Les 2 Kansverdelingen

We hebben in het begin gesteld dat we de kans voor een zekere gunstige uitkomst berekenen als het aantal gunstige uitkomsten gedeelt door het 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 altijd redelijk voor de hand liggend, we eisen alleen maar eigenschappen die heel natuurlijk zijn:

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

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

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

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

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

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

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

2.1 Discrete kansverdelingen

We hebben tot nu toe alleen maar naar voorbeelden gekeken, waarbij de verzameling Ω van mogelijke uitkomsten eindig is. In deze situatie spreken we van

discrete kansverdelingen, in tegenstelling tot *continue* kansverdelingen die we in de volgende paragraaf gaan behandelen.

Een belangrijk voorbeeld van een discrete kansverdeling hebben we al gezien, namelijk de *gelijkverdeling* die vaak ook *Laplace-verdeling* heet:

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

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

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

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

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)

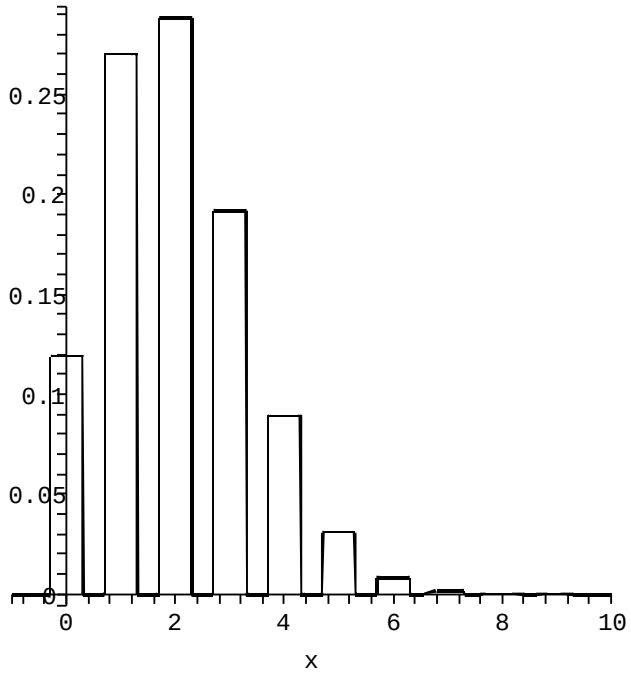
Voorbeeld 2: Bij een qualiteitstoets kiezen we uit een levering van n stukken een steekproef van m stukken die we testen en niet terugleggen. Dit is bijvoorbeeld het geval als de test het object beschadigt, zo als bij het testen van lucifers. We nemen aan dat de levering s slechte stukken bevat en willen 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 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 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.

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*.

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

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

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

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

De Poisson-verdeling

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

Voor heel zeldzame gebeurtenissen zullen we verwachten dat er veel pogingen nodig zijn tot dat 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 $m \cdot p = \lambda$ constant blijft. De constante λ geeft aan hoeveel gunstige uitkomsten we bij m pogingen eigenlijk verwachten.

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

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

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 $m \to \infty$.

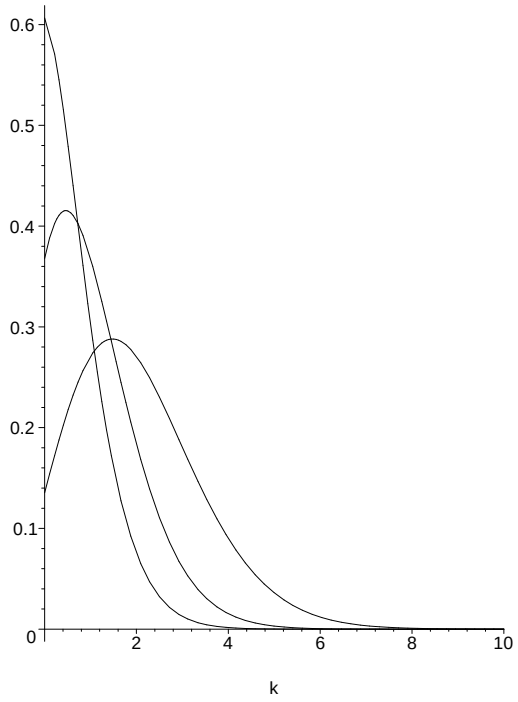
Voor zeldzame gebeurtenissen gaat de binomiale verdeling dus in de limiet tegen de *Poisson-verdeling*

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

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

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

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



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

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

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 verwachten. 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.

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

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

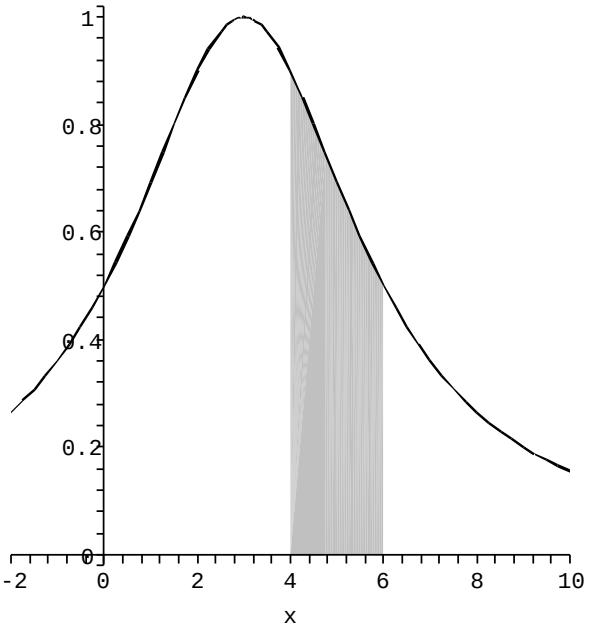
2.2 Continue kansverdelingen

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

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

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

Maar we kunnen ook kansverdelingen met continue uitkomsten beschrijven. Het idee hiervoor is als volgt: We beschrijven de kans dat de uitkomst x van een experiment in het interval $[a, b]$ valt als oppervlakte onder de grafiek van een geschikte functie $f(x)$ op het interval $[a, b]$.



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

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

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

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

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

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

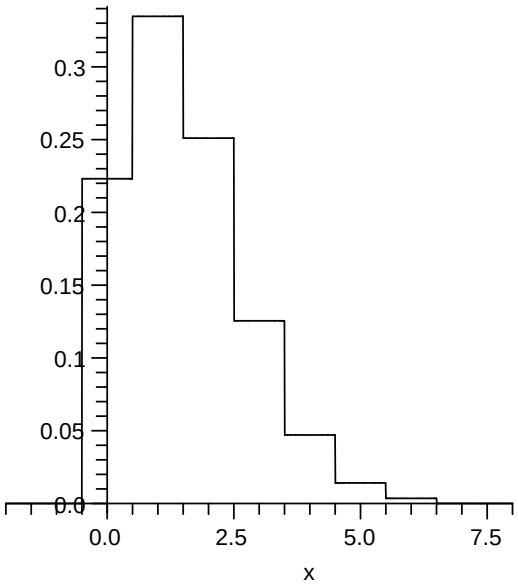
De eerste eis zorgt ervoor dat we steeds niet-negatieve kansen krijgen en de tweede eis zegt dat de totale oppervlakte onder de grafiek 1 is en geeft dus de normering van de kansverdeling weer. We noemen een functie $f(x) : \mathbb{R} \to \mathbb{R}$ die aan deze eisen voldoet een *dichtheidsfunctie*.

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 6 uit. Merk op dat zo'n functie op een *histogram* lijkt, waarmee (relatieve) frequenties van gebeurtenissen in een grafiek weergegeven kunnen worden.



Figuur 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 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.

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

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

De verdelingsfunctie heeft de eigenschappen:

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

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

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

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

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

De uniforme verdeling

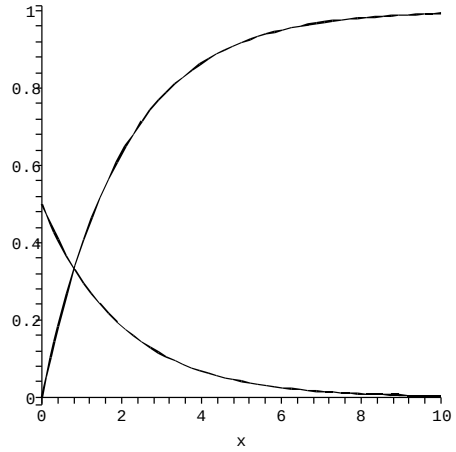
Deze verdeling staat ook bekend als homogene verdeling of rechthoekverdeling en is het continue analoog van de discrete gelijkverdeling. Op een bepaald interval $[a, b]$ (of een vereniging van intervallen) heeft elke punt dezelfde kans en buiten het interval is de kans 0. De normering $\int_{-\infty}^{\infty}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

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



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

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

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 8 afgebeeld is, heeft de vorm van een klok en is gegeven door

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

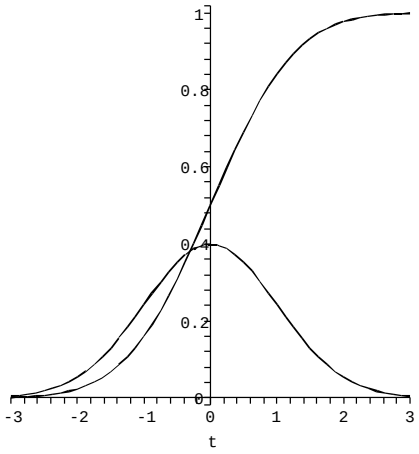
In dit geval kunnen we de verdelingsfunctie $F(x)$ alleen maar door de integraal van $f(x)$ beschrijven, omdat er geen gewone functie $F(x)$ is die $f(x)$ als afgeleide heeft.

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

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

Belangrijke begrippen in deze les

- kansverdeling
- gelijkverdeling (Laplace-verdeling)



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

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

Opgaven

11. 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.

12. 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.

13. 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?

14. 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?

15. 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?

16. 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.)

17. 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 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?

18. 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.

19. 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.