Deel B

Kansrekening

Aanbevolen achtergrondliteratuur met veel opgaven (en oplossingen):

- Murray R. Spiegel, John J. Schiller, R. A. Srinivasan: (Schaum's Outline of Theory and Problems of) Probability and Statistics. McGraw-Hill Companies, 2000, 408 p., ISBN: 0071350047.

Les 6 Combinatoriek

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

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

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

Het Simpson paradox

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

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

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

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

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

6.1 Tellen van uitkomsten

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

Bij het dobbelen met drie dobbelstenen kunnen we ons afvragen of de kans groter is dat de som van de ogen 11 of 12 is. Hiervoor moeten we 11 of 12

schrijven als sommen van drie getallen uit de verzameling $\{1, 2, 3, 4, 5, 6\}$. We hebben

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

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

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

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

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

6.2 Geordende grepen

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

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

Principe van de vermenigvuldiging van uitkomsten: *Het aantal uitkomsten voor een geordende greep is $n_1 \cdot n_2 \cdot \ldots \cdot n_r = \prod_{i=1}^{r} n_i$ als we r keer trekken en er voor de i-de greep n_i mogelijkheden zijn.*

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

Trekken met terugleggen

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

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

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

Trekken zonder terugleggen

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

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

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

> **Notatie:** Het product $1 \cdot 2 \cdot \ldots \cdot m$ van de getallen tot en met m noteren we kort met $m!$ en noemen dit 'm faculteit'. Men spreekt af dat $0! = 1$ is, want het 'lege product' van 1 t/m 0 moet bij het vermenigvuldigen niets veranderen.

Als we zo lang trekken totdat de elementen uit de verzameling op zijn (d.w.z. n keer), vinden we de volgende belangrijke uitspraak:

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

6.3 Ongeordende grepen

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

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

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

Trekken zonder terugleggen

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

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

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

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

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

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

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

dus bijvoorbeeld $(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$.

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

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

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

(ii) $\sum_{r=0}^{n} \binom{n}{r} = 2^n$
Dit volgt uit de binomische formule als we $a = b = 1$ invullen. Maar we kunnen dit ook uit het aftellen van deelverzamelingen zien: Een verzameling Ω van n elementen heeft $\binom{n}{r}$ deelverzamelingen met r elementen, dus is de som over de binomiaalcoëfficiënten het aantal van alle deelverzamelingen van Ω. Maar elk element $a \in \Omega$ is of in een deelverzameling $A \subseteq \Omega$ bevat of is er niet in bevat. Dit geeft 2 mogelijkheden voor elk element en dus 2^n mogelijkheden om de uitkomsten $a \in A$ of $a \notin A$ op de n elementen van Ω te verdelen en dus zijn er 2^n deelverzamelingen van Ω.

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

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

Trekken met terugleggen

Als we na een greep het getrokken element weer terugleggen maar niet op de volgorde letten, willen we het aantal rijen $(a_1, \ldots, a_r)$ bepalen met $a_i \in \{1, \ldots, n\}$ en $a_i \leq a_{i+1}$. Merk op dat we het aantal van dit soort rijen niet

$$\begin{array}{ccccccc}
\binom{0}{0} & & & & & & \\
\binom{1}{0} & \binom{1}{1} & & & & & \\
\binom{2}{0} & \binom{2}{1} & \binom{2}{2} & & & & \\
\binom{3}{0} & \binom{3}{1} & \binom{3}{2} & \binom{3}{3} & & & \\
\binom{4}{0} & \binom{4}{1} & \binom{4}{2} & \binom{4}{3} & \binom{4}{4} & & \\
\binom{5}{0} & \binom{5}{1} & \binom{5}{2} & \binom{5}{3} & \binom{5}{4} & \binom{5}{5} & \\
\binom{6}{0} & \binom{6}{1} & \binom{6}{2} & \binom{6}{3} & \binom{6}{4} & \binom{6}{5} & \binom{6}{6}
\end{array}
\qquad
\begin{array}{ccccccc}
1 & & & & & & \\
1 & 1 & & & & & \\
1 & 2 & 1 & & & & \\
1 & 3 & 3 & 1 & & & \\
1 & 4 & 6 & 4 & 1 & & \\
1 & 5 & 10 & 10 & 5 & 1 & \\
1 & 6 & 15 & 20 & 15 & 6 & 1
\end{array}$$

Figuur B.1: Driehoek van Pascal

zo makkelijk uit het aantal van geordende rijen kunnen bepalen, omdat het aantal permutaties van een rij met herhalingen ervan afhangt hoeveel elementen hetzelfde zijn.

Maar hier komen we met een trucje en het resultaat voor het trekken zonder terugleggen verder: Stel we hebben een rij $(a_1, \ldots, a_r)$ met $a_i \leq a_{i+1}$, dan kunnen we hieruit een rij zonder herhalingen maken door $(i-1)$ bij het element a_i op te tellen. Dit geeft de nieuwe rij $(b_1, \ldots, b_r)$ waarbij

$$b_i = a_i + i - 1 < a_{i+1} + i = b_{i+1}.$$

Voor de elementen b_i geldt $1 \leq b_i \leq n + r - 1$, dus hoort deze rij bij een ongeordende greep zonder terugleggen uit $n + r - 1$ elementen.

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

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

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

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

	geordend	ongeordend
met terugleggen	I	III
zonder terugleggen	II	IV

Deze vier gevallen kunnen we als volgt karakteriseren:

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

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

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

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

Het Verjaardagsparadox

We willen de kans berekenen, dat er in en groep van r mensen twee mensen op dezelfde dag jarig zijn. Als verzameling nemen we de verzameling van verjaardagen, dus $|\Omega| = 365$ (we nemen aan dat niemand op 29 februari jarig is). Voor het aantal mogelijke uitkomsten zijn we in geval I, omdat we de mensen kunnen onderscheiden, dus het aantal is 365^r. Nu gebruiken we een klein trucje: We bepalen de kans van het complement van de gewenste uitkomst, dus we bepalen de kans dat alle r mensen verschillende verjaardagen hebben. Dan zijn we voor de gunstige uitkomsten in geval II, want een verjaardag van één persoon mag niet meer het verjaardag van een andere persoon zijn. Er zijn dus $\binom{365}{r} r!$ gunstige uitkomsten (d.w.z. alle verjaardagen zijn verschillend). Bij elkaar genomen is de kans dat twee mensen op dezelfde dag jarig zijn dus

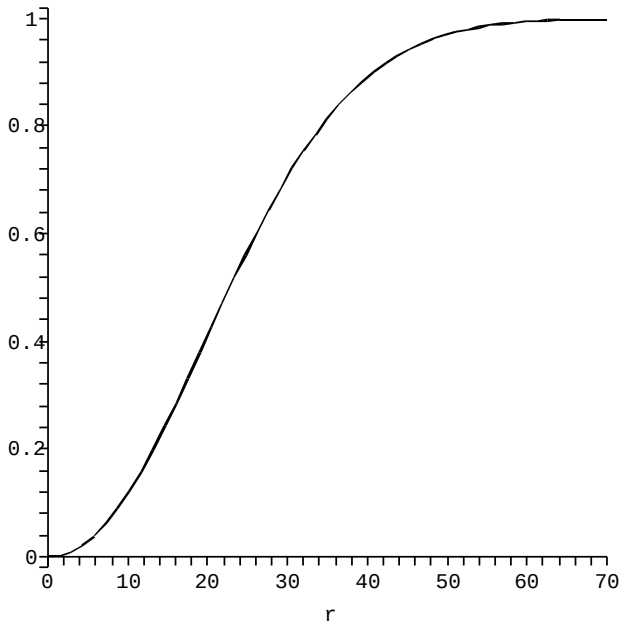
$$p = 1 - \frac{\binom{365}{r} r!}{365^r}.$$

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

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

In Figuur B.2 zie je de functie, die de kans op twee mensen met dezelfde verjaardag afhankelijk van de grootte r van de groep aangeeft. Omdat veel mensen het verrassend vinden dat de kans al voor $r = 23$ groter dan 0.5 is, noemt men dit ook het *verjaardagsparadox*.

> De reden voor de intuïtieve verrassing over dit resultaat ligt eerder op psychologisch dan op wiskundig gebied. Als we erover nadenken, dat twee verjaardagen samenvallen, nemen we vaak stiekem aan dat de verjaardag van iemand uit de groep op de verjaardag van een *bepaalde persoon* valt, bijvoorbeeld op ons eigen verjaardag. Maar in het verjaardagsparadox ligt de dag niet vast, de vraag is of er überhaupt een dag is, waarop twee mensen jarig zijn.



Figuur B.2: Kans op dezelfde verjaardag bij r mensen

Algemeen laat zich aantonen dat voor $r \approx \sqrt{n}$ geldt dat r grepen uit n objecten met kans $\frac{1}{2}$ twee dezelfde resultaten opleveren. Dit is bijvoorbeeld voor de *cryptologie* een belangrijke uitspraak, waar men eist dat de waarden van een *hashing functie* 'botsingvrij' moeten zijn, d.w.z. het moet ondoenlijk zijn om twee boodschappen te produceren die dezelfde hashwaarde opleveren. Als een hashing functie 2^{2n} mogelijke hashwaarden oplevert, moet het dus ondoenlijk zijn om 2^n boodschappen te produceren, want bij ongeveer dit aantal boodschappen kan men verwachten dat een hashwaarde dubbel voorkomt.

Belangrijke begrippen in deze les

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

Opgaven

37. Je hebt 4 verschillende wiskunde boeken, 6 psychologie boeken en 2 letterkundige boeken. Hoeveel manieren zijn er om deze twaalf boeken op een boord te plaatsen als:

    (i) je een genie bent en geen orde nodig hebt,

    (ii) je tenminste de wiskunde boeken naast elkaar plaatst,

    (iii) de boeken van elk vakgebied naast elkaar moeten staan?

38. Hoeveel verschillende getallen van 4 cijfers kan je uit de zestien hexadecimale 'cijfers' $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F\}$ maken?

    (i) Hoeveel van deze getallen zijn 'echte' 4-cijfer getallen, dus hebben de eerste cijfer $\neq 0$?

    (ii) Hoeveel van de getallen uit (i) hebben vier verschillende cijfers?

    (iii) Hoeveel van de getallen uit (ii) eindigen op het cijfer 0?

    (iv) Hoeveel van de getallen uit (ii) hebben opstijgende cijfers?

39. Een Nederlands kentekenplaatje bestaat uit twee groepen van twee letters en een groep van twee cijfers. De groep van cijfers mag voor, tussen of achter de groepen met letters staan. Verder worden bij de letters geen klinkers gebruikt. Bepaal het aantal mogelijke nummerborden.

40. Uit een werkgroep van 8 mannen en 6 vrouwen moet een commissie van 3 mannen en 4 vrouwen gekozen worden. Hoeveel verschillende mogelijkheden bestaan er voor de commissie?

41. Een zekere faculteit heeft 6 hoogleraren, 8 UHD's, 4 UD's en 5 AIO's. In de feestcommissie van de faculteit zitten er 2 hoogleraren, 4 UHD's, 3 UD's en 3 AIO's. Hoeveel mogelijkheden zijn er voor de commissie? Hoe veranderd het aantal als een van de hoogleraren een begenadigde zanger en een van de UHD's (oorspronkelijk) een bierbrouwer is en deze twee per se in de commissie moeten zitten?

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

43. We dobbelen met twee dobbelstenen.

    (a) Bepaal de kansen voor de volgende uitkomsten:

        (i) De som van de twee getallen is 5.

        (ii) Beide dobbelstenen tonen een oneven getal.

        (iii) De eerste dobbelsteen toont een kleiner getal dan de tweede.

        (iv) De som van de twee getallen is even.

        (v) De som van de twee getallen is minstens 4.

        (vi) De som van de twee getallen is of even of minstens 4 (of allebei).

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

44. In een vaas zitten 8 rode, 3 witte en 9 blauwe knikkers. Je trekt drie keer een knikker zonder terugleggen. Bepaal de volgende kansen:

    (i) alle drie getrokken knikkers zijn rood,

(ii) alle drie getrokken knikkers zijn wit,

(iii) twee van de getrokken knikkers zijn rood, de derde is blauw,

(iv) minstens een van de getrokken knikkers is wit,

(v) bij de getrokken knikkers is een van elke kleur,

(vi) de knikkers worden in de volgorde rood, wit, blauw getrokken.

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

    (i) tweeling: twee kaarten van dezelfde soort (bijvoorbeeld twee boeren),

    (ii) dubbele tweeling: twee verschillende tweelingen (bijvoorbeeld twee vrouwen en twee azen),

    (iii) drieling: drie kaarten van dezelfde soort,

    (iv) vierling: vier kaarten van dezelfde soort,

    (v) full house: een tweeling en een drieling,

    (vi) straight: vijf kaarten in de goede volgorde (bijvoorbeeld 9, 10, boer, vrouw, heer),

    (vii) straight flush: een straight van dezelfde kleur.

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

46. Je spreekt met een vriend af om op de volgende dag in de rij te staan om kaarten voor Bruce Springsteen (of AC/DC of Lang Lang) te kopen. Op een gegeven moment staan jullie allebei in de rij, maar hebben elkaar niet gezien.

    (i) Hoe groot is de kans, dat in een rij van n mensen precies r mensen tussen jullie staan?

    (ii) Hoe groot is de kans dat jullie elkaar kunnen zien als er 1000 mensen in de rij staan en je aanneemt dat je je vriend onder de 100 mensen naast je kunt herkennen?

Les 7 Kansverdelingen

We hebben in het begin gesteld dat we de kans voor een zekere gunstige uitkomst berekenen als het aantal gunstige uitkomsten gedeeld door het totale aantal mogelijke uitkomsten. Maar vaak is het handig, dat we verschillende uitkomsten samenvatten en dit als een nieuwe soort uitkomst bekijken. Bijvoorbeeld kunnen we bij het werpen van twee dobbelstenen de som van de twee geworpen getallen als uitkomst nemen. Als we met $P(s)$ de kans op de som s noteren, zien we (door de mogelijke gevallen na te gaan) makkelijk in, dat

$$P(1) = \frac{0}{36},\ P(2) = \frac{1}{36},\ P(3) = \frac{2}{36},\ P(4) = \frac{3}{36},\ P(5) = \frac{4}{36},\ P(6) = \frac{5}{36},$$
$$P(7) = \frac{6}{36},\ P(8) = \frac{5}{36},\ P(9) = \frac{4}{36},\ P(10) = \frac{3}{36},\ P(11) = \frac{2}{36},\ P(12) = \frac{1}{36}.$$

Hieruit laat zich bijvoorbeeld snel aflezen, dat de kans op het dobbelen van een som die een priemgetal is, gelijk is aan $(1 + 2 + 4 + 6 + 2)/36 = 5/12$.

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

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

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

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

nodig, die op de *machtsverzameling* $\mathcal{P}(\Omega)$ van Ω, d.w.z. op de verzameling van alle deelverzamelingen van Ω, gedefinieerd is.

B.1 Definitie Een functie $P : \mathcal{P}(\Omega) \to \mathbb{R}$ op de verzameling van alle deelverzamelingen van Ω heet een *kansverdeling* als P aan de volgende eisen voldoet:

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

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

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

De eerste eigenschap zegt alleen maar, dat kansen niet negatief mogen zijn, en de tweede eigenschap beweert, dat alle mogelijke uitkomsten inderdaad in Ω liggen. De derde eigenschap is een soort van additiviteit, die zegt dat we de kansen voor uitkomsten die niet overlappen (en elkaar dus uitsluiten) gewoon mogen optellen.

We hadden in principe ook nog kunnen eisen, dat $P(A) \leq 1$ is voor alle $A \subseteq \Omega$, maar dit kunnen we inderdaad al uit (i)-(iii) afleiden en we willen graag zo zuinig als mogelijk met onze eisen zijn.

7.1 Discrete kansverdelingen

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

De gelijkverdeling

Een belangrijk voorbeeld van een discrete kansverdeling hebben we al gezien, namelijk de ***gelijkverdeling***:

Elke mogelijke uitkomst $w \in \Omega$ moet dezelfde kans hebben (vandaar de naam), dan is $P(w) = \frac{1}{|\Omega|}$ voor elke $w \in \Omega$. Hieruit volgt met eigenschap (iii) dat $P(A) = \frac{|A|}{|\Omega|}$ en dit is precies de relatieve frequentie.

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

De hypergeometrische verdeling

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

$$P(k) = \frac{\binom{6}{k} \cdot \binom{43}{6-k}}{\binom{49}{6}}.$$

De waarden voor deze kansen zijn:

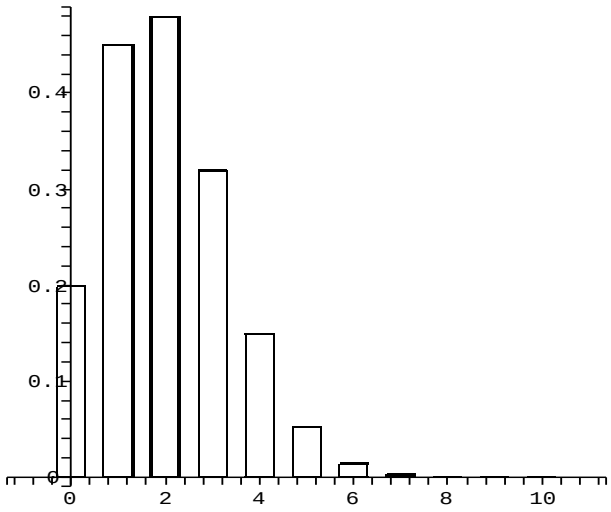
$k = 0:$	43.6%	(1 in 2.3)
$k = 1:$	41.3%	(1 in 2.4)
$k = 2:$	13.2%	(1 in 7.6)
$k = 3:$	1.8%	(1 in 57)
$k = 4:$	0.1%	(1 in 1032)
$k = 5:$	0.002%	(1 in 54201)
$k = 6:$	0.000007%	(1 in 13983816)

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

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

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

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



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

De praktijk van een qualiteitstoets ziet er natuurlijk eigenlijk iets anders uit: We weten niet hoeveel slechte stukken er in de levering zitten, maar de leverancier beweert dat het er minder dan s_0 zijn. Wij kennen wel de waarden

n, m en k en schatten nu de waarde $\hat{s}$ van s zo dat $h(n, m, \hat{s}; k)$ maximaal wordt. Als onze schatting $\hat{s}$ groter dan s_0 is, zullen we de levering waarschijnlijk niet accepteren.

Een andere toepassing van dit soort schatting vinden we in de ecologie. Als we het aantal vissen in een vijver willen bepalen, kunnen we een aantal s van vissen markeren en op de volgende dag het aantal k van gemarkeerde vissen in een greep van m vissen bepalen. We schatten dan het aantal $\hat{n}$ van vissen in de vijver zo dat $h(\hat{n}, m, s; k)$ maximaal wordt.

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

De binomiale verdeling

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

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

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

> We kunnen de binomiale verdeling ook iets anders interpreteren: Stel bij een experiment hebben we iedere keer een kans p op succes. Dan geeft $b(m, p; k)$ de kans aan dat we bij m pogingen k successen boeken. Merk op dat er bij deze interpretatie geen verzameling van n stukken meer is, waaruit we een steekproef nemen, maar dat we van begin af met de kans p op een succes werken.

Intuïtief zullen we zeggen, dat het voor het geval dat n veel groter is dan m bijna geen verschil maakt of we met of zonder terugleggen trekken, want de

kans dat we een element twee keer pakken is heel klein. Er laat zich inderdaad zuiver aantonen, dat voor $n \gg m$ de hypergeometrische verdeling meer en meer op de binomiale verdeling lijkt en in de limiet geldt

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

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

Er laat zich geen algemene regel aangeven, wanneer de binomiale benadering goed genoeg is. Soms leest men iets van $n > 2000$ en $\frac{m}{n} < 0.1$, maar in sommige gevallen heeft de benadering dan al een behoorlijke afwijking. Voor $n = 2000$, $m = 100$, $s = 20$ en $k = 2$ hebben we bijvoorbeeld $h(2000, 100, 20; 2) = 18.95\%$ en de binomiale benadering geeft in dit geval $b(100, \frac{20}{2000}; 2) = 18.49\%$ wat al een tamelijke afwijking is. Als we aan de andere kant naar de kans op 2 goede getallen in de lotto kijken, hebben we $h(49, 6, 6; 2) = 13.24\%$. De binomiale benadering hiervan is $b(6, \frac{6}{49}; 2) = 13.34\%$ en dit is een redelijke benadering terwijl we hier niet aan het criterium voldoen.

De Poisson-verdeling

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

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

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

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

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

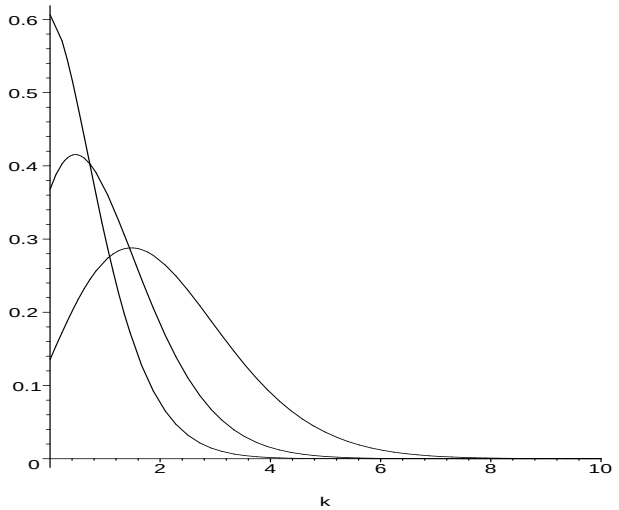
Voor zeldzame gebeurtenissen gaat de binomiale verdeling dus in de limiet over in de ***Poisson-verdeling***

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

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

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

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



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

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

Voor kleine waarden van λ is de grafiek van de Poisson-verdeling strikt dalend, dit geeft weer dat we helemaal geen optreden van het gebeurtenis ver-

wachten. Pas voor waarden $\lambda \gtrsim 0.562$ heeft de functie grotere waarden dan $po_\lambda(0) = e^{-\lambda}$ een heeft dus een maximum.

De precieze positie van het maximum laat zich voor de continue functie alleen maar door een ingewikkelde functie (de Ψ-functie) beschrijven, voor $\lambda = 1$ ligt het ongeveer bij 0.46 en voor $\lambda = 2$ bij 1.48.

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

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

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

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

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

7.2 Continue kansverdelingen

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

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

omdat we de punten van de natuurlijke getallen als gescheiden punten op de reële lijn beschouwen. De Poisson-verdeling is hier in feite een voorbeeld van.

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

Maar we kunnen ook kansverdelingen met continue uitkomsten beschrijven. Om het idee hiervan nader toe te lichten, bekijken we twee bekende voorbeelden.

Voorbeeld 1: Rad van avontuur. Een Rad van avontuur is in een aantal (even grote) segmenten ingedeeld en op sommige van de segmenten maak je een winst als het rad op dit segment stopt. Als we er n segmenten hebben noemen we deze $1, \ldots, n$ en voor elke k met $1 \leq k \leq n$ is de kans dat het rad in het k-de segment stopt gelijk aan $\frac{1}{n}$ (we gaan van een eerlijk rad uit). Maar we kunnen de uitslag dat het rad in het k-de segment stopt ook anders beschrijven, namelijk met behulp van de hoek φ waarop het rad stopt. We hebben namelijk de uitkomst k als voor de hoek φ geldt dat $(k-1)\frac{2\pi}{n} \leq \varphi \leq k\frac{2\pi}{n}$.

Als we nu naar de kans kijken dat het rad van avontuur tussen de hoeken φ_1 en φ_2 stopt dan is deze kans $\frac{\varphi_2-\varphi_1}{2\pi}$ omdat dit het relatieve aandeel van de rand is die tussen de hoeken ligt.

Voorbeeld 2: Dartspel. We gaan nu van het Rad van avontuur naar het dartspel over. Ook hier is de kans om een pijltje tussen de hoeken φ_1 en φ_2 te plaatsen gelijk aan $\frac{\varphi_2-\varphi_1}{2\pi}$, maar dit geldt nu alleen maar omdat de dart schijf een cirkel is. Als we een schijf hebben die niet rond is maar waarvan de straal afhangt van de hoek, dan geven we de straal met een functie $r(\varphi)$ aan. Een segment met een hoek van $\Delta\varphi$ van een cirkel met straal r heeft een oppervlakte van $\frac{\Delta\varphi}{2\pi}\pi r^2 = \frac{1}{2}r^2\Delta\varphi$, dus kunnen we de totale oppervlakte van de schijf door de integraal

$$O = \int_0^{2\pi} \frac{1}{2}\, r(\varphi)^2 \, d\varphi$$

berekenen en krijgen de oppervlakte van het segment tussen φ_1 en φ_2 als

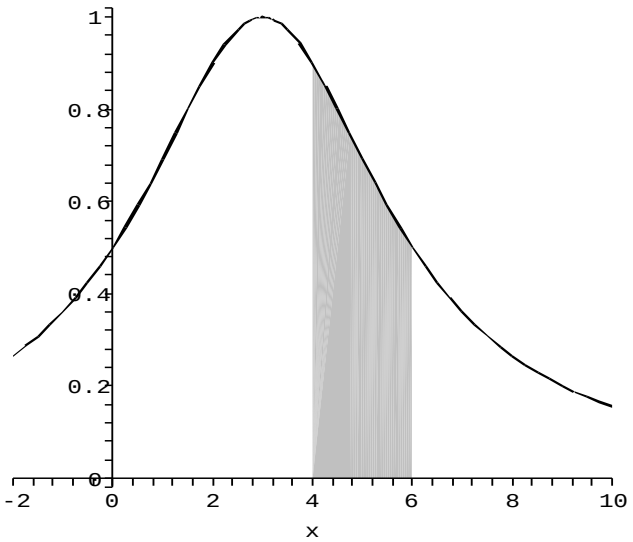
$$S = \frac{1}{2}\int_{\varphi_1}^{\varphi_2} r(\varphi)^2 \, d\varphi.$$

De kans dat een dart-pijltje (bij een toevallige verdeling over de schijf) in het segment tussen φ_1 en φ_2 terecht komt is het aandeel van het segment S aan de totale oppervlakte O van de schijf, dus de integraal

$$P(\varphi_1, \varphi_2) = \frac{S}{O} = \frac{1}{2O}\int_{\varphi_1}^{\varphi_2} r(\varphi)^2 \, d\varphi.$$

Aan de hand van deze twee voorbeelden kunnen we het algemene idee voor continue kansverdelingen makkelijk inzien:

We beschrijven de kans dat de uitkomst x van een experiment in het interval $[a, b]$ valt als oppervlakte onder de grafiek van een geschikte functie $f(x)$ op het interval $[a, b]$ zo als in Figuur B.5 te zien.



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

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

$$P(x \in [a, b]) = \int_a^b f(x)\, dx.$$

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

B.2 Definitie Een functie $f(x) : \mathbb{R} \to \mathbb{R}$ heet een *dichtheidsfunctie* als $f(x)$ aan de volgende eisen voldoet:

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

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

Een dichtheidsfunctie $f(x)$ legt een *continue kansverdeling* vast door de kans op een waarde x in een interval $[a, b]$ te definiëren door

$$P(x \in [a, b]) = \int_a^b f(x)\, dx.$$

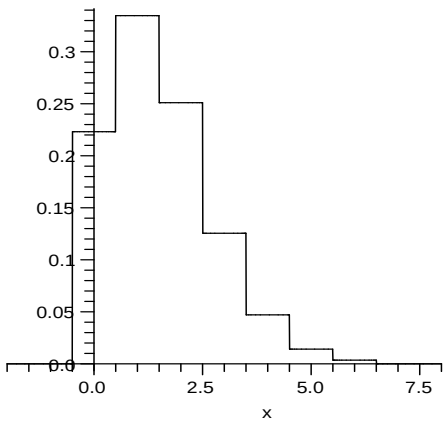
De eerste eis aan de dichtheidsfunctie zorgt ervoor dat we steeds niet-negatieve kansen krijgen en de tweede eis zegt dat de totale oppervlakte onder de grafiek 1 is en geeft dus de normering van de kansverdeling weer.

In principe kunnen we ook discrete kansverdelingen als continue kansverdelingen opvatten. Als de uitkomstenruimte Ω de natuurlijke getallen $0, 1, 2, \ldots$ bevat en we aan de uitkomst i de kans $P(i)$ toekennen, kunnen we de uitkomst i door het interval $I = [i - \frac{1}{2}, i + \frac{1}{2}]$ vervangen. De kans op een uitkomst in het interval I is dan juist de kans op de uitkomst i, want dit is de enige mogelijke uitkomst die in het interval ligt.

Omdat de lengte van het interval I juist 1 is, heeft een rechthoek van hoogte $P(i)$ op dit interval de oppervlakte $1 \cdot P(i) = P(i)$ en geeft dus de kans op de uitkomst i aan.

Als dichtheidsfunctie hebben we dus de functie nodig die op het interval $[i - \frac{1}{2}, i + \frac{1}{2}]$ de constante waarde $P(i)$ heeft.

Voor de Poisson-verdeling met parameter $\lambda = 1.5$ ziet deze functie er bijvoorbeeld zo als in Figuur B.6 uit. Merk op dat zo'n functie op een *histogram* lijkt, waarmee (relatieve) frequenties van gebeurtenissen in een grafiek weergegeven kunnen worden.



Figuur B.6: Dichtheidsfunctie voor de discrete Poisson-verdeling met parameter $\lambda = 1.5$.

Omgekeerd laat zich een continue dichtheidsfunctie als een soort grensgeval van een dichtheidsfunctie zo als in Figuur B.6 opvatten. Als namelijk de mogelijke uitkomsten steeds dichter bij elkaar komen te liggen, worden de rechthoeken steeds smaller en lijkt de functie met stappen steeds meer op een gladde functie.

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

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

> De reden voor de naam *dichtheidsfunctie* ligt in het feit, dat we de kans op een waarde in een interval van breedte Δx voor kleine intervallen kunnen benaderen door $\Delta x \cdot f(x)$. Als we Δx als een *eenheidsinterval* zien, is $f(x)$ de dichtheid van de kansmassa rond x, net zo als we de dichtheid van een stof zien als de massa van een eenheidsvolume van de stof.

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

B.3 Definitie Voor een dichtheidsfunctie $f(x)$ met bijhorende continue kansverdeling $P(x \in [a, b]) = \int_a^b f(x)\, dx$ heet de functie

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

de *verdelingsfunctie* van de kansverdeling.

De verdelingsfunctie $F(a)$ heeft (onder meer) de volgende eigenschappen:

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

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

(iii) $P(x \in [a, b]) = F(b) - F(a)$.

(iv) $F'(a) = f(a)$, dus de verdelingsfunctie is de primitieve van de dichtheidsfunctie.

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

De uniforme verdeling

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

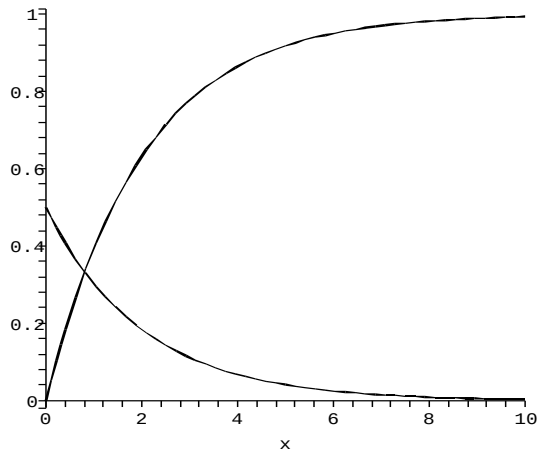
$$f(x) = \begin{cases} 0 & \text{als } x < a \\ \frac{1}{b-a} & \text{als } a \leq x \leq b \\ 0 & \text{als } x > b \end{cases} \quad \text{en} \quad F(x) = \begin{cases} 0 & \text{als } x < a \\ \frac{x-a}{b-a} & \text{als } a \leq x \leq b \\ 1 & \text{als } x > b \end{cases}$$

De exponentiële verdeling

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

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

Merk op dat de constante factor λ bij de exponentiële functie weer door de normering bepaald is, want $\int_0^\infty e^{-\lambda x}\, dx = \frac{-1}{\lambda} e^{-\lambda x}\big|_0^\infty = \frac{1}{\lambda}$.



Figuur B.7: Dichtheidsfunctie en verdelingsfunctie voor de exponentiële verdeling met $\lambda = 0.5$.

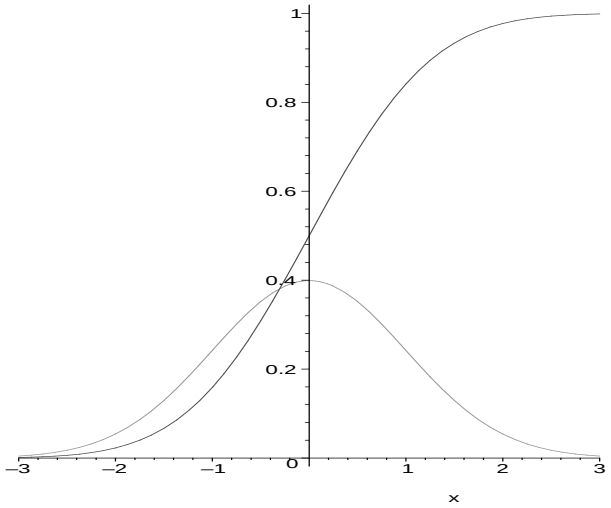
Iets algemener kan men ook een proces bekijken die niet op het tijdstip $x = 0$ begint, maar kans 0 heeft voor $x < c$ en voor $x \geq c$ exponentieel daalt. Dit betekent echter alleen maar een verschuiving op de x-as, de dichtheidsfunctie hiervoor is gewoon $\lambda e^{-\lambda(x-c)}$ in plaats van $\lambda e^{-\lambda x}$.

De normale verdeling (Gauss verdeling)

De belangrijkste continue verdeling is de normale verdeling die centraal in de statistiek staat. De dichtheidsfunctie die in Figuur B.8 afgebeeld is, heeft de vorm van een klok en is gegeven door

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

In dit geval kunnen we de verdelingsfunctie $F(x)$ alleen maar door de integraal van $f(x)$ beschrijven, omdat er geen gewone functie $F(x)$ is die $f(x)$ als afgeleide heeft.



Figuur B.8: Dichtheidsfunctie en verdelingsfunctie voor de standaard-normale verdeling

De normale verdeling met parameters $\mu = 0$ en $\sigma = 1$ noemen we *standaard-normale verdeling*. Voor de standaard-normale verdeling geldt dus

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}} \quad \text{en} \quad F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\, dx$$

De redenen voor de centrale rol van de normale verdeling zijn veelvoudig.

- De misschien belangrijkste reden wordt geformuleerd in de *Centrale limietstelling* die ruwweg zegt dat de combinatie van een aantal (onafhankelijke) toevallige gebeurtenissen bij benadering tot een normale verdeling leidt. We zullen dit in een latere les nader toelichten.

Een verdere reden is, dat voor zekere (voldoende grote) waarden van de parameters ook sommige discrete kansverdelingen goed door de normale verdeling benaderd worden.

- De binomiale verdeling $b(m, p; k)$ wordt door de normale verdeling met $\mu = mp$ en $\sigma^2 = mp(1-p)$ benadert. Deze benadering wordt meestal als redelijk beschouwd als $mp \geq 5$ en $m(1-p) \geq 5$ geldt.

- Voor de Poisson-verdeling geldt iets soortgelijks (omdat deze al een benadering voor de binomiale verdeling is). De Poisson-verdeling met parameter λ wordt benaderd door de normale verdeling met parameters $\mu = \lambda$ en $\sigma^2 = \lambda$. Hierbij wordt vaak de vuistregel $\lambda \geq 5$ voor de toepasbaarheid van de benadering gehanteerd (merk op dat we bij de Poisson-verdeling al veronderstellen dat p klein is, dus is $mp = \lambda \geq 5$ tegenover $m(1-p) \geq 5$ de sterkere eis).

Belangrijke begrippen in deze les

- discrete kansverdeling
- gelijkverdeling
- hypergeometrische verdeling
- binomiale verdeling
- Poisson-verdeling
- continue kansverdeling
- dichtheidsfunctie, verdelingsfunctie
- exponentiële verdeling
- normale verdeling

Opgaven

47. Een oneerlijke dobbelsteen is zo gemaakt dat 3 drie keer zo vaak valt als 4 en 2 twee keer zo vaak als 5. Verder vallen 1, 2, 3 en 6 even vaak.
    (i) Geef een kansverdeling voor het werpen van deze dobbelsteen aan.
    (ii) Bepaal de kans dat bij twee keer werpen van deze dobbelsteen de som minstens 11 is.

48. Bij een hockeytoernooi zijn er 18 teams aangemeld. In de eerste ronde worden de teams in twee groepen van 9 teams geloot. Onder de deelnemers zijn 5 teams uit de hoogste klasse. Hoe groot is de kans dat deze 5 teams in dezelfde groep terecht komen? Hoe groot is de kans dat er in een groep 2 en in de andere 3 teams uit de hoogste klasse terecht komen.

49. In een kast liggen n paren schoenen (dus $2n$ schoenen) willekeurig door elkaar. Je grijpt blindelings $k \leq n$ schoenen. Hoe groot is de kans dat je er minstens één passend paar uit vist? Hoe groot is de kans dat je precies één paar uit vist?

50. De kans dat een eerstejaars student in een bepaald vak afstudeert is 40%. Wat zijn de kansen dat uit een groep van 5 eerstejaars:
    (i) niemand afstudeert,
    (ii) precies 1 afstudeert,
    (iii) minstens 3 afstuderen?

51. Een test bestaat uit 10 ja-nee vragen. Iemand die van toeten nog blazen weet, besluit de vragen op goed geluk te beantwoorden (dit betekent dat hij voor elke vraag een kans van $\frac{1}{2}$ op een goed antwoord heeft). Met 6 goede antwoorden ben je in de test geslaagd. Wat is de kans voor onze kandidaat om de test te halen?

52. In Nijmegen zijn er 800 families met vijf kinderen. Hoeveel families met (a) 3 meisjes, (b) 5 meisjes, (c) 2 of 3 jongens verwacht je? (Je kunt ervan uit gaan dat er even veel jongens als meisjes geboren worden.)

53. In een vaas zitten 7 witte en 1 rode knikkers. Je trekt herhaald een knikker, bekijkt de kleur en legt hem vervolgens terug. Bepaal de kans dat je bij 8 pogingen precies 3 keer de rode knikker pakt. Gebruik hiervoor (a) de binomiale verdeling, (b) de benadering door de Poisson-verdeling.

    Hoe zit het met de resultaten als je 15 witte en 1 rode knikker hebt en 16 pogingen doet? En hoe zit het bij 79 witte en 1 rode knikker en 80 pogingen?

54. Volgens een statistiek vinden in Nederland per jaar 3 op de 100.000 mensen een portemonnee met meer dan 1000 €. Wat is de kans dat in en stad als Nijmegen (met 150.000 inwoners) dit geluk (a) 3, (b) 5, (c) 10, (d) hooguit 2 mensen overkomt.

55. Een Rad van avontuur heeft vier sectoren waarin het rad met dezelfde kans tot stilstand komt. Het rad wordt gedraaid tot dat het in sector I stopt, maar hooguit 10 keer. Bepaal de kansen voor de volgende gebeurtenissen:

    A_i : Het rad stopt bij de i-de draaiing in sector I.

    B : Het rad stopt helemaal niet in sector I.

    C : Het aantal draaiingen is even.

56. De goedkope random-trein vertrekt op een willekeurig tijdstip tussen 10.00 en 10.30 uur. Je beslist zelf ook op een willekeurig tijdstip in dit half uur op het station op te dagen en hooguit 5 minuten te wachten. Als de trein in dit interval niet komt, pak je een taxi om nog op tijd naar het college te komen. Wat is de kans dat je met de trein zult rijden?

Les 8 Verwachtingswaarde en spreiding

8.1 Stochasten

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

Vaak zijn de uitkomsten waarin we geïnteresseerd zijn veel eenvoudiger dan de uitkomstenruimte zelf, bijvoorbeeld kijken we naar het aantal k van defecte stukken in plaats van alle combinaties van m testresultaten, waarvan k negatief zijn. We kunnen dus zeggen, dat we verschillende uitkomsten die een zekere eigenschap gemeenschappelijk hebben in een cluster samenvatten. Zo'n eigenschap laat zich door een functie beschrijven, die aan elk element $\omega \in \Omega$ van de uitkomstenruimte een waarde $X(\omega) \in \mathbb{R}$ toekent. Uiteindelijk willen we dan de kans op alle uitkomsten bepalen, die dezelfde waarde $X(\omega)$ hebben.

B.4 Definitie Een functie

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

die aan de elementen van een uitkomstenruimte Ω waarden toewijst, heet een *random variable* (in het Engels), een *stochastische variabele*, een *kansvariabele* of kort een *stochast*.

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

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

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

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

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

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

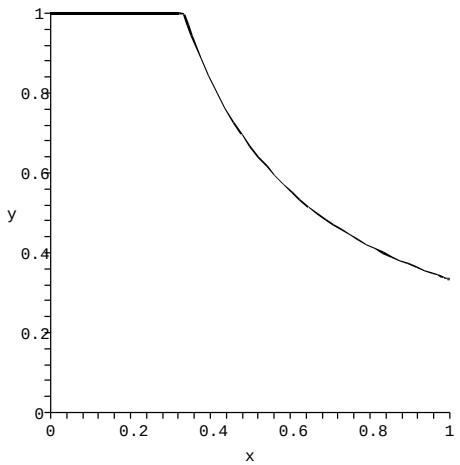
Voor continue kansverdelingen gaat de som over de uitkomsten met $X(\omega) = x$ over in een integraal. Omdat de kans op een enkele uitkomst steeds 0 is, wordt hier de kans bepaald, dat de stochast X een waarde beneden een gegeven grens aanneemt. Voor een continue kansverdeling met dichtheidsfunctie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

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

dus we meten de oppervlakte onder de kromme van $f(x)$ over het interval waar de stochast X een waarde van hoogstens x oplevert.

> **Merk op:** Meestal zijn continue stochasten door hun eigen dichtheidsfunctie aangegeven, dan geldt gewoon
>
> $$P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt.$$

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



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

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

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

8.2 Verwachtingswaarde

Als we in het casino roulette gaan spelen, zijn we er niet in geïnteresseerd of we in het eerste of laatste spel winnen of verliezen en ook niet hoe vaak we winnen of verliezen. Eigenlijk willen we alleen maar weten of we kunnen verwachten dat we aan het eind van de dag (of de nacht) met een winst naar huis komen.

Als we N keer spelen en bij elke keer 10€ op rood zetten, dan is bij elk spel de kans dat we 10€ winnen gelijk aan $\frac{18}{37}$, want er zijn 18 rode en 18 zwarte getallen en de groene 0. De kans dat we de 10€ verliezen is dus $\frac{19}{37}$. Als we heel vaak spelen, kunnen we verwachten dat we $\frac{18 \cdot N}{37}$ keer winnen en $\frac{19 \cdot N}{37}$ keer verliezen. Dit betekent dat we een verlies van $N \cdot \frac{1}{37} \cdot 10$€ kunnen verwachten.

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

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

B.5 Definitie Voor een stochast X definiëren we de *verwachtingswaarde* $E(X)$ (de E staat voor het Engelse *expectation*) door

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

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

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

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

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

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

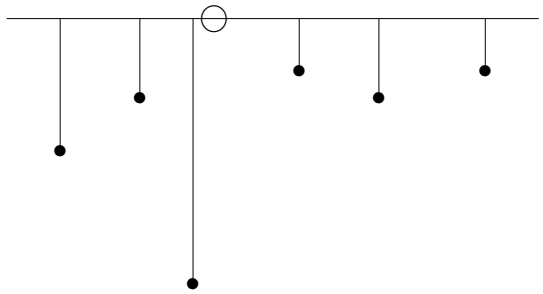
Als we nu de reële lijn in stukjes $[i \cdot \delta, (i+1) \cdot \delta]$ van lengte δ onderverdelen en de uitkomsten $x \in [i \cdot \delta, (i+1) \cdot \delta]$ tot de uitkomst $x = i \cdot \delta$ samenvatten, hebben we alleen maar nog de discrete verzameling $\{i \cdot \delta \mid i \in \mathbb{Z}\}$ van uitkomsten. Voor deze *gediscretiseerde* stochast is de verwachtingswaarde gegeven door

$$\sum_{i \in \mathbb{Z}, x = i \cdot \delta} x \cdot P(X \in [x, x+\delta]) \approx \sum_{i \in \mathbb{Z}, x = i \cdot \delta} x \cdot \delta \cdot f(x)$$

en dit is juist de discrete benadering van de integraal $\int_{-\infty}^{\infty} x \cdot f(x)\, dx = E(X)$.

We kunnen de verwachtingswaarde aanschouwelijk zien als het evenwichtspunt van een balk (oneindig lang, zonder gewicht), waar we in het punt x een

gewicht van massa $P(x)$ aan hangen. Het evenwichtspunt is dan juist het punt $E(X)$. In het plaatje in Figuur B.9 zijn de gewichten gerepresenteerd door de lengten van de verticale ribben.



Figuur B.9: Verwachtingswaarde als evenwichtspunt van een balk

Voordat we de verwachtingswaarde voor de meest belangrijke kansverdelingen bepalen, kunnen we al een aantal elementaire eigenschappen algemeen uit de definitie afleiden.

Als X en Y stochasten zijn, dan geldt:

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

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

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

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

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

Binomiale verdeling

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

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

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

Hypergeometrische verdeling

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

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

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

Poisson-verdeling

We hebben $P(X = k) = po_\lambda(k) = \frac{\lambda^k}{k!}e^{-\lambda}$ en maken gebruik van de relatie $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^\lambda$ (die we hier niet nader kunnen toelichten, in feite is dit een manier om de exponentiële functie te definiëren):

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

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

Uniforme verdeling

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

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

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

Exponentiële verdeling

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

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

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

Normale verdeling

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

8.3 Spreiding

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

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

Als we het traagheidsmoment terug naar de stochast X vertalen, wordt de massa m de kans $P(X = x)$ en de afstand r wordt het verschil $x - E(X)$ met de verwachtingswaarde. Als analogie met het traagheidsmoment definiëren we de *variantie* of *spreiding* van de stochast X:

B.6 Definitie Voor een stochast X heet

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

de *variantie* of *spreiding* van X.

De variantie is de verwachtingswaarde van de kwadratische afstand van de stochast van zijn verwachtingswaarde en is dus een maat ervoor hoe dicht de waarden van een stochast bij de verwachtingswaarde liggen.

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

B.7 Definitie Voor een stochast X met variantie $Var(X)$ heet

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

de *standaardafwijking* van X.

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

Net als voor de verwachtingswaarde kunnen we ook voor de variantie van een stochast X een aantal belangrijke eigenschappen meteen uit de definities afleiden:

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

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

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

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

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

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

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

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

Binomiale verdeling

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

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

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

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

Hypergeometrische verdeling

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

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

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

Poisson-verdeling

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

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

We hebben dus

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

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

Uniforme verdeling

Er geldt

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

dus hebben we

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

Exponentiële verdeling

Er geldt (weer met partiële integratie)

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

want $E(X) = \int_0^\infty x\lambda e^{-\lambda x}\, dx$ en we wisten al dat $E(X) = \frac{1}{\lambda}$. We hebben dus

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

Normale verdeling

Voor de normale verdeling is het helaas iets lastiger om de variantie expliciet te berekenen, en we zullen dat hier ook niet uitwerken. Het resultaat is echter makkelijk te onthouden, de parameters μ en σ in de dichtheidsfunctie $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ van de normale verdeling zijn juist zo gekozen dat σ^2 de variantie aangeeft en dus σ de standaardafwijking.

8.4 De ongelijkheid van Chebyshev

We hebben boven opgemerkt dat de variantie van een stochast aangeeft hoe sterk de uitkomsten van de verwachtingswaarde afwijken. Deze samenhang tussen verwachtingswaarde en spreiding kunnen we heel expliciet aangeven, namelijk in de *Ongelijkheid van Chebyshev*. Hierbij maken we een schatting voor de kans dat een uitkomst een grotere afstand dan $a > 0$ van de verwachtingswaarde $E(X)$ heeft.

Volgens de definitie berekenen we de variantie door

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

Als we de som beperken tot de waarden van x met $|x - E(X)| \geq a$, maken we de waarde van de som kleiner, omdat we niet-negatieve termen weglaten en we krijgen:

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

waarbij we in de tweede stap $|x - E(X)|$ door a naar beneden afschatten. Maar er geldt

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

dus hebben we het volgende bewezen:

B.8 Ongelijkheid van Chebyshev *Voor een stochast X met verwachtingswaarde $E(X)$ en variantie $Var(X)$ geldt voor elke $a > 0$ de ongelijkheid*

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

d.w.z. de kans dat een waarde van de stochast X sterker dan a van de verwachtingswaarde afwijkt neemt met het kwadraat van a af.

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

In de statistiek wordt vaak als vuistregel de zogeheten 2σ-regel gebruikt: Voor een stochast X met standaardafwijking σ_X liggen meestal 95% van de gebeurtenissen in het interval $(E(X) - 2\sigma_X, E(X) + 2\sigma_X)$.

De ongelijkheid van Chebyshev geeft aan dat dit interval minstens 75% van de gebeurtenissen bevat, want $P(|X - E(X)| \geq 2\sigma_X) \leq \frac{1}{4\sigma_X^2} Var(X) = \frac{1}{4}$, omdat $\sigma_X = \sqrt{Var(X)}$. Maar voor de meeste kansverdelingen (in het bijzonder voor de normale verdeling) geldt de sterkere uitspraak van de 2σ-regel.

8.5 Covariantie en correlatie

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

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

B.9 Definitie De grootheid $E(X \cdot Y) - E(X) \cdot E(Y)$ heet de *covariantie* van X en Y en wordt genoteerd met $Cov(X, Y)$.

Volgens de relatie

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

geeft de covariantie aan hoe sterk de variantie van de som van twee stochasten afwijkt van de som van de varianties.

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

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

dus hebben we

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

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

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

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

We hebben dus gezien:

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

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

We hebben gezien dat de covariantie $Cov(X, Y)$ in zekere zin en maat voor de afhankelijkheid van X en Y is. Er laat zich aantonen dat

$$|Cov(X,Y)| \leq \sigma_X \sigma_Y,$$

dus is de covariantie van twee stochasten begrensd door het product van de standaardafwijkingen van de stochasten. Met behulp van de standaardafwijkingen kunnen we dus de covariantie op waarden tussen -1 en 1 normeren.

B.10 Definitie We noemen

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

de *correlatiecoëfficiënt* van X en Y.

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

Belangrijke begrippen in deze les

- stochasten
- verwachtingswaarde
- variantie, standaardafwijking
- ongelijkheid van Chebyshev
- covariantie, correlatiecoëfficiënt

Opgaven

57. Er wordt met twee (eerlijke) dobbelstenen gedobbeld. De stochast X beschrijft het maximale getal in een worp. Bereken $P(X = k)$ voor $k = 1, \ldots, 6$ en de verwachtingswaarde $E(X)$.

    Bekijk hetzelfde probleem voor drie dobbelstenen.

58. Bij een bloedtest van 10 personen is bekend dat 2 een zeker virus in hun bloed hebben. Om het aantal tests in te krimpen wordt te volgende methode toegepast: De 10 personen worden willekeurig in twee groepen van 5 personen ingedeeld. Het bloed van de personen in een groep wordt vermengd en getest. Als het virus in het mengsel gevonden wordt, wordt het bloed van elke persoon in de groep apart getest.

    Beschrijf een geschikte ruimte Ω met een kansverdeling P, zo dat het aantal van bloedtests een stochast op deze kansruimte is. Bereken de verwachtingswaarde voor het aantal bloedtests.

59. Bij een spel met een dobbelsteen win je n€ als je n dobbelt en n even is en je verliest n€ als n oneven is. Wat is de verwachtingswaarde van je winst/verlies.

60. Bij het skaat spel krijg je 10 kaarten uit een kaartspel met 32 kaarten (8 soorten, 4 kleuren). Wat is de verwachtingswaarde voor het aantal boeren dat je krijgt?

61. In een loterij heb je 70% nieten en 30% winnende lotjes. Iemand beslist zo lang lotjes te kopen tot dat hij een winnende lot krijgt, maar hooguit vijf keer. Wat kan hij voor een uitgave verwachten, als een lot 2€ kost?

62. De kans dat een student bij het grote lustrumfeest een bier krijgt is 99.2% (soms is het bier op, soms denkt de baas dat de student geen 16 jaar oud is). Een slimme verzekeringsmaatschappij biedt eenmalig een verzekeringspolis, waar je voor een premie van 10 € tegen bierarmoede verzekerd bent. In het geval dat je inderdaad geen bier op het feest krijgt betaalt de verzekering 1000 €. Wat is de verwachte winst van de verzekeringsmaatschappij bij elke afgesloten polis?

63. Je koopt een nieuwe speelautomaat voor je kroeg. In de automaat draaien twee onafhankelijke wielen die in tien even grote segmenten zijn opgedeeld en volgens een gelijkverdeling in een van de segmenten stoppen. De segmenten hebben de nummers 1 t/m 10. Een speler heeft alleen maar de volgende winstmogelijkheden (bij alle andere uitkomsten verliest hij zijn inzet):

    - Als beide wielen 10 tonen wint hij 5€.
    - Als beide wielen hetzelfde getal maar niet 10 tonen wint hij 2€.
    - Als precies een van de wielen 10 toont wint hij 1€.

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

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

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

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

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

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

Les 9 Voorwaardelijke kansen, de Regel van Bayes en onafhankelijkheid

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

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

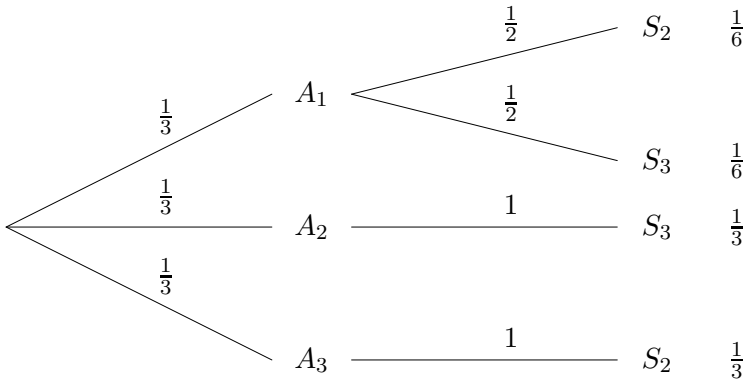
Intuïtief zullen veel mensen denken, dat na het openen van een van de deuren met een geit daarachter de kans 50 : 50 is, dat de auto achter de door de kandidaat gekozen deur staat. Dus zou het niets uitmaken of de kandidaat wisselt of niet. In de VS heeft een journaliste, Marilyn vos Savant, de oplossing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

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

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

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

Deze situatie kunnen we door het volgende boomdiagram beschrijven:



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

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

9.1 Voorwaardelijke kansen

Het idee dat de kans voor een uitkomst kan veranderen als we aanvullende informatie hebben, is zo natuurlijk dat we er meestal niet over nadenken. Bijvoorbeeld kan de kans op vorst op 30 april over de afgelopen 150 jaar eenvoudig afgelezen worden uit de tabellen van de weerkundige dienst. Als er bijvoorbeeld 10 keer in de afgelopen 150 jaren vorst op 30 april was, kunnen we aannemen dat de kans op vorst op 30 april 2007 ongeveer 6.67% is. Als aanvullende informatie kunnen we gebruiken dat er ook 10 keer vorst op 29 april is geweest en dat er in 5 jaren vorst op 29 en 30 april gevallen is. Zo ver maakt dit nog geen verschil voor de kans op vorst op 30 april 2007. Maar als er inderdaad vorst op 29 april 2007 valt, kunnen we zeggen dat de kans op vorst op 30 april 2007 opeens 50% is, want in 5 van de 10 jaren met vorst op 29 april was er ook vorst op 30 april.

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

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

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

B.11 Definitie Voor een kansverdeling P op Ω en $B \subseteq \Omega$ met $P(B) > 0$ noemen we

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

de *voorwaardelijke kans voor A, gegeven B.*

Voor $P(B) = 0$ is het onzin een kans onder de voorwaarde B te bekijken, want B gebeurt nooit.

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

Om te rechtvaardigen, dat we $P(A \mid B)$ een *kans* noemen, moeten we even nagaan dat $P(\cdot \mid B)$ voor $P(B) > 0$ een kansverdeling is, waarbij we natuurlijk erop terug mogen vallen dat $P(\cdot)$ een kansverdeling op Ω is.

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

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

(iii) Voor $A_1, A_2 \subseteq \Omega$ met $A_1 \cap A_2 = \emptyset$ geldt $(A_1 \cup A_2) \cap B = (A_1 \cap B) \cup (A_2 \cap B)$. Verder is in dit geval $(A_1 \cap B) \cap (A_2 \cap B) = \emptyset$ omdat $A_1 \cap B$ een deelverzameling van A_1 en $A_2 \cap B$ een deelverzameling van A_2 is. Daarom geldt:

$$\begin{aligned} P(A_1 \cup A_2 \mid B) &= \frac{P((A_1 \cup A_2) \cap B)}{P(B)} = \frac{P((A_1 \cap B) \cup (A_2 \cap B))}{P(B)} \\ &= \frac{P(A_1 \cap B) + P(A_2 \cap B)}{P(B)} = \frac{P(A_1 \cap B)}{P(B)} + \frac{P(A_2 \cap B)}{P(B)} \\ &= P(A_1 \mid B) + P(A_2 \mid B). \end{aligned}$$

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

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

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

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

We kunnen voorwaardelijke kansen niet alleen maar voor een enkele voorwaarde maar ook algemeen voor n voorwaarden definiëren. Het idee hierbij is hetzelfde, we kijken naar de kans van het gemeenschappelijke optreden van de voorwaarden met een gebeurtenis, gedeeld door de kans voor de voorwaarden en krijgen dus:

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

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

Omgekeerd kunnen we de kans voor het gemeenschappelijke optreden van gebeurtenissen (iteratief) door voorwaardelijke kansen uitdrukken en krijgen zo de zogeheten *kettingregel* die in veel toepassingen handig blijkt:

$$\begin{aligned} P(A_1, A_2) &= P(A_2 \mid A_1) \cdot P(A_1), \\ P(A_1, A_2, A_3) &= P(A_3|A_1, A_2) \cdot P(A_1, A_2) = P(A_3|A_1, A_2) \cdot P(A_2|A_1) \cdot P(A_1) \end{aligned}$$

en in het algemeen

$$P(A_1, ..., A_n) = P(A_n|A_1, ..., A_{n-1}) \cdot P(A_{n-1}|A_1, ..., A_{n-2}) \cdot \ldots \cdot P(A_2|A_1) \cdot P(A_1).$$

9.2 Regel van Bayes

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

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

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

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

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

De Regel van Bayes wordt vaak op een iets slimmere manier toegepast. Hiervoor wordt de deelverzameling $B \subseteq \Omega$ in verschillende gevallen onderverdeeld die elkaar uitsluiten, dus we schrijven $B = \cup_{i=1}^{n} B_i$ met $B_i \cap B_j = \emptyset$ als $i \neq j$.

Een belangrijk speciaal geval hiervoor is $B = B_1 \cup B_2$ met $B_2 = B \setminus B_1 = B_1^c$. We noemen B_2 het *complement* van B_1 in B.

Er geldt:

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

en dus

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

In het bijzonder kunnen we in het geval $A \subseteq B$ de *totale kans* $P(A)$ berekenen door

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

en het belangrijkste geval hiervoor is $B = \Omega$, d.w.z. we delen alle mogelijke uitkomsten in een aantal klassen van uitkomsten op.

In het speciaal geval van de opsplitsing van Ω in een deelverzameling $B_1 \subseteq \Omega$ en zijn complement $B_2 = \Omega \setminus B_1 = B_1^c$ geeft dit

$$P(A) = P(A \mid B_1) \cdot P(B_1) + P(A \mid B_1^c) \cdot P(B_1^c).$$

We kunnen nu de Regel van Bayes algemeen formuleren:

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

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

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

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

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

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

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

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

Voorbeeld 2: Een student moet bij een tentamen een multiple-choice vraag met n mogelijkheden oplossen. Als hij voorbereid is, zal zijn antwoord juist zijn, als niet zal hij willekeurig een antwoord gokken en dus een kans van $\frac{1}{n}$ op een juist antwoord hebben. De kans dat de student voorbereid is, zij p. Voor de docent is het nu interessant om de kans te bepalen, dat de student inderdaad voorbereid was, als hij een juist antwoord heeft gegeven. Als we een juist antwoord met J en een voorbereide student met V noteren, hebben we dus:

$$\begin{aligned} P(V \mid J) &= \frac{P(J \mid V) \cdot P(V)}{P(J \mid V) \cdot P(V) + P(J \mid V^c) \cdot P(V^c)} \\ &= \frac{1 \cdot p}{1 \cdot p + \frac{1}{n}(1-p)} = \frac{np}{np + (1-p)}. \end{aligned}$$

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

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

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

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

als relatieve frequenties op heel grote tekst-corpora (bijvoorbeeld 10 jaar NRC Handelsblad) bepaald.

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

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

$$\begin{aligned} P(A_1 \mid S_2) &= \frac{P(S_2 \mid A_1) \cdot P(A_1)}{P(S_2 \mid A_1) \cdot P(A_1) + P(S_2 \mid A_2) \cdot P(A_2) + P(S_2 \mid A_3) \cdot P(A_3)} \\ &= \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 0 + 1 \cdot \frac{1}{3}} = \frac{1}{3}. \end{aligned}$$

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

$$\begin{aligned} P(A_3 \mid S_2) &= \frac{P(S_2 \mid A_3) \cdot P(A_3)}{P(S_2 \mid A_1) \cdot P(A_1) + P(S_2 \mid A_2) \cdot P(A_2) + P(S_2 \mid A_3) \cdot P(A_3)} \\ &= \frac{1 \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 0 + 1 \cdot \frac{1}{3}} = \frac{2}{3}. \end{aligned}$$

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

9.3 Onafhankelijkheid

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

B.12 Definitie Twee uitkomsten $A \subseteq \Omega$ en $B \subseteq \Omega$ heten *onafhankelijk* als $P(A) = P(A \mid B)$. Equivalent hiermee is dat $P(A \cap B) = P(A) \cdot P(B)$.

De equivalentie van de twee formuleringen volgt uit de definitie van de voorwaardelijke kans, want wegens $P(A \cap B) = P(A \mid B) \cdot P(B)$ is

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

Omdat ook $P(A \cap B) = P(B \mid A) \cdot P(A)$ geldt, volgt hieruit ook dat

$$P(A) = P(A \mid B) \Leftrightarrow P(B) = P(B \mid A),$$

d.w.z. het maakt niets uit welke van de voorwaardelijke kansen $P(A \mid B)$ of $P(B \mid A)$ we bekijken.

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

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

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

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

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

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

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

$P(A \cap B^c) = P(A \cup B) - P(B) = P(A) + P(B) - P(A \cap B) - P(B) = P(A) - P(A \cap B) = P(A) - P(A) \cdot P(B) = P(A)(1 - P(B)) = P(A) \cdot P(B^c)$. Dit werkt evenzo voor $P(A^c \cap B)$.

$P(A^c \cap B^c) = P((A \cup B)^c) = 1 - P(A \cup B) = 1 - P(A) - P(B) + P(A \cap B) = 1 - P(A) - P(B) + P(A) \cdot P(B) = (1 - P(A))(1 - P(B)) = P(A^c) \cdot P(B^c)$.

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

B.13 Definitie Twee stochasten X en Y heten onafhankelijk als

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

voor alle paren (x, y) van uitkomsten voor X en Y geldt.

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

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

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

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

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

A_1 : de eerste dobbelsteen toont een oneven getal,

A_2 : de tweede dobbelsteen toont een oneven getal,

A_3 : de som van de getallen is even.

We hebben $P(A_1) = P(A_2) = P(A_3) = \frac{1}{2}$ en $P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) = \frac{1}{4}$, dus zijn de uitkomsten paarsgewijs onafhankelijk. Maar $P(A_1 \cap A_2 \cap A_3) = P(A_1 \cap A_2)$ omdat de som van twee oneven getallen even is, dus is $P(A_1 \cap A_2 \cap A_3) \neq P(A_1) \cdot P(A_2) \cdot P(A_3) = \frac{1}{8}$ en dus zijn de drie uitkomsten niet onafhankelijk.

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

A_1 : de eerste worp toont kop,

A_2 : er valt vaker kop dan munt,

A_3 : de laatste twee worpen leveren hetzelfde resultaat.

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

9.4 Bernoulli-model

Een belangrijke toepassing van de onafhankelijkheid van uitkomsten is de herhaalde uitvoering van een experiment. We nemen aan dat we in de uitkomstenruimte Ω een deelverzameling $A \subseteq \Omega$ van gunstige uitkomsten hebben. Bij de eenmalige uitvoering van het experiment is de kans op een gunstige uitkomst

gegeven door $p = \frac{|A|}{|\Omega|}$. De kans voor een ongunstige uitkomst is dan $1-p$. Als we het experiment twee keer uitvoeren is de kans dat we twee gunstige uitkomsten hebben de kans van de doorsnede van een gunstige uitkomst bij de eerste keer en een gunstige uitkomst bij de tweede keer. Omdat we ervan uitgaan dat het eerste en het tweede experiment onafhankelijk zijn, kunnen we de kans voor de doorsnede als product van de enkele kansen berekenen, dus als $p \cdot p = p^2$.

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

Als we ervan uitgaan dat herhalingen van een experiment onafhankelijk van elkaar zijn, dan is de kans op k gunstige uitkomsten bij m herhalingen gegeven door de binomiale verdeling:

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

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

De beschrijving van uitkomsten van een stochast door onafhankelijke herhaling van een experiment noemt men ook het *Bernoulli-model* voor de stochast.

Belangrijke begrippen in deze les

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

Opgaven

65. Er wordt met twee dobbelstenen gedobbeld. Gegeven de informatie dat de twee dobbelstenen verschillende getallen tonen (bijvoorbeeld in een spel waar je bij gelijke getallen nog een keer dobbelt), wat is de kans dat de som oneven is?
66. In vaas I zitten 3 rode en 5 witte knikkers, in vaas II zijn er 4 rode en 2 witte. Er wordt een knikker willekeurig uit vaas I gegrepen en in vaas II gelegd. Vervolgens wordt er een knikker uit vaas II getrokken. Wat is de kans dat deze knikker wit is?
67. In een vaas zitten 3 rode en 2 blauwe knikkers, in een tweede vaas zitten 2 rode en 8 blauwe knikkers. Er wordt eerst een munt geworpen om te bepalen uit welke vaas een knikker getrokken wordt: als kop valt uit de eerste en als munt valt uit de tweede.

(i) Bepaal de kans dat een rode knikker getrokken wordt.

(ii) Stel dat je niet hebt gezien of kop of munt gevallen is, maar wel dat een rode knikker getrokken wordt. Wat is de kans dat kop is gevallen, dus dat de knikker uit de eerste vaas is getrokken?

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

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

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

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

71. Een huis is voorzien met een alarminstallatie. Als er een inbraak is, zal er met 96% kans een alarm komen, maar ook als er geen inbraak is, is er (door aardbevingen of andere storingen) met een kans van 0.1% een alarm. In de woonwijk van het huis is de kans op een inbraak 0.3%. Vannacht is er een alarm. Hoe groot is de kans dat er daadwerkelijk een inbraak plaats vindt?

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

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

73. Er wordt twee keer met een eerlijke dobbelsteen gedobbeld. De uitkomsten A, B en C zijn:

    A : Er wordt twee keer hetzelfde getal gedobbeld.

    B : Het eerste getal is 1 of 6.

    C : Het tweede getal is even.

Zijn A, B, C onafhankelijk? Zijn de uitkomsten paarsgewijs onafhankelijk?

74. Er wordt twee keer met een dobbelsteen gedobbeld. De stochast X_1 beschrijft het aantal ogen in de eerste worp, de stochast X_2 het aantal ogen in de tweede. Verder geeft $U := \min(X_1, X_2)$ de kleinste en $V := \max(X_1, X_2)$ de grootste van de twee worpen.

    (i) Zijn U en V onafhankelijk?

    (ii) Bepaal de kansverdeling van U.

    (iii) Bepaal de verwachtingswaarden $E(U)$ en $E(U + V)$.

    (iv) Bepaal de voorwaardelijke kans $P(X_1 = 3 \mid U = 3)$.

Les 10 Schatten en simuleren

10.1 Maximum likelihood schatting

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

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

Een algemene methode om parameters van kansverdelingen te schatten is gebaseerd op het volgende argument: Voor elke keuze van een parameter (of meerdere parameters) heb je een kansverdeling, die aan een waargenomen resultaat een zekere kans geeft. In het voorbeeld is dit

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

Bij onafhankelijke herhaling kunnen we de kans voor een rij observaties als product van de kansen voor de aparte observaties berekenen, in het voorbeeld hebben we dus

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

De kans voor de rij $(k_1, \ldots, k_n)$ van observaties is nu een functie van de parameter p (die we door de index p bij P_p aanduiden) en we noemen deze functie de *aannemelijkheidsfunctie* of *likelihood* functie. We maken nu een schatting voor p door te zeggen, dat we p zo kiezen dat de aannemelijkheidsfunctie aan

maximale waarde heeft, dus dat de kans voor onze observatie maximaal wordt. Deze methode van schatting noemt men de *meest aannemelijke* of *maximum likelihood* schatting van de parameter.

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

Discrete kansverdelingen

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

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

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

We definiëren nu de log-likelihood functie $L(p)$ door

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

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

Een maximum van $L(p)$ vinden we door de nulpunten van de afgeleide $L'(p)$ (met betrekking tot p) te bepalen. Er geldt

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

en we hebben

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

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

Dit betekent dat we de parameter p als de relatieve frequentie van gunstige uitkomsten in alle steekproeven bij elkaar genomen kiezen. Dit komt op de

pragmatische keuze neer, maar we hebben nu een betere onderbouwing voor onze keuze. Het is namelijk de parameter die de observaties het beste verklaart.

Poisson-verdeling: Voor een (zeldzaam) gebeurtenis dat volgens een Poisson-verdeling met parameter λ optreedt is de kans dat we het gebeurtenis k keer waarnemen gegeven door $P(k) = \frac{\lambda^k}{k!}e^{-\lambda}$.

Stel we zien dit gebeurtenis over verschillende onafhankelijke experimenten $k_1, \ldots, k_n$ keer gebeuren, dan krijgen we de likelihood functie

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

We definiëren nu de log-likelihood functie

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

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

De afgeleide (naar λ) hiervan is

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

en we hebben

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

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

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

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

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

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

Als $q(n) \geq 1$ is, stijgt $h(n,m,s;k)$ van $n-1$ naar n, als $q(n) \leq 1$ is, daalt $h(n,m,s;k)$ van $n-1$ naar n. Om de waarde van n te vinden waarvoor $h(n,m,s;k)$ maximaal is, moeten we dus kijken waar $q(n)$ van een waarde ≥ 1 naar een waarde ≤ 1 wisselt. Er geldt:

$$\begin{aligned} q(n) &= \frac{\binom{s}{k}\binom{n-s}{m-k}}{\binom{n}{m}} \cdot \frac{\binom{n-1}{m}}{\binom{s}{k}\binom{n-1-s}{m-k}} = \frac{\frac{(n-s)!}{(m-k)!(n-s-m+k)!} \cdot \frac{(n-1)!}{m!(n-1-m)!}}{\frac{n!}{m!(n-m)!} \cdot \frac{(n-1-s)!}{(m-k)!(n-1-s-m+k)!}} \\ &= \frac{(n-s)!(n-1)!(n-m)!(n-1-s-m+k)!}{(n-s-m+k)!(n-1-m)!n!(n-1-s)!} = \frac{(n-s)(n-m)}{(n-s-m+k)n} \\ &= \frac{n^2 - sn - nm + sm}{n^2 - sn - mn + kn}. \end{aligned}$$

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

Continue kansverdelingen

Ook voor continue kansverdelingen kunnen we op basis van een rij waarnemingen een likelihood (of log-likelihood) functie definiëren. Het verschil tegenover de discrete kansverdelingen is dat in dit geval de likelihood functie op de dichtheidsfunctie gebaseerd is. Dit zien we als volgt in:

Voor een stochast X met dichtheidsfunctie $f(x)$ is de kans voor een gebeurtenis met $X \in [x_0, x_0 + \delta]$ gegeven door

$$P(X \in [x_0, x_0 + \delta]) = \int_{x_0}^{x_0+\delta} f(x)\, dx.$$

Maar als we aannemen dat δ klein is, kunnen we ook aannemen dat $f(x)$ op het interval $[x_0, x_0 + \delta]$ constant is, en dan is de oppervlakte onder de grafiek van $f(x)$ gewoon de rechthoek met breedte δ en hoogte $f(x_0)$, dus is de kans

$$P(X \in [x_0, x_0 + \delta]) \approx f(x_0) \cdot \delta.$$

Als we nu waarnemingen $x_1, \ldots, x_n$ voor een stochast X met dichtheidsfunctie $f(x)$ hebben en altijd met intervallen van dezelfde breedte δ werken, krijgen we als likelihood functie:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} \left(\int_{x_i}^{x_i+\delta} f(x)\, dx\right) \approx \prod_{i=1}^{n} (f(x_i) \cdot \delta) = \delta^n \cdot \prod_{i=1}^{n} f(x_i).$$

Maar de δ hangt niet van de parameters van de dichtheidsfunctie $f(x)$ af, dus kunnen we ook rechtstreeks naar het product

$$F(x_1, \ldots, x_n) := \prod_{i=1}^{n} f(x_i)$$

van de waarden van de dichtheidsfunctie kijken.

Exponentiële verdeling: Voor een gebeurtenis dat volgens een exponentiële verdeling met parameter λ optreedt, is de dichtheidsfunctie $f(x)$ gegeven door

$$f(x) = \lambda e^{-\lambda x}.$$

Stel we maken voor een stochast met een exponentiële verdeling de observaties $x_1, \ldots, x_n$. Dan is de likelihood functie gegeven door

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

We definiëren nu de log-likelihood functie $L(\lambda)$ door

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

De afgeleide (naar λ) hiervan is

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

en we hebben

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

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

Normale verdeling: De normale verdeling wordt in de opgaven behandeld.

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

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

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

10.2 Simulatie

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

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

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

Randomgenerator

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

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

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

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

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

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

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

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

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

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

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

In een iets systematischere test deelt men het interval $[0, 1)$ in d (even grote) deelintervallen, telt hoe veel van $U_1, U_2, \ldots, U_n$ in elk van die deelintervallen ligt en toetst deze verdeling met een χ^2-test tegen de gelijkverdeling.

> De χ^2-toets is een standaardtoets uit de statistiek die toetst of de gevonden verdeling te veel of te weinig van de gelijkverdeling afwijkt. Een te grote afwijking geeft evidentie tegen de hypothese dat alle U_i uniform verdeeld zijn, een zeer kleine afwijking laat aan de onafhankelijkheid twijfelen, want het is ook erg onwaarschijnlijk, dat in elk deelinterval *precies* d/n getallen terecht komen.

Een soortgelijke test kan men in plaats van op enkele toevalsgetallen ook op paren of algemener op k-dimensionale vectoren $(U_1, \ldots, U_k)$, $(U_{k+1}, \ldots, U_{2k})$, $\ldots$, $(U_{(n-1)k+1}, \ldots, U_{nk})$ toepassen, die gelijkverdeeld in de k-dimensionale kubus $[0, 1)^k$ moeten zijn.

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

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

Er zijn een aantal algemene principes, hoe we een gewenste kansverdeling met behulp van een randomgenerator kunnen simuleren. De meest belangrijke zijn de methode van de *inverse verdelingsfunctie* en de *wegwerp* (rejection) methode. In principe zijn deze methoden voor continue kansverdelingen geformuleerd, maar ze zijn net zo goed op discrete kansverdelingen toepasbaar, omdat zich ook een discrete kansverdelingen door een verdelingsfunctie $F(x)$ en

een dichtheidsfunctie $f(x)$ laat beschrijven. Maar voor zekere discrete kansverdelingen zullen we later nog andere (meer directe) methoden aangeven.

Simulatie met behulp van de inverse verdelingsfunctie

Voor een algemene (continue) kansverdeling met dichtheidsfunctie $f(x)$ en verdelingsfunctie $F(x) = \int_{-\infty}^{x} f(t)\,dt$ passen we de inverse F^{-1} van de verdelingsfunctie op de uniforme verdeling toe: Zij U een stochast met uniforme verdeling op $[0,1)$, dan definiëren we een nieuwe stochast X door $X := F^{-1}(U)$, dus $F(X) = U$. Voor de kans $P(X \leq a)$ geldt nu

$$P(X \leq a) = P(F(X) \leq F(a)) = P(U \leq F(a)) = F(a)$$

omdat U uniform verdeeld is. De stochast X heeft dus juist de verdelingsfunctie $F(x)$.

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

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

Voorbeeld 2: Ook voor de *exponentiële verdeling* krijgen we op deze manier een simulatie. Na een mogelijke verschuiving op de x-as heeft de exponentiële verdeling de dichtheidsfunctie $f(x) = \lambda e^{-\lambda x}$ en de verdelingsfunctie $F(x) = 1 - e^{-\lambda x}$. Omdat $y = 1 - e^{-\lambda x} \Leftrightarrow -\lambda x = \log(1-y) \Leftrightarrow x = -\frac{1}{\lambda}\log(1-y)$, hebben we

$$F^{-1}(y) = -\frac{1}{\lambda}\log(1-y).$$

Voor een uniform verdeelde stochast U is dus $X := -\frac{1}{\lambda}\log(1-U)$ exponentieel verdeeld met parameter λ. Maar omdat met U ook $1-U$ gelijkverdeeld op $[0,1)$ is, kunnen we net zo goed

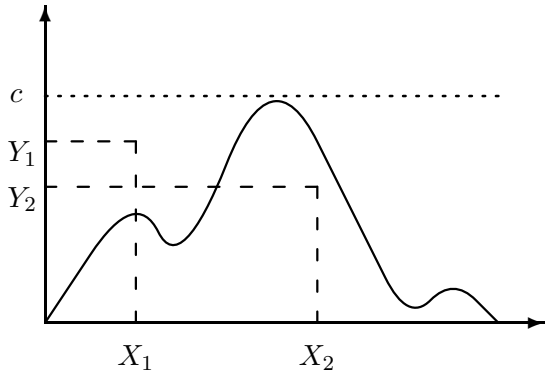
$$X := -\frac{1}{\lambda}\log(U)$$

kiezen om een exponentiële verdeling te simuleren.

Simulatie met behulp van de wegwerp methode

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

Maar we kunnen een kansverdeling met dichtheidsfunctie $f(x)$ op een eindig interval $[a,b]$ als volgt simuleren: Stel de dichtheidsfunctie is op het interval $[a,b]$ door een waarde c begrensd, d.w.z. $f(x) \leq c$ voor alle $x \in [a,b]$. Dan produceren we een rij toevalsgetallen (X_i) volgens een gelijkverdeling op $[a,b]$ en een tweede rij (Y_i) volgens een gelijkverdeling op $[0,c]$. We accepteren nu alleen maar de X_i voor de indices i waarvoor geldt dat $Y_i \leq f(X_i)$ en werpen



Figuur B.10: Simulatie met behulp van de wegwerp methode

de andere X_i weg. In het voorbeeld in Figuur B.10 zou men bijvoorbeeld X_1 verwerpen maar X_2 accepteren.

Het is niet moeilijk om in te zien dat de geaccepteerde toevalsgetallen X_i de dichtheidsfunctie $f(x)$ hebben, want een waarde $X_i = x$ wordt juist met kans $\frac{f(x)}{c}$ geaccepteerd. Om deze methode zo efficiënt mogelijk te maken, wordt de waarde c zo klein mogelijk gekozen, want de kans op het *wegwerpen* van een getal X_i is juist de oppervlakte tussen de grafiek van $f(x)$ en de lijn $y = c$.

Simulatie van speciale verdelingen

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

Discrete gelijkverdeling: Voor een eindige uitkomstenruimte Ω met $|\Omega| = n$ kunnen we aannemen dat $\Omega = \{0, \ldots, n-1\}$. We krijgen een gelijkverdeling op Ω door

$$X := \lfloor n \cdot U_i \rfloor,$$

waarbij $\lfloor x \rfloor$ het grootste gehele getal is dat $\leq x$ is.

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

Voor de binomiale verdeling $b(m, p; k)$ herhalen we m keer een simulatie met kans p en krijgen:

$$X := \sum_{i=1}^{m} \lfloor p + U_i \rfloor.$$

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

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

Poisson-verdeling: Als m groot is, kunnen we met behulp van de simulatie van de binomiale verdeling ook de Poisson-verdeling met parameter $\lambda = m \cdot p$ simuleren.

Maar hiervoor maken we beter gebruik van het idee van een *Poisson-proces* die gewoon de tijdstippen van gebeurtenissen beschrijft die volgens een Poisson-verdeling optreden.

Het cruciale punt is dat de tussentijden tussen twee gebeurtenissen van een Poisson-proces exponentieel verdeeld zijn en de parameter λ van deze exponentiële verdeling noemen we de *intensiteit* van het Poisson-proces. In het bijzonder geldt dat voor een Poisson-proces met intensiteit λ het aantal waarnemingen in het tijdsinterval $[0, t]$ een Poisson-verdeling met parameter λt heeft.

Om een Poisson-verdeling met parameter λ te simuleren, moeten we dus het tijdsinterval $[0, 1]$ volledig overdekken met tussentijden die een exponentiële verdeling met parameter λ hebben. Het aantal benodigde tussentijden is dan één hoger dan het aantal gebeurtenissen die in het interval $[0, 1]$ vallen en dit aantal is een Poisson-verdeelde stochast met parameter λ.

We nemen onafhankelijke stochasten $Y_1, Y_2, \ldots$ die exponentieel verdeeld zijn met parameter λ. Als we nu een stochast X definiëren door de eigenschap

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

dan heeft X een Poisson-verdeling met parameter λ.

Maar we hebben boven gezien dat we een exponentieel verdeelde stochast Y_i met parameter λ kunnen simuleren door

$$Y_i := -\frac{1}{\lambda} \log(U_i)$$

waarbij U_i een randomgenerator is die een uniforme verdeling op het interval $[0, 1]$ voortbrengt.

Voor de stochast X moet dus gelden dat

$$-\sum_{i=1}^{X} \frac{1}{\lambda} \log(U_i) \leq 1 < -\sum_{i=1}^{X+1} \frac{1}{\lambda} \log(U_i) \Leftrightarrow -\sum_{i=1}^{X} \log(U_i) \leq \lambda < -\sum_{i=1}^{X+1} \log(U_i).$$

Door de laatste keten van ongelijkheden met -1 te vermenigvuldigen en vervolgens de e-machten van alle termen te nemen, krijgen we

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

Dit betekent dat we uniform verdeelde toevalsgetallen U_i tussen 0 en 1 moeten vermenigvuldigen tot dat het product kleiner is dan $e^{-\lambda}$. Het aantal X van benodigde getallen waarvoor we het laatst boven $e^{-\lambda}$ hebben gezeten, is dan een Poisson-verdeelde stochast met parameter λ.

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

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

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

onder zwakke verdere voorwaarden aan de X_i een stochast met standaard-normale verdeling. In het bijzonder wordt aan de voorwaarden voldaan als alle X_i dezelfde standaardafwijking σ hebben, in dit geval convergeert

$$\frac{1}{\sqrt{n}\cdot\sigma}\,(\sum_{i=1}^{n} X_i - E(X_i))$$

tegen de standaard-normale verdeling.

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

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

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

Simulatie van het Monty-Hall probleem

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

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

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

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

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

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

(4) Hier zijn er twee versies:

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

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

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

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

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

Belangrijke begrippen in deze les

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

Opgaven

75. Zij X een stochast, waarvan bekend is dat die een uniforme verdeling op een interval $[a, b]$ heeft, maar waarvoor de grenzen van het interval niet bekend zijn.

    Stel je hebt waarnemingen $x_1, x_2, \ldots, x_n$ voor de stochast X. Bepaal de maximum-likelihood schatting voor de grenzen a en b van het interval.

    (Hint: Je hebt hiervoor geen differentiaalrekening nodig.)

76. Voor een gebeurtenis dat volgens een normale verdeling met dichtheidsfunctie

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

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

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

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

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

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

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

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

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

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

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

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

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