Inhoudsopgave Dictaat Kansrekening

Formuleblad Kansrekening voor TBK t.b.v. toetsen in module 1

Verdeling	$\boldsymbol{E(X)}$	$\boldsymbol{var(X)}$
Geometrisch	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hypergeometrisch	$n \cdot \frac{R}{N}$	$n \cdot \frac{R}{N} \cdot \frac{N-R}{N} \cdot \frac{N-n}{N-1}$
Poisson $P(X=x) = \frac{\mu^x e^{-\mu}}{x!}, x = 0, 1, 2, \ldots$	μ	μ
Exponentieel	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Uniform op (a, b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

===

Formuleblad Kansrekening voor INF en BIT t.b.v. toetsen in module 4

Verdeling	$\boldsymbol{E(X)}$	$\boldsymbol{var(X)}$
Geometrisch	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hypergeometrisch	$n \cdot \frac{R}{N}$	$n \cdot \frac{R}{N} \cdot \frac{N-R}{N} \cdot \frac{N-n}{N-1}$
Poisson $P(X=x) = \frac{e^{-\mu} \mu^x}{x!}, x = 0, 1, 2, \ldots$	μ	μ
Uniform op (a, b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponentieel	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Erlang $f_X(x) = \frac{\lambda(\lambda x)^{n-1} e^{-\lambda x}}{(n-1)!}, x \geq 0$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$

$$var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i) + \sum_{i \neq j}\sum cov(X_i, X_j)$$

Formuleblad bij het deel Kansrekening voor TBK in module 3

$E(X) = \sum_x xP(X = x)$ en $var(X) = E(X - \mu)^2 = E(X^2) - (EX)^2$

$E(aX + b) = aE(X) + b$ en $var(aX + b) = a^2 var(X)$

$$var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i) + \sum_{i \neq} \sum_{j} cov(X_i, X_j)$$

Verdeling	Kansfunctie / kansdichtheid	$\boldsymbol{\mu = E(X)}$	$\boldsymbol{\sigma^2 = var(X)}$
Binomiaal $B(n, p)$	$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, 2, \dots, n$	np	$np(1-p)$
Geometrisch (p)	$P(X = x) = (1-p)^{x-1} p$, $x = 1, 2, 3, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson (μ)	$P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$, $x = 0, 1, 2, \dots$	μ	μ
Uniform $U(a, b)$	$f(x) = \frac{1}{b-a}$, $a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponentieel $Exp(\lambda)$	$f(x) = \lambda e^{-\lambda x}$, $x \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Hoofdstuk 1 Experiment, uitkomstenruimte en kans

1.1 Experiment en uitkomstenruimte

In de kansrekening (of waarschijnlijkheidsrekening) houden we ons bezig met (wiskundige) modellen voor het beschrijven van **experimenten** waarin het **toeval** een rol speelt. Men kan daarbij denken aan het werpen met een dobbelsteen of het meten van de levensduur van een bepaald type gloeilamp. In deze gevallen is duidelijk welke handelingen men moet verrichten om een uitkomst van het experiment te verkrijgen. De uitkomst zelf staat echter in het algemeen niet van te voren vast, maar pas nadat het experiment is uitgevoerd. Dergelijke experimenten noemen we stochastisch.
Onder een **toevalsexperiment of stochastisch experiment** verstaan we een experiment dat onder gelijke omstandigheden herhaald niet noodzakelijk tot gelijke uitkomsten leidt.

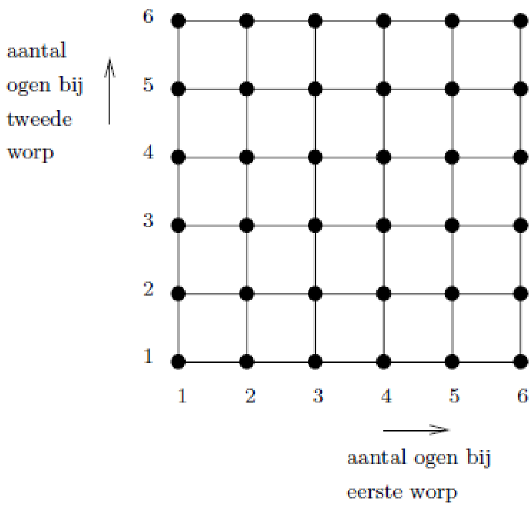
In het vervolg zullen we steeds van **experiment** spreken als we een stochastisch experiment bedoelen.
Het uitvoeren van een experiment leidt altijd tot een resultaat: de **uitkomst** van het experiment. Bij het werpen met een dobbelsteen is de uitkomst het aantal geworpen ogen; de uitkomst bij het meten van de levensduur van een gloeilamp is de vastgestelde levensduur, een positief reëel getal, en het lanceren van de satelliet leidt tot één van de uitkomsten "succes" of "mislukking".
Hoewel de uitkomst van een stochastisch experiment niet a priori vaststaat, kunnen we wel vaststellen welke uitkomsten mogelijk zijn. Deze **mogelijke** uitkomsten horen onverbrekelijk bij het experiment en staan, in tegenstelling tot de uitkomst, ook vast voordat het experiment is uitgevoerd. De verzameling van alle mogelijke uitkomsten noemen we de uitkomstenruimte van het experiment.
Deze verzameling duiden we gewoonlijk aan door S (van het Engelse *Sample space*).

Definitie 1.1.1 De **uitkomstenruimte** S van een experiment is de verzameling van alle mogelijke uitkomsten.

Voorbeeld 1.1.2 We werpen een munt. Er zijn twee mogelijke uitkomsten K (kruis) en M(munt). We nemen dus als uitkomstenruimte de verzameling $S = \{K, M\}$. ■

Voorbeeld 1.1.3 Uit een partij bestaande uit 1000 chips worden er 10 gekozen. Deze 10 chips worden elk getest en goedgekeurd of afgekeurd. De uitkomst van dit experiment is het aantal goedgekeurde chips en de uitkomstenruimte is $S = \{0, 1, \ldots, 10\}$. ■

Voorbeeld 1.1.4
We werpen twee maal met een dobbelsteen. Grafisch kan men deze uitkomstenruimte als hiernaast voorstellen.
Een uitkomst van dit experiment is een getallenpaar, waarbij het eerste getal het aantal geworpen ogen is bij de eerste worp en het tweede getal dat bij de tweede worp.
De uitkomstenruimte is
$S = \{(1,1), .., (1,6), (2,1), .. , (2,6), \ldots, (6,6)\}$
$\quad = \{(i,j) |\ i = 1, 2, \ldots, 6 \text{ en } j = 1, 2, \ldots, 6\}$
S bevat dus 6 × 6 = 36 uitkomsten. ■



Voorbeeld 1.1.5 We werpen een munt zo vaak tot voor het eerst kruis boven komt. Het aantal worpen vatten we op als uitkomst van het experiment en de uitkomstenruimte is $S = \{1, 2, \ldots\}$. Deze uitkomstenruimte is niet eindig, zoals in de voorafgaande voorbeelden, maar aftelbaar oneindig. ■

Voorbeeld 1.1.6 De levensduur van een gloeilamp kan men beschouwen als de uitkomst van een experiment. Aangezien een lamp op ieder willekeurig tijdstip defect kan raken, kunnen we als uitkomstenruimte nemen $S = [0, \infty)$. Deze uitkomstenruimte is niet meer aftelbaar. ■

Een experiment geeft aanleiding tot bepaalde gebeurtenissen. Zo kunnen we in voorbeeld 1.1.3 spreken over de gebeurtenis dat meer dan de helft van de geteste chips afgekeurd wordt. De gebeurtenis “meer dan de helft afgekeurd” treedt op als de uitkomst 0, 1, 2, 3 of 4 is. Daarom identificeren we deze gebeurtenis met de deelverzameling $A = \{0, 1, 2, 3, 4\}$ van de uitkomstenruimte S.

Definitie 1.1.7 Een **gebeurtenis** is een deelverzameling van de uitkomstenruimte S.

Net zoals bij deelverzamelingen zullen we gebeurtenissen veelal aanduiden met hoofdletters $A, B, C, \ldots$. Ook de lege verzameling $\emptyset$ en S zelf zijn gebeurtenissen.
De lege verzameling $\emptyset$ noemen we de **onmogelijke gebeurtenis**, immers geen enkele uitkomst ligt in $\emptyset$ en deze gebeurtenis treedt dus nooit op.
De verzameling S noemen we de **zekere gebeurtenis** omdat elke uitkomst in S ligt en deze gebeurtenis dus steeds optreedt. Een gebeurtenis die uit slechts één uitkomst s bestaat noemen we een **elementaire gebeurtenis**: $\{\boldsymbol{s}\}$.

Voorbeeld 1.1.8 $S = \{(i,j) | i = 1, 2, \ldots, 6 \text{ en } j = 1, 2, \ldots, 6\}$ is de uitkomstenruimte die behoort bij het experiment waarbij we twee maal achtereen een dobbelsteen werpen.
Als A de gebeurtenis is dat beide worpen hetzelfde aantal ogen opleveren dan geldt $A = \{(1,1), (2,2), \ldots, (6,6)\} = \{(i,i) | i = 1, 2, \ldots, 6\}$.
De gebeurtenis B dat het totaal aantal geworpen ogen juist gelijk is aan 5 wordt gegeven door $B = \{(1, 4), (2, 3), (3, 2), (4, 1)\} = \{(i, 5-i) | i = 1, 2, 3, 4\}$. ■

Aangezien gebeurtenissen verzamelingen zijn, kunnen we aan verschillende begrippen uit de verzamelingenleer een interpretatie in de kansrekening geven. We veronderstellen deze begrippen uit de verzamelingenleer bekend.

Als A en B gebeurtenissen zijn, dan geldt:

- $\overline{A}$ (= $\boldsymbol{S - A}$) is de **complementaire gebeurtenis** of het **complement** van A, dus de gebeurtenis die optreedt als A niet optreedt. Spreek uit: “niet A”.
- $\boldsymbol{A \cup B}$**, de vereniging van** $\boldsymbol{A}$ **en** $\boldsymbol{B}$, dus de gebeurtenis die optreedt als **minstens één van de gebeurtenissen** $\boldsymbol{A}$ **en** $\boldsymbol{B}$ optreedt. Spreek uit: “A of B (of beiden)”.
- $\boldsymbol{A \cap B}$**, of kortweg** $\boldsymbol{AB}$**, de doorsnede van** $\boldsymbol{A}$ **en** $\boldsymbol{B}$, de gebeurtenis die optreedt indien **zowel** $\boldsymbol{A}$ **als** $\boldsymbol{B}$ optreedt. Spreek uit: “A èn B”.

Opmerking: De vereniging ∪ en de doorsnede ∩ hebben voor verzamelingen dezelfde functie als de wiskundige of- en èn-tekens voor de elementen van verzamelingen, ∨ en ∧:
$A \cup B = \{s \in S | s \in A \ \vee \ s \in B\}$ en $A \cap B = \{s \in S | s \in A \ \wedge \ s \in B\}$
Indien er onduidelijkheid is over de volgorde waarin de operaties ∩ en ∪ moeten worden uitgevoerd, gaat ∩ vóór ∪, dus $A \cup BC$ betekent $A \cup (B \cap C)$. ■

Voorts betekent

- $A \subset B$: A is een **deelverzameling** van B. A impliceert B, dus als A optreedt, dan treedt B ook op. (In dit dictaat gebruiken we het teken ⊆ niet: $A \subset B$ kan $A = B$ betekenen.)

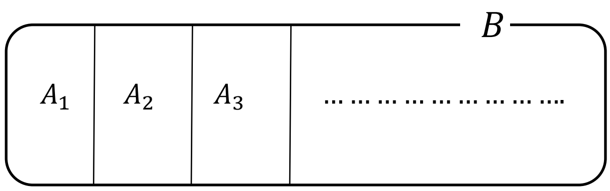
Definitie 1.1.9 A en B zijn **elkaar uitsluitende** (of **disjuncte**) gebeurtenissen als $\boldsymbol{AB} = \emptyset$, dus als A en B kunnen niet gelijktijdig optreden.

We kunnen deze definitie uitbreiden naar een rij gebeurtenissen. Deze bestaat uit een aftelbaar aantal gebeurtenissen A_i, d.w.z. de rij is eindig $(A_1, A_2, \ldots, A_n)$ of de rij is aftelbaar oneindig $(A_1, A_2, \ldots)$. In beide gevallen duiden we de rij aan met $\{A_i\}$.

Definitie 1.1.10 De gebeurtenissen $A_1, A_2, \ldots, A_n$ of $A_1, A_2, \ldots$ noemen we **elkaar uitsluitend** (of disjunct) als $A_i A_j = \emptyset$ voor elk mogelijk paar (i, j) met $i \neq j$.

Zij $\{A_i\}$ een rij gebeurtenissen, dan bedoelen we met $\bigcap_{i=1}^{n} A_i$ in het eindige geval en $\bigcap_i A_i$ in het aftelbaar oneindige geval:
$\bigcap_i A_i$ is de gebeurtenis die optreedt als elk van de gebeurtenissen A_i optreedt en
$\bigcup_i A_i$ is de gebeurtenis die optreedt als tenminste één van de gebeurtenissen A_i optreedt.

Definitie 1.1.11 De rij gebeurtenissen $\{A_i\}$ is een **partitie** van de gebeurtenis B als de gebeurtenissen A_i elkaar uitsluiten en $B = \bigcup_i A_i$



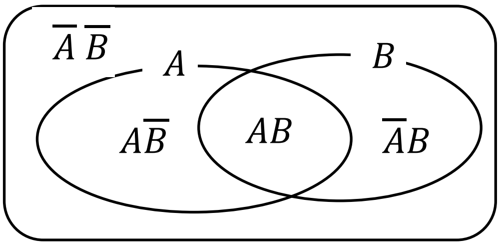
Voorbeeld 1.1.12 Ten behoeve van een communicatiesysteem kunnen we de 26 letters a t/m z en 6 leestekens (. , : ; ? !) coderen m.b.v. de 32 “codewoorden” 00000, 10000, . . . , 11111. Kiezen we het codewoord dat op een willekeurig moment door het communicatiesysteem wordt verzonden, dan is dat een toevalsexperiment met uitkomstemruimte

$$S = \{e_1 e_2 e_3 e_4 e_5 | e_i = 1 \text{ of } e_i = 0, \text{ voor } i = 1, 2, \ldots, 5\}.$$

We kunnen nu A_i voor i = 1, 2, 3, 4, 5 definiëren als de gebeurtenis, dat het codewoord op de i-de plaats een 1 heeft staan en A_0 als de gebeurtenis dat het codewoord geen enen bevat, dus

$A_0 = \{00000\}$. A_1 treedt dus op als het betreffende codewoord begint met 1 en $\overline{A}_1$ bestaat uit alle codewoorden beginnend met 0. A_1 en $\overline{A}_1$vormen dus een partitie van S.
Het is duidelijk dat $S = \bigcup_{i=0}^{5} A_i$, maar A_0 t/m A_5 vormen echter geen partitie van S, daar zij elkaar niet uitsluiten.
Het codewoord 11000 komt bijvoorbeeld in zowel A_1 als A_2 voor ($A_1 A_2 \neq \emptyset$).
Definiëren we echter B_i als de gebeurtenis dat het willekeurig gekozen codewoord (precies) i enen bevat, dan vormen $B_0, B_1, \dots, B_5$ een partitie van S. ■

Eigenschap 1.1.13 (Eigenschappen van gebeurtenissen)



a. $A \cap (B \cup C) = (A \cap B) \cup A \cap C)$ en
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

b. $A \cup B = A \cup (\overline{A}B)$ en
$B = (AB) \cup (\overline{A}B)$.

c. (**regels van De Morgan**)
$\overline{A \cup B} = \overline{A} \cap \overline{B}$ en $\overline{A \cap B} = \overline{A} \cup \overline{B}$ $\left(\text{algemeen:} \overline{\bigcup_i A_i} = \bigcap_i \overline{A_i} \text{ en } \overline{\bigcap_i A_i} = \bigcup_i \overline{A_i}\right)$

De juistheid van de eigenschappen 1.1.13a en b ziet men gemakkelijk door gebruik te maken van een Venndiagram. Zo kan men het eigenschap 1.1.13b verifiëren met het Venndiagram:
Het eerste deel van eigenschap 1.1.13c kan men als volgt beredeneren: $A \cup B$ is de gebeurtenis dat er van A en B tenminste één optreedt, dus als $\overline{A \cup B}$ optreedt, treedt er A niet op 'B ook niet. Dus $\overline{A} \cap \overline{B}$. Op soortgelijke wijze kan men de tweede gelijkheid beredeneren.

1.2 Symmetrische kansruimte

In de vorige paragraaf hebben we gesproken over een experiment en een bijbehorende uitkomstenruimte S en we hebben gezien dat gebeurtenissen deelverzamelingen zijn van S.
We willen nu een uitspraak doen over de kans dat een gebeurtenis optreedt, of kortweg de kans op een gebeurtenis.
We willen een functie P definiëren die aan iedere gebeurtenis A uit S een reëel getal $P(A)$ toekent dat we de kans op de gebeurtenis A zullen noemen. Het paar (S, P) zullen we een kansruimte noemen. Voordat we een algemene wiskundige definitie geven van het begrip kans, beschrijven we in deze paragraaf een aantal eenvoudige situaties.

Voorbeeld 1.2.1 We werpen met een dobbelsteen. Het geworpen aantal ogen is de uitkomst van het experiment, dus de uitkomstenruimte is $S = \{1, 2, 3, 4, 5, 6\}$.
Wordt ons gevraagd naar de kans dat het aantal ogen even is, dan zijn we geneigd om te antwoorden dat deze kans $\frac{3}{6}$ is: drie van de zes mogelijke uitkomsten zijn immers even.
Impliciet veronderstellen we dan dat de dobbelsteen **zuiver** is, d.w.z. dat elke uitkomst even waarschijnlijk is en met kans zal $\frac{1}{6}$ optreden. De gebeurtenis $\{2, 4, 6\}$ is één van de $2^6 = 64$ mogelijke gebeurtenissen (= deelverzamelingen van S). De kans op elk van die gebeurtenissen berekenen we dus door het aantal uitkomsten in die gebeurtenissen te delen door 6. ■

Beschouw, algemener, een experiment met een eindige uitkomstenruimte S. Het totaal aantal uitkomsten noemen we $N(S)$ en het aantal uitkomsten van een gebeurtenis is A duiden we aan met $N(A)$. We veronderstellen dat de verschillende uitkomsten "gelijkelijk mogelijk" of "even waarschijnlijk" zijn. We eisen nu van onze kansdefinitie dat in dit geval de kans op een bepaalde uitkomst, of elementaire gebeurtenis, gelijk is aan $\frac{1}{N(S)}$. Aan deze eis wordt voldaan door de **kansdefinitie van Laplace** (1749 - 1827), die als volgt luidt:

Definitie 1.2.2 Als de uitkomstenruimte S van een experiment $N(S)$ gelijkelijk mogelijke uitkomsten bevat en de gebeurtenis A uit $N(A)$ uitkomsten bestaat, dan is de **kans op een gebeurtenis** A, aangeduid door $P(A)$, gelijk aan

$$\boldsymbol{P(A) = \frac{N(A)}{N(S)}}$$

Indien de gebeurtenis A een elementaire gebeurtenis is, dus slechts één uitkomst bevat, dan is volgens deze definitie de kans op A gelijk aan $P(A) = \frac{N(A)}{N(S)} = \frac{1}{N(S)}$, overeenkomstig onze eis.

Definitie 1.2.3 Indien S een eindige uitkomstenruimte van een experiment is en P gegeven wordt door de kansdefinitie van Laplace, dan heet het paar (S, P) een **symmetrische kansruimte**.

We bespreken nu nog enkele voorbeelden van symmetrische kansruimten.

Voorbeeld 1.2.4 We werpen een zuivere dobbelsteen twee maal.
De uitkomstenruimte is $S = \{(i, j) | i = 1, 2, \ldots, 6 \text{ en } j = 1, 2, \ldots, 6\}$.
We veronderstellen dat iedere uitkomst (i, j) even waarschijnlijk is. De kans op een bepaalde uitkomst, bijv. (4, 2), is dan $\frac{1}{36}$.
Wat is de kans dat het totaal aantal geworpen ogen gelijk is aan 8? De gebeurtenis "het totaal aantal geworpen ogen is 8" is de deelverzameling $A = \{(2,6), (3,5), (4,4), (5,3), (6,2)\}$ en deze bevat 5 uitkomsten. De gevraagde kans is dus $P(A) = \frac{N(A)}{N(S)} = \frac{5}{36}$. ∎

Voorbeeld 1.2.5 Uit een spel van 52 kaarten trekken we lukraak één kaart. Met de definitie van Laplace vinden we:

$P(\text{ruiten}) = \frac{13}{52} = \frac{1}{4}$, $P(\text{aas}) = \frac{4}{52} = \frac{1}{13}$ en $P(\text{ruitenaas}) = \frac{1}{52}$. ∎

In experimenten, zoals de voorgaande, waarin sprake is van het **willekeurig**, **lukraak** of **aselect** kiezen van een element uit een eindige uitkomstenruimte, impliceert dit woordgebruik dat het hanteren van de kansdefinitie van Laplace gerechtvaardigd is. In het volgende voorbeeld laten we zien dat dit niet vanzelfsprekend is als we een aselecte keuze maken uit een aantal dingen, die niet alle onderscheidbaar zijn.

Voorbeeld 1.2.6 We nemen een doos met N knikkers, waarvan er R rood zijn en $B = N - R$ blauw. We kiezen lukraak één knikker uit de doos; er zijn twee mogelijke uitkomsten: r (de knikker is rood) en b (de knikker is blauw), dus de uitkomstenruimte is $S = \{r, b\}$.
Merk op dat het lukraak kiezen betrekking heeft op de knikkers en niet op de kleuren (de elementen van S). Daarom is het gebruik van de kansdefinitie van Laplace, die zou inhouden dat $P(r) = P(b) = \frac{1}{2}$ in het algemeen niet gerechtvaardigd. Nummeren we echter (in

gedachten) de rode knikkers van 1 t/m R en de blauwe van $(R+1)$ t/m $(R+B)=N$, dan zijn nu de nummers 1, 2, ...,N de mogelijke uitkomsten, dus $S` = \{1,2,\ldots,N\}$.
Lukraak kiezen van een knikker betekent dat elke knikker gelijke kans heeft om getrokken te worden, zodat we nu te maken hebben met een symmetrische kansruimte.
Als $ROOD$ de gebeurtenis is dat de getrokken knikker rood is en $BLAUW$ de gebeurtenis dat hij blauw is, d.w.z. $ROOD = \{1,2,\ldots,R\}$ en $BLAUW = \{R+1,\ldots,N\}$, dan geldt

$$P(ROOD) = \frac{N(ROOD)}{N(S`)} = \frac{R}{N} \text{ en } P(BLAUW) = \frac{N(BLAUW)}{N(S`)} = \frac{N-R}{N} \qquad \blacksquare$$

In voorbeeld 1.2.6 zagen we dat alleen in het bijzondere geval dat $R = B = \frac{1}{2}N$ de oorspronkelijke kansruimte $S = \{r, b\}$ symmetrisch is. In alle overige gevallen is een verfijning van de uitkomstenruimte noodzakelijk om toepassing van de kansdefinitie van Laplace mogelijk te maken. De oorspronkelijke uitkomst r is door de nummering van de knikkers verfijnd tot de uitkomsten 1, 2, . . . , R en de uitkomst b is de gebeurtenis bestaande uit de nummers $(R+1)$ t/m N geworden.

Dit voorbeeld leert dat men niet lichtvaardig moet concluderen dat de kansruimte, die bij een experiment hoort, symmetrisch is. Zo zullen alle codewoorden die het communicatiesysteem in voorbeeld 1.1.12 verzendt i.h.a. niet even vaak voorkomen, daar in teksten meer *e*'s dan *x*-en staan. Men dient dus steeds zorgvuldig na te gaan of de verschillende uitkomsten wel alle "gelijkelijk mogelijk" of "even waarschijnlijk" zijn. Als dit inderdaad het geval is, dan bepaalt men de kans op een gebeurtenis A door te tellen hoeveel uitkomsten A bevat. Omdat dit tellen moeilijker kan zijn dan op het eerste gezicht lijkt, komen we hier in hoofdstuk 2 uitgebreid op terug.

Hoewel we de kansdefinitie van Laplace geïntroduceerd hebben na het formuleren van één eis, namelijk dat voor elke gebeurtenis A moet gelden $P(A) = \frac{N(A)}{N(S)}$, is gemakkelijk na te gaan dat deze definitie aan een aantal eisen voldoet die we intuïtief aan een kans zouden stellen.
Zo geldt bijvoorbeeld:

Eigenschap 1.2.7 (Eigenschappen voor een symmetrische kansruimte)

- **a.** $P(A) \geq 0$ voor elke gebeurtenis A,
- **b.** $P(S) = 1$,
- **c.** als $A \subset B$ dan is $P(A) \leq P(B)$,
- **d.** $P(\overline{A}) = 1 - P(A)$,
- **e.** als $A_1, A_2, \ldots, A_n$ elkaar uitsluitende gebeurtenissen zijn, dan is

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

Bewijs:
Deze eigenschappen volgen meteen uit de kansdefinitie van Laplace, bijv. e. volgt daaruit en uit het feit dat voor elkaar uitsluitende gebeurtenissen $A_1, A_2, \ldots, A_n$ geldt dat:

$$N\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} N(A_i)$$

Hieruit volgt: $P\left(\bigcup_{i=1}^{n} A_i\right) = \frac{N(\bigcup_{i=1}^{n} A_i)}{N(S)} = \frac{\sum_{i=1}^{n} N(A_i)}{N(S)} = \sum_{i=1}^{n} P(A_i)$ ■

De kansdefinitie van Laplace heeft twee ernstige beperkingen. Ten eerste veronderstelt de definitie een eindige uitkomstenruimte, terwijl we ook experimenten in onze beschouwingen willen betrekken waarbij de uitkomstenruimte oneindig veel elementen bevat (voorbeelden 1.1.5 en 1.1.6). Ten tweede: al telt de uitkomstenruimte van een experiment eindig veel elementen, de veronderstelling dat iedere elementaire gebeurtenis dezelfde kans heeft, sluit zeker niet altijd aan bij de realiteit.
We zoeken daarom naar een definitie van het begrip kans, die meer algemeen is.

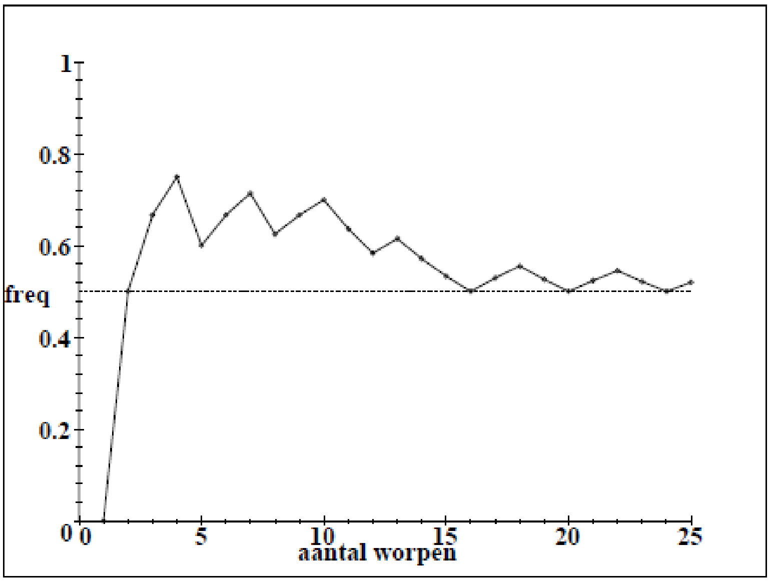
1.3 Frequentiequotiënt en de experimentele wet van de grote aantallen

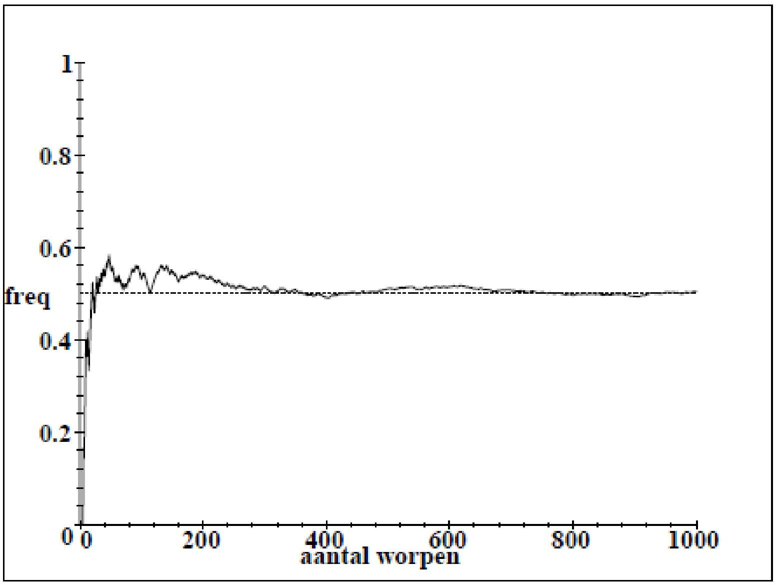
Voorbeeld 1.3.1 We willen weten of een munt "zuiver" is, d.w.z. we willen nagaan of de kans op kruis inderdaad $\frac{1}{2}$ is. Een methode om de kans op kruis "experimenteel te bepalen" is, het zeer vaak werpen van de munt, waarbij zowel het aantal keren kruis als het aantal worpen wordt bijgehouden. Het aantal keren kruis gedeeld door het aantal worpen is dan een schatting van de kans op kruis, bijvoorbeeld $\frac{21}{38}$ bij 21 keren kruis in 38 worpen.
Naarmate we vaker de munt werpen, zal de schatting naar verwachting nauwkeuriger zijn. ■

We maken bij dit soort experimenten gebruik van de frequentie van een gebeurtenis bij herhalingen van een experiment.

Definitie 1.3.2 We veronderstellen dat we een experiment met uitkomstenruimte S willekeurig vaak kunnen herhalen. Als de gebeurtenis A bij n herhalingen $n(A)$ keer is opgetreden, noemen we $f_n(A) = \frac{n(A)}{n}$ het **frequentiequotiënt** (de relatieve frequentie) van A bij deze n herhalingen.

Experimenteel blijkt dat $f_n(A)$ voor toenemende n "convergeert" naar een constante.
Dit verschijnsel noemen we de experimentele wet van de grote aantallen.
Hier is echter geen sprake van convergentie in de gebruikelijke (wiskundige) zin, omdat de uitkomsten bij achtereenvolgende herhalingen niet met zekerheid te voorspellen zijn.
Uitschieters blijven mogelijk, maar worden bij toenemende n steeds "onwaarschijnlijker".
We lichten dit toe aan de hand van de grafiek van het frequentiequotiënt als functie van n voor het werpen van munt uit voorbeeld.





Intuïtief zijn we geneigd om " $\lim_{n\to\infty}$ " $f_n(A)$ de kans op de gebeurtenis A te noemen.
Het probleem is echter dat het oneindig vaak herhalen van het experiment praktisch niet uitvoerbaar is. Bovendien is deze limiet niet wiskundig gedefinieerd (vandaar de aanhalingstekens om de limiet), zodat het frequentiequotiënt niet bruikbaar is voor een wiskundige definitie van het kansbegrip.
Omdat het echter aansluit bij ons intuïtieve kansbegrip dienen de eigenschappen van frequentiequotiënten wel een leidraad te zijn bij het ontwikkelen van een wiskundige kanstheorie.

De **frequentie-interpretatie van kansen** is gebaseerd op een groot aantal herhalingen van het experiment. Deze herhalingen hoeven niet daadwerkelijk te worden uitgevoerd.
Als een arts zijn patiënt vertelt dat de slaagkans van een operatie 95% is, bedoelt hij niet dat de patiënt de operatie vele malen zal moeten ondergaan om in zo'n 95 van de 100 gevallen van een geslaagde operatie te kunnen spreken. Het gaat om een éénmalige gebeurtenis. Wel kunnen we ons er een gedachtenexperiment bij voorstellen: trek een willekeurige knikker uit een bak met 95 witte en 5 zwarte knikkers. Een zwarte knikker staat voor een mislukte

operatie, een witte voor een geslaagde. Helaas kan slechts na de operatie geconstateerd worden, wat voor knikker er getrokken is.

Een vijftal gemakkelijk te verifiëren eigenschappen van frequentiequotiënten verkrijgen we door in eigenschap 1.2.7 P door f_n te vervangen. Het is derhalve wenselijk dat voor onze toekomstige kansdefinitie ook alle eigenschappen in 1.2.7 gelden.
Onze definitie zal gegeven worden in de vorm van een aantal axioma's waaraan een kans moet voldoen, waarbij we onder kans een functie P verstaan die aan iedere gebeurtenis A van een uitkomstenruimte een getal $P(A)$ toevoegt.
Natuurlijk zal het stelsel axioma's zodanig moeten zijn dat elk van de onderdelen van eigenschap 1.2.7 en elke andere, intuïtief gewenste eigenschap op zich een axioma is òf uit de axioma's volgt. Zo'n stelsel wordt gevormd door de axioma's van Kolmogorov.

1.4 De Axioma's van Kolmogorov

Definitie 1.4.1 Beschouw een experiment met een willekeurige niet-lege uitkomstenruimte S. Een functie P die aan elke gebeurtenis $A \subset S$ een reëel getal $P(A)$ toevoegt, noemen we een **kans** of **kansmaat** op S als

1. $\boldsymbol{P(A) \geq 0}$ voor elke gebeurtenis A,
2. $\boldsymbol{P(S) = 1}$ en
3. voor elke aftelbare rij elkaar uitsluitende gebeurtenissen $A_1, A_2, \ldots, A_n$ of $A_1, A_2, \ldots$ uitsluitende gebeurtenissen zijn, dan is
$$\boldsymbol{P\left(\bigcup_i A_i\right) = \sum_i P(A_i)}$$

De eisen 1., 2. en 3. staan bekend als de **axioma's van Kolmogorov**.
Met de formulering van dergelijke axioma's in zijn boekje "*Grundbegriffe der Wahrscheinlichkeitsrechnung*" (1933) legde A.N. Kolmogorov (1903-1987) de basis voor de moderne kansrekening.

Opmerking: Axioma 3 kunnen we in theorie beperken tot een aftelbaar oneindige rij elkaar uitsluitende gebeurtenissen $A_1, A_2, \ldots$. De eigenschap volgt dan ook voor een eindige rij $A_1, A_2, \ldots, A_n$.We kunnen axioma 3 echter niet beperken tot eindige rijen. Een verklaring hiervoor zou een (maat-theoretische) benadering vergen, die gezien het toegepaste karakter van dit vak te theoretisch en weinig verhelderend zou zijn. ∎

Definitie 1.4.2 Wanneer S een uitkomstenruimte is en P is een kans op S dan noemen we het paar (S, P) een **kansruimte**.

We hebben in eigenschap 1.2.7a en 1.2.7b gezien dat voor een symmetrische kansruimte de kansdefinitie van Laplace voldoet aan de eerste twee axioma's.
Eigenschap 1.2.7e betreft een eindig aantal elkaar uitsluitende gebeurtenissen A_i, terwijl volgens axioma 3. van Kolmogorov deze eigenschap ook voor een aftelbaar oneindig aantal gebeurtenissen dient te gelden. We merken op dat bij de kansdefinitie van Laplace sprake is

van een eindig aantal uitkomsten: er zijn dan slechts een eindig aantal verschillende gebeurtenissen mogelijk.
We leiden nu uit de axioma's van Kolmogorov af dat de eigenschappen 1.2.7c en 1.2.7d voor elke kansmaat gelden. Eerst laten we echter zien dat de kans op de onmogelijke gebeurtenis inderdaad 0 is.

Eigenschap 1.4.3 $\boldsymbol{P(\emptyset) = 0}$.

Bewijs: Volgens axioma 3. geldt voor elkaar uitsluitende gebeurtenissen A_1 en A_2:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

Nemen we $A_1 = S$ en $A_2 = \emptyset$, dan geldt:
$A_1 \cap A_2 = S \cap \emptyset = \emptyset$ (A_1 en A_2 sluiten elkaar dus uit) en $A_1 \cup A_2 = S \cup \emptyset = S$.
Dus $1 = P(S) = P(S \cup \emptyset) = P(S) + P(\emptyset) = 1 + P(\emptyset)$.
Ofwel $P(\emptyset) = 0$. ■

Eigenschap 1.4.4 (Complementregel) Voor elke gebeurtenis *A* geldt: $\boldsymbol{P(\overline{A}) = 1 - P(A)}$.

Bewijs: $A \cup \overline{A} = S$ en A en $\overline{A}$ sluiten elkaar uit, zodat volgens axioma's 2. en 3. geldt:
$1 = P(S) = P(A \cup \overline{A}) = P(A) + P(\overline{A})$ dus $P(\overline{A}) = 1 - P(A)$ ■

Eigenschap 1.4.5 Voor twee gebeurtenissen A en B met $\boldsymbol{A \subset B}$ geldt: $\boldsymbol{P(A) \leq P(B)}$.

Bewijs: zie opgave 7.

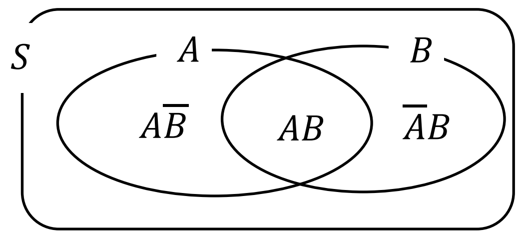
Eigenschap 1.4.6 Voor twee gebeurtenissen A en B (die elkaar niet noodzakelijk uitsluiten) geldt: $\boldsymbol{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$.

Bewijs: volgens eigenschap 1.1.13b geldt:
$A \cup B = A \cup \overline{A}B$, waarbij $A \cap \overline{A}B = \emptyset$ en
$B = AB \cup \overline{A}B$, waarbij $AB \cap \overline{A}B = \emptyset$.
Met axioma 3. volgt dan:

$$P(A \cup B) = P(A) + P(\overline{A}B)$$

en

$$P(B) = P(AB) + P(\overline{A}B) \quad -$$



Dus $P(A \cup B) - P(B) = P(A) - P(AB)$ ■

Als A en B elkaar uitsluiten, dus $P(A \cap B) = 0$, geldt uiteraard $P(A \cup B) = P(A) + P(B)$.

1.5 Vraagstukken

1. A, B en C zijn drie gebeurtenissen. Druk de volgende, met A, B en C samenhangende, gebeurtenissen uit in A, B en C:
- **a.** A en B, maar C niet.
- **b.** Alle drie.
- **c.** Tenminste één van de drie.
- **d.** Tenminste twee.
- **e.** Geen enkele.
- **f.** Precies één van de drie.
- **g.** Niet meer dan twee.

2. We hebben een verzameling van in totaal 1200 bouten en beschouwen de volgende deelverzamelingen.
$A =$ de verzameling van bouten die een lengte van 10 cm hebben;
$B =$ de verzameling van bouten die een gewicht van 1 ons hebben;
$C =$ de verzameling van bouten die een diameter van 20 mm hebben.
Verder is bekend dat
- 400 bouten een lengte van 10 cm en een gewicht van 1 ons hebben;
- 400 bouten een lengte van 10 cm en een diameter van 20 mm hebben;
- 400 bouten een gewicht van 1 ons en een diameter van 20 mm hebben;
- 300 bouten een gewicht van 1 ons en een diameter van 20 mm en een lengte van 10 cm hebben.

Bereken de kans dat een willekeurig gekozen bout uit tenminste twee van de drie verzamelingen A, B en C komt.

3. We werpen met twee munten tegelijk. Men kan dan drie uitkomsten onderscheiden: twee maal kruis òf twee maal munt òf één kruis en één munt. D'Alembert (1717-1783) stelde dat de bijbehorende kansruimte symmetrisch is. Ga experimenteel na, of u zijn bewering gelooft.

4. In zijn roman *Bomber* argumenteert Len Deighton dat een piloot in de Tweede Wereldoorlog 2% kans had om neergeschoten te worden bij iedere vlucht en concludeert: in 20 vluchten heeft hij dus 40% kans om neergeschoten te worden.
Argumenteer of dit juist is (of niet).

5. Gebruik een Venndiagram van de gebeurtenissen A, B en C om $P(A \cup B \cup C)$ uit te drukken in $P(A)$, $P(B)$, $P(C)$, $P(AB)$, $P(BC)$, $P(AC)$ en $P(ABC)$, analoog aan de somregel voor $P(A \cup B)$.

6. Het aselect (willekeurig) kiezen uit geboden mogelijkheden is niet altijd zo makkelijk als dat het lijkt. Zo zal iemand die een *multiple-choice*-tentamen (kies bijv. antwoord a, b, c of d) "op de gok" wil invullen, dit in het algemeen juist niet aselect doen: mogelijkheden b en c zullen door de één vaker gekozen worden dan door de ander; bij sommigen zullen 2, 3 of meer opeenvolgende a's consequent ondervertegenwoordigd zijn. Een manier om zo'n voorkeur te voorkomen, is het invullen te simuleren met een hoed met genummerde balletjes, waaruit men na goed schudden er blindelings één trekt. Een andere manier is gebruik te maken van (zuivere) dobbelstenen. Hoeveel dobbelstenen heeft men minimaal nodig om in één worp aselecte antwoorden te genereren op *multiple-choice*-vragen met **a.** 2 **b**. 4 of **c.** 5 mogelijke antwoorden?

7. Bewijs m.b.v. de axioma's van Kolmogorov dat uit $A \subset B$ volgt dat $P(A) \leq P(B)$ (eigenschap 1.4.5). Teken eerst een Venn-diagram.

8. Gegeven is dat $P(A) = \frac{1}{2}$, $P(AB) = \frac{1}{3}$ en $P(A \cup B) = \frac{8}{9}$
Bereken $P(B)$ en $P(\overline{A} \cap \overline{B})$.

Aanwijzingen/hints bij de opgaven:

1. Schets een Venndiagram met de 3 gebeurtenissen A, B en C, zó dat elk tweetal gebeurtenissen resp. alle drie tegelijk kunnen optreden: de uitkomstenruimte kan dan opgedeeld worden in vlakjes die alle doorsneden van A, B, C, $\overline{A}$, $\overline{B}$ en/of $\overline{C}$ zijn, bijv. $AB\overline{C}$.
2. Zie 1.
3. In plaats van 2 munten tegelijk opwerpen, kunnen we ook één munt twee keer opgooien: welke uitkomsten zijn er nu als we de resultaten van de eerste en de tweede worp onderscheiden?
4. Wat is de kans om bij een vlucht niet te worden neergeschoten? En de kans om op 20 vluchten niet te worden neergeschoten?
5. Zie 1.
6. –
7. Bekijk het Venndiagram, waarin A een deel van B is: hoe kun je het deel van B dat niet deel uitmaakt van A, als doorsnede noteren? Dus $B = A \cup$ .....
8. Gebruik weer het Venndiagram: in welke delen is B op te splitsen? En in welke $A \cup B$?
   Je mag het Venndiagram gebruiken om de kansregels op te schrijven, waarmee je met de gegeven kansen de gevraagde kansen kunt berekenen (een puzzel met "gebiedjes").

Hoofdstuk 2 Combinatorische Kansrekening

2.1 Theorie en voorbeelden

Veel problemen in de kansrekening kunnen worden opgelost door gebruik te maken van de kansdefinitie van Laplace, als het uitgangspunt van "even waarschijnlijke" uitkomsten gerechtvaardigd is. Volgens deze definitie bepaalt men dan de kans op gebeurtenis A door het aantal elementen van A te delen door het totaal aantal mogelijke uitkomsten:

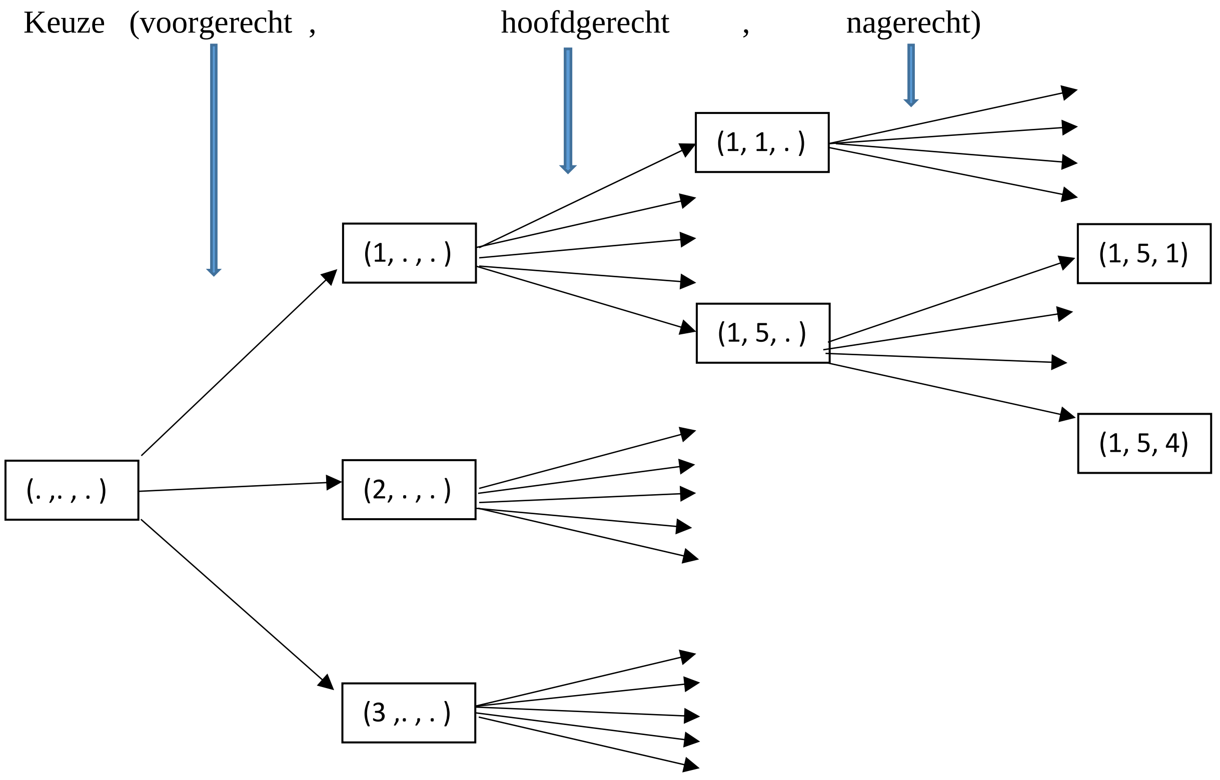
$$P(A) = \frac{N(A)}{N(S)}$$

Derhalve is de theorie van het tellen (combinatoriek) een belangrijk hulpmiddel in de kansrekening. In dit hoofdstuk zullen we een aantal belangrijke resultaten uit de combinatoriek behandelen aan de hand van toepassingen in de kansrekening.

Voorbeeld 2.1.1 Een menukaart omvat 3 voorgerechten, 5 hoofdgerechten en 4 nagerechten.
Van een echtpaar "prikt" de man een willekeurig (3 gangen) menu.
Hoe groot is de kans dat hij hetzelfde menu heeft gekozen als zijn vrouw?
Uit de omschrijving van het probleem leiden we af dat de verzameling van alle mogelijke menu's, die de man kiest, een symmetrische kansruimte vormt. Er zijn 3×5×4 menu's mogelijk. Dus de kans dat de man juist het menu van de vrouw prikt, is $\frac{1}{3\times5\times4} \approx$ 1.7%.
(*Het teken "≈" betekent "is afgerond gelijk aan".*)
Weinigen zullen moeite met bovenstaande berekening hebben.
Vragen we ons af **waarom** we voor het berekenen van het aantal menu's de aantallen voor-, hoofd- en nagerechten moeten vermenigvuldigen (3 × 5 × 4) en niet bijv. optellen (3 + 5 + 4), dan luidt de redenering als volgt:
"Bij elk gekozen voorgerecht kan men steeds 5 hoofdgerechten kiezen, hetgeen resulteert in 3 keer 5 verschillende combinaties van een voor- en een hoofdgerecht.
Bij elk van die combinaties kan men 4 verschillende nagerechten kiezen: er zijn dus (3 × 5) × 4 verschillende menu's."
Elke uitkomst of menu bestaat uit een geordend drietal (voorgerecht, hoofdgerecht, nagerecht). Dus als we de gerechten nummeren dan vinden we

$$S = \{(i, j, k) | i = 1, 2, 3 \text{ en } j = 1, \ldots, 5 \text{ en } k = 1, \ldots, 4\}.$$

Men kan de berekening van het aantal menu's ook weergeven in een "keuzeboom".
De eerste vertakking geeft de keuze van het voorgerecht weer, de tweede het hoofgerecht en de derde het nagerecht. Het aantal eindpunten is het aantal menu's.



3 mogelijkheden 3×5 tweetallen 3×5×4 drietallen ■

In voorbeeld 2.1.1 was het experiment “kiezen van een willekeurig menu” opgedeeld in drie deelexperimenten voor het kiezen van resp. voor-, hoofd- en nagerecht.
Elk deelexperiment heeft een vast aantal mogelijke uitkomsten, ongeacht het resultaat van voorgaande deelexperimenten. Algemener geformuleerd:

Eigenschap 2.1.2 (**de productregel**) Als een experiment bestaat uit het uitvoeren van k deelexperimenten en het i-de deelexperiment n_i mogelijke uitkomsten ($i = 1, 2, \ldots, k$) heeft, ongeacht het resultaat van overige deelexperimenten, dan zijn er in totaal $\boldsymbol{n_1} \times \boldsymbol{n_2} \times \ldots \times \boldsymbol{n_k}$ verschillende uitkomsten (geordende k-tallen) van het experiment mogelijk.

Deze regel is eenvoudig te bewijzen met inductie naar k. Uit het volgende voorbeeld blijkt, dat men moet oppassen met het klakkeloos vermenigvuldigen van aantallen.

Voorbeeld 2.1.3 We willen de kans berekenen dat een willekeurig getal van drie cijfers als laagste cijfer een 2 heeft. Als we afspreken dat bijvoorbeeld 28 = $\boxed{028}$ ook een getal van drie cijfers is dan gaat het om de getallen 0 t/m 999. Omdat sprake is van een **willekeurig** getal nemen we aan dat de uitkomstenruimte bestaat uit 1000 even waarschijnlijke getallen.
We moeten dus nog het aantal getallen bepalen in de gebeurtenis A die optreedt als het laagste cijfer van het getal een 2 is.
Dit betekent dat van zo’n getal elk cijfer 8 mogelijke waarden kan hebben: 2, 3, 4, 5, 6, 7, 8 of 9. Maar $N(A) \neq 8 \times 8 \times 8$, immers: als de eerste twee cijfers 2 zijn, bijv. $\boxed{226}$, dan zijn er 8 mogelijke waarden voor het derde cijfer. Als de eerste twee cijfers groter dan 2 zijn, dan moet het derde cijfer een 2 zijn, bijv. $\boxed{632}$.
Het aantal mogelijkheden bij het derde deelexperiment (het kiezen van het derde cijfer) hangt dus af van het resultaat van de eerste twee deelexperimenten. De productregel is dus niet

van toepassing op deze deelexperimenten.
Een andere (foutieve) aanpak bij het bepalen van $N(A)$ maakt gebruik van een onderscheid naar de plaats, waar een 2 in het getal voorkomt: zij A_i de gebeurtenis dat het willekeurige getal een 2 op plaats i heeft ($i = 1, 2, 3$) en op de overige plaatsen cijfers groter dan of gelijk aan 2, dan geldt:

$$N(A_1) = 1 \times 8 \times 8 = N(A_2) = N(A_3) \quad \text{en} \qquad A = \bigcup_{i=1}^{3} A_i$$

$$\text{Echter:} \quad N(A) \neq \sum_{i=1}^{3} N(A_i) = 3 \times 8 \times 8$$

We kunnen $N(A)$ niet op deze wijze berekenen omdat A_1, A_2 en A_3 elkaar niet uitsluiten, omdat bijvoorbeeld $272 \in A_1 \cap A_3$. Een juiste berekening wordt gevraagd in opgave 2. ■

Het tellen door deelgebeurtenissen te beschouwen heeft dus alleen zin als deze deelgebeurtenissen een partitie vormen. In dat geval kunnen we nuttig gebruik maken van de volgende evidente eigenschap, die we overigens reeds gebruikten in het bewijs van eigenschap 1.2.7.

Eigenschap 2.1.4 Als $A_1, A_2, \ldots, A_k$ **elkaar uitsluitende** gebeurtenissen zijn. dan geldt:

$$N\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} N(A_i)$$

Voorbeeld 2.1.5 Bij een voetbalwedstrijd om de Europa Cup moet de beslissing, na verlenging, vallen door het nemen van 5 strafschoppen. Trainer B. Reker heeft zijn 5 strafschopspecialisten reeds gekozen en bepaalt de volgorde door zijn hulptrainer lootjes met de 5 namen uit zijn pet te laten trekken. Bij het trekken van het eerste lot zijn er 5 kandidaten voor het nemen van de eerste strafschop, bij het trekken van het tweede lot zijn er nog 4 kandidaten over, enz., zoals in nevenstaand voorbeeld:

positie	*1*	*2*	*3*	*4*	*5*
uitkomst	3	2	5	4	1

Er is dus sprake van 5 deelexperimenten met resp. 5, 4, 3, 2 en 1 mogelijke uitkomsten, die volgens de productregel tot 5 × 4 × 3 × 2 × 1 = 5! verschillende volgordes leiden.
De kans op één bepaalde volgorde in deze symmetrische kansruimte is dus $\frac{1}{5!} \approx 0.83\%$. ■

Meer algemeen geldt het volgende plaatje:

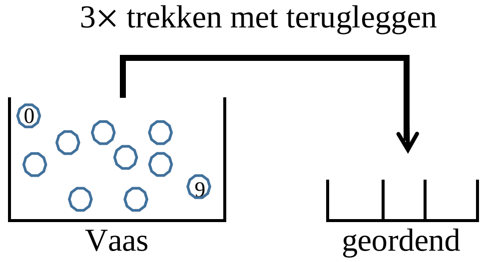
positie	*1*	*2*	*………*	*k*
uitkomst			……..	

Eigenschap 2.1.6 (**De permutatieregel**) Het aantal volgordes of **permutaties** waarin k verschillende dingen gezet kunnen worden is $\boldsymbol{k}$!

Voorbeeld 2.1.7 Gevraagd wordt de kans dat een willekeurig getal van drie cijfers één cijfer 2 bevat. Zoals we in voorbeeld 2.1.1 zagen, gaat het hier om een symmetrische kansruimte met $N(S) = 1000$ mogelijke uitkomsten. We kunnen een willekeurig getal van drie cijfers genereren, door uit een vaas met 10 balletjes, genummerd 0 t/m 9 aselect balletjes te trekken. De eerste trekking levert het eerste cijfer.
We leggen het eerste getrokken balletje terug en herhalen, na goed schudden, het experiment

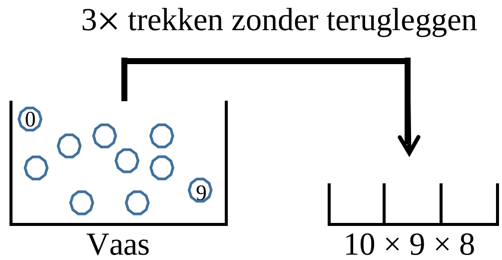
een tweede en derde keer om het tweede en derde cijfer te bepalen.

Schematisch kunnen we de trekkingen uit de vaas als volgt weergeven.



Elk van deze (deel)experimenten heeft 10 mogelijke uitkomsten, wat leidt tot
$N(S) = 10 \times 10 \times 10 = 1000$ mogelijke getallen van drie cijfers.
Zij A de gebeurtenis dat een getal van drie cijfers precies één cijfer 2 heeft, dan kunnen we zo'n getal bepalen door eerst een 2 als eerste, tweede of derde cijfer te kiezen en de overige twee cijfers ongelijk 2 te kiezen. Dit kan volgens de productregel op $N(A) = 3 \times 9 \times 9$ manieren. Dus $P(A) = \frac{N(A)}{N(S)} = 0.243$. ■

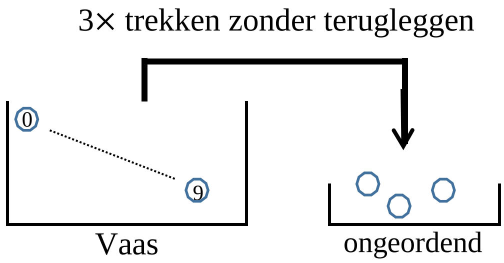
Voorbeeld 2.1.8 Een fabrikant maakt cijfersloten van drie verschillende cijfers.
Elke cijfercode komt even vaak voor.
Een fietsendief heeft 3 seconden nodig om na te gaan of een cijfercode juist is.
Hoe groot is de kans op de gebeurtenis A dat hij binnen 5 minuten het cijferslot los heeft?
Een willekeurig cijferslot kunnen we genereren door weer uit een vaas met tien balletjes genummerd 0 t/m 9 drie keer een balletje te trekken voor resp. het eerste, tweede en derde cijfer van de cijfercode. Om herhaling van cijfers te voorkomen, leggen we een getrokken balletje niet terug in de vaas, zodat zich na elke trekking één balletje minder in de vaas bevindt. Bij de eerste trekking zijn er dus 10, bij de tweede 9 en bij de derde 8 mogelijke uitkomsten, steeds ongeacht het resultaat van voorgaande trekkingen.



Dit levert volgens de productregel dus $N(S) = 10 \times 9 \times 8$ verschillende cijfersloten op, die alle met kans $\frac{1}{10\times9\times8}$ voorkomen.
$N(S) = 10 \times 9 \times 8$ kan geschreven worden als $\frac{10!}{7!}$ en wordt het **aantal variaties** van 3 uit 10 genoemd. In veel boeken ook wel het aantal **permutaties van 3 uit 10** genoemd.
De dief kan in één minuut 20 cijfercodes proberen, dus de kans $P(A)$ dat hij binnen vijf minuten het cijferslot los heeft is $\frac{N(A)}{N(S)} = \frac{100}{10\times9\times8} \approx 0.139$. ■

Bij het simuleren van kans-experimenten maken we vaak gebruik van het zgn. **vaasmodel**.
In bovenstaande voorbeelden trokken we drie keer uit een vaas met 10 genummerde balletjes, de ene keer **met terugleggen**, de andere keer **zonder terugleggen**. In beide gevallen beschouwden we **geordende** drietallen als uitkomsten. Een andere volgorde geeft immers een ander getal of een andere cijfercode.
De twee gevallen waarbij we **ongeordende** drietallen beschouwen, worden in de volgende twee voorbeelden behandeld.

Voorbeeld 2.1.9 We trekken aselect en zonder terugleggen 3 keer een balletje uit een bak met 10 balletjes genummerd 0 t/m 9. Bereken de kans op de gebeurtenis A dat de som van de getrokken nummers groter dan 5 is.



De volgorde van de trekkingen is niet van belang. De uitkomstenruimte S bestaat uit deelverzamelingen van 3 elementen uit de verzameling $\{0, 1, \ldots, 9\}$.

Beschouwen we één deelverzameling, bijv. $\{2, 5, 6\}$ dan kunnen we deze 3 cijfers in 3! volgordes zetten of er 3! verschillende cijfersloten mee maken zoals in het vorige voorbeeld. Dit geldt voor elke deelverzameling. Er zijn dus 3! keer zoveel cijfersloten als deelverzamelingen. Het aantal deelverzamelingen is dus het aantal cijfersloten gedeeld door 3!:

$$N(S) = \frac{10 \times 9 \times 8}{3!} = \frac{10!}{3!\,7!} = \binom{10}{3}$$

Omdat de cijfersloten, dus de geordende drietallen bij het trekken zonder terugleggen, een symmetrische kansruimte vormen, geldt dat ook voor de bijbehorende ongeordende uitkomsten (de deelverzamelingen).
De kans dat de som groter dan 5 is, kunnen we nu als volgt berekenen.

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{N(\overline{A})}{N(S)} = 1 - \frac{4}{120} \approx 0.97,$$

omdat $\overline{A}$ bestaat uit $\{0, 1, 2\}$, $\{0, 1, 3\}$, $\{0, 1, 4\}$ en $\{0, 2, 3\}$. ■

Het aantal $\binom{10}{3}$ berekent niet alleen het aantal deelverzamelingen (of combinaties) van 3 uit 10, maar ook bijvoorbeeld het aantal (geordende!) rijtjes met 3 enen en 7 nullen of het aantal termen a^3b^7 dat ontstaat bij het uitwerken van het binomium $(a + b)^{10}$.
Vanwege het laatste wordt $\binom{10}{3}$ ook wel een **binomiaalcoëfficiënt** genoemd.

We hadden de drie balletjes in voorbeeld 2.1.9 overigens ook tegelijk (in één trekking) uit de vaas kunnen halen. Dit komt op hetzelfde neer als drie keer één balletje trekken zonder terugleggen en zonder op de trekkingsvolgorde te letten.

Voorbeeld 2.1.10 We trekken drie keer aselect en met terugleggen een balletje uit een vaas met tien balletjes, genummerd 0 t/m 9. Bereken de kans op de gebeurtenis A dat er bij de getrokken balletjes één 2 zit. De volgorde van de trekkingen is niet van belang, dus kiezen we als uitkomstenruimte S de verzameling van alle **ongeordende** drietallen (“herhalingscombinaties”). Trekken we bijvoorbeeld twee keer een 2 en één 8, dan noteren we dat met 2 2 8. Dus 2 2 8 = 2 8 2 enz. en $A = \{2\,i\,j \mid i, j = 0, 1, 3, \ldots, 9\}$.
De zo gevormde kansruimte is echter niet symmetrisch en bijgevolg $P(A) \neq \frac{N(A)}{N(S)}$.
De uitkomst 2 2 8 komt bijvoorbeeld vaker voor dan 2 2 2. Dit is in te zien als we S **verfijnen** tot een uitkomstenruimte $S`$, die bestaat uit de bijbehorende **geordende** drietallen.
$S`$ vormt wel een symmetrische kansruimte, zoals we zagen in voorbeeld 2.1.7.
$2\,2\,8 \in S$ wordt verfijnd tot de gebeurtenis $\{(2, 2, 8), (2, 8, 2), (8, 2, 2)\} \subset S`$. En $2\,2\,2 \in S$ tot $(2, 2, 2) \in S`$. Dus $P(2\,2\,8) = 3 \cdot P(2\,2\,2)$. Als we ook $A`$ definiëren als de gebeurtenis dat de geordende uitkomst één 2 bevat, dan geldt:

$$P(A) = P(A`) = \frac{N(A`)}{N(S`)} = \frac{3 \times 9 \times 9}{10 \times 10 \times 10} = 0.243. \qquad ■$$

In de voorbeelden 2.1.7 t/m 2.1.10 hebben we op 4 verschillende manieren drie keer aselect getrokken uit een verzameling van 10 verschillende elementen. Alleen in het laatste geval was er geen sprake van een symmetrische kansruimte: in de andere drie gevallen kunnen we dus meteen kansen op gebeurtenissen m.b.v. de kansdefinitie van Laplace berekenen. Schematisch kunnen we het resultaat als volgt weergeven.

		wijze van trekken	
		zonder terugleggen	met terugleggen
Uitkomst	geordend	$S=$ {getallen van 3 verschillende cijfers} $\boldsymbol{N(S)=10\times 9\times 8=\frac{10!}{7!}}$	$S=$ {getallen van 3 cijfers} $\boldsymbol{N(S)=10^3}$
	ongeordend	$S=$ {deelverzamelingen met 3 (verschillende) cijfers} $\boldsymbol{N(S)=\binom{10}{3}}$	$S=$ {ongeordende drietallen met mogelijke herhalingen} (S,P) is **niet symmetrisch**: overgaan op de symmetrische kansruimte van geordende drietallen.

In het laatste geval hebben we het aantal elementen van de uitkomstenruimte S niet berekend. Dit is voor het berekenen van kansen op gebeurtenissen ook niet van belang, zoals we zagen in voorbeeld 2.1.10. We kunnen deze vier gevallen generaliseren tot het k keer aselect trekken uit een verzameling van n verschillende elementen.

Eigenschap 2.1.11 Als we k keer aselect trekken uit een verzameling van n verschillende elementen, dan is de kansruimte in de volgende 3 gevallen **symmetrisch**:

a. Trekken met terugleggen, geordende uitkomsten: $N(S)=\boldsymbol{n^k}$.

b. Trekken zonder terugleggen, geordende uitkomsten (**variaties/permutaties van $\boldsymbol{k}$ uit $\boldsymbol{n}$**):

$$N(S)=n\cdot(n-1)\cdot\ldots\cdot(n-k+1)=\frac{\boldsymbol{n}!}{(\boldsymbol{n}-\boldsymbol{k})!}$$

c. Trekken zonder terugleggen, ongeordende uitkomsten (**combinaties van $\boldsymbol{k}$ uit $\boldsymbol{n}$**).

$$N(S)=\binom{\boldsymbol{n}}{\boldsymbol{k}}$$

In geval

d. Trekken met terugleggen, ongeordende uitkomsten:
er is sprake van een **niet-symmetrische kansruimte**. We kunnen dan overgaan op een symmetrische kansruimte door de bijbehorende geordende uitkomsten te beschouwen (geval a))

Bewijs: dit verloopt analoog aan de afleiding gegeven in de voorbeelden 2.1.7 t/m 2.1.10 voor het geval dat $k=3$ en $n=10$. Zo kunnen we het aantal combinaties van k uit n (geval c.) berekenen uit het aantal $\frac{n!}{(n-k)!}$ der variaties (geval b.) van k uit n: elke combinatie van k uit n kan in $k!$ volgordes worden geschreven. Er zijn dus $\frac{n!}{(n-k)!}/k! = \binom{n}{k}$ combinaties van k uit n. Tevens volgt hieruit dat dit ook geldt voor de combinaties, daar de variaties een symmetrische kansruimte vormen. ■

Opmerking 2.1.12 We komen nog een keer terug op het voorbeeld van herhalingscombinaties in voorbeeld 2.1.10. Omdat dit geval d. is van bovenstaande

eigenschap is het aantal herhalingscombinaties van 3 uit 10 niet van belang voor het berekenen van kansen, maar volledigheidshalve merken we op dat dit aantal $\binom{12}{3}$ is. Dat kunnen beredeneren door elke herhalingscombinatie kunnen we op unieke wijze representeren met 9 enen en 3 nullen, bijv. 2 2 8 met 110011111101 en moet gelezen als volgt worden: geen o voor de eerste 1 en tussen de eerste en tweede 1 betekent dat 0 en 1 niet voorkomen: twee nullen tussen de tweede en derde 1 betekent twee 2-en, de 0 tussen de achtste en negende 1 betekent een 8. Voor elke herhalingscombinatie is er zo één code en omgekeerd. Dus het aantal herhalingscombinaties is het aantal volgordes van 3 nullen en 9 enen: $\binom{12}{3}$
het aantal herhaling combinaties van k uit n is volgens dezelfde redenering $\binom{n+k-1}{k}$ ■

Er volgen nog enkele voorbeelden waarin we de beschreven situatie met wèl symmetrische kansruimten beschrijven en dus deze eigenschap toepassen

Voorbeeld 2.1.13 We werpen 6 (zuivere) munten op en willen de kans bepalen dat er 2 *K*(ruis) en 4 *M*(unt) opleveren. Een uitkomst is blijkbaar een zestal, waarbij de volgorde niet van belang is. De uitkomstenruimte *S* kunnen we dus laten bestaan uit de 7 (ongeordende) uitkomsten, die we aanduiden met $6{\times}K$, $5{\times}K$, . . . , $0{\times}K$.
Deze uitkomsten zijn echter niet even waarschijnlijk.
We verfijnen de uitkomstenruimte, door de munten in gedachten te nummeren (onderscheiden). Het resultaat is een symmetrische kansruimte van $N = 2^6$ geordende zestallen. De kans op A = "$2{\times}K$" berekenen we uit het aantal rijtjes met 2 *K*'s en 4 *M*-en en uit het totaal aantal mogelijke geordende zestallen.
De berekening van beide aantallen kunnen we toelichten met het volgende plaatjes:

<table>
<tr>
<td>
<table>
<tr><td>positie</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td></tr>
<tr><td>uitkomst</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>
Op elke positie staat een K of een M, dus 2 mogelijkheden voor elke positie: totaal 2×2×2×2×2×2 mogelijkheden.
</td>
<td>
Voorbeeld met twee keer K (en dus 4 keer M):
<table>
<tr><td>positie</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td></tr>
<tr><td>uitkomst</td><td>M</td><td>K</td><td>M</td><td>M</td><td>M</td><td>K</td></tr>
</table>
Het aantal mogelijkheden is dus gelijk aan het aantal manieren waarop je een combinatie van 2 posities voor de K's uit 6 kunt kiezen. Ofwel, de 4 posities van de M's: $\binom{6}{2} = \binom{6}{4}$
</td>
</tr>
</table>

$$\text{Dus } P(A) = \frac{N(A)}{N} = \frac{\binom{6}{2}}{2^6} = \frac{15}{64} \approx \ 0.23$$ ■

Voorbeeld 2.1.14 Hoe groot is de kans dat 2 vrienden bij elkaar in de groep komen als zij deel uitmaken van 15 personen die willekeurig over een groep van 7 (groep 1) en één van 8 (groep 2) personen worden verdeeld? "Willekeurig" wil in dit geval zeggen dat, bijvoorbeeld, uit een bak met 15 briefjes met namen er 7 aselect en zonder terugleggen worden getrokken: deze vormen groep 1, de overige groep 2.
De verzameling van alle mogelijke verdelingen vormt dus een symmetrische kansruimte (S, P) met $N(S) = \binom{15}{7}$.
Dit aantal kan men ook als volgt beredeneren: we kunnen de 15 personen in 15! volgordes zetten. Wijzen we steeds de eerste 7 personen toe aan groep 1 en de overige aan groep 2, dan vinden we zo alle mogelijke verdelingen over de 2 groepen. De volgorde van de eerste 7 en van de laatste 8 personen zijn echter niet van belang voor de groepsindeling. Bij elke

groepsindeling zijn er 7!×8! volgordes van 15 personen aan te wijzen, die tot die groepsindeling leiden: er zijn dus $\frac{15!}{7!8!} = \binom{15}{7}$ groepsindelingen mogelijk.
De gebeurtenis A dat 2 vrienden in dezelfde groep zitten, kan men opdelen in twee (elkaar uitsluitende) gebeurtenissen A_1 en A_2 waarbij A_i de gebeurtenis is dat de twee vrienden in groep *i* zitten. $N(A_1)$ berekent men door, uitgaande van het feit dat de 2 vrienden in groep 1 zitten, van de overige 13 personen nog 5 personen aan groep 1 en 8 personen aan groep 2 toe te voegen. Dus $N(A_1) = \binom{13}{5}$. Evenzo $N(A_2) = \binom{13}{6}$.
Toepassing van eigenschap 2 levert ons de kans die wat kleiner dan 50% is:

$$P(A) = \frac{N(A)}{N(S)} = \frac{N(A_1) + N(A_2)}{N(S)} = \frac{\binom{13}{5} + \binom{13}{6}}{\binom{15}{7}} \approx 46.7\%. \qquad \blacksquare$$

In klassen van ca. 30 leerlingen blijken vaak leerlingen op dezelfde dag jarig te zijn. Is dit toeval? Of is de kans hierop bijvoorbeeld groter dan 50%? Als we een dergelijke kans willen berekenen, zullen we veronderstellingen m.b.t. de verjaardagen moeten maken: een (kans)model van de werkelijkheid. Het ligt in dit geval misschien voor de hand te veronderstellen dat de verjaardag van een willekeurig gekozen persoon op een willekeurige dag van het jaar valt. De dagen van het jaar vormen dan een symmetrische kansruimte.
Bij statistische controle blijkt deze veronderstelling niet helemaal juist. We moeten dus bij de berekening van een kans beseffen dat deze kans geldt voor het **model**, dat we van de werkelijkheid hebben gemaakt.

Voorbeeld 2.1.15 Beschouw een willekeurig gekozen groep van n mensen.
Wat is de kans dat twee of meer van hen op dezelfde dag jarig zijn?
Gemakshalve nemen we aan dat niemand jarig is op 29 februari. Het kiezen van een persoon beschouwen we nu als het kiezen van de dag waarop hij jarig is en we nemen aan dat ieder van de 365 dagen gelijke kans heeft om gekozen te worden. Het kiezen van een groep van n personen is dan het trekken, **geordend en met terugleggen**, van n dagen uit 365, waarbij elke uitkomst, dus een serie van n verjaardagen, even waarschijnlijk is. We hebben dus een symmetrische kansruimte met 365^n uitkomsten.
Als A de gebeurtenis is dat twee of meer van de getrokken dagen samenvallen dan is $\overline{A}$ de gebeurtenis dat alle dagen in de trekking verschillend zijn. De gebeurtenis $\overline{A}$ bestaat uit alle uitkomsten die geen herhalingen bevatten, dus uit alle variaties van n uit 365 dagen. Het aantal van deze variaties is $\frac{365!}{(365-n)!}$, zodat de gevraagde kans gegeven wordt door:

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{365!/(365-n)!}{365^n}$$

Numerieke berekening leert dat reeds voor *n* = 23 personen geldt dat $P(A) > \frac{1}{2}$

n	1	2	3	4	……	22	23	24
$P(A) \approx$	0	0.0027	0.0082	0.0164		0.4757	0.5073	0.5383

■

2.2 Combinatoriek en stochastische variabelen

Voorbeeld 2.2.1 Uit een vaas met 10 rode en 5 witte balletjes trekken we er aselect en zonder terugleggen vier. Bereken de kans dat we precies 3 rode balletjes (en dus 1 witte) trekken.
Schematisch kunnen we dit als volgt weergeven:

	Rood	Wit	Totaal
Vaas	10	5	15
	↓	↓	↓
Trekking	3	1	4

Om een symmetrische uitkomstenruimte te verkrijgen, nummeren we de balletjes (in gedachten) 1 t/m 15, waarbij de rode de nummers 1 t/m 10 krijgen.
Dan bestaat de uitkomstenruimte uit $\binom{15}{4}$ ongeordende uitkomsten.
De gebeurtenis $A =$ "3 rood en 1 wit" treedt op als een (ongeordende) uitkomst bestaat uit drie balletjes met nummers 1 t/m 10 en één van de balletjes met nummers 11 t/m 15. De rode balletjes kunnen we op $\binom{10}{3}$ manieren kiezen, vervolgens de witte op $\binom{5}{1} = 5$ manieren.
De gevraagde kans is dus

$$P(A) = \frac{\binom{10}{3} \cdot \binom{5}{1}}{\binom{15}{4}} \approx 44.0\%$$ ■

Deze laatste uitdrukking wordt, algemener uitgedrukt, wel de hypergeometrische formule genoemd.

Eigenschap 2.2.2 (**hypergeometrische formule**)
We trekken n keer aselect en zonder terugleggen uit een verzameling van N balletjes, bestaande uit R rode en $N - R$ witte balletjes. Dan is de kans op de gebeurtenis A_k dat we k rode balletjes (en dus $n - k$ witte) trekken gelijk aan

$$\boldsymbol{P(A_k) = \frac{\binom{R}{k} \cdot \binom{N-R}{n-k}}{\binom{N}{n}}}$$

	Rood	Wit	Totaal
Vaas	R	$N - R$	N
	↓	↓	↓
Trekking	k	$n - k$	n

Het aantal getrokken rode balletjes ligt tussen 0 en R, dus $0 \leq k \leq R$.
Evenzo geldt voor de witte balletjes: $0 \leq n - k \leq N - R$. ■

Met deze formule kunnen we dus kansen op een specifiek aantal rode balletjes (en het bijbehorende aantal witte balletjes. Zo'n aantal is een **stochastische variabele $\boldsymbol{X}$**, een kwantitatieve variabele bij een toevalsexperiment. In voorbeeld 2.2.1 is X "het aantal rode balletjes bij 4 trekkingen zonder terugleggen" uit de gegeven vaas.
De gebeurtenis $A =$ "3 rode balletjes" kunnen we ook met "$X = 3$" aanduiden, of met "$Y = 1$", waarin Y het aantal witte balletjes is.
$X = 3$ en $Y = 1$ zijn immers equivalent, dus $P(X = 3) = P(Y = 1) = P(A) = \frac{\binom{10}{3} \cdot \binom{5}{1}}{\binom{15}{4}}$.

Indien we voor X alle mogelijke kansen $P(X = k)$ bepalen met de hypergeometrische formule (dus voor $k = 0, 1, 2, 3$ en 4), spreken van de **hypergeometrische kansverdeling** van X.

Dus in dit geval: $P(X = k) = \frac{\binom{10}{k} \cdot \binom{5}{4-k}}{\binom{15}{4}}$, voor $k = 0, 1, 2, 3, 4$.

$N = 15$, $R = 5$ en $n = 4$ worden de **parameters** van de hypergeometrische verdeling genoemd. In hoofdstuk 4 zullen we een algemene formele definitie van stochastische variabelen en hun kansverdelingen geven.

2.3 Vraagstukken

1. Een aantal telproblemen.

- **a.** Bereken het aantal volgordes waarin je de cijfers 1 t/m 7 kunt zetten.
- **b.** Als we met de cijfers 0, 1, …, 9 een getal met 7 verschillende cijfers maken, hoeveel van die getallen kunnen we dan maken?
- **c.** Hoeveel combinaties van 7 cijfers uit de cijfers 0, 1, …, 9 zijn er (het getal 012 kennen wij bijv. als 12)?
- **d.** Een pak kaarten bestaat uit 16 plaatjes en 36 cijferkaarten. Bij het spel bridge krijgt een speler 13 van de 52 kaarten (willekeurig) toebedeeld. Wat is de kans dat hij minstens 2 plaatjes heeft?
- **e.** Op hoeveel manieren kunnen we 30 personen over 4 groepen van 6, 7, 8 en 9 personen verdelen?

2. Hoe groot is de kans dat er in een willekeurig getal van 3 cijfers als laagste cijfer een 2 voorkomt (zie voorbeeld 2.1.3)?

3. We trekken vier keer aselect en zonder terugleggen uit een bak met 3 rode en 7 witte balletjes. Bereken de kans dat het vierde getrokken balletje rood is.

4. We trekken vijf keer zonder terugleggen lukraak een kaart uit een spel met 52 kaarten.

- **a.** Bereken de kans dat alle getrokken kaarten plaatjes (aas, heer, vrouw of boer) zijn.
- **b.** Bereken de kans dat de eerste kaart een aas en de laatste kaart een heer is.
- **c.** Bereken de kans dat van de 5 getrokken kaarten er één een aas en één een heer is.

5. Iemand beweert een wijnkenner te zijn. Men gaat zijn deskundigheid op de proef stellen. Hij krijgt de namen van 6 hem bekende wijnsoorten en 6 glazen wijn, van iedere soort één. Na het proeven moet hij zeggen in welk glas welke wijnsoort zit. Iedere soort moet hij één keer noemen. Men gelooft pas in zijn deskundigheid als hij minstens 4 goede antwoorden geeft.
Hoe groot is de kans dat men in zijn deskundigheid gelooft terwijl hij in het geheel geen kenner is? Geef bij beantwoording aan welke veronderstellingen gemaakt zijn.

6. Een fabrikant van rubberringen garandeert dat niet meer dan 10% van de ringen ondeugdelijk is. Deze ringen worden verkocht in pakjes van 100. Een bepaalde afnemer heeft de gewoonte om uit ieder pakje van 100 ringen er willekeurig 10 te kiezen en deze te testen op deugdelijkheid. Zodra bij het testen een ring ondeugdelijk blijkt te zijn, weigert de afnemer het pakje waaruit die ring komt.

a. Hoe groot is de kans dat een pakje wordt geweigerd terwijl het nog juist aan de garantie voldoet?
b. Hoe groot is de kans dat een pakje, waarvan 20% ondeugdelijk is, toch wordt aanvaard?

7. Een loterij bestaat uit 100 loten. Er zijn 4 hoofdprijzen en 10 troostprijzen. Trekking vindt plaats zonder teruglegging. Als men 5 loten koopt, wat is dan de kans op

a. één hoofdprijs en één troostprijs,
b. precies één prijs,
c. geen prijs,
d. tenminste één prijs?

8. Het kopen van een lot in opgave 7 kan men zien als een trekking uit een bak met 3 soorten dingen: 4 hoofdprijzen, 10 troostprijzen en 86 niet-prijzen.
We kunnen dit als volgt generaliseren: in een doos bevinden zich N dingen. Daarvan zijn er N_1 van soort 1, N_2 van soort 2, . . .,N_k van soort k, met $N_1 + N_2 + \cdots + N_k = N$
We trekken aselect en zonder terugleggen n dingen uit de doos. Wat is de kans op n_1 dingen van soort 1, n_2 dingen van soort 2, . . . , n_k dingen van soort k $(n_1 + \cdots + n_k = n)$

Een paar extra oefenopgaven:

9. Tussen Peter en Paul worden 26 kaarten eerlijk verdeeld. Onder de kaarten bevinden zich precies 6 schoppen. Bepaal de kans dat

a. één van beiden precies 4 schoppen krijgt,
b. beiden 3 schoppen krijgen.

10. (de multinomiaal coëfficiënt)
a. Op hoeveel manieren kunnen we 13 personen over 2 groepen van 6 en 7 personen verdelen?
b. Op hoeveel manieren kunnen we 30 personen over 4 groepen van 6, 7, 8 en 9 personen verdelen?
c. Op hoeveel manieren kunnen we n personen over k groepen van resp. $n_1, n_2, \ldots, n_k$ personen verdelen?
Dit aantal wordt de **multinomiaal coëfficiënt** genoemd (generalisatie van de binomiaalcoëfficiënt).

11. Op willekeurige wijze worden 5 nullen en 6 enen in een rij geplaatst. Een maximale ononderbroken deelrij van identieke symbolen noemen we een run.
Voorbeeld: de rij **0 1 000 11 0 111** telt 6 runs. De lengte van een run is het aantal symbolen dat de run bevat.

a. Bereken de kans dat de rij begint met een run van lengte 3.
b. Bereken de kans dat de rij 5 runs bevat. 7.

12. Beschouw een vaas met 10 ballen, genummerd van 1 t/m 10.
We trekken hieruit aselect 4 ballen zonder teruglegging.
Wat is de kans dat de getrokken ballen toenemende nummers hebben?

Enige aanwijzingen bij de opgaven van hoofdstuk 2:

1. Ga steeds na of de volgorde van belang is (permutaties) of niet (combinaties).
schrijf de correcte formules en gebruik je rekenmachine om faculteiten en aantallen combinaties of permutaties te berekenen.
Bij d.: vergelijk met voorbeeld 2.2.1
Bij e.: beredeneer de uitkomst door eerst het aantal mogelijkheden te bepalen waarop groep 1 gevuld kan worden, hoeveel mogelijkheden er dan nog zijn om groep 2 te vullen, etc.

3. Los dit **niet** op door te "conditioneren", d.w.z. onderscheid maken wat bij de eerste drie trekkingen heeft plaatsgehad (bijv. 1 rode bij de eerste drietrekkingen en vervolgens de kans op een rode bij de 4de trekking). Bepaal het totaal aantal mogelijke trekkingsresultaten en tel vervolgens hoeveel daarvan op de 4de positie een rode hebben.

4. Ga steeds na of de volgorde van belang is (permutaties) of niet (combinaties). Als beide mogelijk is, kies dan voor combinaties.

5. Stel et je voor: 6 glazen en 6 naamkaartjes: wat is het totale aantal volgordes waarin je de naamkaartjes kunt plaatsen, en hoeveel volgordes zijn er met 6 goed? 5 goed? 4 goed?

6. Hoeveel zijn er ondeugdelijk als hij "nog juist voldoet aan de garantie"? Zijn het trekkingen met of zonder terugleggen?

7. Het gaat nu niet om twee verschillende dingen, maar 3 verschillende. Met of zonder terugleggen?

11. **a**. Hoe ziet zo'n rijtje er uit die met een run van (precies!) 3 begint? Hoeveel verschillende van die rijtjes kun je vormen met 5 nullen en 6 enen?

12. Bekijk een combinatie van 4 getallen, bijv. 2, 8, 4 , 6. Hoeveel volgordes kun je hiermee maken? En hoeveel daarvan zijn opklimmend? Hoeveel opklimmende volgordes kun je dus met 10 cijfers maken?

Hoofdstuk 3 Voorwaardelijke kans en onafhankelijkheid

3.1 Voorwaardelijke kans

Het komt dikwijls voor dat over de uitkomst van een experiment van te voren al iets bekend is, of iets verondersteld wordt. Als we een willekeurige voorbijganger in Amsterdam aanspreken dan is de kans dat hij op de VVD stemt een andere dan voor iemand uit Wassenaar. We kunnen ons het experiment zo voorstellen dat we lukraak een Nederlander aanwijzen, echter als extra voorwaarde stellen dat hij in Amsterdam moet wonen. Onder deze voorwaarde kunnen we vragen naar de kans dat hij op de VVD stemt. We kunnen deze "kans" niet berekenen met de hulpmiddelen die we tot nu toe hebben. Er is een nieuwe definitie nodig voor dit begrip. In het voorbeeld vragen we naar de **voorwaardelijke kans** dat een willekeurige Nederlander op de VVD stemt, **gegeven** dat hij in Amsterdam woont.
Intuïtief ligt het voor de hand om deze voorwaardelijke kans gelijk te stellen aan de kans dat een willekeurige **Amsterdammer** op de VVD stemt.
Als we met A de Nederlanders aangeven die op de VVD stemmen en B de Amsterdammers, dan is deze voorwaardelijke kans gelijk aan $\frac{N(AB)}{N(B)}$ waarbij $N(B)$ het aantal Amsterdammers is en $N(AB)$ het aantal daarvan dat op de VVD stemt.
We duiden deze voorwaardelijke kans aan met $P(A|B)$.
Als N het aantal Nederlanders is, dan geldt:

$$P(A|B) = \frac{N(AB)}{N(B)} = \frac{N(AB)/N}{N(B)/N} = \frac{P(AB)}{P(B)}$$

Men moet de kansen $P(A)$, $P(A|B)$ en $P(AB)$ goed onderscheiden. In dit voorbeeld kan men deze kansen als volgt interpreteren:

- $P(A)$: fractie VVD-stemmers onder de Nederlanders
- $P(A|B)$: fractie VVD-stemmers onder de Amsterdammers
- $P(AB)$: fractie VVD-stemmende Amsterdammers onder de Nederlanders.

In bovenstaande konden we de voorwaardelijke kans $P(A|B)$ berekenen met onvoorwaardelijke kansen $P(AB)$ en $P(B)$. Dit intuïtieve resultaat voor deze symmetrische kansruimte gebruiken we nu als definitie van het begrip voorwaardelijke kans voor willekeurige kansruimten:

Definitie 3.1.1 Als A en B gebeurtenissen zijn en $P(B) > 0$ dan noemen we

$$\boldsymbol{P(A|B) = \frac{P(AB)}{P(B)}}$$

Zeg: "**de (voorwaardelijke) kans op $\boldsymbol{A}$ onder de voorwaarde $\boldsymbol{B}$**"
(of: "de (voorwaardelijke) kans op A gegeven B").

Verder volgt uit de definitie dat **voor vaste $\boldsymbol{B}$ met $\boldsymbol{P(B) > 0}$** de voorwaardelijke kans een kans is, d.w.z. dat hij voldoet aan de axioma's van Kolmogorov.
De eis $P(B) > 0$ is geen grote beperking, daar een gebeurtenis met kans 0 niet opgetreden kan zijn! Axioma 2. geldt bijvoorbeeld ook voor de voorwaardelijke kans, want:

$$P(S|B) = \frac{P(SB)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

Dus ook $\big(S, P(\cdot\,|B)\big)$ is een kansruimte. Bijgevolg heeft deze voorwaardelijke kans ook de uit de axioma's afgeleide eigenschappen. Er geldt dus dat $P\big(\overline{A}\big|B\big) = 1 - P(A|B)$ etc.

Voorbeeld 3.1.2 Een bedrijf bezit twee fabrieken die dezelfde soort producten maken. Gedurende een bepaalde periode maakt fabriek 1 1000 producten waarvan er 100 defect zijn en fabriek 2 4000 producten waarvan er 200 defect zijn. Uit de totale productie wordt lukraak 1 product gekozen en dit product blijkt defect te zijn. Wat is de kans dat het afkomstig is van fabriek 1? Als A_1 de productie van fabriek 1 is en A_2 die van fabriek 2, terwijl D de verzameling van defecte producten is, dan is $P(A_1|D)$ de gevraagde kans.
De totale productie is 5000 producten, het aantal defecte producten afkomstig van fabriek 1 is 100 en het totaal aantal defecte producten is 300, dus intuïtief is duidelijk dat één op de drie defecte producten uit fabriek 1 afkomstig zijn.
Controle m.b.v. de definitie:

$$P(A_1D) = \frac{100}{5000}, P(D) = \frac{300}{5000} \text{ en } P(A_1|D) = \frac{P(A_1D)}{P(D)} = \frac{1}{3} \qquad \blacksquare$$

Voorbeeld 3.1.3 Een doos bevat twee munten, waarvan er één zuiver is, d.w.z. kruis en munt worden met gelijke kans geworpen, terwijl de andere aan beide zijden kruis heeft. We kiezen nu lukraak één van de munten en werpen deze. Als het resultaat van deze worp kruis is, wat is dan de kans dat de andere kant van deze munt ook kruis is?
Als A de gebeurtenis is dat de uitkomst kruis is en B de gebeurtenis dat de geworpen munt aan beide zijden kruis heeft dan is $P(B|A)$ de gevraagde voorwaardelijke kans.
Nu geldt $P(A) = \frac{3}{4}$, immers drie van de in totaal 4 zijden van de munten zijn kruis en deze zijden komen alle met gelijke kans boven. Bovendien geldt $P(BA) = P(B) = \frac{1}{2}$, want $B \subset A$ en beide munten hebben gelijke kans om getrokken te worden. Dus $P(B|A) = \frac{P(BA)}{P(A)} = \frac{2}{3}$.

Uit de definitie van een voorwaardelijke kans volgt onmiddellijk: $\boldsymbol{P(AB) = P(A) \cdot P(B|A)}$.
Dit wordt ook wel de **productregel** genoemd.
Substitueren we voor $A = A_1A_2$ en $B = A_3$ dan geldt voor de doorsnede van A_1, A_2 en A_3:

$$P(A_1A_2A_3) = P(A_1A_2) \cdot P(A_3|A_1A_2) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1A_2)$$

Van deze eigenschap maken we vaak intuïtief gebruik, zoals uit het volgende voorbeeld blijkt.

Voorbeeld 3.1.4 We trekken drie keer zonder terugleggen lukraak uit een vaas met 10 knikkers, genummerd 1 t/m 10. We zijn geïnteresseerd in de kans dat we, bijvoorbeeld, achtereenvolgens de knikkers 1, 2 en 3 trekken.
De kansruimte is symmetrisch en $S = \{(i,j,k)|i,j,k = 1,2,\ldots,10 \text{ en } i,j,k \text{ verschillend}\}$ bestaat uit $N(S) = \frac{10!}{(10-3)!}$ uitkomsten

Dus $P\big((1,2,3)\big) = \frac{1}{10\cdot 9\cdot 8}$

Dit antwoord kunnen we ook (intuïtief) afleiden door te stellen dat de kans $\frac{1}{10}$ is dat de eerste trekking knikker 1 oplevert; vervolgens is de kans $\frac{1}{9}$, dat de 2 uit de overige 9 getrokken wordt en ten slotte is de kans $\frac{1}{8}$, dat 3 wordt getrokken, als de eerste twee keren resp. 1 en 2 zijn getrokken.

Veel kansrekenaars stellen nu meteen: "de kans op (1, 2, 3) is dus $\frac{1}{10}\cdot\frac{1}{9}\cdot\frac{1}{8}$."

Waarom is deze vermenigvuldiging juist? We definiëren A_i als de gebeurtenis dat bij de i-de trekking knikker i wordt getrokken (i = 1, 2, 3). Dan geldt:

$$P(A_1) = \frac{1}{10} \quad P(A_2|A_1) = \frac{1}{9} \quad \text{en } P(A_3|A_1A_2) = \frac{1}{8}$$

Volgens bovenstaande regel geldt dan

$$P(A_1A_2A_3) = \frac{1}{10}\cdot\frac{1}{9}\cdot\frac{1}{8}$$

We merken terzijde op dat $A_2 = \{(i,2,k)|i,k = 1,3,4,\ldots,10 \text{ en } i \neq k\}$.

Dus $P(A_2) = \frac{N(A_2)}{N(S)} = \frac{9\cdot 8}{10\cdot 9\cdot 8} = \frac{1}{10} = P(A_1)$.

Evenzo geldt dat $P(A_3) = \frac{1}{10}$, en dus $P(A_1A_2A_3) \neq P(A_1)\cdot P(A_2)\cdot P(A_3)$. ■

Algemener formuleren we:

Eigenschap 3.1.5 (algemene productregel)

Voor n gebeurtenissen $A_1, A_2, \ldots, A_n$ met $n \geq 2$ en $P(A_1A_2\ldots A_{n-1}) > 0$ geldt:

$$\boldsymbol{P(A_1A_2\ldots A_n) = P(A_1)\cdot P(A_2|A_1)\cdot\ldots\cdot P(A_n|A_1A_2\ldots A_{n-1})}.$$

Het bewijs volgt uit de definitie van voorwaardelijke kans met inductie naar n, zoals we hiervoor de formule voor $n = 3$ afgeleid hebben.

3.2 Wet van de totale kans en regel van Bayes

Voorbeeld 3.2.1 Een bedrijf heeft drie fabrieken (1, 2 en 3) die allen dezelfde smartphones maken. Deze fabrieken maken resp. 15%, 35% en 50% van de totale productie. De kans dat een geproduceerde smartphone defect is, is voor deze fabrieken resp. 0.01, 0.05 en 0.02. Als we nu de aankoop van een smartphone beschouwen als het lukraak trekken van een smartphone uit de totale productie, vragen we ons het volgende af:

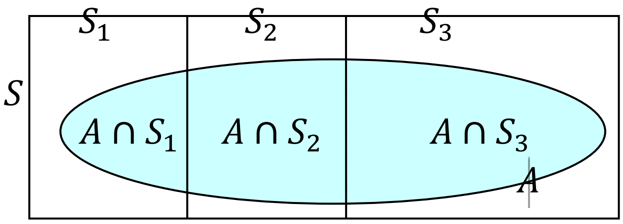
- **a.** Wat is de kans dat de smartphone defect is?
- **b.** Als hij defect blijkt te zijn, wat is dan de kans dat hij afkomstig is van fabriek 1?

Vraag a. is nog gemakkelijk intuïtief te beantwoorden: de kans op een defecte smartphone is het gewogen gemiddelde van de gegeven kansen op defecte smartphones, met als wegingsfactor het aandeel van de betreffende fabriek in de totale productie:

$$0.01\cdot 0.15 + 0.05\cdot 0.35 + 0.02\cdot 0.50 = 2.9\%.$$

Bij vraag b. is een intuïtieve aanpak veel lastiger. Daarnaast streven we er bij dit vak bij voorkeur naar ook intuïtieve berekeningen te onderbouwen via een kansmodel en kansregels. Reden te meer dus om met een kansmodel deze situatie te beschrijven. Daartoe definiëren we

S als de verzameling van alle geproduceerde smartphones, S_i is de gebeurtenis dat de smartphone afkomstig is van fabriek i en A de gebeurtenis dat hij defect is (zie onderstaande Venndiagram).



De aandelen in de productie van de drie fabrieken zijn de gegeven kansen

$$P(S_1) = 0.15, P(S_2) = 0.35 \text{ en } P(S_3) = 0.50.$$

De gegeven kansen op een defecte smartphone zijn voorwaardelijk:

$$P(A|S_1) = 0.01, P(A|S_2) = 0.05 \text{ en } P(A|S_3) = 0.02.$$

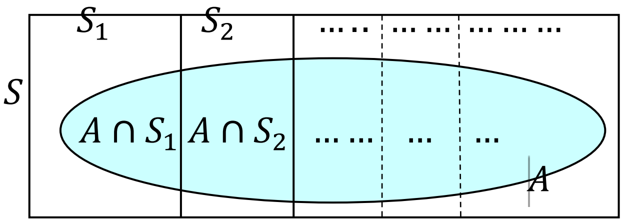
Volgens de productregel geldt $P(AS_i) = P(A|S_i) \cdot P(S_i)$, dus het antwoord op vraag a luidt:

$$\begin{aligned} P(A) &= P(AS_1) + P(AS_2) + P(AS_3) \\ &= P(A|S_1) \cdot P(S_1) + P(A|S_2) \cdot P(S_2) + P(A|S_3) \cdot P(S_3) \\ &= 0.01 \cdot 0.15 + 0.05 \cdot 0.35 + 0.02 \cdot 0.5 = 2.9\%. \end{aligned}$$

Bij vraag b. gaat het om de voorwaardelijke kans, dat een smartphone in fabriek 1 is geproduceerd, gegeven dat hij defect is, dus:

$$P(S_1|A) = \frac{P(S_1A)}{P(A)} = \frac{P(A|S_1) \cdot P(S_1)}{P(A)} = \frac{0.01 \cdot 0.15}{0.029} \approx 5.2\% \qquad \blacksquare$$

In dit voorbeeld vormen S_1, S_2 en S_3 een partitie van S. De kansen op de delen S_i en de voorwaardelijke kansen op een defecte smartphone zijn gegeven. Bij a. berekenden we daaruit de “overall” kans op een defecte smartphone. We formuleren nu een eigenschap die geldt voor een partitie $\{S_i\}$ van S: $\{S_i\}$ stelt weer een eindige reeks delen S_1 t/m S_n, zoals in voorbeeld 3.2.1, of een aftelbaar oneindige reeks delen S_1, S_2, . . . voor.



Eigenschap 3.2.2 (De wet van de totale kans)

Als $\{S_i\}$ een partitie is van S met $P(S_i) > 0$ voor alle i, dan geldt voor iedere gebeurtenis A:

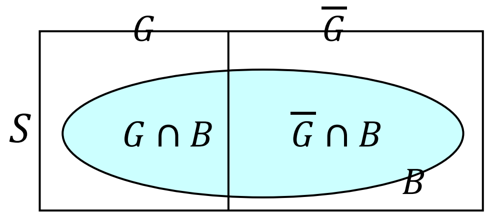
$$\boldsymbol{P(A)} = P(A|S_1) \cdot P(S_1) + P(A|S_2) \cdot P(S_2) + \cdots = \sum_{\boldsymbol{i}} \boldsymbol{P(A|S_i) \cdot P(S_i)}$$

De berekening van de kans bij vraag b. van voorbeeld 3.2.1 kunnen we als volgt generaliseren:

Eigenschap 3.2.3 (Regel van Bayes)
Als $\{S_i\}$ een partitie is van S met $P(S_i) > 0$ voor alle i, dan geldt voor iedere gebeurtenis A met $P(A) > 0$ en voor iedere n:

$$\boldsymbol{P(S_n|A) = \frac{P(AS_n)}{P(A)} = \frac{P(A|S_n)P(S_n)}{\sum_i P(A|S_i) \cdot P(S_i)}}$$

Voorbeeld 3.2.4 (oude tentamenopgave)
Volgens wielerkenners gebruikt 10% van alle wielerprofs verboden stimulerende middelen. Met een test wordt het gebruik gecontroleerd en als een wielrenner betrapt wordt ("positief" is), wordt hij geschorst. Uit proeven met de test blijkt dat gebruikers van stimulerende middelen in 85% van de gevallen betrapt worden. De test is echter ook positief bij 5% van de niet-gebruikers.
Bereken op grond van deze gegevens de kans dat iemand die positief is bevonden, ten onrechte wordt geschorst.
Uitwerking:
Definiëren we in de uitkomstenruimte S van alle wielerprofs, G als de gebeurtenis dat een prof verboden stimulerende middelen gebruikt en
B de gebeurtenis dat een prof betrapt wordt (dus positief op de test reageert).
Gegeven is dus $P(G) = 0.1$, $P(B|G) = 0.85$ en $P(B|\overline{G}) = 0.05$.
G en $\overline{G}$ vormen een partitie van S.



Dan volgt de gevraagde voorwaardelijke kans
op een niet-gebruiker, gegeven een positieve test, uit de regel van Bayes:

$$P(\overline{G}|B) = \frac{P(B|\overline{G}) \cdot P(\overline{G})}{P(B|G) \cdot P(G) + P(B|\overline{G}) \cdot P(\overline{G})} = \frac{0.05 \cdot (1 - 0.1)}{0.85 \cdot 0.1 + 0.05 \cdot 0.9} \approx \ 34.6\% \qquad \blacksquare$$

3.3 Onafhankelijkheid en stochastische variabelen

Als ik willekeurig een Nederlander aanwijs, zal het voor de kans dat ik iemand uit Friesland aanwijs, geen verschil maken als ik weet dat ik een vrouw aan zal wijzen. Anders wordt het als ik zou weten dat degene die ik aanwijs, blond haar heeft. Er zijn immers relatief meer blonde Friezen dan blonde Nederlanders. Als het optreden van een gebeurtenis B geen invloed uitoefent op de kans van optreden van een andere gebeurtenis A zullen we zeggen dat A onafhankelijk is van B. Dan zou moeten gelden, dat

$$P(A|B) = P(A)$$

Geldt dan, omgekeerd, ook dat het optreden van gebeurtenis A geen invloed heeft op de kans van optreden van gebeurtenis B? Het antwoord luidt bevestigend, want uit $P(A|B) = P(A)$ volgt:

$$\frac{P(AB)}{P(B)} = P(A), \text{mits } P(B) > 0$$

Dus

$$P(AB) = P(A) \cdot P(B)$$

Als $P(A) > 0$ volgt hieruit dat $P(AB) = P(A)P(B)$ en dus

$$P(B|A) = \frac{P(AB)}{P(A)} = P(B).$$

De gelijkheid $P(AB) = P(A) \cdot P(B)$ is ook gedefinieerd als $P(A) = 0$ of $P(B) = 0$, dus als $P(A|B)$ of $P(B|A)$ niet bestaat. Daarom definiëren we

Definitie 3.3.1 De gebeurtenissen A en B zijn **onderling onafhankelijk** (afgekort **o.o.**) als

$$\boldsymbol{P(AB) = P(A) \cdot P(B)}$$

Als twee gebeurtenissen A en B elkaar uitsluiten, dus $P(AB) = 0$, dan kunnen A en B alleen o.o. zijn als $P(A) = 0$ of $P(B) = 0$.

Voorbeeld 3.3.2 Uit een spel van 52 kaarten trekken we lukraak één kaart.
H is de gebeurtenis dat de getrokken kaart een harten is en B is de gebeurtenis dat de getrokken kaart een boer is. Dan geldt $P(H) = \frac{13}{52} = \frac{1}{4}$ en $P(B) = \frac{4}{52} = \frac{1}{13}$.
Omdat HB de gebeurtenis is dat we de hartenboer trekken, geldt $P(HB) = \frac{1}{52} = P(H)P(B)$.
De gebeurtenissen H en B zijn dus o.o. ■

In het vorige voorbeeld konden we de onafhankelijkheid van twee gebeurtenissen bewijzen. Vaak kennen we de kansen niet en besluiten we op andere gronden tot onafhankelijkheid van twee gebeurtenissen.

Voorbeeld 3.3.3 We werpen tweemaal achtereen met een dobbelsteen. Zij A de gebeurtenis dat we de eerste keer een 5 werpen en B de gebeurtenis dat we de tweede keer een 3 of meer werpen. Als we uitgaan van een zuivere dobbelsteen, geldt $P(A) = \frac{1}{6}$ en $P(B) = \frac{2}{3}$.
Algemeen geldt: $P(AB) = P(A) \cdot P(B|A)$. De voorwaardelijke kans $P(B|A)$ kunnen we echter bepalen door aan te nemen dat het resultaat van de eerste worp niet van invloed is op het resultaat van de tweede worp, d.w.z. we nemen aan dat A en B o.o. zijn. Dan geldt $P(AB) = P(A) \cdot P(B)$. De kans dat we de eerste keer een 5 werpen en de tweede keer 3 of meer is dus

$$P(AB) \overset{\text{o.o.}}{=} P(A)P(B) = \frac{1}{6} \cdot \frac{2}{3} = \frac{1}{9}$$ ■

Merk op dat we in het bovenstaande voorbeeld twee veronderstellingen deden t.a.v. ons kansmodel, namelijk dat de dobbelsteen zuiver is en dat het resultaat van de tweede worp onafhankelijk is van het resultaat van de eerste worp. Deze veronderstellingen samen zijn equivalent met de veronderstelling dat we met een symmetrische kansruimte van doen hebben. Immers zij A_i de gebeurtenis dat we de eerste keer *i* ogen gooien, en B_j de gebeurtenis dat we de tweede keer *j* ogen gooien ($i, j = 1, 2, \dots, 6$), dan geldt op grond van de zuiverheid van de dobbelsteen dat $P(A_i) = P(B_j) = \frac{1}{6}$ en op grond van de onafhankelijkheid dat $P(A_iB_j) \overset{\text{o.o.}}{=} P(A_i)P(B_j) = \frac{1}{36}$. We hadden dus de kans op de gebeurtenis in voorbeeld

3.3.3 ook kunnen bepalen door van de kansdefinitie van Laplace gebruik te maken. De beschreven methode van berekening, die direct gebruik maakt van de onafhankelijkheid, is echter veelal eenvoudiger. In het vervolg zullen we bij soortgelijke experimenten (het herhaalde malen gooien van een munt, het herhaalde malen trekken **met** terugleggen van een knikker uit een vaas) onafhankelijkheid veronderstellen zonder zulks expliciet te vermelden. Als we twee **experimenten onafhankelijk** noemen, dan bedoelen we daarmee, dat elk tweetal gebeurtenissen A en B, waarbij A alléén betrekking heeft op het eerste experiment en B alléén op het tweede, o.o. verondersteld mag worden.

Voorbeeld 3.3.4 Een apparaat bestaat uit twee componenten. A_1 is de gebeurtenis dat de ene component werkt en A_2 de gebeurtenis dat de tweede werkt. Het apparaat werkt slechts als beide componenten werken en we hebben goede redenen om aan te nemen dat het werken van de ene component geen invloed heeft op het werken van de andere, d.w.z. A_1 en A_2 zijn o.o. Onder deze modelveronderstelling geldt voor de kans dat het apparaat werkt:

$$P(A_1A_2) \overset{\text{o.o.}}{=} P(A_1)P(A_2). \qquad \blacksquare$$

We willen de definitie van onafhankelijkheid uitbreiden naar het geval van méér dan twee gebeurtenissen. We noemen de gebeurtenissen $A_1, A_2, \ldots$ **paarsgewijs onafhankelijk** als elk tweetal onafhankelijk is. Paarsgewijze onafhankelijkheid sluit echter niet uit dat er toch een zekere afhankelijkheid bestaat tussen de gebeurtenissen, zoals blijkt uit het volgende voorbeeld.

Voorbeeld 3.3.5 We werpen tweemaal achtereen met een zuivere munt. De uitkomstenruimte is $S = \{KK, KM, MK, MM\}$ en elk van deze uitkomsten treedt met gelijke kans op (omdat we onafhankelijkheid veronderstellen). Zij A de gebeurtenis “de eerste keer kruis”, B de gebeurtenis “de tweede keer kruis” en C de gebeurtenis “beide keren hetzelfde”,
dus $A = \{KK, KM\}$, $B = \{KK, MK\}$ en $C = \{KK, MM\}$.
A, B en C zijn paarsgewijs onafhankelijk, immers
$P(AB) = 1/4 = P(A)P(B)$, $P(AC) = \frac{1}{4} = P(A)P(C)$ en $P(BC) = 1/4 = P(B)P(C)$.
Maar $P(C|AB) = \frac{P(ABC)}{P(AB)} = 1 \neq P(C)$, dus AB geeft “informatie” over C. $\blacksquare$

Om iedere vorm van afhankelijkheid tussen de gebeurtenissen A, B en C uit te sluiten, zullen we dus meer moeten eisen dan paarsgewijze onafhankelijkheid. Er zal bijvoorbeeld ook moeten gelden dat $P(A|BC) = P(A)$ ofwel $P(ABC) = P(A) \cdot P(B) \cdot P(C)$.
Deze laatste eis blijkt op zich weer niet voldoende om paarsgewijze onafhankelijkheid af te dwingen, zoals uit het volgende voorbeeld blijkt.

Voorbeeld 3.3.6 We werpen tweemaal met een zuivere dobbelsteen. Zij A de gebeurtenis dat de eerste worp 1, 2 of 3 ogen en B de gebeurtenis dat de eerste worp 3, 4 of 5 ogen oplevert. C is de gebeurtenis dat de som van het aantal ogen 9 is, dus C bestaat uit 4 uitkomsten: (3,6), (4,5), (5,4) en (6,3). $P(A) = P(B) = \frac{1}{2}$ en $P(C) = \frac{4}{36}$.
De doorsnede van A, B en C bestaat uit één uitkomst, n.l. (3,6). Dus

$$P(ABC) = \frac{1}{36} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{9} = P(A) \cdot P(B) \cdot P(C)$$

terwijl

$$P(AB) = \frac{1}{6} \neq \frac{1}{2} \cdot \frac{1}{2} = P(A) \cdot P(B)$$
$$P(BC) = \frac{3}{36} \neq \frac{1}{2} \cdot \frac{1}{9} = P(B) \cdot P(C)$$
$$P(AC) = \frac{1}{36} \neq \frac{1}{2} \cdot \frac{1}{9} = P(A) \cdot P(C)$$

A, B en C zijn dus niet paarsgewijs onafhankelijk. ■

A, B en C kunnen we dus pas o.o. noemen als voor elk tweetal èn voor het drietal de kans op de doorsnede te schrijven is als het product van de afzonderlijke kansen.
Voor een eindige of aftelbaar oneindige rij gebeurtenissen $A_1, A_2, A_3, \ldots$ zullen we deze eis dus moeten opleggen aan elk tweetal, drietal, viertal, enz.

Definitie 3.3.7 De gebeurtenissen $A_1, A_2, A_3, \ldots$ zijn o.o. als voor elke eindige deelrij $A_{i1}, A_{i2}, \ldots, A_{ik}$ met $k \geq 2$, geldt dat

$$\boldsymbol{P(A_{i1}A_{i2}\ldots A_{ik}) \overset{\text{o.o.}}{=} (A_{i1}) \cdot P(A_{i2}) \cdot \ldots \cdot P(A_{ik})}$$

Als twee gebeurtenissen A en B o.o. zijn, dan zijn ook A en $\overline{B}$ o.o. en $\overline{A}$ en $\overline{B}$ o.o. (zie opgave 8.a). Soortgelijke eigenschappen kan men ook voor meer dan twee gebeurtenissen bewijzen. Als A, B, C en D o.o. zijn dan zijn ook AB en $C \cup D$ o.o., immers

$$\begin{aligned} P(AB(C \cup D)) &= P(ABC \cup ABD) \\ &= P(ABC) + P(ABD) - P(ABCD) \\ &= P(A)P(B)P(C) + P(A)P(B)P(D) - P(A)P(B)P(C)P(D) \\ &= P(A)P(B)[P(C) + P(D) - P(CD)] \\ &= P(AB)P(C \cup D). \end{aligned}$$

Evenzo geldt dat $\overline{A}B$ en CD o.o. zijn, etc.
Als we zeggen dat gebeurtenissen "onafhankelijk" zijn bedoelen we "onderling onafhankelijk". Analoog aan de (onderlinge) onafhankelijkheid van twee experimenten is die van n experimenten gebaseerd op de (onderlinge) onafhankelijkheid van n bijbehorende gebeurtenissen.

Voorbeeld 3.3.8 We werpen tien keer met een zuivere dobbelsteen en noteren het aantal keren dat we 6 werpen; we willen de kans berekenen op de gebeurtenis B_k, dat van de 10 worpen er k een 6 opleveren ($k = 0, 1, \ldots, 10$). Dit experiment bestaat uit tien deelexperimenten met steeds twee mogelijke uitkomsten 6 en niet-6, aangeduid met A en $\overline{A}$. Daar het resultaat van een worp andere worpen niet beïnvloedt, mogen we deze deelexperimenten onafhankelijk veronderstellen. Definiëren we A_i als de gebeurtenis, dat de i-de worp 6 oplevert met $1 \leq i \leq 10$, dan geldt wegens de onafhankelijkheid bijvoorbeeld dat

$$P(A_1A_2\overline{A}_3) \overset{\text{o.o.}}{=} P(A_1) \cdot P(A_2) \cdot P(\overline{A}_3) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{5}{6}$$

B_3 treedt op als bijvoorbeeld de eerste drie worpen 6 en de overige zeven worpen niet-6 zijn. Daarvoor geldt:

$$P(A_1A_2A_3\overline{A}_4 \,..\overline{A}_{10}) \overset{\text{o.o.}}{=} P(A_1)\cdot P(A_2)\cdot P(A_3)\cdot P(\overline{A}_4)\ \cdot \ldots \cdot\ P(\overline{A}_{10}) = \left(\frac{1}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^7$$

Elke volgorde met drie zessen en zeven niet-zessen treedt op met deze kans en daar er $\binom{10}{3}$ van deze volgordes zijn, geldt dat

$$P(B_3) = \binom{10}{3}\left(\frac{1}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^7 \approx 15.5\%.$$

Evenzo: $P(B_k) = \binom{10}{k}\left(\frac{1}{6}\right)^k \cdot \left(\frac{5}{6}\right)^{10-k}$

Zoals we eerder bij de combinatorische kansrekening stochastische variabelen gebruikten om numerieke grootheden te beschrijven en hun kansverdeling te bepalen, kunnen we hier definiëren: $X =$ “het aantal zessen bij 10 worpen”. Omdat $B_k = \{X = k\}$ wordt de kansverdeling van X gegeven door:

$$P(X = k) = \binom{10}{k}\left(\frac{1}{6}\right)^k \cdot \left(\frac{5}{6}\right)^{10-k} \ \textit{waarin}\ k = 0, 1, \ldots, 10$$

X heet dan **binomiaal verdeeld** te zijn, met **parameters** $\boldsymbol{n} = 10$ (aantal pogingen) en $\boldsymbol{p} = \frac{1}{6}$ ■

De gevonden kansformule is een voorbeeld van de binomiale formule, waarmee men in soortgelijke situaties kansen kan berekenen.
Voorbeelden hiervan zijn: het lukraak invullen van *multiple choice* vraagstukken, waarbij men het aantal juiste antwoorden noteert; het bepalen van het aantal getrokken rode balletjes, als we aselect en met terugleggen tien balletjes trekken uit een doos met 7 rode en 13 witte balletjes; het aantal juiste uitslagen als een toto-formulier (met 1, 2 of 3) volledig willekeurig wordt ingevuld.
In al deze gevallen gaat het om het herhalen van een experiment met, in essentie, twee mogelijke uitkomsten: 6 of niet-6, goed of fout, rood of wit. De kans op een bepaalde uitkomst is bij elk experiment gelijk en wordt niet beïnvloed door het resultaat bij andere experimenten. Dit soort experimenten worden Bernoulli-experimenten of -pogingen genoemd.

Definitie 3.3.9 Bij een serie van experimenten spreken we van **Bernoulli-experimenten** (of Bernoulli-pogingen) als

1) elk experiment slechts 2 mogelijke uitkomsten heeft (vaak aangeduid met “succes” en “mislukking”),
2) de experimenten onafhankelijk zijn en
3) de kans op succes voor alle experimenten steeds gelijk is.

De succeskans wordt meestal aangeduid met p en de kans op mislukking met $1 - p$.
We kunnen nu voorbeeld 3.3.8 als volgt generaliseren (het bewijs verloopt analoog aan de afleiding bij voorbeeld 3.3.8).

Eigenschap 3.3.10 (**de binomiale formule**)
Als X het aantal successen is bij n Bernoulli-experimenten met succeskans p, dan is

$$\boldsymbol{P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}}, \ \text{met } k = 0, 1, 2, \ldots, n$$

Onthoud:
"Hierin is $p^k(1-p)^{n-k}$ de kans op (eerst) k successen en (dan) $n-k$ mislukkingen en $\binom{n}{k}$ het aantal volgordes waarin we k successen en $n-k$ mislukkingen kunnen rangschikken."

Een andere vraag, die we bij het werpen met een dobbelsteen kunnen stellen, is:
hoe groot is de kans dat we pas met de tiende worp er in slagen een 6 te werpen?
Een ander soortgelijk voorbeeld: hoe groot is de kans dat pas de tiende passerende auto een Audi is?
Of, algemener gesteld: hoe groot is de kans op de gebeurtenis $\{X=k\}$, dat bij het uitvoeren van Bernoulli-experimenten pas bij het k-de experiment het eerste succes optreedt?
De stochastische variabele X is dus het aantal worpen (Bernoulli-pogingen), dat we nodig hebben om 6 te werpen (een succes te bereiken)
Als we A_i definiëren als het optreden van succes bij het i-de experiment, dan is bijvoorbeeld

$$\{X=4\}=\overline{A}_1\cap\overline{A}_2\cap\overline{A}_3\cap A_4 \quad (=\overline{A}_1\overline{A}_2\overline{A}_3A_4)$$

En algemener: $\{X=k\}=\overline{A}_1\overline{A}_2\ldots\overline{A}_{k-1}A_k$

En dus, wegens de onafhankelijkheid van de experimenten (voor k = 1, 2, 3, . . .):

$$P(X=k)=P\left(\overline{A}_1\overline{A}_2\ldots\overline{A}_{k-1}A_k\right)\overset{\text{o.o.}}{=}P\left(\overline{A}_1\right)P\left(\overline{A}_2\right)\ldots P\left(\overline{A}_{k-1}\right)P(A_k)$$
$$=(1-p)(1-p)\ldots(1-p)p=(1-p)^{k-1}p$$

We hebben hiermee afgeleid:

Eigenschap 3.3.11 (de geometrische formule)
Als we achtereenvolgens Bernoulli-experimenten met succeskans p uitvoeren en $\{X=k\}$ is de gebeurtenis dat het eerste succes optreedt in het k-de experiment, dan is

$$\boldsymbol{P(X=k)=(1-p)^{k-1}p}, \text{voor } k=1,2,3,\ldots$$

X heeft dan een **geometrische verdeling met parameter p** (de succeskans bij elke poging).

3.4 Vraagstukken

1. **a.** Bereken $P(A|B)$ als bekend is dat $P(A\cup B)=0.8$, $P(A\cap B)=0.3$ en $P(A)=0.35$.
b. Bereken $P(ABC)$ als bekend is dat $P(B|AC)=\frac{1}{2}$, $P(C)=\frac{4}{5}$ en $P(A|C)=\frac{3}{4}$.

2. Van een routine-eindtest van een productielijn is bekend dat 98% van de productie goedgekeurd wordt. Of een product echt aan de eisen voldoet weet men pas later. Uit de statistieken blijkt dat van de goedgekeurde producten 97% ook daadwerkelijk aan de eisen voldoet, en dat van de afgekeurde producten toch nog 5 % wel degelijk aan de eisen voldoet. Definieer voor een aselect gekozen product de gebeurtenissen A en B als volgt.
$A=$ "het product wordt goedgekeurd" en $B=$ "het product voldoet aan de eisen".

a. Druk de gegeven kansen uit in A en B en bereken $P(B)$.
b. Bereken de kans dat een product wordt afgekeurd gegeven dat het niet aan de eisen voldoet.
c. Zijn de gebeurtenissen A en B onderling onafhankelijk? Motiveer je antwoord.

3. Een automobilist die een aanrijding veroorzaakt, moet zich onderwerpen aan een bloedproef. De ervaring heeft geleerd dat, als zo iemand “onder invloed” verkeert, er 75% kans is dat het resultaat van de bloedproef positief is. Is hij niet onder invloed dan is er maar 2% kans dat de bloedproef toch positief uitvalt. Aangenomen mag worden dat 5% van de automobilisten die een aanrijding veroorzaken onder invloed is.
Hoe groot is de kans dat iemand die een aanrijding veroorzaakt onder invloed is als de bloedproef positief uitvalt? (Beantwoord de vraag door gebeurtenissen te definiëren, de kansen daarin uit te drukken en de rekenregels voor kansen te gebruiken zie voorbeeld).

4. In het West-Afrikaanse land Gambia zijn 3 mobiele netwerken actief: *Africel*, *Gamcel* en *Comium*. Het marktaandeel van *Africel* is twee maal zo groot als dat van *Gamcel*. Het percentage vaste abonnementen is bij *Africel* 10% (de rest heeft een *prepaid* kaart), bij *Gamcel* 20% en bij *Comium* 30%.Volgens de overheid heeft 15% van alle mobiele telefoongebruikers in Gambia een vast abonnement.
Bereken het marktaandeel van *Comium* (wederom door expliciet de kansregels te gebruiken).

5. Een kast heeft drie laden. In de eerste la liggen twee gouden munten, in de tweede twee zilveren en in de derde één gouden en één zilveren. Blindelings wordt een la gekozen en hieruit wordt willekeurig een munt genomen: deze blijkt van goud te zijn.
Hoe groot is de kans dat ook de andere munt in die la van goud is?

6. In een vaas bevinden zich 5 rode en 7 witte knikkers. We werpen met een zuivere dobbelsteen en pakken daarna zonder terugleggen lukraak zoveel knikkers uit de vaas als we ogen gegooid hebben.

a. Bepaal de voorwaardelijke kans op 3 rode knikkers als we 5 ogen gegooid hebben.
b. Bepaal de kans op precies 3 rode knikkers.

7. Een student maakt een meerkeuzetoets met steeds twee mogelijke antwoorden. Als hij het goede antwoord niet weet, gokt hij door met een zuivere munt te gooien. Op 3/5 van het aantal vragen weet hij het antwoord (ga er vanuit dat het antwoord dan ook goed is).
Wat is de kans dat hij het antwoord wist op een goed beantwoorde vraag?

8. Bewijs:
a. Als A en B o.o. zijn, dan zijn A en $\overline{B}$ o.o. en ook $\overline{A}$ en $\overline{B}$.
b. Als A, B en C o.o. zijn, dan zijn ook A en BC o.o.

9. Gebeurtenissen die elkaar uitsluiten zijn onafhankelijk. Is dit juist of onjuist? Motiveer je antwoord.

10. Bereken de kans dat we meer dan zes worpen nodig hebben om 6 te werpen met een zuivere dobbelsteen.

11. Bij de toto moet men voor 12 wedstrijden voorspellen of het winst, gelijk spel of verlies wordt voor de thuisspelende ploeg. Bereken de kans op minstens 10 goede voorspellingen als iemand het totoformulier volkomen willekeurig invult.

Extra oefenopgave voorwaardelijke kansen:

12. ELISA-tests worden gebruikt om donorbloed te onderzoeken op aanwezigheid van het AIDS-virus. De test detecteert in feite antilichamen, stoffen die door het lichaam worden geproduceerd wanneer het virus aanwezig is. Als er antilichamen zijn, is ELISA positief met kans 0.997 en negatief met kans 0.003. Indien het onderzochte bloed niet met AIDS-antilichamen is besmet, is de kans dat ELISA een positief resultaat geeft gelijk aan 0.015, en de kans dat er een negatief resultaat komt is 0.985. (Omdat ELISA tot doel heeft het AIDS-virus buiten de bloedbanken te houden, is de betrekkelijk grote kans (0.015) op een foutief positief antwoord acceptabel, tegenover de geringe kans (0.003) op het niet herkennen van besmet bloed. Deze kansen zijn afhankelijk van de deskundigheid van het laboratorium dat de tests uitvoert.) Veronderstel dat 1% van een grote populatie de AIDS-antilichamen in het bloed heeft.

a. Bereken de kans dat ELISA voor een aselect gekozen persoon uit deze populatie een positief resultaat laat zien.

b. Bereken de kans dat iemand de antilichamen heeft, gegeven dat ELISA positief was. (Deze opgave illustreert een feit dat belangrijk wordt wanneer men overweegt voorstellen te doen voor grootschalig testen op AIDS of illegale drugs: als het verschijnsel waarop getest wordt in de populatie ongewoon is, zullen de meeste positieve uitkomsten foutief zijn ("vals positieven"), zelfs als de test een zeer grote kans heeft om per geval een correct resultaat te geven.)

Een aantal aanwijzingen voor het oplossen van de vraagstukken:

1. **a.** Gebruik een Venndiagram om snel de kansregels die je kunt gebruiken af te lezen uit het diagram.
b. Bij voorwaardelijke kansen helpt een Venndiagram niet veel: de definitie van voorwaardelijke kans en de daaruit afgeleide productregel gebruiken.

2. Schets het Venndiagram zo dat de delen van de partitie die je gebruikt bekende kansen hebben. De wet van de totale kans kun je eruit af lezen door op de doorsnedes de productregel toe te passen. De regel van Bayes volgt meteen uit de definitie van voorwaardelijke kans (op de teller kun je de productregel toepassen en de noemer uitschrijven met de wet van de totale kans.)

3. Kies handige namen van de gebeurtenissen (bijv. $O =$ "onder invloed") en geef eerst de gegeven en gevraagde kansen m.b.v. die gebeurtenissen.

4. Een puzzel: kijk of je eruit komt door de wet van de totale kans toe te passen op de gebeurtenis $V =$ "klant heeft vast telefoonabonnement"

5. Onderscheid de lades en de soort munt die je als eerste kiest (tweede goud impliceert lade …)

6. **a.** Gebruik de hypergeometrische formule (want het zijn trekkingen zonder terugleggen).
b. Welke gevallen kun je hier onderscheiden?

7. Aanpak als bij opgaven 2 en 3.

8. Schrijf eerst op wat per definitie gegeven is en wat (per definitie) gevraagd wordt. Vervolgens kun je bij a. het Venndiagram gebruiken om de link tussen die twee te leggen.

9. Geef eerst de definities van beide begrippen!

10. Beredeneer hoe je dit met één eenvoudige formule kunt berekenen (kan ook met een som, maar dat is meer werk).

11. Kun je hier onafhankelijkheid van pogingen toepassen, en, zo ja, de geometrische of de binomiale formule?

Hoofdstuk 4 Discrete stochastische variabelen

4.1 Stochastische variabele

In voorgaande hoofstukken hebben we ons beziggehouden met experimenten en de bijbehorende kansruimten. De uitkomst van een experiment is soms een reëel getal, zoals bij het werpen met een dobbelsteen: $S = \{1, 2, 3, 4, 5, 6\}$.
Of het meten van de levensduur van een gloeilamp: $S = [0, \infty)$.
Het al dan niet winnen van de hoofdprijs bij een loterij is echter een experiment met als uitkomst “succes” of “mislukking”: $S = \{s, m\}$. Ook komt het voor dat uitkomsten zijn samengesteld uit meerdere getallen, zoals bij het communicatiekanaal dat codewoorden van 5 nullen of enen verzendt. Als we op een willekeurig moment kijken welk woord verzonden wordt, is de uitkomst een geordend rijtje van 5 nullen of enen.
Ook bij dit soort experimenten zullen we vaak aan elke uitkomst een getalswaarde willen toekennen. Bij de loterij bijvoorbeeld het uit te keren bedrag: 1 miljoen bij succes en 0 bij mislukking. Bij het verzenden van een codewoord tellen we bijvoorbeeld het aantal enen in dat codewoord. Op deze wijze wordt aan elke uitkomst van het experiment een reëel getal toegekend. Er is hier sprake van een functie X die aan elk codewoord een getal (= het aantal enen) toekent. X wordt een stochastische variabele (**toevalsvariabele**) genoemd. Als het experiment wordt uitgevoerd, verkrijgen we door een kans-mechanisme een uitkomst, bijv. 01101, waaraan de stochastische variabele X een functiewaarde toekent: bij 01101 dus 3. Het getal 3 wordt dan de **realisatie** van X genoemd.

Definitie 4.1.1 Als S de uitkomstenruimte van een experiment is, dan is een (reële) functie $X: S \rightarrow \mathbb{R}$, die aan elke uitkomst $s \in S$ een reëel getal $X(s)$ toekent, **een stochastische variabele**.

($\mathbb{R}$ is de verzameling reele getallen). Voor stochastische variabelen gebruiken we steeds hoofdletters: X_1, X_2, X_3 , $Y, Z,$ ... of (bij aantallen) N. De realisaties zijn de getalswaarden zelf.

Voorbeeld 4.1.2 Voor demografisch onderzoek wordt een willekeurige Nederlander gekozen en gevraagd naar zijn leeftijd. Bij dit experiment vormen de Nederlanders een symmetrische kansruimte met $S = \{\text{alle Nederlanders}\}$. We definiëren nu X als “de leeftijd van de gekozen persoon”. De stochastische variabele X krijgt de waarde $X(s)$ voor persoon $s \in S$.
De leeftijden variëren van 0 tot 110 jaar, dus de verzameling van alle realisaties $X(s)$ is $\{0, 1, 2, \ldots, 110\}$: dit is het **waardenbereik** S_X van X.
Aan dezelfde, willekeurig gekozen persoon kunnen we ook zijn “gewicht in kg” of zijn “lengte in cm” vragen, waarmee we stochastische variabele Y resp. Z introduceren, met bijvoorbeeld realisaties $Y(s) = 80$ kg en $Z(s) = 185$ cm voor persoon s.
Bij één experiment kunnen we dus vele stochastische variabelen introduceren, voor elk gewenst **kwantitatief** aspect van dit experiment. ■

Definitie 4.1.3 Het **waardenbereik** S_X van een stochastische variabele X, gedefinieerd op een uitkomstenruimte S is de verzameling van alle mogelijke realisaties $X(s)$.

Dus $S_X = \{X(s)|s \in S\}$.
Het waardenbereik van een stochastische variabele kan

- **eindig** zijn zoals bij X = "het aantal ogen bij een worp met een dobbelsteen": $S_X = \{1, 2, 3, 4, 5, 6\}$,
- **aftelbaar oneindig** (met de natuurlijke getallen te nummeren) zijn, zoals bij Y = "het aantal worpen dat nodig is om 6 te werpen met een dobbelsteen": $S_Y = \{1, 2, 3, 4, \dots\}$ of
- **overaftelbaar** zijn, zoals bij Z = "de levensduur (in uren) van een willekeurige processor": $S_Z = [0, \infty)$.

X en Y zijn voorbeelden van discrete stochastische variabelen, waartoe we ons in dit hoofdstuk zullen beperken. Z is een voorbeeld van een continue variabele (hoofdstuk 6).

Definitie 4.1.4 Een **discrete stochastische variabele** X is een stochastische variabele met een aftelbaar waardenbereik S_X.

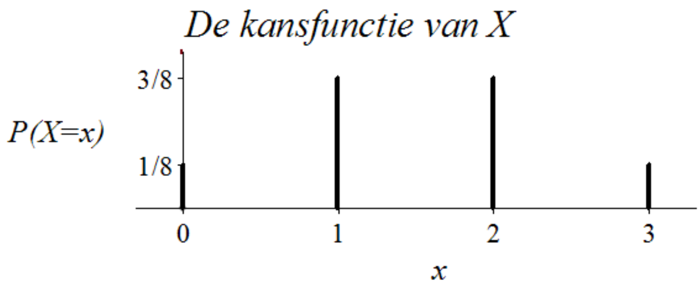
In dat geval is S_X dus van de vorm $\{x_1, x_2, \dots, x_n\}$ of $\{x_1, x_2, x_3, \dots\}$.

4.2 De kansfunctie van een discrete stochastische variabele

Voorbeeld 4.2.1 We werpen driemaal met een zuivere munt en definiëren X als het aantal keren, dat bij die drie worpen kruis boven ligt. Bij dit experiment vormen de $2^3 = 8$ uitkomsten in $S = \{KKK, KKM, KMK, MKK, KMM, MKM, MMK, MMM\}$ een symmetrische kansruimte, dus $P(A) = \frac{N(A)}{8}$ voor elke gebeurtenis A.
$X =$ "het aantal kruizen in drie worpen" kan 0, 1, 2 of 3 zijn, dus $S_X = \{0, 1, 2, 3\}$.
Als we ons bijvoorbeeld afvragen hoe groot de kans is op 1 kruis in drie worpen, vragen we naar de kans op de gebeurtenis $\{X = 1\} = \{s \in S|X(s) = 1\}$. Deze gebeurtenis treedt op bij drie uitkomsten, namelijk $\{X = 1\} = \{KMM, MKM, MMK\}$ dus $P(\{X = 1\}) = \frac{3}{8}$.
Voor een compactere notatie laten we de accolades weg: $P(X = 1) = \frac{3}{8}$
Op dezelfde wijze kunnen we de kansen op X = 0, 2 of 3 bepalen:

$$P(X = 0) = \frac{1}{8}, \;\; P(X = 1) = \frac{3}{8}, P(X = 2) = \frac{3}{8} \;\; \text{en} \;\; P(X = 3) = \frac{1}{8}$$

Dit wordt de kansfunctie $P(X = x)$ van X genoemd, die we grafisch weergeven door op de X-as de waarden x van S_X uit te zetten en op de Y-as de bijbehorende kansen $P(X = x)$.



Omdat de grafiek bestaat uit 4 geïsoleerde punten, trekken we lijnen van de punten naar de X-as, waarvan de lengte gelijk is aan de betreffende kans. De totale lengte van de getrokken lijnen is 1. Dit is niet verwonderlijk want $\{X=0\},\{X=1\},\{X=2\}$ en $\{X=3\}$ vormen een partitie van S, zodat:

$$\sum_{x=0}^{3} P(X=x) = P(X \in \{0,1,2,3\}) = P(S) = 1. \qquad \blacksquare$$

Definitie 4.2.2 Als X een discrete stochastische variabele is, noemen we de functie die aan elke $x \in S_X$ de kans $P(X=x)$ toevoegt, de **kansfunctie** van X.

In voorbeeld 4.2.1 zagen we al dat de som van de kansen $P(X=x)$ gelijk is aan 1.
Dit geldt algemeen:

Eigenschap 4.2.3 Voor de kansfunctie van een discrete stochastische variabele X geldt

1) $\boldsymbol{P(X=x) \geq 0}$ voor $x \in S_X$ en

2) $\boldsymbol{\sum_{x \in S_X} P(X=x) = 1}$

Omgekeerd zullen we een functie die aan 1) en 2) voldoet een kansfunctie noemen.
Kansuitspraken met betrekking tot discrete stochastische variabelen kunnen we nu uitdrukken in de betreffende kansfunctie. In voorbeeld 4.2.1 geldt bijvoorbeeld:

$$P(X>1) = P(X=2) + P(X=3) = \frac{3}{8} + \frac{1}{8} = \frac{1}{2}$$

$$\text{Dus } P\big(X \in (1,\infty)\big) = \frac{1}{2}$$

Algemeen geldt voor elke deelverzameling B van reële getallen, $B \subset \mathbb{R}$:

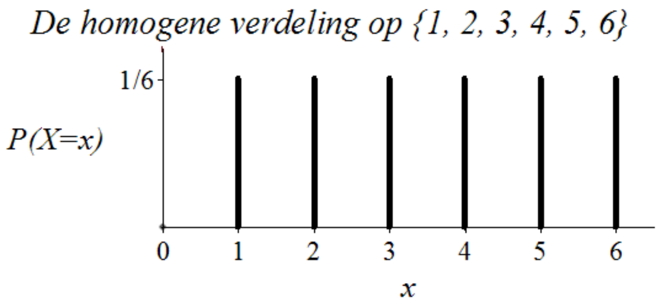
$$P(X \in B) = \sum_{x \in B} P(X=x)\,,$$

waarbij gesommeerd wordt over waarden x uit S_X, dus eigenlijk $x \in B \cap S_X$.
Op deze wijze hebben we op S_X weer een kans P gedefinieerd (aan de axioma's van Kolmogorov is voldaan), zodat $(\boldsymbol{S_X, P})$ weer **een kansruimte** is.
De kansen $P(X \in B)$ voor $B \subset S_X$ worden tezamen wel de **(kans)verdeling** van de stochastische variabele X genoemd. Daar deze kansen berekend kunnen worden uit de kansfunctie van X, kunnen we de kansverdeling ook geven d.m.v. die kansfunctie $P(X=x)$ voor $x \in S_X$.

Voorbeeld 4.2.4 De kansverdeling van X = "aantal ogen bij een worp met een zuivere dobbelsteen" wordt gegeven door:

$$P(X = x) = \frac{1}{6}, \text{voor } x \in S_X = \{1, 2, 3, 4, 5, 6\}.$$

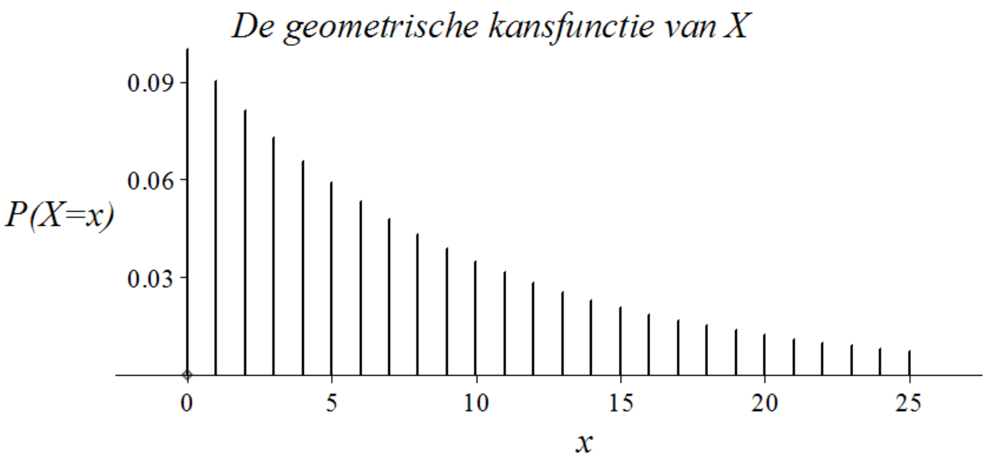
De kansverdeling kan ook grafisch gegeven worden als een staafdiagram van kansen:



We zeggen wel dat X een **homogene** verdeling heeft op $\{1, 2, 3, 4, 5, 6\}$, vanwege de gelijke kansen voor alle mogelijke waarden van X. ■

Voorbeeld 4.2.5 Een handelsreiziger verkoopt aan gemiddeld 1 op de 10 klanten een pannenset. Op een dag bezoekt hij net zo lang klanten totdat hij één pannenset heeft verkocht. Als we nu aannemen dat de klanten onafhankelijk van elkaar met kans $p = \frac{1}{10}$ besluiten tot aankoop, dan is hier sprake van Bernoulli-experimenten.
Zij X het aantal bezochte klanten op die dag, dan wordt de kans op de gebeurtenis $X = k$ dat hij zijn pannenset pas aan de k-de klant weet te slijten, gegeven door de geometrische formule (eigenschap 3.3.11). We zeggen dan ook dat X een **geometrische verdeling** heeft met succeskans $p = \frac{1}{10}$:

$$P(X = k) = \left(\frac{9}{10}\right)^{k-1}\left(\frac{1}{10}\right), \ \ voor\ k \in 1, 2, 3, ...$$



Voor deze kansfunctie kunnen we eigenschap 1 verifiëren.

1) $P(X = k) = \left(\frac{9}{10}\right)^{k-1}\left(\frac{1}{10}\right) > 0$ voor alle $k = 1, 2, 3, ...$

2) $\sum_{k \in S_X} P(X = k) = \sum_{k=1}^{\infty}\left(\frac{9}{10}\right)^{k-1}\left(\frac{1}{10}\right) = \left(\frac{1}{10}\right)\sum_{k=1}^{\infty}\left(\frac{9}{10}\right)^{k-1} = \left(\frac{1}{10}\right) \cdot \frac{1}{1-\frac{9}{10}} = 1$

Ga na dat deze laatste sommatie volgt uit de formule voor een **meetkundige of geometrische reeks:** $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$, voor $|x| < 1$ (zie ook de bijlage Wiskundige Technieken). ■

4.3 De verwachtingswaarde van een discrete stochastische variabele

Voorbeeld 4.3.1 Als we voor een mogelijk onzuivere dobbelsteen willen bepalen hoeveel ogen we **gemiddeld** werpen, kunnen we dit doen door een groot aantal malen met die dobbelsteen te werpen en de resultaten bij te houden in een frequentietabel.
Zij $f_{1000}(x)$ het frequentiequotiënt van de gebeurtenis "x ogen in 1000 worpen" en het resultaat van het uitgevoerde experiment is:

x	1	2	3	4	5	6	totaal
$f_{1000}(x)$	$\frac{180}{1000}$	$\frac{163}{1000}$	$\frac{164}{1000}$	$\frac{161}{1000}$	$\frac{162}{1000}$	$\frac{170}{1000}$	$\frac{1000}{1000}$

Dan wordt het gemiddelde $\overline{x}$ van de aantallen ogen in 1000 worpen gegeven door

$$\overline{x} = \sum_{x=1}^{6} x \cdot f_{1000}(x) = 1 \cdot \frac{180}{1000} + 2 \cdot \frac{163}{1000} + 3 \cdot \frac{164}{1000} + 4 \cdot \frac{161}{1000} + 5 \cdot \frac{162}{1000} + 6 \cdot \frac{170}{1000} = 3.472$$

Het gemiddelde $\overline{x}$ is dus berekend door het **gewogen gemiddelde** van de mogelijke waarden 1, 2, 3, 4, 5, 6 te nemen met als wegingsfactoren de frequentiequotiënten van die waarden. ■

Als we X definiëren als "het aantal ogen bij een worp met een dobbelsteen" en we kennen de kansfunctie $P(X = x)$ voor $x \in S_X = \{1, 2, \ldots, 6\}$, dan kunnen we naar analogie van het herhaald werpen met de dobbelsteen nu het "gemiddelde" van de waarden uit S_X berekenen door deze waarden x te wegen met de kansen $P(X = x)$.
Immers, $f_{1000}(x)$ kunnen we beschouwen als een experimenteel bepaalde schatting van $P(X = x)$. Dit "gemiddelde" m.b.t. de waarden van de stochastische variabele X wordt de verwachtingswaarde $E(X)$ van de stochastische variabele X genoemd:

$$E(X) = \sum_{x=1}^{6} x \cdot P(X = x)$$

Voor een zuivere dobbelsteen geldt: $E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$ ■

Definitie 4.3.2 De **verwachting** of **verwachtingswaarde** $E(X)$ van een discrete stochastische variabele X wordt gegeven door

$$\boldsymbol{E(X) = \sum_{x \in S_X} xP(X = x)}$$

mits deze sommatie absoluut convergent is (d.w.z. $\sum_{x \in S_X} |x| \cdot P(X = x) < \infty$)

Aan de voorwaarde van absolute convergentie is in de praktijk veelal voldaan. In voorbeeld 4.3.4 zullen we een geval zien waarin de sommatie niet absoluut convergeert, zodat in dat

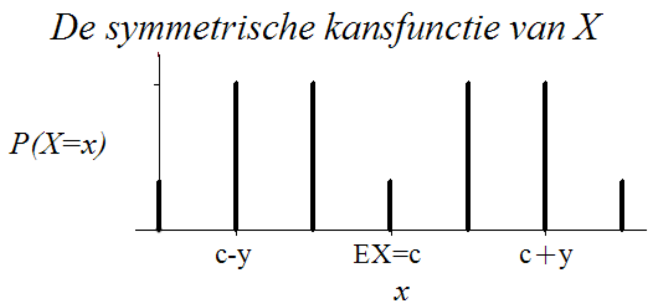
geval de **verwachtingswaarde niet bestaat**.
In plaats van het symbool $E(X)$ (E van het Engelse *Expectation)* wordt ook EX, μ of μ_X gebruikt, en in de natuurkunde $< X >$.
Verder zullen we in plaats van de "som over $x \in S_X$" kortweg x vermelden:

$$E(X) = \sum_x xP(X = x)$$

De verwachtingswaarde of kortweg **verwachting** $E(X)$ kan dus opgevat worden als het gewogen gemiddelde van de mogelijke waarden x van X, met als wegingsfactoren de bijbehorende kansen $P(X = x)$: de som van de wegingsfactoren is uiteraard 1.
Omdat de variabele X met waardenbereik S_X en kansfunctie $P(X = x)$ een kansmodel geven van een populatie, wordt $E(X)$ wel het **populatiegemiddelde** genoemd, en daarom ook aangeduid met μ (de Griekse letter m voor *mean*, gemiddelde). Dit is de tegenpool van wat we in de statistiek het **steekproefgemiddelde** $\overline{\boldsymbol{x}}$ noemen, het gemiddelde van de waarden die X aanneemt bij een beperkte steekproef uit de hele populatie.
Frequentie-interpretatie van het populatiegemiddelde $E(X)$: als een experiment met een numerieke variabele zeer vaak wordt herhaald (onder gelijke omstandigheden) ligt de gemiddelde waargenomen waarde dicht in de buurt van het $E(X) = \mu$.

$E(X)$ hoeft dus niet "de middelste waarde van het waardenbereik" te zijn, maar kan wel als een **maat voor het midden (centrum)** van de verdeling worden gezien.
Een meer natuurkundige interpretatie van $E(X)$ is die van een "**evenwichtspunt**": als we de X-as als (gewichtsloze) balk opvatten en de kansen $P(X = x)$ als gewichten die aangrijpen op de punten x op de X-as, dan is $E(X)$ juist het punt waar de balk ondersteund moet worden om hem in evenwicht te houden.

Als de kansfunctie symmetrisch is t.o.v. een punt c, zoals in onderstaande grafiek, kan men beredeneren dat $E(X) = c$.



In de uitdrukking $E(X) = \sum_{x \in S_X} xP(X = x)$ komen namelijk steeds de waarden $c - y$ en $c + y$ voor met gelijke kans, dus het "gemiddelde" is $E(X) = c$. In voorbeeld 4.3.1 zagen we dat de symmetrisch verdeelde stochastische variabele X inderdaad het symmetriepunt 3.5 als verwachtingswaarde had.

Voorbeeld 4.3.3 Het aantal bezochte klanten, X, van de handelsreiziger in voorbeeld 4.2.5 was geometrisch verdeeld met succeskans $p = \frac{1}{10}$.
Het verwachte aantal klanten, die hij moet bezoeken om één pannenset te verkopen wordt dus gegeven door:

$$EX = \sum_{k \in S_X} kP(X = k) = \sum_{k=1}^{\infty} k \cdot \left(\frac{9}{10}\right)^{k-1} \frac{1}{10} = \frac{1}{10} \sum_{k=1}^{\infty} k \cdot \left(\frac{9}{10}\right)^{k-1} = \frac{1}{10} \cdot \left(\frac{1}{1 - \frac{9}{10}}\right)^2 = 10$$

De gelijkheid volgt uit de sommatie van een meetkundige reeks met differentiëren (zie bijlage Wiskundige technieken). $E(X)$ wordt nogal eens verward met de **mediaan**: dat is de waarde M van X, zodanig dat $P(X \leq M) \geq 50\%$ en $P(X \geq M) \geq 50\%$.
Na wat rekenwerk vinden we voor dit voorbeeld dat het mediane aantal klanten 7 bedraagt, terwijl $E(X) = 10$. Zie ook de betreffende opgave 16. ■

Voorbeeld 4.3.4 A en B werpen om de beurt met een zuivere munt. Wie het eerst kruis gooit, is winnaar. Beiden zetten één gulden in en A begint. Steeds als er geen kruis wordt gegooid, wordt de inzet verdubbeld. Voor A lijkt dit aantrekkelijk: al bij de eerste worp is de kans 50% dat hij wint, dus $P(\text{"}A \text{ wint"}) > \frac{1}{2}$.
Als we X definiëren als de winst van A, dan is de winst 1 als hij zelf meteen kruis gooit, -2 als B na de eerste keer munt bij zijn poging kruis gooit, etc.:

$$S_X = \{1, -2, 4, -8, \ldots\} = \{(-2)^n | n = 0, 1, 2, \ldots\}$$

En omdat de kans dat er k worpen nodig zijn om voor het eerst kruis te werpen $\left(\frac{1}{2}\right)^k$ is, geldt:

$$P[X = (-2)^n] = \left(\frac{1}{2}\right)^{n+1}, \text{voor } n = 0, 1, 2, \ldots$$

Dus de "verwachte winst van A" berekenen we als volgt:

$$\sum_{n=0}^{\infty} (-2)^n \cdot \left(\frac{1}{2}\right)^{n+1} = \frac{1}{2} - \frac{1}{2} + \frac{1}{2} - \frac{1}{2} + \cdots,$$

Dus een altererende reeks, die niet convergeert. Deze is ook niet absoluut convergent:
$\sum_{n=0}^{\infty} 2^n \cdot \left(\frac{1}{2}\right)^{n+1} = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots = \infty$. $E(X)$ bestaat dus niet: X is onbegrensd groot ■

4.4 Functies van een discrete stochastische variabele; variantie

Voorbeeld 4.4.1 Iemand daagt mij uit om 4 Euro in te zetten bij het volgende spel: als ik X ogen met een zuivere dobbelsteen werp krijg ik $(X - 3)^2$ Euro uitbetaald.
Gooi ik bijvoorbeeld 6, dan is mijn winst de uitbetaling minus de inzet: $(6 - 3)^2 - 4 = 5$ Euro. En gooi ik 3, dan is de winst $(3 - 3)^2 - 4 = -4$ gulden (4 Euro verlies).
Ik besluit mee te doen als mijn verwachte winst positief is. Dus als de verwachtingswaarde van $Y = (X - 3)^2 - 4$ positief is. Y is een functie van X en zelf ook weer een stochastische variabele: Y kan de waarden 0, −3, −4, −3, 0, 5 aannemen als het aantal ogen X respectievelijk gelijk is aan 1, 2, 3, 4, 5 resp. 6: $S_Y = \{-4, -3, 0, 5\}$ en de kansverdeling is:

$P(Y = 0) = P(X = 1) + P(X = 5) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$

Evenzo $P(Y = -3) = \frac{1}{3}$

en $P(Y = -4) = P(Y = 5) = \frac{1}{6}$

Voor de verwachte winst vinden we

$$E(Y) = \sum_{y \in S_Y} yP(Y = y)$$
$$= (-4) \cdot \frac{1}{6} + (-3) \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 5 \cdot \frac{1}{6} = -\frac{5}{6} \text{ Euro,}$$

een negatieve winstverwachting, die volgens de frequentie-interpretatie betekent dat “bij een groot aantal herhalingen van het gokspel het gemiddelde verlies dicht bij $\frac{5}{6}$ Euro per spel zal zijn.”
We hadden de verwachte winst ook direct met de kansverdeling van X kunnen berekenen door, voor elke waarde van x, de winst $(x-3)^2 - 4$ te wegen met de kans dat $X = x$, dus:

$$E[(X-3)^2 - 4] = \sum_{x \in S_X} [(x-3)^2 - 4] \cdot P(X = x)$$
$$= (-4) \cdot \frac{1}{6} + (-3) \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 5 \cdot \frac{1}{6} = -\frac{5}{6} \text{ Euro} \qquad \blacksquare$$

In voorgaande is $Y = (X-3)^2 - 4$ een voorbeeld van een functie $Y = g(X)$ van een discrete stochastische variabele X. In voorbeeld 4.4.1 zagen we hoe we vrij eenvoudig de kansfunctie van Y af kunnen leiden uit de gegeven kansfunctie van X, door naar de corresponderende waarden en kansen van X en Y te kijken. Vervolgens kunnen dan $E(Y)$ bepalen met de verdeling van Y. Maar we konden $E(Y) = E(g(X))$ ook met de kansfunctie van X bepalen, zoals algemeen in de volgende eigenschap (zonder bewijs) is verwoord.

Eigenschap 4.4.2 Voor een discrete stochastische variabele X en een (reële) functie g geldt:

$$\boldsymbol{E\big(g(X)\big) = \sum_{x \in S_X} g(x)P(X = x)}$$

(mits deze sommatie absoluut convergent is).

Met $Eg(X)$ bedoelen we $E(g(X))$, bijvoorbeeld: EX^2 betekent $E(X^2)$ en **niet** $(EX)^2$.

Als ***Y* een lineaire functie van *X*** is, dus $Y = aX + b$ met reële constanten $a, b \in \mathbb{R}$, dan geldt volgens eigenschap 4.4.2

$$E(aX + b) = \sum_{x \in S_X} (ax + b) \cdot P(X = x)$$
$$= \sum_{x \in S_X} ax \cdot P(X = x) + \sum_{x \in S_X} b \cdot P(X = x)$$
$$= a \cdot \sum_{x \in S_X} x \cdot P(X = x) + b \cdot \sum_{x \in S_X} P(X = x)$$
$$= a \cdot E(X) + b \cdot 1$$

Hiermee hebben we het eerste deel bewezen van de volgende eigenschap:

Eigenschap 4.4.3 Voor een discrete stochastische variabele X en reële functies g en h en voor reële constanten $a, b \in \mathbb{R}$ geldt:

$$\mathbf{1)}\ \boldsymbol{E(aX + b) = aE(X) + b}$$
$$\mathbf{2)}\ \boldsymbol{E[ag(X) + bh(X)] = aEg(X) + bEh(X)}.$$

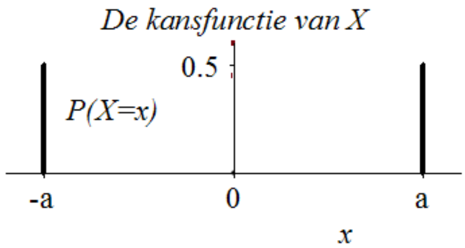
Het bewijs van 2) verloopt analoog aan dat van 1).
Blijkbaar geldt voor voorbeeld 4.4.1 volgens deze eigenschap:
$E(Y) = E((X-3)^2 - 4) = E(X^2 - 6X + 5) = E(X^2) - 6E(X) + 5.$
Met (wegens symmetrie) $E(X) = 3.5$ en $E(X^2) = (1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) \cdot \frac{1}{6} = \frac{91}{6}$
Zo vinden we inderdaad weer $E(Y) = \frac{91}{6} - 6 \cdot \frac{7}{2} + 5 = -\frac{5}{6}$
Het komt vaak voor in de kansrekening dat we van functies van de vorm $g(X) = X^k$ de verwachtingswaarde willen berekenen. Met eigenschap 4.4.2 is dat mogelijk:

$$\boldsymbol{E(X^k) = \sum_x x^k P(X = x)}$$

Definitie 4.4.4 $\boldsymbol{E(X^k)}$ is het ***k*-de moment** van de stochastische variabele X ($k = 1, 2, 3, ...$).

Uiteraard is het k-de moment alléén gedefinieerd als de bijbehorende sommatie absoluut convergent is. Omdat hieraan in de praktijk bijna altijd voldaan is, zullen we in het vervolg deze voorwaarde niet steeds controleren.
Het eerste moment $E(X^1)$ kennen we als de verwachtingswaarde $E(X) = \mu_X$ van een stochastische variabele X. Dit gewogen gemiddelde noemden we al een **maat voor het midden** van de verdeling van X (met de mediaan als alternatieve maat voor het midden).
$E(X)$ zegt echter niets over de **spreiding** in de waarden van X, zoals ook uit het volgende voorbeeld blijkt.

Voorbeeld 4.4.5 We beschouwen de verdeling $P(X = a) = P(X = -a) = 1$ met $a > 0$

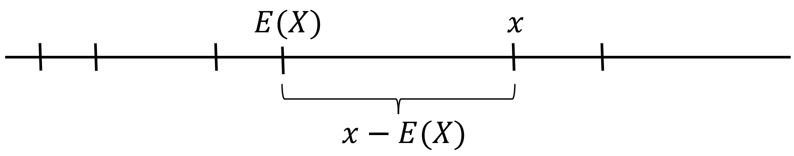


Op grond van symmetrie is meteen duidelijk dat $E(X) = 0$, ongeacht de waarde van a.
De spreiding van X, in dit geval de mate waarin de twee punten van S_X afwijken van 0, wordt groter naarmate a groter wordt. Voor de momenten van X geldt:

$$E(X^k) = a^k \cdot \frac{1}{2} + (-a)^k \cdot \frac{1}{2} = \begin{cases} a^k & \text{voor even } k \\ 0 & \text{voor oneven } k \end{cases}$$

De "gemiddelde gekwadrateerde waarden" van X, dus het tweede moment $E(X^2) = a^2$, wordt blijkbaar groter naarmate a groter wordt, evenals alle andere even momenten. ■

Op zoek naar een maat voor de spreiding bekijken we eerst $X - E(X)$: "de afwijking van X t.o.v. $E(X)$". Dan is $E(X - E(X))$ het gewogen gemiddelde van die afwijkingen.
Als $X = x$, dan krijgt de afwijking $x - E(X)$ de wegingsfactor $P(X = x)$:



Volgens eigenschap 4.4.3 geldt: $E[X - E(X)] = E(X) - E(X) = 0$. Dit is niet verwonderlijk daar de afwijkingen $x - E(X)$ zowel positief als negatief kunnen zijn: $E(X)$ is juist zó gedefinieerd dat de "gewogen afwijkingen" $[x - E(X)] \cdot P(X = x)$ gesommeerd nul zijn.
Het ligt voor de hand om nu het gewogen gemiddelde van de absolute waarde der afwijkingen als **maat voor de spreiding** te kiezen: $E\{|X - E(X)|\}$, of als alternatief $E\left[\left(X - E(X)\right)^2\right]$, de gemiddelde gekwadrateerde afwijkingen.
Doorgaans wordt voor deze laatste gekozen, vanwege zijn prettige wiskundige eigenschappen. De haken […] worden gemakshalve weggelaten en $E(X)$ ofwel EX vaak vervangen door μ_X.

Definitie 4.4.6 De **variantie** van X (notatie: $\boldsymbol{var(X)}$ of $\boldsymbol{\sigma_X^2}$) wordt gedefinieerd door

$$\boldsymbol{var(X) = E(X - \mu_X)^2}$$

Volgens eigenschap 4.4.2 kunnen we $var(X)$ dus als volgt berekenen:

$$var(X) = \sum_x (x - \mu_X)^2 \cdot P(X = x)$$

We kunnen echter ook met eigenschap 4.4.3 $var(X)$ uitdrukken in het eerste en tweede moment:

$$\begin{aligned} var(X) &= E(X - \mu_X)^2 = E(X^2 - 2\mu_X \cdot X + \mu_X^2) = E(X^2) - 2\mu_X \cdot E(X) + \mu_X^2 \\ &= \boldsymbol{E(X^2) - \mu_X^2} \end{aligned}$$

Merk op dat μ_X een getal is, en dat dus $E(\mu_X) = \mu_X$ ofwel $E\big(E(X)\big) = E(X)$.
Evenzo geldt: $E(\mu_X^2) = \mu_X^2$

Veelal zullen we deze rekenformule voor $var(X) =$ "het 2de moment minus het kwadraat van het 1ste moment" prefereren boven directe berekening uit de definitie. Deze en andere eigenschappen van verwachting en variantie dient men paraat te hebben om deze vlot te kunnen toepassen, zoals ook blijkt uit het volgende voorbeeld.

Voorbeeld 4.4.7 Zij X het aantal zessen bij 2 worpen met een zuivere dobbelsteen, dan geldt:

$$P(X = 2) = \frac{1}{36}, P(X = 1) = \frac{10}{36} \;\; en \; P(X = 0) = \frac{25}{36}$$

De berekening van $E(X) = \sum_x xP(X = x)$ en $var(X) = E(X^2) - \mu_X^2$ voeren we nu uit met de volgende overzichtelijke tabel:

x	0	1	2	Totaal
$P(X = x)$	$\frac{25}{36}$	$\frac{10}{36}$	$\frac{1}{36}$	1
$x \cdot P(X = x)$	$0 \cdot \frac{25}{36}$	$1 \cdot \frac{10}{36}$	$2 \cdot \frac{1}{36}$	$\frac{1}{3} = E(X) = \mu$
$x^2 \cdot P(X = x)$	$0 \cdot \frac{25}{36}$	$1 \cdot \frac{10}{36}$	$4 \cdot \frac{1}{36}$	$\frac{14}{36} = E(X^2)$
$(x - \mu_X)^2 \cdot P(X = x)$	$\left(\frac{1}{3}\right)^2 \cdot \frac{25}{36}$	$\left(\frac{2}{3}\right)^2 \cdot \frac{10}{36}$	$\left(\frac{5}{3}\right)^2 \cdot \frac{1}{36}$	$\frac{10}{36} = var(X)$

Dus:
$var(X) = E(X^2) - \mu_X^2$
$= \frac{14}{36} - \left(\frac{1}{3}\right)^2 = \frac{10}{36}$

Ter vergelijking hebben we in de laatste rij de directe berekening van de variantie met de definitie uitgevoerd.

We zien dat in dit voorbeeld dat $\frac{14}{36} = E(X^2) \neq (EX)^2 = \left(\frac{1}{3}\right)^2$: in het algemeen geldt dat het gemiddelde van de kwadraten van getallen groter is dan het kwadraat van het gemiddelde! ■

Vanwege het kwadraat in de definitie kan $var(X) = E(X^2) - \mu_X^2$ niet negatief kan zijn, dus $E(X^2) \geq \mu_X^2$
De gelijkheid geldt als $var(X) = 0$, ofwel als alle termen in $\sum_x (x - \mu_X)^2 \cdot P(X = x)$ 0 zijn.
Dus als $P(X = x) > 0$, moet $x = \mu_X$ zijn: X kan dan dus slechts één waarde, μ_X, aannemen: $P(X = \mu_X) = 1$.
We spreken dan van een **ontaarde** verdeling van X is (in μ_X). Het stochastische experiment leidt dan tot één en dezelfde waarde voor X.
Daar $var(X)$ een gemiddelde is van gekwadrateerde afwijkingen $(x - \mu_X)^2$, is de eenheid van $var(X)$ het kwadraat van de eenheid van X: als X in cm is, dan is $var(X)$ in cm^2.
Om terug te gaan naar de oorspronkelijke eenheid trekken we de wortel uit de variantie
(Let op: $\sqrt{E(X - \mu_X)^2} \neq E|X - \mu_X|$).

Definitie 4.4.8 De **standaardafwijking** van X (notatie: σ_X) is de wortel uit de variantie:

$$\boldsymbol{\sigma_X = \sqrt{var(X)}}$$

$var(X) = \sigma_X^2$ en $\sigma_X = \sqrt{var(X)}$ zijn dus uitwisselbare maten voor de spreiding.

Eigenschap 4.4.9 (Eigenschappen van variantie en standaardafwijking)

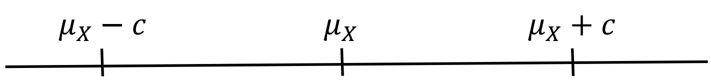
- **a.** $\boldsymbol{var(X) \geq 0}$ en $\sigma_X \geq 0$.
- **b.** $\boldsymbol{var(X) = E(X^2) - \mu_X^2}$ (de rekenformule).
- **c.** als $var(X) > 0$, d.w.z. X is niet ontaard, geldt $E(X^2) > (EX)^2$.
- **d.** $\boldsymbol{var(aX + b) = a^2 \cdot var(X)}$ en $\sigma_{aX+b} = |a| \cdot \sigma_X$.

Bewijs: a. t/m c. volgen uit het voorgaande
d. houdt in dat bij een lineaire transformatie van X een verschuiving ($+b$) de maten voor de spreiding niet beïnvloedt:

$$\begin{aligned} var(aX + b) &= E[aX + b - E(aX + b)]^2 \\ &= E[aX + b - aE(X) - b]^2 \\ &= a^2 \cdot E(X - \mu_X)^2 \\ &= a^2 \cdot var(X) \end{aligned}$$

En dus $\sigma_{aX+b} = \sqrt{var(aX + b)} = \sqrt{a^2 \cdot var(X)} = |a| \cdot \sigma_X$ ■

Met de maten voor het midden en de spreiding, μ_X en σ_X^2 (of σ_X), kunnen we uitspraken doen over de kans dat X waarden aanneemt in een interval met μ_X als midden.



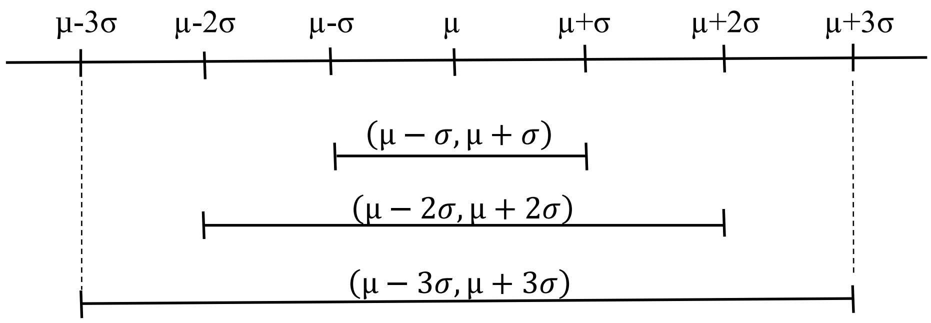
Eigenschap 4.4.10 (**de ongelijkheid van Chebyshev**)

$$\text{Voor elke } c > 0 \text{ geldt: } \boldsymbol{P(|X - \mu_X| \geq c) \leq \frac{var(X)}{c^2}}$$

We zullen dit theoretische resultaat niet bewijzen, maar wel op de interpretatie ten aanzien van de standaardafwijking ingaan. De ongelijkheid geldt voor **elke** verdeling en geeft een **bovengrens** voor de kans dat de variabele X meer dan c afwijkt van μ_X, dus een waarde aanneemt buiten het interval $(\mu_X - c, \mu_X + c)$.
De link met de standaardafwijking wordt duidelijk als we voor het getal $c = k \cdot \sigma_X$ kiezen. Dan is het interval dus $(\mu_X - k \cdot \sigma_X, \mu_X + k \cdot \sigma_X)$ en de bovengrens voor de kans een waarde buiten het interval voor X waar te nemen $\frac{var(X)}{c^2} = \frac{var(X)}{k^2\sigma_X^2} = \frac{1}{k^2}$.
Men kan nagaan dat voor $c < \sigma_X$ de bovengrens voor de kans groter dan 1 wordt (en dus geen informatie geeft).
Kiezen we bijv. $k = 2$, dan geldt $P(|X - \mu_X| \geq 2\sigma_X) \leq \frac{1}{2^2} = 25\%$

En voor $k = 3$, dan geldt $P(|X - \mu_X| \geq 3\sigma_X) \leq \frac{1}{3^2} \approx 11\%$

Het laatste betekent dus dat de kans dat X met een kans van maximaal 11% waarden aanneemt buiten $(\mu_X - 3 \cdot \sigma_X, \mu_X + 3 \cdot \sigma_X)$ en met kans van minimaal 89% erbinnen.



De regel van Chebyshev geldt voor elke verdeling, maar de zo geheten **Empirische Regel** geldt alleen voor symmetrische, klokvormige (of heuvelvormige) kansverdelingen.
Deze luidt:

Indien de kansverdeling van X een klokvorm heeft, dan geldt bij benadering:

- De kans dat X een waarde heeft in het interval $(\mu - \sigma, \mu + \sigma)$ is ongeveer **68%.**
- De kans dat X een waarde heeft in het interval $(\mu - 2\sigma, \mu + 2\sigma)$ is ongeveer **95%.**
- De kans dat X een waarde heeft in het interval $(\mu - 3\sigma, \mu + 3\sigma)$ is ongeveer **99.7%.**

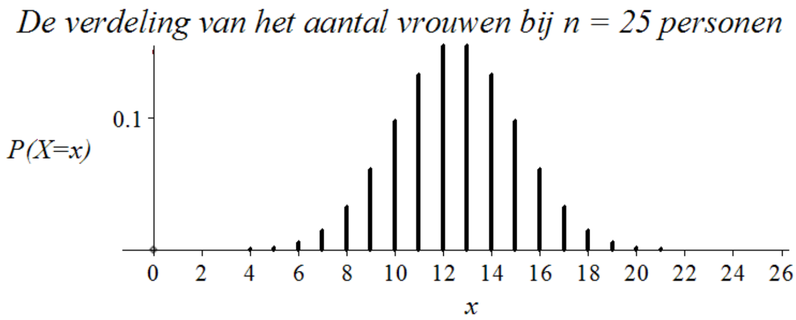
Deze regel wordt ook wel de 68-95-99.7%-regel genoemd en is gebaseerd op de klokvormige verdeling bij uitstek, de normale verdeling, die we in hoofdstuk 6 bespreken.

Voorbeeld 4.4.11 In Enschede is 50% van de volwassenen man. We kiezen er 25 willekeurig uit voor een onderzoek. In de praktijk zal dat zonder terugleggen gebeuren, maar gezien de omvang van de populatie is dat nagenoeg hetzelfde als trekken met terugleggen. Dan is de verdeling van X, het aantal vrouwen in de steekproef, te geven met de binomiale formule met

$n = 25$ en
succeskans $p = \frac{1}{2}$:

$$P(X = k) = \binom{25}{k} \left(\frac{1}{2}\right)^{25}$$

met $k = 0, 1, 2, 3, \ldots$



De kansverdeling van X ziet er inderdaad klokvormig/heuvelvormig uit, dus de Empirische regel geldt.
Om dit te controleren hebben we $\mu = E(X)$ en $\sigma^2 = var(X)$ nodig:
$E(X) = \frac{1}{2} \cdot 25 = 12.5$ (intuïtief: we verwachten dat de helft van de steekproef vrouw is).
$var(X) = np(1-p) = 6.25$ (deze formule bespreken we in de volgende paragraaf).
We vergelijken nu voor de drie intervallen, de werkelijke kans met de kansen volgens de empirische regel en de ongelijkheid van Chebyshev:

Interval	**$P(X$ in interval$)$**	Kans volgens Empirische regel	Kans volgens Chebyshevs regel
$(\mu - \sigma, \mu + \sigma) = (10, 15)$	**57.6%**	68%	$\geq 0\%$
$(\mu - 2\sigma, \mu + 2\sigma) = (7.5, 17,5)$	**95.7%**	95%	$\geq 75\%$
$(\mu - 3\sigma, \mu + 3\sigma) = (5, 20)$	**99.6%**	99.7%	$\geq 89\%$

De werkelijke kansen liggen veelal dicht bij de waarden volgens de Empirische regel en zijn veel groter dan de ondergrens volgens Chebyshev. ■

4.5 De binomiale, hypergeometrische, geometrische en Poisson-verdeling

De Binomiale verdeling

In de kansrekening en de statistiek spelen **Bernoulli-experimenten** (ook wel **Bernoulli-pogingen** genoemd) een grote rol: het al dan niet optreden van een bepaald fenomeen wordt gekarakteriseerd als een succes (= 1) of een mislukking (= 0). We noemen dit alléén Bernoulli-experimenten als ze o.o. zijn. Bij een lange reeks Bernoulli-experimenten kunnen we de kans op succes schatten, door de fractie van het aantal succesvolle experimenten te bepalen. De verdeling van het aantal successen bij n Bernoulli-experimenten is dus van belang.

Voorbeeld 4.5.1 Uit een populatie trekken we n keer met terugleggen een kiezer en vragen die persoon of hij op partij A stemt. Zij X het aantal A-stemmers onder de n gekozenen, dan is de gebeurtenis $\{X = k\}$ juist de gebeurtenis k successen optreden bij n Bernoulli-experimenten met succeskans p = “de kans dat een willekeurig persoon partij A stemt”.

Dus volgens eigenschap 3.3.10 geldt:

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}, \text{met } k \in S_X=\{0,1,2,\dots,n\}$$

Aantal volgordes van k successen en $n-k$ mislukkingen"

kans op "eerst k successen en dan $n-k$ mislukkingen"

Dit is inderdaad een kansverdeling, want

1) $P(X = k) \geq 0$ voor $k=0,1,\dots,n$ (en natuurlijk $0 \leq p \leq 1$)
2) $\sum_{k\in S_X} P(X=k)=\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}=(p+(1-p))^n=1$
   wegens het Binomium van Newton: $(a+b)^n=\sum_{k=0}^{n}\binom{n}{k}a^kb^{n-k}$ (zie ook bijlage) ∎

Definitie 4.5.2 (**de binomiale verdeling**)

We noemen X **binomiaal verdeeld met parameters n en p**, met $n=1,2,\dots$ en $p\in[0,1]$, als de kansfunctie van X gegeven wordt door:

$$\boldsymbol{P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}}, \quad \text{voor } k=0,1,2,\dots,n$$

Korte schrijfwijze: X is $B(n,p)$-verdeeld of: $\boldsymbol{X \sim B(n,p)}$

De binomiale verdeling geeft een goede beschrijving van de werkelijkheid als er sprake is van een serie van n Bernoulli-pogingen, dus:

- al dan niet optreden van een fenomeen met steeds dezelfde kans (p resp. $1-p$) en
- onafhankelijkheid van de pogingen

De onafhankelijkheid is gewaarborgd in situaties waarin sprake is van **aselecte trekkingen met terugleggen** uit een populatie, zoals de aselecte keuze van kiezers in voorbeeld 4.5.1, kiezers of als het gaat om herhalingen van experimenten met twee mogelijke uitkomsten onder gelijke omstandigheden: bij het drie keer werpen met een zuivere munt in voorbeeld 4.2.1 is het aantal keren kruis dus $B\left(3,\frac{1}{2}\right)$-verdeeld. Het aantal zessen bij 2 worpen met een dobbelsteen in voorbeeld 4.4.7 is binomiaal verdeeld met parameters $n=2$ en $p=\frac{1}{6}$.
We vonden $E(X)=\frac{1}{3}=n\cdot p$ en $var(X)=\frac{10}{36}=np(1-p)$
Het verwachte aantal zessen in bijvoorbeeld 60 worpen is intuïtief gelijk aan één zesde deel daarvan, dus $60\cdot\frac{1}{6}=10$. Dit is dan ook het verwachte aantal.

Als X $B(n,p)$-verdeeld is, geldt inderdaad altijd $\boldsymbol{E(X)=np}$.
Verder geldt: $\boldsymbol{var(X)=np(1-p)}$.

Deze formules kunnen we afleiden uit de definities van EX en $var(X)$.
Zo geldt:

$$E(X)=\sum_{k\in S_X}k\cdot P(X=k)=\sum_{k=0}^{n}k\cdot\binom{n}{k}p^k(1-p)^{n-k}=\dots=np$$

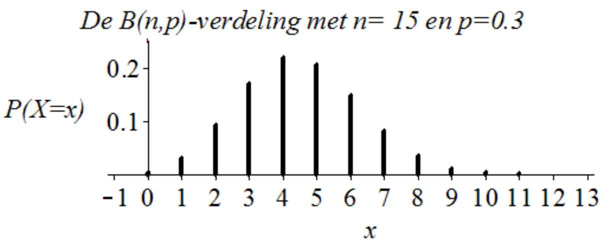
Deze afleiding vergt enig nauwkeurig analytisch werk om wederom van het binomium van Newton gebruik te kunnen maken. In hoofdstuk 5 zullen we echter van een meer inzichtelijke aanpak gebruikmaken om de formules voor $E(X)$ en $var(X)$ aan te tonen.

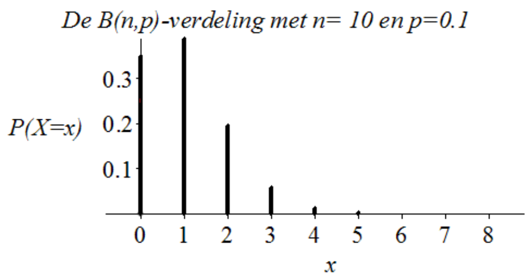
Een paar bijzondere gevallen bij de keuze van n en p van de binomiale verdeling:

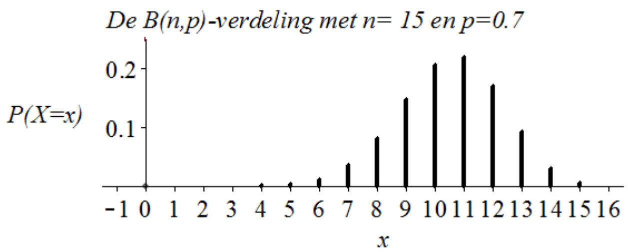
- Als $\boldsymbol{p = 1}$ ("altijd succes"), dan is $P(X = n) = 1$ en $E(X) = n$: X heeft een **ontaarde verdeling** in n. Analoog geldt voor $p = 0$, dat $P(X = 0) = 1$ en $E(X) = 0$.

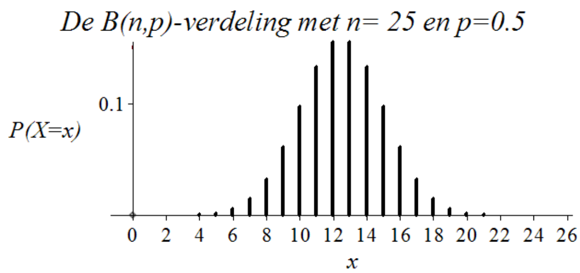
- Als er slechts één poging wordt uitgevoerd (één schot op goal, één product op kwaliteit beoordeeld, etc.), spreekt men wel van een **alternatieve verdeling met succeskans $\boldsymbol{p}$,** een $B(1, p)$-verdeling dus, waarvoor geldt:
  $P(X = 1) = p$ en $P(X = 0) = 1 - p$, dus
  $$E(X) = \sum_x xP(X = x) = 1 \cdot p + 0 \cdot (1 - p) = p$$
  Ook $$E(X^2) = \sum_x x^2 P(X = x) = 1^2 \cdot p + 0^2 \cdot (1 - p) = p$$
  Zodat $var(X) = E(X^2) - (EX)^2 = p(1 - p)$, de variantie van de $B(1, p)$-verdeling.

Hieronder geven we nog enkele grafieken voor verschillende waarden van *n* en *p*:









Om het berekenen van kansen m.b.t. de binomiale verdeling te vergemakkelijken zijn voor verschillende waarden van n en p de zgn. **cumulatieve binomiale tabellen** gegeven.
Voor, bijvoorbeeld, $n = 15$ en $p = 0.3$ leiden we daaruit af:

- $P(X \leq 5) = 0.7216$
- $P(X = 5) = P(X \leq 5) - P(X \leq 4) = 0.7216 - 0.5155 \ = \ 0.2061$
  Controle van de uitkomst met de binomiale kansformule: $P(X = 5) = \binom{15}{5} 0.3^5 0.7^{10}$
- $P(X > 5) = 1 - P(X \leq 5) = 1 - 0.7216 = 0.2784$.
- Soms (zoals bij de tabellen achter in het dictaat) hebben we slechts de beschikking over kanstabellen met succeskans $p \leq 0.5$. Voor $\boldsymbol{p > 0.5}$ kunnen we dan toch kansen berekenen door het bijbehorend aantal mislukkingen te beschouwen.
  Als Y bijvoorbeeld $B(15, 0.7)$-verdeeld is, is de kans op een mislukking 0.3. We bedenken dan dat het aantal mislukkingen $X = 15 - Y \quad B(15, 0.3)$-verdeeld is, dus $P(Y = 10) = P(X = 15 - 10) = 0.2061$.

De Hypergeometrische verdeling

Voorbeeld 4.5.3 Een hotelmanager wil voor 5 van zijn hotelkamers flatscreens aanschaffen door een deel van een faillissementspartij van 20 flatscreens op te kopen. Van deze partij is bekend dat 5 flatscreens ernstige gebreken vertonen. Hoeveel van deze (verpakte) flatscreens dient de hotelmanager (minimaal) aan te schaffen opdat hij met een kans van minstens 90%, vijf (of meer) goed werkende flatscreens heeft?
Stel dat hij 8 flatscreens opkoopt. Dan komt dat neer op 8 keer aselect trekken zonder terugleggen uit 20 flatscreens waarvan er 15 goed en 5 slecht werken.
De kans dat hij (precies) $X = 5$ goede flatscreens kiest, wordt gegeven door de hypergeometrische formule (eigenschap 2.1.15):

$$P(X = 5) = \frac{\binom{15}{5}\binom{5}{3}}{\binom{20}{8}} \approx 23.8\%$$

$$\text{En } P(X \geq 5) = \sum_{k=5}^{8} \frac{\binom{15}{k}\binom{5}{8-k}}{\binom{20}{8}} \approx 94\%$$

Conclusie: bij 8 aangekochte flatscreens aan de voorwaarde is voldaan.
(Controleer dat dit bij 7 aangekochte flatscreens niet het geval is)
Het verwachte aantal goed werkende flatscreens onder de gekozen acht is intuïtief 75% van deze 8, dus 6 flatscreens. Immers, 15 van de beschikbare 20 flatscreens werkt goed.
Merk op dat het verwachte aantal weer van de vorm np is, namelijk $8 \cdot \frac{15}{20} = 6$ ■

Als de kansfunctie van X gegeven wordt door de hypergeometrische formule, spreken we van een hypergeometrische verdeling. Deze kunnen we toepassen indien we een aantal keren **aselecte trekkingen** uitvoeren, **zonder terugleggen** en uit een **dichotome** populatie (die bestaat uit 2 "soorten" elementen, zoals rode en witte ballen in eigenschap 2.1.15).

$X =$ "aantal rode ballen in de steekproef"

	Rood	Wit	Totaal
Populatie	R	$N - R$	N
	$\downarrow$	$\downarrow$	$\downarrow$
Trekking	k	$n - k$	n

Definitie 4.5.4 (de hypergeometrische verdeling)
X is hypergeometrisch verdeeld (met parameters N, R en n) als

$$\boldsymbol{P(X = k)} = \frac{\binom{\boldsymbol{R}}{\boldsymbol{k}}\binom{\boldsymbol{N-R}}{\boldsymbol{n-k}}}{\binom{\boldsymbol{N}}{\boldsymbol{n}}}, \qquad \text{voor } k = 0, 1, 2, \ldots, n$$

Voor het aantal X in voorbeeld 4.5.3 geldt dan: $P(X=2)=\frac{\binom{15}{2}\binom{5}{6}}{\binom{20}{8}}=?$
Twee werkende flatscreens impliceert bij een keuze van 8 dus zes niet-werkende. Maar er zijn slechts 5 niet-werkende op voorraad. In de formule verschijnt ook de (onbekende) binomiaalcoëfficiënt $\binom{5}{6}$, het “aantal keuzes van 6 uit 5”.
Door af te spreken dat dit aantal $\binom{5}{6}=0$, is ook deze kans 0.
De verwachte fractie werkende flatscreens in de steekproef is gelijk aan de fractie werkende in de populatie, dus $E(X)=8\cdot\frac{15}{20}=6$.

Algemeen geldt: $\boldsymbol{E(X)=np}$ en $\boldsymbol{var(X)=np(1-p)\cdot\frac{N-n}{N-1}}$, waarin $\boldsymbol{p=\frac{R}{N}}$

In het volgende hoofdstuk tonen we de juistheid van deze uitdrukkingen aan.

Aselecte trekkingen uit een dichotome populatie leiden enerzijds tot de hypergeometrische verdeling bij trekken **zonder** terugleggen en anderzijds tot de binomiale verdeling bij trekken **met** terugleggen. In het laatste geval is er namelijk sprake van **onafhankelijkheid** (steeds weer dezelfde samenstelling van de populatie waaruit getrokken wordt), in het eerste geval zijn de trekkingen **afhankelijke** experimenten.
In termen van de figuur bij definitie 4.5.4:

$$P(\text{"}1^{\text{ste}}\text{ getrokken bal is rood"})=P(\text{“}2^{\text{de}}\text{ getrokken bal is rood”})=\frac{R}{N}$$

$$\text{maar: } P\left(2^{\text{de}}\text{ rood }\middle|1^{\text{ste}}\text{rood}\right)=\frac{R-1}{N-1}$$

De kans op “de tweede rood” hangt af van het resultaat van de eerste trekking.
Echter, als de samenstellende delen van de populatie, waaruit getrokken wordt, zeer omvangrijk zijn (denk bijv. aan het aantal Nederlandse stemgerechtigden en de aanhang van een grote partij), geldt bij benadering: $\frac{R}{N}\approx\frac{R-1}{N-1}$
Bij relatief kleine aantallen trekkingen (bijv. $n=1000$ trekkingen uit $N=10$ miljoen kiezers) veel kleiner dan R of $N-R$) is er bij benadering sprake van onafhankelijkheid en geldt (we geven geen bewijs):

Eigenschap 4.5.5 Voor relatief grote R en $N-R$ en relatief kleine n gaat de hypergeometrische verdeling met parameters N, R en n bij benadering over in een $\boldsymbol{B\left(n,\frac{R}{N}\right)}$**-verdeling.**

Merk op dat ook de varianties onder deze voorwaarden nagenoeg aan elkaar gelijk zijn:
$np(1-p)\cdot\frac{N-n}{N-1}\approx np(1-p)$

Een (zeer veilige) vuistregel voor toepassing van eigenschap 4.5.5 luidt: $\boldsymbol{N>5n^2}$
Als $n=1000$ kan bij trekkingen zonder terugleggen de $B\left(n,\frac{R}{N}\right)$-verdeling worden gebruikt als $N>5\cdot1000^2=5\,000\,000$

De Geometrische verdeling

De geometrische verdeling is in voorbeelden 4.2.5 en 4.3.3 ter sprake gekomen bij de handelsreiziger die net zo lang klanten bezoekt totdat hij een pannenset verkoopt.
De onafhankelijke verkooppogingen hadden elk een succeskans $p = \frac{1}{10}$.
Het aantal Bernoulli-pogingen bepalen totdat er een succes optreedt, komt o.m. ook voor bij mens-erger-je-niet: 6 proberen te werpen ($p = \frac{1}{6}$). Of met tussenpozen proberen in te loggen op een druk bezet computersysteem, etc. In het algemeen, als er sprake is van een aantal pogingen tot het eerste succes.

Definitie 4.5.6 X is **geometrisch verdeeld met parameter** $p \in (0, 1]$, als

$$\boldsymbol{P(X = k) = (1 - p)^{k-1} p}\,, \quad \text{voor } k = 1, 2, \ldots.$$

Voor $p = 1$ hebben we de ontaarde verdeling $P(X = 1) = 1$.
Met behulp van de meetkundige reeks en zijn afgeleiden (zie bijlage “Wiskundige technieken”) kunnen we bewijzen:

$$\boldsymbol{E(X) = \frac{1}{p}} \quad \text{en} \quad \boldsymbol{var(X) = \frac{1-p}{p^2}}$$

De Poisson-verdeling

Voorbeeld 4.5.7 Uit onderzoek van de PTT in een bepaalde regio blijkt dat er op een willekeurig moment (op werkdagen) gemiddeld 4 lijnen voor autotelefoons in een systeem bezet zijn. Hoe groot moet nu het aantal lijnen (minimaal) zijn, opdat de kans op overbelasting op een willekeurig moment kleiner dan 0.001 is?
Om deze vraag te beantwoorden moeten we een kansmodel opstellen: we kiezen X als het aantal bezette lijnen (bij zeer grote capaciteit) op een willekeurig moment.
Wat is dan de kansverdeling van X?
Vervolgens kunnen we de vraag beantwoorden wat het minimale aantal van m lijnen moet zijn opdat $P(X > m) \leq 0.001$.
Als er zich in de regio 1000 autotelefoonbezitters bevinden, ligt het voor de hand te veronderstellen dat elk van hen, onafhankelijk van elkaar, met kans $\frac{4}{1000}$ op een willekeurig moment een telefoongesprek voert.
Het aantal telefoongesprekken (= aantal bezette lijnen) is dan $B\left(1000, \frac{4}{1000}\right)$-verdeeld met het gewenste “gemiddelde” van $\mu = np = 4$. Afgezien van rekenproblemen door bijv. binomiaalcoëfficiënten van de vorm $\binom{1000}{k}$ kennen we in dit soort gevallen ook het aantal autotelefoonbezitters op dat moment in de regio niet erg precies. Als we aannemen dat het er n zijn, die onafhankelijk van elkaar met kans $p = \frac{4}{n} = \frac{\mu}{n}$ op een willekeurig moment telefoneren, dan geldt:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} = \frac{n(n-1)\ldots(n-k+1)}{k!} \cdot \left(\frac{\mu}{n}\right)^k \left(1 - \frac{\mu}{n}\right)^{n-k}$$

Herschrijven van de factoren leidt tot:

$$\lim_{n\to\infty} P(X=k) = \lim_{n\to\infty} \frac{\mu^k}{k!} \cdot \frac{n(n-1)\ldots(n-k+1)}{n \cdot n \cdot \ldots\ldots\ldots \cdot n} \cdot \left(1-\frac{\mu}{n}\right)^{n-k} = \frac{\mu^k e^{-\mu}}{k!}$$

Dit resultaat kan berekend worden door de limiet van de factoren te nemen:
$\lim_{n\to\infty} \frac{n}{n} = \cdots = \lim_{n\to\infty} \frac{n-k+1}{n} = 1$ en $\lim_{n\to\infty} \left(1-\frac{\mu}{n}\right)^{n-k} = \lim_{n\to\infty} \left(1-\frac{\mu}{n}\right)^{n} = e^{-\mu}$.
Dit laatste volgt uit de standaardlimiet $\lim_{n\to\infty} \left(1+\frac{x}{n}\right)^n = e^x$ (zie je calculus boek).

De limietverdeling van X geldt grote *n* geldt bij benadering: $P(X=k) = \frac{\mu^k e^{-\mu}}{k!}, k = 0,1,2$ ....

Terugkerend naar de vraag wat het minimale aantal van m lijnen moet zijn opdat $P(X > m) \leq 0.001$, kunnen we deze vraag met bovengenoemde kansverdeling beantwoorden. ■

Definitie 4.5.8 X is **Poisson-verdeeld** met parameter $\mu > 0$ als

$$\boldsymbol{P(X=k) = \frac{\mu^k e^{-\mu}}{k!}}, \qquad \text{voor } k = 0,1,2 \ldots.$$

Dat dit inderdaad een kansfunctie definieert (voor μ > 0), volgt uit de Taylor-reeksontwikkeling (om 0) van de *e*-macht: $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$
(zie ook de bijlage wiskundige technieken)

$$\sum_k P(X=k) = \sum_{k=0}^{\infty} \frac{\mu^k e^{-\mu}}{k!} = e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^k}{k!} = e^{-\mu} \cdot e^{\mu} = 1$$

Voor de verwachtingswaarde ligt het “gemiddelde” μ voor de hand, en dat klopt:

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

Voor de formule van de variantie $var(X) = \mu$ hebben we enige rekenkundige trucs nodig.
We gebruiken: $var(X) = E(X^2) - (EX)^2 = E\big(X(X-1)\big) + E(X) - (EX)^2$
In de laatste uitdrukking is $E(X) = EX = \mu$ en

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

Dus: $\boldsymbol{var(X)} = E\big(X(X-1)\big) + E(X) - (EX)^2 = \mu^2 + \mu - \mu^2 = \boldsymbol{\mu}$

Ook voor de Poisson-verdeling zijn (cumulatieve) kans-tabellen beschikbaar. Daaruit kunnen we voor voorbeeld 4.5.7, het minimum m voor het aantal beschikbare lijnen zodat

$P(X\ > m) \leq 0.001$, snel aflezen (μ = 4): dit minimum is 11.

De eigenschap die we in dit voorbeeld hebben toegepast, kunnen we als volgt formuleren.

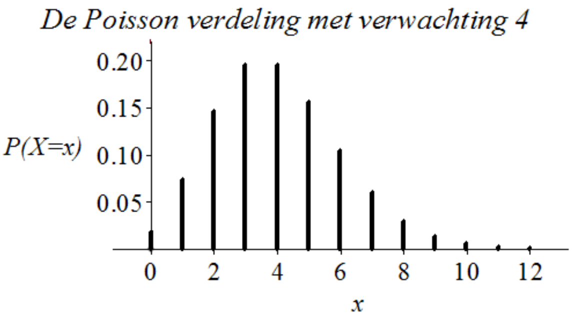
Eigenschap 4.5.9 Als $\boldsymbol{X}$ $\boldsymbol{B(n,p)}$**-verdeeld is met "grote** $\boldsymbol{n}$ **en kleine** $\boldsymbol{p}$**"**, dan is X **bij benadering Poisson-verdeeld met parameter** $\boldsymbol{\mu = np}$.

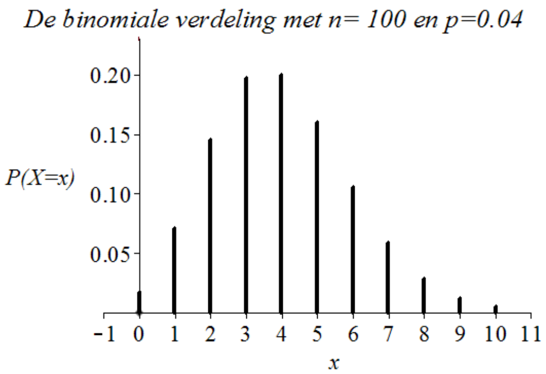
Als **vuistregel** voor deze benadering gebruiken we $\boldsymbol{n > 25}$ en $\boldsymbol{np < 10}$ **of** $\boldsymbol{n(1-p) < 10}$.

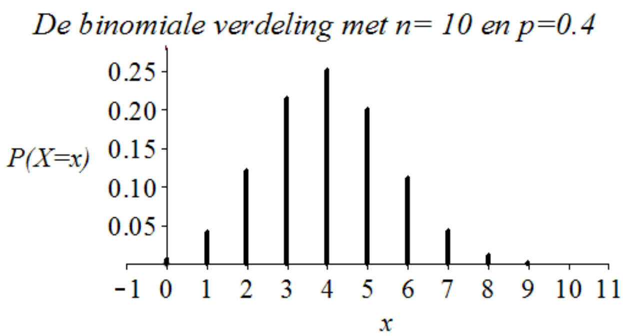
Overigens zijn dit soort benaderingen dus ook toepasbaar voor grote n en "p dicht bij 1".
Want als het aantal successen $X \sim B(n,p)$, dan geldt voor het aantal mislukkingen:
$(n - X) \sim B(n, 1-p)$.

Ter illustratie van eigenschap 4.5.9 vergelijken we in de volgende tabel de Poisson-verdeling met μ = 4 met de $B\left(10,\frac{2}{5}\right)$- en de $B\left(100,\frac{1}{25}\right)$-verdeling (steeds geldt $E(X) = 4$):

Verdeling	$P(X{=}0)$	$P(X{=}1)$	$P(X{=}2)$	$P(X{=}3)$	$P(X{=}4)$	$P(X{=}5)$	$P(X = 6)$	$P(X = 7)$	$P(X = 8)$
$B\left(10,\frac{2}{5}\right)$	0.006	0.040	0.121	0.215	0.251	0.201	0.111	0.042	0.011
$B\left(100,\frac{1}{25}\right)$	0.017	0.070	0.145	0.197	0.199	0.160	0.105	0.059	0.029
Poisson μ= 4	0.018	0.073	0.147	0.195	0.195	0.156	0.104	0.060	0.030



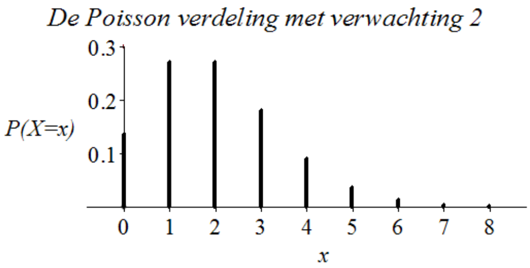


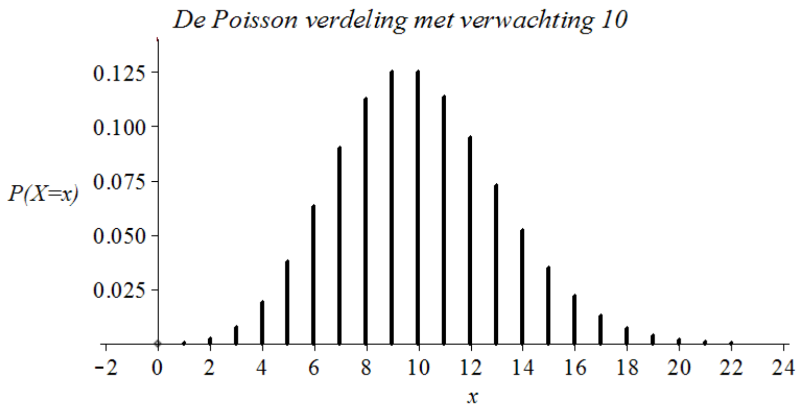


Eigenschap 4.5.9 geeft de mogelijke toepassingen van de Poisson-verdeling reeds aan: vaak gaat het om het aantal schaars optredende (**zeldzame**) gebeurtenissen per tijdseenheid of in een gebied, zoals het aantal autodiefstallen op een dag in een grote stad, het aantal acute blindedarmoperaties in een ziekenhuis in een dag, het aantal vrijkomende deeltjes per

tijdseenheid van een stuk radioactief materiaal of het aantal paddenstoelen op een are bosgebied.
In veel van dit soort gevallen is wel een "gemiddelde" (= verwachtingswaarde μ) bekend en wordt de kansverdeling goed beschreven met de Poisson-kansfunctie met betreffende gemiddelde als parameter. Vergroting of verkleining van tijdeenheid of gebied heeft alléén een evenredige vergroting of verkleining van dat gemiddelde tot gevolg.

Ter illustratie nog enige Poisson-kansfuncties met verschillende waarden van μ.





Merk op dat de Poisson-verdeling in zijn algemeenheid scheef is (niet-symmetrisch), zoals duidelijk het geval is bij μ = 2. Voor μ =10 lijkt de verdeling redelijk symmetrisch en klokvormig: d empirische regel is bij benadering van toepassing. Hoe groter de waarde van μ, des te sterker dit geldt. ■

We vatten de verdelingen, die in dit hoofdstuk zijn besproken, samen in de volgende

Eigenschap 4.5.10 (Bekende discrete verdelingen met hun karakteristieken)

Verdeling	**Kansfunctie**	$\boldsymbol{E(X)}$	$\boldsymbol{var(X)}$	**Voorbeeld**
Homogeen op 1, 2, …, N	$P(X=x)=\frac{1}{N}, x=1,2,..,N$	$\frac{N+1}{2}$	- -	resultaat van één worp dobbelsteen
Alternatief (p)	$P(X=1)=p$ $P(X=0)=1-p$	p	$p(1-p)$	Dobbelsteen 6 (X=1) of niet (0)
Binomiaal $B(n,p)$	$P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}$, $x=0,\ 1,\ 2,\ \ldots,n$	np	$np(1-p)$	Aantal zessen bij 30 worpen met een dobbelsteen
Geometrisch (p)	$P(X=x)=(1-p)^{x-1}p$, $x=1,\ 2,\ 3,\ \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	Aantal worpen met dobbelsteen tot de eerste 6
Poisson (μ)	$P(X=x)=\frac{\mu^x e^{-\mu}}{x!}$, $x=0,1,\ldots$	μ	μ	Aantal klanten dat in 10 minuten kantoor in komt
Hyper-geometrisch (R,N,n)	$P(X=x)=\frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}}$, $x=0,1,\ldots,n$	np $p=\frac{R}{N}$	$np(1-p)\cdot\frac{N-n}{N-1}$	Aantal meisjes als we 5 personen kiezen uit 10 jongens en 12 meisjes

4.6 Vraagstukken

1. De kansfunctie van X wordt gegeven door de volgende tabel.

x	-5	-2	0	1	3
$P(X=x)$	0.1	0.2	0.2	0.4	0.1

Schets de grafiek van de kansfunctie en bereken

a. $P(X > 0)$
b. $E(X)$
c. $E(X^2)$ en
d. De variantie en de standaardafwijking van X (door de resultaten van b. en c. te gebruiken).

2. Geef de kansfunctie van de variabele X = "het eerste cijfer van het nummerbord van een willekeurige passerende auto".
Schets de kansfunctie en bepaal $E(X)$, $E(X^2)$ en $var(X)$

3. ("**One-armed bandit**") Bij een simpel model fruitmachine hebben we 3 schijven, elk met 10 symbolen, waarvan één J (van *Jackpot*). Na inworp van één munt draaien de schijven een tijdje rond. Als ze tot stilstand gekomen zijn, kan men door een venster 3 naast elkaar liggende symbolen - één van elke schijf - zien. Als er geen J bij zit, is men de inworp kwijt, is er één J dan krijgt men de inworp terug, zijn er twee J's dan krijgt men in totaal 10 munten van de machine (zodat de winst 9 munten is). Hoeveel zou de uitbetaling bij 3 J's moeten zijn om er een "eerlijk" spel ("gemiddeld quitte spelen") van te maken?

4. ("**Chuck-a-luck**") Bij dit kermisspelletje zet je een bepaald bedrag in op één van de getallen 1, 2, 3, 4, 5 of 6. Dan wordt één keer gegooid met drie zuivere dobbelstenen en wordt gekeken hoeveel dobbelstenen het ogen aantal tonen waarop je hebt ingezet.
Is dat 0, dan ben je je inzet kwijt. Is het positief dan, krijg je je inzet terug en win je bovendien je inzet maal dat aantal dobbelstenen.
Bereken de verwachte winst bij inzet van 1 Euro.

5. In een vaas bevinden zich N knikkers genummerd $1, 2, \ldots, N$. Lukraak worden uit de vaas, zonder teruglegging, n ($n \leq N$) knikkers getrokken. We definiëren X als het nummer van de knikker die het hoogste nummer heeft van de getrokken knikkers.

a. Bepaal $P(X = 7)$ als $n = 4$ en $N = 10$.
b. Bepaal de kansfunctie van X voor willekeurige $1 \leq n \leq N$.

6. Een stochastische variabele X heeft een binomiale verdeling. Bereken met behulp van binomiale tabel (achter in het dictaat) de volgende kansen.

a. $P(X \leq 7)$, als $n = 10$ en $p = 0.3$
b. $P(X \geq 7)$**,** als $n = 10$ en $p = 0.3$
c. $P(X = 9)$, als $n = 15$ en $p = 0.6$. Controleer deze waarde met de binomiale formule
d. $P(X < 12)$, als $n = 15$ en $p = 0.6$

7. Een stochastische variabele X heeft een Poisson verdeling met $\mu = 3$.
Bereken met behulp van de tabel voor Poisson-kansen:

a. $P(X = 5)$. Controleer de gevonden waarde met de Poisson kansfunctie.
b. $P(X < 2)$.
c. $P(X > 3)$.

8. Bepaal voor elk van de volgende voorbeelden of de stochastische variabele X een binomiale, hypergeometrische, geometrische of Poisson-verdeling heeft of géén van allen.
Geef steeds 1. een korte argumentatie voor keuze van de verdeling (inclusief parameters)
2. De berekening van de kans $P(X = 2)$
3. $E(X)$

a. Een softwareontwikkelaar zorgt voor een support hotline voor klanten, die kunnen bellen om vragen te stellen over het gebruik van de software. De ervaring leert dat op een normale werkdag dat er gemiddeld 30 per uur zijn. X is het aantal telefoontjes dat de support hotline in een periode van 10 minuten krijgt
b. Veronderstel dat een bedrijf twee vacatures heeft op directieniveau. Er zijn 5 even geschikte kandidaten, waarvan twee vrouwelijke. Men besluit tot loting en $X =$ "het aantal gekozen vrouwen".
c. Een fabrikant van computerchips neemt een aselecte steekproef van 100 chips uit de productie van elk uur om het aantal defecte chips te schatten.
X is het aantal defecte chips in de steekproef van 100 chips in een uur waarin 2% van de totale productie defect is.
d. Iemand probeert met een sleutelbos van 10 sleutels een deur open te krijgen. Op die deur past slechts één van de sleutels. X is het aantal pogingen, dat nodig om de deur te openen als hij steeds een willekeurige sleutel uit de bos kiest ("met terugleggen")
e. Bekijk het sleutelprobleem in d. nog eens, maar laat nu X het aantal pogingen zijn als hij een sleutel die niet past van de bos afhaalt en vervolgens uit de overige sleutels er willekeurig één kiest ("zonder terugleggen").

9. We gaan na hoe goed de binomiale verdeling met parameters $n = 25$ en $p = 0.05$ te benaderen is met de Poisson-verdeling.

a. Bereken $P(X = 0)$ met de binomiale verdeling.
b. Benader $P(X = 0)$ met de geschikte Poisson-verdeling.

10. De variabele X heeft een eenvoudige verdeling: $P(X = c) = \frac{1}{2}$ en $P(X = 0) = \frac{1}{2}$.

a. Bepaal een formule voor de momenten $E(X^k)$.
b. Gebruik a. om de verwachting en de variantie van X te bepalen.

11. De variabele X heeft de volgende kansfunctie: $P(X = i) = c \cdot \left(\frac{1}{3}\right)^i$, voor $i = 0, 1, 2, ...$

a. Ga na dat $c = \frac{2}{3}$ en schets de kansfunctie.
b. Ga na dat $Y = X + 1$ een bekende verdeling heeft. Welke incl. parameter(s)?
c. Gebruik het verband om de verwachting en de variantie van X te bepalen.

12. (oude tentamenopgave)
Een bedrijf met 150 medewerkers wil voor al haar medewerkers telefoontoestellen laten plaatsen. Zij kunnen via een centrale automatisch "naar buiten" bellen. Het aantal

(kostbare) buitenlijnen kan echter veel kleiner dan 150 zijn, daar niet iedereen tegelijk belt: gedurende kantooruren verwacht men dat op een willekeurig tijdstip één op de vijftig werknemers een buitengesprek voert.

a. Wat is de kansverdeling van het aantal buitengesprekken op zo'n willekeurig tijdstip? Geef aan welke veronderstellingen daarvoor noodzakelijk zijn.

b. Bepaal het kleinste aantal buitenlijnen zodat de kans op overbelasting op een willekeurig tijdstip bij benadering kleiner dan 5% is. Gebruik de Poisson-tabel.

13. (oude tentamenopgave) Geef voor elk van de drie hieronder beschreven situaties aan welke verdeling van X volgens u het meest voor de hand ligt en bepaal $P(X > EX)$.

a. Een colporteur verkoopt gemiddeld aan 15% van de door hem bezochte klanten een energiecontract. Op een zekere dag bezoekt hij 12 willekeurig gekozen potentiële klanten. X is het aantal verkochte sets op die dag.

b. Een ziekenhuis heeft de beschikking over twee couveuses. In het verzorgingsgebied van het ziekenhuis is de "vraag" naar couveuses op een willekeurig tijdstip **gemiddeld** eveneens twee. X is de vraag naar couveuses op een willekeurig tijdstip.

c. Van een groep van 10 kandidaat-astronauten, waaronder 3 Nederlanders, worden 4 personen door een (eerlijke) loting uitverkoren om een vlucht met een ruimteveer te maken. X is het aantal Nederlanders onder de uitverkorenen.

14. Het is bekend dat 4% van de in een supermarkt verkochte eieren niet in de aangegeven gewichtsklasse 2 vallen.

a. Bereken de kans dat in een doos van 10 eieren er één of meer buiten de gewichtsklasse 2 vallen. Welke veronderstelling(en) dient men voor deze berekening te maken?

b. Bereken of benader (onder dezelfde veronderstellingen) de kans dat er in 10 dozen van 10 eieren zich vier of meer eieren buiten de gewichtsklasse 2 bevinden.

c. We kopen net zo lang dozen eieren van gewichtsklasse 2, totdat een doos één (of meer) eieren bevat die niet tot die gewichtsklasse behoort. Wat is dan het verwachte aantal gekochte dozen?

15. **a.** Bij een spel moeten we met een zuivere dobbelsteen drie keer een 6 werpen. Bereken de kans dat we de derde 6 in de tiende worp gooien.

b. (Generalisatie van a). We herhalen een Bernoulli-experiment met succeskans p net zo lang totdat we m successen hebben. Als X het benodigd aantal experimenten is, bepaal dan S_X en de kansfunctie van X.
(Deze kansverdeling wordt de **negatief binomiale verdeling** genoemd.)

16. M heet een **mediaan** van de verdeling van X als $P(X \geq M) \geq \frac{1}{2}$ en $P(X \leq M) \geq \frac{1}{2}$

Bepaal de mediaan (medianen) van

a. de geometrische verdeling met parameter $p = \frac{1}{3}$,

b. de Poisson-verdeling met parameter 2,

c. de Poisson-verdeling met parameter 2.5 en

d. de $B(7, ½)$-verdeling.

Enige aanwijzingen bij de opgaven van hoofdstuk 4:

1. Schrijf eerst de formules van $E(X)$, $E(X^2)$ en $var(X)$ $(= E(X^2) = (EX)^2$) op.
2. Idem
3. Bepaal de kansen op elk aantal J's en de bijbehorende winst (uitbetaling – inleg).
4. Idem
5. Los dit combinatorisch op: wat is het totale aantal en wat is het gunstige aantal trekkingsresultaten?
6. Bedenk dat X alleen gehele waarden kan aannemen en dat kansen van de vorm $P(X \leq k)$ in de tabel staan.
Bij c. en d. : $p > 0.5$. Als de kans op een succes groter dan ½ is, is de kans op een mislukking < ½ , dus ga over op het bijbehorende aantal mislukkingen.
7. Poisson-tabellen werken hetzelfde als die van de binomiale verdeling.
8. Zorg dat je de typeringen van de verdelingen en hun verwachtingswaarden paraat hebt.
- Geometrisch: tel het aantal onafhankelijke pogingen totdat je succes hebt (steeds kans p)
- Binomiaal: aantal successen (p) bij n onafhankelijke pogingen (trekken met terugleggen)
- Hypergeometrisch: n trekkingen zonder terugleggen, tel het aantal "successen" ($p = \frac{R}{N}$)
- Poisson: aantal zeldzame gebeurtenissen in gebied/periode, gemiddeld μ
- Homogeen: gelijke kansen op een eindig aantal uitkomsten.

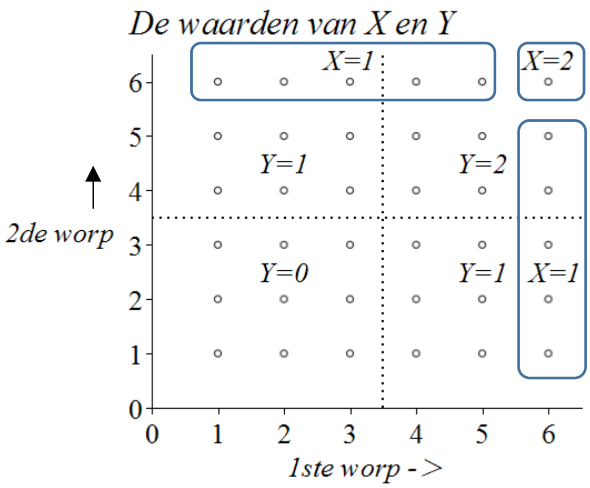
9. De formule van de Poisson-kansfunctie hoef je niet te kennen: staat op formuleblad.
10. Formule van $E(X^k)$ onthouden als gewogengemiddelde van x^k
11. a. herken in de sommatie een meetkundige reeks (zie bijlage wiskundige technieken).
b. Formules voor $E(aX + b)$ en $var(aX + b)$ moet je paraat hebben.
12. Zie 8.
13. Zie 8.
14. Zie 8.
15. Formule is op een zelfde manier af te leiden als die van geometrisch en binomiaal: bedenk dat de laatste poging een succes moet zijn, want dan is het derde succes "binnen".
16. Bij de Poisson- en binomiale verdelingen kun je gebruik maken van de tabellen.

Hoofdstuk 5: Twee of meer discrete variabelen

5.1 Simultane kansfuncties

In hoofdstuk IV hebben we steeds de verdeling van één stochastische variabele bestudeerd. Bij een experiment kunnen echter meerdere kwantitatieve aspecten een rol spelen: we kunnen meerdere stochastische variabelen definiëren. Als bijvoorbeeld de stochastische variabelen X en Y gedefinieerd zijn op dezelfde kansruimte, dan zijn we geïnteresseerd in de kans op het gelijktijdig optreden van de gebeurtenissen $\{X \in B\}$ en $\{Y \in C\}$, waarbij $B \subset \mathbb{R}$ en $C \subset \mathbb{R}$, en in de samenhang van de verdelingen van X en van Y. We beperken ons eerst tot twee discrete stochastische variabelen.

Voorbeeld 5.1.1 We werpen tweemaal met een zuivere dobbelsteen en definiëren X en Y als "het aantal geworpen zessen" resp. "het aantal worpen met een ogenaantal groter dan 3". Deze aantallen kunnen 0, 1 of 2 zijn, dus $S_X = S_Y = \{0, 1, 2\}$.
De symmetrische kansruimte bestaat uit 36 uitkomsten en bij elke uitkomst krijgt zowel X als Y een waarde, zoals schematisch in de volgende figuur wordt aangegeven:



$i \backslash j$	0	1	2
0	$\frac{9}{36}$	$\frac{12}{36}$	$\frac{4}{36}$
1	0	$\frac{6}{36}$	$\frac{4}{36}$
2	0	0	$\frac{1}{36}$

Tabel van simultane kansen P(X = i en Y = j)

De gebeurtenis dat beide worpen groter dan 3 zijn en bovendien één worp gelijk aan 6 is, dus $\{X = 1 \text{ en } Y = 2\}$, treedt op als de uitkomst (4,6), (5,6), (6,4) of (6,5) is.
Dus $P(X = 1 \text{ en } Y = 2) = \frac{4}{36}$.
Zo kunnen we voor alle waarden van i en j de zogeheten simultane kansen $P(X = i \text{ en } Y = j)$ bepalen. Voor $i \in S_X$ en $j \in S_Y$ vinden we de bovenstaande tabel.

We zien dat de som van alle kansen één is: $\sum_{i=0}^{2} \sum_{j=0}^{2} P(X = i \text{ en } Y = j) = 1$
De gebeurtenissen $\{X = i \text{ en } Y = j\}$ vormen immers voor de verschillende waarden van i en j een partitie van de uitkomstenruimte.

Gebruikmakend van bovenstaande tabel van simultane kansen, kunnen we de kans op één zes (dus $X = 1$) bepalen door de gebeurtenis $\{X = 1\}$ op te delen in $\{X = 1 \text{ en } Y = 0\}$, $\{X = 1 \text{ en } Y = 1\}$ en $\{X = 1 \text{ en } Y = 2\}$:

$$\begin{aligned} P(X = 1) &= P(X = 1 \text{ en } Y = 0) + P(X = 1 \text{ en } Y = 1) + P(X = 1 \text{ en } Y = 2). \\ &= 0 + \tfrac{6}{36} + \tfrac{4}{36} = \tfrac{10}{36} \end{aligned}$$

Evenzo vinden we $P(X = 0) = \sum_{j=0}^{2} P(X = 0 \text{ en } Y = j) = \frac{25}{36}$.

En $P(X = 2) = \sum_{j=0}^{2} P(X = 2 \text{ en } Y = j) = \frac{1}{36}$.

Hiermee hebben we de verdeling van X afgeleid uit de tabel van $P(X = i \text{ en } Y = j)$, door kansen per rij op te tellen. Uiteraard kunnen we in dit geval ook de kansfunctie van X meteen uit de omschreven situatie afleiden (zie figuur): het gaat hier immers om het aantal successen (zessen) bij twee Bernoulli-pogingen (worpen).
De verdeling van Y vinden we door de kansen kolomsgewijs op te tellen. ■

Definitie 5.1.2 Voor een paar discrete stochastische variabelen (X, Y) gedefinieerd op dezelfde kansruimte en met een waardenbereik $S_X \times S_Y$, noemen we $\boldsymbol{P(X = x \text{ en } Y = y)}$ voor $(x, y) \in S_X \times S_Y$ de **simultane kansfunctie** van X en Y.

Eigenschap 5.1.3 Voor de simultane kansfunctie $P(X = x \text{ en } Y = y)$ van X en Y geldt:

$$\mathbf{1)}\quad \boldsymbol{P(X = x \text{ en } Y = y) \geq 0}.$$

$$\mathbf{2)}\quad \sum_{\boldsymbol{x \in S_X}} \sum_{\boldsymbol{y \in S_Y}} \boldsymbol{P(X = x \text{ en } Y = y) = 1}.$$

Omgekeerd geldt dat elke tweedimensionale functie die aan deze twee eisen voldoet een simultane kansfunctie is. De kansfunctie van X alleen of Y alleen noemen we in dit verband de **marginale** kansfunctie van X of van Y. Generaliseren we de afleiding van de marginale kansfunctie van X, gegeven in voorbeeld 5.1.1, dan vinden we:

Eigenschap 5.1.4

$$\boldsymbol{P(X = x) = \sum_{y \in S_Y} P(X = x \text{ en } Y = y)} \qquad \text{en} \qquad \boldsymbol{P(Y = y) = \sum_{x \in S_X} P(X = x \text{ en } Y = y)}$$

In voorbeeld 5.1.1 komt dit neer op het sommeren van de rijen (X) resp. de kolommen (Y) van de tabel:

$i \backslash j$	0	1	2	$P(X = i)$
0	$\frac{9}{36}$	$\frac{12}{36}$	$\frac{4}{36}$	$\frac{25}{36}$
1	0	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{10}{36}$
2	0	0	$\frac{1}{36}$	$\frac{1}{36}$
$P(Y = j)$	$\frac{9}{36}$	$\frac{18}{36}$	$\frac{9}{36}$	1

Vooruitlopend op de definitie van onafhankelijkheid van X en Y in paragraaf 5.3 merken we op dat de gebeurtenissen $\{X = i\}$ en $\{Y = j\}$ in het algemeen niet o.o. zijn.

Zo is $\frac{9}{36} = P(X = 0 \text{ en } Y = 0) \neq P(X = 0) \cdot P(Y = 0) = \frac{25}{36}$
Worpen groter dan 3 kunnen ook niet-zessen opleveren. De kans dat het aantal zessen kleiner is dan het aantal worpen groter dan 3 berekenen we als volgt:

$$\begin{aligned} P(X < Y) &= P(X = 0 \text{ en } Y = 1) + P(X = 0 \text{ en } Y = 2) + P(X = 1 \text{ en } Y = 2) \\ &= \frac{12}{36} + \frac{4}{36} + \frac{4}{36} = \frac{20}{36} \end{aligned}$$ ■

De gebeurtenis $\{X < Y\}$ bestaat in dit voorbeeld uit 3 paren van waarden die X en Y kunnen aannemen.
Algemeen: als B een deelverzameling is van het platte vlak, dus $B \subset \mathbb{R}^2$, dan kunnen we de kans dat het paar (X, Y) waarden uit B aanneemt, als volgt in de simultane kansfunctie van X en Y uitdrukken:

$$P((X, Y) \in B) = \sum_{(x,y) \in B} P(X = x \text{ en } Y = y)$$

Voorbeeld 5.1.5 De schakers Timman en Karpov spelen een match van zes partijen. Uit alle vroegere onderlinge partijen leiden we af dat Timman met kans $p_1 = \frac{2}{10}$ een partij van Karpov wint, met kans $p_2 = \frac{3}{10}$ verliest (en dus met kans $p_3 = \frac{5}{10}$ remise overeenkomt), ongeacht wie er met de witte stukken speelt.
Hoe groot is dan de kans dat Timman de match wint (dus meer winstpartijen op zijn naam brengt dan Karpov), als we aannemen dat de resultaten van partijen elkaar niet beïnvloeden? (Is dat een redelijke modelveronderstelling?)
Als we X het aantal winstpartijen van Timman en Y het aantal winstpartijen van Karpov noemen dan is de gevraagde kans:

$$P(X > Y) = \sum_{i > j} P(X = i \text{ en } Y = j)$$

De kans bijvoorbeeld dat Timman drie keer wint ($X = 3$) en één keer verliest ($Y = 1$) van Karpov, en dus twee keer remise overeenkomt, kunnen we als volgt berekenen. Als *w* winst, *v* verlies en *r* remise voor Timman voorstelt, dan treedt $\{X = 3 \text{ en } Y = 1\}$ bijvoorbeeld op als (*v*, *r*, *r*, *w*, *w*, *w*) de uitkomst van de match is, met kans:

$$P\big((v, r, r, w, w, w)\big) = p_2 \cdot p_3 \cdot p_3 \cdot p_1 \cdot p_1 \cdot p_1 = p_2 (1 - p_1 - p_2)^2 p_1^3,$$

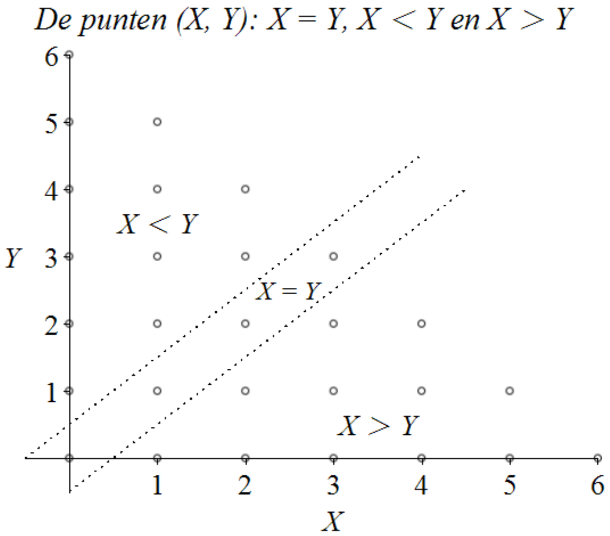
wegens de onafhankelijkheid van de partijen.
Het aantal volgordes met 3 winst-, 1 verlies- en 2 remisepartijen is $\binom{6}{3}\binom{3}{1}\binom{2}{2} = \frac{6!}{3!1!2!}$
dus (elke van die volgordes heeft gelijke kans):

$$P(X = 3 \text{ en } Y = 1) = \frac{6!}{3!\,1!\,2!} \left(\frac{2}{10}\right)^3 \left(\frac{3}{10}\right) \left(\frac{5}{10}\right)^2 \approx 3.6\%.$$

Algemener kunnen we de simultane kansfunctie van X en Y bij n partijen met winstkansen p_1 en p_2 voor resp. Timman en Karpov als volgt opschrijven:

$$P(X = i \text{ en } Y = j) = \frac{n!}{i!\,j!\,(n - i - j)!} p_1^i p_2^j (1 - p_1 - p_2)^{n-i-j}$$

Dit geldt voor $i \geq 0$ en $j \geq 0$ en $i + j \leq n$.
Voor de overige waarden van $(i, j) \in S_X \times S_Y$ geldt $P(X = i \text{ en } Y = j) = 0$.
Deze simultane kansfunctie van X en Y noemen we ook wel de **trinomiale kansfunctie**.
(Vergelijk met de binomiale kansfunctie: er is hier sprake van 6 o.o. experimenten met **drie** mogelijke uitkomsten: een bijzonder geval van de multinomiale verdeling).
In onderstaande figuur hebben we de waarden (i, j), die het paar (X, Y) met positieve kans aanneemt, in de X-Y-vlak uitgezet:



De gevraagde kans kunnen we dus als volgt berekenen:

$$P(X > Y) = \sum_{i>j} P(X = i \text{ en } Y = j)$$

$$= \sum_{i=1}^{6} P(X = i \ \text{ en } Y = 0) + \sum_{i=2}^{5} P(X = i \ \text{ en } Y = 1) + \sum_{i=3}^{4} P(X = i \ \text{ en } Y = 2)$$

$$= \ 0.25956 \approx 26\%.$$

(De kansen $P(X = Y)$ en $P(X < Y)$ kunnen op soortgelijke wijze worden berekend).
Uit de trinomiale kansfunctie van X en Y zouden we de marginale kansfuncties van X en van Y met behulp van eigenschap 5.1.4 kunnen bepalen.
We kunnen echter beredeneren dat X $B(n, p_1)$-verdeeld is: X is immers het aantal winstpartijen van Timman bij n gespeelde partijen, die onafhankelijk van elkaar met kans p_1 tot winst voor Timman leiden. Evenzo kunnen we afleiden dat Y $B(n, p_2)$ -verdeeld is en dat het aantal remises $Z = n - X - Y$ $B(n, 1 - p_1 - p_2)$-verdeeld is. ■

De simultane kansfunctie $P(X = i \text{ en } Y = j)$ in voorbeeld 5.1.5 kunnen we dus ook opvatten als de simultane kansfunctie $P(X = i \text{ en } Y = j \text{ en } Z = k)$, die alléén positieve waarden aanneemt als $i + j + k = n$.
Definitie 5.1.2 en bijbehorende eigenschappen 5.1.3 en 5.1.4 kunnen we dan ook eenvoudig uitbreiden naar simultane kansfuncties $P(X_1 = x_1 \text{ en } X_2 = x_2 \text{ en} \ldots \text{en } X_n = x_n)$ van n (> 2) discrete stochastische variabelen. Uit die simultane kansfunctie kunnen we op soortgelijke wijze als in eigenschap 5.1.4 de marginale kansfunctie
van bijvoorbeeld X_1 afleiden:

$$P(X_1 = x_1) = \sum_{x_2 \in S_{X_2}} \dots \sum_{x_n \in S_{X_n}} P(X_1 = x_1 \text{ en } X_2 = x_2 \text{ en} \dots \text{en } X_n = x_n)$$

Voorbeeld 5.1.6 Een bedrijf produceert megachips in hoeveelheden van 20 stuks. Door een ontwerpfout blijken er 7 van elke hoeveelheid van 20 megachips niet te voldoen aan de specificaties. Van één hoeveelheid van 20 stuks zijn er reeds (willekeurig) in totaal 4 geleverd aan 4 verschillende bedrijven (genummerd 1 t/m 4).
Het al dan niet voldoen aan de specificaties kunnen we aangeven met een "turf"-variabele, die de waarde 1 (voldoet) of de waarde 0 (voldoet niet) aanneemt.
Voor elk van de 4 bedrijven hebben we dus zo'n turfvariabele. Voor bedrijf i geldt:

$$X_i = \begin{cases} 1 & \text{als de aan het bedijf } i \text{ geleverde chip voldoet} \\ 0 & \text{als dat niet het geval is} \end{cases} \qquad (\text{voor } i = 1, 2, 3, 4)$$

Verder geldt: $P(X_i = 1) = \frac{13}{20}$ en $P(X_i = 0) = \frac{7}{20}$, ongeacht de volgorde van levering (!)
Zie voor een motivatie hiervan o.m. opgave 2.3.
Elke X_i is dus $B\left(1, \frac{13}{20}\right)$ -verdeeld: we spreken ook wel van een **alternatieve verdeling met succeskans $p = \frac{13}{20}$**.
Uit deze kansverdelingen van de X_i's kunnen we de simultane kansen niet berekenen:
er geldt bijvoorbeeld **niet**, dat de kans dat alle 4 chips voldoen gelijk is aan $\left(\frac{13}{20}\right)^4$
Passen we namelijk de **productregel voor afhankelijke gebeurtenissen** (eigenschap 3.1.5) toe op de gebeurtenissen $\{X_1 = 1\}, \{X_2 = 1\}, \{X_3 = 1\}$ en $\{X_4 = 1\}$ dan vinden we:

$$\begin{aligned} &P(X_1 = 1 \text{ en } X_2 = 1 \text{ en } X_3 = 1 \text{ en } X_4 = 1) \\ &\quad = P(X_1 = 1) \cdot P(X_2 = 1 | X_1 = 1) \cdot P(X_3 = 1 | X_1 = X_2 = 1) \cdot P(X_4 = 1 | X_1 = X_2 = X_3 = 1) \\ &\quad = \frac{13}{20} \cdot \frac{12}{19} \cdot \frac{11}{18} \cdot \frac{10}{17} \end{aligned}$$

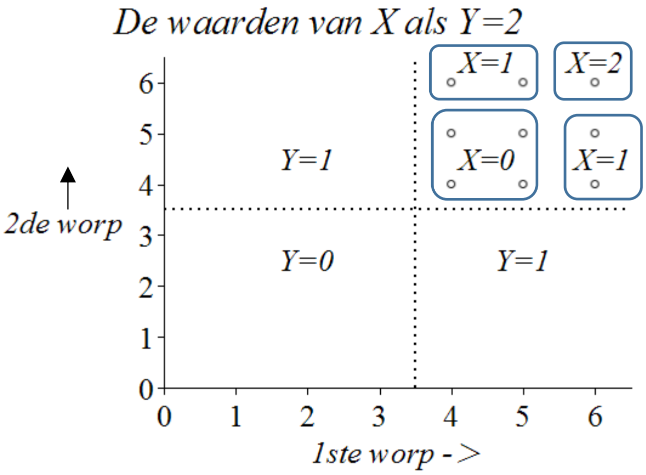
Op dezelfde wijze kunnen we de simultane kansfunctie $P(X_1 = x_1 \text{ en} \dots \text{en } X_4 = x_4)$ van X_1 t/m X_4 afleiden voor andere waarden van x_1 t/m x_4. ■

5.2 Voorwaardelijke verdelingen

Tot nu toe hebben we in dit hoofdstuk steeds simultane kansen berekend. Soms is echter de waarde van één van de variabelen bekend en vragen we ons af in hoeverre die wetenschap de kansverdeling van de overige variabelen beïnvloedt. Als in voorbeeld 5.1.6 bedrijf 1 meldt dat zijn megachip niet voldoet ($X_1 = 0$), heeft die informatie gevolgen voor de kansverdeling van X_2, X_3 en X_4.

Voorbeeld 5.2.1 Als we in voorbeeld 5.1.1 (2 keer werpen met een dobbelsteen) weten dat beide worpen groter dan 3 zijn ($Y = 2$) is dat gegeven van invloed op de "kans" dat er 0, 1 of 2 zessen zijn geworpen: we kunnen ons in dat geval beperken tot negen uitkomsten, zoals onderstaande figuur illustreert:

Het ligt voor de hand aan te nemen dat deze 9 uitkomsten weer een symmetrische kansruimte vormen. De kans dat er twee zessen geworpen worden, als gegeven is dat beide worpen groter dan 3 zijn, is dan blijkbaar $\frac{1}{9}$.



Deze voorwaardelijke kans op de gebeurtenis $\{X = 2\}$, gegeven de gebeurtenis $\{Y = 2\}$, noteren we met $P(X = 2|Y = 2)$.
Als we de definitie van voorwaardelijke kans gebruiken om de uitkomst $\frac{1}{9}$ te verifiëren, dan vinden we inderdaad:

$$P(X = 2|Y = 2) = \frac{P(X = 2 \text{ en } Y = 2)}{P(Y = 2)} = \frac{1/36}{9/36} = \frac{1}{9}$$

evenzo

$$P(X = 1|Y = 2) = \frac{4}{9}$$
$$P(X = 0|Y = 2) = \frac{4}{9}$$

Dit is een kansverdeling: X neemt de waarden 0, 1 en 2 aan met deze kansen, die gesommeerd 1 zijn, echter wel onder de voorwaarde dat beide worpen groter dan 3 zijn.
We spreken dan ook van de **voorwaardelijke verdeling van X, gegeven $Y = 2$**.
We kunnen bij deze verdeling ook een verwachtingswaarde bepalen door het gewogen gemiddelde van de waarden van X te bepalen en vinden zo:

$$0 \cdot \frac{4}{9} + 1 \cdot \frac{4}{9} + 2 \cdot \frac{1}{9} = \frac{2}{3}$$

Deze verwachtingswaarde kunnen we niet aanduiden met $E(X)$: hij geldt immers onder de voorwaarde $Y = 2$. Daarom noteren we $E(X|Y = 2) = \frac{2}{3}$.
Op gelijke wijze kunnen we uit de simultane en marginale verdelingen van X en Y de verdeling en verwachtingswaarde van X bepalen, als gegeven is dat $Y = 1$:

$$P(X = 0|Y = 1) = \frac{P(X = 0 \text{ en } Y = 1)}{P(Y = 1)} = \frac{2}{3}$$
$$P(X = 1|Y = 1) = \frac{P(X = 1\ en\ Y = 1)}{P(Y = 1)} = \frac{1}{3}$$
$$P(X = 2|Y = 1) = 0.$$

Dus $E(X|Y = 1) = \sum_{i=0}^{2} i \cdot P(X = i|Y = 1) = 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} + 2 \cdot 0 = \frac{1}{3}$.
In dit geval is het verwachte aantal zessen dus kleiner dan wanneer $Y = 2$.
Tot slot bekijken we $Y = 0$ als voorwaarde:

$$P(X = 0|Y = 0) = \frac{P(X = 0 \text{ en } Y = 0)}{P(Y = 0)} = \frac{9/36}{9/36} = 1$$

X kan alléén de waarde 0 aannemen als $Y = 0$, dus $E(X|Y = 0) = 0$ ∎

Algemeen definiëren we voor twee stochastische variabelen (die gedefinieerd zijn op dezelfde kansruimte):

Definitie 5.2.2 Als X en Y discrete stochastische variabelen zijn, dan wordt de **voorwaardelijke kansfunctie van $\boldsymbol{X}$ gegeven $\boldsymbol{Y = y}$** gedefinieerd door

$$\boldsymbol{P(X = x|Y = y) = \frac{P(X = x \text{ en } Y = y)}{P(Y = y)}} \text{ voor } x \in S_X.$$

Deze voorwaardelijke verdeling is alléén gedefinieerd voor elk element y van S_Y, waarvoor $P(Y = y) > 0$.

Voor vaste $y \in S_Y$ is $P(X = x|Y = y)$ dan inderdaad een kansfunctie, want:

1) $P(X = x|Y = y) = \frac{P(X=x \text{ en } Y=y)}{P(Y=y)} \geq 0$ voor $x \in S_X$ en
2) $\sum_{x \in S_X} P(X = x\,|Y = y) = \sum_{x \in S_X} \frac{P(X=x \text{ en } Y=y)}{P(Y=y)} = \frac{P(Y=y)}{P(Y=y)} = 1$

De verwachtingswaarde van deze verdeling noteren we met $E(X|Y = y)$ en berekenen we met de volgende eigenschap

Eigenschap 5.2.3 De **voorwaardelijke verwachtingswaarde van $\boldsymbol{X}$, gegeven $\boldsymbol{Y = y}$,** is

$$\boldsymbol{E(X|Y = y) = \sum_{x \in S_X} x \cdot P(X = x|Y = y).}$$

Uiteraard gelden ook de gebruikelijke eigenschappen voor verwachtingswaarde zoals:

$$E(aX + b|Y = y) = aE(X|Y = y) + b.$$

Voorbeeld 5.2.4 We komen terug op voorbeelden 5.1.1 en 5.2.1 over het twee keer werpen met een dobbelsteen. We hebben gezien dat het aantal worpen groter dan 3 (Y) het aantal worpen met als resultaat 6 (X) beïnvloedt. We vonden:

$$E(X) = \frac{1}{3} \quad \text{versus} \quad \begin{cases} E(X|Y = 0) = 0 \\ E(X|Y = 1) = \frac{1}{3} \\ E(X|Y = 2) = \frac{2}{3} \end{cases}$$

Wat is nu het verband tussen de onvoorwaardelijke verwachting $E(X)$ en de drie voorwaardelijke verwachtingen?

Welnu: als $Y = 0$, geldt $E(X|Y = 0) = 0$, met kans $P(Y = 0) = \frac{1}{4}$, . Evenzo treedt $E(X|Y = 1) = \frac{1}{3}$ op met kans $P(Y = 1) = \frac{1}{2}$ en $E(X|Y = 2) = \frac{2}{3}$ met kans $P(Y = 2) = \frac{1}{4}$.

Het gewogen gemiddelde van deze voorwaardelijke verwachtingen is dus:

$$\begin{aligned} &E(X|Y = 0) \cdot P(Y = 0) + E(X|Y = 1) \cdot P(Y = 1) + E(X|Y = 2) \cdot P(Y = 2) \\ &\quad = 0 \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{1}{2} + \frac{2}{3} \cdot \frac{1}{4} = \frac{1}{3} = E(X) \end{aligned}$$

We kunnen $E(X)$ dus opvatten als het gewogen gemiddelde van de voorwaardelijke verwachtingen $E(X|Y = y)$ met als wegingsfactor de kansen op de voorwaarden, $P(Y = y)$.■

Voorbeeld 5.2.4. laat zien dat we de numerieke waarden $E(X|Y=y)$ kunnen opvatten als de waarden die een stochastische variabele V kan aannemen, en wel zó dat:

$$P[V=E(X|Y=y)]=P(Y=y)$$

In het voorbeeld: $P\left(V=\frac{2}{3}\right)=\frac{1}{4}$, $P\left(V=\frac{1}{3}\right)=\frac{1}{2}$ en $P(V=0)=\frac{1}{4}$.
V wordt genoteerd als $E(X|Y)$.

Definitie 5.2.5 Als X en Y discrete stochastische variabelen zijn, dan is de **voorwaardelijke verwachting $E(X|Y)$ de stochastische variabele, die met kans $P(Y=y)$ de waarde $E(X|Y=y)$ aanneemt.**

Spreek uit: **"de voorwaardelijke verwachting van *X*, gegeven *Y*."**
Let op: $E(X|Y)$ is een s.v., terwijl $E(X|Y=y)$ een getal is voor een specifieke waarde van Y.

Met de verdeling van $E(X|Y)$ kunnen we de verwachting van die variabele berekenen:

$$E[\,E(X|Y)\,]=\sum_y E(X|Y=y)\cdot P(Y=y)$$

In voorbeeld 5.2.5 zagen we dat deze sommatie $E(X)$ oplevert. Dat is algemeen zo.
Om deze gelijkheid aan te tonen gaan we in het rechterlid de definities van $E(X|Y=y)$ en van $P(X=x|Y=y)$ toepassen:

$$\begin{aligned}
E[E(X|Y)] &= \textstyle\sum_y E(X|Y=y)\cdot P(Y=y)\\
&= \sum_y\left(\sum_x x\cdot P(X=x|Y=y)\right)\cdot P(Y=y)\\
&= \sum_x\sum_y x\cdot\frac{P(X=x \text{ en } Y=y)}{P(Y=y)}\cdot P(Y=y)\\
&= \sum_x x\cdot\sum_y P(X=x \text{ en } Y=y)=\sum_x xP(X=x)=E(X)
\end{aligned}$$

Daarmee hebben we bewezen:

Eigenschap 5.2.6 $\quad \boldsymbol{E[\,E(X|Y)\,]=E(X)}$

Deze regel kunnen we soms handig gebruiken om de verwachting van een variabele met een onbekende verdeling uit de (wel) bekende voorwaardelijke verdeling af te leiden.

Voorbeeld 5.2.7 We nemen aan dat het aantal fietsendiefstallen per dag in een stad Poisson-verdeeld is met een "gemiddelde" van 8 gestolen fietsen per dag. De ervaring leert dat slechts één op de twee diefstallen bij de politie gemeld wordt.
De stochastische variabelen X en N stellen resp. het aantal gemelde diefstallen en het werkelijke aantal diefstallen voor: dus $X\le N$. We weten dat $N\sim$ Poisson(8).
Als we aannemen dat er $N=10$ diefstallen zijn gepleegd en dat deze onafhankelijk van elkaar met kans ½ gemeld worden, is het aantal meldingen X, **gegeven *N* = 10,** dus $B\left(10,\frac{1}{2}\right)$-verdeeld.
Algemener, is ***X*, gegeven $\boldsymbol{N=n}$, $\boldsymbol{B\left(n,\frac{1}{2}\right)}$**-verdeeld, dus: $E(X|N=n)=\frac{1}{2}n$.
De stochastische variabele $E(X|N)$ neemt steeds waarden $E(X|N=n)=\frac{1}{2}n$ aan met kans

$P(N = n)$, dus $E(X|N)$ is een functie van N : $E(X|N) = \frac{1}{2}N$.
Maar dan geldt volgens eigenschap 5.2.6:

$$E(X) = E[E(X|N)] = E\left[\frac{1}{2}N\right] = \frac{1}{2}E(N) = \frac{1}{2} \cdot 8 = 4,$$

dus het verwachte aantal meldingen is de helft van het verwachte aantal diefstallen.
Op een dag krijgt de politie 6 meldingen binnen en men vraagt zich af wat de kansverdeling van het werkelijk aantal diefstallen is en wat het verwachte aantal diefstallen is, in het licht van deze 6 meldingen.
Bij 6 meldingen kunnen er $N = 6, 7, 8, \ldots$ diefstallen hebben plaatsgevonden:

$$P(N = n|X = 6) = \frac{P(X = 6 \text{ en } N = n)}{P(X = 6)} = \frac{P(X = 6|N = n) \cdot P(N = n)}{P(X = 6)}$$

In deze formule is

- X, gegeven $N = n$, $B\left(n, \frac{1}{2}\right)$-verdeeld,
- N Poisson-verdeeld met parameter $\mu = 8$ en
- $P(X = 6)$ te berekenen met behulp van de **wet van de totale kans**:

$$P(X = 6) = \sum_{n=6}^{\infty} P(X = 6 \text{ en } N = n) = \sum_{n=6}^{\infty} P(X = 6|N = n) \cdot P(N = n)$$

$$= \sum_{n=6}^{\infty} \binom{n}{6} \left(\frac{1}{2}\right)^n \cdot \frac{8^n e^{-8}}{n!}$$

$$= \frac{e^{-4}}{6!} \sum_{n=6}^{\infty} \frac{\left(\frac{1}{2} \cdot 8\right)^n e^{-4}}{(n-6)!}$$

$$\overset{n-6=k}{=} \frac{e^{-4}}{6!} \cdot 4^6 \sum_{k=0}^{\infty} \frac{4^k e^{-4}}{k!} = \frac{4^6 e^{-4}}{6!} \cdot 1$$

$$= \frac{4^6 e^{-4}}{6!}$$

Meer algemeen vinden we: $P(X = k) = \frac{4^k e^{-4}}{k!}$, voor $k = 0, 1, 2, \ldots$.
Dus X is Poisson-verdeeld met parameter $\mu = E(X) = 4$.
De eerder gevonden formule voor $P(N = n|X = 6)$ werken we dus uit met Bayes' regel:

$$P(N = n|X = 6) = \frac{P(X = 6|N = n) \cdot P(N = n)}{P(X = 6)} = \frac{\binom{n}{6}\left(\frac{1}{2}\right)^n \cdot \frac{8^n e^{-8}}{n!}}{\frac{4^6 e^{-4}}{6!}} = \frac{4^{n-6} e^{-4}}{(n-6)!},$$

voor $n = 6, 7, 8, \ldots$

We herkennen hierin een "verschoven" Poisson-verdeling: nemen we $j = n - 6$, dan geldt:

$$P(N - 6 = j|X = 6) = \frac{4^j \, e^{-4}}{j!} \text{ voor } j = 0, 1, 2, \ldots$$

$N - 6$ is, gegeven $X = 6$, dus Poisson-verdeeld met parameter $\mu = 4$.
Dus $E(N - 6|X = 6) = 4$ en dus $E(N\,|X = 6) = 6 + 4 = 10$.

Hoewel het verwachte aantal meldingen de helft is van het gegeven aantal diefstallen ($E(X|N = n) = \frac{1}{2}n$), is het verwachte aantal diefstallen niet het dubbele van een gegeven aantal meldingen, want $E(N|X = x) = x + 4$.
Dus $E(N|X) = X + 4$, dus (eig. 6.2.6): $E(N) = E[E(N|X)] = E(X + 4) = E(X) + 4 = 8$ ■

Opmerking: het valt buiten de stof van dit vak, maar naast de voorwaardelijke verwachting kunnen we uiteraard ook de voorwaardelijke variantie berekenen: $var(X|Y = y)$. Er geldt dan echter **niet** dat $var(X) = E[var(X|Y)]$, zoals eerder voor de verwachting wel gold.
In voorgaand voorbeeld geldt $var(X|N = n) = np(1 - p) = \frac{1}{4}n$, dus $var(X|N) = \frac{1}{4}N$.
Maar $E[var(X|N)] = E\left[\frac{1}{4}N\right] = \frac{1}{4}E(N) = \frac{1}{4} \cdot 8 = 2$, hetgeen niet gelijk is aan $var(X) = \mu = 4$, zoals we uit de Poisson-verdeling van X kunnen afleiden. ■

5.3 Onafhankelijke stochastische variabelen

Twee gebeurtenissen A en B zijn per definitie o.o. als $P(AB) = P(A) \cdot P(B)$.
In voorbeeld 5.1.1 zagen we dat deze gelijkheid niet altijd hoeft te gelden voor gebeurtenissen $\{X = i\}$ en $\{Y = j\}$ als X en Y discrete stochastische variabelen zijn.
In voorbeeld 5.1.6 hebben we gezien dat voor een viertal stochastische variabelen $\{X_1 = x_1\}$ t/m $\{X_4 = x_4\}$ afhankelijk kunnen zijn. In dergelijke gevallen noemen we de **stochastische variabelen** zelf ook **afhankelijk.**

Definitie 5.3.1 Twee discrete stochastische variabelen X en Y noemen we **onderling onafhankelijk (o.o.)** als

$$\boldsymbol{P(X = x \text{ en } Y = y) = P(X = x) \cdot P(Y = y)} \text{ voor } \textbf{alle paren } (\boldsymbol{x, y}) \in S_X \times S_Y$$

Deze definitie is eenvoudig te generaliseren voor een o.o. n-tal $X_1, \dots, X_n$:

$$P(X_1 = x_1 \text{ en} \dots \text{en } X_n = x_n) = P(X_1 = x_1) \cdot \dots \cdot P(X_n = x_n), \text{ voor alle } n\text{-tuples } (x_1, \dots, x_n).$$

We merken op dat, als X en Y o.o. zijn, ook gebeurtenissen als $\{X > a\}$ en $\{Y > b\}$ o.o. zijn, immers:

$$P(X > a \text{ en } Y > b) = \sum_{x>a} \sum_{y>b} P(X = x \text{ en } Y = y) = \sum_{x>a} \sum_{y>b} P(X = x) \cdot P(Y = y)$$
$$= \sum_{x>a} P(X = x) \cdot \sum_{y>b} P(Y = y) = P(X > a) \cdot P(Y > b)$$

Bovendien is bij onafhankelijkheid de voorwaardelijke verdeling van X, gegeven $Y = y$, gelijk aan de (onvoorwaardelijke) verdeling van X:

$$P(X = x|Y = y) = \frac{P(X = x \text{ en } Y = y)}{P(Y = y)} \overset{o.o.}{=} \frac{P(X = x) \cdot P(Y = y)}{P(Y = y)} = P(X = x)$$

(*Interpretatie: de wetenschap dat $Y = y$ beïnvloedt de kans op $X = x$ niet*)

Uit de definitie van onafhankelijkheid blijkt dat de simultane verdeling bij onafhankelijkheid vast ligt indien de marginale kansfuncties bekend zijn.
Voor afhankelijke stochastische variabelen geldt dit niet. De simultane verdeling legt de marginale verdelingen wel vast en daarmee kunnen we de onafhankelijkheid dus controleren.

$i \backslash j$	-1	0	1	$P(X=i)$
-1	$\frac{1}{9}$	$\frac{2}{9}$	0	$\frac{1}{3}$
0	0	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{3}$
1	$\frac{2}{9}$	0	$\frac{1}{9}$	$\frac{1}{3}$
$P(Y=j)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1

Voorbeeld 5.3.2 De simultane en marginale kansfuncties van X en Y worden gegeven in nevenstaande tabel.
Weliswaar geldt voor $i = -1$, 0 en 1:

$$P(X = i \text{ en } Y = i) = \frac{1}{9} = \frac{1}{3} \cdot \frac{1}{3}$$
$$= P(X = i) \cdot P(Y = i)$$

Maar voor $i \neq j$ geldt dit niet, bijvoorbeeld voor $i = -1$ en $j = 0$ geldt:

$$P(X = -1 \text{ en } Y = 0) = \frac{2}{9}$$
$$\neq \frac{1}{3} \cdot \frac{1}{3} = P(X = -1) \cdot P(Y = 0)$$

(voor de simultane kansen die 0 zijn is de ongelijkheid ook evident).
Conclusie: X en Y zijn afhankelijk.
Tevens kan worden nagegaan dat $E(X|Y = j) \neq E(X)$ voor $j = -1, 0, 1$. ■

In de meeste gevallen hoeven we de onafhankelijkheid van stochastische variabelen niet te bewijzen, maar volgt dit uit de onafhankelijkheid van de experimenten, waarop zij betrekking hebben. Simultane kansen volgen dan uit de marginale verdelingen, zoals in het volgende voorbeeld.

Voorbeeld 5.3.3 X en Y zijn respectievelijk het aantal acute blindedarmoperaties en het aantal niersteenoperaties op een dag in een streekziekenhuis. Gemiddeld zijn dat er 3 resp. 4. In verband met de operatieroosters wil de directie weten hoe groot de kans is dat er van elk van beide operaties minstens 5 spoedgevallen binnenkomen. Om deze vraag te kunnen beantwoorden maken we de volgende modelveronderstellingen: X en Y zijn o.o. (optreden van niersteen en blindedarmontsteking hebben niets met elkaar te maken) en beide zijn Poisson-verdeeld, met parameters 3 en 4. De gevraagde kans berekenen we nu als volgt, m.b.v. de tabel voor de Poisson-verdeling:

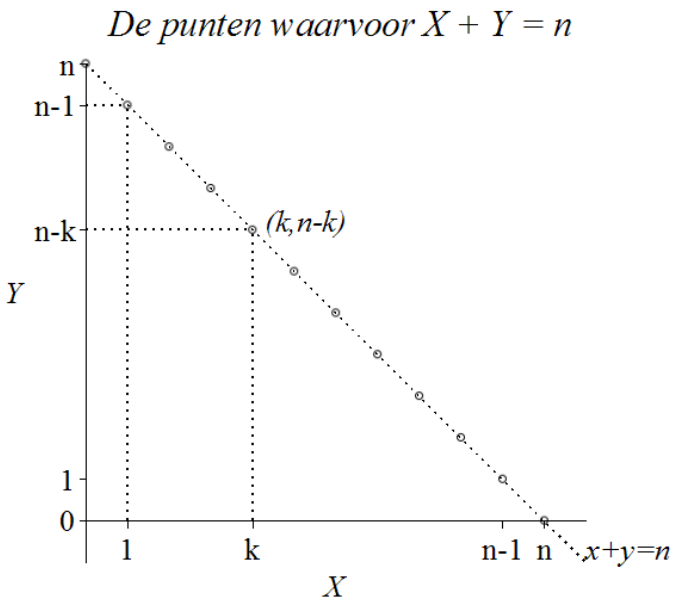
$P(X \geq 5 \text{ en } Y \geq 5) = P(X \geq 5) \cdot P(Y \geq 5)$, wegens onafhankelijkheid
$= (1 - 0.815) \cdot (1 - 0.629)$, m.b.v. de betreffende Poisson-tabellen
$\approx 6.9\%$ ■

5.4 Functies van discrete stochastische variabelen

In voorbeeld 5.3.3 waren we geïnteresseerd in de kans dat zowel X als Y een waarde van minstens 5 heeft. Dit kunnen we ook noteren als $P(min(X,Y) \geq 5)$.
Als we kansen van de vorm $P(min(X,Y) = k)$ voor meerdere waarden van k willen berekenen, zijn we in feite geïnteresseerd in de verdeling van $W = min(X,Y)$.

W is ook weer een stochastische variabele die het getal $min(x, y)$ realiseert als X en Y de waarden x en y aannemen. Een andere functie $g(X, Y)$ van X en Y die in voorbeeld 5.3.3 van belang kan zijn, is het totaal aantal operaties (van beide types) op een dag, $X + Y$.

Voorbeeld 5.4.1 X en Y zijn, net zoals in voorbeeld 5.3.3, o.o. en Poisson-verdeeld met verwachte waarden (parameters) 3 en 4. Het totaal aantal operaties op die dag is $Z = X + Y$. Omdat $S_X = S_Y = \{0, 1, 2, \dots\}$ is het waardenbereik van Z ook $S_Z = \{0, 1, 2, \dots\}$.



De kansfunctie $P(Z = n)$ van Z bepalen we nu door te bedenken dat de gebeurtenis $\{X + Y = k\}$ opgedeeld kan worden in de gebeurtenissen $\{X = 0 \text{ en } Y = n\}, \{X = 1 \text{ en } Y = n - 1\}, \dots, \{X = n \text{ en } Y = 0\}$, dus

$$P(Z = n) = \sum_{k=0}^{n} P(X = k \text{ en } Y = n - k)$$

$$\overset{o.o.}{=} \sum_{k=0}^{n} P(X = k) \cdot P(Y = n - k) \text{, wegens onafhankelijkheid van } X \text{ en } Y.$$

$$= \sum_{k=0}^{n} \frac{3^k\, e^{-3}}{k!} \cdot \frac{4^{n-k}\, e^{-4}}{(n-k)!} = e^{-7} \sum_{k=0}^{n} \frac{1}{k!\,(n-k)!} 3^k 4^{n-k} = \frac{e^{-7}}{n!} \sum_{k=0}^{n} \binom{n}{k} 3^k 4^{n-k}$$

Op deze sommatie passen we het Binomium van Newton toe: $\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = (a + b)^n$

$$P(Z = n) = \frac{e^{-7}}{n!} \cdot (3 + 4)^n = \frac{7^n e^{-7}}{n!} \text{, voor } n = 0, 1, 2, \dots$$

We herkennen in de kansfunctie van Z de Poisson-verdeling met parameter $\mu = 7$, dus alle mogelijke kansen m.b.t. Z kunnen we berekenen en we weten dat $E(Z) = var(Z) = 7$. We merken op dat blijkbaar $7 = E(X + Y) = E(X) + E(Y) = 3 + 4$. ■

Als X en Y waarden bereiken S_X en S_Y hebben, dan kunnen we de verdeling van $X + Y$ afleiden uit de simultane kansfunctie van X en Y.

$$P(X + Y = n) = \sum_{k \in S_X \text{ met } n - k \in S_Y} P(X = k \text{ en } Y = n - k)$$

Zijn X en Y bovendien o.o., dan geldt:

Eigenschap 5.4.2 (Convolutiesom)
Als X en Y discrete en o.o. s.v.-en zijn met een geheeltallig waardenbereik, dan geldt:

$$\boldsymbol{P(X+Y=n) = \sum_{k\in S_X} P(X=k)\cdot P(Y=n-k)}$$

De sommatie in het rechterlid wordt aangeduid met “Convolutiesom”.
Indien we de waarden van het paar (X,Y) uitzetten in een grafiek, zoals bij voorbeeld 5.4.1, dan sommeert de Convolutiesom over de roosterpunten $(k, n-k)$ die op de lijn $x+y=n$ liggen, voor zover de waardenbereiken S_X en S_Y dit toelaten.
$X+Y$ is een voorbeeld van een (reële) functie $g(X,Y)$ van twee stochastische variabelen. Ook van functies als $X\cdot Y$, $min(X,Y)$ of $max(X,Y)$ zullen we op soortgelijke wijze de verdeling wel eens bepalen. Met die verdeling kunnen we bijvoorbeeld $Eg(X,Y)$ afleiden. Analoog aan de berekening van $Eg(X)$ in hoofdstuk 4 kunnen we $Eg(X,Y)$ ook direct berekenen door de functiewaarden $g(x,y)$ te wegen met de simultane kansen $P(X=x \text{ en } Y=y)$, zonder bewijs:

Eigenschap 5.4.3 Voor een paar (X,Y) van discrete stochastische variabelen geldt:

$$\boldsymbol{Eg(X,Y) = \sum_{x\in S_X}\sum_{y\in S_Y} g(x,y)P(X=x \textbf{ en } Y=y)}$$

Voorbeeld 5.4.4 In voorbeeld 5.1.6 bekeken we een aantal aselecte trekkingen zonder terugleggen uit 20 megachips waarvan er 13 wél en 7 niet aan bepaalde specificaties voldoen. We beperken ons in dit voorbeeld tot de “turfvariabelen” (alternatieven) X_1 en X_2, die 1 of 0 zijn al naar gelang de eerste resp. de tweede gekozen megachip voldoet of niet.
X_1 en X_2 zijn beide $B\left(1,\frac{13}{20}\right)$-verdeeld, zodat $EX_1 = EX_2 = \frac{13}{20}$.
X_1 en X_2 zijn afhankelijk en hun simultane kansfunctie berekenen we daarom als volgt, met de productregel voor afhankelijke gebeurtenissen:

$$P(X_1=i \text{ en } X_2=j) = P(X_1=i)\cdot P(X_2=j|X_1=i)$$

bijvoorbeeld:

$$\begin{aligned} P(X_1=1 \text{ en } X_2=1) &= P(X_1=1)\cdot P(X_2=1|X_1=1) \\ &= \frac{13}{20}\cdot\frac{12}{19} \end{aligned}$$

Volgens eigenschap 5.4.3 geldt nu

$$\begin{aligned} E(X_1X_2) &= \sum_{i=0}^{1}\sum_{j=0}^{1} i\cdot j\cdot P(X_1=i \text{ en } X_2=j) \\ &= 1\cdot 1\cdot P(X_1=1 \text{ en } X_2=1) = \frac{13}{20}\cdot\frac{12}{19} \end{aligned}$$

En

$$\begin{aligned} E(X_1+X_2) &= \sum_{i=0}^{1}\sum_{j=0}^{1}(i+j)\cdot P(X_1=i \text{ en } X_2=j) \\ &= (0+0)\cdot P(X_1=0 \text{ en } X_2=0) + (0+1)\cdot P(X_1=0\ en\ X_2=1) \\ &\quad +(1+0)\cdot P(X_1=1 \text{ en } X_2=0) + (1+1)\cdot P(X_1=1 \text{ en } X_2=1) \end{aligned}$$

$$= 0 \cdot \frac{7}{20} \cdot \frac{6}{19} + 1 \cdot \frac{7}{20} \cdot \frac{13}{19} + 1 \cdot \frac{13}{20} \cdot \frac{7}{19} + 2 \cdot \frac{13}{20} \cdot \frac{12}{19} = \mathbf{\frac{26}{20}}$$

Omdat $EX_1 = EX_2 = \frac{13}{20}$, is blijkbaar in dit voorbeeld:

$$E(X_1 + X_2) = EX_1 + EX_2,$$

maar

$$\frac{13}{20} \cdot \frac{12}{19} = E(X_1 \cdot X_2) \neq EX_1 \cdot EX_2 = \left(\frac{13}{20}\right)^2 \qquad \blacksquare$$

De gelijkheid $E(X + Y) = EX + EY$ kunnen we ook algemeen bewijzen met behulp van eigenschap 5.4.3.

Eigenschap 5.4.5 Voor (discrete) stochastische variabelen X en Y geldt:

a. $\boldsymbol{E(X + Y) = E(X) + E(Y)}$.
b. **Als X en Y o.o. zijn, geldt: $\boldsymbol{E(XY) = E(X) \cdot E(Y)}$.**

Bewijs:

a. We merken op dat volgens eigenschap 5.4.3 voor $g(x, y) = x$ betekent:

$$E(X) = \sum_{x \in S_X} \sum_{y \in S_Y} x \cdot P(X = x \text{ en } Y = y)$$

Dus:
$$E(X + Y) = \sum_{x \in S_X} \sum_{y \in S_Y} (x + y) \cdot P(X = x \text{ en } Y = y)$$
$$= \sum_{x \in S_X} \sum_{y \in S_Y} x \cdot P(X = x \text{ en } Y = y) + \sum_{x \in S_X} \sum_{y \in S_Y} y \cdot P(X = x \text{ en } Y = y)$$
$$= E(X) + E(Y)$$

b. Leid deze gelijkheid zelf af m.b.v. eigenschap 5.4.3 en de onafhankelijkheid van X en Y. ■

Eigenschap 5.4.5b., toegepast op voorbeeld 5.4.1 (*X* en *Y* zijn o.o.!) geeft
$E(XY) = E(X) \cdot E(Y) = 3 \cdot 4 = 12$.
In dit zelfde voorbeeld toonden we aan dat de som van deze twee o.o. en Poisson-verdeelde stochastische variabelen ook weer Poisson-verdeeld is en wel zodanig dat
$E(X + Y) = E(X) + E(Y)$.
Een zelfde soort eigenschap zouden we willen afleiden voor het totaal aantal niersteenoperaties in een week, dus voor de som van 7 o.o. en Poisson-verdeelde stochastische variabelen. Daartoe kunnen we eigenschap 5.4.2 eenvoudig uitbreiden voor een n-tal $X_1, X_2, \ldots, X_n$.
Verder kunnen we eigenschap 5.4.5a. generaliseren (bewijs met inductie naar *n*).

Eigenschap 5.4.6 $\boldsymbol{E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n)}$.

Eigenschap 5.4.7 Als $X_1, X_2, \ldots, X_n$ o.o. zijn, dan zijn ook $\boldsymbol{g(X_1, \ldots, X_{n-1})}$ en $\boldsymbol{X_n}$ **o.o.**

Dus als $X_1, X_2, \ldots, X_n$ o.o. zijn, dan zijn ook $\sum_{i=1}^{n-1} X_i$ en X_n o.o.
Deze eigenschap kunnen we gebruiken om algemeen te bewijzen:

Eigenschap 5.4.8 Als $X_1, X_2, \ldots, X_n$ o.o. zijn en X_i is Poisson-verdeeld met parameter μ_i voor $i = 1, 2, \ldots, n$, dan is

$$\sum_{i=1}^{n} \boldsymbol{X_i} \textbf{ Poisson} - \textbf{verdeeld met parameter } \boldsymbol{\mu} = \sum_{i=1}^{n} \boldsymbol{\mu_i}$$

Voorbeeld 5.4.9 (de verwachting van de binomiale en de hypergeometrische verdeling)
Een populatie met N elementen bestaat voor het deel $p = \frac{R}{N}$ uit elementen met een bepaald kenmerk (succes) en voor het overige deel $1 - p$ ontbreekt dit kenmerk (mislukking). Op deze populatie voeren we n aselecte trekkingen uit. Voor zowel trekkingen **met** als **zonder** terugleggen kunnen we voor de i-de trekking een "turfvariabele" definiëren.

$$X_i = \begin{cases} 1 & \text{als de } i^{\text{de}} \text{ trekking een succes oplevert} \\ 0 & \text{als de } i^{\text{de}} \text{ trekking een mislukking oplevert} \end{cases} \quad (i = 1, 2, \ldots, n)$$

De marginale verdeling van X_i is voor beide gevallen (met en zonder terugleggen) $B(1, p)$, dus $E(X_i) = p$ en $var(X_i) = p(1 - p)$.
Het aantal successen X kan voor beide gevallen worden uitgedrukt in de X_i's: $X = \sum_{i=1}^{n} X_i$
Dus het verwachte aantal successen is volgens eigenschap 5.4.6

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = n \cdot p$$

We merken op dat het voor de berekening niet van belang is, dat in het ene geval $X_1, X_2, \ldots, X_n$ o.o. zijn en $\sum_{i=1}^{n} X_i$ dus $B(n, p)$-verdeeld is; en in het andere geval $X_1, X_2, \ldots, X_n$ afhankelijk zijn en $\sum_{i=1}^{n} X_i$ dus hypergeometrisch verdeeld is met parameters N, R en n. ■

In voorgaand voorbeeld hebben we binomiaal of hypergeometrisch verdeelde variabelen gemodelleerd als een som van turfvariabelen, die het optreden van succes of mislukking per trekking symboliseren. Deze aanpak zullen we in het vervolg vaker gebruiken o.m. om de formules van variantie af te leiden.

5.5 Correlatie

In voorbeeld 5.2.4 zagen we dat X, het aantal gemelde fietsendiefstallen op een dag in een stad, samenhang vertoont met Y, het aantal van werkelijk gepleegde diefstallen: $X \leq Y$. Grote waarden van X gaan samen met grote waarden van Y en kleine waarden van X zullen in de regel samengaan met kleine waarden van Y. Een dergelijke samenhang willen we "vangen" in één karakteristiek getal (maat voor de samenhang). We proberen dit te doen dit door de afwijkingen t.o.v. de verwachtingswaarden μ_X en μ_Y te beschouwen. De covariantie is gedefinieerd als "het gemiddelde product van de afwijkingen van X en van Y, als volgt:

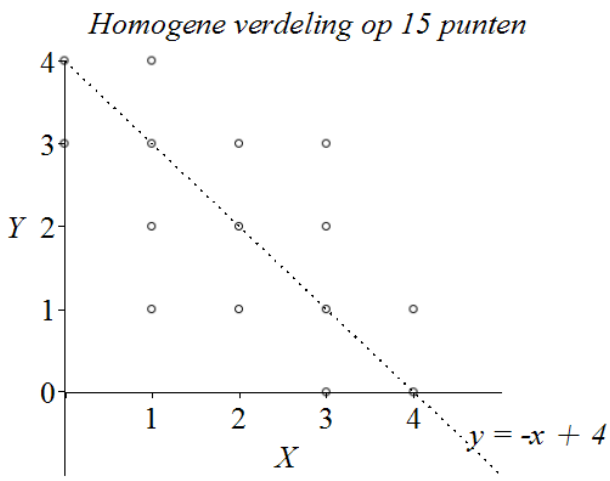
Definitie 5.5.1 De **covariantie** van twee stochastische variabelen X en Y wordt gedefinieerd door $\boldsymbol{cov(X, Y) = E(X - \mu_X)(Y - \mu_Y)}$

Volgens eigenschap 5.4.2 geldt, met $g(X,Y) = (X - \mu_X)(Y - \mu_Y)$, kan dit getal als een gewogen gemiddelde van producten (van afwijkingen) volgt berekend worden:

$$E(X - \mu_X)(Y - \mu_Y) = \sum_{x \in S_X} \sum_{y \in S_Y} (x - \mu_X)(y - \mu_Y) \cdot P(X = x \text{ en } Y = y)$$

Waarom het zo berekende getal informatie verschaft over de samenhang van X en Y moge blijken uit het volgende voorbeeld.

Voorbeeld 5.5.2 Het waardenbereik van (X,Y) bestaat uit 15 punten, die elk met kans $\frac{1}{15}$ optreden en gegroepeerd zijn rond de lijn $y = -x + 4$ liggen, zoals in onderstaande figuur is geschetst.



Een dergelijk verband tussen X en Y wordt wel gekenschetst worden als "negatief gecorreleerd": als X grote waarden aanneemt, neemt Y meest kleinere waarden aan en, omgekeerd als X kleine waarden aanneemt, neemt Y juist grotere waarden aan.
De lijn door de puntenwolk (die de afstanden tot de punten minimaliseert) heeft ook een negatieve richtingscoëfficiënt.
De marginale kansverdelingen van X en Y zijn hetzelfde (symmetrie!).

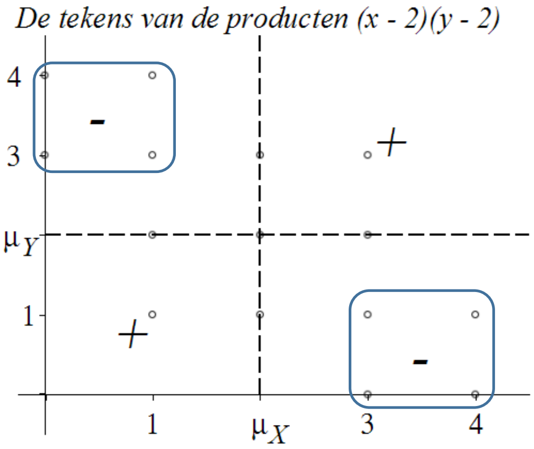
x	0	1	2	3	4	som
$P(Y = x) = P(X = x)$	$\frac{2}{15}$	$\frac{4}{15}$	$\frac{3}{15}$	$\frac{4}{15}$	$\frac{2}{15}$	1

Wederom wegens symmetrie kunnen we stellen dat $\mu_X = \mu_Y = 2$
X en Y zijn niet o.o. want bijv. $P(X = 4 \text{ en } Y = 4) = 0$ terwijl $P(X = 4)P(Y = 4) = \left(\frac{2}{15}\right)^2$.
We berekenen nu de $cov(X,Y)$ als gewogen gemiddelde van de producten $(x - \mu_X)(y - \mu_Y)$ met kansen $P(X = x \text{ en } Y = y) = \frac{1}{15}$, dus

$$cov(X,Y) = \sum_{(x,y)} (x - \mu_X)(y - \mu_Y) \cdot \frac{1}{15}$$

Delen we het platte vlak op in 4 delen met de lijnen $x = \mu_X$ en $y = \mu_Y$, dan zien we dat bijv.

- Voor $(x,y) = (3,3)$: $(x - \mu_X)(y - \mu_Y) = (3 - 2) \cdot (3 - 2) = +1$
- Voor $(x,y) = (0,4)$: $(x - \mu_X)(y - \mu_Y) = (0 - 2) \cdot (4 - 2) = -4$
- Voor $(x,y) = (3,1)$: $(x - \mu_X)(y - \mu_Y) = (3 - 2) \cdot (1 - 2) = -1$



Blijkbaar zal de uitkomst van $cov(X, Y)$ negatief zijn, want er zijn 8 punten met een negatief product en slechts 2 met een positief product, die in absolute waarde ook nog eens klein zijn.

Berekening levert inderdaad op dat:

$$cov(X, Y) = -\frac{16}{15}$$

Dus X en Y zijn negatief gecorreleerd zijn. ■

We zien in voorbeeld 5.5.2 dat de covariantie negatief is, indien de punten gegroepeerd zijn in de buurt van een lijn met negatieve richtingscoëfficiënt.
Op analoge wijze, bijvoorbeeld door in voorbeeld 5.5.1 Y te vervangen door $Z = 4 - Y$ (de grafiek van voorbeeld 5.5.2 wordt dan gespiegeld t.o.v. de lijn $y = 2$), zien we dat de covariantie positief is als de punten uit het waardenbereik “met grote kans in de buurt” van een lijn met positieve richtingscoëfficiënt liggen.
In voorbeeld 5.2.4 over de fietsendiefstallen verwachten we dus een positieve covariantie tussen meldingen en diefstallen: $cov(X, Y) = 4$ (zie de opgaven).
We tonen nu aan dat bij onafhankelijkheid deze maat voor samenhang inderdaad 0 is. Als $cov(X, Y) = 0$ noemen we X en Y **ongecorreleerd**. In de volgende eigenschap geven we bovendien een rekenformule waarmee we i.h.a. de covariantie makkelijker kunnen berekenen.

Eigenschap 5.5.3 **a.** $\boldsymbol{cov(X, Y) = E(XY) - \mu_X\mu_Y}$.
b. Als $\boldsymbol{X}$ **en** $\boldsymbol{Y}$ **o.o.** zijn dan geldt: $\boldsymbol{cov(X, Y) = 0}$.

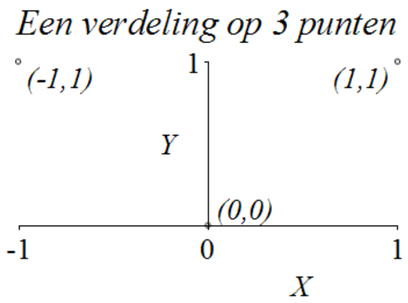
Bewijs:

a. $cov(X, Y) = E(X - \mu_X)(Y - \mu_Y)$
$= E(XY - X \cdot \mu_Y - \mu_X \cdot Y + \mu_X\mu_Y)$
$= E(XY) - \mu_Y \cdot E(X) - \mu_X \cdot E(Y) + \mu_X\mu_Y$ (eigenschappen 4.4.3 en 5.4.5)
$= E(XY) - \mu_X\mu_Y$

b. Volgt uit a. en uit $E(XY) = E(X) \cdot E(Y)$ (eigenschap 5.4.5b) ■

De bewering in eigenschap mogen we niet omkeren: ongecorreleerdheid impliceert geen onafhankelijkheid, zoals blijkt uit het volgende tegenvoorbeeld.

Voorbeeld 5.5.4
De simultane verdeling van X en Y worden gegeven door de drie kansen:
$P(X = -1 \text{ en } Y = 1) = P(X = 0 \text{ en } Y = 0) = P(X = 1 \text{ en } Y = 1) = \frac{1}{3}$



De marginale verdelingen zijn eenvoudig af te leiden:

i	-1	0	1
$P(X = i)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$P(Y = i)$	0	$\frac{1}{3}$	$\frac{2}{3}$

Dus $\mu_X = 0$ (symmetrie) en $\mu_Y = \sum_i i \cdot P(Y = i) = 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3}$
X en Y zijn afhankelijk: bijv. $P(X = 0 \text{ en } Y = 1) = 0$, maar $P(X = 0) \cdot P\,(Y = 1) = \frac{1}{3} \cdot \frac{2}{3}$
(intuïtief: als $Y = 1$ kan X alléén -1 of 1 zijn en als $Y = 0$ kan X alléén 0 zijn, dus afhankelijk)
$E(XY) = -1 \cdot 1 \cdot \frac{1}{3} + 0 \cdot 0 \cdot \frac{1}{3} + 1 \cdot 1 \cdot \frac{1}{3} - 0$, dus volgens eigenschap 5.5.3a is

$$cov(X, Y) = E(XY) - \mu_X \mu_Y = 0 - 0 \cdot \frac{2}{3} = 0$$

X en Y zijn dus **niet gecorreleerd, maar wel afhankelijk.** ∎

De reden, dat afhankelijke stochastische variabelen toch ongecorreleerd kunnen zijn, is gelegen in het feit, dat het "gemiddelde" van het product van de "afwijkingen" $X - \mu_X$ en $Y - \mu_Y$ ook bij afhankelijke stochastische variabelen nul kan zijn.
Dit "gemiddelde" blijkt vooral af te wijken van nul als de punten in het waardenbereik van (X, Y) "in de buurt van" of op een rechte lijn liggen. In voorbeeld 5.5.4 is dit duidelijk niet het geval: er is geen sprake van **lineaire samenhang** ("welke lijn ligt het dichtst in de buurt van deze drie punten?"). De samenhang is eerder van de vorm $Y = |X|$ of $Y = X^2$, terwijl $cov(X, Y)$ slechts lineaire samenhang "meet".

We kunnen ons afvragen in hoeverre de grootte van $cov(X, Y)$ iets zegt over de mate van onafhankelijkheid. Dit is **niet** het geval zoals blijkt uit onderdeel c van de volgende eigenschap.

Eigenschap 5.5.5 (Eigenschappen van covariantie)

- **a.** $\boldsymbol{cov(X, X) = var(X)}$.
- **b.** $\boldsymbol{cov(X, Y) = cov(Y, X)}$.
- **c.** $\boldsymbol{cov(aX + b, Y) = a \cdot cov(X, Y)}$**,** voor $a \in \mathbb{R}$ en $b \in \mathbb{R}$.
- **d.** $\boldsymbol{cov(X + Y, Z) = cov(X, Z) + cov(Y, Z)}$**.**
- **e.** $\boldsymbol{var(X + Y) = var(X) + var(Y) + 2cov(X, Y)}$ en
  $\boldsymbol{var(X - Y) = var(X) + var(Y) - 2cov(X, Y)}$

Bewijs:

a. $cov(X, X) = E(X - \mu_X)(X - \mu_X) = E(X - \mu_X)^2 = var(X)$.

b. Volgt uit symmetrie van de definitie van $cov(X, Y)$.

c.
$$\begin{aligned} cov(aX + b, Y) &= E[\big(aX + b - E(aX + b)\big) \cdot (Y - \mu_Y)] \\ &= E[a \cdot (X - \mu_X) \cdot (Y - \mu_Y)] \\ &= a \cdot E(X - \mu_X)(Y - \mu_Y) \\ &= a \cdot cov(X, Y) \end{aligned}$$

d. Gebruik de definitie van covariantie om dit te bewijzen.

e. $var(X + Y) = cov(X + Y, X + Y)$, volgens a.
$= cov(X, X) + cov(X, Y) + cov(Y, X) + cov(Y, Y)$, volgens d. en b.
$= var(X) + 2 \cdot cov(X, Y) + var(Y)$, volgens a. en b.

Dus $var(X - Y) = var\big(X + (-Y)\big)$
$= var(X) + var(-Y) + 2cov(X, -Y)$
$= var(X) + var(Y) - 2cov(X, Y)$,
omdat $var(-Y) = (-1)^2 var(X) = var(Y)$ ∎

Eigenschap 5.5.5c houdt in dat $cov(X, Y)$ **afhangt** van de gekozen **eenheid**.
Als X gedefinieerd is in meters en we gaan over op de eenheid centimeter (X wordt $100X$), dan wordt de covariantie een factor 100 groter: $cov(100X, Y) = 100 \cdot cov(X, Y)$.

Een **eenheidsonafhankelijke maat voor lineaire samenhang** verkrijgen we door de covariantie te delen door de standaardafwijkingen van X en Y (als $\sigma_X > 0$ en $\sigma_Y > 0$).

Definitie 5.5.6 De **correlatiecoëfficiënt** $\rho(X,Y)$ van twee stochastische variabelen X en Y wordt gedefinieerd door

$$\boldsymbol{\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}}$$

.

Voorbeeld 5.5.7 Bij een grootscheeps onderzoek naar het verband tussen fluortandpasta en het optreden van gaatjes bij de halfjaarlijkse controle door de tandarts, definieerde men $X = 1$ of 0 al naar gelang er wel of niet fluortandpasta wordt gebruikt en $Y = 1$ of 0 als er gaatjes waren of niet. Men bepaalde (schattingen van) de kansen $P(X = i \text{ en } Y = j)$:

i \ j	0	1
0	$\frac{1}{9}$	$\frac{1}{3}$
1	$\frac{1}{3}$	$\frac{2}{9}$

Hoewel het hier om schattingen gaat, zullen we aannemen dat, vanwege de omvang van het onderzoek, de kansen nauwelijks afwijken van de werkelijke waarden.
De kansfuncties van X en Y zijn blijkbaar gelijk (wegens symmetrie) en:
$P(X = 1) = \frac{5}{9}$ en $P(X = 0) = \frac{4}{9}$, zodat $\mu_X = \mu_Y = \frac{5}{9}$ en $\sigma_X = \sigma_Y = \sqrt{\frac{20}{81}}$

Voor het berekenen van $cov(X,Y) = E(XY) - \mu_X\mu_Y$ hebben we de waarde van $E(XY)$ nodig:

$$E(XY) = \sum_x \sum_y x \cdot y \cdot P(X = x \text{ en } Y = y) = 1 \cdot 1 \cdot \frac{2}{9} = \frac{2}{9}$$

De correlatiecoëfficiënt kunnen we dus als volgt berekenen:

$$\rho(X,Y) = \frac{E(XY) - \mu_X\mu_Y}{\sigma_X\sigma_Y} = \frac{\frac{2}{9} - \frac{5}{9} \cdot \frac{5}{9}}{\sqrt{\frac{20}{81}} \cdot \sqrt{\frac{20}{81}}} = -\frac{7}{20}$$

Covariantie en correlatiecoëfficiënt hebben hetzelfde teken: bij een negatieve waarde van ρ kunnen we spreken van een negatieve correlatie van X en Y: gebruik van fluoride ($X = 1$) gaat wat meer samen met "geen gaatjes" ($Y = 0$), en omgekeerd. ■

Als de correlatiecoëfficiënt van X en Y niet gelijk aan 0 is, zoals in bovenstaand voorbeeld, kunnen we concluderen dat er afhankelijkheid bestaat tussen beide fenomenen "fluorgebruik" en "gaatjes". Daaruit mogen we niet zo maar een **oorzakelijk verband** leggen tussen fluorgebruik en het (niet) optreden van gaatjes.
Zo kan het best zijn dat mensen die bewuster met gebitshygiëne omgaan en dus vaker hun tanden poetsen, ook vaker kiezen voor fluortandpasta vanwege vermeende voordelen. Minder gaatjes kan dan veroorzaakt worden door het vaker poetsen en niet of in mindere mate door fluorgebruik.
Beroemd is in dit verband de positieve correlatie tussen het afnemende aantal geboorten en het afnemend aantal ooievaars in Nederland gedurende de jaren zeventig. Zeker na de seksuele revolutie in de jaren '60 zag men in dat hier geen sprake is van een oorzakelijk verband….

De mate van samenhang kan men uit de (absolute) waarde van $\rho(X,Y)$ afleiden: $\rho(X,Y)$ is genormeerd en bereikt zijn uiterste waarden $+1$ en -1 als er sprake is van een strikt lineair verband van X en Y , zoals uit het volgende blijkt.

Eigenschap 5.5.8 (Eigenschappen van een correlatiecoëfficiënt)

a. $\boldsymbol{\rho(aX+b,Y)} = \begin{cases} \boldsymbol{\rho(X,Y)} & \text{als } a>0 \\ -\boldsymbol{\rho(X,Y)} & \text{als } a<0 \end{cases}$

b. $\boldsymbol{-1 \le \rho(X,Y) \le 1}$

c. Als $Y = aX+b$, dan $\boldsymbol{\rho(X,Y)} = \begin{cases} \boldsymbol{+1} & \text{als } a>0 \\ \boldsymbol{-1} & \text{als } a<0 \end{cases}$

En omgekeerd, als $\rho(X,Y)=1$, geldt $Y=aX+b$, met $a>0$
en als $\rho(X,Y)=-1$, geldt $Y=aX+b$, met $a<0$

Bewijs:

a. $cov(aX+b,Y) = a\cdot cov(X,Y)$ volgens eigenschap 5.5.5c en $var(aX+b) = a^2 var(X)$ volgens eigenschap 4.4.9, dus: $\rho(aX+b,Y) = \frac{cov(aX+b,Y)}{\sqrt{var(aX+b)\cdot var(Y)}} = \frac{a cov(X,Y)}{|a|\cdot\sigma_X\cdot\sigma_Y} = \frac{a}{|a|}\rho(X,Y)$

b. Om dit te bewijzen, gebruiken we dat $var\left(\frac{X}{\sigma_X}\right) = \left(\frac{1}{\sigma_X}\right)^2 var(X) = 1$ en eigenschap 5.5.5e voor $\frac{X}{\sigma_X}+\frac{Y}{\sigma_Y}$:

$$var\left(\frac{X}{\sigma_X}+\frac{Y}{\sigma_Y}\right) = var\left(\frac{X}{\sigma_X}\right) + var\left(\frac{Y}{\sigma_Y}\right) + 2cov\left(\frac{X}{\sigma_X},\frac{Y}{\sigma_Y}\right) = 1+1+2\cdot\rho(X,Y)$$

Omdat $var\left(\frac{X}{\sigma_X}+\frac{Y}{\sigma_Y}\right) \ge 0$, geldt $2+2\rho(X,Y)\ge 0$, dus $\rho(X,Y)\ge -1$

Evenzo volgt uit $var\left(\frac{X}{\sigma_X}-\frac{Y}{\sigma_Y}\right) \ge 0$, dat $\rho(X,Y)\le 1$.

c. Met $Y=aX+b$ wordt bedoeld: $P(Y=aX+b)=1$. De paren van waarden (x,y) die met positieve kans voorkomen, liggen op de lijn. Zie verder de opgave 11. ■

De mate van samenhang (correlatie) van X en Y wordt i.h.a. als volgt geclassificeerd:

- $\rho = 0$: *ongecorreleerd*
- $\rho > 0$: *positief gecorreleerd* en $\rho < 0$: *negatief gecorreleerd*
- $|\rho| = 1$: *strikt lineair verband*
- $0.9 \le |\rho| < 1$: *sterk gecorreleerd*
- $0 < |\rho| < 0.9$: *licht* of *matig gecorreleerd (evt.* ≤ 0.3 *en* ≥ 0.3 *onderscheiden).*

In voorbeeld 5.5.7 vonden we $\rho(X,Y) = -0.35$: X en Y zijn matig negatief gecorreleerd.

In eigenschap 5.5.5e hebben we reeds aangetoond dat de variantie van $X+Y$ i.h.a. niet gelijk is aan de som van $var(X)$ en $var(Y)$. Dit geldt wèl als X en Y o.o. zijn, omdat dan de covariantie 0 is. Deze eigenschappen kunnen we generaliseren voor een n-tal stochastische variabelen (we geven het bewijs niet).

Eigenschap 5.5.9

$$\textbf{a. } \boldsymbol{var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i) + \sum_{i\neq}\sum_{j} cov(X_i,X_j)}$$

$$\textbf{b. Als } \boldsymbol{X_1,\dots,X_n} \textbf{ o.o. zijn, geldt: } \boldsymbol{var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i)}$$

Opmerking: deze eigenschappen voor de variantie zijn aanzienlijk eenvoudiger dan wanneer we voor $E|X-\mu|$ hadden gekozen als maat voor de spreiding.
De $n \times n$ termen in het rechterlid van a. worden vaak in een zgn. covariantiematrix gezet, met op de hoofddiagonaal de varianties ($= cov(X_i, X_i)$)

$$\begin{bmatrix} var(X_1) & cov(X_1,X_2) & \cdot\ \cdot\ \cdot\ \cdot\ \cdot & cov(X_1,X_n) \\ cov(X_2,X_1) & var(X_2) & & cov(X_2,X_n) \\ \cdot & & \ddots & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ cov(X_n,X_1) & cov(X_n,X_2) & \cdot\ \cdot\ \cdot\ \cdot\ \cdot & var(X_n) \end{bmatrix}$$

■

We passen nu eigenschap 5.5.9 toe om de variantie van de binomiale en de hypergeometrische verdeling af te leiden (paragraaf 4.5).
Bij onafhankelijke experimenten, zoals bij trekkingen met terugleggen uit een dichotome populatie, zagen we in voorbeeld 5.4.9 reeds dat bij n Bernoulli-experimenten met succeskans p evenzoveel o.o. “turfvariabelen” of alternatieven X_i kunnen worden gedefinieerd, die de waarde 1 bij succes en 0 bij mislukking aannemen. Het aantal successen wordt gegeven door

$$X = \sum_{i=1}^{n} X_i,$$

een $B(n,p)$-verdeelde grootheid, waarvoor $E(X) = \sum_{i=1}^{n} E(X_i) = np.$
Wegens de onafhankelijkheid van X_i’s geldt (eigenschap 5.5.9b):

$$var(X) = var\left(\sum_{i=1}^{n} X_i\right) \overset{\text{o.o.}}{=} \sum_{i=1}^{n} var(X_i) = n \cdot p(1-p)$$

(gebruikmakend van de verdeling van X_i: $var(X_i) = E(X_i^2) - (EX_i)^2 = p - p^2$)

Bij n trekkingen **zonder terugleggen** uit een dichotome populatie met R “successen” en $N - R$ “mislukkingen” kunnen we voor elke trekking i weer een “turf”-variabele X_i definiëren, die waarde 1 en 0 aannemen bij resp. succes en mislukking: alle X_i’s blijven $B(1,p)$-verdeeld, met $p = \frac{R}{N}$. Maar de X_i’s zijn wel **afhankelijk**, waardoor we eigenschap 5.5.9a moeten gebruiken om de variantie van het aantal successen $X = \sum_{i=1}^{n} X_i$ af te leiden.

$$var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i) + \sum_{i\ \neq}\sum_{j} cov\left(X_i, X_j\right)$$

In deze uitdrukking zijn alle varianties gelijk $var(X_i) = p(1-p) = \frac{R}{N} \cdot \left(1 - \frac{R}{N}\right)$
Uit symmetrie-overwegingen zijn ook alle covariantietermen ($n^2 - n$ in aantal) gelijk.
Het is dus voldoende om één covariantieterm te bepalen:

$$cov(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2)$$

waarin

$E(X_1) = E(X_2) = \frac{R}{N}$ en

$E(X_1X_2) = \sum\sum i \cdot j \cdot P(X_1 = i \text{ en } X_2 = j) = 1 \cdot 1 \cdot P(X_1 = 1 \text{ en } X_2 = 1)$

$\qquad = P(X_1 = 1) \cdot P(X_2 = 1 | X_1 = 1) = \frac{R}{N} \cdot \frac{R-1}{N-1}$ (de kans op "eerste 2 keer succes")

Deze resultaten invullend vinden we:

$$var(X) = var\left(\sum_{i=1}^{n} X_i\right) = n \cdot var(X_1) + (n^2 - n)cov(X_1, X_2)$$

$$= n \cdot \frac{R}{N}\left(1 - \frac{R}{N}\right) + (n^2 - n)\left[\frac{R}{N} \cdot \frac{R-1}{N-1} - \frac{R}{N} \cdot \frac{R}{N}\right]$$

$$= \cdots = n \cdot \frac{R}{N}\left(1 - \frac{R}{N}\right) \cdot \frac{N-n}{N-1}$$

Zoals reeds opgemerkt in hoofdstuk 4 is met $p = \frac{R}{N}$ de variantie van hypergeometrische verdeling op een factor $\frac{N-n}{N-1}$ na gelijk aan die van de $B(n, p)$-verdeling.
Deze factor wordt wel de **correctiefactor voor eindige populaties** genoemd: hij tendeert naar 1 bij zeer grote populaties ($N \to \infty$). Dan wordt de hypergeometrische ook goed door de binomiale verdeling benaderd (eigenschap 4.5.5).

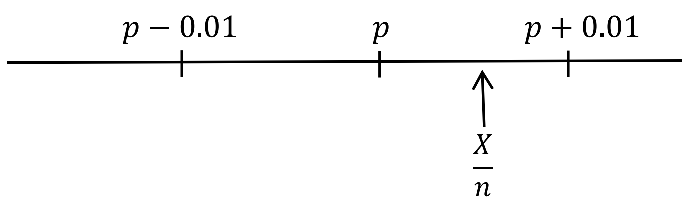
5.6 De zwakke wet van de grote aantallen

Voorbeeld 5.6.1 Bij de massaproductie van bijv. chips, weerstanden, sensoren wordt de kwaliteitscontrole veelal verricht door een aselecte steekproef uit de productie te nemen en de fractie van afgekeurde producten te bepalen. De **experimentele wet** van de grote aantallen (hoofdstuk 1) zegt ons dat de steekproeffractie "op den duur" de kans p op een afgekeurd product zeer dicht nadert. Dit is een ervaringsfeit en wat "zeer dicht bij p" is en hoe groot de steekproef moet zijn, is daarmee niet beantwoord. Inmiddels hebben we kansmodellen ontwikkeld waarmee we dit soort situaties goed kunnen beschrijven. In dit geval definiëren we de stochastische variabele X als "het aantal afgekeurde producten bij een aselecte steekproef van n producten". De $B(n, p)$-verdeling voor X is een correcte keuze als er uit de productie met terugleggen wordt getrokken, maar ook een goede benaderende verdeling bij trekken zonder terugleggen, daar de totale productie zeer omvangrijk is.
Het frequentiequotiënt wordt nu gegeven door $\frac{X}{n}$ die in verwachtingswaarde (dus "gemiddeld genomen") inderdaad gelijk is aan p en waarvan de spreiding afneemt met toenemende n, want:

- $E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n} \cdot np = p$
- $var\left(\frac{X}{n}\right) = \left(\frac{1}{n}\right)^2 var(X) = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n}$

Hoe groot moet n nu zijn, zodat we met een kans van minstens 90% een waarde voor $\frac{X}{n}$ vinden die minder dan 1% afwijkt van de werkelijke waarde van p?

Geschetst:



$$\text{Dus: } P\left(\left|\frac{X}{n}-p\right|<0.01\right)\geq 0.90$$

Ofwel: $P\left(\left|\frac{X}{n}-p\right|\geq 0.01\right)\leq 0.10$

Volgens de regel van Chebyshev geldt: $P\left(\left|\frac{X}{n}-p\right|\geq 0.01\right)\leq\frac{var\left(\frac{X}{n}\right)}{0.01^2}$

Dus aan het vereiste is voldaan als $\frac{var\left(\frac{X}{n}\right)}{0.01^2}=\frac{p(1-p)}{0.01^2 n}\leq 0.10$, dus als $n\geq 100000\cdot p(1-p)$

Omdat $f(p)=p(1-p)$ maximaal $\frac{1}{4}$ is voor $0\leq p\leq 1$, vinden we $n\geq 25000$.

Iets algemener gesteld geldt voor elk (klein) interval van de vorm $(p-c,p+c)$ volgens Chebyshev:

$$P\left(\left|\frac{X}{n}-p\right|\geq c\right)\leq\frac{p(1-p)}{c^2 n}$$

We kunnen dus altijd een steekproefomvang vinden om de kans, dat $\frac{X}{n}$ in bepaalde mate afwijkt van p, zo klein te maken als wenselijk ■

We merken op dat $\frac{X}{n}$ ook als een gemiddelde kan worden opgevat: een $B(n,p)$-verdeelde variabele kunnen we immers ook opvatten al een som van o.o. alternatieven $X_1,\ldots,X_n$.
Dus $\frac{X}{n}=\frac{\sum_{i=1}^{n}X_i}{n}=\frac{1}{n}\sum_{i=1}^{n}X_i$, die we ook als $\overline{X}$ of $\overline{X}_n$ noteren.

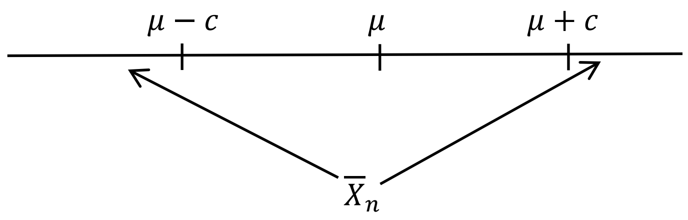
Op een dergelijk gemiddelde van o.o. en gelijk verdeelde stochastische variabelen met verwachtingswaarde μ (in het voorbeeld p), is de volgende eigenschap van toepassing.

Eigenschap 5.6.2 De zwakke wet van de grote aantallen

Indien $X_1,X_2,\ldots$ o.o. zijn en allen dezelfde verdeling hebben met verwachtingswaarde μ en variantie σ^2, dan geldt voor het gemiddelde $\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$ en elke constante $c>0$

$$\lim_{n\to\infty}P(|\overline{X}_n-\mu|\geq c)=0$$

In een schets:



Bewijs: Het bewijs maakt gebruik van de ongelijkheid van Chebyshev, waarin we voor X nu $\overline{X}_n$ kiezen. Voor $\overline{X}_n$ geldt:

$$E(\overline{X}_n)=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}\cdot n\cdot\mu=\mu \quad \text{en}$$

$$var(\overline{X}_n) = var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 var\left(\sum_{i=1}^{n} X_i\right) \overset{\text{o.o.}}{=} \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} var(X_i) = \frac{1}{n^2}\cdot n\cdot\sigma^2 = \frac{\sigma^2}{n}$$

Chebyshev: $P\left(\left|\overline{X}_n - \mu\right| \geq c\right) \leq \dfrac{var(\overline{X}_n)}{c^2} = \dfrac{\sigma^2}{nc^2}$

Dus $\lim\limits_{n\to\infty} P\left(\left|\overline{X}_n - \mu\right| \geq c\right) = 0$ ■

We zeggen ook wel: $\overline{X}_n$ **convergeert in kans** naar μ.

Toegepast op voorbeeld 5.6.1 zien we dat het frequentiequotiënt $f_n(A) = \frac{X}{n} = \overline{X}_n$ van de gebeurtenis A = "product voldoet niet" in kans convergeert naar de onbekende p.
De zwakke wet van de grote aantallen geeft dus in die zin een theoretische bevestiging voor het ervaringsfeit dat $f_n(A)$ "convergeert" naar p (de experimentele wet van de grote aantallen, paragraaf 1.3).

5.7 Vraagstukken

1. We gooien viermaal met een zuivere munt. Zij X het totaal aantal malen kruis en Y het aantal malen kruis in de laatste twee worpen.

- **a.** Bepaal de simultane kansfunctie van X en Y.
  Maak daartoe een lijst van alle 2^4 uitkomsten van de 4 worpen en noteer bij elke uitkomst de waarde van zowel X als Y.
- **b.** Bepaal de kansfunctie van X, gegeven $Y = 1$ en $E(X|Y = 1)$.
- **c.** Bepaal $P(Y = 1|X = 3)$.

2. De simultane kansfunctie van X en Y wordt gegeven door

$$P(X = i \text{ en } Y = j) = \left(\frac{1}{3}\right)^{i-j}\left(\frac{2}{3}\right)^{1+j}, \qquad \text{voor } i = 1,2,3,\ldots \text{ en } j = 0,1$$

- **a.** Schets het waardenbereik van X en Y met roosterpunten in het xy-vlak.
- **b.** Laat zien dat de marginale verdeling van X geometrisch is en geef $E(X)$.
- **c.** Bepaal de marginale kansfunctie van Y, dus $P(Y = 0)$ en $P(Y = 1)$.
- **d.** Zijn X en Y o.o.?

3. Een kleine fabriek werkt met een ochtendploeg en een avondploeg. Definieer voor een willekeurige werkdag de stochastische variabelen X en Y als volgt:
X = het aantal afwezigen bij de ochtendploeg
Y = het aantal afwezigen bij de avondploeg

Op basis van jarenlange statistieken is de simultane kansverdeling van X en Y bepaald. De kansen $P(X = x \text{ en } Y = y)$ zijn als volgt:

$x \backslash y$	0	1	2	3
0	0.05	0.05	0.10	0
1	0.05	0.10	0.25	0.10
2	0	0.15	0.10	0.05

a. Bepaal de (marginale) kansverdelingen van X en van Y.
b. Bereken verwachting en variantie voor zowel X als Y.
c. We bekijken $Z = 8 \cdot Y$, het aantal verloren arbeidsuren door afwezigheid in de avondploeg. Volgens de rekenregels geldt nu:$E(Z) = 8E(Y)$ en $var(Z) = 64var(Y)$
Ga dit na door verwachting en variantie van Z te berekenen.
d. Bepaal voor $T = X + Y$ de kansverdeling, d.w.z. bepaal voor T de mogelijke uitkomsten en de bijbehorende kansen en bereken $E(T)$ en $var(T)$
e. Ga na dat $E(X + Y) = E(X) + E(Y)$ geldt.
f. Ga na dat $var(X + Y) \neq var(X) + var(Y)$ in dit geval.
Waarom is $var(X + Y)$ in dit geval niet gelijk aan $var(X) + var(Y)$?

4. Zij N het aantal worpen met een zuivere munt dat nodig is om voor het eerst munt te werpen. Als we voor N de waarde n (bijv. 10) hebben waargenomen, gooien we nog eens n keer. De stochastische variabele X is het aantal keren munt in deze n (bijv. 10) worpen.

a. Bepaal $P(N = 10)$, $P(X = 4|N = 10)$ en $P(X = 4 \text{ en } N = 10)$.
b. Bepaal de verdeling van N, de voorwaardelijke verdeling van X gegeven $N = n$ en de simultane verdeling van X en N (met “verdeling” bedoelen we: geef “de kansfunctie”).
c. Bepaal $E(X|N = 10)$, $E(X|N = n)$ en $E(X)$, de laatste m.b.v. $E(X) = E[E(X|N)]$.
d. Bepaal $P(X = 0)$.
e. Bepaal de voorwaardelijke kansfunctie van N gegeven $X = 0$ en $E(N|X = 0)$.

5. Veronderstel dat het aantal ongelukken N in de avondspits Poisson-verdeeld is met parameter µ. Bij elk ongeluk bedraagt de schade € 1000, € 2000, € 3000 of € 4000 met kansen (resp.) 0.1, 0.3, 0.4 en 0.2.
Zij S de totale schade. Doel van deze opgave is om ES te berekenen.
Laat X_i de schade zijn bij het i^{de} ongeluk, $i = 1, 2, \ldots$.
We veronderstellen dat de X_i’s en N onafhankelijke s.v.-en zijn.

a. Beschrijf S als functie van de X_i’s als gegeven is dat $N = n$.
b. Bereken EX_i.
c. Bepaal $E(S|N = n)$.
d. Bereken ES (gebruik eigenschap 5.2.6).

6. Laat X het aantal klanten zijn dat in een bank gedurende een half uur arriveert. Veronderstel dat X een Poisson-verdeling heeft met parameter µ = 10.
Laat Y het aantal in een half uur arriverende klanten zijn die een vraag hebben die meer dan 3 minuten bedieningstijd vergt. Veronderstel verder dat de kansverdeling van Y, gegeven $X = x$, een binomiale verdeling is en wel $B(x, 0.3)$.

a. Bereken (in drie decimalen nauwkeurig) $P(X = 8 \text{ en } Y = 2)$.

b. Bereken $E(Y|X=8)$ en druk $E(Y|X=x)$ uit in x.
c. Bereken $E(Y)$.

7. Van twee lopende banden A en B worden de erop liggende producten gecontroleerd, die per band op volgorde zijn genummerd met 1, 2, 3, …
Steeds geldt $P(\text{goed}) = 0.9$ en het al of niet goed zijn der producten is onafhankelijk van elkaar.
Zij X_1 en X_2 het nummer van het eerste product dat niet goed is op resp. band A en B.

a. Bereken $P(X_1 = 10)$.
b. Bereken $P(20 \leq X_1 \leq 30)$.
c. Bereken $P(X_1 = X_2)$.
d. Bereken $P(X_1 + X_2 = 20)$, via de convolutieformule.

8. X en Y zijn o.o. s.v.-en met ieder een geometrische kansverdeling met parameter p.
We gaan stapsgewijs laten zien dat het minimum van X en Y ook een geometrische verdeling heeft.

a. Bereken $P(X > i \text{ en } Y > i)$.
b. Bereken $P(min(X,Y) > i)$.
c. Bereken $P(min(X,Y) = i)$.
d. Welke (geometrische) verdeling heeft $min(X,Y)$? Bepaal $E[min(X,Y)]$.

9. X en Y zijn o.o. en beide geometrisch verdeeld met parameter p.

a. Druk $E(X+Y)$ en $var(X+Y)$ uit in p.
b. Geef de simultane kansfunctie van X en Y en schets het waardenbereik $S_X \times S_Y$ in $\mathbb{R}^2$ (het xy-vlak)
c. Bepaal de kansfunctie van $X+Y$, door de Convolutiesom toe te passen.

10. Bekijk de onderstaande vier verschillende simultane verdelingen van X en Y, geven door de simultane kansfunctie $P(X = i \text{ en } Y = j)$ in een kanstabel.

Verdeling 1

$i \backslash j$	0	1	2
0	0.2	0.1	0
1	0.1	0.2	0.1
2	0	0.1	0.2

Verdeling 2

$i \backslash j$	0	1	2
0	$\frac{1}{20}$	$\frac{1}{5}$	$\frac{1}{20}$
1	$\frac{1}{5}$	0	$\frac{1}{5}$
2	$\frac{1}{20}$	$\frac{1}{5}$	$\frac{1}{20}$

Verdeling 3

$i \backslash j$	0	1	2
0	0.09	0.12	0.09
1	0.12	0.16	0.12
2	0.09	0.12	0.09

Verdeling 4

$i \backslash j$	0	1	2
0	0	0	0.3
1	0	0.4	0
2	0.3	0	0

a. Ga na dat de marginale verdelingen van X en van Y voor alle 4 simultane verdelingen gelijk zijn en bepaal $E(X)$, $var(X)$, $E(Y)$ en $var(Y)$.
b. Bekijk de verdelingen 1 t/m 4 eens goed door bijvoorbeeld de punten met kans > 0 in het xy-vlak te tekenen: kies voor elke verdeling een waarde voor $\rho(X,Y)$ uit de verzameling $\{-2, -1, -\frac{2}{3}, 0, \frac{2}{3}, 1, 2\}$.
c. Ga na bij welke van de vier verdelingen X en Y o.o. zijn. Motiveer je antwoorden kort.

d. Bepaal voor verdeling 1: $E(XY)$, $cov(X,Y)$ en $\rho(X,Y)$
e. Bepaal ook voor verdeling 1: $cov(3X, 2-Y)$ en $\rho(3X, 2-Y)$
f. Laat zien dat voor verdeling 2 geldt dat $\rho(X,Y) = 0$.
Kunnen we hieruit concluderen dat X en Y o.o. zijn? Waarom (niet)?
g. Bepaal voor verdeling 1 de verdeling van X, gegeven $Y = 0$, en $E(X|Y = 0)$.
Herhaal dit voor de voorwaarden $Y = 1$ en $Y = 2$.
Ga vervolgens na dat $E(X) = \sum_y E(X|Y = y) \cdot P(Y = y)$.

11. Gebruik de eigenschappen voor covariantie en correlatiecoëfficiënt om $\rho(X,Y)$ te bepalen als $Y = -3X + 4$ (ofwel $P(Y = -3X + 4) = 1$).

12. $X_1, X_2, \ldots, X_n$ zijn o.o. en gelijk verdeeld met verwachtingswaarde 1 en variantie 2.

a. Bepaal met de rekenregels $cov(X_1, X_1 + X_2)$ en de correlatiecoëfficiënt $\rho(X_1, X_1 + X_2)$.
b. Bereken de kleinste waarde van n waarvoor $\rho(X_1, X_1 + X_2 + \ldots + X_n) < \frac{1}{3}$.

13. (oude tentamenopgave)
Tien ballen, genummerd 1 t/m 10, worden in een willekeurige volgorde gezet: posities 1 t/m 10.
Definieer: $X_i = \begin{cases} 1 & \text{als de bal met nimmer } i \text{ op positie } i \text{ staat} \\ 0 & \text{anders} \end{cases}$, $i = 1, 2, \ldots, 10$
dus $S = X_1 + X_2 + \cdots + X_{10}$ is het aantal ballen op de “juiste” plaats.

a. Bereken $E(X_1)$, $var(X_1)$ en $cov(X_1, X_2)$.
b. Bereken $E(S)$ en $var(S)$.

14. (oude tentamenopgave)
Een chirurg opereert per week gemiddeld twee mensen met acute blindedarmontsteking en gemiddeld drie mensen met nierstenen. We weten dat hij in de afgelopen week in totaal 7 van dergelijke operaties heeft verricht.

a. Als X en Y o.o. en Poisson-verdeelde stochastische variabelen met parameters μ_1 resp. μ_2 zijn, geef dan de verdeling van $X + Y$ (niet afleiden!).
b. Toon aan dat X, gegeven $X + Y = n$, $B\left(n, \frac{\mu_1}{\mu_1 + \mu_2}\right)$-verdeeld is.
c. Bepaal het verwachte aantal blindedarmoperaties in de afgelopen week, gegeven de 7 operaties in totaal. Welke (redelijke) veronderstellingen gebruik je daarbij om a. toe te passen?

Enige aanwijzingen bij de opgaven van hoofdstuk 5:

1. Noteer een 1 voor Kruis en een 0 voor Munt en maak een tabel met 3 kolommen: kolom 1 bevat de uitkomsten (0010 e.d.), kolom 2 de waarde van X en kolom 3 de waarde van Y. Vervolgens kun je de simultane kansen $P(X = i \text{ en } Y = j)$ bepalen door te tellen.
2. Let er op dat Y slechts de waarden 0 en 1 kan aannemen: als je $j = 0$ substitueert, vind je de kansen $P(X = i \text{ en } Y = 0)$ (de punten op de X-as vanaf (1, 0)): deze kansen bij elkaar opgeteld leveren je de kans $P(Y = 0)$. Bij die sommatie moet je de meetkundige reeks toepassen, zie zo nodig de bijlage wiskundige technieken. Idem voor $Y = 1$.
3. Marginale verdeling van X: de kansen in de rijen van de tabel optellen. Zet de kansen in een tabel en voeg er 2 rijen met $x \cdot P(X = x)$ en $x^2 \cdot P(X = x)$ aan toe voor het berekenen van EX en $E(X^2)$.
   d. als bijvoorbeeld $T = 3$ kan (X, Y) òf (1, 2) òf (2, 1) zijn ((0, 3) heeft kans 0).
4. Merk op dat X van N afhangt, maar dat N wel variabel is.
5. Het probleem in deze som is dat S de som van X_1 t/m X_N is maar dat het aantal N niet vast is maar variabel. Het blijkt in de opgave dat je $E(S)$ wel kunt berekenen door de verwachte schade per geval, $E(X)$, te vermenigvuldigen met het verwachte aantal $E(N)$.
6. Vergelijkbaar met opgave 5.
7. In b. moet (kun) je gebruik maken van de eigenschap van de geometrische verdeling dat $P(X > k) = (1 - p)^k$ en verdere weer van de meetkundige reeks om c. op te lossen.
8. –
9. –
10. –
11. Vul $Y = -3X + 4$ in de teller èn de noemer van de formule van $\rho(X, Y)$..
    Gebruik niet $cov(X, Y) = E(XY) - EX \cdot E(Y)$, maar $cov(aX + b, Y) = \cdots$.
12. Vul $X = X_1$ en $Y = X_1 + X_2$ in de teller èn de noemer van de formule van $\rho(X, Y)$.
    Gebruik niet $cov(X, Y) = E(XY) - EX \cdot EY$, maar $cov(X, Y + Z) = \cdots$.
13. Wat is $P(X_i = 1)$?
14. –

Hoofdstuk 6 Continue stochastische variabelen

6.1 Kansdichtheid, verwachting en variantie van een continue variabele

In hoofdstukken 4 en 5 hebben we vooral de verdeling van aantallen onderzocht, meer algemeen van grootheden die een eindig of aftelbaar oneindig aantal waarden kunnen aannemen. Bij veel toevalsexperimenten is de uitkomstenruimte echter de reële rechte of een deel daarvan. Enige voorbeelden:

- Meet het IQ van een willekeurige UT-student: $S = [50{,}200]$.
- Meet de duur (in seconden) van een willekeurig telefoongesprek: $S = [0\,,\infty)$.
- Meet de waardeverandering (in %) van een aandelenpakket na een jaar: $S = \mathbb{R}$.
- Laat een randomgenerator (t.b.v. simulaties) een willekeurig getal tussen 0 en 1 produceren: $S = [0{,}1]$.
- Bepaal de afstand op tijdstip *t* van een vrij elektron t.o.v. de plaats op tijdstip 0: $S = [0\,,\infty)$.

Bij dergelijke experimenten speelt de eindige meetnauwkeurigheid natuurlijk ook een rol. Zo zal de telefoongespreksduur in hele of tienden van seconden gemeten kunnen worden: er zijn dan aftelbaar veel uitkomsten mogelijk. Maar voor het opstellen van een kansmodel gaan we in dit soort gevallen uit van de fysische werkelijkheid, en niet van de (veranderlijke) meetnauwkeurigheid.

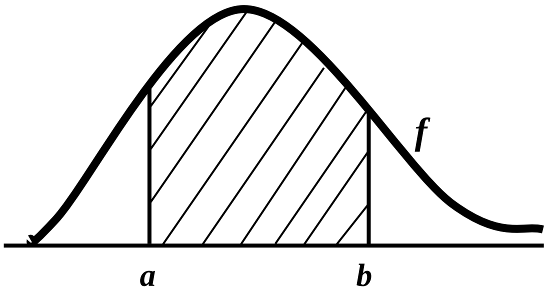
Voor continue stochastische variabelen kunnen we geen kansfunctie geven, omdat in zijn algemeenheid $\boldsymbol{P(X = x) = 0}$ voor elke reële waarde x.

Voorbeeld 6.1.1 Leeftijd, gemeten in jaren, is in het algemeen een continue variabele met een waardenbereik $[0, \infty)$. Als je je afvraagt wat de kans is dat een willekeurig persoon even oud is als jij (20 jaar), dan zal dat een positieve kans opleveren, bijv. $P(X = 20) = 1.75\%$. Echter, impliciet hebben we hier een discrete leeftijdsverdeling van X aangenomen: $S_X = \{0, 1, 2, 3, \ldots, 130\}$.
Voor een continue leeftijdsvariabele X komt “20 jaar” overeen met een interval $20 \leq X < 21$. Kiezen we het tijdsinterval kleiner, bijv. de kans op leeftijdsgenoten die in dezelfde maand, op dezelfde dag, in hetzelfde uur of zelfs in de zelfde minuut geboren zijn, dan is het duidelijk dat de kansen steeds dichter bij 0 komen te liggen. ■

We kunnen dus niet meer spreken van een kansfunctie van een continue grootheid, hoewel in de “meetwerkelijkheid” we altijd afronden op een bepaald aantal decimalen nauwkeurig. Hierdoor kunnen de gemeten waarden (vaak met zeer kleine) toch met positieve kans optreden. Los van die meetwerkelijkheid zijn we op zoek naar een modellering van kansen voor continue grootheden. Dat model wordt gegeven door een kansdichtheid.

Definitie 6.1.2 De **kansdichtheid** van een continue stochastische variabele X is een

$$\text{niet} - \text{negatieve functie } \boldsymbol{f}, \text{zodat } \boldsymbol{P(a \leq X \leq b) = \int_a^b f(x)dx}$$



Volgens de hoofdstelling van de algebra kunnen we deze kans ook uitdrukken in een primitieve $F(x)$ van f:

$$\boldsymbol{P(a \leq X \leq b) = \int_a^b f(x)dx = F(x)|_{x=a}^{x=b} = F(b) - F(a)}$$

Hiermee zijn kansen dus te interpreteren als oppervlakten: de kans dat X een waarde aanneemt tussen a en b is gelijk aan de oppervlakte onder de kansdichtheid f boven het interval $[a, b]$. Hiervoor zouden we ook het open interval (a, b) kunnen nemen: dat beïnvloedt de oppervlakte (dus de kans) niet:

$$P(a \leq X \leq b) = P(a < X < b)$$

voor een continue variabele X geldt immers dat de oppervlakte, boven $X = a$ op de X-as, 0 is:

$$P(X = a) = \int_a^a f(x)dx = F(a) - F(a) = 0$$

Het is duidelijk dat $f(x)$ niet negatief mag zijn op een interval (de bijbehorende kans zou dan negatief zijn) en dat de totale oppervlakte onder de grafiek van f de totale kans 1 is. Omgekeerd kunnen we elke functie $f(x)$, die voldoet aan deze twee voorwaarden, beschouwen als kansdichtheid:

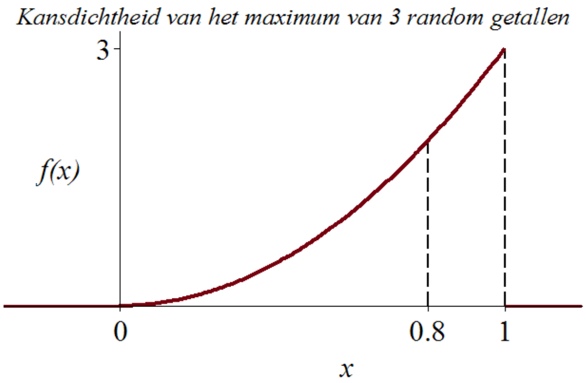
Eigenschap 6.1.3 f is een kansdichtheid als **a.** $\boldsymbol{f(x) \geq 0}$ en

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

Overigens mag f in een eindig aantal punten niet gedefinieerd (of zelfs negatief) zijn: dat beïnvloedt immers de kansen (de oppervlakten) niet.

Voorbeeld 6.1.4 Indien een random generator (bijv. via de randomknop op je rekenmachine) drie “random” getallen tussen 0 en 1 oplevert, kunnen we daaruit de grootste kiezen. Dit maximum X is een continue s.v. met een waarden bereik $S_X = [0, 1]$. De kansdichtheid van X wordt als volgt gegeven (later zullen we zien hoe je deze kansdichtheid kunt afleiden):

$$f(x) = \begin{cases} 3x^2 & \text{als } 0 \le x \le 1 \\ 0 & \text{anders} \end{cases}$$



Dit definieert inderdaad een kansdichtheid
want (eigenschap 6.1.3):

$$1. f(x) \ge 0 \quad \text{en} \quad 2. \int_{-\infty}^{\infty} f(x)dx = \int_0^1 3x^2 dx = x^3|_{x=0}^{x=1} = 1$$

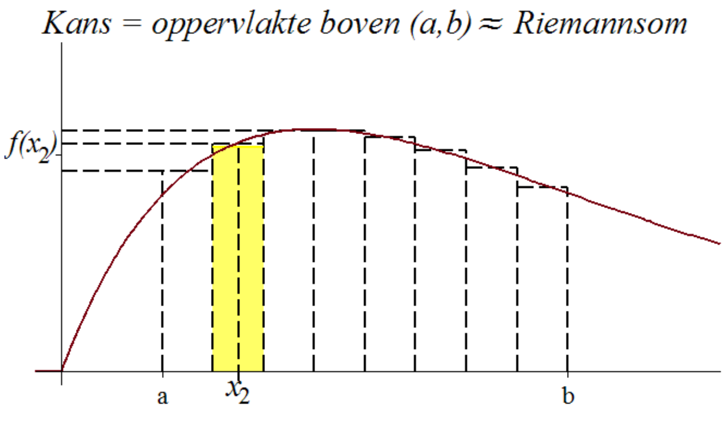
De kans dat het maximum van 3 random getallen groter is dan 0.8 wordt als volgt berekend:

$$P(X > 0.8) = \int_{0.8}^{1} 3x^2 dx = x^3|_{x=0.8}^{x=1} = 1^3 - 0.8^3 = 48.8\%$$ ■

Voorbeeld 6.1.4 illustreert ook nog eens dat $f(x)$ geen kans is maar een kans***dichtheid***:
anders dan bij kansen, kan $f(x)$ waarden groter dan 1 kan aannemen.
$f(x)$ is geen kans, maar bepaalt wel de grootte van de kans bij even brede intervallen.
Met behulp van de grafiek in voorbeeld 6.1.4 is eenvoudig in te zien dat

$$P(0.8 \le X \le 1) > P(0 \le X \le 0.2)$$

Het begrip voor kansdichtheden en kansen als oppervlakten wordt ondersteund door de definitie van een integraal als de “limiet van een Riemann-som” (zie onderstaande toelichtende grafiek):
We delen dan het interval $[a, b]$ in n deelintervallen met gelijke breedte $\frac{b-a}{n}$, ook wel aangeduid met Δx of dx:
de intervallen $[a, a + \Delta x), [a + \Delta x, a + 2 \cdot \Delta x), \dots, [a + (n-1) \cdot \Delta x, b]$ hebben als middens (midpoints) $x_1, x_2, \dots, x_n$.

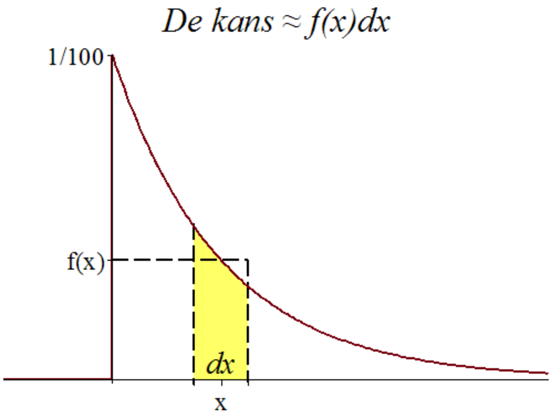


De kans dat X in het i-de intervalletje ligt, is bij benadering de oppervlakte van de rechthoek erboven: oppervlakte rechthoek = lengte×breedte = $f(x_i) \times \frac{b-a}{n} = f(x_i) \times \Delta x$.

Dus er geldt: de Riemann-som $\sum_{i=1}^{n} f(x_i) \times \frac{b-a}{n}$ gaat in de limiet ($n \to \infty$) naar $\int_a^b f(x)dx$ (indien de limiet bestaat).

Voor een kleine intervalbreedte $dx = \frac{b-a}{n}$ geldt:

$$P\left(x - \tfrac{1}{2}dx \le X \le x + \tfrac{1}{2}dx\right) \approx f(x)dx$$



Daarmee kunnen we de analogie van de discrete en continue verdeling inzien:

$$\sum_{a \le x \le b} P(X = x) \text{ komt overeen met } \int_a^b f(x)dx$$

$$\sum_{x \in S_X} P(X = x) = 1 \text{ komt overeen met } \int_{-\infty}^{\infty} f(x)dx = 1$$

$$\text{En: } E(X) = \sum_{x \in S_X} \boldsymbol{x \cdot P(X = x)} \text{ komt overeen met } \boldsymbol{E(X)} = \int_{-\infty}^{\infty} \boldsymbol{x \cdot f(x)dx}$$

Dus ook voor continue variabelen kunnen we $E(X)$ zien als een gewogen gemiddelden van de waarden x met wegingsfactor de "kans" $f(x)dx$. De som vervangen we door een integraal.

Definitie 6.1.5 De **verwachting(swaarde)** van een continue stochastische variabele X is

$$\boldsymbol{E(X)} = \int_{-\infty}^{\infty} \boldsymbol{xf(x)dx},$$

mits deze integraal absoluut convergent is: $\int_{-\infty}^{\infty} |x| \cdot f(x)dx < \infty$

We zullen weer de notaties $E(X)$, EX, μ en μ_X hanteren.
Het *label* X (of Y, Z) zullen we in geval van mogelijke verwarring toevoegen: dat geldt ook voor de kansdichtheden: $f_X(x)$ en f_Z etc.

Voorbeeld 6.1.6 In voorbeeld 6.1.4 gebruikten we als kansdichtheid van s.v. X:

$$f(x) = 3x^2 \text{ als } 0 \le x \le 1 \text{ en } f(x) = 0 \text{ voor andere waarden van } X$$

Wat is nu de verwachte waarde en de variantie van zo'n maximum van 3 random getallen?

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_{0}^{1} x \cdot 3x^2 dx = \tfrac{3}{4}x^4\Big]_{x=0}^{x=1} = \frac{3}{4}$$

Voor de berekening van de variantie gebruiken we dezelfde definitie als bij discrete verdelingen en de eigenschap die ook voor continue verdelingen geldig blijft:

$$var(X) = E(X^2) - (EX)^2$$

Hierin is $E(X^2)$ het gewogen gemiddelde van de kwadraten van de waarden van X:

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx = \int_{0}^{1} x^2 \cdot 3x^2 dx = \tfrac{3}{5}x^5\Big]_{x=0}^{x=1} = \frac{3}{5}$$

Dus: $var(X) = E(X^2) - (EX)^2 = \frac{3}{5} - \left(\frac{3}{4}\right)^2 = \frac{3}{80}$ ■

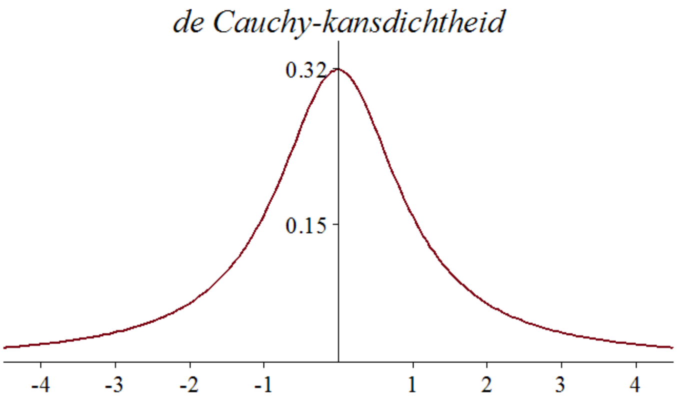
Eigenschap 6.1.7 Voor een reële functie g geldt: $\boldsymbol{Eg(X) = \int_{-\infty}^{\infty} g(x)f(x)dx}$

Bij de berekening van de variantie kunnen we van deze eigenschap gebruikmaken:

- Direct met de definitie: $\boldsymbol{var(X) = E(X-\mu)^2} = \int_{-\infty}^{\infty}(x-\mu)^2 \cdot f(x)dx$
- Maar meestal leidt het gebruik van de rekenformule $\boldsymbol{var(X) = E(X^2) - \mu^2}$ tot eenvoudiger berekeningen: we passen dan $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx$ toe.

Voorbeeld 6.1.8 De **Cauchy-verdeling** wordt gedefinieerd door de volgende kansdichtheid

$$f(x) = \frac{1}{\pi(1+x^2)}, \qquad \text{voor } x \in \mathbb{R}$$



Dit is inderdaad een kansdichtheid want:
1) $f(x) \geq 0$ en
2) $\int_{-\infty}^{\infty} f(x)dx = 1$.

Dit laatste volgt uit het feit dat $\frac{1}{1+x^2}$ de afgeleide is van de inverse tangens-functie

$$arctan(x): \int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)}dx = \frac{1}{\pi} \cdot \arctan(x)\Big]_{x\to-\infty}^{x\to\infty} = \frac{1}{\pi} \cdot \left[\frac{1}{2}\pi - \left(-\frac{1}{2}\pi\right)\right] = 1$$

Op grond van de symmetrie van de kansdichtheid zou je verwachten dat $E(X) = 0$, maar in dit geval klopt dat niet (!), omdat EX niet bestaat: de integraal is niet (absoluut) convergent:

$$\int_{-\infty}^{\infty} |x| \cdot \frac{1}{\pi(1+x^2)} dx = 2\int_{0}^{\infty} x \cdot \frac{1}{\pi(1+x^2)} dx \ \text{(omdat het een even functie is)}$$
$$= \frac{1}{\pi} \cdot \ln(1+x^2)\Big]_{x=0}^{x\to\infty} = \infty$$

Omdat $\mu = EX$ niet bestaat, bestaat $var(X) = E(X-\mu)^2$ ook niet. ■

6.2 Verdelingsfunctie

Voorbeeld 6.2.1 Een grote verzekeraar uit Apeldoorn wil de kansverdeling van de duur X (in seconden) van telefoongesprekken met cliënten modelleren. Een verdeling van X in hele seconden ligt niet voor de hand gezien het grote aantal mogelijkheden. Uit metingen blijkt echter dat de fractie van gesprekken, die langer dan t seconden duren, exponentieel afneemt met de tijd t, en wel als volgt:

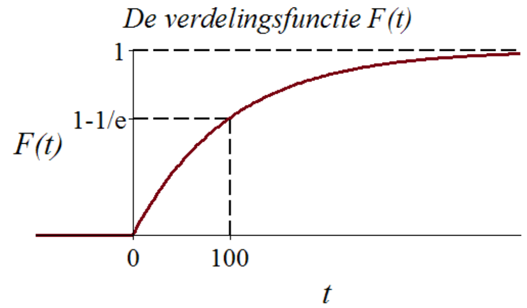
$$P(X > t) = e^{-0.01t}, \ \text{voor } t \geq 0.$$

Uit dit functionele verband, dat de werkelijkheid goed bleek te beschrijven, kunnen we nu allerlei kansen afleiden:

- $P(X > 100) = e^{-1} \approx 36.8\%$
- $P(X > 0) = e^0 = 1$
- $P(100 < X \leq 200) = P(X > 100) - P(X > 200) = e^{-1} - e^{-2} \approx 23.3\%$.
- En wegens de complementregel: $P(X \leq t) = 1 - P(X > t) = 1 - e^{-0.01t}$ $(t \geq 0)$.

De laatste functie $F: t \to P(X \leq t)$ wordt wel de **verdelingsfunctie** van X genoemd. M.b.v. de frequentie-interpretatie kunnen we, uitgaande van een groot aantal van die gesprekken, $F(t)$ interpreteren als de fractie van gesprekken die t seconden of korter duren.

$$F(t) = P(X \leq t) = 1 - P(X > t) = \begin{cases} 1 - e^{-0.01t} & \text{voor } t \geq 0 \\ 0 & \text{voor } t < 0 \end{cases}$$



We kunnen kansen op intervallen dus in F uitdrukken, zoals

$$\begin{aligned} P(100 < X \leq 200) &= P(X \leq 200) - P(X \leq 100). \\ &= F(200) - F(100) \\ &= (1-e^{-2}) - (1-e^{-1}) \approx 23.3\% \end{aligned}$$

■

Voor elke stochastische variabele X, gedefinieerd op een kansruimte (S, P) kunnen we zo'n verdelingsfunctie F definiëren.

Definitie 6.2.2 De functie F, gedefinieerd door $F(x) = P(X \leq x)$ met $x \in \mathbb{R}$, heet de **verdelingsfunctie** van de stochastische variabele X.

Als er sprake is van meer dan één stochastische variabele, "labelen" we de verdelingsfuncties: $F_X(x)$ en $F_Y(y)$ zijn de marginale verdelingsfuncties van de stochastische variabelen X en Y.

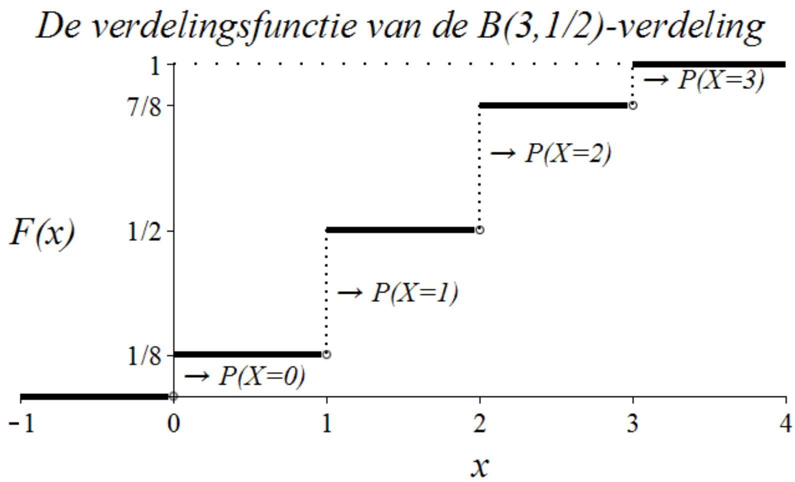
Voorbeeld 6.2.3 In voorbeeld 4.2.1 beschreven we de $B\left(3, \frac{1}{2}\right)$-verdeelde stochastische variabele X. Ook voor een discrete stochastische variabele kunnen we de verdelingsfunctie bepalen.
Er geldt hier bijvoorbeeld

$$F(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}$$

en ook

$$F(1.7) = P(X \leq 1.7) = P(X = 0) + P(X = 1) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}$$

Dus voor elke $x \in [1, 2)$ geldt $F(x) = P(X \leq x) = P(X \leq 1) = \frac{1}{2}$.



Bij discontinuïteitspunten x wordt de hoogte van de "sprong" gegeven door de kans $P(X = x)$ op de betreffende waarde. ■

We merken op dat cumulatieve binomiale kans-tabellen in wezen de verdelingsfunctie van de binomiale verdeling geven: als we voor de stochastische variabele X, die $B\left(20, \frac{3}{10}\right)$-verdeelde is, $F_X(4.5)$ willen bepalen, vinden we dat als volgt in $B\left(20, \frac{3}{10}\right)$-tabel:

$$F(4.5) = P(X \leq 4.5) = P(X \leq 4) = \sum_{k=0}^{4} P(X = k) \approx 23.75\%$$

De grafieken van de verdelingsfuncties in voorbeelden 6.2.1 en 3 laten niet-dalende functies $F(x)$ zien, die voor kleine x-waarden beginnen bij de waarde 0 en op de positieve X-as eindigen (of naderen naar) de waarde 1. Bovendien is één van de grafieken overal continu en de andere minstens rechtscontinu. Deze eigenschappen kunnen bewezen worden m.b.v. de axioma's van Kolmogorov ($F(x)$ is een kans!). Dit technische bewijs laten we achterwege.

Eigenschap 6.2.4 (Karakteristieke eigenschappen van een verdelingsfunctie)
Voor F, de verdelingsfunctie van de stochastische variabele X, geldt:

- **a.** F is niet-dalend (als $x_2 > x_1$, dan $F(x_2) \geq F(x_1)$).
- **b.** $\lim\limits_{x\to\infty} F(x) = 1$ en $\lim\limits_{x\to-\infty} F(x) = 0$
- **c.** F is rechtscontinu ($\lim\limits_{h\to 0^+} F(x+h) = F(x)$)

Deze 3 eigenschappen zijn karakteristiek voor verdelingsfuncties: elke functie die hieraan voldoet noemen we een verdelingsfunctie. De volgende eigenschappen kunnen we uit de definitie van $F(x)$ en de eigenschappen voor kansen afleiden. Ze zijn vooral van belang bij het berekenen van kansen.

Eigenschap 6.2.5 Voor de verdelingsfunctie F van een stochastische variabele X geldt:

- **a.** $P(a < X \leq b) = F(b) - F(a)$.
- **b.** $P(X > x) = 1 - F(x)$.
- **c.** $P(X < x) = \lim\limits_{h\to 0^+} F(x-h) = F(x)$
- **d.** $P(X = x) = F(x) - P(X < x)$

Bewijs: a. $\{X \leq a\}$ en $\{a < X \leq b\}$ vormen een partitie van $\{X \leq b\}$.
Dus volgens axioma (3) van Kolmogorov: $P(X \leq b) = P(X \leq a) + P(a < X \leq b)$
Ofwel $F(b) = F(a) + P(a < X \leq b)$.
b. volgt uit de complementregel uit Hoofdstuk I met $A = \{X > x\}$.
c. Bewijzen we niet. $\lim\limits_{h\to 0^+}$ betekent de "limiet van de functie voor h **daalt** naar 0" ($\lim\limits_{h\downarrow 0}$).
d. Wegens $\{X \leq x\} = \{X < x\} \cup \{X = x\}$. ■

In voorbeeld 6.2.3 zagen we dat de verdelingsfunctie van een discrete stochastische variabele een "trapfunctie" is, waarvan de som van de hoogten van de "treden" 1 is. In voorbeeld 6.2.1 zagen we een continue verdelingsfunctie waarvoor volgens eigenschap 6.2.5d dus geldt: $P(X = x) = 0$ voor elke $x \in$ R.

Definitie 6.2.6 Een stochastische variabele X is **continu** als de verdelingsfunctie F van X een continue functie is.

Uiteraard zijn er nog meer typen stochastische variabelen denkbaar, zoals "mengvormen" van discrete en continue verdelingen. Deze laten we echter buiten beschouwing.

Het verband tussen kansdichtheid en verdelingsfunctie is nu m.b.v. de definitie van de verdelingsfunctie gemakkelijk te geven:

$$\boldsymbol{F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du}$$

Daarmee is de verdelingsfunctie $F(x)$ dus een primitieve van de kansdichtheid $f(x)$. We zullen in de kansrekening de veredelingsfunctie steeds met $F(x)$ aanduiden. Bedenk dus, dat $F(x)$ een **specifieke primitieve** is. De primitieven van $f(x)$ zijn van de vorm $F(x)+c$. Omgekeerd volgt uit de hoofdstelling van de algebra

$$\boldsymbol{F'(x) = f(x)}$$

Ofwel, voor een klein intervalletje $\left(x-\frac{1}{2}dx, x+\frac{1}{2}dx\right)$, met een lengte dx (zie de grafiek op pagina 6-3), geldt met de eigenschappen van $F(x)$:

$$P\left(x-\tfrac{1}{2}dx < X \leq x+\frac{1}{2}dx\right) = F\left(x+\frac{1}{2}dx\right) - F\left(x-\frac{1}{2}dx\right) \approx F'(x)dx = f(x)dx$$

We zien dat de kans, dat X in een interval met gegeven lengte dx in de buurt van x ligt, bepaald door **de afgeleide van de verdelingsfunctie** (aannemende dat de afgeleide van F in x bestaat). Hieruit blijkt dat we de verdeling van een continue s.v. X kunnen geven met de kansdichtheid f of met de verdelingsfunctie F: uit de één volgt de nader.

Voorbeeld 6.2.7 In voorbeeld 6.2.1 hebben we de volgende verdelingsfunctie voor een de gespreksduur X gebruikt:

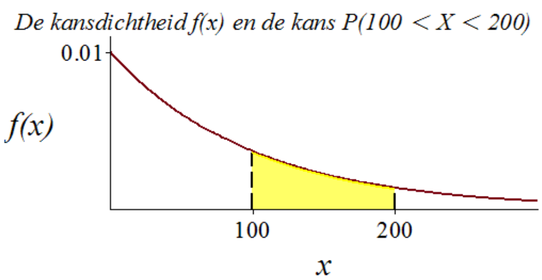
$$F(t) = \begin{cases} 1-e^{-0.01t} & \text{voor } t \geq 0 \\ 0 & \text{voor } t < 0 \end{cases}$$

Daarmee berekenden we de kans

$$P(100 < X \leq 200) = F(200) - F(100) = (1-e^{-2}) - (1-e^{-1}) \approx 23.3\%$$

De kansdichtheid van X is dus:

$$f(t) = \frac{d}{dt}F(t) = \begin{cases} 0.01e^{-0.01t} & \text{voor } t \geq 0 \\ 0 & \text{voor } t < 0 \end{cases}$$



Dit wordt de **exponentiele kansdichtheid met parameter 0.01** genoemd.

Ook hiermee kunnen we de kans $P(100 < X \leq 200)$ berekenen:

$$P(100 < X \leq 200) = \int_{100}^{200} 0.01e^{-0.01t}dt = -e^{-0.01t}\Big|_{t=100}^{t=200} = -e^{-2} + e^{-1} \approx 23.3\% \qquad \blacksquare$$

Voor discrete X geldt ook dat de verdeling gegeven kan worden door de verdelingsfunctie F en door de kansfunctie $P(X=x)$, maar bijna altijd zullen we de kansfunctie gebruiken om de verdeling te definiëren.

Eigenschap 6.2.8 (Eigenschappen van continue verdelingen)
Voor een continue stochastische variabele X met kansdichtheid f en verdelingsfunctie F geldt:

a. $P(X=x) = 0$ voor $x \in \mathbb{R}$.

b. $P(X \in [a,b]) = \int_a^b f(x)dx = F(b) - F(a)$
het gesloten interval $[a,b]$ kunnen we hierin vervangen door het open (a,b)

c. $F(x) = \int_{-\infty}^{x} f(u)du$
d. $f(x) = \frac{d}{dx}F(x)$
e. Als de kansdichtheid $f(x)$ van X symmetrisch is t.o.v. $x = c$ is $E(X) = c$ (mits $E(X)$ bestaat)

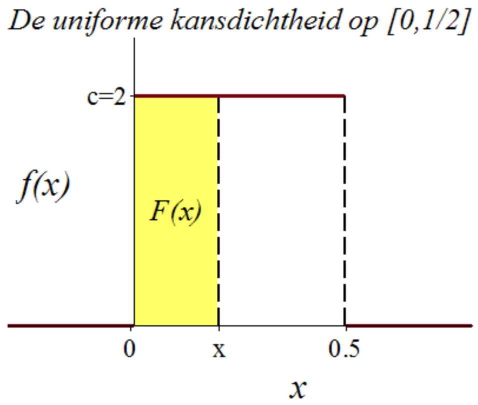
6.3 De uniforme, de exponentiële en de standaardnormale verdeling

De uniforme verdeling op het interval $[a, b]$

Voorbeeld 6.3.1 Stel dat een grootheid slechts gemeten kan worden in hele eenheden. De waarde 387.84 wordt dus waargenomen als 388 (afleesfout 0.16) en 238.435 als 238 (afleesfout 0.435). (We nemen dus aan dat waarnemen afronden op een geheel aantal “impliceert”.)
Zij de afleesfout X het (absolute) verschil tussen een willekeurige optredende waarde van de grootheid en de gemeten waarde, dan is $S_X = \left[0, \frac{1}{2}\right]$. Voor de kansdichtheid f van X zal moeten gelden, dat $f(x) = 0$ voor $x \notin \left[0, \frac{1}{2}\right]$
Binnen dit interval zal elke waarde “even vaak” voorkomen.
Daar X een continu verdeelde stochastische variabele is, zal $f(x)dx$ dus, bij constante dx, overal op $\left[0, \frac{1}{2}\right]$ gelijk moeten zijn, dus f moet constant zijn op $\left[0, \frac{1}{2}\right]$: $f(x) = c$
Voor $c = 2$ is f een kansdichtheid volgens eigenschap 6.1.3:

$$\int_{-\infty}^{\infty} f(x)dx = 1 = \text{de oppervlakte van de rechthoek} = 0.5 \cdot c, \text{dus } c = 2$$

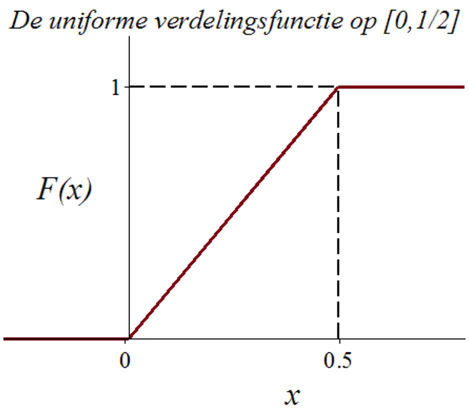
$$f(x) = \begin{cases} 2, & \text{als } x \in \left[0, \frac{1}{2}\right] \\ 0, & \text{als } x \notin \left[0, \frac{1}{2}\right] \end{cases}$$



We zeggen wel dat X een **uniforme kansdichtheid** heeft op het **interval** $\left[\mathbf{0}, \frac{\mathbf{1}}{\mathbf{2}}\right]$

De verdelingsfunctie $F(x) = P(X \leq x)$ kunnen we uit deze grafiek gemakkelijk afleiden.
Immers als $x \in \left[0, \frac{1}{2}\right]$, geldt (zie grafiek): $F(x) = x \cdot 2 = 2x$, etc.

$$F(x) = \begin{cases} 0, & \text{als } x < 0 \\ 2x, & \text{als } 0 \le x \le \frac{1}{2} \\ 1, & \text{als } x > \frac{1}{2} \end{cases}$$



Ga na dat inderdaad $f(x) = \frac{d}{dx}F(x)$
De "gemiddelde" afleesfout is intuïtief (op grond van symmetrie van f) gelijk aan $\frac{1}{4}$, het midden van het interval $\left[0, \frac{1}{2}\right]$.
Met de definitie van $E(X)$ kunnen we dit controleren:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_{0}^{1/2} x \cdot 2dx = x^2\Big]_{x=0}^{x=1/2} = \frac{1}{4}$$

Evenzo

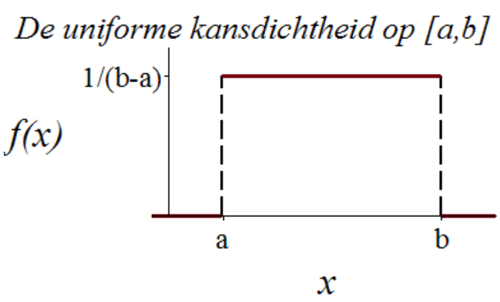
$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx = \int_{0}^{1/2} x^2 \cdot 2dx = \frac{2}{3}x^3\Big]_{x=0}^{x=1/2} = \frac{1}{12}$$

Dus $var(X) = E(X^2) - (EX)^2 = \frac{1}{12} - \left(\frac{1}{4}\right)^2 = \frac{1}{48}$ ■

Algemeen geldt:

Definitie 6.3.2 De stochastische variabele X is **uniform verdeeld op het interval $[a, b]$,** als

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{als } x \in [a, b] \\ 0, & \text{als } x \notin [a, b] \end{cases}$$



Korte notatie: $\boldsymbol{X \sim U(a, b)}$

Soms wordt in plaats van het gesloten interval een open interval (a, b) gekozen.
Eenvoudig is in te zien dat dit inderdaad een kansdichtheid is: de oppervlakte onder de (niet-negatieve) $f(x)$ is $\frac{1}{b-a} \cdot (b-a) = 1$.
De verwachte waarde van "random getal uit het interval $[a, b]$" is op grond van symmetrie uiteraard $\frac{a+b}{2}$, het midden van het interval.

Eigenschap 6.3.3 De verwachting en de variantie van de uniforme verdeling op $[a, b]$ zijn:

a. $\boldsymbol{E(X) = \frac{a+b}{2}}$

b. $\boldsymbol{var(X) = \frac{(b-a)^2}{12}}$

Bewijs:

a. volgt uit eigenschap 6.2.8.e ofwel uit

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_a^b x \cdot \frac{1}{b-a}dx = \frac{1}{2} \cdot \frac{x^2}{b-a}\Big]_{x=a}^{x=b} = \frac{1}{2} \cdot \frac{b^2-a^2}{b-a} = \frac{b+a}{2}$$

b. $E(X^2) = \int_a^b x^2 \cdot \frac{1}{b-a}dx = \frac{1}{3} \cdot \frac{x^3}{b-a}\Big]_{x=a}^{x=b} = \frac{1}{3} \cdot \frac{b^3-a^3}{b-a} = \frac{b^2+ab+a^2}{3}$, dus

$$var(X) = E(X^2) - (EX)^2 = \frac{b^2+ab+a^2}{3} - \left(\frac{b+a}{2}\right)^2 = \frac{4b^2+4ab+4a^2-(3b^2+6ab+3a^2)}{12} = \frac{(b-a)^2}{12}$$

■

De meest gebruikte uniforme verdeling is de $U(0,1)$-verdeling, die een kansmodel biedt voor random getallen tussen 0 en 1. Op rekenmachines zit meestal zo'n random knop: men spreekt wel van "pseudo random generatoren", daar de productie van deze getallen niet echt random is, maar gebaseerd op een functie. We zullen later zien dat random getallen tussen 0 en 1 ook voor simulatiedoeleinden gebruikt kunnen worden.

De exponentiële verdeling met parameter λ

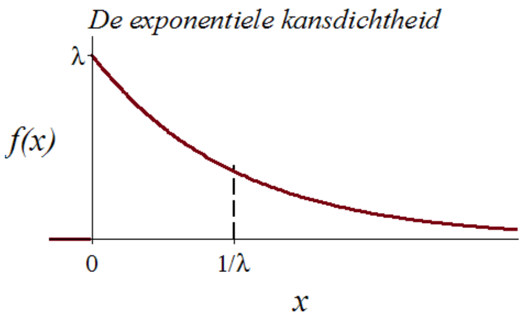
In de voorbeelden 6.2.1 en 6.2.7 hebben we al een voorbeeld van de exponentiele verdeling besproken. Voor de gespreksduur X (in seconden) gold:

- $f(t) = 0.01e^{-0.01t}$ (als $t \geq 0$) en
- $P(X > t) = e^{-0.01t}$, de kans dat het gesprek t seconden "overleeft" neemt exponentieel af met toenemende t.
- $F(t) = P(X \leq t) = 1 - e^{-0.01t}$, voor positieve t omdat de gespreksduur niet negatief kan zijn.

We spreken dan ook van een **exponentiele verdeling met parameter $\boldsymbol{\lambda = 0.01}$**.

Definitie 6.3.4 De stochastische variabele X heeft een **exponentiële verdeling met parameter λ** (> 0) als

$$\boldsymbol{f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{voor } x \geq 0 \\ 0 & \text{voor } x < 0 \end{cases}}$$



Korte notatie: $\boldsymbol{X \sim Exp(\lambda)}$

Eigenschap 6.3.5 Voor een exponentieel verdeelde variabele X (met parameter **λ**) geldt:

a. $\boldsymbol{P(X > x) = e^{-\lambda x}}$, voor $x \geq 0$

b. $\boldsymbol{F(x)} = \begin{cases} \boldsymbol{1 - e^{-\lambda x}} & \text{voor } x \geq 0 \\ 0 & \text{voor } x < 0 \end{cases}$

c. $\boldsymbol{E(X) = \frac{1}{\lambda}}$

d. $\boldsymbol{var(X) = \frac{1}{\lambda^2}}$

Bewijs:

a. $P(X > x) = \int_x^\infty \lambda e^{-\lambda t}dt = -e^{-\lambda t}\Big|_{t=x}^{t\to\infty} = 0 - \left(-e^{-\lambda x}\right) = e^{-\lambda x}$
(voor $x \geq 0$, anders $P(X > x) = 1$
b. Volgt uit a.: $F(x) = P(X \leq x) = 1 - P(X > x)$
c. Om de verwachting te bepalen gebruiken we partiële integratie (zie ook de bijlage “Wiskundige Technieken”)

$$E(X) = \int_0^\infty x \cdot \lambda e^{-\lambda x}dx = x \cdot -e^{-\lambda x}\Big|_{x=0}^{x\to\infty} - \int_0^\infty 1 \cdot -e^{-\lambda x}dx = 0 + -\frac{1}{\lambda}e^{-\lambda x}\Big|_{x=0}^{x\to\infty} = \frac{1}{\lambda}$$

d. Analoog aan b. vinden we $E(X^2) = \int_0^\infty x^2 \cdot \lambda e^{-\lambda x}dx = \ldots = \frac{2}{\lambda^2}$
Dus $var(X) = E(X^2) - (EX)^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$ ■

De exponentiele verdeling vindt vooral zijn toepassing in het modelleren van wachttijden, bedieningsduren, tussenaankomsttijden van klanten en levensduur van apparaten (als de levensduur niet of nauwelijks bepaald wordt door slijtage, maar vooral door “invloeden van buitenaf”).

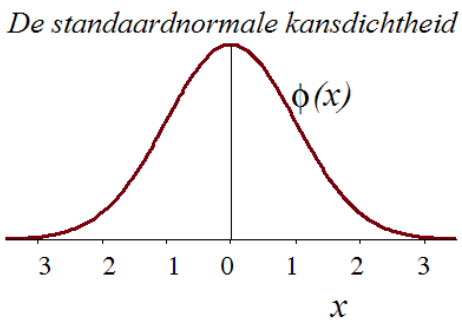
De standaardnormale verdeling

De standaardnormale verdeling is een bijzonder geval van de normale verdeling die we in paragraaf 6 uitvoerig aan de orde stellen. Vanwege het belang van deze verdeling in de kansrekening en de statistiek zullen we voor deze verdeling een aparte notatie invoeren:

- Een standaardnormale s.v. noteren we meestal met $\boldsymbol{Z}$, tenzij anders vermeld.
- De standaardnormale kansdichtheid noteren we met $\boldsymbol{\varphi(z)}$, dus $\varphi = f_Z$.
- De standaardnormale verdelingsfunctie is $\boldsymbol{\Phi(z)}$, dus $\Phi(z) = P(Z \leq z)$.
  (φ en Φ, uitgesproken als “phi”, zijn de Griekse versies van f en F)

Definitie 6.3.6 De continue s.v. Z heeft een **standaardnormale verdeling** als

$$\boldsymbol{\varphi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}}, \text{met } \boldsymbol{z} \in \mathbb{R}$$



Dat dit een kansdichtheid is, dus $\int_{-\infty}^\infty \varphi(z)dz = \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz = 1$, kan bewezen worden m.b.v. overgang op poolcoördinaten (valt buiten het bestek van dit vak).
Uit de symmetrie van de grafiek van φ volgt onmiddellijk dat $E(Z) = 0$, mits $E(Z)$ bestaat.
Dat dit klopt, blijkt uit de berekening:

$$E(Z) = \int_{-\infty}^{\infty} z \cdot \varphi(z) dz = \int_{-\infty}^{\infty} z \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = -\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \Big|_{z\to-\infty}^{z\to\infty} = 0$$

Om in de variantie-formule $var(Z) = E(Z^2) - (EZ)^2$ de term $E(Z^2)$ te bepalen, moeten we z^2 wegen met de kans $\varphi(z)dz$.
Met partiële integratie vinden we

$$E(Z^2) = \int_{-\infty}^{\infty} z^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = z \cdot -\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \Big|_{z\to-\infty}^{z\to\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 0 + 1$$

Merk op dat de laatste integraal precies de totale oppervlakte onder de grafiek van φ is: $\int_{-\infty}^{\infty} \varphi(z) dz = 1$
Dus $var(Z) = E(Z^2) - (EZ)^2 = 1 - 0^2 = 1$
Omdat $\mu = 1$ en $\sigma^2 = 1$, wordt de standaardnormale verdeling ook kort als volgt aangeduid:

$$Z \sim N(0{,}1)$$

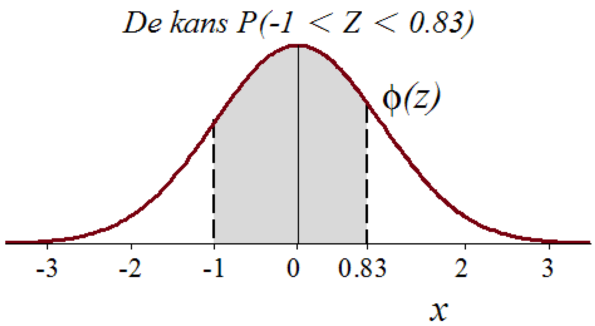
Voorbeeld 6.3.7 Kansen als $P(-1 \le Z \le 0.83)$ zouden we analytisch kunnen bepalen met een integraal:

$$P(-1 \le Z \le 0.83) = \int_{-1}^{0.83} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = ???$$

Maar een primitieve van $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$ bestaat niet! Daarom moeten we dit soort kansen dus bepalen met een numerieke benadering (Riemann-som over het interval!).
Om dit arbeidsintensieve karwei te voorkomen, is de standaardnormale verdelingsfunctie $\Phi(z) = P(Z \le z)$ getabelleerd in de **standaardnormale tabel** (zie achter in het dictaat), voor **positieve *z* en in twee decimalen nauwkeurig**.

We vinden hiermee (zonder integralen):

$$\begin{aligned} P(-1 \le Z \le 0.83) &= P(Z \le 0.83) - P(Z \le -1.00) \\ &= \Phi(0.83) - \Phi(-1.00) \\ &= 0.7967 - 0.1587 \\ &= 63.80\% \end{aligned}$$



De tweede kans $P(Z \le -1.00) = \Phi(-1.00)$ bepalen we door gebruik te maken van de symmetrie van φ rond 0:

$$P(Z \le -1.00) = P(Z \ge 1.00) = 1 - \Phi(1.00) = 1 - 0.8413 = 0.1587 \qquad \blacksquare$$

6.4 Functies van een continue stochastische variabele

Stel X is een **stochastisch** ingangssignaal bij een apparaat (bijvoorbeeld een spanning), waarvan de waarde op een bepaald moment dus niet bekend is, maar waarvan we wel weten dat het signaal X een bekende continue verdeling heeft.
Het apparaat transformeert het signaal tot een uitgangssignaal Y, waarbij de waarde van Y volgens een vervormingsfunctie g uit die van X wordt bepaald.

Schematisch: $X \longrightarrow \boxed{Y = g(X)} \longrightarrow Y$

In het voorbeeld van een spanning zou het apparaat een versterker kunnen zijn met $Y = g(X) = a \cdot X (a > 0)$ of een “gelijkrichter” $Y = g(X) = |X|$.
Er is geen voor de hand liggende methode om de kansdichtheid van Y direct uit die van X af te leiden. Zo definieert $f_Y(y) = a \cdot f_X(y)$ bij de versterker geen kansdichtheid als $a \neq 1$
Wel kunnen we **kansen op gebeurtenissen m.b.t. *Y* uitdrukken in kansen op gebeurtenissen m.b.t. *X***.

Voorbeeld 6.4.1 Voor de gelijkrichter $Y = |X|$ geldt bijvoorbeeld

$$P(Y \leq 3) = P(-3 \leq X \leq 3) = F_X(3) - F_X(-3),$$

waarin F_X de bekende verdelingsfunctie van X is – en voor $P(Y \leq 3)$ kunnen we $F_Y(3)$
Meer algemeen geldt voor de verdelingsfunctie $F_Y(y) = P(Y \leq y)$:

- Als $y < 0$ dan is de gebeurtenis $\{Y \leq y\} = \{|X| \leq y\} = \emptyset$, dus $F_Y(y) = 0$ als $y < 0$.
- Als $y \geq 0$ geldt: $F_Y(y) = P(|X| \leq y) = P(-y \leq X \leq y) = F_X(y) - F_X(-y)$ (eigenschap 6.2.8b.)

De kansdichtheid van Y vinden we door te differentiëren (kettingregel!):

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \begin{cases} 0 & \text{als } y < 0 \\ f_X(y) + f_X(-y) & \text{als } y \geq 0 \end{cases}$$

Vervolgens kunnen we de bekende kansdichtheid van X gebruiken om die van Y te bepalen: de kansverdeling van Y is dan bekend en, met de definitie, $E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy$. ■

In bovenstaand voorbeeld hebben we gezien dat we de verdeling van $Y = g(X)$ kunnen afleiden uit de gegeven verdeling van X door de verdelingsfunctie $F_Y(y)$ uit te drukken in de verdelingsfunctie $F_X(x)$ van X, en vervolgens via de afgeleide de relatie van de kansdichtheden te bepalen.

Voorbeeld 6.4.2 Stel dat de vervormingsfunctie $g(Z) = Z^2$ en dat Z een standaardnormale verdeling heeft, dus: $f_Z(z) = \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$, met $z \in \mathbb{R}$

We zijn nu geïnteresseerd in de verdeling van het uitgangssignaal $Y = Z^2$ en in de verwachtingswaarde $E(Y) = E(Z^2)$.
Om bij dat laatste te beginnen: Dat hebben we in feite al besproken bij de introductie van de standaardnormale verdeling vorige paragraaf:

$$E(Y) = E(Z^2) = \int_{-\infty}^{\infty} z^2 \varphi(z) dz = \cdots = 1$$

Nu het bepalen van de kansverdeling van $Y = Z^2$:

1. Druk eerst $F_Y(y)$ uit in F_Z ($= \Phi$, in dit geval):

   $$F_Y(y) = P(Y \leq y) = P(Z^2 \leq y) = \begin{cases} 0 & \text{als } y \leq 0 \\ P(-\sqrt{y} \leq Z \leq \sqrt{y}) & \text{als } y > 0 \end{cases}$$

   Dus als $y > 0$: $F_Y(y) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y})$
2. Druk via de afgeleide van $f_Y(y)$ uit in f_Z ($= \varphi$).
   als $y \leq 0$, geldt $f_Y(y) = \frac{d}{dy} F_Y(y) = 0$

als $y > 0$, geldt $f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy}\left[\Phi\left(\sqrt{y}\right) - \Phi\left(-\sqrt{y}\right)\right]$
$= \frac{1}{2\sqrt{y}}\left[\varphi\left(\sqrt{y}\right) + \varphi\left(-\sqrt{y}\right)\right]$

3. Gebruik de gegeven verdeling van Z om de formule voor $f_Y(y)$ te bepalen.
   Voor $y \leq 0$ is $f_Y(y) = 0$ en
   $$f_Y(y) = \frac{1}{2\sqrt{y}}\left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\sqrt{y})^2} + \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\sqrt{y})^2}\right] = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y} \text{voor } y > 0.$$
   (*merk op dat de laatste uitdrukking niet bestaat voor $y = 0$: daarom is voor deze waarde $f_Y(y) = 0$ gedefinieerd. We hadden het echter ook ongedefinieerd kunnen laten*)

Berekening van $E(Y)$, nu met de verdeling moet hetzelfde resultaat opleveren als $E(Z^2) = 1$:
$$E(Y) = \int_{-\infty}^{\infty} y \cdot f_Y(y) dz = \int_0^{\infty} y \cdot \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y} dy = \int_0^{\infty} \frac{\sqrt{y}}{\sqrt{2\pi}} e^{-\frac{1}{2}y} dy$$
Geen eenvoudige integraal: we moeten substitutie toepassen. Probeer $y = z^2$, dus $dy = 2z dz$
$$E(Y) = \int_0^{\infty} \frac{z}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \cdot 2z dz = \int_0^{\infty} 2z^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$
Hierin herkennen we de integraal $\int_0^{\infty} 2z^2 \cdot \varphi(z) dz$, die wegens symmetrie (even functie) hetzelfde is als $\int_{-\infty}^{\infty} z^2 \cdot \varphi(z) dz \ = E(Z^2)$, dus inderdaad $E(Y) = E(Z^2) = 1$ ∎

In voorbeeld 6.4.2 hebben we de verdeling van het kwadraat van een standaardnormaal verdeelde variabele bepaald: $f_Y(y) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y}$ $(y > 0)$.
Dit wordt de **chi-kwadraatverdeling met één vrijheidsgraad** genoemd.
Chi-kwadraatverdelingen spelen een belangrijke rol in de statistiek.

Verder hebben we aan de hand van dit voorbeeld de volgende methode behandeld:
Als $Y = g(X)$ waarin X een bekende continue verdeling heeft,

- Dan is de kansverdeling van Y uit die van X vaak af te leiden met de volgende aanpak:

  1. Druk eerst $F_Y(y)$ uit in F_X. $\quad (F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \cdots)$
  2. Druk via de afgeleide van $f_Y(y)$ uit in f_X. $\left(f_Y(y) = \frac{d}{dy} F_Y(y)\right)$
  3. Gebruik de gegeven kansdichtheid f_X om de formule voor $f_Y(y)$ te bepalen

- En $E(Y)$ in dat geval op twee manieren is te bepalen:
  - Met de verdeling van Y: $E(Y) = \int_{-\infty}^{\infty} y \cdot f_Y(y) dy$
  - Met de verdeling van X: $E(Y) = Eg(X) = \int_{-\infty}^{\infty} g(x) \cdot f_X(x) dx$

  We kiezen de methode die analytisch het makkelijkst is.

We illustreren bovenstaande methode met het lineaire verband $Y = aX + b$, waarin a en b reële constanten zijn. We weten al dat geldt: $E(Y) = E(aX + b) = aE(X) + b$, maar hoe leid je de kansdichtheid van Y af uit die van X?
Het is duidelijk dat de lineaire transformatie van X, dus $Y = aX + b$, voor $a = 0$ leidt tot een ontaarde verdeling van Y: dan geldt $P(Y = b) = 1$ en $E(Y) = b = aE(X) + b$.

Eigenschap 6.4.3 Als continue s.v. X een kansdichtheid f_X heeft, dan geldt voor $Y = aX + b$, met $a \neq 0$:

$$\boldsymbol{f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)}$$

Bewijs: we bewijzen het hier met voornoemde aanpak voor $a > 0$ (voor $a < 0$ zie opgaven):

1. $F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right)$
2. De laatste uitdrukking voor $F_Y(y)$ differentiëren naar y (kettingregel!):

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{a} \cdot f_X\left(\frac{y-b}{a}\right) \qquad \blacksquare$$

Voorbeeld 6.4.4 Een randomgenerator levert willekeurige getallen X tussen 0 en 1, dus X heeft een uniforme verdeling: $f_X(x) = \begin{cases} 1 & \text{als } 0 \leq x \leq 1 \\ 0 & \text{elders} \end{cases}$

Hoe kunnen we met zo'n randomgenerator nu een random getal tussen 4 en 7 geven?
Met gezond verstand kiezen we voor $Y = 3X + 4$: als X tussen 0 en 1 ligt, ligt Y tussen 4 en 7. Maar is $Y = 3X + 4$ nu werkelijk $U(4,7)$-verdeeld?
Volgens eigenschap 6.4.3 geldt: $f_Y(y) = \frac{1}{3} \cdot f_X\left(\frac{y-4}{3}\right)$
$f_X\left(\frac{y-4}{3}\right) = 1$, als $0 \leq \frac{y-4}{3} \leq 1$, ofwel als $0 \leq y - 4 \leq 3$, ofwel als $4 \leq y \leq 7$, dus

$$f_Y(y) = \begin{cases} \frac{1}{3} & \text{als } 4 \leq y \leq 7 \\ 0 & \text{elders} \end{cases}, \qquad \text{dus } Y \sim U(4,7) \qquad \blacksquare$$

Eenvoudig is in te zien dat we voorbeeld 6.4.4 kunnen generaliseren:
Als $X \sim U(0,1)$, dan is $Y = (b-a)X + a \sim U(a,b)$.

Een andere toepassing van random getallen tussen 0 en 1 is het genereren van waarnemingen, die afkomstig zijn uit een gegeven kansverdeling. Bij simulaties van ingewikkelde wachttijdsystemen wordt bijvoorbeeld gebruik gemaakt van tussenaankomsttijden en bedieningsduren, die exponentieel verdeeld zijn. Met behulp van de simulaties worden dan de prestaties van het systeem bepaald (geschat).

Eigenschap 6.4.5 Als X uniform verdeeld is op $(0,1)$,
dan is $\boldsymbol{Y = -\frac{\ln(X)}{\lambda}}$ **exponentieel** verdeeld met parameter λ (> 0).

Bewijs: we maken weer gebruik van de 3-staps aanpak voor het afleiden van f_Y:

1. $F_Y(y) = P(Y \leq y) = P\left(-\frac{\ln(X)}{\lambda} \leq y\right) = P(\ln(X) \geq -\lambda y) = P\left(X \geq e^{-\lambda y}\right)$
$= 1 - F_X\left(e^{-\lambda y}\right)$

2. Differentiëren: $f_Y(y) = \frac{d}{dy} F_Y(y) = \lambda e^{-\lambda y} f_X\left(e^{-\lambda y}\right)$

3. $X \sim U(0,1)$, dus $f_X\left(e^{-\lambda y}\right) = 1$ als $0 \leq e^{-\lambda y} \leq 1$, dus voor $y \geq 0$

Conclusie: $f_Y(y) = \begin{cases} \lambda e^{-\lambda y} & \text{als } y \geq 0 \\ 0 & \text{elders} \end{cases}$, dus $Y \sim Exp(\lambda)$ $\blacksquare$

6.5 De normale verdeling $N(\mu, \sigma^2)$

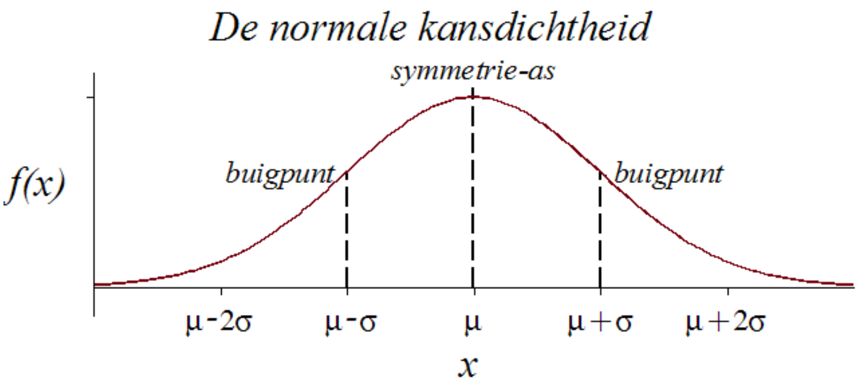
Als we van een willekeurige 20-jarige Nederlandse man de lengte, het gewicht of de longinhoud meten, kunnen we deze grootheden opvatten als stochastische variabelen.
Uit onderzoek blijkt dat vele van dergelijke in de natuur voorkomende grootheden een symmetrische kansdichtheid hebben.
Als X de lengte van een willekeurige 20-jarige Nederlandse man is en als we aannemen dat de "gemiddelde" lengte 180 cm is, dus $E(X) = 180$, dan zullen relatief de meeste mannen een lengte in de buurt van 180 cm hebben. De symmetrie van de kansdichtheid houdt in dat er bijvoorbeeld evenveel mannen met een (afgeronde) lengte van 170 cm als mannen met een lengte van 190 cm zijn.
Hoe groter de afwijking t.o.v. het gemiddelde van 180 cm is, des te kleiner is de kans mensen met zo'n lengte aan te treffen. Deze en vele andere in de biologie, techniek en economie voorkomende grootheden blijken een "klokvormige" kansdichtheid te hebben, die door de zogenaamde normale kansdichtheid zeer goed benaderd wordt. Deze verdeling (ook wel **Gauss-verdeling** genoemd) neemt een centrale plaats in de kansrekening en de statistiek.

Definitie 6.5.1 De stochastische variabele X heet **normaal verdeeld met parameters μ en σ²** als de kansdichtheid van X wordt gegeven door

$$\boldsymbol{f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}} \quad \text{met } x \in \mathbb{R}$$

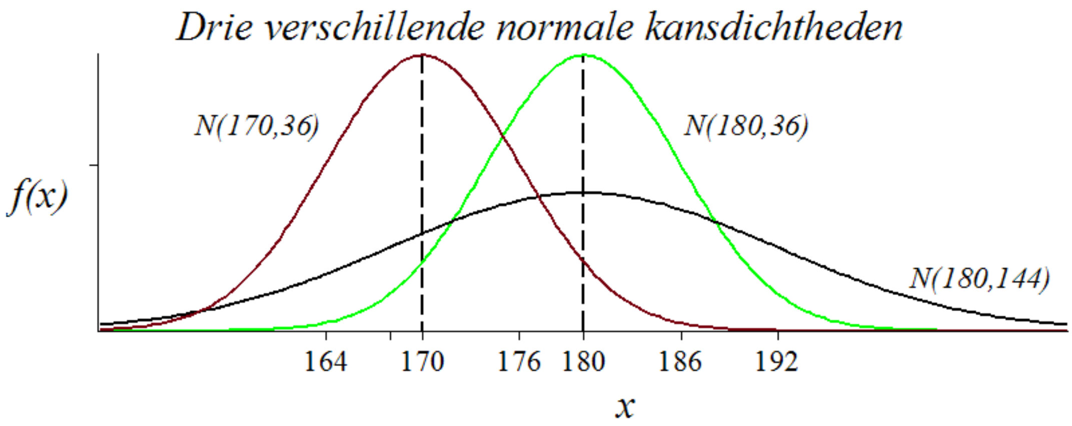
Korte schrijfwijze: $\boldsymbol{X \sim N(\mu, \sigma^2)}$ ofwel: X is $N(\mu, \sigma^2)$-verdeeld.

In paragraaf 6.3 hebben we de standaardnormale ofwel $N(0,1)$-verdeling beschouwd, dus $\mu = 0$ en $\sigma^2 = 1$, en gezien dat in dit speciale geval f, genoteerd met φ, inderdaad een kansdichtheid is.
In het algemeen gelden voor μ geen restricties en $\sigma^2 > 0$.
Als we de grafiek van f willen schetsen, merken we meteen op dat, vanwege het kwadraat in de exponent, $\boldsymbol{f}$ **symmetrisch is t.o.v. μ** en dat f een maximum $\frac{1}{\sqrt{2\pi\sigma^2}}$ aanneemt in het symmetriepunt $x = \mu$.



Verder heeft **f twee buigpunten, namelijk voor $x = \mu - \sigma$ en $x = \mu + \sigma$** en de X-as is een horizontale asymptoot. Het symbool μ voor de eerste parameter wordt verklaard uit het feit dat het symmetriepunt $x = \mu$ van f ook de verwachtingswaarde is (men kan nagaan dat $E(X)$ bestaat). De tweede parameter σ^2 is de variantie.

De standaardafwijking $\sigma = \sqrt{\sigma^2}$ is een zogeheten schaalparameter: naarmate σ groter is wordt de kans op waarden in de buurt van μ kleiner (want $f(\mu)$ wordt kleiner) en evenzo wordt de kans op grote afwijkingen t.o.v. μ groter.
We lichten dit toe met de kansdichtheden van de lengte X van een willekeurige man, als X, voor verschillende populaties, respectievelijk $N(180,144)$-, $N(180,36)$- of $N(170,36)$-verdeeld is.
Merk op dat dit betekent dat de buigpunten op afstand $\sigma = 12$ en $\sigma = 6$ cm van de gemiddelde lengte $\mu = 180$ cm resp. afstand $\sigma = 6$ van $\mu = 170$ liggen (zie onderstaande grafieken).



Hoe groot is nu, bijvoorbeeld, de kans dat een willekeurig gekozen man uit de populatie met een $N(180,36)$ kleiner dan 190 cm is? Net als bij de standaardnormale verdeling kunnen kansen bij de normale verdeling niet door integratie worden berekend, maar moeten we deze numeriek benaderen. Echter, op de schaalverdeling na is de vorm van de normale verdelingen hetzelfde. Van die gelijkvormigheid van de verschillende normale verdelingen maken we gebruik door te herschalen ("standaardiseren") naar een standaardnormaal verdeelde variabele. Vervolgens kunnen we gebruik maken de standaardnormale verdelingstabel, die de kansen $\Phi(z) = P(Z \leq z)$ bevat. Dit blijkt uit de volgende eigenschap.

Eigenschap 6.5.2 Als $X \sim N(\mu, \sigma^2)$ en $Z \sim N(0,1)$, geldt:

a. $\boldsymbol{\sigma Z + \mu \sim N(\mu, \sigma^2)}$
b. De **gestandaardiseerde** $\boldsymbol{\frac{X-\mu}{\sigma} \sim N(0,1)}$
c. $\boldsymbol{E(X) = \mu}$ en $\boldsymbol{var(X) = \sigma^2}$

Bewijs:

a. als $Y = \sigma Z + \mu$, geldt volgens de formule $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$ voor $Y = aX + b$ (eigenschap 6.4.3): $f_Y(y) = \frac{1}{\sigma} \cdot \varphi\left(\frac{y-\mu}{\sigma}\right)$, waarin $f_Z(z) = \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$.
Dus $f_Y(y) = \frac{1}{\sigma} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$, de kansdichtheid van de $N(\mu, \sigma^2)$-verdeling.

b. Schrijf $Y = \frac{X-\mu}{\sigma} = \frac{1}{\sigma} \cdot X - \frac{\mu}{\sigma}$ en pas eigenschap 6.4.3 toe voor $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, om te laten zien dat Y een $N(0,1)$-kansdichtheid heeft.

c. Volgens b. geldt dat als $X \ \ N(\mu, \sigma^2)$-verdeeld is, $Z \ \ N(0,1)$-verdeeld is.
En voor een standaardnormale verdeelde Z geldt $E(Z) = 0$ en $var(Z) = 1$, dus:

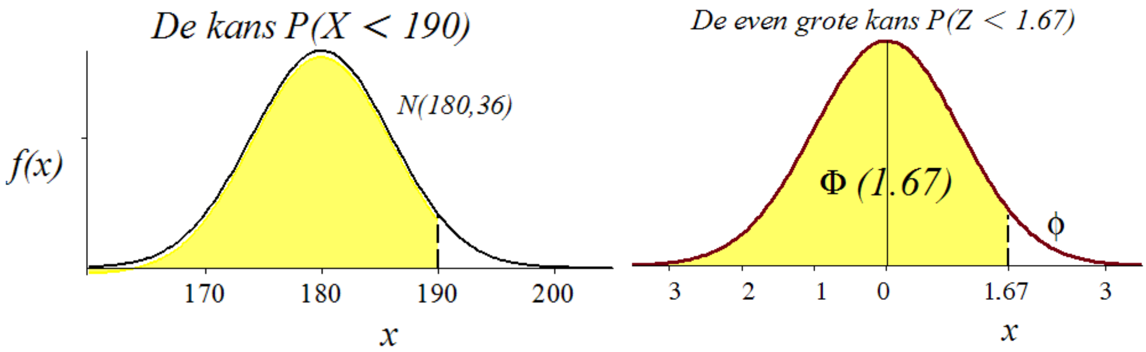
$$E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = E\left(\frac{1}{\sigma} \cdot X - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma} \cdot E(X) - \frac{\mu}{\sigma} = 0 \ \Leftrightarrow \ E(X) = \mu.$$

evenzo: $var(Z) = var\left(\frac{1}{\sigma} \cdot X - \frac{\mu}{\sigma}\right) = \left(\frac{1}{\sigma}\right)^2 \cdot var(X) = 1 \ \Leftrightarrow \ var(X) = \sigma^2$ ∎

Eigenschap 6.5.2b zullen we veelvuldig toepassen om kansen voor normaal verdeelde grootheden te berekenen.

Voorbeeld 6.5.3 De lengte X van een willekeurige man uit een populatie is $N(180,36)$, dus

- De standaardafwijking van de lengtes in de populatie is $\sigma = \sqrt{36} = 6$
- $Z = \frac{X-180}{6}$ is standaardnormaal verdeeld
- $P(X \leq 190) = P\left(\frac{X-180}{6} \leq \frac{190-180}{6}\right) \approx P(Z \leq 1.67) = \Phi(1.67) = 0.9525$
  We noemen $\frac{190-180}{6} \approx 1.67$ de **z-score** van de lengte 190 cm: deze is op 2 decimalen nauwkeurig afgerond, want dat dat is precisie van de standaardnormale tabel.



- $P(X > 200) = P\left(Z > \frac{200-180}{6}\right) = 1 - \Phi(3.33) = 1 - 0.9996 = 0.04\%$
  Een lengte van meer dan 2 meter komt in deze populatie dus zelden voor.
- $P(X \leq 170) = P\left(\frac{X-180}{6} \leq \frac{170-180}{6}\right) \approx \Phi(-1.67) = 1 - \Phi(1.67) = 1 - 0.9525$
  $= 4.75\%$
  Vergelijk met $P(X \leq 190)$ in de grafiek: $P(X \leq 170) = 1 - P(X \leq 190)$
- $P(180 \leq X \leq 190) = P\left(\frac{180-180}{6} \leq Z \leq \frac{190-180}{6}\right) \approx \Phi(1.67) - \Phi(0)$
  $= 0.9525 - 0.5 = 45.25\%$ ∎

De empirische regel, *z*-scores en percentielen.
Bij de introductie van de variantie σ^2 en standaardafwijking hebben we gewezen op de interpretatie met behulp van de **empirische regel voor klokvormige verdelingen**: de kansen dat X ligt in intervallen van de vorm $(\mu - k\sigma, \mu + k\sigma)$ zijn resp. (ongeveer) 68% ($k = 1$), 95% ($k = 2$) en 99.7% ($k = 3$).
Deze percentages zijn gebaseerd op de normale verdeling - en kunnen we nu controleren:

Voor $k = 2$ geldt: $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = P\left(-2 \leq \frac{X-\mu}{\sigma} \leq 2\right)$

$$= \Phi(2) - \Phi(-2)$$
$$= \Phi(2) - \big(1 - \Phi(2)\big)$$
$$= 2 \cdot 0.9772 - 1 = 0.9544 \approx 95\%$$

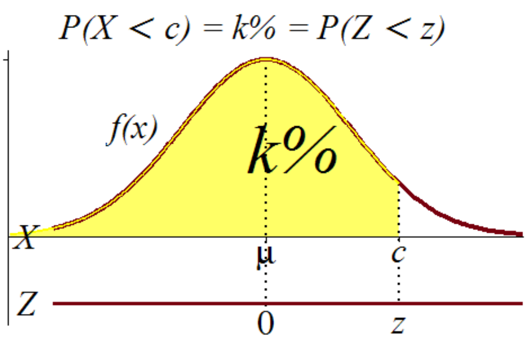
Bij het berekenen van kansen m.b.t. een $N(\mu, \sigma^2)$-verdeelde X maken we gebruik van "standaardisatie": $P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$

Hierin wordt $\frac{x-\mu}{\sigma}$ ook wel de $\boldsymbol{z}$ **−score** (z-waarde) van de grenswaarde x genoemd.

Het **$\boldsymbol{k}$-de percentiel** is een zodanige waarde c dat $\boldsymbol{P(X \leq c) = k\%}$.

Als we de z-score $\boldsymbol{z}$ opzoeken in de N(0, 1)-tabel, zodat $\Phi(z) = k\%$, volgt uit

$P(X \leq c) = \Phi\left(\frac{c-\mu}{\sigma}\right) = k\%$, dat $\frac{c-\mu}{\sigma} = z$, ofwel: $c = \mu + z \cdot \sigma$ is het k-de percentiel.



Voorbeeld 6.5.4 (Procescontrole)

Bij productieprocessen wordt regelmatig gecontroleerd of de productie niet ontregeld raakt, en dus opnieuw ingeregeld moet worden. Indien de maatvoering (lengte, gewicht, inhoud) op een bepaald niveau μ is vastgesteld en de precisie gemeten is en uitgedrukt is in de spreidingsmaat σ (standaardafwijking), kan de ontregeling van het proces met bijv. de 3σ-regel gecontroleerd worden: producten waarvan de maat buiten het interval $(\mu - 3\sigma, \mu + 3\sigma)$ zijn een indicatie dat het proces ontregeld is. Omdat bij productiegewichten e.d. de normale verdeling veelal een goed model is, weten we van de empirische regel dat waarden buiten de "tolerantiegrenzen" klein is, namelijk 0.3%. Overigens, indien je de procescontrole zou baseren op een grote steekproef, bijv. n = 1000, is de kans groot dat je waarden buiten het interval vindt, namelijk $1 - 0.997^{1000} \approx 95\%$ en naar verwachting 3 op de 1000. ■

Voor **lineaire transformaties van de vorm $\boldsymbol{Y = aX + b}$** van de $N(\mu, \sigma^2)$-verdeelde variabele X kunnen we ook weer aantonen dat deze normaal verdeeld zijn (zoals bij eigenschap 6.5.2 is uitgevoerd). De parameters zijn eenvoudig met de rekenregels voor verwachting en variantie te bepalen:

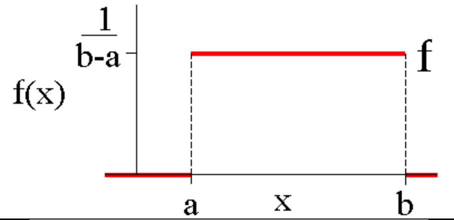
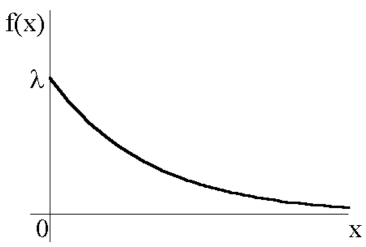
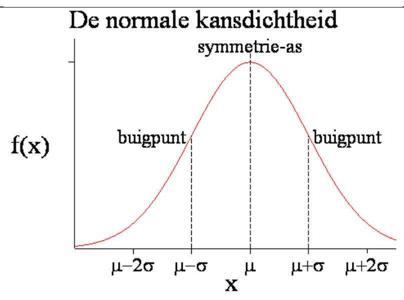
- $E(Y) = E(aX + b) = aE(X) + b = a\mu + b$ en
- $var(Y) = var(aX + b) = a^2 var(X) = a^2\sigma^2$. Dus:

Eigenschap 6.5.5 Voor een $N(\mu, \sigma^2)$-verdeelde stochastische variabele X geldt:
$Y = aX + b$ is $N(a\mu + b, a^2\sigma^2)$-verdeeld (voor $a \neq 0$).

In het volgende hoofdstuk breiden we de eigenschappen van de normale verdeling verder uit door ook lineaire combinaties van 2 o.o. normaal verdeelde stochastische variabelen X en Y te beschouwen.

6.6 Overzicht veel gebruikte continue verdelingen

In de loop van dit hoofdstuk hebben we verschillende continue kansverdelingen de revue laten passeren. De meest gebruikte geven we hieronder in het overzicht weer, met hun verwachting en variantie. In situaties waar sprake is van **random getallen** uit een interval, kunnen we de uniforme verdeling als model gebruiken. Voor **wachttijden**, tussenaankomsttijden en levensduren is vaak de exponentiele verdeling een goed model van de werkelijkheid.
En ten slotte is de normale verdeling veelal een goed model voor meetbare **grootheden in de natuur**, die symmetrisch variëren rondom een gemiddelde. Ook in de techniek, foutenleer, de economie, bedrijfskunde, etc. zijn er vele toepassingen.
In de tabel is de kansdichtheid gegeven voor de waarden van x, waar $f(x)$ *niet* gelijk is aan 0.

Verdeling	*Kansdichtheid*	$E(X)$	$var(X)$	*Grafiek*
Uniform U(a,b)	$f(x) = \frac{1}{b-a}$, $a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	
Exponentieel Exp(λ)	$f(x) = \lambda e^{-\lambda x}$, $x \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	
Normaal $N(\mu, \sigma^2)$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, $x \in \mathbb{R}$	μ	σ^2	

De volgende verbanden zijn ter sprake gekomen:

- Standaardiseren: als $X \sim N(\mu, \sigma^2)$, dan is $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$
- Verband tussen standaardnormaal en normaal:
  als $Z \sim N(0,1)$, dan is $X = \sigma Z + \mu \sim N(\mu, \sigma^2)$
- Een exponentiële verdeling simuleren met random getallen tussen 0 en 1:
  als $X \sim U(0,1)$, dan is $Y = -\frac{\ln(X)}{\lambda} \sim Exp(\lambda)$
- Verband tussen $U(0,1)$ en $U(a,b)$: als $X \sim U(0,1)$ dan is $Y = (b-a)X + a \sim U(a,b)$

6.7 Vraagstukken

1. De kansdichtheid van X wordt gegeven door $f(x) = \begin{cases} 1 - \frac{1}{2}x, & \text{voor } 0 \leq x \leq 2 \\ 0, & \text{elders} \end{cases}$

a. Schets de kansdichtheid en de kans $P(X > 1)$. Bereken vervolgens $P(X > 1)$.
b. Bereken $E(X)$, $E(X^2)$ en $var(X)$.
c. Bepaal de verdelingsfunctie $F(x)$ (let op het definitiegebied!) en bereken opnieuw de kans $P(X > 1)$, nu m.b.v. F.

2. X is exponentieel verdeeld met parameter λ.

a. Geef de kansdichtheid f van X, toon aan dat $\int_{-\infty}^{\infty} f(x)\,dx = 1$ en schets de grafiek.
b. Toon aan dat $E(X) = \frac{1}{\lambda}$ en ga na of $P(X > EX) = \frac{1}{2}$.
c. Bereken de mediaan M, waarvoor geldt dat $P(X > M) = 0.50$.
d. Bepaal de modus van X, d.i. de waarde uit S_X waarvoor f maximaal is.
Markeer de positie van de verwachtingswaarde, de mediaan en modus in de grafiek.

3. X is uniform verdeeld op het interval [0, 4]

a. Beantwoord de vragen van opgave 2 voor deze verdeling.
b. Bepaal ook de verdelingsfunctie van X en schets deze.

4. De kansdichtheid van X wordt gegeven door $f(x) = \frac{c}{x^3}$ als $x > 1$, en $f(x) = 0$ elders.

a. Bepaal c, schets in de grafiek van f de kans $P(X > 2)$ en bereken deze kans.
b. Bereken $E(X)$ en de mediaan M (zodat $P(X \leq M) = P(X \geq M) = \frac{1}{2}$).
c. Bepaal de verdelingsfunctie F_X van X.

5. **a**. Gebruik de afleiding (in 3 stappen) voor de kansdichtheid van $Y = g(X)$ om aan te tonen dat, als $X \sim U(0,1)$, dan is $Y = 5 - 2X \sim U(3,5)$
b. Hoe kun je met een random getal X tussen 0 en 1 (dus $X \sim U(0,1)$) een random getal Y van een gegeven interval (a, b) "genereren"?

6. X is een willekeurig getal tussen 0 en 1 en Y zijn reciproque,
dus X is uniform verdeeld op (0, 1) en $Y = \frac{1}{X}$

a. Bepaal de kansdichtheid van Y.
b. Bepaal op twee manieren, met de verdeling van X resp. Y: $P(Y > 2)$.
c. Bepaal, zo mogelijk, op twee manieren, met de verdeling van X resp. Y: $E(Y)$.

7. X exponentieel verdeeld is met parameter $\lambda = 1$ en $Y = \sqrt{|X|}$

a. Bepaal de verdelingsfunctie van X en gebruik deze om de verdelingsfunctie van Y te bepalen.
b. Bepaal de kansdichtheid van Y en $E(Y)$

8. X is $N(1, 4)$-verdeeld. (*gebruik de N(0, 1)-tabel bij het oplossen van deze opgave*)

a. Schets de grafiek van X en bepaal $P(X > 2)$, $P(|X| > 2)$ en $P(|X - 1| < 2)$.
b. Bepaal het **90ste percentiel** c van de verdeling van X, d.w.z.: c zó dat $P(X \leq c)$ = 90%
c. Bepaal ook het 10de percentiel.

9. (**De empirische regel**). X is $N(\mu, \sigma^2)$-verdeeld.
Ga na dat de kans $P(|X - \mu| < k \cdot \sigma)$ niet afhankelijk is van de waarden van μ en σ en ga na dat de kans voor de waarden k = 1, 2 en 3 resp. 68.3%, 95.4% en 99.7% is.

10. Op een boerderij produceren de kippen eieren waarvan het gewicht normaal verdeeld is met een "gemiddelde" van 50 gram en een standaardafwijking $\sigma = 5$ gram. (Schets de kansdichtheid.)
De boer wil de eieren verkopen in 5 gewichtsklassen, zodat in elke gewichtsklasse evenveel eieren voorkomen. Bereken de klassegrenzen die de boer moet aanhouden.

11. $E(X - \mu)^3$ wordt voor **maten voor de scheefheid** van een verdeling gebruikt: bij symmetrie heeft deze grootheid de waarde 0, bij scheefheid naar rechts is hij positief en bij scheefheid naar links negatief.

a. Druk $E(X - \mu)^3$ uit in het eerste, tweede en derde moment: $E(X)$, $E(X^2)$ en $E(X^3)$.
b. Bepaal de eerste 3 momenten en $E(X - \mu)^3$, als $X \sim U(0,1)$
c. Bepaal de eerste 3 momenten en $E(X - \mu)^3$, als $X \sim Exp(\lambda = 1)$
d. Is de waarde $E(X - \mu)^3$ in b. resp. c. inderdaad 0 resp. positief?

12. (extra opgave m.b.t. functies van een variabele van de vorm $Y = g(X)$.)
a. Laat voor $Z \sim N(0,1)$ zien dat de kansdichtheid van $Y = e^Z$ is:
$$f_Y(y) = \frac{1}{y\sqrt{2\pi}} e^{-\frac{1}{2}(\ln(y))^2}$$
b. (Algemeen) Stel: $Y = g(X)$, waarin X een bekende kansdichtheid f_X heeft en g een monotoon stijgende functie met inverse $u = g^{-1}$, laat dan zien dat
$f_Y(y) = f_X(u(y)) \cdot u`(y)$ (deze formule wordt in module 4 van TBK toegepast)

Enige aanwijzingen bij de opgaven van hoofdstuk 6:

1. Schrijf eerst de formules voor $E(X)$, $E(X^2)$, $var(X)$ en $F(x)$ op.
2. Verwachtingswaarde, Mediaan en modus zijn verschillende "maten voor het midden".
3. idem
4. Zie opgave 1.
5. Vergelijk met de uitwerkingen in voorbeelden 6.4.4. en 6.4.5
6. Idem
7. Idem
8. Vergelijk met de uitgewerkte voorbeelden op pag. 20 en 21
9. Idem
10. Idem
11. a. Werk de macht $(X - \mu)^3$ uit en neem termsgewijs de verwachtingswaarde.
b./c. Pas $E(X^k) = \int_{-\infty}^{\infty} x^k \cdot f(x)dx$ toe.

Hoofdstuk 7 Twee of meer continue variabelen

7.1 Onafhankelijkheid

Bij de **discrete** verdelingen hebben we gezien dat bij afhankelijkheid van X en Y de simultane verdeling bepalend is voor de mate van afhankelijkheid en dat de correlatiecoëfficiënt ρ als maat voor (lineaire) samenhang van die simultane verdeling is. Bij onafhankelijkheid kunnen we de simultane kansen onmiddellijk berekenen uit de marginale kansfuncties:

$$P(X = x \text{ en } Y = y) = P(X = x) \cdot P(Y = y), \qquad \text{voor elk paar } (x, y)$$

Bij **continue** verdelingen is een soortgelijke aanpak mogelijk, maar in dit dictaat zullen we de simultane continue verdelingen slechts aanstippen en ons beperken tot het berekenen van simultane kansen m.b.t. 2 of meer onderling onafhankelijke (o.o.) continue variabelen en het bepalen van de verdeling van de som en het gemiddelde van die variabelen.

We beginnen met het geven van een **algemene definitie voor onafhankelijkheid**:

Definitie 7.1.1 De stochastische variabelen X en Y zijn onderling onafhankelijk als voor elk tweetal verzamelingen $A \subset \mathbb{R}$ en $B \subset \mathbb{R}$ geldt dat

$$\boldsymbol{P(X \in A \text{ en } Y \in B) = P(X \in A) \cdot P(Y \in B)}$$

Deze definitie leidt bij **o.o. discrete** variabelen, door $A = \{x\}$ en $B = \{y\}$ te kiezen, tot de eerder gebruikte definitie, want: $P(X \in \{x\} \text{en } Y \in \{y\}) = P(X \in \{x\}) \cdot P(Y \in \{y\})$ is hetzelfde als

$$P(X = x \text{ en } Y = y) = P(X = x) \cdot P(Y = y).$$

Omgekeerd, als we uitgaan van deze gelijkheid voor alle paren (x, y), dan kunnen we de gelijkheid in definitie 7.1.1 daaruit afleiden: de twee definities zijn voor discrete variabelen equivalent.

Bij **o.o. continue** variabelen X en Y kunnen we simultane kansen direct berekenen uit de gegeven kansverdelingen van X en Y, bijv. $P(X \leq 4 \text{ en } Y > 3) = P(X \leq 4) \cdot P(Y > 3)$.

Op deze wijze kunnen we ook de kansverdeling van het maximum en het minimum van twee o.o. continue variabelen afleiden.

Voorbeeld 7.1.2 Een helpdesk heeft twee medewerkers en als deze allebei bezet zijn worden de volgende klanten in een wachtrij geplaatst. Op zeker moment is een klant als eerste aan de beurt indien één van de klanten die geholpen worden klaar is.
Voor deze situatie gebruiken we de volgende modelveronderstellingen: de bedieningsduren van de twee klanten in bediening zijn o.o. en exponentieel verdeelde variabelen X en Y

(beiden met dezelfde parameter $\lambda = \frac{1}{4}$, dus $E(X) = E(Y) = 4$.
De wachttijd W van de eerste in de wachtrij is de kortste van de twee bedieningsduren, dus $W = \min(X, Y)$.
Wat is de verdeling van W? En hoeveel bedraagt de verwachte wachttijd $E(W)$?

Omdat W een functie is van X en Y kunnen we in ieder geval de verdelingsfunctie als simultane kans opschrijven en dan van de onafhankelijkheid gebruik maken (λ is steeds $\frac{1}{4}$)

$F_W(w) = P(\min(X,Y) \leq w)$
$= 1 - P(\min(X,Y) > w)$ wegens de complementregel
$= 1 - P(X > w \text{ en } Y > w)$
$= 1 - P(X > w) \cdot P(Y > w)$ wegens onafhankelijkheid van
$= 1 - e^{-\lambda w} \cdot e^{-\lambda w}$ omdat $P(X > x) = e^{-\lambda x}$ bij de exponentiele verd.
$= 1 - e^{-2\lambda w}$, voor $w \geq 0$ want $F_W(w) = 0$ voor negatieve waarden van w

Dus $f_W(w) = \frac{d}{dw} F_W(w) = 2\lambda e^{-2\lambda w}$, voor $w \geq 0$
Hierin herkennen we de exponentiële kansdichtheid met parameter $2\lambda = 2 \cdot \frac{1}{4} = \frac{1}{2}$
Dus de verwachte wachttijd is $E(X) = \frac{1}{2\lambda} = 2$. ■

Op soortgelijke wijze kun je de verdeling van het maximum afleiden X en Yafleiden (zie opgave 2) en voor beide afleidingen een generalisatie naar n o.o. variabelen $X_1, X_2, \ldots, X_n$ geven, waarbij de onafhankelijkheid bijvoorbeeld als volgt wordt gebruikt:

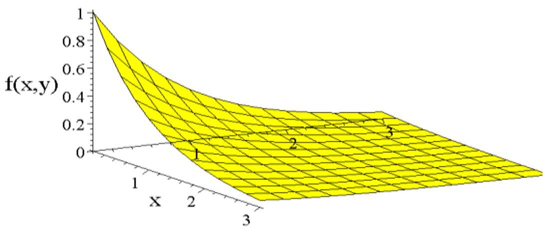
$$P(X_1 \leq x_1 \text{ en } \ldots \text{ en } X_n \leq x_n) = P(X_1 \leq x_1) \cdot \ldots \cdot P(X_n \leq x_n)$$

De aanpak die we in voorbeeld 7.1.2 hanteerden, "werkt" niet bij functies als $X \cdot Y$ en $X + Y$. De gebeurtenis $\{X + Y \leq w\}$ kunnen we namelijk niet uitdrukken in onafhankelijke gebeurtenissen van de vorm $X \in A$ en $Y \in B$. In zijn algemeenheid gebruiken we voor dat soort problemen simultane continue verdelingen. Hieronder geven we daarvan kort een overzicht dat niet tot de leerstof van dit vak behoort om vervolgens in paragraaf 7.2 de afleiding van de kansdichtheid van $X + Y$ voor o.o. s.v.-en te bespreken.

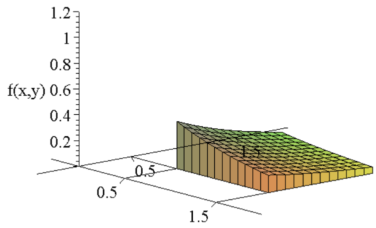
Intermezzo: de continue simultane verdeling van X en Y aan de hand van een voorbeeld.
Analoog aan de kansdichtheid van één continue variabele kunnen we een simultane kansdichtheid $f(x, y)$ definiëren die we grafisch in 3 dimensies kunnen schetsen. Een kans is dan een volume boven een gebied in het xy-vlak onder de simultane kansdichtheid, bijvoorbeeld:

$$f(x,y) = \begin{cases} e^{-x-y} & \text{als } x \geq 0 \text{ en } y \geq 0 \\ 0 & \text{elders} \end{cases}$$

Hieronder is f geschetst en ook de kans $P(0.5 \leq X \leq 1.5 \;\; en \;\; 0.5 \leq Y \leq 1.5)$



de simultane kans $P(0.5<X<1.5 \text{ en } 0.5<Y<1.5)$



De feitelijke berekening van de kans betreft de dubbele integraal $\int_{0.5}^{1.5}\int_{0.5}^{1.5} e^{-x-y}\,dxdy$, een techniek die in elk calculusboek beschreven staat. Op die wijze kunnen we ook de **simultane verdelingsfunctie $\boldsymbol{F(x,y)}$** van het paar (X,Y) bepalen. Deze definiëren we als

$$F(x,y) = P(X \le x \text{ en } Y \le y)$$

De integraalberekening in ons voorbeeld levert

$$F(x,y) = 1 - e^{-x} - e^{-y} + e^{-x-y}, \quad \text{als } x \ge 0 \text{ en } y \ge 0$$

Omdat bijvoorbeeld $P(X \le x) = \lim\limits_{y\to\infty} P(X \le x \; en \; Y \le y)$, kunnen we hiermee de **marginale verdeling** van X bepalen: $F_X(x) = \lim\limits_{y\to\infty} F(x,y) = 1 - e^{-x}$, voor $x \ge 0$

En dus $\quad f_X(x) = \frac{d}{dx}F_X(x) = e^{-x}$, voor $x \ge 0$

We herkennen de exponentiële kansdichtheid met $\lambda = 1$, dus $E(X) = 1 = var(X)$

Voor Y vinden we analoog $f_Y(y) = e^{-y}$, voor $y \ge 0$: dezelfde verdeling als voor X.

Ook bij continue verdelingen kunnen we voorwaardelijke verdelingen definiëren, analoog aan het discrete geval.

De **voorwaardelijke kansdichtheid** van X, gegeven $Y = y > 0$, is toegepast op het voorbeeld: $\quad f_X(x|Y=y) = \frac{f(x,y)}{f_Y(y)} = \frac{e^{-x-y}}{e^{-y}} = e^{-x}$, voor $x \ge 0$

En **voorwaardelijke verwachting** $E(X|Y=y) = \int_{-\infty}^{\infty} x f_X(x|Y=y)dx = \int_{-\infty}^{\infty} xe^{-x}dx = 1$

In de berekening van het voorbeeld is blijkbaar $f_X(x|Y=y) = f_X(x)$, dus het gegeven dat $Y = y$ beïnvloedt de verdeling van X niet: X en Y zijn in het voorbeeld onafhankelijk.

Dit blijkt ook uit de volgende gelijkheden bij onafhankelijkheid (voor $x \ge 0$ en $y \ge 0$):

$$F(x,y) = 1 - e^{-x} - e^{-y} + e^{-x-y} = (1-e^{-x})(1-e^{-y}) = F_X(x) \cdot F_Y(y)$$

En: $\quad f(x,y) = e^{-x-y} = e^{-x} \cdot e^{-y} = f_X(x) \cdot f_Y(y)$

Dus bij onafhankelijkheid geldt ook voor de **simultane verdelingsfunctie** en de **simultane kansdichtheid** dat zij het product van de **marginale verdelingsfuncties** resp. **kansdichtheden** zijn. ■

7.2 De convolutie-integraal

De aanpak zoals beschreven in het intermezzo maakt het mogelijk om de kansdichtheid van $X+Y$ voor twee o.o. continue stochastische variabelen te bepalen, via de verdelingsfunctie: $F_{X+Y}(z) = P(X+Y \le z)$, te bepalen m.b.v. een dubbele integraal van $f_X(x) \cdot f_Y(y)$ op het half-vlak $x + y \le z$.

De afgeleide van $F_{X+Y}(z)$ levert vervolgens een uitdrukking voor de kansdichtheid $f_{X+Y}(z)$.

Het resultaat is:

Eigenschap 7.2.1 (de convolutie-integraal)

Als X en Y o.o. en continue variabelen zijn, dan geldt:

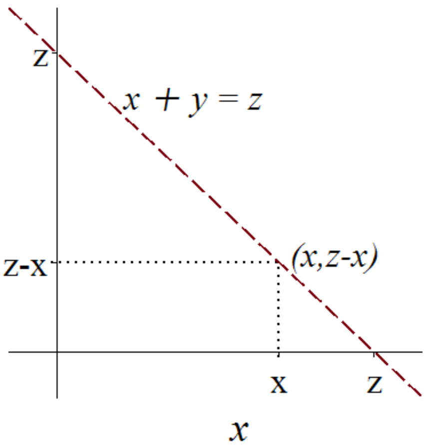
$$\boldsymbol{f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)dx}$$

De analogie met de convolutiesom dringt zich op:

- Bij o.o. en variabelen X en Y bepaalden we de kans op $X + Y = z$ door, voor de roosterpunten (x, y) op de lijn $x + y = z$, de kansen $P(X = x)P(Y = y)$ bij elkaar op te tellen (zie figuur).
- Bij continue X en Y integreren we (Riemann-som!) de *kansdichtheden* $f_X(x) \cdot f_Y(y)$ over de lijn $x + y = z$:

$$P(X + Y = z) = \sum_x P(X = x)P(Y = z - x)$$

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)dx$$



In plaats van de som van o.o. stochastische variabelen spreken we ook wel van de "convolutie" van die variabelen

Voorbeeld 7.2.2 (convolutie van o.o. exponentieel verdeelde variabelen)

"Er is nog één wachtende voor u" .Wat betekent dat voor de totale tijd die ik nog moet wachten en wat is de verwachte wachttijd?

Laten we veronderstellen dat er één medewerker is die klanten op volgorde van binnenkomst bedient. "Één wachtende voor ons" betekent dat er bij binnenkomst één in bediening is (bedieningsduur X) en een tweede nog bediend moet worden (duur Y), alvorens ik aan de beurt ben. Een kansmodel voor deze situatie zou kunnen luiden:

Model: de bedieningsduren X en Y zijn o.o. en beiden exponentieel verdeeld met parameter λ.

Dit model houdt onder meer in dat $f_X(x) = f_Y(x) = \lambda e^{-\lambda x}$ $(x \geq 0)$ en dat de convolutie-integraal van toepassing is op mijn totale wachttijd X + Y:

$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)dx$ Hierin is $f_X(x) = 0$ als $x < 0$
en $f_Y(z-x) = 0$ als $x > z$

$= \int_0^z \lambda e^{-\lambda x} \lambda e^{-\lambda(z-x)} dx$

$= \int_0^z \lambda^2 e^{-\lambda z} dx$ Merk op dat de te integreren functie een constante is.

$= \lambda^2 e^{-\lambda z} \cdot x|_{x=0}^{x=z}$

$= \lambda^2 z e^{-\lambda z}$, voor $z \geq 0$
En $f_{X+Y}(z) = 0$ als $z < 0$.

Deze verdeling staat bekend als de Erlang verdeling met parameters $n = 2$ en λ.
De som $X + Y + Z$ van o.o. Exp(λ)-verdeelde variabelen is Erlang-verdeeld met parameters $n = 3$ en λ: de verdeling kunnen we weer met de convolutie-integraal voor $X + Y$ en Z te bepalen. Dit wordt verder besproken in het hoofdstuk over wachttijden.

Ten aanzien van de berekening van $E(X + Y)$ en $var(X + Y)$ kunnen we gebruik maken van de zojuist bepaalde kansdichtheid $f_{X+Y}(z)$, maar eenvoudiger is gebruik te maken van de rekenregels: $E(X + Y) = E(X) + E(Y) = \frac{1}{\lambda} + \frac{1}{\lambda} = \frac{2}{\lambda}$
En, wegens o.o.-heid $var(X + Y) = var(X) + var(Y) = \frac{1}{\lambda^2} + \frac{1}{\lambda^2} = \frac{2}{\lambda^2}$ ■

7.3 Sommen van o.o. en normaal verdeelde variabelen

Als X en Y o.o. en beiden normaal verdeeld zijn, is dan de som $X + Y$ ook normaal verdeeld? Ja! En we kunnen dat aantonen met de convolutie-integraal, zoals we in deze paragraaf in een relatief eenvoudig voorbeeld ook zullen uitwerken. Het antwoord op de vraag wat dan de parameters μ en σ^2 van de normale verdeling zijn, is eenvoudiger te geven:

$$\mu = E(X + Y) = E(X) + E(Y) \quad \text{ofwel} \quad \mu_{X+Y} = \mu_X + \mu_Y$$

En $$\sigma^2 = var(X + Y) = var(X) + var(Y) \quad \text{ofwel} \quad \sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

Voorbeeld 7.3.1 Voor een gepland benzinestation gaat men er vanuit dat de dagelijkse vraag naar superbenzine normaal verdeeld is met verwachte vraag 600 *l* en een standaardafwijking van 100 *l*. Verder neemt men aan dat de vraag voor verschillende dagen onderling onafhankelijke stochastische variabelen zijn. Het benzinestation zal eens per week bevoorraadt moeten worden. Hoe groot moet de capaciteit van de tank voor superbenzine zijn opdat de kans dat de voorraad opraakt (in één week tijd) 5% is?
Kansmodel:
de dagelijkse vragen om superbenzine $X_1, X_2, \ldots, X_7$ zijn o.o. en $N(600, 100^2)$-verdeeld.

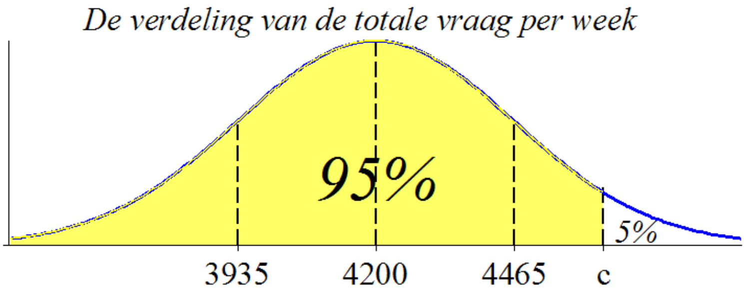
De totale vraag $T = X_1 + X_2 + \cdots + X_7$ is dan ook normaal verdeeld met

$$\mu_T = E(X_1 + \cdots + X_7) = E(X_1) + \cdots + E(X_7) = 7 \cdot 600 = 4200 \text{ en}$$

$$\sigma_T^2 = var(X_1 + \cdots + X_7) = var(X_1) + \cdots + var(X_7) = 7 \cdot 100^2 = 70000$$

Dus $\sigma_T = \sqrt{70000} \approx 265\ l$

Hieronder is de grafiek van T geschetst met de capaciteit *c* van de tank die aan de eis voldoet



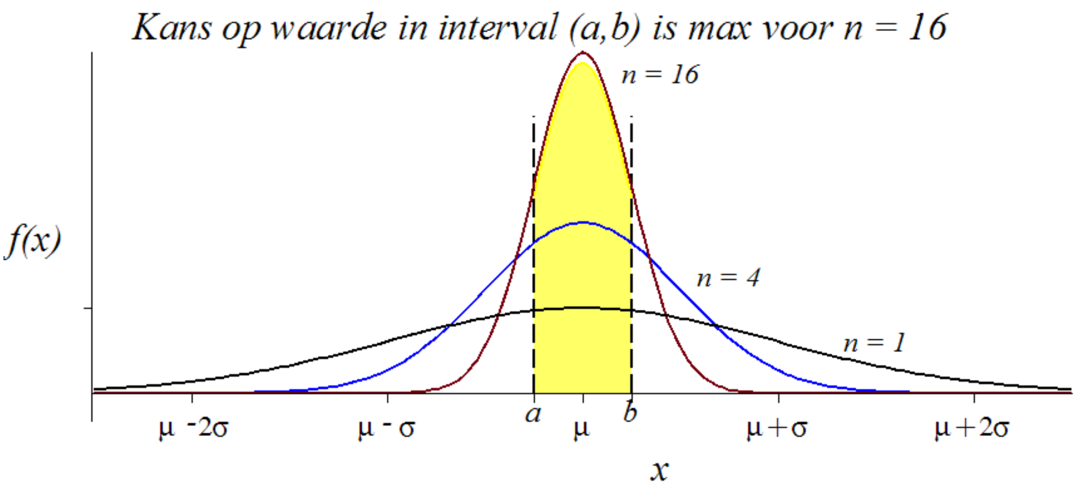
De tankcapaciteit c moet zodanig zijn dat $P(T > c) \leq 5\%$ ofwel (ga over op de N(0,1)-verd.):
$P(T \leq c) = P\left(Z \leq \frac{c-4200}{265}\right) \geq 95\%$
Uit de N(0, 1)-tabel volgt dat $\frac{c-4200}{265} \geq 1.645$, dus: $c \geq 4200 + 1.645 \cdot 265 = 4636\ l$
Opmerking: binnen het vakgebied van "Voorraad Beheer" (*Supply chain management*) spreekt men van een *safety stock* van 4636 liter bij een *service level* van 95%. ■

Naast de som is het **steekproefgemiddelde** van o.o. en $N(\mu, \sigma^2)$-verdeelde $X_1, \ldots, X_n$ belangrijk, zoals in de statistiek bij een aselecte steekproef uit $N(\mu, \sigma^2)$:

$$\boldsymbol{\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i}$$

Omdat de som $X_1 + \cdots + X_n$ normaal verdeeld is, geldt dat ook voor $\overline{X}_n$, met parameters:
$\mu_{\overline{X}_n} = E\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n}E(X_1 + \cdots + X_n) = \frac{1}{n} \cdot n\mu = \mu$ en
$\sigma^2_{\overline{X}_n} = var\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n^2}var(X_1 + \cdots + X_n) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$, dus $\sigma_{\overline{X}_n} = \frac{\sigma}{\sqrt{n}}$

Doordat $var(\overline{X}_n) = \frac{\sigma^2}{n}$ afneemt als n, het aantal metingen, toeneemt wordt de kans groter om voor $\overline{X}_n$ een realisatie te verkrijgen die dicht in de buurt van μ ligt. In onderstaande figuur is de verdeling van het gemiddelde geschetst, voor $n = 1$, $n = 4$ en $n = 16$. Voor $n = 16$ is de kans dat het gemiddelde in een klein interval (a, b) rondom μ ligt het grootst.



Als de variabelen normaal verdeeld en onderling onafhankelijk zijn (zoals bij “aselecte steekproeven uit een normaal verdeelde populatie” het geval is), zijn sommen als $X+Y$ en $X_1+\cdots+X_n$ en ook de bijbehorende gemiddelden als $\frac{X+Y}{2}$ en $\overline{X}_n$ normaal verdeeld.
Dit is in zijn algemeenheid analytisch lastig aan te tonen. We beperken ons daarom tot het eenvoudigste geval van twee o.o. en standaardnormale variabelen: dan is het al lastig genoeg. Voorbeeld 7.3.2 hoort echter niet tot de examenstof.

Voorbeeld 7.3.2 Als X en Y o.o. zijn en beide $N(0,1)$-verdeeld zijn, dan is $X+Y$ volgens de convolutie-integraal (eigenschap 7.2.1) ook normaal verdeeld, want

$$f_{X+Y}(z)=\int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\ \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-x)^2}dx$$

In de exponent kunnen we als volgt “kwadraat afsplitsen”:

$$-\tfrac{1}{2}x^2-\tfrac{1}{2}(z-x)^2=-\left[x^2-zx+-\tfrac{1}{2}z^2\right]=-\left(x-\tfrac{1}{2}z\right)^2-\tfrac{1}{4}z^2$$

Dus

$$f_{X+Y}(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{4}z^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-(x-\frac{1}{2}z)^2}dx=\frac{1}{\sqrt{2\pi\cdot 2}}e^{-\frac{1}{2}\cdot\frac{z^2}{2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\cdot\frac{1}{2}}}e^{-\frac{1}{2}\cdot\frac{\left(x-\frac{1}{2}z\right)^2}{\frac{1}{2}}}dx$$

Onder de laatste integraal staat nu een normale kansdichtheid, namelijk van de $N\left(\frac{1}{2}\mathrm{z},\frac{1}{2}\right)$-verdeling: de totale oppervlakte onder een kansdichtheid is 1. Zo vinden we:

$$f_{X+Y}(z)=\frac{1}{\sqrt{2\pi\cdot 2}}e^{-\frac{1}{2}\cdot\frac{z^2}{2}},\ \text{ de N(0,2)} - \text{kansdichtheid}$$

Dus als X en Y o.o. en $N(0,1)$-verdeeld zijn, is $X+Y$ $\mathrm{N}(0+0,1+1)$-verdeeld.
Met eigenschap 6.4.3 is eenvoudig na te gaan dat ook het gemiddelde $\frac{X+Y}{2}$ normaal verdeeld is, volgens een $N\left(0,\frac{1}{2}\right)$-verdeling. Merk op dat de variantie $\frac{\sigma^2}{n}=\frac{1}{2}$ is. ■

Wat voor twee o.o. en standaardnormaal verdeelde variabelen geldt, kun je ook bewijzen voor twee normaal verdeelde variabelen: zowel de som $X+Y$ als het verschil $X-Y$ is ook normaal verdeeld. Met de rekenregels voor verwachting en variantie kunnen we eenvoudig de parameters (μ en σ^2) bepalen:

$$E(X+Y)=E(X)+E(Y)\quad \text{en} \quad E(X-Y)=E(X)-E(Y)$$

$$var(X+Y)\overset{\text{o.o.}}{=}var(X)+var(Y)\quad \text{en} \quad var(X-Y)\overset{\text{o.o.}}{=}var(X)+var(Y)$$

Eigenschap 7.3.3 Als $X\sim N(\mu_X,\sigma_X^2)$ en $Y\sim N(\mu_Y,\sigma_Y^2)$ en X en Y zijn o.o., dan geldt:

a. $\boldsymbol{X+Y\ \sim N\left(\mu_X+\mu_Y,\sigma_X^2+\sigma_Y^2\right)}$
b. $\boldsymbol{X-Y\ \sim N\left(\mu_X-\mu_Y,\sigma_X^2+\sigma_Y^2\right)}$

Voorbeeld 7.3.4
Aandelenmarkt-deskundigen hebben het volgende model opgesteld voor het jaarrendement van twee aandelenfondsen (jaarrendement X voor agrarische producten en Y voor olieproducten):

- $X \sim N(8, 36)$ en $Y \sim N(12, 64)$
- X en Y zijn o.o.

Het verwachte rendement voor de olieproducten is 50% hoger dan dat van agrarische producten, maar hoe groot is de kans dat na één jaar blijkt dat het rendement van de agrarische producten hoger is?
Oplossing: we moeten de kans $P(X > Y)$ berekenen: dat lukt niet door bijv. X en Y afzonderlijk te standaardiseren, maar wel door te stellen dat $P(X > Y) = P(X - Y > 0)$.
Volgens eigenschap 7.3.3b geldt: $X - Y \sim N(8 - 12, 36 + 64)$, dus:

$$P(X > Y) = P(X - Y > 0) = P\left(Z > \frac{0 - (-4)}{\sqrt{100}}\right) = 1 - \Phi(0.4) = 1 - 0.6554 = 34.46\% \blacksquare$$

Algemeen geldt voor de som van n variabelen:

Eigenschap 7.3.5
Als $X_i \sim N(\mu_i, \sigma_i^2)$ voor $i = 1, 2, \dots, n$ en $X_1, X_2, \dots, X_n$ zijn o.o., dan geldt voor $S_n = \sum_{i=1}^n X_i$

a. $\boldsymbol{E(S_n) = \sum_{i=1}^n \mu_i}$ en $\boldsymbol{var(S_n) = \sum_{i=1}^n \sigma_i^2}$
b. $\boldsymbol{S_n \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)}$

Eigenschap 7.3.5a. geldt uiteraard ook voor niet-normale verdelingen.
We merken op dat $\sigma_{S_n} \neq \sigma_1 + \cdots + \sigma_n$, maar $\sigma_{S_n} = \sqrt{(\sigma_1)^2 + \cdots + (\sigma_n)^2}$.
We moeten dus onthouden dat de som van o.o. en normaal verdeelde variabelen ook weer normaal verdeeld is, met de som van verwachtingen (de μ's) en de som van varianties (de σ^2-en) als parameters.
Eigenschap 7.3.5 is eenvoudig toe te passen op sommen en gemiddelden bij aselecte steekproeven uit een normale verdeling:

Eigenschap 7.3.6
Als $X_i \sim N(\mu, \sigma^2)$ voor $i = 1, 2, \dots, n$ en $X_1, X_2, \dots, X_n$ zijn o.o., dan geldt

a. voor $S_n = \sum_{i=1}^n X_i$ dat $\boldsymbol{S_n \sim N(n\mu, n\sigma^2)}$ en
b. voor $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ dat $\boldsymbol{\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)}$

Voorbeeld 7.3.7
Een lift mag maximaal 1000 kg vervoeren. Hoe groot is het maximale aantal personen dat deze lift mag vervoeren, opdat de kans op overbelasting maximaal 1% is, als we aannemen dat het gewicht van een liftgebruiker $N(75{,}100)$-verdeeld is?
We modelleren deze situatie als volgt. Als het maximale aantal personen n is, dan zijn hun gewichten $X_1, X_2, \dots, X_n$ o.o. en alle $N(75{,}100)$-verdeeld. Dus volgens eigenschap 7.3.6 heeft het totale gewicht $\sum_{i=1}^n X_i$ dan de $N(n \cdot 75, n \cdot 100)$-verdeling.

We willen n zodanig kiezen dat de kans op overbelasting $P(\sum_{i=1}^{n} X_i > 1000)$ maximaal 1% is, ofwel zodat $P(\sum_{i=1}^{n} X_i \leq 1000) > 0.99$.
We kunnen nu door standaardisatie overgaan op de standaardnormale verdeling:

$$P\left(\sum_{i=1}^{n} X_i \leq 1000\right) = P\left(\frac{\sum_{i=1}^{n} X_i - 75n}{\sqrt{100n}} \leq \frac{\sum_{i=1}^{n} X_i - 75n}{\sqrt{100n}}\right) = \Phi\left(\frac{1000-75n}{\sqrt{100n}}\right) > 0.99$$

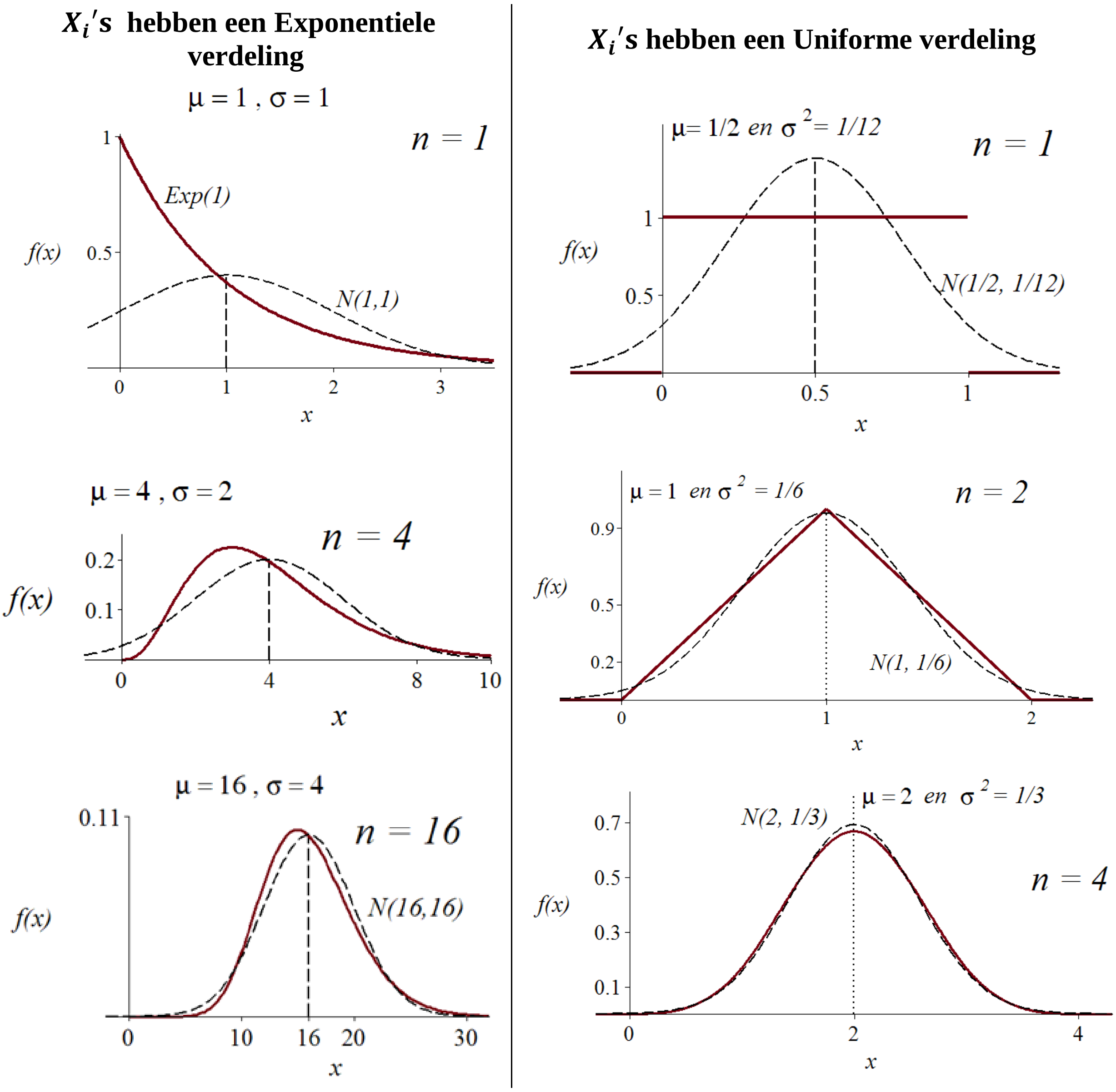
gebruik makend van de N(0,1)-tabel vinden we: $\frac{1000-75n}{\sqrt{100n}} > 2.33$.
De oplossing kan gevonden worden door te kwadrateren, maar we kunnen ook de gehele waarden van n invullen ($n < \frac{1000}{75}$) . Zo vinden we $\boldsymbol{n = 12.}$
Bij 12 personen is het verwachte totale gewicht $12 \cdot 75 = 900$ en de kans op overbelasting

$$1 - \Phi\left(\frac{1000-900}{\sqrt{1200}}\right) \approx 1 - \Phi(2.89) = 0.19\% \quad (< 1\%) \qquad \blacksquare$$

7.4 De Centrale Limietstelling

In de voorgaande paragraaf hebben we gezien dat de som van n o.o. en $N(\mu, \sigma^2)$- verdeelde variabelen ook normaal verdeeld is met verwachtingswaarde nμ en variantie $n\sigma^2$.
Als $X_1, X_2, \ldots, X_n$ niet normaal verdeeld zijn, maar **wel o.o. en gelijk verdeeld met verwachtingswaarde μ en variantie $\boldsymbol{\sigma^2}$**, geldt weliswaar voor $S_n = \sum_{i=1}^{n} X_i$ dat $E(S_n) = n\mu$ en $var(S_n) = n\sigma^2$, maar de normale verdeling geldt niet S_n.
De gestandaardiseerde $\frac{S_n - n\mu}{\sqrt{n\sigma^2}}$ is in dat geval dus **niet standaardnormaal** verdeeld

Kansen m.b.t. S_n zullen we in dat geval moeten berekenen door eerst de verdeling van S_n te bepalen. Als bijvoorbeeld n klanten door een loketbeambte bediend moeten worden en de bedieningsduren $X_1, X_2, \ldots, X_n$ als o.o. en exponentieel verdeelde stochastische variabelen (alle met parameter $\lambda = 1$), dan kunnen we de kansdichtheid totale bedieningsduur $S_n = \sum_{i=1}^{n} X_i$ bepalen door herhaald toepassen van de convolutie-integraal. (We komen hierop terug in het volgende hoofdstuk.)
De verkregen kansdichtheden hebben we in onderstaande figuur (linker kolom) vergeleken met de normale kansdichtheden, waarvoor $\mu = E(S_n) = n$ en $\sigma^2 = var(S_n) = n$.
In de rechter kolom is hetzelfde gedaan voor o.o. en $U(0,1)$-verdeelde getallen $X_1, X_2, \ldots, X_n$



Het beeld dat uit deze plaatjes opduikt is dat de verdeling van de som meer neigt naar een normale verdeling naarmate het aantal n groter wordt: het is een belangrijk resultaat van de kansrekening dat voor elke willekeurige verdeling van de X_i's, de som $S_n = \sum_{i=1}^{n} X_i$ "voor grote n" bij benadering normaal verdeeld is, of, preciezer gezegd, dat de **gestandaardiseerde** $\boldsymbol{\frac{S_n - n\mu}{\sqrt{n\sigma^2}}}$ **bij benadering** ***N*****(0, 1)-verdeeld** is.
Dit is een gevolg van de volgende limiet-uitspraak (die we niet bewijzen).

Eigenschap 7.4.1 De Centrale Limiet Stelling (CLS)
Indien $X_1, X_2, \ldots$ een rij o.o. en gelijk verdeelde stochastische variabelen is, met verwachtingswaarde μ en variantie $\sigma^2 > 0$, geldt voor $S_n = \sum_{i=1}^{n} X_i$

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq z\right) = \Phi(z),$$

waarin **Φ de standaardnormale verdelingsfunctie** is.

Gevolg eigenschap 7.4.1: als n "voldoende groot" is, geldt:

- $\frac{S_n - n\mu}{\sqrt{n\sigma^2}}$ is bij benadering $N(0,1)$-verdeeld
- S_n is bij benadering $N(n\mu, n\sigma^2)$-verdeeld
- $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is bij benadering $N\left(\mu, \frac{\sigma^2}{n}\right)$-verdeeld

Of n "voldoende groot" is, is afhankelijk van de gewenste nauwkeurigheid en van het type verdeling van de X_i's: uit de grafieken op voorgaande bladzijden blijkt dat f_{S_n} sneller "convergeert" naar de normale kansdichtheid bij de (symmetrische) uniforme verdeling dan bij de (asymmetrische) exponentiële verdeling.
Voor praktische toepassing zullen de volgende algemene regel hanteren:

Vuistregel voor normale benadering m.b.v. de CLS: $\boldsymbol{n \geq 25}$

Voorbeeld 7.4.2
Hoe groot is de kans dat de som van 50 random getallen tussen 0 en 1 kleiner is dan 24?
Het **kansmodel** voor de 50 random getallen wordt gevormd door de variabelen $X_1, X_2, \ldots, X_{50}$, die o.o. en alle uniform verdeeld is op [0,1].
Omdat $n = 50$ voldoende groot is om de CLS toe te passen, kunnen we bij benadering een normale verdeling voor de som $S_{50} = \sum_{i=1}^{50} X_i$ gebruiken: de parameters worden gegeven door $n\mu = 50 \cdot \frac{1}{2} = 25$ en $n\sigma^2 = 50 \cdot \frac{1}{12} = \frac{25}{6}$, gebruikmakend van de μ en σ^2 van de $U(0,1)$-verdeling. Dus de normale benadering van de kans $P(S_{50} < 24)$ is:

$$P(S_{50} < 24) = P\left(\frac{S_{50} - 25}{\sqrt{25/6}} \leq \frac{24 - 25}{\sqrt{25/6}}\right) \overset{\text{CLS}}{\approx} \Phi(-0.49) = 1 - \Phi(0.49) = 31.21\%$$

We spreken ook wel van de **normale benadering van de kans** $P(S_{50} < 24)$.
Overigens kan men de vraag in dit voorbeeld ook als volgt formuleren: hoe groot is de kans dat het gemiddelde van 50 random getallen tussen 0 en 1 kleiner is dan 0.48 (= $\frac{24}{50}$)?
Immers, ook op $\overline{X}_{50} = \frac{1}{50} S_{50}$ is de CLS van toepassing (je kunt $\overline{X}_{50}$ ook als som $\sum_{i=1}^{n}\left(\frac{X_i}{50}\right)$ van gelijk verdeelde variabelen schrijven): $\overline{X}_{50}$ is bij benadering $N\left(\mu, \frac{\sigma^2}{n}\right)$-verdeeld, dus $N\left(\frac{1}{2}, \frac{\frac{1}{12}}{50}\right)$-verdeeld. Dus:

$$P(S_{50} < 24) = P\left(\overline{X}_{50} < 0.48\right) \overset{\text{CLS}}{\approx} \mathrm{P}\left(Z \leq \frac{0.48 - 0.50}{\sqrt{1/600}}\right) \approx \Phi(-0.49) = 31.21\% \qquad \blacksquare$$

De CLS verklaart waarom de normale verdeling niet alleen als kansmodel voor grootheden "uit de natuur" veelvuldig wordt gebruikt, maar ook in de statistiek waar het gaat om gemiddelden van steekproefwaarden en in de fysica, waar meetfouten veelal m.b.v. de normale verdeling worden gemodelleerd. Een meetfout kan namelijk vaak beschouwd worden als de som van een groot aantal onafhankelijke elementaire foutjes van diverse origine, waarop de CLS van toepassing is (of de gegeneraliseerde versie van de CLS, die Bessel in 1838 bewees voor niet gelijk verdeelde elementaire foutjes).

De CLS geldt ook voor discreet verdeelde stochastische variabelen. Als X binomiaal verdeeld is met parameters n en p, kunnen we X opvatten als een som van n o.o. en $B(1,p)$-verdeelde stochastische variabelen X_i ("alternatieven"). Dus $X = \sum_{i=1}^{n} X_i$.
Indien n voldoende groot is, is X volgens de CLS bij benadering normaal verdeeld met de binomiale verwachtingswaarde $E(X) = n \cdot p$ en variantie $var(X) = n \cdot p(1-p)$. Hierin zijn p en $p(1-p)$ de verwachting van de $B(1,p)$-verdeling van de X_i's.

Eigenschap 7.4.3 (**gevolg van CLS: normale benadering van de binomiale verdeling**)

Als $\boldsymbol{X \sim B(n,p)}$**,** dan is $\boldsymbol{X}$ voor **voldoende grote** $\boldsymbol{n}$ bij benadering $\boldsymbol{N(np, np(1-p))}$

Voor waarden van n en p, waarvoor tabellen beschikbaar zijn, gebruiken we uiteraard deze exacte waarden. Eerder zagen we al dat voor grote waarden van n en waarden van p dichtbij 0 of 1 een benadering m.b.v. de Poisson-verdeling gerechtvaardigd is (zie hoofdstuk 4). De normale benadering van de binomiale kansfunctie kan gebruikt worden indien n voldoende groot is en p niet te dicht bij 0 of 1.

Vuistregels voor de benadering van de binomiale verdeling:

- $\boldsymbol{n \geq 25}$
- Gebruik de **Poisson-benadering** met $\mu = np$, als $\boldsymbol{np \leq 10}$ **of** $\boldsymbol{n(1-p) \leq 10}$
- Gebruik de **normale benadering** volgens de CLS met $\mu = np$ en $\sigma^2 = np(1-p)$
  als $\boldsymbol{np > 5}$ **èn** $\boldsymbol{n(1-p) > 5}$

We zien dat binnen een bepaalde bandbreedte beide benaderingen mogelijk zijn.
In alle andere gevallen ($n < 25$) gebruiken we de binomiale verdeling exact of de binomiale tabel indien beschikbaar.
Een goede normale benadering dient met continuïteitscorrectie plaats te vinden, zoals in het volgende voorbeeld aannemelijk wordt gemaakt. We zullen daarbij n op de grens van de voorwaarde $n \geq 25$ zodat we ter vergelijking ook de exacte waarde in de tabel kunnen opzoeken.

Voorbeeld 7.4.4 Iemand tracht het theorie-examen voor het rijbewijs te halen. Tijdens zijn examen blijkt hij 45 van de 70 ja/nee vragen met zekerheid te weten. We gaan er vanuit dat die 45 antwoorden ook correct zijn. De overige 25 vragen besluit hij "op de gok" te beantwoorden (bijvoorbeeld door een zuivere munt op te werpen).
Hoe groot is de kans dat hij slaagt, als daarvoor minstens 60 juiste antwoorden nodig zijn?
Zij X het aantal goede antwoorden op de 25 gegokte vragen: dan is X B(25, 0.5)-verdeeld.
We kunnen een normale benadering toepassen omdat $n = 25 \geq 25$ en $np = n(1-p) = 12.5 > 5$.
Dus volgens de CLS is X bij benadering $N(12.5, 6.25)$-verdeeld ($\sigma^2 = np(1-p) = 6.25$).
Hij slaagt als hij minstens 15 van de 25 gegokte antwoorden juist heeft, dus:

$$P(X \geq 15) \overset{\text{CLS}}{\approx} P\left(Z \geq \frac{15 - 12.5}{\sqrt{6.25}}\right) = 1 - \Phi(1) = 0.1587$$

Als alternatieve berekening hadden we kunnen kiezen: hij slaagt indien meer dan 14 antwoorden goed gegokt zijn ($X \geq 15$ is hetzelfde als $X > 14$):

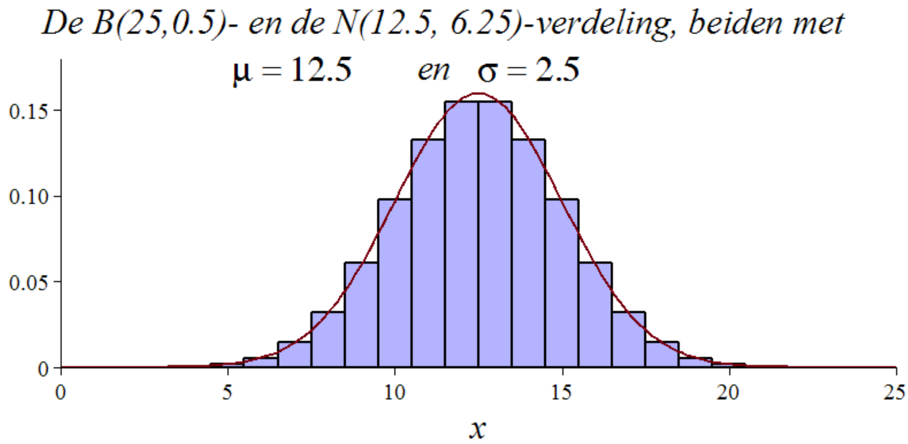
$$P(X > 14) \overset{\text{CLS}}{\approx} P\left(Z \geq \frac{14 - 12.5}{\sqrt{6.25}}\right) = 1 - \Phi(0.60) = 0.2743$$

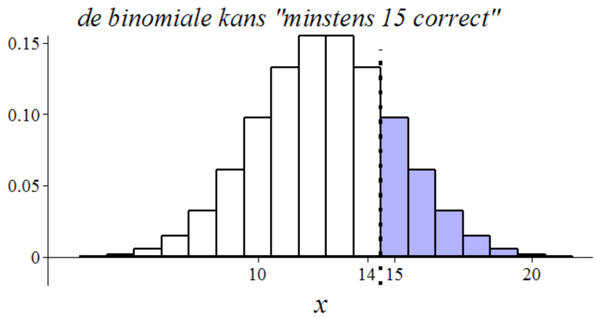
Beide kansen benaderen de gevraagde binomiale kans, maar het verschil is maar liefst 11.5%! De verschillende berekeningen nalopend is dit een gevolg van het feit dat X slechts gehele waarden kan aannemen en dus $P(X \geq 15) = P(X > 14)$, maar de bijbehorende z-scores 1 en 0.6 geven bij de continue standaardnormale verdeling aanleiding voor het grote verschil.

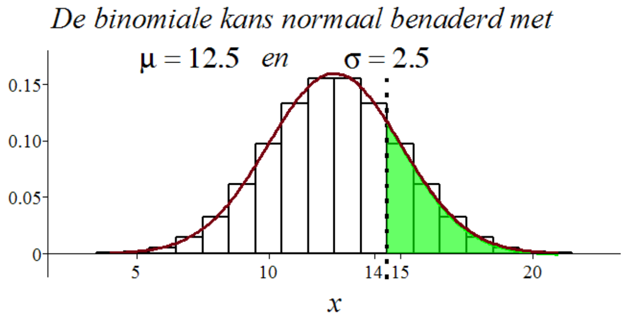
Uit de *B*(25, 0.5)-tabel blijkt de exacte waarde:

$$P(X \geq 15) = 1 - P(X \leq 14) = 1 - 0.7878 = 0.2121$$

Bestudering van de volgende grafieken leidt tot de conclusie dat voor de beste benadering niet de *z*-score van 15 of van 14 moet worden bepaald maar van de tussengelegen waarde 14.5.







$$P(X \geq 15) \overset{\text{cont. corr.}}{=} P(X \geq 14.5) \overset{\text{CLS}}{\approx} P\left(Z \geq \frac{14.5 - 12.5}{\sqrt{6.25}}\right) = 1 - \Phi(0.80) = 0.2119$$

Inderdaad: continuïteitscorrectie geeft in dit geval (verreweg) de beste benadering. ■

Hierboven hebben we gezien hoe en waarom continuïteitscorrectie (c.c.) wordt toegepast: we gaan over van een discrete (binomiale) op een continue (normale) verdeling.
Voor benadering van de kans op $X \geq 15$, hebben we het interval (14.5, 15.5) rondom de grenswaarde 15 meegenomen in de berekening, dus $P(X \geq 15) \overset{\text{c.c.}}{=} P(X \geq 14.5)$.
De gebeurtenis $X > 14$ impliceert eveneens dat 15 de grenswaarde is, omdat X slechts gehele waarden kan aannemen: $P(X > 14) \overset{\text{c.c.}}{=} P(X \geq 14.5)$

Normale benadering van de binomiale verdeling met continuïteitscorrectie:
Indien

- X $B(n,p)$-verdeeld is voor voldoende grote $n \geq 25$ met $np > 5$ en $n(1-p) > 5$
- Y heeft de benaderende $N(np, np(1-p))$-verdeling,

Dan berekenen we benaderingen van de binomiale kansen altijd **met continuïteitscorrectie**:

- $P(X \leq k) \overset{\text{c.c.}}{=} P\left(X \leq k + \frac{1}{2}\right) \overset{\text{CLS}}{\approx} P\left(Y \leq k + \frac{1}{2}\right) = \Phi\left(\frac{k+\frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$
- $P(X < k) \overset{\text{c.c.}}{=} P\left(X \leq k - \frac{1}{2}\right) \overset{\text{CLS}}{\approx} P\left(Y \leq k - \frac{1}{2}\right) = \Phi\left(\frac{k-\frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$
- $P(X = k) \overset{\text{c.c.}}{=} P\left(X \leq k + \frac{1}{2}\right) - P\left(X \leq k - \frac{1}{2}\right) \overset{\text{CLS}}{\approx} \Phi\left(\frac{k+\frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-\frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$

Het effect van continuïteitscorrectie is overigens kleiner naarmate n groter is.

Voorbeeld 7.4.5 In voorbeeld 5.6.1 hebben we met de ongelijkheid van Chebyshev aangetoond dat voor $n \geq 25000$ geldt dat

$$P\left(\left|\frac{X}{n} - p\right| < 0.01\right) \geq 0.90$$

Interpretatie van deze kansuitspraak: je moet de steekproefomvang n minstens 25000 nemen om te zorgen dat de kans dat de **steekproeffractie** $\frac{X}{n}$ minder dan 1% afwijkt van de werkelijke fractie p afwijkt minstens 90% is
Bovengenoemde kansuitspraak kunnen we herschrijven tot een kans m.b.t. X:

$$P(\, n(p - 0.01) \leq X \leq n(p + 0.01)\,) \;\geq 0.90$$

Vervolgens gaan we opnieuw n bepalen zodat aan de ondergrens 0.90 is voldaan, maar nu met de normale benadering van de binomiale verdeling van X.
Daarbij kunnen we *geen continuïteitscorrectie* toepassen, omdat $n(p \pm 0.01)$ geen gehele getallen hoeven te zijn:

$$\begin{aligned} P(\, n(p - 0.01) \leq X \leq n(p + 0.01)\,) &\overset{\text{CLS}}{\approx} \Phi\left(\frac{n(p+0.01) - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{n(p-0.01) - np}{\sqrt{np(1-p)}}\right) \\ &= \Phi\left(\frac{0.01n}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{-0.01n}{\sqrt{np(1-p)}}\right) \\ &= 2\Phi\left(\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 \geq 0.90 \end{aligned}$$

Hieruit volgt dat $\Phi\left(\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right) \geq 0.95$, dus $\frac{0.01\sqrt{n}}{\sqrt{p(1-p)}} \geq 1.645$, dus $n \geq \left(\frac{1.645}{0.01}\right)^2 p(1-p)$.
n hangt af van de (onbekende) p, maar we kunnen gebruiken dat $p(1-p) \leq \frac{1}{4}$, voor alle $p \in [0,1]$. Dus $n \geq \left(\frac{1.645}{0.01}\right)^2 \cdot \frac{1}{4} \approx 6765$ voor alle mogelijke waarden van de kans p.
We zien dat n aanzienlijk kleiner is dan bij gebruik van de regel van Chebyshev. ∎

De normale benadering hebben we in bovenstaand voorbeeld uitgevoerd voor X, maar we hadden ook direct de kans $P\left(\left|\frac{X}{n} - p\right| < 0.01\right)$, die betrekking heeft op $\frac{X}{n}$, normaal kunnen benaderen (met hetzelfde resultaat):

omdat het **aantal $\boldsymbol{X}$** bij benadering $N\big(np, np(1-p)\big)$ is, is de **steekproeffractie** $\boldsymbol{\frac{X}{n}} = \frac{1}{n} \cdot X$ bij benadering ook normaal, en wel volgens een $\boldsymbol{N\left(p, \frac{p(1-p)}{n}\right)}$**-verdeling**.
Immers: $\mu = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n} \cdot np = p$ en

$$\sigma^2 = var\left(\frac{X}{n}\right) = \left(\frac{1}{n}\right)^2 var(X) = \left(\frac{1}{n}\right)^2 \cdot np(1-p) = \frac{p(1-p)}{n}$$

In de statistiek zullen we de steekproeffractie $\frac{X}{n}$ noteren met $\boldsymbol{\hat{p}}$.

Tot slot nog een opmerking over de **normale benadering van de Poisson-verdeling** die, zoals bekend, voor grote n ook een benadering is van de binomiale verdeling. Een benadering komt aan de orde voor grote waarden van μ. Ook uit praktische overwegingen: voor $\mu > 10$ hebben we geen tabellen meer. Dat hier ook de CLS van toepassing is in eerste instantie misschien niet evident, maar aangezien een Poisson-verdeelde variabele X betrekking heeft op het aantal gebeurtenissen in een bepaalde periode en/of in een bepaald gebied, is duidelijk dat we periode of gebied in kleinere eenheden kunnen opdelen: als X het Poisson verdeelde aantal klanten is dat een verzekeringsbedrijf in een uur belt, met verwacht aantal $\mu = 90$, dan is het aantal bellers in een minuut ook Poisson verdeeld met parameter $\mu = \frac{90}{60} = 1.5$.
Voor elke minuut kunnen we een X_i definiëren.
Dus $X = \sum_{i=1}^{60} X_i$, waarbij we aannemen dat de X_i's o.o. zijn.
Volgens de CLS is dan X bij benadering normaal verdeeld met verwachting $60 \cdot 1.5 = 90$ en variantie eveneens $60 \cdot 1.5 = 90$.

Dus in zijn algemeenheid kunnen we voor **grote $\boldsymbol{\mu}$ de Poisson verdeling benaderen met de $\boldsymbol{N(\mu, \mu)}$-verdeling**. Bij normale benadering van Poisson-kansen kunnen we weer continuïteitscorrectie toepassen.

7.5 Vraagstukken

1. Twee mannen worden willekeurig gekozen uit de populatie van 20-jarige mannen van wie de gewichten $N(80, 100)$ zijn verdeeld: hun gewichten zijn X en Y.

a. Bereken $P(X > 90 \text{ en } Y > 90)$
b. Bereken $P(X + Y > 180)$
c. Waarom is de kans in onderdeel b. groter dan die van onderdeel a.?

2. In de voorbeelden 6.1.4 en 6.1.6 hebben we de kansdichtheid van het maximum M van 3 random getallen tussen 0 en 1 gebruikt:

$$f_M(m) = \begin{cases} 3m^2 & \text{als } 0 \le m \le 1 \\ 0 & \text{anders} \end{cases}$$

We gaan deze kansdichtheid nu daadwerkelijk afleiden uit de $U(0,1)$-verdeling van 3 random getallen X_1, X_2 en X_3. Dus $M = \max(X_1, X_2, X_3)$.

a. Geef eerst de kansdichtheid $f(x)$ en verdelingsfunctie $F(x)$ van de $U(0,1)$-verdeling.

b. Druk eerst $F_M(m) = P(max(X_1, X_2, X_3) \leq m) = \cdots$ uit in F.
Bepaal vervolgens de $f_M(m)$ als afgeleide van $F_M(m)$.

3. X en Y zijn twee o.o. en exponentieel verdeelde wachttijden met parameters λ_1 en λ_2.

a. Bepaal (met de convolutie-integraal) de kansdichtheid van $X + Y$ als $\lambda_1 = \lambda_2 = 1$.
b. Bepaal de kansdichtheid van $X + Y$ als $\lambda_1 = 1$ en $\lambda_2 = 2$.
c. Bereken $P(X > 1 \text{ en } Y < 1)$ als $\lambda_1 = 1$ en $\lambda_2 = 2$.

4. Als Z_1 en Z_2 o.o. en beide $N(0,1)$-verdeeld zijn en $X = Z_1^2$ en $Y = Z_2^2$, dan heeft $X + Y = Z_1^2 + Z_2^2$ (per definitie) een **Chi-kwadraatverdeling met 2 vrijheidsgraden**.
In deze opgave gaan we de kansdichtheid van die verdeling afleiden.
In voorbeeld 6.4.2 hebben we al gezien dat $X = Z_1^2$ (en dus ook Y) een chi-kwadraat-verdeling heeft met 1 vrijheidsgraad: $f(x) = \frac{1}{\sqrt{2\pi x}} e^{-\frac{1}{2}x}$, voor $x > 0$.
Pas de convolutie-integraal toe om de kansdichtheid van $X + Y$ af te leiden.
In de verkregen integraal moet je gebruiken dat $\int_0^z \frac{1}{\sqrt{x(z-x)}} dx = \pi$ (zie het calculusboek).
Welke verdeling herken je in de Chi-kwadraatverdeling met 2 vrijheidsgraden?

5. Een studie over tweeverdienersgezinnen meet zowel het inkomen X van de man als het inkomen Y van de vrouw voor een groot aantal echtparen, waarvan zowel de man als de vrouw een baan hebben. Stel dat u de verwachtingen μ_X en μ_Y en de varianties σ_X^2 en σ_Y^2 van de beide variabelen in de populatie kent.

a. Is het redelijk om $\mu_X + \mu_Y$ als verwachting van het totale inkomen $X + Y$ te nemen? Licht uw antwoord toe.
b. Is het redelijk om $\sigma_X^2 + \sigma_Y^2$ te nemen als de variantie van het totale inkomen? Licht uw antwoord toe.

6. We vervoeren 100 breekbare flessen. De tijdens het transport uitgeoefende kracht (door schokken, enz.) wordt verondersteld $N(50, 100)$-verdeeld te zijn. De breeksterkte van de flessen is $N(60, 36)$-verdeeld. Een fles breekt als de uitgeoefende kracht groter is dan de breeksterkte. Deze kracht en de breeksterkte worden verondersteld onafhankelijk te zijn. Bereken het verwachte aantal flessen dat tijdens het transport breekt.

7. X en Y zijn o.o.: X is $N(4,1)$- en Y is $N(2,4)$-verdeeld.

a. Bereken $P(3X \leq 2Y - 1)$.
b. Bereken $\rho(X, 3X - 2Y)$.

8. (oude tentamenopgave) Druk de volgende kansen uit in (of geef een benadering met) de standaardnormale verdelingsfunctie Φ en geef daarbij aan welke eigenschap(pen) je gebruikt:

a. de kans dat een willekeurige jongen en een willekeurig meisje samen meer dan 150 kg wegen, als het gewicht van jongens $N(75, 250)$- en dat van meisjes $N(65, 150)$-verdeeld is.
b. $P(\sum_{i=1}^{100} X_i \leq 58)$, als $X_1, X_2, \ldots, X_{100}$ o.o. en alle exponentieel verdeeld zijn met parameter $\lambda = 2$.

9. (oude tentamenopgave) Een machine verwerkt jobs. Uit vroeger onderzoek is gebleken dat de gemiddelde verwerkingsduur 95 tijdseenheden en de standaardafwijking van de verwerkingsduur 20 tijdseenheden bedraagt. Op een dag worden 100 jobs aangeboden en we willen de kans berekenen dat de gemiddelde verwerkingsduur op die dag meer dan 100 tijdseenheden bedraagt.

- **a.** Welke veronderstellingen m.b.t. verwerkingsduren van de 100 jobs zijn nodig om de CLS toe te kunnen passen?
- **b.** Benader de gevraagde kans.

10. Een opiniepeiler houdt in Nederland een enquête onder 250 willekeurig gekozen (verschillende) personen, om de aanhang te meten van een partij A, die bij de laatste verkiezingen 25% scoorde.
Zij X het aantal personen onder deze 250, die op partij A zeggen te stemmen.

- **a.** Wat voor verdeling heeft X en door welke verdeling(en) kan men deze benaderen, als we er vanuit gaan dat de aanhang nog steeds 25% is?
- **b.** Benader, met continuïteitscorrectie, de kans dat de opiniepeiling wijst op een verlies van minstens 3% (de aanhang is dan 22% of minder), terwijl in werkelijkheid de aanhang gelijk is gebleven?
- **c.** De opiniepeiler houdt ook bij hoeveel van de ondervraagde kiezers op partij B stemmen. Benader de kans dat er in de steekproef meer dan 5 B-stemmers waren, indien de aanhang van partij B in werkelijkheid 1% is.

11. Om een indruk te krijgen van het percentage fietsen met ondeugdelijke verlichting in den lande, wordt de verlichting van 100 willekeurige gekozen fietsen gecontroleerd. Laat X het aantal fietsen met ondeugdelijke verlichting in deze steekproef zijn. De kans op een ondeugdelijke verlichting geven we aan met p.

- **a.** Geef de verdeling van X en druk $E(X)$ en $var(X)$ uit in p.
- **b.** Geef een ondergrens voor $P\left(\left|\frac{X}{100} - p\right| \leq 0.05\right)$, m.b.v. een normale benadering van de steekproeffractie $\frac{X}{100}$.
  Gebruik daarbij dat voor alle waarden van p geldt dat $p(1-p) \leq \frac{1}{4}$

12. Het aantal i-Phones dat een dealer verkoopt per dag is Poisson verdeeld met verwachting 6. We nemen aan dat de aantallen per dag variëren volgens die verdeling en o.o. zijn. Eens per week vult hij zijn voorraad aan, in de hoop dat het er voldoende zijn voor die week (6 verkoopdagen).

- **a.** Hoe groot is de kans dat hij voldoende op voorraad heeft als hij een beginvoorraad van 40 hanteert? *(deze kans wordt "service level" genoemd)*
- **b.** Hoe groot moet de beginvoorraad (de *safety stock*) zijn, opdat de kans dat deze voldoende is minstens 99% is?

13. (oude tentamenopgave) De kwaliteitscontrole van de massaproductie van spijkers is georganiseerd door de afmetingen van de spijker op te meten voor een aselecte steekproef van n spijkers. De producent garandeert dat hoogstens 1% van de spijkers buiten de

specificatiemarge vallen. Neem, om de onderstaande vragen te beantwoorden, aan dat precies aan dit percentage van 1% “foute” spijkers wordt voldaan.
X is het aantal foute spijkers in een aselecte steekproef van *n* spijkers.

a. Bereken $P(X \geq 1)$, de kans dat er minstens één foute spijker zit in een aselecte steekproef van $n = 15$ spijkers.
b. Bereken $P(X \leq 3)$ voor een aselecte steekproef van $n = 200$ spijkers.
c. Als de steekproefomvang $n = 4000$ spijkers is, bereken dan $E(X)$, $var(X)$ en $P(X \geq 50)$

Enkele aanwijzingen bij de opgaven van hoofdstuk 7:

1. **b**. Welke verdeling heeft $X + Y$?
2. Vergelijk met de aanpak in voorbeeld 7.1.2.
3. Schrijf eerst de formule van de convolutie integraal op (integraal over de lijn $x + y = z$).
4. Idem.
5. Zou het inkomen van de vrouw, direct of indirect, beïnvloed worden door dat van de man?
6. Herschrijf $X > Y$ tot $X - Y > 0$ en bepaal eerst de verdeling van $X - Y$.
7. **a.** Zelfde soort aanpak als in 6.
b. herhaling rekenregels voor $cov(X, Y)$ en $\rho(X, Y)$ (en $var(X)$).
8. Ga voor beide onderdelen na of je hier een exacte normale verdeling kunt toepassen of een benadering (met CLS).
9. Er is geen normaliteit van de jobs gegeven? De vraag is of je dat wel redelijkerwijs kunt veronderstellen.
10. 22% van 250 = … kiezers
11. $\frac{X}{n}$ is bij benadering normaal. Gebruik de binomiale verdeling van X om $\mu = E\left(\frac{X}{n}\right)$ en $\sigma^2 = var\left(\frac{X}{n}\right)$ te bepalen.
12. $n = 6$ lijkt te klein voor het toepassen van de CLS, maar het verwachte aantal $\mu = 6 \cdot 6 = 36$ is voldoende groot om normaal te benaderen! Zie de laatste alinea van dit hoofdstuk.

Hoofdstuk 8: Wachttijden

8.1 De geheugenloosheid van de exponentiële en de geometrische verdeling

Wachttijden, levensduren en bedieningsduren spelen een steeds belangrijker rol door o.m. technische en bedrijfskundige ontwikkelingen. Wachtrijen treffen we aan bij loketten in het postkantoor, kassa's in de supermarkt, inlichtingsdiensten, telefonische klantenservice van o.m. verzekeringsmaatschappijen, toegangswegen van minirotondes en allerlei elektronische loketten, zoals telefooncentrales, telecommunicatienetwerken en computersystemen.
We zijn dan geïnteresseerd in het stochastisch gedrag van de wachtrijlengte, de gemiddelde wachttijd en/of bedieningsduur van een klant, de maximale capaciteit van het systeem e.d. In dit hoofdstuk beschouwen we de meest eenvoudige wachttijdverdelingen en hun eigenschappen.

Voorbeeld 8.1.1 De gemiddelde bedieningsduur van een klant bij een loket van een postkantoor bedraagt twee minuten. We nemen aan dat we de bedieningsduur X (in minuten) van een willekeurige klant kunnen modelleren als een exponentieel verdeelde stochastische variabele.
Als parameter kiezen we $\lambda = \frac{1}{2}$, want dan is de gemiddelde bedieningsduur $E(X) = \frac{1}{\lambda} = 2$ minuten. $\lambda = \frac{1}{2}$ is het aantal klanten dat gemiddeld per minuut wordt bediend en wordt ook wel de **intensiteit** van de bediening genoemd. (In tien minuten kunnen zo'n $10 \cdot \lambda = 5$ klanten worden geholpen.)
Stel nu dat er 2 loketten in een postkantoor geopend zijn. Bij binnenkomst zie ik dat bij het ene loket, klant 1 wordt bediend en naar het andere loopt juist klant 2 heen. Word ik nu sneller geholpen indien ik voor het loket kies waar een klant reeds in bediening is?
Als de bedieningsduren X_1 en X_2 van klanten 1 en 2 beide dezelfde exponentiële verdeling hebben, is dit niet het geval. De kans dat ik bij klant 2 meer dan t minuten moet wachten, is

$$P(X_2 > t) = 1 - F_{X_2}(t) = e^{-\lambda t}$$

En als klant 1 reeds s minuten in bediening is, dan is de kans de resterende bedieningsduur (d.i. de bedieningsduur vanaf het moment dat ik binnenkom) nog eens *t* minuten of meer bedraagt, een voorwaardelijke kans:

$$P(X_1 > t + s | X_1 > s) = \frac{P(X_1 > t + s)}{P(X_1 > s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X_2 > t)$$

In woorden: de resterende bedieningsduur heeft, ongeacht de tijd die de bediening reeds in beslag heeft genomen, dezelfde kansverdeling als de bedieningsduur van een klant gemeten vanaf het begin van de bediening. Dit wordt wel de **geheugenloosheid** van de exponentiële verdeling genoemd. ■

Bovenstaande en soortgelijke situaties geven aanleiding tot de bekende **wachttijd-paradox**. Als bijvoorbeeld de levensduur X van een lamp exponentieel verondersteld wordt met een "gemiddelde" van $E(X) = 1000$ branduren, dan is de resterende levensduur van zo'n lamp, die reeds 500 uren brandt, ook weer exponentieel verdeeld. De verwachte totale levensduur van de betreffende lamp is dus 500 + 1000 = 1500 branduren, 500 branduren méér dan gemiddeld. Als de lampen na een defect vervangen worden, dan is de verwachte totale levensduur van een lamp, die op een willekeurig moment brandt, dus groter dan 1000 uur, want de verwachte **resterende** levensduur is al 1000 uur. De schijnbare tegenstelling met de gemiddelde levensduur van 1000 uur kunnen we intuïtief verklaren uit het feit, dat het willekeurig gekozen tijdstip met relatief grotere kans zal liggen in een langere levensduur. Uitgedrukt met een voorwaardelijke verwachting geldt:

$$E(X|X > 500) = 500 + E(X) = 500 + 1000$$

Als we de levensduur van een apparaat m.b.v. de exponentiële verdeling modelleren, betekent de geheugenloosheid dat het apparaat niet door slijtage stuk gaat, maar door "toevallige oorzaken van buitenaf".

Voorbeeld 8.1.2 We kunnen de situatie in voorbeeld 8.1.1 als volgt discretiseren: in plaats van de exacte bedieningsduur X meten we in welke minuut Y de bediening wordt voltooid. Voltooien in de eerste minuut, dus de gebeurtenis $\{Y = 1\}$, is hetzelfde als de gebeurtenis $\{0 \leq X < 1\}$. En $\{Y = n\} = \{n - 1 \leq X < n\}$. Dus voor $n = 1, 2, 3, ...$ geldt:

$$\begin{aligned} P(Y = n) &= P(n - 1 \leq X < n) \\ &= P(X \geq n - 1) - P(X \geq n) \\ &= e^{-(n-1)\lambda} - e^{-n\lambda} \\ &= \left(e^{-\lambda}\right)^{n-1}(1 - e^{-\lambda}). \end{aligned}$$

In deze kansfunctie van Y herkennen we met enige moeite de geometrische kansfunctie $P(Y = n) = (1 - p)^{n-1}p$ met succeskans $p = 1 - e^{-\lambda}$, de kans dat een bediening binnen een minuut voltooid wordt.

Deze kansfunctie kunnen we ook uit de geheugenloosheid van X beredeneren: volgens de exponentiele verdeling geldt $P(X > 1) = e^{-\lambda}$. Als de bediening reeds een (willekeurig) aantal minuten n aan de gang is, dan wordt de kans dat de bediening niet in de eerst volgende minuut wordt voltooid gegeven door $e^{-\lambda}$, ongeacht het aantal reeds verstreken minuten (n). Dus onder meer geldt:

$$P(X > n + 1|X > n) = P(X > 1) = e^{-\lambda}$$

En met de complementregel: $P(X \leq n + 1|X > n) = 1 - P(X > 1) = 1 - e^{-\lambda}$

Omdat $\{X > 2\} = \{X > 2$ en $X > 1\}$ geldt met de productregel ($P(A \cap B) = P(A|B)P(B)$):

- $P(X > 2) = P(X > 2 \text{ en } X > 1) = P(X > 2|X > 1) \cdot P(X > 1) = P(X > 1)^2 = \left(e^{-\lambda}\right)^2$
- $P(X > 3) = P(X > 3 \text{ en } X > 2) = P(X > 3|X > 2) \cdot P(X > 2) = P(X > 1)^3 = \left(e^{-\lambda}\right)^3$
- Met inductie vinden we:
  $P(X > n) = P(X > n|X > n - 1) \cdot P(X > n - 1) = P(X > 1)^n = \left(e^{-\lambda}\right)^n$

Dus $P(Y = n) = P(X > n - 1) - P(X > n) = \left(e^{-\lambda}\right)^{n-1} - \left(e^{-\lambda}\right)^n = \left(e^{-\lambda}\right)^{n-1}(1 - e^{-\lambda})$

We vergelijken nu $E(X)$ en $E(Y)$ voor het geval dat $\lambda = \frac{1}{2}$:
X is exponentieel verdeeld met $\lambda = \frac{1}{2}$, dus $E(X) = \frac{1}{\lambda} = 2$ en Y is geometrisch verdeeld met

$p = 1 - e^{-\lambda} = 1 - e^{-\frac{1}{2}}$, dus $E(Y) = \frac{1}{p} = \frac{1}{1-e^{-0.5}} \approx 2.54$. Dat $E(Y) > E(X)$ is logisch doordat Y steeds de naar boven afgeronde gehele waarde van X is. ■

De geheugenloosheidseigenschap geldt dus ook voor de geometrische verdeling: immers als $P(Y = n) = (1-p)^{n-1}p$, voor $n = 1, 2, 3, \ldots$, dan geldt $P(X > n) = (1-p)^n, n = 0, 1, 2, \ldots$

$$\text{Dus: } P(Y > m + n | Y > n) = \frac{P(Y > m + n)}{P(X > n)} = \frac{(1-p)^{m+n}}{(1-p)^n} = (1-p)^m = P(Y > m)$$

Opmerking 8.1.3
Bij de discretisatie in voorbeeld 8.1.2 hebben we gekozen voor het registreren van de **minuut** van voltooiing. Kiezen we voor het registreren van de seconde of milliseconde van voltooiing, dan vinden we weer een geometrische verdeling, met kleinere kans p op voltooiing in die tijdseenheid. Door de eenheid steeds kleiner te kiezen gaat de verdeling in de limiet van een geometrische in een exponentiele over. We werken dat hier niet verder uit. ■

In voorbeelden 8.1.1 en 8.1.2 zagen we dat een exponentiële en een geometrische verdeling geheugenloos waren. Deze eigenschap blijkt karakteristiek voor deze twee verdelingen te zijn.

Definitie 8.1.4 De verdeling van een stochastische variabele X noemen we **geheugenloos** op zijn waardenbereik S_X, indien voor alle $t, s \in S_X$:

$$\boldsymbol{P(X > t + s | X > s) = P(X > t)}.$$

In de voorbeelden hebben we de volgende eigenschappen gedemonstreerd. Daarom geven we de volgende eigenschappen zonder formeel bewijs:

Eigenschap 8.1.5 Voor een continue stochastische variabele X met waardenbereik $S_X = [0, \infty)$ zijn de volgende drie beweringen gelijkwaardig:

- **a.** X is exponentieel verdeeld met parameter λ.
- **b.** $P(X > t) = e^{-\lambda t}$ voor $t \geq 0$.
- **c.** De verdeling van X is geheugenloos op S_X en $E(X) = \frac{1}{\lambda}$.

Eigenschap 8.1.6 Voor een discrete stochastische variabele X met waardenbereik $S_X = \{1, 2, \ldots\}$ zijn de volgende beweringen gelijkwaardig:

- **a.** X is geometrisch verdeeld met parameter p.
- **b.** $P(X > n) = (1-p)^n$ voor $n = 0, 1, 2, \ldots$.
- **c.** De verdeling van X is geheugenloos op S_X en $p = P(X = 1)$.

De geometrische verdeling is dus de enige discrete geheugenloze verdeling (op het gegeven waardenbereik), zoals de exponentiële verdeling de enige continue geheugenloze verdeling is.

8.2 Sommen van onderling onafhankelijke wachttijden

Een lokettist van een postkantoor moet een rij van 10 personen bedienen. We kunnen de totale bedieningsduur dan modelleren als de som S van tien o.o. en identiek verdeelde bedieningsduren $X_1, X_2, \ldots, X_{10}$, dus $S = \sum_{i=1}^{10} X_i$
Eenzelfde modellering kunnen we kiezen indien via een telecommunicatiekanaal een tiental (willekeurige) boodschappen moeten worden verzonden. Indien we voor de kansverdeling van de X_i's de exponentiële met parameter λ kiezen, kunnen we m.b.v. de Convolutie-integraal aantonen dat S een zogenaamde Erlang-verdeling heeft.

Definitie 8.2.1 X heeft een **Erlang-verdeling met parameters $\boldsymbol{n}$ en λ,** indien

$$\boldsymbol{f_X(x)} = \frac{\boldsymbol{\lambda(\lambda x)^{n-1}e^{-\lambda x}}}{\boldsymbol{(n\,-\,1)!}} \text{voor } x \geq 0 \text{ en } f_X(x) = 0 \text{ voor } x < 0$$

Korte notatie: $X \sim \text{Erlang}(n, \lambda)$.

Voor $n = 1$ is $f_X(x) = \lambda e^{-\lambda x} (x \geq 0)$: X is exponentiële verdeeld met parameter λ.
In hoofdstuk 7 hebben we aangetoond dat de som van 2 o.o. en exponentieel verdeelde stochastische variabelen deze Erlang-verdeling met $n = 2$ heeft.

Algemeen geldt:

Eigenschap 8.2.2
Als $X_1, X_2, \ldots.$ o.o. en exponentieel verdeeld zijn met parameter λ, dan geldt:

$$\boldsymbol{S_n} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{X_i} \;\sim\; \textbf{Erlang}(\boldsymbol{n}, \boldsymbol{\lambda}).$$

Bewijs (met inductie naar n):
De bewering is waar voor $n = 1$
Inductieveronderstelling: stel S_n is Erlang-verdeeld met parameters n en λ.
We willen nu aantonen dat ook $S_{n+1} = S_n + X_{n+1}$ een Erlang verdeling heeft.
Wegens (onderlinge) onafhankelijkheid van $S_n = \sum_{i=1}^{n} X_i$ en X_{n+1} (eigenschap 5.4.7) m.b.v. de Convolutie-integraal:
Als $s < 0$, dan is $f_{S_{n+1}}(s) = 0$ *(omdat* $S_n \geq 0$ *en* $X_{n+1} \geq 0)$ en als $s \geq 0$, dan is

$$f_{S_{n+1}}(s) = \int_{-\infty}^{\infty} f_{S_n}(x)\, f_{X_{n+1}}(s - x)dx = \int_0^s \frac{\lambda(\lambda x)^{n-1}e^{-\lambda x}}{(n\,-\,1)!} \cdot \lambda e^{-\lambda(s-x)}dx$$

$$= \int_0^s \frac{\lambda^{n+1}x^{n-1}e^{-\lambda s}}{(n\,-\,1)!}dx = \frac{\lambda^{n+1}x^n e^{-\lambda s}}{n!}\bigg|_{x=0}^{x=s} = \frac{\lambda^{n+1}s^n e^{-\lambda s}}{n!}$$

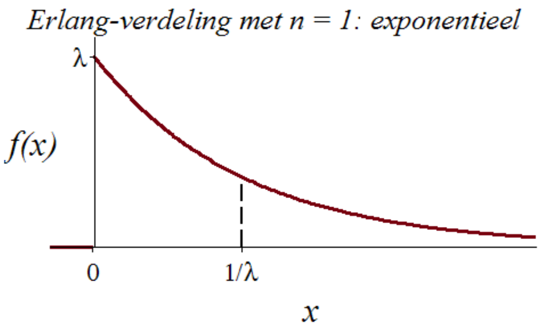
Dit is juist de Erlang-kansdichtheid met parameters $n + 1$ en λ. ■

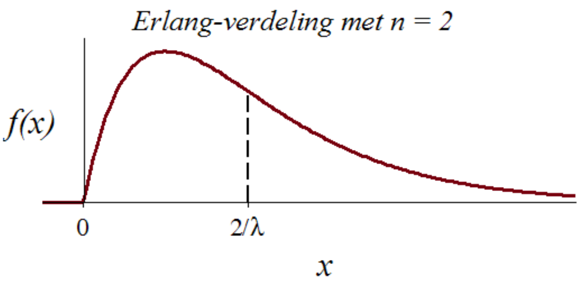
Met behulp van eigenschap 8.2.2 kunnen we snel de verwachtingswaarde van de Erlang-verdeelde stochastische variabele X berekenen door deze te zien als een som van o.o. exponentieel verdeelde stochastische variabelen:

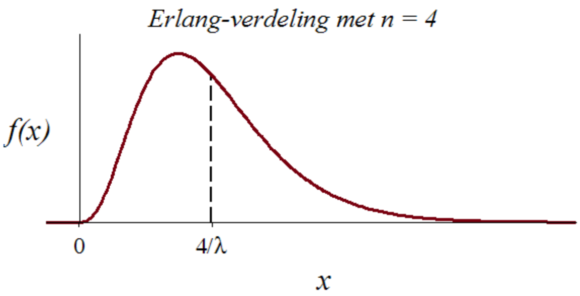
$$\boldsymbol{E(S_n)} = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = n \cdot \frac{1}{\lambda} = \frac{\boldsymbol{n}}{\boldsymbol{\lambda}} \quad \text{en}$$

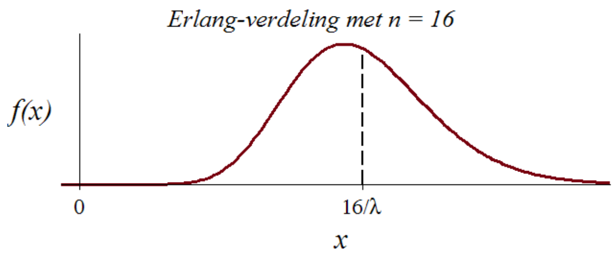
$$\boldsymbol{var(S_n)} = var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i) = n \cdot \frac{1}{\lambda^2} = \boldsymbol{\frac{n}{\lambda^2}}$$

Enkele grafieken van Erlang-verdelingen (voor $n = 1$, 2, 4 en 16):









Als we kansen m.b.t. de totale wachttijd (bedieningsduur) S_n willen berekenen, zoals de kans dat de lokettist in 15 minuten een rij van 10 klanten heeft weggewerkt, $P(S_{10} \leq 15)$, dan kunnen we deze kansen berekenen door eerst de verdelingsfunctie van S_{10} af te leiden, immers $P(S_{10} \leq 15) = F_{S_{10}}(15)$. Voordat we dit toepassen zullen we eerst in zijn algemeenheid de verdelingsfunctie $F_{S_n}(s)$ van de som S_n afleiden:

Als $X_1, X_2, \ldots, X_n$ o.o. en alle $Exp(\lambda)$-verdeeld zijn en $S_n = \sum_{i=1}^{n} X_i$, dan geldt:

$$F_{S_n}(s) = \int_{-\infty}^{\infty} f_{S_n}(x)dx = \int_0^s \frac{\lambda(\lambda x)^{n-1}e^{-\lambda x}}{(n-1)!}dx \quad (we\ moeten\ partieel\ integreren!)$$

$$= \frac{(\lambda x)^{n-1}}{(n-1)!} \cdot -e^{-\lambda x}\Big|_{x=0}^{x=s} + \int_0^s \frac{\lambda(\lambda x)^{n-2}e^{-\lambda x}}{(n-2)!}dx$$

$$= \int_0^s \frac{\lambda(\lambda x)^{n-2}e^{-\lambda x}}{(n-2)!}dx - \frac{(\lambda s)^{n-1}e^{-\lambda s}}{(n-1)!} = \ldots \quad (opnieuw\ partieel\ integreren)$$

Herhaald partieel integreren levert:

$$F_{S_n}(s) = 1 - \sum_{k=0}^{n-1} \frac{(\lambda s)^k e^{-\lambda s}}{k!}$$

In de som in het rechterlid herkennen we een som van Poisson-kansen met parameter $\mu = \lambda s$:
$F_{S_n}(s) = 1 - P(Y \leq n - 1)$ waarin $Y \sim$ Poisson(λs)

Voorbeeld 8.2.3 Bereken de kans $P(S_{10} \leq 15)$, dat de totale bedieningsduur hoogstens 15 minuten bedraagt als de bedieningstijden (in min.) $X_1, X_2, \ldots, X_{10}$ o.o. en exponentieel verdeeld zijn met parameter $\lambda = \frac{1}{2}$.
Oplossing: Hier is $n = 10$ en $\mu = \lambda s = \frac{1}{2} \cdot 15 = 7.5$, dus:

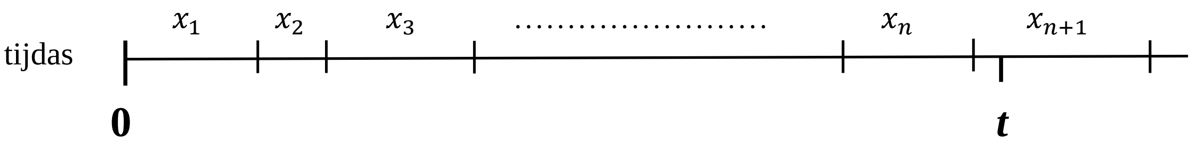
$$P(S_{10} \leq 15) = F_{S_{10}}(15) = 1 - \sum_{k=0}^{9} \frac{7.5^k e^{-7,5}}{k!}$$

We zoeken $P(Y \leq 9)$ op in de Poisson-tabel met parameter μ = 7.5 en vinden:

$$P(S_{10} \leq 15) = 1 - 0.776 = 22.4\%$$

■

De verdelingsfunctie van S_n kunnen we ook gebruiken om aan te tonen dat het aantal klanten dat na t minuten geholpen is Poisson-verdeeld is met parameter λt.
We gaan nu het aantal klanten bekijken dat gedurende het tijdsinterval $[0, t]$ wordt bediend.
Als we aannemen dat zich voor het loket een zeer lange rij klanten bevindt, dan is $\{N(t) = n\}$ de gebeurtenis dat het aantal klanten, dat na t minuten bediend is, precies gelijk is aan n.
Dit is juist de gebeurtenis, dat de totale bedieningstijd voor de eerste *n* klanten **hoogstens** *t* minuten, maar die van de eerste $n + 1$ klanten **meer dan** *t* minuten bedraagt.
Deze situatie, met gerealiseerde bedieningsduren x_i voor klant i en totale bedieningsduur $s_n = \sum_{i=1}^{n} x_i$ voor de eerste n klanten, is in onderstaande figuur op een tijd-as geschetst:



De gebeurtenis $\{N(t) = n\}$ is niet hetzelfde als $\{S_n \leq t\}$: dan kan ook $\{S_{n+1} \leq t\}$ ook nog optreden (en dus $N(t) > n$). $\{N(t) \geq n\}$ en $\{S_n \leq t\}$ zijn wèl hetzelfde.
Omdat $P(N(t) = n) = P(N(t) \geq n) - P(N(t) \geq n + 1)$, vinden we m.b.v. het voorgaande:

$$\begin{aligned} P(N(t) = n) &= P(S_n \leq t) - P(S_{n+1} \leq t) \\ &= 1 - \sum_{k=0}^{n-1} \frac{(\lambda t)^k e^{-\lambda t}}{k!} - \left(1 - \sum_{k=0}^{n} \frac{(\lambda t)^k e^{-\lambda t}}{k!}\right) \\ &= \frac{(\lambda t)^n e^{-\lambda t}}{n!} \end{aligned}$$

Dus $N(t)$ is Poisson-verdeeld met "gemiddelde" λt. We zien dus dat het gemiddeld aantal bediende klanten λt in een tijdsinterval $[0, t]$ toeneemt, als de intensiteit λ, het aantal bediende klanten per minuut, toeneemt.
Een dergelijke reeks bedieningen, waarbij de bedieningsduren exponentieel verdeeld zijn, noemen we wel een **Poisson-proces**. Het Poisson-proces geldt ook voor aankomsten van klanten, elektronische boodschappen e.d. in een bedieningssysteem, als we kunnen aannemen dat de periodes tussen twee opeenvolgende aankomsten o.o. en exponentieel verdeelde stochastische variabelen zijn.

Eigenschap 8.2.4 Indien de tussenaankomsttijden van klanten in een systeem o.o. en exponentieel verdeelde stochastische variabelen zijn, alle met parameter λ, dan is het aantal aankomsten $N(t)$ in het tijdsinterval $(0, t)$ Poisson-verdeeld met parameter $\mu = \lambda t$.

Kansen m.b.t. de som S_n van bedieningsduren, tussenaankomsttijden en levensduren, kunnen we niet alleen via de Erlang-verdeling of de Poisson-verdeling (voorbeeld 8.2.3) bepalen. Voor grote n kunnen we ook de CLS gebruiken:

$$\boldsymbol{S_n \overset{\text{CLS}}{\sim} N\left(\frac{n}{\lambda}, \frac{n}{\lambda^2}\right)}$$

Voorbeeld 8.2.5 Via een communicatiekanaal worden boodschappen verzonden, die dit kanaal gemiddeld 1 milliseconde bezetten. Om capaciteitsproblemen te voorkomen wil men gedurende elke seconde zoveel boodschappen via het kanaal verzenden, dat de kans dat de totale bezettingsduur meer dan 1 seconde bedraagt zeer klein, minder dan 0.1% is.
Om deze vraag te beantwoorden, modelleren we de verzendduren $X_1, X_2, \ldots, X_n$ (in milliseconden) van de n, in een seconde te verzenden boodschappen als o.o. exponentieel verdeelde stochastische variabelen met parameter $\lambda = 1$, zodat $E(X_i) = 1$ milliseconde.
De totale bezettingsduur noteren we met $S_n = \sum_{i=0}^{n} X_i$ en is dus Erlang-verdeeld met parameters n en $\lambda = 1$. Volgens de CLS (n ligt in de buurt van 1000 en is dus voldoende groot) is S_n echter bij benadering $N(n, n)$-verdeeld ($\lambda = 1$) . Dus we willen n bepalen zodat:

$$P(S_n > 1000) < \frac{1}{1000}$$

Ofwel:

$$P\left(\frac{S_n - n}{\sqrt{n}} > \frac{1000 - n}{\sqrt{n}}\right) = \frac{1}{1000}$$

Uit de standaardnormale tabel blijkt dat $\Phi(3.09) = 0.999$, dus volgens de CLS

$$\frac{1000 - n}{\sqrt{n}} \approx 3.09$$

Na enig rekenwerk (kwadrateren en kwadratisch ongelijkheid oplossen, of "proberen" door mogelijke waarden van n in te vullen) volgt dat hieraan wordt voldaan, als het aantal boodschappen per seconde (n) maximaal 891 bedraagt. ■

8.3 Vraagstukken

1. (oude tentamenopgave n.a.v. PTT-gids 1992, pag. 34)
Het tarief voor bepaalde 06-nummers is 15 cent voor elke 30 seconden of een gedeelte ervan: een gesprek van 70 seconden kost dus 3×15 cent (“3 tikken”). De PTT stelt dat dit soort gesprekken “dus circa 30 cent per minuut kosten”.
We gaan deze bewering na door aan te nemen dat de gespreksduur X in seconden exponentieel verdeeld is met parameter $\lambda = \frac{1}{60}$.
We definiëren N als het bijbehorend aantal tikken, dus

$$P(N = n) = P[30(n - 1) < X \leq 30n], \text{voor } n = 1, 2, \ldots .$$

- **a.** Bereken achtereenvolgens: de kans $P(X \geq 30)$, de gemiddelde gespreksduur $E(X)$ en de voorwaardelijke kans $P(X \geq 90|X \geq 60)$.
- **b.** Toon aan dat N geometrisch verdeeld is met parameter $p = 1 - e^{-\frac{1}{2}} \approx 0.39$.
- **c.** Bereken de “gemiddelde” gesprekskosten uit $E(15N)$ en vergelijk deze met het door de PTT opgegeven “gemiddelde”.
- **d.** Bereken de variantie in de gesprekskosten, dus $var(15N)$.

2. In het vakgebied Verkeer wordt onder meer op een aantal telpunten het aantal passerende auto’s geteld om de verkeersdrukte in kaart te brengen. Het is een gebruikelijke aanname dat dit soort aantallen (X) een Poisson-verdeling hebben met een verwachting evenredig met de lengte van de periode van tellen. Als de periode van tellen begint op tijdstip 0 en stopt bij tijd t is het verwachte aantal gelijk aan $\mu = a \cdot t$, met a een nog te schatten constante. In plaats van aantallen per tijdsperiode kun je ook tijdsduren Y tot de eerstvolgende passerende auto registreren. Als we aannemen X, het aantal in een tijdsperiode, Poisson verdeeld zijn, wat voor type verdeling heeft de tijdsduur Y dan? We gaan een formule voor $P(Y > t)$ opstellen voor een willekeurige waarde t. Daarbij gaan we ervan dat het aantal passerende auto’s in het tijdsinterval $[0, t]$ een Poisson verdeling heeft met verwachting $\mu = a \cdot t$.

- **a.** Ga na dat de gebeurtenis $\{Y > t\}$ ook weergegeven kan worden als de gebeurtenis $\{X = 0\}$.
- **b.** Geef een formule voor $P(X = 0)$ in termen van a en t.
- **c.** Bepaal $P(Y > t)$ voor willekeurige waarden t. Bij welke kansverdeling hoort deze kans?

3. Het stochastisch gedrag van ingewikkelde wachtsystemen wordt soms statistisch bepaald m.b.v. computersimulaties. Daarbij worden wachttijden (bedieningsduren e.d.) gegenereerd met zogenaamde randomgeneratoren, die “willekeurige” getallen tussen 0 en 1 produceren.
Stel dat X zo’n willekeurig getal is (uniform verdeeld op $(0, 1)$), dan kunnen we een willekeurige bedieningsduur met de exponentiële verdeling met parameter $\lambda = 1$ genereren door $Y = \ln\left(\frac{1}{X}\right)$ te nemen.

- **a.** Toon aan dat Y exponentieel verdeeld is met parameter $\lambda = 1$.

b. Ga na of $E(Y) = \ln\left(\frac{1}{E(X)}\right)$

4. We kunnen de tijd tussen opeenvolgende klanten die inloggen op het intranet van een bedrijf modelleren als o.o. en exponentieel verdeelde variabelen. Gemiddeld blijkt die tijd (gedurende kantooruren) 12 seconden te zijn. We nemen nu de tussenaankomsttijden van klanten gedurende 1 minuut waren. X_1 is de tijd (in sec) tussen het begin van de meting en de eerste aankomst, X_2 de tijd die vervolgens gemeten wordt tot de tweede aankomst, etc.

a. Bereken $E(X_i)$, $var(X_i)$ en $E(\sum_{i=1}^{6} X_i)$
b. Bereken $P(X_1 > 12)$ en $P(X_1 > 15|X_1 > 3)$.
c. Geef de verdeling van $N =$ "het aantal aankomsten gedurende de minuut (60 seconden)" en bereken $P(N \geq 6)$.
d. Geef de verdeling van $\sum_{i=1}^{6} X_i$ (naam en parameters) en bereken $P(\sum_{i=1}^{6} X_i \leq 60)$

5. (oude tentamenopgave)
$X_1, X_2, \ldots, X_n$ zijn o.o. wachttijden, die exponentieel verdeeld zijn met parameter $\lambda = \frac{1}{4}$.
De som van de wachttijden is $S_n = \sum_{i=1}^{n} X_i$ en de gemiddelde wachttijd $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

a. Geef (zonder bewijs) het type van verdeling van S_n, $E(S_n)$ en $var(S_n)$.
b. Leid de kansdichtheid van S_2 af uit die van X_1 en X_2 (m.b.v. de convolutie-integraal).
c. Bereken (voor $n = 2$): $P\left(\overline{X}_2 > 5\right)$.
d. Benader (voor $n = 100$): $P\left(\overline{X}_{100} > 5\right)$.

Enige aanwijzingen bij de opgaven van hoofdstuk 8:

1. a. Gebruik de geheugenloosheid voor de voorwaardelijke kans
2. Gebruik $P(Y > t) = P(X = 0)$.
3. Vergelijk, indien nodig, met de afleiding in 6.4.5.
4. Wat betekent $N \geq 6$ voor $\sum_{i=1}^{6} X_i$? Gebruik dit bij d. om niet de Erlang kansdichtheid te hoeven integreren.
5. Gebruik bij c. dat $\overline{X}_2 = \frac{S_2}{2}$
Bij d. kun je direct met de benaderende verdeling voor het gemiddelde werken, of eerst de kansuitspraak "vertalen" naar S_{100}.

Antwoorden bij de opgaven

Hoofdstuk 1

1. **a.** $AB\overline{C}$ **b.** ABC **c.** $A \cup B \cup C$ **d**. $AB \cup AC \cup BC$
e. $\overline{A}\,\overline{B}\,\overline{C} = \overline{A \cup B \cup C}$ **f.** $A\overline{B}\,\overline{C} \cup \overline{A}B\overline{C} \cup \overline{A}\,\overline{B}C$ **g.** $\overline{ABC} = \overline{A} \cup \overline{B} \cup \overline{C}$ (De Morgan)

2. $\frac{600}{1200}$

3. De kansen zijn niet gelijk. De kans op “één kruis en één munt” is bijvoorbeeld ½.

4. Niet juist, de kans is $1 - 0.98^{20} \approx 33.2\%$

5. $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$

6. **a.** 1 **b.** 2 **c.** niet mogelijk

7. $A \subset B$ betekent $B = A \cup (\overline{A}B)$, waarbij A en $\overline{A}B$ elkaar uitsluiten, dus (axioma 3):
$P(B) = P\left(A \cup \overline{A}B\right) = P(A) + P\left(\overline{A}B\right)$.
Omdat $P\left(\overline{A}B\right) \geq 0$ (axioma 1), volgt hieruit: $P(B) \geq P(A)$

8. $P(B) = \frac{13}{18}$ en $P\left(\overline{A} \cap \overline{B}\right) = \frac{1}{9}$

Hoofdstuk 2

1. **a.** 7! = 5040 **b.** $10 \cdot 9 \cdot 8 \cdot \ldots \cdot 5 \cdot 4 = \frac{10!}{3!} = 604800$ **c.** $\binom{10}{3} = \frac{10!}{3!7!} = 120$
d. $1 - \dfrac{\binom{36}{13} + 16 \cdot \binom{36}{12}}{\binom{52}{13}} \approx 1 - 0.0352 = 96.48\%$ **e.** $\binom{30}{6} \cdot \binom{24}{7} \cdot \binom{17}{8} \cdot \binom{9}{9} = \frac{30!}{6!7!8!9!}$

2. $\dfrac{169}{1000} = \dfrac{8^3 - 7^3}{10^3}$

3. $\dfrac{3 \cdot 9 \cdot 8 \cdot 7}{10 \cdot 9 \cdot 8 \cdot 7} = \dfrac{3}{10}$

4. a. $\dfrac{\binom{16}{5}}{\binom{52}{5}} = \dfrac{16!/11!}{52!/47!} \approx 0.0017$ **b**. $\dfrac{4 \cdot 4 \cdot (50 \cdot 49 \cdot 48)}{52!/47!} \approx 0.0060$

c. $\dfrac{\binom{4}{1}\binom{4}{1}\binom{44}{3}}{\binom{52}{5}} = \dfrac{5 \cdot 4 \cdot 4 \cdot 4 \cdot (44 \cdot 43 \cdot 42)}{52!/47!} \approx 0.0815$

5. $\dfrac{\binom{6}{4} + 1}{6!} \approx 0.0222$

6. a. $1-\dfrac{\binom{90}{10}}{\binom{100}{10}}\approx 0.6695$ **b.** $\dfrac{\binom{80}{10}}{\binom{100}{10}}\approx 0.0951$

7. a. $\dfrac{\binom{4}{1}\cdot\binom{10}{1}\cdot\binom{86}{3}}{\binom{100}{5}}\approx 0.0544$ **b.** $\dfrac{\binom{14}{1}\cdot\binom{86}{4}}{\binom{100}{5}}\approx 0.3949$ **c.** $\dfrac{\binom{86}{5}}{\binom{100}{5}}\approx 0.4626$ **d.** 0.5374

8. $\dfrac{\binom{N_1}{n_1}\cdot\binom{N_2}{n_2}\cdot\ldots\cdot\binom{N_k}{n_k}}{\binom{N}{n}}$

9. a. $2\cdot\dfrac{\binom{6}{4}\cdot\binom{20}{9}}{\binom{26}{13}}\approx 0.4845$ **b.** $\dfrac{\binom{6}{3}\cdot\binom{20}{10}}{\binom{26}{13}}\approx 0.3553$

10. a. $\binom{13}{6}$ **b.** $\binom{30}{6}\cdot\binom{24}{7}\cdot\binom{17}{8}\cdot\binom{9}{9}=\frac{30!}{6!7!8!9!}$

c. $\binom{n}{n_1}\cdot\binom{n-n_1}{n_2}\cdot\ldots\cdot\binom{n_k}{n_k}=\dfrac{n!}{n_1!\,n_2!\ldots n_k!}$.

11. a. $\dfrac{\binom{7}{2}+\binom{7}{3}}{\binom{11}{5}}\approx 0.1212$ **b.** $\dfrac{\binom{4}{2}\binom{5}{1}+\binom{5}{2}\binom{4}{1}}{\binom{11}{5}}=\dfrac{70}{462}\approx 0.1515$

12. $\dfrac{\binom{10}{4}}{10!/6!}=\dfrac{1}{4!}=\dfrac{1}{24}$

Hoofdstuk 3

1. a. $\frac{0.30}{0.75}=\frac{2}{5}$ (= 40%) **b.** $\frac{3}{10}$

2.

a. $P(B)=0.97*0.98+0.05*0.02=0.9516$.

b. $P(\overline{A}\mid\overline{B})=39.3\%$.

c. A en B zijn afhankelijk.

3. $\frac{0.75\cdot 0.05}{0.75\cdot 0.05+0.02\cdot 0.95}\approx 0.6673$

4. $P(\text{Comium})=0.1$

5. $\frac{2}{3}$

6. a. $\dfrac{\binom{5}{3}\cdot\binom{7}{2}}{\binom{12}{5}}\approx 0.2652$ **b.** $\displaystyle\sum_{i=3}^{6}\frac{\binom{5}{3}\cdot\binom{7}{i-3}}{\binom{12}{i}}\cdot\frac{1}{6}\approx 0.1385$

7. $\frac{3}{4}$

8. **a.** $P(A\overline{B}) = P(A) - P(AB) = P(A) - P(A)P(B) = P(A)[1 - P(B)] = P(A) \cdot P(\overline{B})$
A en $\overline{B}$ zijn dus ook o.o., waaruit dan ook weer volgt dat $\overline{A}$ en $\overline{B}$ o.o. zijn.
b. $P(A \cap BC) = P(ABC) = P(A)P(B)P(C) = P(A) \cdot P(BC)$.

9. alléén mogelijk als $P(A) = 0$ of als $P(B) = 0$.

10. $\left(\frac{5}{6}\right)^6 \approx 33.49\%$

11. $\sum_{k=10}^{12} \binom{12}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{12-k} \approx 0.054\%$

12. **a.** $P(B) = 0.997 \cdot 0.01 + 0.015 \cdot 0.99 = 0.02482$ **b.** 40.2%

Hoofdstuk 4

1. **a**. $P(X > 0) = 0.5$ **b.** $E(X) = -0.2$ **c.** $E(X^2) = 4.6$ **d**. $var(X) = 4.56$ en $\sigma_X \approx 2.14$

2. $P(X = k) = \frac{1}{10}, k = 0, 1, 2, \dots, 9$, $E(X) = 4.5$, $E(X^2) = 28.5$ en $var(X) = 8.25$

3. 487

4. $-\frac{17}{216} \approx -0.0787$

5. **a.** $P(X = 7) = \frac{\binom{6}{3}}{\binom{10}{4}}$ **b**. $P(X = k) = \frac{\binom{k-1}{n-1}}{\binom{N}{n}}$, $k = n, n+1, \dots, N$

6. **a**. $P(X \leq 7) = 0.998$ **b.** $P(X \geq 7) = 0.011$
c. $P(X = 9) = 20.7\%$ **d**. $P(X < 12) = 0.909$

7. **a**. $P(X = 5) = 0.101$ **b**. $P(X < 2) = 0.199$. **c**. $P(X > 3) = 0.353$

8.
a. Poisson($\mu = 5$), $P(X = 2) \approx 0.0842$. $E(X) = \mu = 5$.
b. hypergeometrische verdeling: $P(X = 2) = \frac{\binom{2}{2}\binom{3}{0}}{\binom{5}{2}} = \frac{1}{10}$ en $E(X) = \frac{4}{5}$
c. $X \sim B(100, 0.02)$, $P(X = 2) \approx 27.3\%$ en $E(X) = np = 2$.
d. $X \sim geometrisch\left(p = \frac{1}{10}\right)$, $P(X = 2) = 0.09$ en $E(X) = 10$
e. X is homogeen verdeeld op {1,2,..,10}, dus $P(X = 2) = \frac{1}{10}$ en $E(X) = 5.5$

9. **a**. 0.277 **b**. 0.287

10. **a.** $E(X^k) = \frac{1}{2}c^k$ **b**. $E(X) = \frac{1}{2}c$ en $E(X^2) = \frac{1}{2}c^2$, $var(X) = \frac{1}{4}c^2$

11. a. $c = \frac{2}{3}$ **b.** $Y = X + 1$ is geometrisch ($p = \frac{2}{3}$) **c**. $E(X) = \frac{1}{2}$ en $var(X) = \frac{3}{4}$

12. a. $B\left(150, \frac{1}{50}\right)$-verdeeld **b.** $c = 6$ buitenlijnen.

13. a. $B(12, 0.15)$-verdeling; $P(X > 1.8) \approx 0.5565$
b. Poisson-verdeeld met parameter $\mu = 2$: $P(X > 2) \approx 32.33\%$
c. Hypergeometrische verdeling met parameter $N = 10, R = 3$ en $n = 4$: $P(X > 1.2) = \frac{1}{3}$

14. a. $1 - (0.96)^{10} \approx 0.3352$.
b. Benaderd met Poisson verdeling ($\mu = 4$): 0.567 (exact: 0.5705) **c**. 2.984

15. a. $\binom{9}{2}\left(\frac{5}{6}\right)^7\left(\frac{1}{6}\right)^3 \approx 4.65\%$
b. $S_X = \{m, m+1, \dots\}$ en $P(X = k) = \binom{k-1}{m-1}(1-p)^{k-m}p^m$, met $k \in S_X$.

16. a. $M = 2$ **b**. $M = 2$ **c.** $M = 2$ **d**. $M \in [3, 4]$

Hoofdstuk 5

1. a. **b.** $P(X = 1|Y = 1) = \frac{1}{4}$, $P(X = 2|Y = 1) = \frac{1}{2}$ en $P(X = 3|Y = 1) = \frac{1}{4}$.
Dus $E(X|Y = 1) = 2$.
c. $P(Y = 1|X = 3) = \frac{1}{2}$

2. b. $P(X = i) = \left(\frac{1}{3}\right)^{i-1}\frac{2}{3}$, voor $i = 1, 2, 3, \dots$, dus $X \sim geometrisch\left(\frac{2}{3}\right)$ en $E(X) = \frac{1}{p} = \frac{3}{2}$
c. $P(Y = 0) = \frac{1}{3}$ en $P(Y = 1) = \frac{2}{3}$ ($P(Y = j) = \left(\frac{1}{3}\right)^{1-j}\left(\frac{2}{3}\right)^j$, voor $j = 0{,}1$)
d. ja

3.
a. Kansverdeling van X: kansen per rij optellen.
Kansverdeling van Y: kansen per kolom optellen.

i	0	1	2	Som
$P(X = i)$	0.2	0.5	0.3	1

j	0	1	2	3	som
$P(Y = j)$	0.10	0.30	0.45	0.15	1

b. $E(X) = 1.10$ en $var(X) = 0.490$, $E(Y) = 1.65$ en $var(Y) = 0.7281$
c. Voor $Z = 8Y$ geldt $E(Z) = 8 \times E(Y) = 8 \times 1.65 = 13.20$ en
$\text{var}(8Y) = 8^2\,\text{var}(Y) = 64 \times \text{var}(Y) = 46.56$ (afgerond).
d. De waarden van $T = X + Y$ lopen van 0 t/m 5. (kansen “diagonaal” optellen)

t	0	1	2	3	4	5	Som
$P(T = t)$	0.05	0.10	0.20	0.40	0.20	0.05	1

$E(T) = 2.75$ en $var(T) = 1.3875$
e. $E(T) = E(X + Y) = 2.75$ is in overeenstemming met de som van de verwachtingen $E(X) = 1.10$ en $E(Y) = 1.65$.

f. $\text{var}(T) = \text{var}(X+Y) = 1.3875$ correspondeert niet met $var(X) + var(Y) = 0.490 + 0.7281 = 1.218$, dit heeft te maken met de afhankelijkheid van X en Y.

4. a. $P(N = 10) = \left(\frac{1}{2}\right)^{10}$, $P(X = 4|N = 10) = \binom{10}{4}\left(\frac{1}{2}\right)^{10}$ en

$P(X = x \text{ en } N = n) = \binom{10}{4}\left(\frac{1}{4}\right)^{10}$

b. $P(N = n) = \left(\frac{1}{2}\right)^{n}$, $n = 1,2, \ldots$

$P(X = x|N = n) = \binom{n}{x}\left(\frac{1}{2}\right)^{n}$, $x = 0, 1, 2, \ldots, n$

$P(X = x \text{ en } N = n) = \binom{n}{x}\left(\frac{1}{4}\right)^{n}$, $x = 0, 1, 2, \ldots, n$ en $n = 1, 2, \ldots$

c. 5, $\frac{1}{2}n$, $E(X) = E[E(X|N)] = E\left(\frac{1}{2}N\right) = \frac{1}{2}E(N) = \frac{1}{2} \cdot 2 = 1$

d. $P(X = 0) = \frac{1}{3}$

e. N is gegeven $X = 0$ geometrisch verdeeld met parameter $p = \frac{3}{4}$ dus $E(N|X = 0) = \frac{4}{3}$

5. a. als $N = n$, dan is $S = X_1 + \cdots + X_n$

b. $E(X_i) = \sum_x x \cdot P(X_i = x) = 1000 \cdot \frac{1}{10} + 2000 \cdot \frac{3}{10} + 3000 \cdot \frac{4}{10} + 4000 \cdot \frac{2}{10} = 2700$

c. $E(S|N = n) = 2700n$

d. $E(S) = E[\,E(S|N)] = E(2700N) = 2700 \cdot E(N) = 2700\mu$

6. a. $P(X = 8 \text{ en } Y = 2) = 0.033$.

b. $E(Y|X = 8) = 2.4$ en $E(Y|X = x) = 0.3x$.

c. $E(Y) = E[\,E(Y|X)] = E[\,0.3X] = 0.3 \cdot E(X) = 3$

7. a. $P(X_1 = 10) = 0.9^9 \cdot 0.1 = 3.87\%$

b. Gebruik de eigenschap $P(X > x) = (1-p)^x$ van de geometrische verdeling:

$P(20 \leq X_1 \leq 30) = P(X_1 > 19) - P(X_1 > 30) = 0.9^{19} - 0.9^{30} \approx 9.27\%$

c. $P(X_1 = X_2) \approx 5.26\%$

d. $P(X_1 + X_2 = 20) \approx 2.85\%$

8. a. $P(X > i \text{ en } Y > i) \overset{\text{o.o.}}{=} P(X > i)P(Y > i) = (1-p)^i \cdot (1-p)^i = [(1-p)^2]^i$

b. $P(min(X,Y) > i) = P(X > i \text{ en } Y > i) = [(1-p)^2]^i$, voor $i = 0, 1, 2, \ldots$

c. $P(min(X,Y) = i) = P(min(X,Y) > i-1) - P(min(X,Y) > i)$

$= [(1-p)^2]^{i-1} - [(1-p)^2]^i = [(1-p)^2]^{i-1}[1-(1-p)^2]$

voor $i = 1, 2, \ldots$

d. Dus $min(X,Y)$ is geometrisch verdeeld met parameter $1-(1-p)^2 = 2p - p^2$

$E[min(X,Y)] = \frac{1}{2p-p^2}$

9. **a**. $E(X+Y) = E(X) + E(Y) = \frac{2}{p}$ en $var(X+Y) \overset{\text{o.o.}}{=} var(X) + var(Y) = 2 \cdot \frac{1-p}{p^2}$

b. $P(X = i \text{ en } Y = j) \overset{\text{o.o.}}{=} P(X = i) \cdot P(Y = j) = (1-p)^{i-1}p \cdot (1-p)^{j-1}p$
$= (1-p)^{i+j-2}p^2 \qquad (i, j = 1, 2, \dots)$

c. $P(X+Y = n) = \sum_{i=1}^{n-1}(1-p)^{i-1}p \cdot (1-p)^{n-i-1}p = \sum_{i=1}^{n-1}(1-p)^{n-2}p^2$
$= (n-1) \cdot (1-p)^{n-2}p^2, n = 1, 2, 3, \dots$

10. **a**. X en Y hebben dezelfde nevenstaande verdeling $E(X) = 1$ en
$var(X) = E(X^2) - \mu^2 = 1.6 - 1 = 0.6$

x	0	1	2	som
$P(X = x)$	0.3	0.4	0.3	1
$x \cdot P(X = x)$	0	0.4	0.6	$1 = E(X)$
$x^2 \cdot P(X = x)$	0	0.4	1.2	$1.6 = E(X^2)$

b. Resp. $+\frac{2}{3}$, 0, 0 en -1

c. Alléén bij verdeling 3 zijn X en Y o.o.:
$P(X = i \text{ en } Y = j) = P(X = i) \cdot P(Y = j)$, voor alle (i, j)

d. $E(XY) = \sum_i \sum_j \; i \cdot j \cdot P(X = i \text{ en } Y = j) = 1 \cdot 0.2 + 2 \cdot 0.1 + 2 \cdot 0.1 + 4 \cdot 0.2 = 1.4$
$cov(X,Y) = E(XY) - E(X)E(Y) = 1.4 - 1 \cdot 1 = 0.4$
$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{0.4}{\sqrt{0.6}\sqrt{0.6}} = +\frac{2}{3}$

e. $cov(3X, 2-Y) = 3 \cdot -1 \cdot cov(X,Y) = -3 \cdot 0.4 = -1.2$,
dus $\rho(3X, 2-Y) = -\rho(X,Y) = -\frac{2}{3}$

f. $E(XY) = 1$, dus $cov(X,Y) = 0$ en $\rho(X,Y) = 0$

g. $E(X|Y=0) = \frac{2}{3}$, $E(X|Y=0) = 1$ en $E(X|Y=2) = \frac{4}{3}$
$E(X) = 1 = \frac{2}{3} \times 0.3 + 1 \times 0.4 + \frac{4}{3} \times 0.3 = \sum_y E(X|Y=y) \times P(Y=y)$

11. $\rho(X,Y) = \rho(X, -3X+4) = \frac{cov(X,-3X+4)}{\sigma_X \sigma_{-3X+4}}$
Omdat $var(-3X+4) = (-3)^2 \cdot var(X) = 9 \cdot var(X)$, dus $\sigma_{-3X+4} = 3\sigma_X$, en
$cov(X, -3X+4) = -3 \cdot cov(X,X) = -3 \cdot var(X) = -3\sigma_X^2 = -1.2$
vinden we $\rho(X,Y) = \frac{-3\sigma_X^2}{\sigma_X \cdot 3\sigma_X} = -1$

12. **a**. $cov(X_1, X_1+X_2) = cov(X_1,X_1) + cov(X_1,X_2) \overset{\text{o.o.}}{=} var(X_1) + 0 = 2$
$\rho(X_1, X_1+X_2) = \frac{cov(X_1,X_1+X_2)}{\sigma_{X_1}\sigma_{X_1+X_2}} = \frac{var(X_1)}{\sqrt{2}var(X_1)} = \frac{1}{\sqrt{2}}$

b. $\rho(X_1, X_1+X_2+\ldots+X_n) = \frac{1}{\sqrt{n}} < \frac{1}{3}$, als $n > 9$

13. **a**. $E(X_1) = 1 \cdot \frac{1}{10} + 0 \cdot \frac{9}{10} = \frac{1}{10}$, $E\left({X_1}^2\right) = \frac{1}{10}$ en $var(X_1) = \frac{1}{10} - \left(\frac{1}{10}\right)^2 = \frac{9}{100}$
$cov(X_1,X_2) = E(X_1X_2) - E(X_1)E(X_2) = 1 \cdot 1 \cdot \frac{1}{10} \cdot \frac{1}{9} - \left(\frac{1}{10}\right)^2 = \frac{1}{900}$

b. $E(S) = E\left(\sum_{i=1}^{10} X_i\right) = \sum_{i=1}^{10} E(X_i) = 10 \cdot E(X_1) = 1$
$var(S) = var\left(\sum_{i=1}^{10} X_i\right) = \sum_{i=1}^{10} var(X_i) + \sum \; \sum_{i \neq j} \; cov(X_i, X_j)$
$= 10 \cdot var(X_1) + 90 \cdot cov(X_1,X_2) = 10 \cdot \frac{9}{100} + 90 \cdot \frac{1}{900} = 1$

14. **a**. $X + Y \sim Poisson(\mu_1 + \mu_2)$, met $\mu_1 + \mu_2 = 2 + 3$

b. $P(X = k|X+Y=n) = \frac{P(X=x \text{ en } Y=n-k)}{P(X+Y=n)} = \frac{\frac{2^k e^{-2}}{k!} \cdot \frac{3^{n-k} e^{-3}}{(n-k)!}}{\frac{5^n e^{-5}}{n!}} = \frac{n!}{k!(n-k)!}\left(\frac{2}{5}\right)^k \left(\frac{3}{5}\right)^{n-k}$
(hierin is $\mu_1 = 2$en $\mu_2 = 3$). Dit is voor $k = 0,1,2,..,n$ de binomiale verdeling

c. $E(X|X+Y=7) = 7 \cdot \frac{2}{2+3} = 2.8$, omdat X, gegeven $X+Y=n$, $B\left(n, \frac{2}{5}\right)$- verdeeld is. Veronderstellingen: X en Y resp. het aantal blindedarm- en het aantal niersteenoperaties zijn o.o. en beide Poisson-verdeeld met parameters 2 resp. 3.

Hoofdstuk 6

1. a. $P(X>1) = \int_1^\infty f(x)dx = \int_1^2 \left(1-\frac{1}{2}x\right)dx = x - \frac{1}{4}x^2\Big|_{x=1}^{x=2} = 1 - \frac{3}{4} = \frac{1}{4}$,

(of door de oppervlakte van het driehoekje te bepalen: $\frac{1}{2} \cdot 1 \cdot \frac{1}{2} = \frac{1}{4}$)

b. $E(X) = \int_0^2 x \cdot \left(1-\frac{1}{2}x\right)dx = \frac{1}{2}x^2 - \frac{1}{6}x^3\Big|_{x=0}^{x=2} = \frac{2}{3}$

$E(X^2) = \int_0^2 x^2 \cdot \left(1-\frac{1}{2}x\right)dx = \frac{1}{3}x^3 - \frac{1}{8}x^4\Big|_{x=0}^{x=2} = \frac{2}{3}$

$var(X) = E(X^2) - (EX)^2 = \frac{2}{3} - \left(\frac{2}{3}\right)^2 = \frac{2}{9}$

c. $F(x) = P(X \le x)$, dus $F(x) = 0$ als $x < 0$ en $F(x) = 1$ als $x > 2$.

als $0 \le x \le 2$ is $F(x) = \int_0^x \left(1-\frac{1}{2}u\right)du = u - \frac{1}{4}u^2\Big|_{u=0}^{u=x} = x - \frac{1}{4}x^2$

Dus $P(X>1) = 1 - F(1) = 1 - \frac{3}{4} = \frac{1}{4}$

2. a. $f(x) = \lambda e^{-\lambda x}$ voor $x \ge 0$ en $f(x) = 0$ als $x < 0$

$\int_{-\infty}^\infty f(x)\,dx = \int_0^\infty \lambda e^{-\lambda x}\,dx = -e^{-\lambda x}|_0^\infty = -0-(-1) = 1$ (grafiek: zie pag. VI. 22).

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

$P(X > EX) = \int_{1/\lambda}^\infty \lambda e^{-\lambda x}\,dx = -e^{-\lambda x}|_{\frac{1}{\lambda}}^\infty = e^{-\lambda \cdot \frac{1}{\lambda}} = e^{-1} \approx 36.8\%$ ($< \frac{1}{2}$)

c. $P(X > M) = \int_M^\infty \lambda e^{-\lambda x}\,dx = -e^{-\lambda x}|_M^\infty = e^{-\lambda M} = \frac{1}{2}$, dus $M = \frac{\ln(2)}{\lambda}$

d. De modus = 0 (zie grafiek)

3. a. - $f(x) = \frac{1}{4}$, voor $0 < x < 4$: $\int_{-\infty}^\infty f(x)\,dx = 4 \cdot \frac{1}{4} = 1$

- $E(X) = 2 =$ mediaan (wegens symmetrie van f. Dus $P(X > EX) = P(X \le M) = \frac{1}{2}$
- modus: alle waarden uit $[0, 4]$

b. $F_X(x) = \begin{cases} 0 & \text{als} \quad x \le 0 \\ \frac{x}{4} & \text{als } 0 < x < 4 \\ 1 & \text{als} \quad x \ge 4 \end{cases}$

4. a. $\int_{-\infty}^\infty f(x)\,dx = \int_1^\infty \frac{c}{x^3}dx = -c \cdot \frac{1}{2}x^{-2}|_1^\infty = c \cdot \frac{1}{2} = 1$, dus $c = 2$

$P(X > 2) = \int_2^\infty \frac{2}{x^3}dx = -2 \cdot \frac{1}{2}x^{-2}|_2^\infty = \frac{1}{4}$

b. $E(X) = \int_1^\infty x \cdot \frac{2}{x^3}dx = -2x^{-1}|_1^\infty = 2$

$P(X > m) = \int_m^\infty \frac{2}{x^3}dx = -2 \cdot \frac{1}{2}x^{-2}|_m^\infty = \frac{1}{m^2} = \frac{1}{2}$, dus $m = \sqrt{2}$

c. $F(x) = \int_1^x \frac{2}{u^3}du = -u^{-2}|_1^x = 1 - x^2$, voor $x \ge 1$, en $F(x) = 0$ als $x < 1$

5. **a**. 1. $F_Y(y) = P(5-2X \le y) = P(-2X \le y-5) = P\left(X \ge \frac{y-5}{-2}\right) = 1 - F_X\left(\frac{y-5}{-2}\right)$

2. $f_Y(y) = \frac{d}{dy}F_y(y) = \frac{1}{2}f_X\left(\frac{y-5}{-2}\right)$

3. $f_X\left(\frac{y-5}{-2}\right) = 1$ als $0 < \frac{y-5}{-2} < 1$, dus $-2 < y-5 < 0$ ofwel: $3 < y < 5$

$f_Y(y) = \begin{cases} \frac{1}{2}\cdot 1, \text{voor } 3 < y < 5 \\ 0 \quad \text{elders} \end{cases}$, dus $Y \sim U(3,5)$

b. Neem $Y = a + (b-a)\cdot X$

6. **a**. 1. $F_Y(y) = P\left(\frac{1}{X} \le y\right) = P\left(X \ge \frac{1}{y}\right) = 1 - F_X\left(\frac{1}{y}\right)$, $y > 0$

(En $F_Y(y) = P\left(\frac{1}{X} < y\right) = 0$ als $y < 0$)

2. $f_Y(y) = \frac{d}{dy}F_y(y) = -\frac{1}{y^2}\cdot -f_X\left(\frac{1}{y}\right)$

3. $f_X\left(\frac{1}{y}\right) = 1$ als $\frac{1}{y} > 0$, dus als $y > 1 \rightarrow f_Y(y) = \frac{1}{y^2}\cdot 1 = \frac{1}{y^2}$ als $y > 1$

b. $P(Y > 2) = \int_2^\infty f_Y(y)\,dy = \int_2^\infty \frac{1}{y^2}dy = -y^{-1}|_2^\infty = \frac{1}{2}$

$P(Y > 2) = P\left(\frac{1}{X} > 2\right) = P\left(X < \frac{1}{2}\right) = \frac{1}{2}$

c. $E(Y) = \int_{-\infty}^\infty y f_Y(y)dy = \int_1^\infty y\cdot\frac{1}{y^2}dy = \ln(y)\,|_1^\infty = \infty$, dus $E(Y)$ bestaat niet.

Ook $E(Y) = E\left(\frac{1}{X}\right) = \int_{-\infty}^\infty \frac{1}{x}f_X(x)dx = \int_0^1 \frac{1}{x}dx = \ln(x)\,|_0^1$ bestaat ook niet.

7. **a**. als $y > 0$ geldt: $F_Y(y) = P\left(\sqrt{|X|} \le y\right) = P(|X| \le y^2) = P(-y^2 \le X \le y^2)$

$= F_X(y^2) - F_X(-y^2)$

Daar $F_X(x) = 1 - e^{-x}$, al $x > 0$ en $F_X(x) = 0$ als $x < 0$, is

$F_Y(y) = \left(1 - e^{-y^2}\right) - 0 = 1 - e^{-y^2}$, voor $y > 0$.

b. $f_Y(y) = \frac{d}{dy}F_y(y) = 2ye^{-y^2}$ voor $y > 0$

$E(Y) = \int_{-\infty}^\infty y f_Y(y)dy = \int_0^\infty y\cdot 2ye^{-y^2}dy = \cdots partieel \ldots = \int_0^\infty e^{-y^2}dy$,

Hier herkennen we de standaardnormale verdeling in, dus $\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx = 1$,

M.b.v. substitutie $x = \sqrt{2}\cdot y$, vinden we $\int_{-\infty}^\infty e^{-y^2}dy = \sqrt{\pi}$, dus $E(Y) = \frac{1}{2}\sqrt{\pi}$

8. **a.** 0.3085; 0.3753; 0.6826.

bijv. $P(|X-1| < 2) = P(-2 < X-1 < +2) = P\left(-\frac{2}{2} < \frac{X-1}{2} < \frac{2}{2}\right)$

$= \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 = 2\cdot 0.8417 - 1 = 68.34\%$.

b. $P(X \le c) = P\left(\frac{X-1}{2} \le \frac{c-1}{2}\right) = \Phi\left(\frac{c-1}{2}\right) = 90\%$, dus $\frac{c-1}{2} = 1.28$.

$c = 1 + 2\cdot 1.28 = 3.56$ is het 90$^{\text{ste}}$ percentiel van X

c. $c = 1 - 2\cdot 1.28 = -1.56$

9. Bijvoorbeeld: $P(-2\sigma < X - \mu < 2\cdot\sigma) = P\left(-2 < \frac{X-\mu}{\sigma} < 2\right) = \Phi(2) - \Phi(-2)$

$= 2\Phi(1) - 1 = 2\cdot 0.9772 - 1 = 0.9544 \approx 95.4\%$

10. De klassegrenzen voor de eieren worden 50 ± 1.27 gram en 50 ± 4.21 gram.

11. a. Daar $E(X) = \mu$, geldt:
$E(X-\mu)^3 = E(X^3 - 3\cdot X^2\cdot\mu + 3\cdot X\cdot\mu^2 - \mu^3)$
$= E(X^3) - 3\mu E(X^2) + 3\mu^2 E(X) - \mu^3 = E(X^3) - 3\mu E(X^2) + 2\mu^3$

b. $E(X) = \frac{1}{2}$, $E(X^2) = \int_0^1 x^2\cdot dx = \frac{1}{3}x^3|_0^1 = \frac{1}{3}$ en $E(X^3) = \int_0^1 x^3\cdot dx = \frac{1}{4}x^4|_0^1 = \frac{1}{4}$.
En met a. $E(X-\mu)^3 = \frac{1}{4} - 3\cdot\frac{1}{2}\cdot\frac{1}{3} + 2\cdot\left(\frac{1}{2}\right)^3 = 0$
(of direct: $E(X-\mu)^3 = \int_0^1\left(x-\frac{1}{2}\right)^3 dx = \frac{1}{4}\left(x-\frac{1}{2}\right)^4|_0^1 = \frac{1}{4}\left(\frac{1}{16}-\frac{1}{16}\right) = 0$).

c. $E(X) = \frac{1}{\lambda} = 1$ en omdat $var(X) = E(X^2) - (EX)^2 = 1$, is $E(X^2) = 1+1 = 2$.
$E(X^3) = \int_0^\infty x^3\cdot e^{-x}dx = x^3\cdot -e^{-x}|_0^1 + 3\int_0^\infty x^2\cdot e^{-x}dx = 3E(X^2) = 6.$
$E(X-\mu)^3 = 6 - 3\cdot 1\cdot 2 + 2\cdot 1 = 2$

d. Klopt: de uniforme verdeling is symmetrisch: $E(X-\mu)^3 = 0$.
en de exponentiele verdeling is scheef naar rechts: $E(X-\mu)^3 = 2$.

Hoofdstuk 7

1. **a**. $P(X > 90$ en $Y > 90) = P(X>90)\cdot P(Y>90) = \left(1-\Phi\left(\frac{90-80}{10}\right)\right)^2 \approx 2.52\%$

b. $X+Y$ is N(80+80, 100+100)-verdeeld, dus
$P(X+Y > 180) = P\left(Z > \frac{180-160}{\sqrt{200}}\right) \approx 1 - \Phi(1.41) = 7.93\%$

c. De gebeurtenis "$X > 90$ en $Y > 90$" is een deel van de gebeurtenis "$X + Y > 180$"

2. **a.** $f(x) = 1$ en $F(x) = x$, beiden als $0 \le x \le 1$

b. $F_M(m) = P(max(X_1, X_2, X_3) \le m) \overset{\text{o.o.}}{=} P(X_1 \le m)P(X_2 \le m)P(X_3 \le m) = [F(m)]^3$
$f_M(m) = 3F(m)^2\cdot f(m) = 3m^2\cdot 1$, voor $0 \le m \le 1$

3. **a**. $f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx = \int_0^z e^{-x}e^{-(z-x)}dx = \int_0^z e^{-z}dx = e^{-z}\cdot x|_{x=0}^{x=z}$
$= ze^{-z}$ voor $z \ge 0$
(En $f_{X+Y}(z) = 0$ als $z < 0$)

b. $f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx = \int_0^z e^{-x}\cdot 2e^{-2(z-x)}dx = \int_0^z 2e^{-2z}e^x dx$
$= 2e^{-2z}\cdot e^x|_{x=0}^{x=z} = 2e^{-z} - 2e^{-2z}$ voor $z \ge 0$
(En $f_{X+Y}(z) = 0$ als $z < 0$)

c. $P(X > 1$ en $Y < 1) = P(X>1)\cdot P(Y<1) = e^{-1}\cdot(1 - e^{-2\cdot 1}) \approx 31.8\%$

4. $\mathrm{f}_{X+Y}(\mathrm{z}) = \int_{-\infty}^{\infty}\mathrm{f}_X(\mathrm{x})\mathrm{f}_Y(\mathrm{z}-\mathrm{x})\mathrm{dx} = \int_0^z \frac{1}{\sqrt{2\pi x}}e^{-\frac{1}{2}x}\cdot\frac{1}{\sqrt{2\pi x}}e^{-\frac{1}{2}(z-x)}\mathrm{dx} = \frac{\mathrm{e}^{-\frac{1}{2}\mathrm{z}}}{2\pi}\int_0^z\frac{1}{\sqrt{x(z-x)}}dx$

De laatste integraal is gelijk aan π (gegeven): $\mathrm{f}_{X+Y}(\mathrm{z}) = \frac{1}{2}\mathrm{e}^{-\frac{1}{2}\mathrm{z}}$ voor $\mathrm{z} > 0$.

De Chi-kwadraat verdeling met 2 vrijheidsgraden is dus gelijk aan de Exp $\left(\frac{1}{2}\right)$-verdeling.

5. **a**. Ja, de regel $E(X+Y) = E(X) + E(Y)$ geldt altijd

b. Nee, dat geldt alléén bij onafhankelijkheid, hetgeen hier een onredelijke veronderstelling is de hoogte van inkomens zijn afhankelijk (denk maar aan opleidingsniveau, aan elkaar aangepaste deeltijdbanen e.d.)

6. $P(X>Y)=P(X-Y>0)=1-\Phi\left(\frac{0-(-10)}{\sqrt{136}}\right)\approx 0.1965$, dus naar verwachting zullen er (afgerond) 20 van de 100 breken.

7. **a**. $\Phi(-1.80)=0.0359$
b. $\frac{3}{5}$

8. **a**. $X+Y\sim N(75+65,250+150)$, dus $P(X+Y>150)=1-\Phi\left(\frac{150-140}{\sqrt{400}}\right)=1-\Phi\left(\frac{1}{2}\right)$
b. $\sum_{i=1}^{100}X_i \overset{\text{CLS}}{\sim} N\left(100\cdot\frac{1}{2},100\cdot\frac{1}{4}\right)$, dus $P(\sum_{i=1}^{100}X_i\leq 58)=\Phi\left(\frac{58-50}{\sqrt{25}}\right)=\Phi(1.6)$

9. **a**. onafhankelijkheid van de 100 verwerkingsduren en alle dezelfde verdeling met $\mu=95$ en $\sigma = 20$.
b. $\overline{X}=\frac{1}{100}\sum_{i=1}^{100}X_i \overset{\text{CLS}}{\sim} N\left(\mu,\frac{\sigma^2}{100}\right)$, dus $N(95,4)$ en

$$P(\overline{X}>100)\overset{\text{CLS}}{\approx} 1-\Phi\left(\frac{100-95}{\sqrt{4}}\right)=1-\Phi(2.5)=0.62\%$$

10. **a**. $X\sim B(250,0.25)$ dus X is bij benadering $N(np,np(1-p))=N(62.5,46.5)$ (aan de vuistregel $n>25$, $np>5$ en $n(1-p)>5$ is voldaan)
b. 22% van 250 is 55 kiezers:
$P(X\leq 55)\overset{\text{c.c.}}{=}P(X\leq 55.5)\overset{\text{CLS}}{\approx}\Phi\left(\frac{55.5-62.5}{\sqrt{46.5}}\right)\approx\Phi(-1.03)=0.1515$
c. Nu is $\mu=np=250\cdot 0.01=2.5<5$, dus Poisson benaderen met $\mu = 2.5$.
$P(X>5)=1-P(X\leq 5)\approx 4.2\%$

11. **a**. $X\sim B(100,p)$, dus $\mu = 100p$ en $\sigma^2=100p(1-p)$
b. $\frac{X}{100}\overset{\text{CLS}}{\sim}N\left(p,\frac{p(1-p)}{100}\right)$, dus $P\left(-0.05\leq\frac{X}{100}-p\leq 0.05\right)$

$$=P\left(-\frac{0.05}{\sqrt{\frac{p(1-p)}{100}}}\leq Z\leq\frac{0.05}{\sqrt{\frac{p(1-p)}{100}}}\right)=\Phi\left(\frac{0.5}{\sqrt{p(1-p)}}\right)-\Phi\left(-\frac{0.5}{\sqrt{p(1-p)}}\right)$$

$p(1-p)\leq\frac{1}{4}$, dus deze kans is minimaal $\Phi(1)-\Phi(-1)=68.21\%$

12. $X=$ "vraag naar i-pads in 6 werkdagen" is Poisson($\mu=6\cdot 6=36$)-verdeeld.
Deze verdeling wordt benaderd met de $N(36,36)$-verdeling.
a. $P(X\leq 40)\overset{\text{c.c.}}{=}P(X\leq 40.5)=\Phi\left(\frac{40.5-36}{\sqrt{36}}\right)=\Phi(0.75)=77.34\%$
b. $P(X\leq s)\overset{\text{c.c.}}{=}P(X\leq s+0.5)=\Phi\left(\frac{s+0.5-36}{\sqrt{36}}\right)\geq 99\%$, dan $\frac{s+0.5-36}{\sqrt{36}}=2.33$
dus $s=35.5+2.33\cdot 6=49.48$. de safety stock s moet (minstens) 50 zijn.

13. **a.** $P(X\geq 1)=1-P(X=0)=1-0.99^{15}=14.0\%$
b. $P(X\leq 3)=0.857$ (Poisson –table with $\mu=2$)
c. $P(X\geq 50)=1-P(Z\leq 1.51)=1-0.9345\approx 6.5\%$

Hoofdstuk 8

1. **a**. $E(X) = 60$ en $P(X \geq 30) = P(X \geq 90 | X \geq 60) = e^{-\frac{1}{2}}$.
c. 38.12 $cent$ (> 30 $cent$) **d**. 881.48 $cent^2$

2. **a**. $P(Y > t) = P(X = 0)$ en **b**. $P(X = 0) = \frac{(at)^0 e^{-at}}{0!} = e^{-at}$
c. $P(Y > t) = e^{-at}$ hoort bij een exponentiele verdeling met $\lambda = a$,

3. **b.** $E(Y) = 1$ en $E(X) = \frac{1}{2}$, dus $1 = E(Y) \neq \ln\left(\frac{1}{EX}\right) = \ln(2)$

4. **a**. $E(X_i) = 12$, $var(X_i) = 144$ en $E(\sum_{i=1}^{6} X_i) = 12 \cdot 6 = 72$
b. $P(X_1 > 12) = P(X_1 > 15 | X_1 > 3) = e^{-\frac{1}{12} \cdot 12} = e^{-1}$.
c. $N \sim Poisson(\lambda t)$, met $\lambda t = \frac{1}{12} \cdot 60 = 5$, dus $P(N \geq 6) = 1 - P(X \leq 5) = 38.4\%$.
d. $P(\sum_{i=1}^{6} X_i) = P(N \geq 6) = 38.4\%$

5. **a**. S_n heeft een Erlang verdeling met parameters n en $\lambda = 1/4$. $E(S_n) = \frac{n}{\lambda} = 4n$ en
$var(S_n) = \frac{n}{\lambda^2} = 16n$
c. $P(\overline{X}_2 > 5) = \frac{7}{2} e^{-\frac{3}{2}}$ **d.** $P(\overline{X}_{100} > 5) \approx 0.62\%$.

Bijlage Wiskundige Technieken voor Kansrekening

(zie voor meer gedetailleerde informatie het calculus boek)

Reeksen

1. **Binomium van Newton:**

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k\, b^{n-k}$$

Toepassing: de som van alle binomiale kansen is 1:

$$\sum_{k=0}^{n} \binom{n}{k} p^k\, (1-p)^{n-k} = (p+1-p)^n = 1$$

2. **Meetkundige of geometrische reeks:**

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

Toepassing: de som van alle kansen van de geometrische verdeling is 1

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

Afgeleide van de Meetkundige reeks (naar x):

$$\sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2}$$

Toepassing: de verwachting van de geometrische verdeling is $1/p$:

$$E(X) = \sum_{i=1}^{\infty} i \cdot (1-p)^{i-1}p \overset{k=i}{=} p \cdot \sum_{k=0}^{\infty} k \cdot (1-p)^{k-1} = p \cdot \frac{1}{\big(1-(1-p)\big)^2} = \frac{1}{p}$$

Eindige meetkundige reeks:

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

Toepassing: som van geometrische kansen:

$$P(X \le 10) = \sum_{i=1}^{10}(1-p)^{i-1}p \overset{k=i-1}{=} p \cdot \sum_{k=0}^{9}(1-p)^k = p \cdot \frac{1-(1-p)^{10}}{1-(1-p)}$$

3. **Taylor-reeksontwikkeling van een functie $\boldsymbol{f(x)}$ om $\boldsymbol{x=0}$**:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k, \qquad \text{toegepast op } f(x) = e^x\text{:}\;\; e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

Toepassing: de som van de kansen en verwachting van de Poisson-verdeling:

$$\sum_{k=0}^{\infty} \frac{\mu^k}{k!} \cdot e^{-\mu} = e^{\mu} \cdot e^{-\mu} = 1 \quad \text{en}$$

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

Differentiëren en Integreren

4. Kettingregel:

$$\frac{d}{dx}\left[f\big(g(x)\big)\right] = f'\big(g(x)\big) \cdot g'(x)$$

Productregel:

$$\frac{d}{dx}[f(x)g(x)] = f'(x) \cdot g\,(x) + f(x) \cdot g'(x)$$

5. Hoofdstelling van de Algebra:

$$\int_a^b f(x)dx = F(b) - F(a), \quad \text{waarin } F \text{ een primitieve van } f \text{ is, dus } F'(x) = f(x)$$

Toepassing: kansen berekenen voor een continue variabele X: f_X is de kansdichtheid en F_X de verdelingsfunctie. $P(a < X \leq b) = \int_a^b f_X(x)dx = F_X(b) - F_X(a)$

6. Partiële Integratie:

$$\int_a^b f(x)g'(x)dx = f(x)g(x)\Big|_{x=a}^{x=b} - \int_a^b f'(x)g(x)dx$$

Toepassing: berekening van de verwachting van de exponentiele verdeling:

$$E(X) = \int_0^{\infty} x \cdot \lambda e^{-\lambda x} dx\,, waarin\, f(x) = x\ en\ g\,`(x) = \lambda e^{-\lambda x}, zodat\ g\,(x) = -e^{-\lambda x},$$

Dus: $E(X) = \int_0^{\infty} x \cdot \lambda e^{-\lambda x} dx = x \cdot -e^{-\lambda x}\Big|_{x=0}^{\infty} + \int_0^{\infty} e^{-\lambda x} dx = 0 - \frac{1}{\lambda} e^{-\lambda x}\Big|_{x=0}^{\infty} = \frac{1}{\lambda}$

Trefwoordenregister

Engelse woordenlijst Kansrekening

Namen van verdelingen e.d. zijn i.h.a. hetzelfde, zoals binomial, geometric, exponential etc.

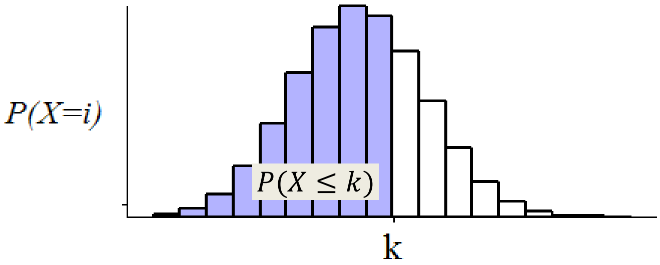
(Bernoulli) trial	(Bernoulli) poging of experiment
Central Limit Theorem	Centrale Limiet Stelling
conditional distribution	voorwaardelijke verdeling
conditional probability	voorwaardelijke kans
correlation (coefficient)	correlatie(coëfficiënt)
covariance	covariantie
disjoint	disjunct, elkaar uitsluitend
distribution function	verdelingsfunctie
event	gebeurtenis
expectation	verwachting
expected value	verwachtingswaarde
independent	onderling) onafhankelijk
joint distribution	simultane verdeling
mean	steekproefgemiddelde òf verwachting
marginal distribution	marginale verdeling
mode	modus
mutually exclusive	elkaar uitsluitend
population mean	Verwachting, populatiegemiddelde
population proportion	populatiefractie
(probability) density function	kansdichtheid
(probability) distribution	(kans)verdeling
probability (mass) function	kansfunctie
probability (measure) random	kans(maat)
sample	aselecte steekproef
random variable	stochastische variabele
sample mean	steekproefgemiddelde
sample proportion	steekproeffractie
sample size	steekproefuitgebreidheid
sample space	uitkomstenruimte
simple event	elementaire gebeurtenis
standard deviation	Standaardafwijking, standaaddeviatie
variance	variantie
Weak Law of Large Numbers	Zwakke wet van grote aantallen

Tabel van binomiale kansen

De tabellen bevatten cumulatieve kansen

$$P(X \leq k) = \sum_{i=0}^{k} P(X = i)$$

(in drie decimalen nauwkeurig)



n **= 5**

k \ *p*	**0,01**	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**	**1/6**	**1/3**
0	0,951	0,774	0,590	0,444	0,328	0,237	0,168	0,116	0,078	0,050	0,031	0,402	0,132
1	0,999	0,977	0,919	0,835	0,737	0,633	0,528	0,428	0,337	0,256	0,188	0,804	0,461
2	1,000	0,999	0,991	0,973	0,942	0,896	0,837	0,765	0,683	0,593	0,500	0,965	0,790
3		1,000	1,000	0,998	0,993	0,984	0,969	0,946	0,913	0,869	0,813	0,997	0,955
4				1,000	1,000	0,999	0,998	0,995	0,990	0,982	0,969	1,000	0,996

n **= 6**

k \ *p*	**0,01**	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**	**1/6**	**1/3**
0	0,941	0,735	0,531	0,377	0,262	0,178	0,118	0,075	0,047	0,028	0,016	0,335	0,088
1	0,999	0,967	0,886	0,776	0,655	0,534	0,420	0,319	0,233	0,164	0,109	0,737	0,351
2	1,000	0,998	0,984	0,953	0,901	0,831	0,744	0,647	0,544	0,442	0,344	0,938	0,680
3		1,000	0,999	0,994	0,983	0,962	0,930	0,883	0,821	0,745	0,656	0,991	0,900
4			1,000	1,000	0,998	0,995	0,989	0,978	0,959	0,931	0,891	0,999	0,982
5					1,000	1,000	0,999	0,998	0,996	0,992	0,984	1,000	0,999

n **= 7**

k \ *p*	**0,01**	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**	**1/6**	**1/3**
0	0,932	0,698	0,478	0,321	0,210	0,133	0,082	0,049	0,028	0,015	0,008	0,279	0,059
1	0,998	0,956	0,850	0,717	0,577	0,445	0,329	0,234	0,159	0,102	0,063	0,670	0,263
2	1,000	0,996	0,974	0,926	0,852	0,756	0,647	0,532	0,420	0,316	0,227	0,904	0,571
3		1,000	0,997	0,988	0,967	0,929	0,874	0,800	0,710	0,608	0,500	0,982	0,827
4			1,000	0,999	0,995	0,987	0,971	0,944	0,904	0,847	0,773	0,998	0,955
5				1,000	1,000	0,999	0,996	0,991	0,981	0,964	0,938	1,000	0,993
6						1,000	1,000	0,999	0,998	0,996	0,992		1,000

n **= 8**

k \ *p*	**0,01**	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**	**1/6**	**1/3**
0	0,923	0,663	0,430	0,272	0,168	0,100	0,058	0,032	0,017	0,008	0,004	0,233	0,039
1	0,997	0,943	0,813	0,657	0,503	0,367	0,255	0,169	0,106	0,063	0,035	0,605	0,195
2	1,000	0,994	0,962	0,895	0,797	0,679	0,552	0,428	0,315	0,220	0,145	0,865	0,468
3		1,000	0,995	0,979	0,944	0,886	0,806	0,706	0,594	0,477	0,363	0,969	0,741
4			1,000	0,997	0,990	0,973	0,942	0,894	0,826	0,740	0,637	0,995	0,912
5				1,000	0,999	0,996	0,989	0,975	0,950	0,912	0,855	1,000	0,980
6					1,000	1,000	0,999	0,996	0,991	0,982	0,965		0,997
7							1,000	1,000	0,999	0,998	0,996		1,000

Tab-2

***n* = 9**

k \ *p*	**0,01**	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**	**1/6**	**1/3**
0	0,914	0,630	0,387	0,232	0,134	0,075	0,040	0,021	0,010	0,005	0,002	0,194	0,026
1	0,997	0,929	0,775	0,599	0,436	0,300	0,196	0,121	0,071	0,039	0,020	0,543	0,143
2	1,000	0,992	0,947	0,859	0,738	0,601	0,463	0,337	0,232	0,150	0,090	0,822	0,377
3		0,999	0,992	0,966	0,914	0,834	0,730	0,609	0,483	0,361	0,254	0,952	0,650
4		1,000	0,999	0,994	0,980	0,951	0,901	0,828	0,733	0,621	0,500	0,991	0,855
5			1,000	0,999	0,997	0,990	0,975	0,946	0,901	0,834	0,746	0,999	0,958
6				1,000	1,000	0,999	0,996	0,989	0,975	0,950	0,910	1,000	0,992
7						1,000	1,000	0,999	0,996	0,991	0,980		0,999
8								1,000	1,000	0,999	0,998		1,000

***n* = 10**

k \ *p*	**0,01**	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**	**1/6**	**1/3**
0	0,904	0,599	0,349	0,197	0,107	0,056	0,028	0,013	0,006	0,003	0,001	0,162	0,017
1	0,996	0,914	0,736	0,544	0,376	0,244	0,149	0,086	0,046	0,023	0,011	0,485	0,104
2	1,000	0,988	0,930	0,820	0,678	0,526	0,383	0,262	0,167	0,100	0,055	0,775	0,299
3		0,999	0,987	0,950	0,879	0,776	0,650	0,514	0,382	0,266	0,172	0,930	0,559
4		1,000	0,998	0,990	0,967	0,922	0,850	0,751	0,633	0,504	0,377	0,985	0,787
5			1,000	0,999	0,994	0,980	0,953	0,905	0,834	0,738	0,623	0,998	0,923
6				1,000	0,999	0,996	0,989	0,974	0,945	0,898	0,828	1,000	0,980
7					1,000	1,000	0,998	0,995	0,988	0,973	0,945		0,997
8							1,000	0,999	0,998	0,995	0,989		1,000
9								1,000	1,000	1,000	0,999		

***n* = 15**

k \ *p*	**0,01**	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**	**1/6**	**1/3**
0	0,860	0,463	0,206	0,087	0,035	0,013	0,005	0,002	0,000	0,000	0,000	0,065	0,002
1	0,990	0,829	0,549	0,319	0,167	0,080	0,035	0,014	0,005	0,002	0,000	0,260	0,019
2	1,000	0,964	0,816	0,604	0,398	0,236	0,127	0,062	0,027	0,011	0,004	0,532	0,079
3		0,995	0,944	0,823	0,648	0,461	0,297	0,173	0,091	0,042	0,018	0,768	0,209
4		0,999	0,987	0,938	0,836	0,686	0,515	0,352	0,217	0,120	0,059	0,910	0,404
5		1,000	0,998	0,983	0,939	0,852	0,722	0,564	0,403	0,261	0,151	0,973	0,618
6			1,000	0,996	0,982	0,943	0,869	0,755	0,610	0,452	0,304	0,993	0,797
7				0,999	0,996	0,983	0,950	0,887	0,787	0,654	0,500	0,999	0,912
8				1,000	0,999	0,996	0,985	0,958	0,905	0,818	0,696	1,000	0,969
9					1,000	0,999	0,996	0,988	0,966	0,923	0,849		0,991
10						1,000	0,999	0,997	0,991	0,975	0,941		0,998
11							1,000	1,000	0,998	0,994	0,982		1,000
12									1,000	0,999	0,996		
13										1,000	1,000		

Tab-3

***n* = 20**

p / k	0,01	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50	1/6	1/3
0	0,818	0,358	0,122	0,039	0,012	0,003	0,001	0,000	0,000	0,000	0,000	0,026	0,000
1	0,983	0,736	0,392	0,176	0,069	0,024	0,008	0,002	0,001	0,000	0,000	0,130	0,003
2	0,999	0,925	0,677	0,405	0,206	0,091	0,035	0,012	0,004	0,001	0,000	0,329	0,018
3	1,000	0,984	0,867	0,648	0,411	0,225	0,107	0,044	0,016	0,005	0,001	0,567	0,060
4		0,997	0,957	0,830	0,630	0,415	0,238	0,118	0,051	0,019	0,006	0,769	0,152
5		1,000	0,989	0,933	0,804	0,617	0,416	0,245	0,126	0,055	0,021	0,898	0,297
6			0,998	0,978	0,913	0,786	0,608	0,417	0,250	0,130	0,058	0,963	0,479
7			1,000	0,994	0,968	0,898	0,772	0,601	0,416	0,252	0,132	0,989	0,661
8				0,999	0,990	0,959	0,887	0,762	0,596	0,414	0,252	0,997	0,809
9				1,000	0,997	0,986	0,952	0,878	0,755	0,591	0,412	0,999	0,908
10					0,999	0,996	0,983	0,947	0,872	0,751	0,588	1,000	0,962
11					1,000	0,999	0,995	0,980	0,943	0,869	0,748		0,987
12						1,000	0,999	0,994	0,979	0,942	0,868		0,996
13							1,000	0,998	0,994	0,979	0,942		0,999
14								1,000	0,998	0,994	0,979		1,000
15									1,000	0,998	0,994		
16										0,998	0,994		
17										1,000	1,000		

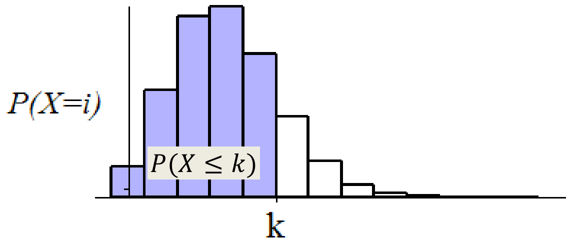
***n* = 25**

p / k	0,01	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50	1/6	1/3
0	0,778	0,277	0,072	0,017	0,004	0,001	0,000	0,000	0,000	0,000	0,000	0,010	0,000
1	0,974	0,642	0,271	0,093	0,027	0,007	0,002	0,000	0,000	0,000	0,000	0,063	0,001
2	0,998	0,873	0,537	0,254	0,098	0,032	0,009	0,002	0,000	0,000	0,000	0,189	0,004
3	1,000	0,966	0,764	0,471	0,234	0,096	0,033	0,010	0,002	0,000	0,000	0,382	0,015
4		0,993	0,902	0,682	0,421	0,214	0,090	0,032	0,009	0,002	0,000	0,594	0,046
5		0,999	0,967	0,838	0,617	0,378	0,193	0,083	0,029	0,009	0,002	0,772	0,112
6		1,000	0,991	0,930	0,780	0,561	0,341	0,173	0,074	0,026	0,007	0,891	0,222
7			0,998	0,975	0,891	0,727	0,512	0,306	0,154	0,064	0,022	0,955	0,370
8			1,000	0,992	0,953	0,851	0,677	0,467	0,274	0,134	0,054	0,984	0,538
9				0,998	0,983	0,929	0,811	0,630	0,425	0,242	0,115	0,995	0,696
10				1,000	0,994	0,970	0,902	0,771	0,586	0,384	0,212	0,999	0,822
11					0,998	0,989	0,956	0,875	0,732	0,543	0,345	1,000	0,908
12					1,000	0,997	0,983	0,940	0,846	0,694	0,500		0,958
13						0,999	0,994	0,975	0,922	0,817	0,655		0,984
14						1,000	0,998	0,991	0,966	0,904	0,788		0,994
15							1,000	0,997	0,987	0,956	0,885		0,998
16								0,999	0,996	0,983	0,946		1,000
17								1,000	0,999	0,994	0,978		
18									1,000	0,998	0,993		
19										1,000	0,998		
20											1,000		

Tabel Poisson-kansen

De tabellen bevatten cumulatieve kansen

$$P(X \leq k) = \sum_{i=0}^{k} \frac{\mu^i e^{-\mu}}{i!}$$



(Afgerond op drie decimalen)

μ \ k	0	1	2	3	4	5	6	7	8	9	10
0.02	0.980	1.000									
0.04	0.961	0.999	1.000								
0.06	0.942	0.998	1.000								
0.08	0.923	0.997	1.000								
0.10	0.905	0.995	1.000								
0.15	0.861	0.990	0.999	1.000							
0.20	0.819	0.982	0.999	1.000							
0.25	0.779	0.974	0.998	1.000							
0.30	0.741	0.963	0.996	1.000							
0.35	0.705	0.951	0.994	1.000							
0.40	0.670	0.938	0.992	0.999	1.000						
0.45	0.638	0.925	0.989	0.999	1.000						
0.50	0.607	0.910	0.986	0.998	1.000						
0.55	0.577	0.894	0.982	0.998	1.000						
0.60	0.549	0.878	0.977	0.997	1.000						
0.65	0.522	0.861	0.972	0.996	0.999	1.000					
0.70	0.497	0.844	0.966	0.994	0.999	1.000					
0.75	0.472	0.827	0.959	0.993	0.999	1.000					
0.80	0.449	0.809	0.953	0.991	0.999	1.000					
0.85	0.427	0.791	0.945	0.989	0.998	1.000					
0.90	0.407	0.772	0.937	0.987	0.998	1.000					
0.95	0.387	0.754	0.929	0.984	0.997	1.000					
1.00	0.368	0.736	0.920	0.981	0.996	0.999	1.000				
1.1	0.333	0.699	0.900	0.974	0.995	0.999	1.000				
1.2	0.301	0.663	0.879	0.966	0.992	0.998	1.000				
1.3	0.273	0.627	0.857	0.957	0.989	0.998	1.000				
1.4	0.247	0.592	0.833	0.946	0.986	0.997	0.999	1.000			
1.5	0.223	0.558	0.809	0.934	0.981	0.996	0.999	1.000			
1.6	0.202	0.525	0.783	0.921	0.976	0.994	0.999	1.000			
1.7	0.183	0.493	0.757	0.907	0.970	0.992	0.998	1.000			
1.8	0.165	0.463	0.731	0.891	0.964	0.990	0.997	0.999	1.000		
1.9	0.150	0.434	0.704	0.875	0.956	0.987	0.997	0.999	1.000		
2.0	0.135	0.406	0.677	0.857	0.947	0.983	0.995	0.999	1.000		
2.2	0.111	0.355	0.623	0.819	0.928	0.975	0.993	0.998	1.000		
2.4	0.091	0.308	0.570	0.779	0.904	0.964	0.988	0.997	0.999	1.000	
2.6	0.074	0.267	0.518	0.736	0.877	0.951	0.983	0.995	0.999	1.000	
2.8	0.061	0.231	0.469	0.692	0.848	0.935	0.976	0.992	0.998	0.999	1.000

Poissonkansen (Vervolg)

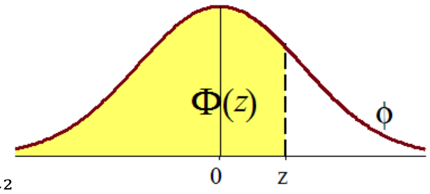
μ \ k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
3.0	0.050	0.199	0.423	0.647	0.815	0.916	0.966	0.988	0.996	0.999	1.000					
3.2	0.041	0.171	0.380	0.603	0.781	0.895	0.955	0.983	0.994	0.998	1.000					
3.4	0.033	0.147	0.340	0.558	0.744	0.871	0.942	0.977	0.992	0.997	0.999	1.000				
3.6	0.027	0.126	0.303	0.515	0.706	0.844	0.927	0.969	0.988	0.996	0.999	1.000				
3.8	0.022	0.107	0.269	0.473	0.668	0.816	0.909	0.960	0.984	0.994	0.998	0.999	1.000			
4.0	0.018	0.092	0.238	0.433	0.629	0.785	0.889	0.949	0.979	0.992	0.997	0.999	1.000			
4.2	0.015	0.078	0.210	0.395	0.590	0.753	0.867	0.936	0.972	0.989	0.996	0.999	1.000			
4.4	0.012	0.066	0.185	0.359	0.551	0.720	0.844	0.921	0.964	0.985	0.994	0.998	0.999	1.000		
4.6	0.010	0.056	0.163	0.326	0.513	0.686	0.818	0.905	0.955	0.980	0.992	0.997	0.999	1.000		
4.8	0.008	0.048	0.143	0.294	0.476	0.651	0.791	0.887	0.944	0.975	0.990	0.996	0.999	1.000		
5.0	0.007	0.040	0.125	0.265	0.440	0.616	0.762	0.867	0.932	0.968	0.986	0.995	0.998	0.999	1.000	
5.2	0.006	0.034	0.109	0.238	0.406	0.581	0.732	0.845	0.918	0.960	0.982	0.993	0.997	0.999	1.000	
5.4	0.005	0.029	0.095	0.213	0.373	0.546	0.702	0.822	0.903	0.951	0.977	0.990	0.996	0.999	1.000	
5.6	0.004	0.024	0.082	0.191	0.342	0.512	0.670	0.797	0.886	0.941	0.972	0.988	0.995	0.998	0.999	1.000
5.8	0.003	0.021	0.072	0.170	0.313	0.478	0.638	0.771	0.867	0.929	0.965	0.984	0.993	0.997	0.999	1.000
6.0	0.002	0.017	0.062	0.151	0.285	0.446	0.606	0.744	0.847	0.916	0.957	0.980	0.991	0.996	0.999	0.999
6.2	0.002	0.015	0.054	0.134	0.259	0.414	0.574	0.716	0.826	0.902	0.949	0.975	0.989	0.995	0.998	0.999
6.4	0.002	0.012	0.046	0.119	0.235	0.384	0.542	0.687	0.803	0.886	0.939	0.969	0.986	0.994	0.997	0.999
6.6	0.001	0.010	0.040	0.105	0.213	0.355	0.511	0.658	0.780	0.869	0.927	0.963	0.982	0.992	0.997	0.999
6.8	0.001	0.009	0.034	0.093	0.192	0.327	0.480	0.628	0.755	0.850	0.915	0.955	0.978	0.990	0.996	0.998
7.0	0.001	0.007	0.030	0.082	0.173	0.301	0.450	0.599	0.729	0.830	0.901	0.947	0.973	0.987	0.994	0.998
7.2	0.001	0.006	0.025	0.072	0.156	0.276	0.420	0.569	0.703	0.810	0.887	0.937	0.967	0.984	0.993	0.997
7.4	0.001	0.005	0.022	0.063	0.140	0.253	0.392	0.539	0.676	0.788	0.871	0.926	0.961	0.980	0.991	0.996
7.6	0.001	0.004	0.019	0.055	0.125	0.231	0.365	0.510	0.648	0.765	0.854	0.915	0.954	0.976	0.989	0.995
7.8	0.000	0.004	0.016	0.048	0.112	0.210	0.338	0.481	0.620	0.741	0.835	0.902	0.945	0.971	0.986	0.993
8.0	0.000	0.003	0.014	0.042	0.100	0.191	0.313	0.453	0.593	0.717	0.816	0.888	0.936	0.966	0.983	0.992
8.5	0.000	0.002	0.009	0.030	0.074	0.150	0.256	0.386	0.523	0.653	0.763	0.849	0.909	0.949	0.973	0.986
9.0	0.000	0.001	0.006	0.021	0.055	0.116	0.207	0.324	0.456	0.587	0.706	0.803	0.876	0.926	0.959	0.978
9.5	0.000	0.001	0.004	0.015	0.040	0.089	0.165	0.269	0.392	0.522	0.645	0.752	0.836	0.898	0.940	0.967
10.0	0.000	0.000	0.003	0.010	0.029	0.067	0.130	0.220	0.333	0.458	0.583	0.697	0.792	0.864	0.917	0.951

μ \ k	16	17	18	19	20	21	22
6.0	1.000						
5.2	1.000						
6.4	1.000						
6.6	0.999	1.000					
6.8	0.999	1.000					
7.0	0.999	1.000					
7.2	0.999	1.000					
7.4	0.998	0.999	1.000				
7.6	0.998	0.999	1.000				
7.8	0.997	0.999	1.000				
8.0	0.996	0.998	0.999	1.000			
8.5	0.993	0.997	0.999	0.999	1.000		
9.0	0.989	0.995	0.998	0.999	1.000		
9.5	0.982	0.991	0.996	0.998	0.999	1.000	
10.0	0.973	0.986	0.993	0.997	0.998	0.999	1.000

De standaardnormale verdeling

De tabel geeft de verdelingsfunctie Φ voor een N(0, 1)-verdeelde Z

$$\Phi(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}}\, dx$$



Laatste kolom: de dichtheid van de N(0,1)-verdeling: $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

	Tweede decimaal z										
z	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**	φ(z)
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359	0,3989
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753	0,3970
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141	0,3910
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517	0,3814
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879	0,3683
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224	0,3521
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549	0,3332
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852	0,3123
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133	0,2897
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389	0,2661
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621	0,2420
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830	0,2179
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015	0,1942
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177	0,1714
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319	0,1497
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441	0,1295
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545	0,1109
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633	0,0940
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706	0,0790
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767	0,0656
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817	0,0540
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857	0,0440
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890	0,0355
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916	0,0283
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936	0,0224
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952	0,0175
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964	0,0136
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974	0,0104
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981	0,0079
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986	0,0060
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990	0,0044
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993	0,0033
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995	0,0024
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997	0,0017
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998	0,0012
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,0009
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,0006