steekproef zitten (n stuks) of niet ($N - n$ stuks). De vier aantallen n, m, $N - n$ en $N - m$ zijn de randgetallen van een 2×2 tabelletje

	wit	rood	
In steekproef	X	$n-X$	n
Niet in steekproef	$m-X$	$N-n-m+X$	$N-n$
	m	$N-m$	

De tweede indeling wordt aangebracht onafhankelijk van de eerste. De aantallen in de 4 hokjes (cellen) hebben alle vier de hypergeometrische verdeling.
Voor simulatie komt het vakje met de kleinste randgetallen het eerst in aanmerking. het aantal randomgetallen is dan gelijk aan het kleinste randgetal.

11.7 Simulatie in de statistiek. *11.7.1 Twee gemiddelden*

Een test meet zoiets als kennis, begrip, woordenschat van het Engels. 8 jongens en 10 meisjes hebben die test afgelegd en daarbij de volgende scores behaald:

Jongens:	48, 58, 62, 72, 74, 80, 88, 98	gemiddeld: $\frac{580}{8} = 72{,}5$
Meisjes:	63, 72, 74, 74, 74, 80, 80, 85, 92, 94	gemiddeld: 78,8

Het gemiddelde van de meisjes ligt 6,3 hoger dan het gemiddelde van de jongens. Hoe beoordeel je zoiets?
Ook als er in het algemeen geen verschil zou bestaan tussen de prestaties van jongens en meisjes op deze test dan zul je toch bijna nooit twee groepen met precies hetzelfde gemiddelde tegenkomen. Toeval speelt hier een rol.
Laten we uitgaan van de veronderstelling dat er geen wezenlijk verschil is. Deze aanname noemt men in de statistiek de "nulhypothese". Deze houdt in dat geconstateerde verschillen op toeval berusten. De invloed van het toeval kunnen we onderzoeken d.m.v. simulatie.
Dat gaat zo. Doe alle uitslagen bij elkaar in één rijtje

$$\mathcal{P} = (x_1, \ldots, x_{18}) = (48, 58, \ldots, 98, 63, \ldots, 94).$$

Trek steekproeven van 8 uit $\mathcal{P}$: $(X_1, \ldots, X_8)$. Vergelijk het gemiddelde

$$\overline{X} = \tfrac{1}{8}(X_1 + \ldots + X_8)$$

(het gemiddelde van de jongens) met het gemiddelde $\overline{Y}$ van de rest (de meisjes).
$8\overline{X} + 10\overline{Y} = 1368$ (het totaal van alle 18 uitslagen), dus $\overline{Y} = 136{,}8 - \frac{8}{10}\overline{X}$.
Het verschil van de gesimuleerde gemiddelden is dus

$$V = \overline{X} - \overline{Y} = 1{,}8\overline{X} - 136{,}8.$$

Door een groot aantal van zulke verschillen te simuleren krijgen we een indruk of het feitelijk waargenomen verschil uitzonderlijk is. Krijg je bijvoorbeeld in 1000 herhalingen maar 20 keer (2%) V-waarde kleiner dan $-6{,}3$ (het verschil tussen gemiddelden zoals oorspronkelijk waargenomen) dan hebben we hier toch wel een sterke aanwijzing dat het waargenomen verschil niet toevallig is (een *significant verschil* heet zoiets in de statistiek).

11.7.2 Correlatie.

n personen hebben ieder aan twee tests meegedaan. Persoon i heeft score x_i behaald op de 1^e test (Engels) en score y_i op de 2^e test (wiskunde).
Dit geeft n punten (x_i, y_i). Deze "puntenwolk" (scatter diagram) kan al een indruk

geven van een zekere samenhang tussen de prestaties door beide tests gemeten. Een kwantitatieve maat voor die samenhang is de zogenaamde *correlatiecoëfficiënt* r. Deze wordt als volgt berekend (vergelijk ook paragraaf 5.12):
Bepaal eerst de gemiddelden

$$\overline{x} = \tfrac{1}{n}\sum x_i \text{ en } \overline{y} = \tfrac{1}{n}\sum y_i.$$

Dan:

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left(\sum(x_i - \overline{x})^2\right)\left(\sum(y_i - \overline{y})^2\right)}}.$$

Het is niet moeilijk in te zien dat r altijd tussen -1 en $+1$ ligt. $r = 1$ als alle punten op een lijn $y = ax + b$ met $a > 0$ liggen. $r = -1$ als alle punten op een lijn $y = ax + b$ met $a < 0$ liggen.
Ook als er in werkelijkheid geen enkel verband is tussen x en y dan zal r toch, af en toe, een tamelijk hoge waarde kunnen aannemen.
Om te kunnen beoordelen of de correlatiecoëfficiënt die we uit onze waarnemingen berekend hebben echt iets betekent, moeten we de toevalsspreiding van r kennen. Daar kunnen we een goede indruk van krijgen met een simulatie-experiment. We maken de verbinding tussen de x-waarden en de y-waarden in paren (x_i, y_i). We zetten de x-waarden in een vaste volgorde en maken een randompermutatie van de y-waarden: $Y_1, \ldots, Y_n$.
Voor de puntenwolk (x_i, Y_i), $i = 1, \ldots, n$, berekenen we r.
Dat doen we een groot aantal keren, 1000 keer bijvoorbeeld.
Dan kun je zien of de feitelijke waarde van r uitzonderlijk hoog (of laag) is.
Men kan aantonen dat voor de verwachting en spreiding van r (over alle $N!$ gelijk-waarschijnlijke permutaties) geldt:

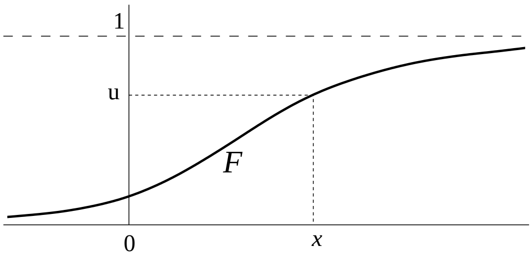
$$\mathbb{E}(r) = 0 \text{ en } \mathbb{V}\mathrm{ar}(r) = \frac{1}{n-1}$$

(Dit geldt ongeacht de waarden van de getallen x_i, y_i. De vorm van de verdeling van r hangt natuurlijk *wel* van deze waarden af).

11.8 De methode met de inverse verdelingsfunctie. Een stochast X met een strikt stijgende continue verdelingsfunctie F, kunnen we simuleren met

$$X = F^{-1}(U).$$

F^{-1} is de inverse functie van F. Het plaatje maakt dit duidelijk: Als $x = F^{-1}(u)$ (dus $u = F(x)$), dan is $X \leq x$ precies dan als $U \leq u$, en dit gebeurt met kans $\mathbb{P}[X \leq x] = \mathbb{P}[U \leq u] = u = F(x)$.



Om de gewenste uitdrukking $X = F^{-1}(U)$ te krijgen moeten we X oplossen uit de vergelijking $F(X) = U$. Je mag ook de vergelijking $F(X) = 1 - U$ nemen, want $1 - U$ is ook uniform-$(0,1)$-verdeeld.

Voorbeeld 1. De exponentiële verdeling. $F(x) = 1 - e^{-\lambda x}$. Op te lossen: $F(X) = 1 - U$, $e^{-\lambda X} = U$, dus

$$X = -\tfrac{1}{\lambda} \ln U.$$

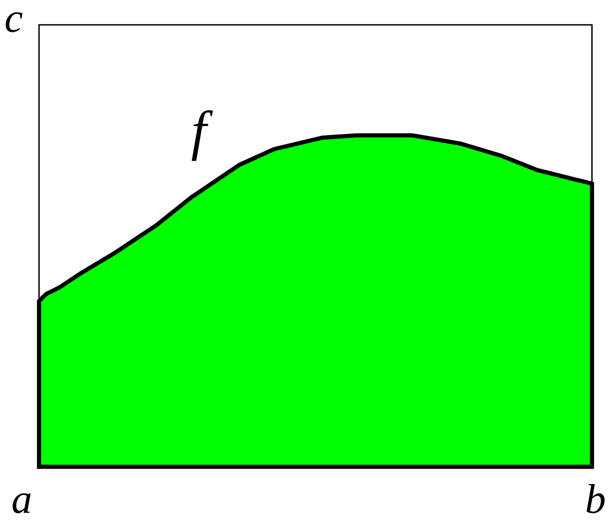
Voorbeeld 2. De Cauchy-verdeling. $F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x$, dichtheid $f(x) = \frac{1}{\pi(1+x^2)}$. Uit $F(X) = U$ vinden we:

$$X = \tan(\pi(U - \tfrac{1}{2})).$$

11.9 De wegwerpmethode (rejection method). Een continue verdelingsfunctie heeft niet altijd een eenvoudige inverse. De normale verdelingsfunctie, bijvoorbeeld. In zulke gevallen is de methode van paragraaf 11.8 minder praktisch. Er is gelukkig een andere algemene methode waarbij we alleen de dichtheidsfunctie nodig hebben.

We demonstreren het principe voor het geval van een dichtheidsfunctie die begrensd is en nul wordt buiten een interval (a, b).

Begrensd wil zeggen dat er een c is zodat $f(x) \le c$ voor alle x.
Een stochast Z met dichtheid f laat zich als volgt simuleren:
Met $X = a + (b-a)U_1$ en $Y = cU_2$ maken we een punt (X, Y) uniform verdeeld over de rechthoek $(a, b) \times (0, c)$. Valt dit punt in het gearceerde gebied onder de grafiek van f dan is het een goed punt en accepteren we de x-coördinaat als een gesimuleerde waarde van X. Punten die buiten het gearceerde gebied vallen gooien we weg. De x-coördinaten van de geaccepteerde punten zijn verdeeld volgens de dichtheidsfunctie f.



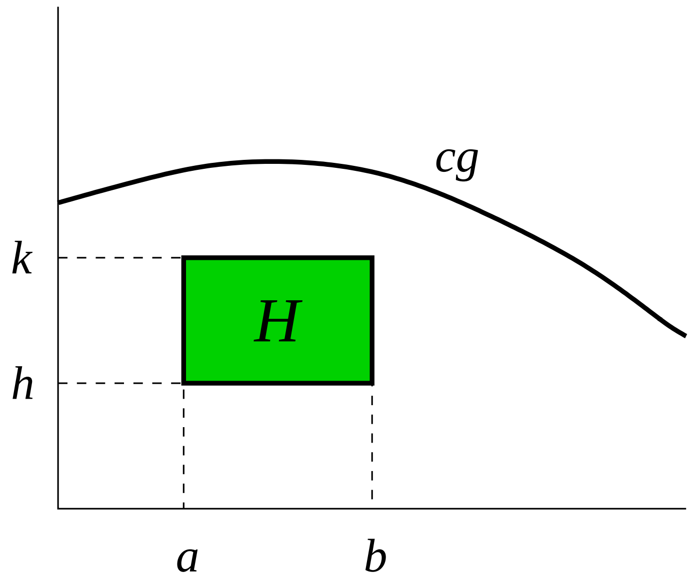
11.10 Wat te doen wanneer je een stochast wilt simuleren (normaal verdeeld, bijvoorbeeld) waarvan de dichtheidsfunctie f niet binnen een begrensde rechthoek te vangen is? Stel je hebt een stochast met een dichtheid g die je gemakkelijk kunt simuleren. Neem verder aan dat voor zekere $c > 0$ geldt: $f(x) \leq cg(x)$ voor alle x. We kunnen dan een punt (X, Y) genereren, uniform over

$$A = \{(x, y) : -\infty < x < \infty, 0 < y < cg(x)\}.$$

Als dat in orde is dan accepteren we het punt (X, Y) wanneer het in

$$B = \{(x, y) : -\infty < x < \infty, 0 < y < f(x)\}$$

valt.
De x-coördinaat van het geaccepteerde punt heeft dan dichtheid f, beweren we.
Hoe maken we een punt (X, Y) uniform over A? Wel, dat gaat zo:
Maak X volgend bekend recept met dichtheid g (bv. $X = tan(\pi(U_1 - \frac{1}{2}))$ heeft dichtheid $g(x) = (\pi(1 + x^2))^{-1}$).
Neem $Y = cg(x)U_2$, dan is (X, Y) uniform over A.



BEWIJS: Neem een rechthoekje $H = (a, b) \times (h, k)$ binnen A. Bij gegeven x is de kans dat Y in (h, k) valt gelijk aan $\frac{k-h}{cg(x)}$. Dit geïntegreerd van a tot b geeft:

$$\begin{aligned} \mathbb{P}[(X, Y) \in H] &= \int_a^b \frac{k - h}{cg(x)} g(x) dx \\ &= \frac{(k - h)(b - a)}{c} \\ &= \frac{\text{Opp}(H)}{\text{Opp}(A)}. \end{aligned}$$

Zie opgave 2 voor een toepassing op de normale verdeling. We geven echter eerst twee andere manieren om normaal verdeelde stochasten te simuleren.

11.11 De normale verdeling. De *eerste methode* berust op de centrale limietstelling: Als n groot is dan lijkt de verdeling van $S_n = U_1 + \ldots + U_n$ sterk op een normale verdeling (met $\mathbb{E}(S_n) = \frac{n}{2}$ en $\mathbb{V}\mathrm{ar}(S_n) = \frac{n}{12}$).
Reeds bij $n = 4$ is de benadering fraai. Door standaardiseren:

$$Z = \sqrt{3}(S_4 - 2)$$

krijg je een stochast die moeilijk te onderscheiden is van een standaardnormale.
Wil je het helemaal mooi maken, neem dan $n = 12$.

$$Z = S_{12} - 6$$

doet het fantastisch.

De *tweede methode* (de poolcoördinatenmethode) maakt tegelijk twee onafhankelijke standaardnormale stochasten X en Y zodat (X, Y) een punt in het vlak is met dichtheid

$$f(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}.$$

Dit gaat via de poolcoördinaten: straal R en hoek H:

$$X = R\cos H \text{ en } Y = R\sin H.$$

Het is duidelijk dat we H uniform $(0, 2\pi)$ moeten nemen.
Wat R betreft:

$$\mathbb{P}[R \le r] = \iint_C \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} dxdy,$$

waarbij

$$C = \{(x, y) : x^2 + y^2 \le r^2\}.$$

Met behulp van poolcoördinaten reken je die integraal uit:

$$\begin{aligned} \mathbb{P}[R \le r] &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^r e^{-\frac{1}{2}u^2} udud\varphi \\ &= 1 - e^{-\frac{1}{2}r^2}. \end{aligned}$$

Dus $\mathbb{P}[R^2 \le x] = \mathbb{P}[R \le \sqrt{x}] = 1 - e^{-\frac{1}{2}x}$. Dus R^2 is exponentieel verdeeld met parameter $\frac{1}{2}$: $R^2 \sim \mathrm{Exp}(\frac{1}{2})$.
Volgens het recept in paragraaf 11.8, voorbeeld 2, maken we R^2 dus met $r^2 = -2\ln U$

en dus $R = \sqrt{-2 \ln U}$.
Samenvattend: Met

$$H = 2\pi U_1 \text{ en } R = \sqrt{-2 \ln U_2}$$

en

$$Z_1 = R \cos H \text{ en } Z_2 = R \sin H$$

maak je tegelijk twee standaardnormale stochasten die nog onafhankelijk zijn ook.
Een voorbeeld. De gemiddelde lengte van jonge mannen in Nederland is 182 cm; standaardafwijking 8 cm. Zulke lengten kunnen we simuleren met $X = 182 + 8Z$, waarbij Z een standaardnormale stochast is. Hoe de verdeling van de lengten in een aselecte steekproef van 1000 personen uit deze verdeling er uit zou kunnen zien kunnen we door simulatie aan het licht brengen (gesimuleerde waarden worden afgerond op hele centimeters en in klassen van 1 cm geteld).

Opgaven

1. X is het aantal worpen met een dobbelsteen dat nodig is om iedere zijde minstens één keer boven te krijgen. Beschrijf een methode om X te simuleren, waarbij slechts 5 randomgetallen $U_1, \ldots, U_5$ gebruikt worden.

2. Een foute methode:
   In het boek *Understanding and Learning Statistics by Computer* van M.C.K. Yang en D.H. Robinson (World scientific 1986) wordt op blz. 142-143 een recept gegeven voor het trekken van een steekproef van 10 uit $\{1, \ldots, 20\}$.
   Het resultaat kan weergegeven worden door een rijtje $(Y_1, \ldots, Y_{20})$ met 10 enen en 10 nullen ($Y_i = 1$ als element i getrokken wordt).
   Yang en Robinson doen het zo: Te beginnen bij $k = 1$, om vast te stellen of $Y_k = 1$ of $Y_k = 0$, gooi je een munt op. Ga ook zo te werk voor $k = 2$, $k = 3$, enzovoorts. Om te zorgen dat je het juiste aantal enen en nullen krijgt houd je bij hoeveel je er al van ieder hebt. Zodra één van die aantallen de waarde 10 bereikt stop je en vul je de rij aan met de ontbrekende nullen of enen.
   Bijvoorbeeld, wanneer bij $k = 17$ de tiende één krijgt, dan zet je op de laatste drie plaatsen drie nullen.

   (a) Toon aan dat bij deze methode niet alle $\binom{20}{10}$ mogelijke steekproeven dezelfde kans krijgen. Bereken bijvoorbeeld

   $$\mathbb{P}[(X_{(1)}, \ldots, X_{(10)}) = (1, 2, \ldots, 10)]$$

   en

   $$\mathbb{P}[(X_{(1)}, \ldots, X_{(10)}) = (1, 3, 5, \ldots, 17, 19)]$$

   (zie voor de notatie paragraaf 11.4 over de lotto).

   (b) Is het wel waar dat ieder element dezelfde kans heeft om in de steekproef voor te komen?

   (c) Als $(Y_1, \ldots, Y_{20})$ weer de rij voorstelt met 10 enen en 10 nullen, bepaald zoals boven beschreven, wat is dan $\mathbb{P}[Y_1 = Y_2]$? En $\mathbb{P}[Y_{19} = Y_{20}]$? Hoe groot zou $\mathbb{P}[Y_i = Y_j]$ moeten zijn bij een aselecte steekproef?

3. Toppen tellen.
   $U_0, U_1, \ldots, U_n, U_{n+1}$ is een rij uniforme randomgetallen. We hebben een top bij k ($k = 1, \ldots, n$) als $U_k > U_{k-1}$ en $U_k > U_{k+1}$. Het aantal toppen in de rij is $T = T_1 + \ldots + T_n$, waarbij $T_k = 1$ als er een top is bij k (en 0 als dat niet zo is).

Een te grote of te kleine waarde van T kan een aanwijzing zijn dat er iets mis is met de randomgenerator. Bij een ideale randomgenerator geldt $\mathbb{E}(T) = \frac{n}{3}$, $\mathbb{V}\mathrm{ar}(T) = \frac{2(n+3)}{45}$ (zie onderdeel c).
Bijvoorbeeld als $n = 99$, dan: $\mathbb{E}(T) = 33$, $\mathbb{V}\mathrm{ar}(T) = 4{,}53$. De standaardafwijking is slechts 2,13.

(a) Wat is, bij gegeven n, de maximale waarde van T?

(b) Toon aan dat $\mathbb{E}(T) = \frac{n}{3}$.

(c) Bepaal $\mathbb{V}\mathrm{ar}(T)$.
Aanwijzing: $\mathbb{V}\mathrm{ar}(T) = \sum \mathbb{V}\mathrm{ar}(T_i) + \sum \mathbb{C}\mathrm{ov}(T_i, T_j) =$
$= n\mathbb{V}\mathrm{ar}(T_1) + 2(n-1)\mathbb{C}\mathrm{ov}(T_1, T_2) + 2(n-2)\mathbb{C}\mathrm{ov}(T_1, T_3) + \ldots$

(d) Schrijf een programma dat het aantal toppen telt in rijen randomgetallen van gegeven lengte. Laat ook de frequentie bijhouden waarmee de verschillende waarden van T optreden.

4. U_1, U_2, U_3 zijn onafhankelijke uniforme randomgetallen. Gerangschikt naar grootte: $U_{(1)}$, $U_{(2)}$, $U_{(3)}$.

(a) Bepaal de verdelingsfunctie en de dichtheidsfunctie van $U_{(k)}$, $k = 1, 2, 3$. Bepaal ook $\mathbb{E}(U_{(k)})$.

(b) * Hoe wordt dit bij $U_{(1)}, \ldots, U_{(n)}$?

5. $g(x) = \frac{1}{2}e^{-|x|}$ is de dichtheidsfunctie van de zogenaamde *Laplace-verdeling*. Er zijn verschillende manieren om deze verdeling te simuleren.

(a) Laat zien dat de methode met de inverse verdelingsfunctie leidt tot:

$$X = \mathrm{sgn}(1 - 2U) \ln(1 - |2U - 1|)$$

($\mathrm{sgn}(x)$, het teken van x, is gedefinieerd door $\mathrm{sgn}(x) = 1$ als $x > 0$, $\mathrm{sgn}(x) = -1$ als $x < 0$, en $\mathrm{sgn}(0) = 0$).

(b) Een eenvoudiger formule krijg je als je twee randomgetallen gebruikt: $X = \mathrm{sgn}(U_1) \ln U_2$. Laat zien dat dit juist is.

(c) Laat zien dat

$$X = \ln \frac{U_2}{U_1}$$

ook deze verdeling genereert.

6. De normale verdeling met de verwerpingsmethode.

(a) Bepaal de kleinste c zo dat

$$e^{-\frac{1}{2}x^2} \leq ce^{-|x|}$$

voor alle x (uitkomst: $c = \sqrt{e}$).

(b) Laat zien dat je met de volgende methode een standaardnormale stochast simuleert:

$$\text{Maak } X = \ln \tfrac{U_1}{U_2} \text{ en } Y = e^{\frac{1}{2}-|x|} \cdot U_3.$$
$$\text{Accepteer } X \text{ als } Y < e^{-\frac{1}{2}x^2} \text{ (dus als } U_3 < e^{|x|-\frac{1}{2}(1+x^2)}\text{).}$$

(c) Bepaal de acceptatiekans p bij de bovenstaande methode (het aantal randomgetallen dat gemiddeld nodig is om een X-waarde te produceren is dan $\frac{3}{p}$).

7. Beschrijf een methode om een punt (X, Y) te genereren, uniform verdeeld over $C = \{(x, y) : x^2 + y^2 \leq a^2\}$

(a) Met het wegwerpprincipe;

(b) Met behulp van poolcoördinaten: $X = R\cos H$ en $Y = R\sin H$.

Hoofdstuk 12

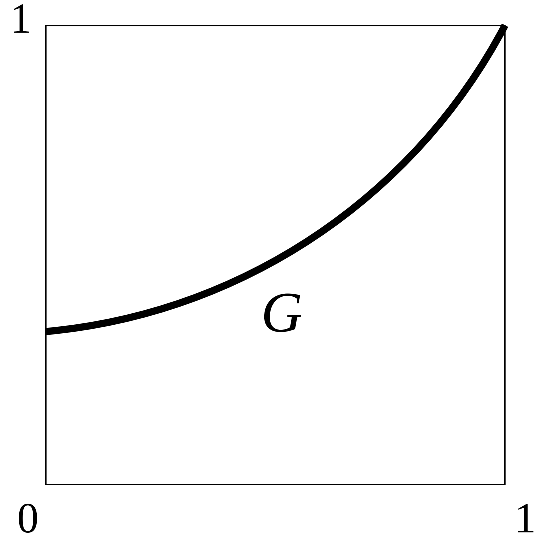
Kansgenererende functies

12.1 Definitie. Als een stochast X alleen natuurlijke getallen als waarden aanneemt, dan kunnen we bij X een functie $G = G_X : [0,1] \to [0,1]$ maken op de volgende manier:

$$G(s) = \mathbb{E}(s^X) = \sum_{n=0}^{\infty} s^n \mathbb{P}[X{=}n].$$

Deze *kansgenererende functie* G_X is een bijzonder handig instrument voor het berekenen en bestuderen van de kansverdeling van X.

12.2 Hieronder geven we enkele **eigenschappen:**



(a) G is stijgend en convex;

(b) $G(0) = \mathbb{P}[X = 0]$ en $G(1) = 1$;

(c) $\mathbb{E}(X) = G'(1)$ (eventueel beide ∞);

(d) $\mathbb{V}\mathrm{ar}(X) = G''(1) + G'(1) - G'(1)^2$;

(e) $X \perp\!\!\!\perp Y \Rightarrow G_{X+Y}(s) = G_X(s) \cdot G_Y(s)$.

Bewijs:

(a) Omdat $s^n \leq t^n$ voor $0 < s < t$ en $n \in \mathbb{N}$, is G stijgend; omdat $s \mapsto s^n$ convex is voor alle $n \in \mathbb{N}$, is G convex.

(b) Dit zie je direct.

(c) Als $\mathbb{E}(X) < \infty$, dan is

$$\begin{aligned} G'(1) &= \lim_{s\uparrow 1} \frac{G(1) - G(s)}{1 - s} \\ &= \lim_{s\uparrow 1} \sum_{n=0}^{\infty} \frac{1 - s^n}{1 - s} \cdot \mathbb{P}[X{=}n] \\ &= \lim_{s\uparrow 1} \sum_{n=0}^{\infty} \left(1 + s + s^2 + \ldots + s^{n-1}\right) \mathbb{P}[X{=}n] \\ &= \sum_{n=0}^{\infty} n\mathbb{P}[X{=}n] \\ &= \mathbb{E}(X). \end{aligned}$$

Ga zelf na dat de limiet $s\uparrow 1$ hier achter het somteken uitgevoerd mag worden[1].

(d)

$$\begin{aligned} G''(1) &= \frac{d^2}{ds^2} \sum_{n=0}^{\infty} s^n \mathbb{P}[X{=}n] \bigg|_{s=1} \\ &= \sum_{n=2}^{\infty} n(n{-}1)\mathbb{P}[X{=}n] \\ &= \mathbb{E}(X^2) - \mathbb{E}(X). \end{aligned}$$

Dus als $G''(1) < \infty$, dan is ook $G'(1) < \infty$, en

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = G''(1) + \mathbb{E}(X) - \mathbb{E}(X)^2 = G''(1) + G'(1) - G'(1)^2.$$

(e) Als X en Y onafhankelijk zijn, dan is

$$G_{X+Y}(s) = \mathbb{E}(s^{X+Y}) = \mathbb{E}(s^X \cdot s^Y) = \mathbb{E}(s^X)\mathbb{E}(s^Y) = G_X(s)G_Y(s).$$

12.3 Kansgenererende functies van bekende verdelingen. Laten we nu eens de kansgenererende functies uitrekenen van de belangrijkste geheelwaardige stochasten die we tot dusver zijn tegengekomen.

De alternatieve verdeling. Als $X \sim \mathrm{Alt}(p)$, dan is

$$G_X(s) = \mathbb{P}[X{=}0] \cdot s^0 + \mathbb{P}[X{=}1] \cdot s^1 = q + ps.$$

[1] In het vervolg zullen we zulke limieten en oneindige sommen zonder veel omhaal verwisselen. Zolang de coëfficiënten van s^n in de som positief zijn, is dit toegestaan op grond van de stelling van Abel. Een voorbeeld volgt bij (d)

De binomiale verdeling. Als $Y \sim \text{Bin}(n,p)$, dan is

$$G_Y(s) = \sum_{k=0}^{n} \mathbb{P}[Y{=}k]s^k = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} s^k = (q+ps)^n.$$

Merk op dat $G_Y(s) = G_X(s)^n$. Dit hadden we direct kunnen zeggen op grond van eigenschap (e) hierboven. Immers de som van n onafhankelijke $\text{Alt}(p)$-verdeelde stochasten is $\text{Bin}(n,p)$-verdeeld.

De Poisson-verdeling. Als $Z \sim \text{Poisson}(\lambda)$, dan is

$$G_Z(s) = \sum_{n=0}^{\infty} \mathbb{P}[Z{=}n]s^n = \sum_{n=0}^{\infty} \left(\frac{\lambda^n}{n!} e^{-\lambda} \right) s^n = e^{\lambda(s-1)}.$$

Ook deze kansgenererende functie hadden we uit de voorgaande af kunnen leiden: Vul in G_Y in: $p = \frac{\lambda}{n}$ en neem de limiet $n \to \infty$:

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n} + \frac{\lambda s}{n}\right)^n = e^{\lambda(s-1)}.$$

12.4 Een toepassing: Veel dobbelstenen. Wat is de kans op m ogen bij een worp met n dobbelstenen?

In het voorgaande hebben we gezien dat, als X_1 kansgenererende functie G_1 heeft, en $X_1, X_2, \ldots, X_n$ onafhankelijk en gelijk verdeeld zijn, $S_n = X_1 + \ldots + X_n$ de kansgenererende functie

$$G_n(s) = G_1(s)^n$$

heeft.

Laten we voor X_1 nu het aantal ogen van een dobbelsteen nemen. Dan is

$$G_1(s) = \tfrac{1}{6}\left(s + s^2 + s^3 + s^4 + s^5 + s^6\right).$$

Het totale aantal ogen bij een worp met n dobbelstenen heeft dus kansgenererende functie $G_n(s) = G_1(s)^n$.

De kans op m ogen is de coëfficiënt van s^m hierin. Op het eerste gezicht is het bepalen van deze coëfficiënt net zo moeilijk als de oorspronkelijke vraag. Nu kunnen we echter iets doen: We kunnen schrijven

$$G_1(s) = \frac{s}{6} \cdot \frac{1-s^6}{1-s},$$

dus

$$G_n(s) = \left(\frac{s}{6}\right)^n \cdot \frac{(1-s^6)^n}{(1-s)^n}.$$

De factor $(1-s^6)^n$ kunnen we uitschrijven met de binomiaalformule van Newton, en de factor $(1-s)^{-n}$ kunnen we in een machtreeks ontwikkelen door de formule

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

links en rechts $n-1$ maal te differentiëren, en te delen door $(n-1)!$. Er komt

$$\begin{aligned}
\frac{1}{(1-s)^n} &= \frac{1}{(n-1)!}\frac{d^n}{ds^n}\sum_{k=0}^{\infty} s^k \\
&= \frac{1}{(n-1)!}\sum_{k=n-1}^{\infty} k(k-1)\cdots(k-n+2)s^{k-n+1} \\
&= \sum_{k=n-1}^{\infty}\binom{k}{n-1}s^{k-n+1} \\
&= \sum_{l=0}^{\infty}\binom{l+n-1}{n-1}s^l.
\end{aligned}$$

Resultaat:

$$G_n(s) = \frac{s^n}{6^n}\left(\sum_{k=0}^{n}\binom{n}{k}(-s^6)^k\right)\left(\sum_{l=0}^{\infty}\binom{l+n-1}{n-1}s^l\right).$$

De coëfficiënt van s^m hierin vinden we door die termen bijeen te rapen waarvoor $n+6k+l=m$:

$$\mathbb{P}[S_n{=}m] = \frac{1}{6^n}\sum_{k=0}^{\left[\frac{m-n}{6}\right]}(-1)^k\binom{n}{k}\binom{m-6k-1}{n-1}.$$

Zo is bijvoorbeeld de kans om 24 ogen te gooien met 8 dobbelstenen gelijk aan

$$\begin{aligned}
\mathbb{P}[S_8{=}24] &= \frac{1}{6^8}\left(\binom{8}{0}\binom{23}{7}-\binom{8}{1}\binom{17}{7}+\binom{8}{2}\binom{11}{7}\right) \\
&= \frac{98813}{1679616} \sim 0,05883.
\end{aligned}$$

Het valt niet mee, dit resultaat door direct tellen te verkrijgen.

12.5 Een tweede toepassing: Het Sinterklaasprobleem. Een gezelschap van n personen trekt lootjes om te bepalen wie er voor wie een Sinterklaassurprise zal maken. We willen de kansverdeling weten van M, het aantal deelnemers dat hun eigen naam trekt (De kans dat $M=0$ heb je al berekend in opgave 2.5.).
Onze kansruimte Ω_n is de verzameling van de $n!$ permutaties $\omega:\{1,2,\ldots,n\}\to\{1,2,\ldots,n\}$, waarbij $\omega(j)=i$ staat voor: persoon j trekt de naam van persoon i. Elke permutatie heeft evenveel kans; de kansmaat $A\mapsto\frac{\#(A)}{n!}$ geven we aan met $\mathbb{P}_n$, en middelen over $(\Omega_n,\mathbb{P}_n)$ geven we aan met $\mathbb{E}_n$.

Verder beschouwen we bij elke deelverzameling K van ons gezelschap (zeg $K \subset \{1, 2, \ldots, n\}$) de gebeurtenis A_K dat deze personen allemaal hun eigen naam trekken (geheel afgezien van wat de rest doet):

$$A_K = \{\omega \in \Omega_n | \forall_{j \in K} : \omega(j) = j\}.$$

Merk op dat $\mathbb{P}_n(A_K) = \frac{(n-\#(K))!}{n!}$, omdat er voor de overige $n - \#(K)$ personen $(n-\#(K))!$ mogelijkheden overblijven.
Zij verder $V(\omega)$ de vaste-punten-verzameling van ω:

$$V(\omega) = \{j \in \{1, 2, \ldots, n\} | \omega(j) = j\}.$$

We zijn geïnteresseerd in de stochast

$$M(\omega) = \#(V(\omega)).$$

We berekenen nu de kansgenererende functie G_n van M (bij deelname van n personen) met behulp van de volgend truc. Als een verzameling V m elementen heeft, dan is (ga na!)

$$(1+x)^m = \sum_{k=0}^{m} \binom{m}{k} x^k = \sum_{K \subset V} x^{\#(K)}.$$

Laten we dit los op de vaste-punten-verzameling $V(\omega)$ van $\omega \in \Omega_n$, dan vinden we

$$(1+x)^{M(\omega)} = \sum_{K \subset V(\omega)} x^{\#(K)} = \sum_{K \subset \{1,2,\ldots,n\}} \mathbf{1}_{A_K}(\omega) x^{\#(K)}.$$

En dus is

$$\begin{aligned} G_n(1+x) &= \mathbb{E}\left((1+x)^M\right) \\ &= \sum_{K \subset \{1,2,\ldots,n\}} \mathbb{P}(A_K) x^{\#(K)} \\ &= \sum_{k=0}^{n} \binom{n}{k} \cdot \frac{(n-k)!}{n!} x^k \\ &= \sum_{k=0}^{n} \frac{x^k}{k!} \end{aligned}$$

Conclusie:

$$\boxed{G_n(s) = \sum_{k=0}^{n} \frac{(s-1)^k}{k!}.}$$

De kans dat niemand zijn eigen naam trekt, vinden we terug als

$$\mathbb{P}_n[M{=}0] = G_n(0) = \sum_{k=0}^{n} \frac{(-1)^k}{k!}.$$

Ook zien we dat, voor $n \geq 2$,

$$\mathbb{E}_n(M) = G_n'(1) = 1;$$

$$\mathbb{V}\mathrm{ar}_n(M) = G_n''(1) + G_n'(1) - G_n'(1)^2 = 1.$$

Verder blijkt dat $\lim_{n\to\infty} G_n(s) = e^{s-1}$, de kansgenererende functie van de Poisson-verdeling met parameter 1.
Tenslotte kunnen we de hele kansverdeling van M uit onze formule voor G_n aflezen:

$$\begin{aligned}
\mathbb{P}_n[M{=}m] &= \sum_{k=0}^{n} \frac{1}{k!}\Big(\text{coëfficiënt van } s^m \text{ in } (s{-}1)^k\Big) \\
&= \sum_{k=m}^{n} \frac{1}{k!} \cdot (-1)^{k-m} \binom{k}{m} \\
&= \frac{1}{m!} \sum_{l=0}^{n-m} \frac{(-1)^l}{l!} \\
&= \frac{1}{m!} \left(1 - 1 + \frac{1}{2} - \frac{1}{6} + \frac{1}{24} - \ldots \pm \frac{1}{(n{-}m)!}\right).
\end{aligned}$$

Opgaven

1. Zij N de wachttijd op het eerste succes in een rij onafhankelijke pogingen met slaagkans $p \in (0,1)$ en faalkans $q = 1 - p$.

   (a) Bereken de kansgenererende functie G_N van N.

   (b) Bereken hieruit $\mathbb{E}(N)$ en $\mathbb{V}\mathrm{ar}(N)$ (vergelijk met eerdere resultaten!).

   (c) Zij M de wachttijd op het tweede succes.
   Laat zien dat $G_M(s) = G_N(s)^2$.
   Kun je dit resultaat ook op een andere manier verklaren?

2. Een bepaalde soort bacteriën plant zich voort volgens het volgende schema: ieder uur loot ieder individu tussen twee mogelijkheden: zich delen in twee individuen (kans p) of sterven (kans $q = 1 - p$). Op tijdstip 0 is in ons kweekschaaltje één bacterie van deze soort aanwezig. Het aantal individuen na k uur noemen we: N_k.

   (a) Bereken de kansgenererende functies G_0, G_1 en G_2 van respectievelijk N_0, N_1, en N_2.

   (b) Laat zien dat voor alle $k \in \mathbb{N}$ geldt:

   $$G_{k+1}(s) = G_k(G_1(s))$$

   door te conditioneren naar N_k.

   (c) Toon aan dat de kans op uitsterven van de bacteriekolonie in ons kweekschaaltje gelijk is aan

   $$\lim_{k\to\infty} G_k(0) = \lim_{k\to\infty} \underbrace{G_1 \circ G_1 \circ \ldots \circ G_1}_{k \text{ keer}}(0).$$

   (d) Teken de grafiek van G_1. Laat in het plaatje zien wat de uitsterfkans is. Maak onderscheid tussen de gevallen $p > \frac{1}{2}$ en $p \leq \frac{1}{2}$.

Hoofdstuk 13

Markovketens

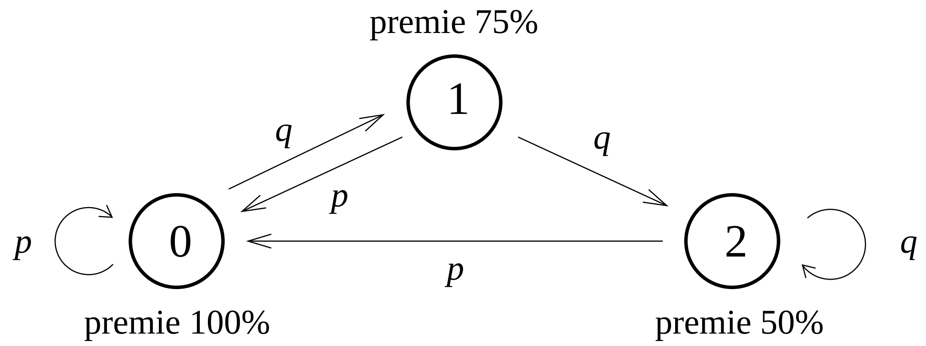
13.1 Inleiding. Een rij van verschijnselen kan meer of minder systematisch verlopen.
Het minst systematisch is een rij onafhankelijke proeven zoals bijvoorbeeld een rij worpen met een dobbelsteen. Om het resultaat van de volgende worp te voorspellen hebben we geen enkel houvast aan hetgeen de voorgaande worpen hebben opgeleverd.
Wat meer systeem vind je in een rij letters van een tekst. Zelfs als de taal waarin die tekst geschreven is je onbekend is, zie je toch dat na bepaalde letters bepaalde andere letters meer of minder waarschijnlijk zijn.
Dergelijke voorbeelden brachten Markov (Andrej Andrejewitsj Markov, 1856-1922) ertoe het volgende model voor te stellen:
Er is een verzameling van mogelijke toestanden (de *toestandsruimte* S, bijvoorbeeld de letters van het alfabet). Als het systeem zich op tijdstip n (nu) in toestand x bevindt, dan zal het zich op tijdstip $n+1$ in toestand y bevinden met kans $P(x, y)$. De overgangskans $P(x, y)$ hangt af van x, de toestand nu, maar niet van het verleden (toestanden op tijdstippen vóór n).
Deze aanname (de Markov-aanname) geldt niet zo goed voor de letters van een tekst (Als we onder de toestand niet de laatste letter maar de laatste drie letters verstaan, dan klopt het al wat beter). Er zijn echter veel voorbeelden waarbij de Markov-aanname een redelijke veronderstelling is. Er is een uitgebreide literatuur over de theorie en de toepassingen van het Markov-model. In dit hoofdstuk geven we een korte, enigszins intuïtieve inleiding.

13.2 Wat voorbeelden.

(a) No-claim-korting. Als je je auto verzekert bij General Accident N.V., dan betaal je voor het eerste jaar de volle premie (toestand 0). Heb je in een jaar geen schade gemeld dan betaal je voor het volgende jaar slechts 75% (25%

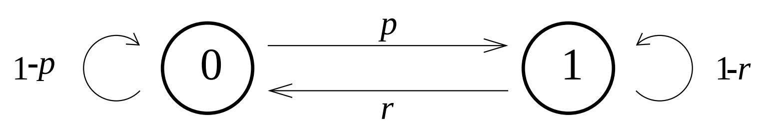
korting). Je bent dan in toestand 1. Heb je het jaar daarop weer geen schade te melden dan krijg je de maximale korting van 50% (toestand 2).

We nemen aan dat er ieder jaar een kans p is op meldenswaardige schade en dat je na een schademelding weer de volle premie moet betalen (je komt dan weer in toestand 0). Het volgend schema geeft de situatie weer ($q = 1 - p$):



We hebben hier een systeem met 3 toestanden. Toestandsruimte $S = \{0, 1, 2\}$. De pijlen geven de mogelijke overgangen aan. De bijbehorende overgangskansen staan er bij.

(b) Twee toestanden. Bij dit eenvoudige model met $S = \{0, 1\}$ hoort het volgende schema:

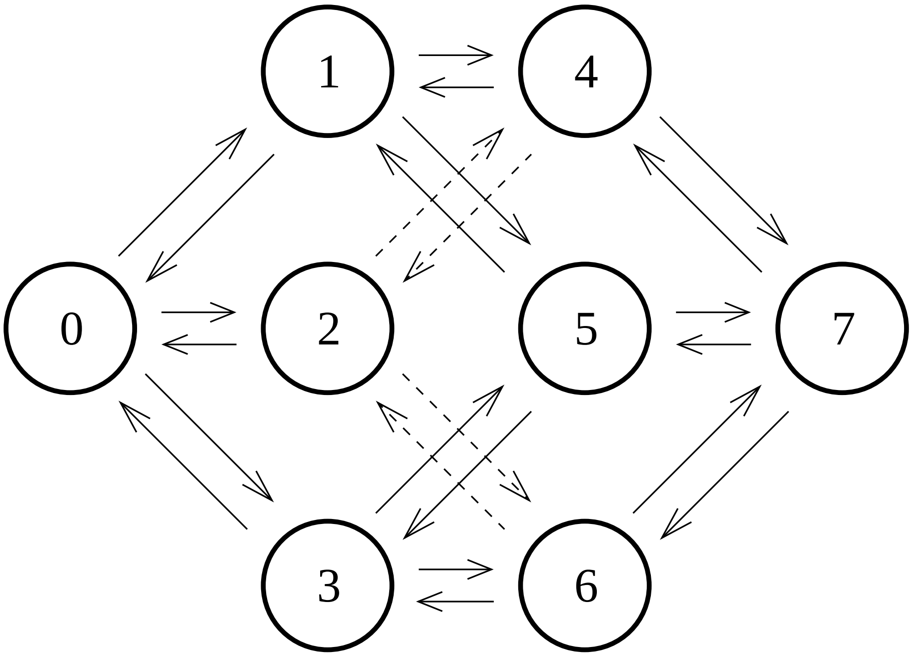


Er zijn allerlei interpretaties mogelijk:

1. Een apparaat (machine, voertuig,...) kan in goede conditie zijn: toestand 0, of niet: toestand 1. Er is iedere dag een kans p dat er een mankement zal optreden. Als dat gebeurt gaan we er direct mee naar de reparatieafdeling en dan is er iedere dag een kans r dat we het weer goed terugkrijgen.
2. Het weer. Toestand 0: droog. Toestand 1: regen. Zie hoofdstuk 4, opgave 11, pagina 91.

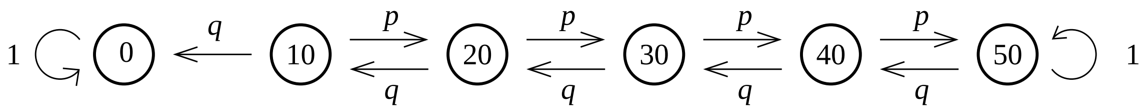
(c) Toevalswandeling op de hoekpunten van een kubus.
Maak via de ribben van een kubus een wandeling langs hoekpunten. Iedere keer wandel je van een hoekpunt naar een willekeurig aangrenzend hoekpunt (elk

aangrenzend hoekpunt heeft evenveel kans). $S = \{0, 1, \ldots, 7\}$. Langs iedere ribbe van de kubus lopen pijlen die de overgangskansen (steeds $\frac{1}{3}$) aangeven.



(d1) Gokken.

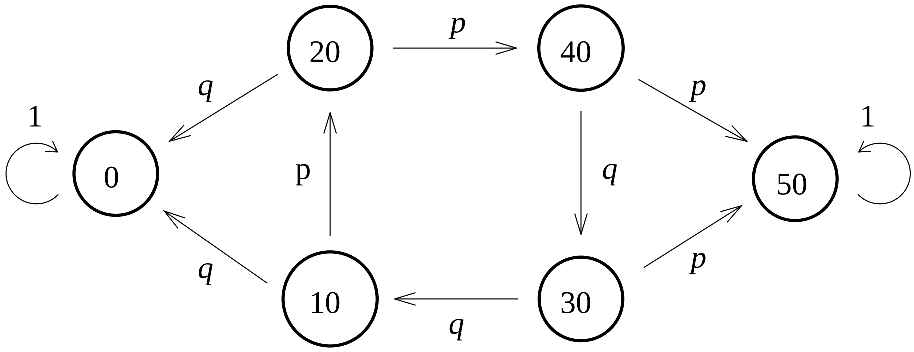
Je gaat het casino binnen met 20 gulden op zak. Je bent van plan te stoppen met spelen wanneer je 50 gulden hebt, of niets. Je zet telkens een tientje in op rood. Met kans p win je er een tientje bij; met kans $q = 1 - p$ ben je dat tientje kwijt.



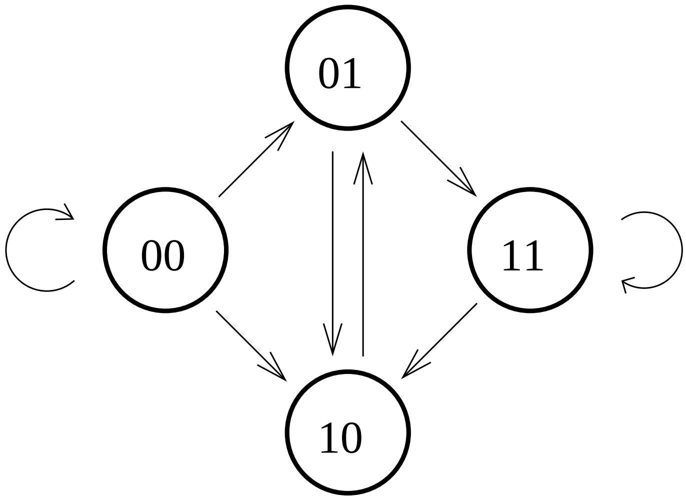
Hoe groot is de kans dat je de 50 gulden haalt?

(d2) Bold Play.

Volgens Dubins en Savage (How to Gamble if You Must 1965) heb je meer kans de 50 te halen wanneer je dat zo snel mogelijk probeert te bereiken (Hoe langer je speelt hoe meer het casino aan je verdient). Dus je zet meteen je twee tientjes in en vervolgens (zolang je geld hebt) zet je net zoveel in als nodig is om 50 te bereiken, of al je geld, als dat niet in één keer kan. Schema:



(e) In een rij worpen met een munt letten we steeds op de laatste twee uitkomsten. Schrijf 0 voor munt, 1 voor kop. $S = \{00, 01, 10, 11\}$ Schema:



Bij iedere pijl hoort een overgangskans van $\frac{1}{2}$.
Teken zelf een dergelijk schema voor het geval dat de drie laatste uitkomsten de toestand bepalen ($S = \{000, 001, \ldots, 111\}$).

13.3 Grondbeginselen, terminologie.

De toestandsruimte S**.** We beperken ons hier tot een eindige of hoogstens aftelbaar oneindige verzameling toestanden. Nummeren we de toestanden $0, 1, 2 \ldots$ dan krijgen we $S = \{0, 1, \ldots, M\}$ of $S = \{0, 1, 2, \ldots\}$.

Discrete tijd. Verandering van toestand kan optreden tussen de tijdstippen 0, 1, 2,.... De toestand op tijdstip n geven we aan met X_n. De begintoestand wordt aangegeven door X_0. De rij stochasten $X_0, X_1, \ldots$ beschrijft de evolutie van het systeem (zoiets heet: stochastisch proces).

Overgangskansen. Vanuit toestand x kan het systeem overgaan naar toestand y met kans $P(x,y)$:

$$P(x,y) = \mathbb{P}(X_{n+1} = y | X_n = x).$$

De overgangskansen vormen een matrix P, de *overgangsmatrix.* In voorbeeld (a) hebben we bijvoorbeeld

$$P = \begin{pmatrix} P(0,0) & P(0,1) & P(0,2) \\ P(1,0) & P(1,1) & P(1,2) \\ P(2,0) & P(2,1) & P(2,2) \end{pmatrix} = \begin{pmatrix} p & q & 0 \\ p & 0 & q \\ p & 0 & q \end{pmatrix}$$

Merk op dat voor iedere $x \in S$ geldt:

$$\sum_{y \in S} P(x,y) = 1.$$

Beginverdeling. De kansverdeling van X_0 geven we aan met π_0 (dit is een rijvector met elementen $\pi_0(x)$, $x \in S$). $\pi_0(x) = \mathbb{P}[X_0{=}x]$.
Beginverdeling en overgangsmatrix leggen de Markovketen vast. De kansverdeling van X_1 (aan te duiden met π_1) krijgen we als volgt:

$$\pi_1(x) = \mathbb{P}[X_1{=}x] = \sum_{u \in S} \pi_0(u) P(u,x).$$

In matrixnotatie:

$$\pi_1 = \pi_0 P.$$

Net zo krijgen we de verdeling π_2 van X_2 uit die van X_1:

$$\mathbb{P}[X_2{=}x] = \sum_{u \in S} \pi_1(u) P(u,x)$$

en dus

$$\pi_2 = \pi_1 P = (\pi_0 P) P = \pi_0 P^2.$$

Meerstapsovergangskansen. Het element $P^2(x,y)$ van de matrix P^2 geeft de kans op in twee stappen van x naar y te komen:

$$P^2(x,y) = \sum_{z \in S} P(x,z) P(z,y).$$

De regels van de matrixvermenigvuldiging en die van de kansrekening zijn hier helemaal met elkaar in overeenstemming.
Fundamenteel voor de theorie van de Markovketens is de volgende formule:

$$\mathbb{P}[X_0{=}x_0, X_1{=}x_1, \ldots, X_n{=}x_n] = \pi_0(x) P(x_0,x_1) \cdots P(x_{n-1},x_n).$$

Deze formule berust op de *Markovaanname*, die we zo kunnen formuleren: Voor ieder rijtje toestanden $x_0, \ldots, x_n$ geldt:

$$\mathbb{P}(X_n{=}x_1|X_0{=}x_0, \ldots, X_{n-1}{=}x_{n-1}) = \mathbb{P}(X_n{=}x_n|X_{n-1}{=}x_{n-1}) = P(x_{n-1}, x_n).$$

Neem bijvoorbeeld $n = 2$, dan geldt altijd (zie de rekenregels voor voorwaardelijke kansen)

$$\mathbb{P}[X_0{=}x_0, X_1{=}x_1, X_2{=}x_2] =$$

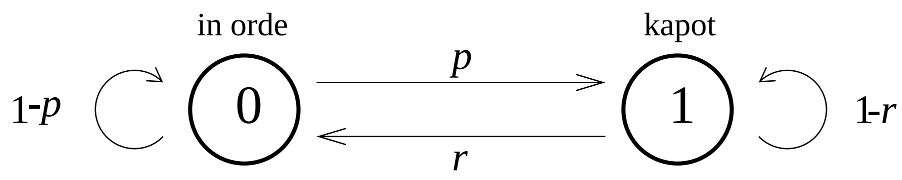
$$\mathbb{P}[X_0{=}x_0]\mathbb{P}(X_1{=}x_1|X_0{=}x_0)\mathbb{P}(X_2{=}x_2|X_0{=}x_0, X_1{=}x_1).$$

Dankzij de Markovaanname kunnen we de laatste factor

$$\mathbb{P}(X_2{=}x_2|X_0{=}x_0, X_1{=}x_1)$$

vervangen door $\mathbb{P}(X_2{=}x_2|X_1{=}x_1) = P(x_1, x_2)$.

Kansen bij een gegeven begintoestand. In het bijzondere geval dat de beginverdeling π_0 geconcentreerd is in één enkele toestand x zullen we de kans op iets, zeg A, aangeven met $\mathbb{P}_x(A)$, dus: $\mathbb{P}_x(A) = \mathbb{P}(A|X_0{=}x)$. Zo krijg je bijvoorbeeld $\mathbb{P}_x(X_1{=}y) = P(x, y)$ en $\mathbb{P}_x(X_n{=}y) = P^n(x, y)$.



13.4 De keten met twee toestanden.(zie interpretatie 2 bij voorbeeld b uit paragraaf 13.2)

In dit geval kunnen we de kans $\mathbb{P}[X_n{=}0]$ dat de machine na n dagen nog (of weer) heel is expliciet berekenen, gegeven de kans $\pi_0(0) = \mathbb{P}[X_0{=}0]$. Dat gaat zo:

$$\begin{aligned}
\mathbb{P}[X_{n+1}{=}0] &= \mathbb{P}[X_n{=}0]P(0,0) + \mathbb{P}[X_n{=}1]P(1,0) \\
&= \mathbb{P}[X_n{=}0](1-p) + (1 - \mathbb{P}[X_n{=}0])r \\
&= (1-p-r)\mathbb{P}[X_n{=}0] + r.
\end{aligned}$$

Voor het gemak schrijven w: $a_n = \mathbb{P}[X_n{=}0]$; $c = 1 - p - r$. Dan staat er:

$$a_{n+1} = ca_n + r.$$

Deze recursie heeft als vaste waarde $\frac{r}{1-c}$ (als $a_n = \frac{r}{1-c}$ dan is ook $a_{n+1} = \frac{r}{1-c}$). Voor de afwijking van a_n t.o.v. deze vaste waarde geldt:

$$a_{n+1} - \frac{r}{c-1} = c\left(a_n - \frac{r}{1-c}\right)$$

waaruit volgt dat

$$a_n = \frac{r}{1-c} + c^n \left(a_0 - \frac{r}{1-c}\right).$$

Met $1-c = p+r$, $a_n = \mathbb{P}[X_n{=}0]$ en $a_0 = \pi_0(0)$ wordt dit:

$$\mathbb{P}[X_n{=}0] = \frac{r}{p+r} + c^n \left(\pi_0(0) - \frac{r}{p+r}\right) \tag{1}$$

Op dezelfde manier vinden we voor $\mathbb{P}[X_n{=}1]$ (verwissel r en p, en ook 1 en 0):

$$\mathbb{P}[X_n{=}1] = \frac{p}{p+r} + c^n \left(\pi_0(1) - \frac{p}{p+r}\right) \tag{2}$$

Schrijven we $\pi = \left(\frac{r}{p+r}, \frac{p}{p+r}\right)$ en, zoals afgesproken, π_n voor de verdeling van X_n dan kunnen de formules (1) en (2) in vectornotatie gecombineerd worden tot

$$\pi_n = \pi + c^n(\pi_0 - \pi) \tag{3}$$

Hieruit blijkt het volgende:

(a) Bij beginverdeling $\pi_0 = \pi$ krijgen we: $\pi_n = \pi$ voor iedere n. De verdeling π noemt men de *evenwichtsverdeling* (ook wel *stationaire verdeling*). Deze verdeling voldoet aan

$$\pi P = \pi.$$

In lineaire-algebra-taal: de rijvector π is een links-eigenvector met eigenwaarde 1 voor de matrix P (Een andere eigenvector is $(1,-1)$. Deze heeft eigenwaarde $c = 1-p-r$. Ga na dat $\pi_0 - \pi$ een veelvoud is van $(1,-1)$).

(b) Convergentie naar de evenwichtsverdeling. Omdat $|c| = |1-p-r| < 1$ (tenzij $p = r = 0$ of $p = r = 1$) geldt dat $\pi_n \to \pi$ als $n \to \infty$. Dus, onafhankelijk van de beginverdeling, zullen we het systeem op den duur in toestand x aantreffen met kans $\pi(x)$ in overeenstemming met de evenwichtsverdeling.

(c*) Voor de n-staps overgangsmatrix P^n krijgen we:

$$P^n = \frac{1}{p+r}\left[\begin{pmatrix} r & p \\ r & p \end{pmatrix} + c^n \begin{pmatrix} p & -p \\ -r & r \end{pmatrix}\right].$$

Afleiding: De bovenste rij van P^n:

$$(P^n(0,0), P^n(0,1)) = (\mathbb{P}_0(X_n{=}0), \mathbb{P}_0(X_n{=}1))$$

krijgen we door in formule (3) in te vullen: $\pi_0 = (1,0)$. De onderste rij van P^n wordt gegeven door formule (3) met $\pi_0 = (0,1)$.
We zien dat P^n convergeert naar een matrix met rijen die gelijk zijn aan de evenwichtsverdeling.

13.5 Praktische betekenis van de evenwichtsverdeling. Stel dat in een fabriek N machines in gebruik zijn die zich ieder gedragen volgens het schema van voorbeeld (b) uit paragraaf 13.2 met kans $p = 0{,}1$ om kapot te gaan en $r = 0{,}5$ om, kapotzijnd, weer na één dag gerepareerd te worden: Van de machines in de reparatieafdeling wordt gemiddeld per dag de helft weer in goede staat terug gebracht.
Na zekere tijd zijn $n(0)$ machines in bedrijf (toestand 0) en $n(1)$ zijn er kapot (toestand 1). We kunnen van evenwicht spreken als de in- en uitgaande stromen van de fabriekshal naar de reparatieafdeling en terug even groot zijn. Dus het aantal machines dat op een dag kapot gaat (gemiddeld $n(0)p$) is gelijk aan het aantal machines dat op een dag gerepareerd terugkomt (gemiddeld $n(1)r$).
De verhouding $n(0) : n(1) = r : p$ is dan in overeenstemming met de evenwichtsverdeling. Als $p = 0{,}1$ en $r = 0{,}5$ dan zal gemiddeld een fractie $\frac{p}{r+p} = \frac{1}{6}$ van de machines in reparatie zijn.
In voorbeeld (a) (verzekering met no-claim-korting) zegt de evenwichtsverdeling ons iets over welk deel van de verzekerden op den duur hoeveel procent korting krijgt.

Opgave 1. Bepaal de evenwichtsverdeling bij voorbeeld (a) als $p = 0{,}2$.

Opgave 2. Voorbeeld (e). Schrijf de overgangsmatrix P op. Bepaal P^n voor $n = 2, 3, \ldots$, en de evenwichtsverdeling.

13.6 Evenwicht en convergentie naar evenwicht. In deze paragraaf bespreken we een paar fundamentele stellingen over evenwichtsverdelingen bij eindige Markovketens (Markovketens met eindige toestandsruimte). De uitspraken van deze stellingen zijn eenvoudig. De bewijzen zijn minder eenvoudig. Daarvoor verwijzen we naar de literatuur.

Stelling 1. Een eindige Markovketen heeft minstens één evenwichtsverdeling.

In lineaire-algebra-taal (laat $S = \{1, 2, \ldots, n\}$):

Bij een willekeurige $n \times n$ overgangsmatrix P is er minstens één $1 \times n$ rijvector $\pi = (\pi(1), \ldots, \pi(n))$ met $\sum_{k=1}^{n} \pi(k) = 1$, $\pi(k) \geq 0$ voor $k = 1, 2, \ldots, n$, zó dat

$$\pi = \pi P.$$

Voorbeeld 1. $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $\pi = (\frac{1}{2}, \frac{1}{2})$ is de enige evenwichtsverdeling.

Voorbeeld 2. $P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Voor iedere p, $0 \leq p \leq 1$, voldoet $\pi = (p, 1-p)$ aan $\pi = \pi P$.

De volgende voorwaarde zorgt ervoor dat er hoogstens één evenwichtsverdeling is.
Definitie: Een Markovketen heet *irreducibel* als bij ieder tweetal toestanden x en y er een n is zó dat $P^n(x, y) \neq 0$.
Dit betekent dat iedere toestand vanuit iedere andere toestand bereikbaar moet zijn.

Stelling 2. Een eindige Markovketen die irreducibel is heeft precies één evenwichtsverdeling.

13.7 Convergentie. In paragraaf 13.4 hebben we gezien dat bij de Markovketen met 2 toestanden de verdeling van X_n convergeert naar de evenwichtsverdeling: $\pi_n(x) = \mathbb{P}[X_n{=}x] \to \pi(x)$ als $n \to \infty$, ongeacht de beginverdeling.
In het bijzonder: $P^n(z, x) = \mathbb{P}_z(X_n{=}x) \to \pi(x)$ voor iedere $z \in S$. De n-staps overgangsmatrix P^n convergeert dus naar de matrix waarvan alle rijen overeenstemmen met de evenwichtsverdeling.

Zoiets geldt niet altijd. Bijvoorbeeld als $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ dan is $\pi = (\frac{1}{2}, \frac{1}{2})$ maar P^n blijft altijd heen en weer flippen tussen $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ en $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Nog een voorbeeld: $P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$, $S = \{1, 2, 3, 4\}$.
Teken het pijlenschema. Je ziet: deze Markovketen is irreducibel. De evenwichtsverdeling heeft de vorm $\pi = (\frac{1}{3}, \cdot, \cdot, \cdot)$. P^n convergeert niet. Bijvoorbeeld: $P^n(1, 1) = 1$ als n een 3-voud is, en $P^n(1, 1) = 0$ als n geen drievoud is.
Dit voorbeeld lijdt aan *periodiciteit*.

Definitie: Een Markovketen met toestandsruimte S en overgangsmatrix P noemen we *periodiek* als S is op te splitsen in $d \geq 2$ disjuncte delen $S_1, S_2, \ldots, S_d$, zó dat

$P(x,y) \neq 0$ alleen kan voorkomen als $x \in S_i$ en $y \in S_{i+1}$ ($i = 1, 2, \ldots, d-1$), of $x \in S_d$ en $y \in S_1$.
De grootste d waarbij zo'n opsplitsing mogelijk is noemen we de *periode* van de Markovketen.

De toestand maakt steeds een rondreis door $S_1, S_2, \ldots, S_d, S_1, \ldots$. Dit impliceert dat terugkeer in dezelfde toestand alleen mogelijk is in een aantal stappen dat een veelvoud is van d.
Noem k een *terugkeertijd* van toestand x als $P^k(x,x) \neq 0$. Dan zijn alle terugkeertijden (van alle toestanden) veelvouden van d. Anders gezegd: De periode d van een Markovketen is de grootste gemene deler van de terugkeertijden.
Een Markovketen heet *aperiodiek* als er voor geen enkele $d > 1$ een opsplitsing zoals zojuist beschreven mogelijk is. De grootste gemene deler van de terugkeertijden is dan 1.

Stelling 3. In een eindige Markovketen die irreducibel en aperiodiek is, geldt: $\pi_n \to \pi$ als $n \to \infty$, d.w.z. de verdeling van X_n convergeert naar de evenwichtsverdeling en P^n convergeert naar de matrix met rijen die allemaal gelijk zijn aan de evenwichtsverdeling.

Opgave 3. Voorbeeld (c) van paragraaf 13.2. Is deze Markovketen aperiodiek?

13.8 Hoe vinden we de evenwichtsverdeling? Laat P de $n \times n$ overgangsmatrix zijn van een irreducibele Markovketen. De eerste manier om π te vinden ligt voor de hand: Los op: $\pi P = \pi$, n vergelijkingen met n onbekenden.
De tweede manier is soms handig als het pijlenschema van de toestandsruimte niet te ingewikkeld is. Denk aan de toestandsruimte als een verzameling plaatsen (honderuggen bijvoorbeeld). Op plaats x zitten $n(x)$ vlooien. Het gaat in totaal over $N = \sum_{x \in S} n(x)$ vlooien, behoorlijk veel.
Op de tijdstippen $0, 1, 2, \ldots$ (of liever: net na ieder van die tijdstippen) springen al die vlooien op. Een vlo, opgesprongen uit x landt in y met kans $P(x,y)$.
Verdeel S in twee stukken S_0 en S_1 (disjunct). Als er evenwicht is, dan zullen bij zo'n springpartij gemiddeld evenveel vlooien uit S_0 vertrekken naar S_1 als er aankomen uit S_1: *Instroom* = *Uitstroom*.
Van S_0 vertrekken gemiddeld

$$\sum_{x \in S_0, y \in S_1} n(x) P(x,y)$$

vlooien naar S_1, en van S_1 vertrekken gemiddeld

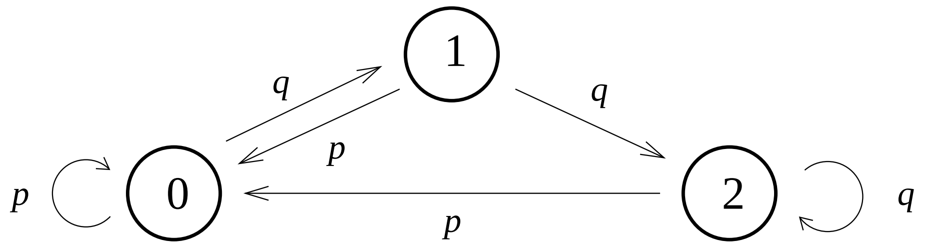
$$\sum_{x \in S_1, y \in S_0} n(x)P(x,y)$$

vlooien naar S_0. Dit levert de relatie

$$\sum_{x \in S_0, y \in S_1} n(x)P(x,y) = \sum_{x \in S_1, y \in S_0} n(x)P(x,y).$$

Door een aantal van zulke opsplitsingen geschikt te kiezen krijg je een aantal relaties waaruit de aantallen $n(x)$ te bepalen zijn, en daaruit, door normeren, de evenwichtskansen.

Voorbeeld: Voorbeeld (a) uit paragraaf 13.2. $S = \{0, 1, 2\}$, schema:



Met $S_0 = \{1\}$ krijg je: $n(1)p + n(1)q = n(0)q$, dus:

$$n(1) = n(0)q.$$

Met $S_0 = \{0, 1\}$:

$$n(1)q = n(2)p.$$

Uit deze twee relaties volgt dat

$$n(0) : n(1) : n(2) = p : pq : q^2.$$

13.9 Gemiddelde terugkeertijd. In een irreducibele Markovketen kun je vanuit iedere toestand naar iedere andere toestand. De stochast T_x die het aantal stappen telt waarin x voor het eerst bereikt wordt, heet de *aankomsttijd* in x (arrival time, hitting time). Algemener: Voor een deelverzameling A van S is de aankomsttijd in A:

$$T_A = \min\{n : n \geq 1, X_n \in A\}.$$

Wordt A nooit bereikt ($X_n \notin A$ voor alle $n \geq 1$) dan $T_A = \infty$.
Bestaat A uit een enkele toestand, $A = \{x\}$, dan schrijven we T_x in plaats van $T_{\{x\}}$.

Bij begintoestand x noemen we T_x de *terugkeertijd* in x.

Voor een irreducibele eindige Markovketen geldt:

Stelling 4. Voor iedere x en y in S geldt $\mathbb{P}_x(T_y{<}\infty) = 1$.
In het bijzonder: $\mathbb{P}_x(T_x,\infty) = 1$ en voor de verwachting van T_x bij begintoestand $X_0 = x$ geldt:

$$m(x) := \mathbb{E}_x(T_x) = \sum_{k=1}^{\infty} k\mathbb{P}_x(T_x{=}k) = \sum_{k=0}^{\infty} \mathbb{P}_x(T_x{>}k)$$

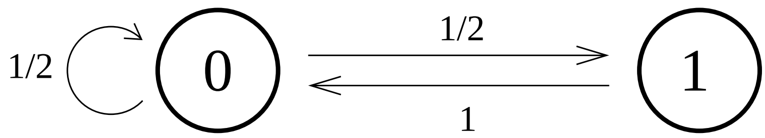
en:

$$m(x) = \frac{1}{\pi(x)},$$

waarbij $\pi(x)$ de kans op toestand x bij de evenwichtsverdeling is.

Dit is een prachtige eigenschap van de eindige irreducibele Markovketen: Iedere toestand komt steeds weer terug. De gemiddelde terugkeertijd is korter of langer naarmate de evenwichtskans van die toestand groter of kleiner is.

Voorbeeld: Twee toestanden, schema:

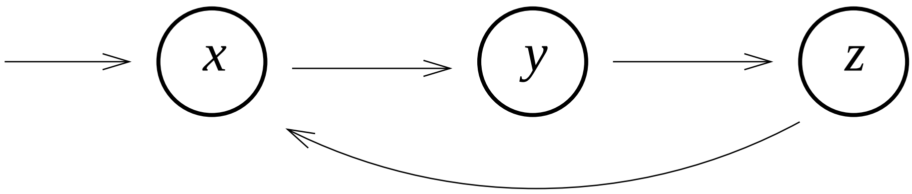


Terugkeer in toestand 0 kan alleen in 1 of 2 stappen. De gemiddelde terugkeertijd van 0 is dus $m(0) = \frac{3}{2}$. Dus: $\pi(0) = \frac{2}{3}$. Dan blijft er $\pi(1) = \frac{1}{3}$ over voor toestand 1. Dit zou betekenen dat toestand 1 een gemiddelde terugkeertijd van $m(1) = 3$ heeft. Ga na dat dit klopt.

We merken nog op dat de uitspraken van stelling 4 niet altijd waar zijn bij een irreducibele Markovketen met (aftelbaar) oneindige toestandsruimte. Daar gelden ze alleen als er een evenwichtsverdeling is. Die hoeft er echter niet te zijn. Het kan dan gebeuren dat terugkeer niet zeker is: $\mathbb{P}_x(T_x{<}\infty) < 1$. Dan heet x *voorbijgaand* (transient). Als terugkeer zeker is, $\mathbb{P}_x(T_x{<}\infty) = 1$, toestand x is *terugkerend* (recurrent), dan hoeft $m(x)$ nog niet eindig te zijn.
Voorbeeld (zonder bewijs): Een stochastische wandeling in $\mathbb{Z}$. $P(x, x{+}1) = p$, $P(x, x{-}1) = q$, $p + q = 1$. Als $p \neq q$ dan is iedere toestand voorbijgaand. Als $p = \frac{1}{2}$ dan is iedere toestand terugkerend met oneindig gemiddelde.

13.10 Fuiken en vangstkansen. Als je je geld verspeeld hebt (in het casino bijvoorbeeld) kun je niet verder spelen, want je hebt niets meer om in te zetten. Vanuit toestand 0 (in de voorbeelden (d1) en (d2) van paragraaf 13.2) kun je nergens anders heen: $P(0,x) = 0$ voor $x \neq 0$ en $P(0,0) = 1$. Zo'n toestand heet *absorberend.* x is een absorberende toestand betekent dus: $P(x,x) = 1$.
In het diagram kun je voor het gemak de pijl (lus) van x naar x weglaten, dus in plaats van

Een groepje toestanden $F \subset S$ wordt een *fuik* genoemd (Engels: closed set) als $P(x,y) = 0$ voor iedere $x \in F$, $y \notin F$.
Een *priemfuik* (irreducible closed set) is een fuik die geen kleinere fuik bevat. Een fuik die maar één toestand bevat (een absorberende toestand) is vanzelf een priemfuik. Een priemfuik die uit meerdere toestanden bestaat kan er bijvoorbeeld zó uitzien:



Bij een irreducibele Markovketen is de hele toestandsruimte S de enige priemfuik. In een eindige Markovketen die niet irreducibel is zijn er altijd een of meer priemfuiken. Het systeem komt dan vroeg of laat in een van die priemfuiken terecht.
In de voorbeelden (d1) en (d2) van paragraaf 13.2 komt de speler op den duur in één van de priemfuiken $\{0\}$ of $\{50\}$.

Het is van belang de bijbehorende kansen te weten. De kans om, startend in x, uiteindelijk in priemfuik F terecht te komen noemen we de *vangstkans* van F, notatie $v(x,F)$, of korter $v(x)$ als we weten over welke F we het hebben.

De functie v voldoet aan:

(a) $v(x) = 1$ als $x \in F$, $v(x) = 0$ als x in een andere priemfuik. Dit zijn de "randvoorwaarden".

(b) $v(x) = \sum_y P(x,y)v(y)$ (matrixnotatie: $v = Pv$).

Bij een eindige Markovketen is v door (a) en (b) eenduidig vastgelegd. We passen dit toe op het probleem van de gokker uit paragraaf 13.2:

13.11 Ruïneringsprobleem (Gamblers Ruin).

Neem het schema van voorbeeld (d1).
De kans om te winnen, de succeskans, is (bij begintoestand x) de vangstkans $v(x)$ van toestand $\{50\}$.
Er geldt:

$$\begin{cases} \text{(a)} & v(0) = 0,\ v(5) = 1 \\ \text{(b)} & v(x) = qv(x{-}10 + pv(x{+}10) \end{cases}$$

Als $p\frac{1}{2}$ dan volgt uit $v(x) = \frac{1}{2}v(x{-}10) + \frac{1}{2}v(x{+}10)$ dat de grafiek van de functie v rechtlijnig is en dus $v(x) = \frac{x}{50}$. In het bijzonder: $v(20) = \frac{20}{50} = 0{,}4$.
Dit is eerlijk spel. De verwachting aan het eind is gelijk aan het kapitaal van de speler in het begin.
Voor $p \neq \frac{1}{2}$ krijgen we

$$lv(x) = (p{+}q)v(x) = pv(x{+}1) + qv(x{-}1)$$

dus

$$v(x{+}1) - v(x) = \frac{q}{p}(v(x) - v(x{-}1))$$

Schrijf even $a(x) = v(x) - v(x{-}1)$ en $r = \frac{q}{p}$, dan staat hier:

$$a(x) = ra(x{-}1).$$

Dit betekent dat

$$\begin{aligned} v(x) &= v(0) + a(1) + a(2) + \ldots + a(x) \\ &= 0 + a(1) + ra(1) + \ldots + r^{x-1}a(1) \\ &= a(1) \cdot \frac{r^x - 1}{r - 1}. \end{aligned}$$

Uit de randvoorwaarde $v(5) = 1$ volgt: $a(1) = \frac{r-1}{r^5-1}$ en dus:

$$v(x) = \frac{r^x - 1}{r^5 - 1} = \frac{\left(\frac{q}{p}\right)^x - 1}{\left(\frac{q}{p}\right)^5 - 1}.$$

Bij $p = \frac{18}{37}$, $q = \frac{19}{37}$ (roulette: 18 rode en 18 zwarte nummers waar je op kunt inzetten en één nul, waarbij alle inzetten naar het casino gaan) krijgen we $v(2) = \frac{(19/18)^2-1}{(19/18)^5-1} =$

0,368.
Bij de tweede manier van spelen (schema (d2)) krijgen we

$$\begin{cases} v(2) = pv(4) \\ v(4) = p + qv(3) \\ v(3) = p + qv(1) \\ v(1) = pv(2) \end{cases} \text{randvoorwaarden:} v(0) = 0, v(5) = 1$$

Voor $v(2)$ krijgen we dan $v(2) = \frac{p^2+p^2q}{1-p^2q^2}$.
Met $p = \frac{18}{37}$ en $q = \frac{19}{37}$: $v(2) = 0{,}382$.
Volgens Dubins en Savage is dit de maximale kans om te winnen (Bold play is optimal in subfair casinos).
Het verschil tussen de speelwijzen van (d1) en (d2) wordt duidelijker wanneer het om grotere bedragen gaat (zie opgave 5).

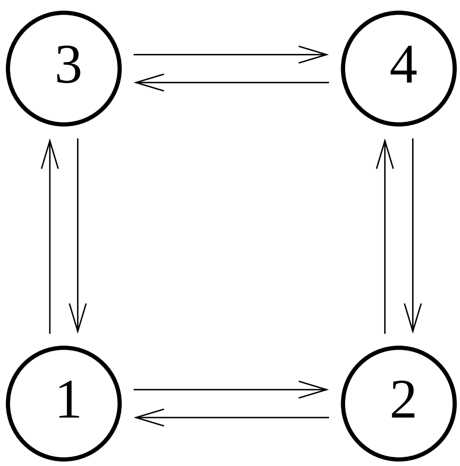
13.12 Hoe lang gaat het spel duren? Beginnend in toestand x zal het systeem na T stappen in één van de absorberende priemfuiken terechtkomen.
Algemener: Laat $T := T_A$ de aankomsttijd zijn in een verzameling $A \subset S$ (bij het gokprobleem bijvoorbeeld: $A = \{0, 50\}$).
Voor de verwachting $u(x) := \mathbb{E}_x(T)$ geldt:

(a) $u(x) = 0$ als $x \in A$;

(b) $u(x) = 1 + \sum_{y \in S} P(x,y)u(y)$ als $x \notin A$.

Uit deze relatie kan $u(x)$ berekend worden. In voorbeeld (d1) krijgen we als $p = \frac{1}{2}$: $u(x) = \frac{x(50-x)}{100}$.

Ander voorbeeld: Toevalswandeling op het vierkant.



Hoe lang duurt het tot je in 4 aankomt, als je begint in 1 (de overgangskansen zijn allemaal $\frac{1}{2}$)?
Uit

$$\begin{cases} u(1) = 1 + \frac{1}{2}u(2) + \frac{1}{2}u(3) \\ u(2) = 1 + \frac{1}{2}u(1) \\ u(3) = 1 + \frac{1}{2}u(1) \end{cases}$$

volgt: $u(1) = 4$; $u(2) = u(3) = 3$.

Met de volgende truc, bedacht door David van Danzig (Amsterdam (1948)) kunnen we ook de voortbrengende functie $G(s)$ van T afleiden.
Het idee is $G(s) := \mathbb{E}_x(s^T) = \sum_{k=1}^{\infty} s^k \mathbb{P}_x(T{=}k)$ te interpreteren als een kans. Hierbij is s de kans dat een stap veilig verloopt (s = safe). Bij iedere stap van het proces kan er een ongeluk gebeuren. Er ontploft een bom, of iemand krijgt een woedeaanval en slaat alles kort en klein.
$G(s)$ is dan de kans dat het goed afloopt. Immers, wanneer er $T = k$ stappen nodig zijn om A te bereiken, dan is er kans s^k dat die alle k veilig gedaan worden.
Met deze interpretatie krijgen we voor $a_x := \mathbb{E}_x(s^T)$:

(a) $a_x = 1$ als $x \in A$;

(b) $a_x = \sum_{y \in S} P(x, y) s a_y$ als $x \notin A$.

Voor de toevalswandeling op het vierkant krijgen we voor de aankomsttijd T in 4, beginnende in x:

$$\begin{cases} a_1 = \frac{1}{2}sa_2 + \frac{1}{2}sa_3 \\ a_2 = \frac{1}{2}sa_1 + \frac{1}{2}s \\ a_3 = \frac{1}{2}sa_1 + \frac{1}{2}s \end{cases}$$

Hieruit volgt $a_1 = G(s) = \frac{\frac{1}{2}s^2}{1-\frac{1}{2}s^2}$ waarmee we de hele kansverdeling van T te pakken hebben.

Opgaven

4. $S = \{1,2,3\}$, $P = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$ Is de Markovketen met deze overgangsmatrix irreducibel?

   Toch is er maar één evenwichtsverdeling. Bepaal die.

5. Rood en zwart.

   Het verschil tussen de beide manieren van spelen (voorbeelden (d1) en (d2)) wordt groter naarmate het om meer geld gaat. Bepaal de succeskansen als je begint met 200 gulden, stoppen bij 500 of 0

   (a) als je iedere keer een tientje inzet;

   (b) met "bold play".

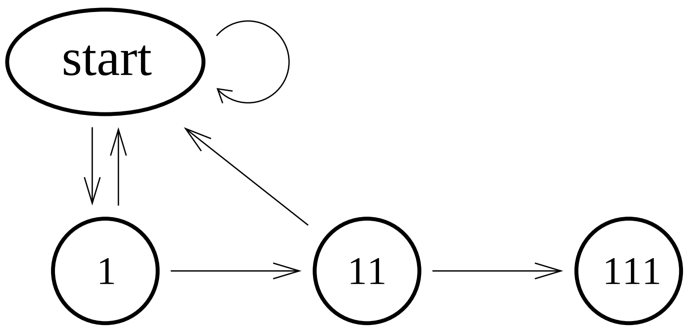
6. Voorbeeld (c) uit paragraaf 13.2. Bij iedere sprong verplaatst het beestje zich van een hoekpunt van de kubus naar een van de drie aangrenzende hoekpunten (met gelijke kansen). De reis begint in hoekpunt 1. Na T sprongen wordt voor het eerst het tegenoverliggende hoekpunt bezocht.

   (a) Bepaal $\mathbb{E}(T)$.

   (b) Bepaal de voortbrengende functie van T.

7. Een reeks worpen met een munt.

   Na T worpen heb ik voor het eerst een rijtje van 3 keer achter elkaar kop. Bepaal $\mathbb{E}(T)$. Schema (munt = 0, kop = 1, bij iedere pijl hoort kans $\frac{1}{2}$):



   Bepaal ook de verwachting van het aantal worpen dat nodig is om voor het eerst een rijtje 010 te krijgen.

8. Speler A wint als in de rij worpen met de munt het rijtje 111 eerder komt dan het rijtje 010 van speler B. Bepaal de kans dat A wint.

Appendix A

Tabel voor de normale verdeling

Voor de cijfers moet "0," worden gezet. 637 betekent dus: 0,637.

x	0	1	2	3	4	5	6	7	8	9
0,0	500	504	508	512	516	520	524	528	532	536
0,1	540	544	548	552	556	560	564	567	571	575
0,2	579	583	587	591	595	599	603	606	610	614
0,3	618	622	626	629	633	637	641	644	648	688
0,4	655	659	663	666	670	674	677	681	684	688
0,5	691	695	698	702	705	709	712	716	719	722
0,6	726	729	732	736	739	742	745	749	752	755
0,7	758	761	764	767	770	773	776	779	782	785
0,8	788	791	794	797	800	802	805	808	811	813
0,9	816	819	821	824	826	829	831	834	836	839
1,0	841	844	846	848	851	853	855	858	860	862
1,1	864	867	869	871	873	875	877	879	881	883
1,2	885	887	889	891	893	894	896	898	900	901
1,3	903	905	907	908	910	911	913	915	916	918
1,4	919	921	922	924	925	926	928	929	931	932
1,5	933	934	936	937	938	939	941	942	943	944
1,6	945	946	947	948	949	951	952	953	954	954
1,7	955	956	957	958	959	960	961	962	962	963
1,8	964	965	966	966	967	968	969	969	970	971
1,9	971	972	973	973	974	974	975	976	976	977
2,0	977	978	978	979	979	980	980	981	981	982
2,1	982	983	983	983	984	984	985	985	985	986
2,2	9861	9864	9868	9871	9875	9878	9881	9884	9887	9890
2,3	9893	9896	9898	9901	9904	9906	9909	9911	9913	9916
2,4	9918	9920	9922	9925	9927	9929	9931	9932	9934	9936
2,5	9938	9940	9941	9943	9945	9946	9948	9949	9951	9952
2,6	9952	9955	9956	9957	9959	9960	9961	9962	9963	9964
2,7	9965	9966	9967	9968	9969	9970	9971	9972	9973	9974
2,8	9974	9975	9976	9977	9977	9978	9979	9979	9980	9981
2,9	9981	9982	9982	9983	9984	9984	9985	9985	9986	9986

Index