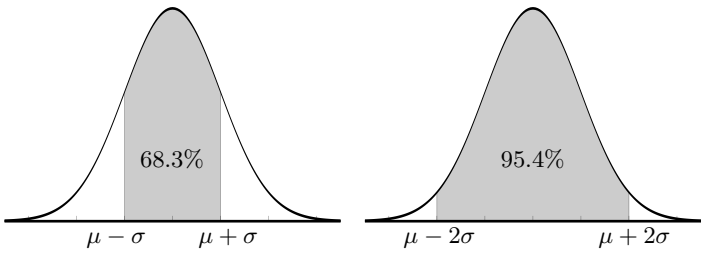
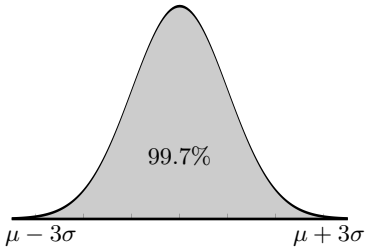
KANSREKENING EN STATISTIEK

Jan van de Craats





Collegedictaat – augustus 2002

Inhoudsopgave

Hoofdstuk 1

Kansen en kansmodellen

1.1 Introductie

In het dagelijks leven hebben we het vaak over kansen, bijvoorbeeld

1 de kans op regen of zon,
2 de kans op kruis of munt bij het opgooien van een geldstuk,
3 de kans op het winnen van een hoofdprijs in de Staatsloterij,
4 de kans dat een vliegtuig neerstort,
5 de kans op een jaar lang schadevrij autorijden,
6 de kans op genezing dankzij het gebruik van een bepaald medicijn,
7 de kans op een overstroming als de rivierdijken op deltahoogte zijn gebracht,
8 de kans dat het record op de 10 kilometer sneuvelt op het eerstkomende WK schaatsen,
9 de kans op het aftreden van een minister naar aanleiding van een geruchtmakende politieke affaire,
10 de kans op escalatie van een diplomatiek incident tot een gewapend conflict.

In al deze gevallen gaat het om verwachtingen die men koestert over mogelijke gebeurtenissen in de toekomst. Welke gebeurtenis daadwerkelijk zal optreden is onzeker. Maar naast deze overeenkomst zijn er ook verschillen. Neem de voorbeelden 1 en 2. Die verschillen hemelsbreed. In voorbeeld 1 hebben we geen helder idee wat er met 'kans' bedoeld wordt. Is dat de kans op regenval, van hoe korte duur ook, binnen een etmaal ? En geldt hetzelfde voor de kans op zon? Zijn regen en zon categorieën die elkaar uitsluiten? Zijn ze 'uitputtend', dat wil zeggen, zijn er geen andere toestanden dan regen en zon? En valt er iets zinnigs te zeggen over de grootte van de kans? Vergeleken met voorbeeld 1 is voorbeeld 2 veel minder problematisch. Om te beginnen hebben we hier ogenblikkelijk een idee – in ieder geval intuïtief – wat er met 'kans' bedoeld wordt. Kruis en munt zijn categorieën die elkaar uitsluiten en hun uitputtendheid is evident: bij het tossen met een munt is er geen derde mogelijkheid. Ook de grootte van de kansen is niet problematisch: als men aanneemt dat er niet met de munt geknoeid is, zal men veronderstellen dat de kans op kruis net zo groot is als de kans op munt. Omdat de twee kansen samen 1 zijn (of honderd procent, maar dat is hetzelfde) luidt de conclusie dat beide kansen gelijk moeten zijn aan een half.

Van het begin van het rijtje stappen we over naar het eind ervan, en beschouwen de laatste twee voorbeelden. Daarbij is het duidelijk dat het in beide gevallen gaat om twee mogelijke gebeurtenissen die elkaar uitsluiten en die uitputtend zijn. De minister treedt af of niet; het incident escaleert of niet. En in beide gevallen

geldt: andere mogelijkheden zijn er niet. Anderzijds hebben beide voorbeelden ook een problematische kant. We hebben namelijk geen idee van de grootte van de betreffende kans, behalve dan dat die klein zal zijn. Immers, ministers treden meestal niet af en de meeste diplomatieke incidenten escaleren niet. Maar verder komen we niet.

1.2 Discrete kansmodellen

Wat maakt nu dat sommige uitspraken waarin kansen voorkomen slechts slagen in de lucht zijn, terwijl andere als verantwoord gelden? Wij zullen een uitspraak over kansen slechts verantwoord noemen als er een wiskundig model achter zit.

wiskundig model

Niet alle situaties waarin men over kansen spreekt, laten zich echter gemakkelijk in wiskundige modellen vertalen.
In dit hoofdstuk gaan we in op een aantal voorbeelden waarin wiskundig modelleerbare kansen centraal staan. In de loop van die bespreking komen relevante aspecten die in het voorgaande op intuïtief niveau werden aangestipt uitgebreider terug, en tegen het einde ook in een meer formeel kader.

Voorbeeld 1.1 *Het gooien met een dobbelsteen*

Een van de schoolvoorbeelden uit de kansrekening is het gooien met een dobbelsteen. Als uitkomst van zo'n 'kansexperiment' nemen we het aantal ogen dat na het werpen boven ligt; de mogelijke uitkomsten zijn dus de getallen 1 tot en met 6. We kunnen het experiment net zo vaak herhalen als we willen, en zo een rij uitkomsten genereren, bijvoorbeeld

kansexperiment

$$3, 1, 6, 6, 2, 4, 1, 5, 3, \ldots$$

In een concreet geval kennen we de uitkomst van het experiment niet voordat we het gedaan hebben. Toch zijn er wel verstandige dingen over te zeggen. Basis daarvoor is een wiskundig model.

Wanneer we een wiskundig model willen construeren van het werpen met een dobbelsteen, beginnen we met een uitkomstenruimte U, die model staat voor de verzameling van de mogelijke uitkomsten van het experiment. In dit geval ligt het voor de hand om $U = \{1, 2, 3, 4, 5, 6\}$ te nemen. Vervolgens willen we aan elke uitkomst een kans toekennen. Men zal in het geval van de dobbelsteen meestal aannemen dat het ding volkomen homogeen van samenstelling is en een kubusvorm heeft die bij alle hoeken op dezelfde wijze is afgerond. Het ligt dan voor de hand om in het model aan de zes uitkomsten dezelfde kans toe te kennen. In het model hebben we het dan over een zuivere dobbelsteen. Is de echte dobbelsteen waarmee gegooid wordt niet homogeen van samenstelling, dan kan het voorkomen dat het wiskundige model van de zuivere dobbelsteen niet goed toepasbaar is. Voor een passend model moet men dan verschillende kansen voor de afzonderlijke uitkomsten nemen. Welke? Dat is dan natuurlijk de vraag.

uitkomstenruimte

Bij het dobbelsteenexperiment kan men ook uitkomsten samen nemen, en bijvoorbeeld spreken over de kans op een even uitkomst. De even uitkomsten vormen de deelverzameling $E = \{2, 4, 6\}$ van de uitkomstenruimte U. Het is gebruikelijk om zo'n deelverzameling van U met de algemene term gebeurtenis aan te duiden. Iedere uitkomst op zichzelf kan men ook opvatten als een deelverzameling van U, en dus is iedere uitkomst ook een gebeurtenis. We noemen uitkomsten in dit verband

gebeurtenis

ook wel 'elementaire' gebeurtenissen. Bij de dobbelsteen zijn die elementaire gebeurtenissen dus de uitkomsten $\{1\}$ tot en met $\{6\}$.

Het ligt in het zojuist genoemde voorbeeld voor de hand om de kans op E te definiëren als de som van de kansen op de afzonderlijke uitkomsten. In het algemeen kan men voor een gebeurtenis G de kans definiëren als de som van de kansen op de afzonderlijke elementaire gebeurtenissen waaruit G is samengesteld. In het bijzonder is de kans van de gehele uitkomstenruimte $U = \{1, 2, 3, 4, 5, 6\}$ de som van de kansen op de zes gebeurtenissen afzonderlijk. Natuurlijk moet die kans 1 zijn, want als men het experiment uitvoert, is men er honderd procent zeker van dat één van de zes mogelijke uitkomsten op zal treden. Het is dus duidelijk wat in het model van de zuivere dobbelsteen de kans op elk van de zes uitkomsten afzonderlijk moet zijn (voor zover u dat al niet vermoedde): ze zijn onderling gelijk en samen 1, dus elke uitkomst afzonderlijk heeft kans 1/6.

Zoals gezegd, het kan voorkomen dat men het idee heeft dat het model van de zuivere dobbelsteen niet goed toepasbaar is, bijvoorbeeld omdat de dobbelsteen niet homogeen is, of omdat hij niet een zuiver symmetrische kubusvorm heeft. In dat geval ligt het voor de hand om een model te kiezen waarin niet alle uitkomsten dezelfde kans krijgen. Bij het kiezen van de kansen in zo'n model kan men zich laten leiden door de zogenaamde *experimentele wet van de grote aantallen*. Deze wet drukt het *ervaringsfeit* uit dat de relatieve frequentie van het optreden van een gebeurtenis in een lange reeks onder dezelfde omstandigheden uitgevoerde herhalingen van hetzelfde experiment, op den duur steeds minder schommelingen gaat vertonen. Zo heeft men al vaak opgemerkt dat bij het herhaaldelijk tossen met een munt de relatieve frequentie van het aantal malen kop steeds meer de waarde 1/2 gaat benaderen. Tabel 1.1, ontleend aan het boek Introduction to Mathematical Statistics van E. Kreyszig, geeft de resultaten van drie van zulke lange experimentele series worpen, uitgevoerd door resp. Buffon en Pearson. Wat Buffon en Pearson

experimentele wet van de grote aantallen

Tabel 1.1: Werpen met een geldstuk

experiment door	aantal worpen	aantal malen kop	rel. freq.
G. Buffon	4040	2048	0.5069
K. Pearson	12000	6019	0.5016
K. Pearson	24000	12012	0.5005

met een munt hebben gedaan, kan men ook met een dobbelsteen doen: in een zeer lange serie worpen kan men voor elk van de zes uitkomsten de relatieve frequentie bepalen, dat wil zeggen het quotiënt van het aantal malen dat die uitkomst voorkwam en het totale aantal worpen. Wijken de uitkomsten erg af van 1/6, dan geeft dit steun aan het vermoeden dat er iets met de dobbelsteen mis is. Men zou dan deze experimenteel bepaalde quotiënten als kansen kunnen gebruiken in een kansmodel voor deze dobbelsteen. Merk daarbij op dat de som van de zes kansen (quotiënten) nog steeds 1 is.

relatieve frequentie

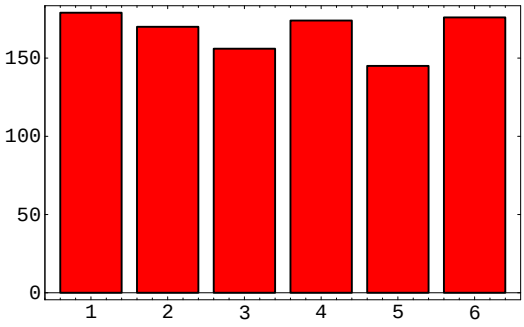
Dat men hierbij voorzichtig moet zijn, en niet te snel conclusies mag trekken, zelfs bij grote aantallen, kunt u zelf ervaren als u gebruik maakt van een *randomgenerator* zoals die in de meeste wiskundige software-pakketten aanwezig is. Vaak is dat een functie die bij aanroep een 'willekeurig' getal tussen 0 en 1 produceert. In de praktijk kan een computer, een deterministische automaat, natuurlijk nooit 'echte' toevalsgetallen produceren. Trouwens, over wat nu precies een rij 'toevalsgetallen'

randomgenerator

is, valt nog een uitgebreide verhandeling te schrijven. We laten die vraag hier rusten. Feit is dat de meeste randomgeneratoren slechts 'pseudo-randomgetallen' kunnen produceren, dat wil zeggen periodieke rijen met een zeer lange periode.
Het is eenvoudig om met zo'n randomgenerator het werpen met een dobbelsteen te simuleren: neem als uitkomst 1 als de waarde van het randomgetal tussen 0 en 1/6 ligt, 2 als de waarde tussen 1/6 en 2/6 ligt, enzovoort. In Tabel 1.2 ziet u de resultaten van zo'n serie van duizend 'worpen'. Onder de tabel zijn ze ook in de vorm van een staafdiagram weergegeven. U ziet dat de afwijkingen van de 'ideale' waarde 1/6 nog steeds aanzienlijk zijn!

Tabel 1.2: Duizend worpen met een gesimuleerde dobbelsteen

aantal ogen	aantal worpen	rel. freq.
1	179	0.179
2	170	0.170
3	156	0.156
4	174	0.174
5	145	0.145
6	176	0.176



Staafdiagram bij de tabel

Voorbeeld 1.2 *Massaproductie van gloeilampen.*

Bij de massaproductie van gloeilampen zullen naast 'goede' ook 'defecte' exemplaren van de band komen. Wanneer de machines goed zijn afgesteld, kan het redelijk zijn om aan te nemen dat de defecte exemplaren 'toevallig' optreden met een zekere vaste kans p $(0 < p < 1)$. Als wiskundig kansmodel kiest men dan dus een uitkomstenruimte $U = \{$*'defect', 'goed'*$\}$ met kansen resp. p en $1 - p$. Dit model kan men vervolgens gebruiken om een systeem van kwaliteitscontrole op te zetten, waarbij men bijvoorbeeld aan de hand van periodiek uitgevoerde steekproeven beslist of de instelling van de machines moet worden gecontroleerd.

Ook in dit geval kan men een randomgenerator gebruiken voor computersimulaties: een randomwaarde tussen 0 en p stelt dan een defect exemplaar voor, en een waarde tussen p en 1 een goed exemplaar. We hebben weer een simulatie uitgevoerd, en daarbij $p = 0.03$ gekozen. In een serie van duizend randomgetallen vonden we 24 waarden kleiner dan p en 976 waarden groter dan p.

Het tossen met een munt, het werpen met een dobbelsteen of het in massaproductie vervaardigen van gloeilampen zijn voorbeelden van kansexperimenten, dat wil zeggen experimenten die (in theorie) willekeurig vaak onder gelijke omstandigheden herhaald kunnen worden. Bij kansexperimenten kan men proberen bruikbare wiskundige modellen op te stellen. In de drie behandelde gevallen hebben we als model een discreet kansmodel gekozen. We verstaan daaronder een model waarbij de uitkomstenruimte U bestaat uit een verzameling van discrete, dat wil zeggen los van elkaar staande uitkomsten. Hier volgt de formele definitie van een discreet kansmodel.

kansexperiment

discreet kansmodel

Definitie 1.1 *Een discreet kansmodel bestaat uit een discrete uitkomstenruimte U en een kansfunctie P die aan elke deelverzameling van U een kans toekent zo, dat*
1. $P(A) \geq 0$ voor elke $A \subset U$,
2. $P(A \cup B) = P(A) + P(B)$ als $A \cap B = \phi$,
3. $P(U) = 1$.
(Met ϕ wordt de lege verzameling bedoeld.) De deelverzamelingen van U noemt men in dit verband ook gebeurtenissen.

De drie voorwaarden uit de definitie komen overeen met voor de hand liggende eisen: kansen zijn niet-negatief, kansen zijn additief bij disjuncte gebeurtenissen en de kans op de gehele uitkomstenruimte is 1.

Opgave 1.1 *U heeft een munt in handen waarvan u aan mag nemen dat de 'kansen op kruis en munt' even groot zijn. U moet met behulp van die munt op een eerlijke wijze een taart verloten onder tien mensen. Hoe doet u dat? U mag meerdere malen werpen.*

Opgave 1.2 *In deze opgave werken we met het model van de zuivere dobbelsteen.*
a. Hoe groot is de kans om géén 6 te werpen?
b. Hoe groot is de kans om in twee achtereenvolgende worpen twee maal een 6 te werpen?
c. En in twee opvolgende worpen geen enkele zes?
d. Hoe groot is de kans om met twee dobbelstenen in totaal 8 te werpen?

Opgave 1.3 *Stel dat u in een café het aanbod krijgt dat u een inzet van een tientje kunt verdubbelen als u met een dobbelsteen in vier worpen minstens een keer 6 gooit; zo niet, dan bent u uw tientje kwijt. Is het verstandig om op zo'n aanbod in te gaan?*

Opgave 1.4 *Beschrijf hoe u met een 'verdachte munt', dat wil zeggen een munt waarvan u vermoedt dat de 'kansen op kruis en munt' niet gelijk zijn, toch eerlijk kunt tossen. U mag daarbij meerdere malen met de munt werpen.*

1.3 Rekenen met gebeurtenissen

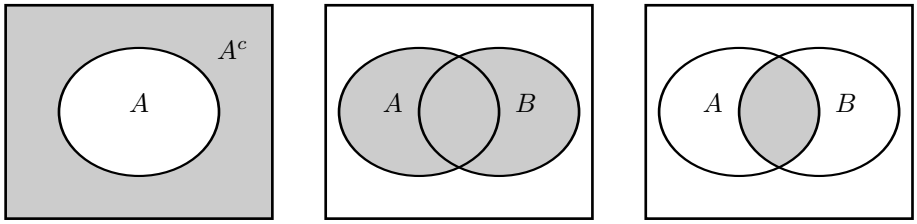
We hebben een gebeurtenis gedefinieerd als een deelverzameling van de uitkomstenruimte U van een kansexperiment. We kunnen met gebeurtenissen dus ook rekenen zoals we dat met verzamelingen doen. Daartoe definiëren we drie bewerkingen op zulke deelverzamelingen van U:

1. *Het complement nemen*: $A^c = U - A$.
   Dit kan men interpreteren als de gebeurtenis dat A *niet* optreedt.

2. *De vereniging nemen*: $A \cup B$.
   Dit kan men interpreteren als de gebeurtenis dat A of B (of beide) optreden.

3. *De doorsnede nemen*: $A \cap B$.
   Dit kan men interpreteren als de gebeurtenis dat A en B tegelijkertijd optreden.

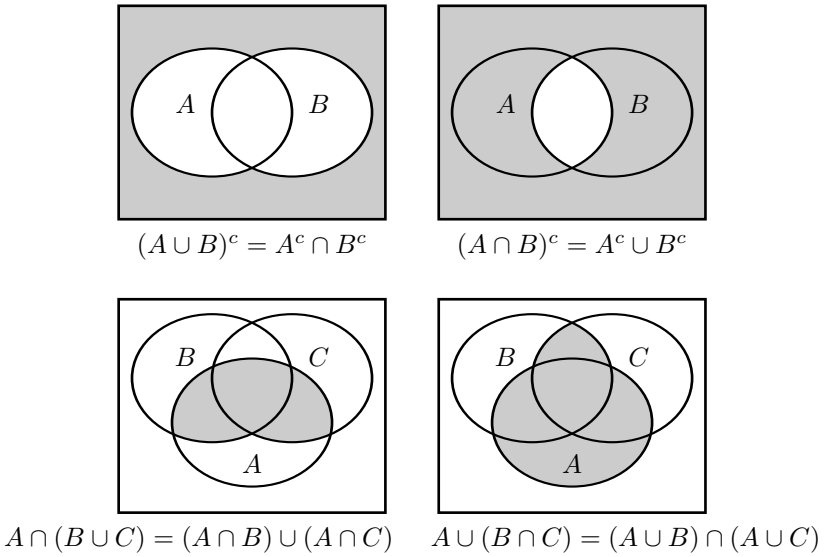
Men kan deze drie bewerkingen visualiseren met behulp van zogenaamde *Venndiagrammen*, symbolische tekeningen waarin men de verzamelingen voorstelt als deelverzamelingen van een rechthoek in het vlak die U representeert (zie Figuur 1.1).



Figuur 1.1: Venn-diagrammen voor het complement, de vereniging en de doorsnede.

Voorbeeld 1.3 *Het gooien met een dobbelsteen.*

Bij het gooien met een dobbelsteen neemt men $U = \{1, 2, 3, 4, 5, 6\}$. Stel nu $A = \{1, 3, 5\}$, dat wil zeggen dat A de gebeurtenis is dat er een oneven aantal ogen wordt gegooid, en stel dat $B = \{1, 6\}$, dat wil zeggen dat B de gebeurtenis is dat er een 1 of een 6 wordt gegooid. Dan is A^c de gebeurtenis dat er een even aantal ogen wordt gegooid, $A \cup B = \{1, 3, 5, 6\}$ en $A \cap B = \{1\}$.



Figuur 1.2: Venn-diagrammen voor rekenregels met gebeurtenissen.

Voor het rekenen met verzamelingen gelden de volgende regels:

$$\begin{aligned}
(A^c)^c &= A \\
A \cup A^c &= U \\
A \cap A^c &= \phi \qquad \text{(de lege verzameling)} \\
(A \cup B)^c &= A^c \cap B^c \\
(A \cap B)^c &= A^c \cup B^c \\
A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \\
A \cup (B \cap C) &= (A \cup B) \cap (A \cup C)
\end{aligned}$$

In Figuur 1.2 ziet u illustraties van de laatste vier relaties met behulp van Venn-diagrammen.
Men noemt twee gebeurtenissen A en B *complementair* als $A = B^c$ (en dan is dus ook $B = A^c$). Wanneer A en B een lege doorsnede hebben, spreekt men over *disjuncte gebeurtenissen*, of ook wel over *elkaar wederzijds uitsluitende gebeurtenissen*.

complementaire gebeurtenissen
disjuncte gebeurtenissen

1.4 Onafhankelijkheid en voorwaardelijke kansen

Stel weer dat een discreet kansmodel gegeven is, dat wil zeggen een uitkomstenruimte U en een kansfunctie P op de deelverzamelingen van U. Volgens definitie 1.2 voldoet die kansfunctie aan de eisen

1. $P(A) \geq 0$ voor elke $A \subset U$,
2. $P(A \cup B) = P(A) + P(B)$ als $A \cap B = \phi$,
3. $P(U) = 1$.

Door in eis (2) $A = B = \phi$ te nemen, ziet men dat

$$P(\phi) = 0$$

moet zijn. En door vervolgens in (2) $B = A^c$ te nemen en (3) te gebruiken, ziet men dat $1 = P(U) = P(A \cup A^c) = P(A) + P(A^c)$, met andere woorden:

$$P(A^c) = 1 - P(A).$$

Dit is de zogenaamde complementregel voor kansen.

complementregel

Het is verder niet moeilijk om aan te tonen dat in het algemeen voor willekeurige (niet noodzakelijkerwijze disjuncte) gebeurtenissen A en B geldt dat

$$P(A \cup B) + P(A \cap B) = P(A) + P(B)$$

(zie ook de Venn-diagrammen in Figuur 1.1). Dit staat bekend als de algemene somregel.

algemene somregel

Een belangrijk begrip in de kansrekening is het begrip onafhankelijkheid.

Definitie 1.2 *Men noemt twee gebeurtenissen (stochastisch) onafhankelijk wanneer*

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

onafhankelijke gebeurtenissen

Men kan onafhankelijkheid van gebeurtenissen op de volgende wijze interpreteren: de kans op het optreden van gebeurtenis A wordt niet beïnvloed door het gelijktijdig optreden van gebeurtenis B.

Voorbeeld 1.4 *Kwaliteitscontrole van schroeven.*

Bij de massaproductie van schroeven worden de schroefkop en de schroefdraad beide aan een kwaliteitsonderzoek onderworpen. Stel dat de kans

op een defecte schroefkop gelijk is aan 0.003 en dat de kans op een defecte schroefdraad gelijk is aan 0.005. Dat betekent intuïtief gezien dat men verwacht dat ongeveer 0.3 procent van de schroeven een defecte kop, en ongeveer 0.5 procent een defecte schroefdraad heeft. Zijn die kansen onafhankelijk, dan verwacht men dat het voor het defect zijn van de kop niet uitmaakt of de schroefdraad al dan niet defect is. In definitie 1.2 vinden we een precieze uitwerking van deze intuïtieve gedachtengang. Als er onafhankelijkheid is, is de kans op een schoef met zowel een defecte kop als een defecte schroefdraad gelijk aan $0.003 \times 0.005 = 0.000015$.

Nauw gerelateerd aan het begrip onafhankelijkheid is het begrip voorwaardelijke kans. Vaak wordt dat in de volgende context gebruikt. Stel dat men bij een kansexperiment slechts de situaties bekijkt waarin een gebeurtenis A optreedt. Men wil nu de kans weten dat onder deze voorwaarde tevens de gebeurtenis B optreedt. Dit noemt men de *voorwaardelijke kans* op B onder voorwaarde A.
Een simpel voorbeeld maakt duidelijk wat we bedoelen. Wat is bij het werpen met een zuivere dobbelsteen de kans op een even aantal ogen, als men slechts die worpen bekijkt waarbij het ogenaantal 1, 2 of 3 is? Hier is dus $A = \{1, 2, 3\}$ en $B = \{2, 4, 6\}$. Er is binnen A slechts één even uitslag, dus de voorwaardelijke kans zal $\frac{1}{3}$ zijn.
Hier is de algemene definitie:

voorwaardelijke kans

Definitie 1.3 *Als $P(A) > 0$ is, verstaat men onder de kans op B onder voorwaarde A het quotient*

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

Wanneer A en B onafhankelijke gebeurtenissen zijn, is $P(A \cap B) = P(A) \cdot P(B)$, en dan is $P(B \mid A) = P(B)$. In dat geval is de voorwaardelijke kans op B dus gelijk aan de gewone kans op B, in overeenstemming met onze intuïtie.

Opgave 1.5 *Waar of niet waar: disjuncte gebeurtenissen zijn onafhankelijk.*

Opgave 1.6 *Gegeven is: $P(A) = 0.15$, $P(B) = 0.33$, $P(A \cup B) = 0.45$. Zijn A en B disjunct? Zijn ze onafhankelijk? Indien niet, bereken dan de voorwaardelijke kansen $P(B \mid A)$ en $P(A \mid B)$.*

Opgave 1.7 *In voorbeeld 1.4 blijkt de kans op een defecte schroef (dat wil zeggen een defecte kop of een defecte schroefdraad) gelijk te zijn aan 0.795 procent. Bereken de kans dat een schroef met een defecte schroefdraad tevens een defecte kop heeft.*

Opgave 1.8 *Waar of niet waar: de voorwaardelijke kans $P(B \mid A)$ is altijd kleiner dan of gelijk aan de onvoorwaardelijke kans $P(B)$.*

Opgave 1.9 *Stel dat $P(B \mid A) = P(B)$. Zijn A en B dan noodzakelijkerwijze onafhankelijk?*

1.5 De regel van Bayes

Uit de definitie van voorwaardelijke kansen volgt onmiddellijk:

Stelling 1.1 *Als $P(A) > 0$ en $P(B) > 0$ dan is*

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

Met deze stelling kan men een voorwaardelijke kans $P(A \mid B)$ berekenen zodra men de 'omgekeerde' voorwaardelijke kans $P(B \mid A)$ en de kansen $P(A)$ en $P(B)$ kent. We beginnen met een fictief voorbeeld.

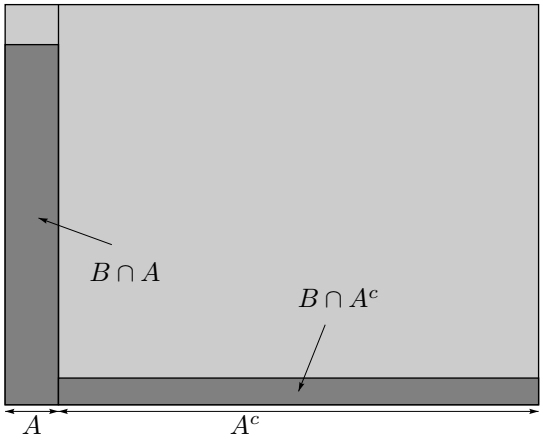
Voorbeeld 1.5 *De VIH-test*

Stel dat bekend is dat 0.1 procent van de bevolking besmet is met het VIH-virus dat de ernstige ziekte SDIA kan veroorzaken. Er is een test om uit te maken of iemand met het VIH-virus besmet is. Die test is niet helemaal betrouwbaar: in 2 procent van de gevallen waarin de betrokkene besmet is, geeft de test de uitslag 'niet besmet', en in 1 procent van de gevallen waarin de betrokkene niet besmet is, geeft de test de uitslag 'besmet'.

Persoon X laat zich testen. De uitslag van de test is positief. Hoe groot is de kans dat X werkelijk besmet is met het VIH-virus?

Voordat we ons over deze vraag buigen, is het goed om op te merken dat de vraagstelling eigenlijk niet met kansrekening kan worden opgelost, Voor persoon X is er geen sprake van een kansexperiment: hij is besmet of niet. Toch zal X, al is het alleen maar om emotionele redenen, wel degelijk in een kanstheoretisch antwoord geïnteresseerd zijn. Om er toch een kansexperiment van te maken, nemen we aan dat zich vele malen een dergelijke situatie voordoet: dat er vele malen 'at random' een persoon Y uit de populatie getrokken wordt, en dat die persoon daarna onderworpen wordt aan de VIH-test.

Laat A de gebeurtenis zijn dat Y besmet is. We weten dat $P(A) = 0.001$. Laat B de gebeurtenis zijn dat de test positief uitvalt. We vragen naar de voorwaardelijke kans $P(A \mid B)$.



Figuur 1.3: Schematische weergave (niet op schaal) bij Voorbeeld 1.5.

We kennen de 'omgekeerde' voorwaardelijke kansen: $P(B \mid A^c) = 0.01$ en $P(B \mid A) = 1 - P(B^c \mid A) = 0.98$ en dus is

$$\begin{aligned} P(B) &= P(B \cap A^c) + P(B \cap A) \\ &= P(B \mid A^c) \times P(A^c) + P(B \mid A) \times P(A) \\ &= 0.01 \times 0.999 + 0.98 \times 0.001 = 0.01097. \end{aligned}$$

Volgens Stelling 1.1 is dan

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} = \frac{0.00098}{0.01097} = 0.0893345.$$

De kans dat persoon X ook werkelijk besmet is, ondanks de positief uitgevallen VIH-test, dus toch nogal klein: ze bedraagt minder dan 9 procent!

We hebben in dit voorbeeld gezien dat we de kans $P(B)$ in de noemer niet direct kenden, maar moesten uitrekenen met behulp van de voorwaardelijke kansen $P(B \mid A)$ en $P(B \mid A^c)$ en de kansen $P(A)$ en $P(A^c)$, die wél bekend waren. Dat is een situatie die zich meestal voordoet. Iets algemener nog, stel dat de uitkomstenruimte U is opgedeeld in n elkaar wederzijds uitsluitende ('onderling disjuncte') gebeurtenissen A_1, A_2, ..., A_n. In formule:

$$\bigcup_{i=1}^{n} A_i = U \quad \text{en} \quad A_i \cap A_j = \phi \quad \text{als } i \neq j.$$

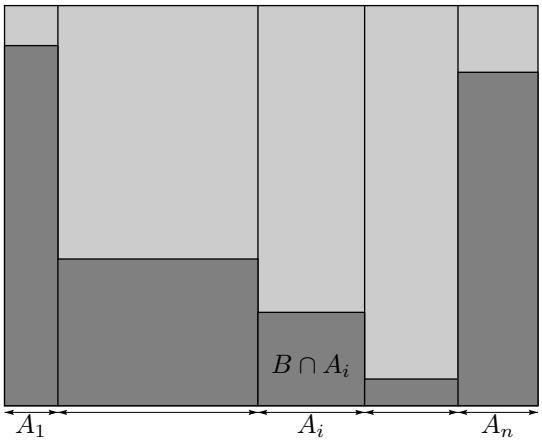
Dan geldt

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

Substitueren we dit in Stelling 1.1, dan krijgen we een beroemde stelling, die genoemd is naar de Engelse dominee Thomas Bayes (1702-1761):

Stelling 1.2 (Regel van Bayes) *Als A_1, A_2, ..., A_n onderling disjuncte gebeurtenissen zijn met positieve kansen, en hun vereniging is de gehele uitkomstenruimte U, dan geldt voor iedere k dat*

$$P(A_k \mid B) = \frac{P(B \mid A_k) \cdot P(A_k)}{\sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i)}$$



Figuur 1.4: Bij de Regel van Bayes. B is het donkere gebied.

We wijzen nog op een theoretisch aanvechtbare, maar veel verbreide terminologie die in dit verband vaak gebruikt wordt. Men noemt de kansen $P(A_k)$ dan de *a priori* kansen, en de voorwaardelijke kansen $P(A_k \mid B)$ de *a posteriori* kansen, en interpreteert ze dan als kansen vóór respectievelijk ná het optreden van gebeurtenis B. In termen van het voorbeeld: voor het uitvoeren van de VIH-test heeft persoon X een (a priori) kans $P(A)$ op besmetting, en na de positieve testuitslag een (a posteriori) kans $P(A \mid B)$. Het is duidelijk dat dit een aanvechtbare wijze van spreken is, wanneer men ze op één individu toepast. Ze is slechts verantwoord bij

een kansexperiment dat in principe willekeurig vaak onder dezelfde omstandigheden kan worden uitgevoerd. Maar natuurlijk, in veel praktijkgevallen heeft men niet de mogelijkheid om veel experimenten uit te voeren ...
Er is nog een ander probleem met de toepassing van de Regel van Bayes: in veel gevallen zijn de kansen $P(A_i)$ niet goed gedefinieerd of heel moeilijk te schatten. Kijk maar weer naar het gegeven voorbeeld. Men kan met zekere methoden best een redelijke schatting maken voor het percentage van de bevolking dat besmet is met het VIH-virus. Maar dat virus zal niet *at random* over de bevolking verspreid zijn, bijvoorbeeld omdat het alleen op bepaalde manieren overgedragen kan worden: er zijn groepen met een hoog risico en groepen met een besmettingsrisico dat vrijwel nul is. Bij een verantwoorde kansanalyse wil men hier ook rekening mee houden: beschouwt men echter zekere deelpopulaties, dan zal men de kansen binnen die populaties moeten schatten, en dat is vaak helemaal niet eenvoudig.

Opgave 1.10 *Stel dat in de situatie van Voorbeeld 1.5 de test verbeterd wordt, zodat de kans op een 'vals alarm' van* 1 *procent teruggaat naar* 0.5 *procent. Wat is hiervan de invloed op de voorwaardelijke kans* $P(A \mid B)$*?*

Opgave 1.11 *En wat is het effect op diezelfde voorwaardelijke kans als het percentage besmetten in de gehele bevolking oploopt tot* 0.3 *procent?*

Opgave 1.12 *Vier schutters* a*,* b*,* c *en* d *hebben verschillende graden van geoefendheid: de kans dat* a *een doel raakt is* 10 *procent, bij* b*,* c*, en* d *zijn die percentages respectievelijk* 20*,* 15 *en* 5*. Eén van de vier schutters (we weten niet wie) vuurt op een doel, en treft het. Wat is de kans dat dit schutter* a *was? En wat zijn de kansen voor de andere schutters? Geef ook commentaar op de vraagstelling.*

Opgave 1.13 *Nu komen de schutters uit een populatie* P *die verdeeld is in vier deelpopulaties* A*,* B*,* C *en* D*. De leden van groep* A *zijn allemaal even bedreven: de kans dat een schutter* a *uit* A *het doel treft, is* 10 *procent. Evenzo heeft elk lid* b *uit groep* B *een trefkans van* 20 *procent, en bij* C *en* D *zijn die kansen respectievelijk* 15 *en* 5 *procent. De groepen* A*,* B*,* C *en* D *zijn niet even groot. Ze verhouden zich in grootte als* $A : B : C : D = 2 : 1 : 3 : 2$*.*

1. *Er wordt* at random *een schutter uit de totale populatie gekozen, die één schot afvuurt. Het is raak. Wat zijn de respectievelijke voorwaardelijke kansen dat de schutter uit groep* A*,* B*,* C *of* D *afkomstig is?*

2. *Men kiest opnieuw* at random *een schutter, en laat die drie schoten afvuren. Geen ervan is raak. Wat zijn nu die kansen?*

Hoofdstuk 2

Continue kansmodellen

We keren terug naar de 'ideale' randomgenerator die op afroep een 'willekeurig' getal uit het interval $[0, 1]$ produceert, waarbij alle getallen uit dit interval even waarschijnlijk zijn. Maar met die laatste uitspraak stuiten we al direct op een fundamenteel probleem. Het is duidelijk dat de uitkomstenruimte U in dit model bestaat uit het gehele interval $[0, 1]$, en dat de uitkomsten, de elementaire gebeurtenissen in dit model, de afzonderlijke getallen uit dit interval moeten zijn. Maar wat zou de kans moeten zijn op zo'n getal? Enig nadenken leert dat die kans alleen maar nul kan zijn. Het interval bevat immers oneindig veel getallen, en als alle uitkomsten 'even waarschijnlijk' zijn, zou een positieve kans op zo'n getal resulteren in kans oneindig voor het gehele interval, hetgeen absurd is, want een kans is nooit groter dan 1.

Hetzelfde probleem doet zich voor bij ieder continu kansmodel, dat wil zeggen bij elk kansmodel waarbij de uitkomstenruimte een interval of zelfs de gehele $\mathbb{R}$ beslaat. *continu kansmodel*

We geven nog twee voorbeelden waarbij het gebruik van continue kansmodellen voor de hand ligt.

Voorbeeld 2.1 *De brandtijd van een gloeilamp*

We keren terug naar de massaproductie van gloeilampen, en nemen nu aan dat alle defecte exemplaren verwijderd zijn. De brandtijd totdat zo'n goede lamp stuk gaat, zal van lamp tot lamp verschillen. De ervaring leert echter dat de brandtijden een zekere concentratie rond een 'gemiddelde' waarde vertonen; bij een bepaald productieproces is die gemiddelde waarde bijvoorbeeld 100 branduren.

Bij het opstellen van een wiskundig model kan men het bepalen van de brandtijd van een lamp opvatten als een kansexperiment. Het ligt dan voor de hand een continu kansmodel te gebruiken, waarbij de uitkomstenruimte U een geheel interval beslaat, bijvoorbeeld het interval $[60, 140]$. Bepaalt men van een groot aantal lampen de brandtijd, dan zal blijken dat die uitkomsten niet gelijkmatig verdeeld liggen binnen dit interval, maar dat ze een grote concentratie rond de 100 vertonen. In het wiskundige model zal dit ook tot uitdrukking gebracht moeten worden. We komen hier later nog op terug.

Voorbeeld 2.2 *Wachttijden in het postkantoor*

Aan de loketten van een postkantoor verschijnen met onregelmatige tussenpozen klanten. Ook de bedieningstijd per klant is onregelmatig. Als men hiervoor wiskundige modellen op wil stellen, bijvoorbeeld omdat men iets te weten wil komen over gemiddelde wachttijden of de kans op

grote drukte door lange wachtrijen, ligt het voor de hand te gaan werken met continue kansmodellen.

In de zojuist genoemde voorbeelden lag het voor de hand om als uitkomstenruimte U telkens een deelinterval van $\mathbb{R}$ te nemen. Maar het definiëren van een kansfunctie daarop geeft problemen: de kans op een individuele uitkomst uit U kan niet anders dan 0 zijn. De definitie van een kans op een gebeurtenis A als de som van de kansen op de elementaire gebeurtenissen waaruit A is samengesteld, zoals we dat bij discrete kansmodellen deden, is hier dus onmogelijk. Een uitweg uit deze impasse vinden we als we ons indenken hoe we als het ware een geleidelijke overgang kunnen maken van een discreet kansmodel naar een continu kansmodel.

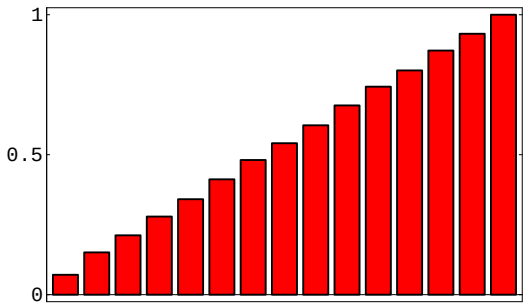
2.1 Van discrete naar continue kansmodellen

We keren terug naar Tabel 1.2, die de resultaten liet zien van een computersimulatie van duizend worpen met een dobbelsteen met behulp van een randomgenerator. We kunnen dat experiment op twee manieren verfijnen. In de eerste plaats kunnen we het aantal worpen groter nemen. We verwachten dan dat de relatieve frequenties van de zes uitkomsten op den duur steeds beter de waarde 1/6 gaan benaderen. Maar we kunnen ook het aantal worpen op duizend houden, en de onderverdeling van het eenheidsinterval verfijnen, om zo een 'dobbelsteen' te simuleren met meer dan 6 zijkanten. In Tabel 2.1 ziet u het resultaat van een onderverdeling in 15 deelintervallen van gelijke lengte.
We hebben in Tabel 2.1 een extra kolom toegevoegd, die van de cumulatieve relatieve frequenties. Daarin worden de relatieve frequenties uit de kolom ernaast van bovenaf bij elkaar opgeteld. Bij dobbelstenen lijkt dat niet erg zinvol, maar als we toe willen naar een soort 'continue kansverdeling' op het interval $[0, 1]$ is dat anders. Ook de laatste kolom heeft hier namelijk een heel duidelijke interpretatie: op de k-de rij staat de relatieve frequentie van de uitkomsten tussen 0 en $k/15$. In de grafiek onder de tabel ziet u van zo'n kolom van cumulatieve relatieve frequenties een staafdiagram.
Wat gebeurt er met de drie laatste kolommen van Tabel 2.1 als we het aantal deelintervallen verder opvoeren, en daarbij steeds dezelfde collectie van duizend randomwaarden hanteren? De kolom met aantallen zal dan op den duur voornamelijk nullen bevatten. Als het aantal deelintervallen n maar groot genoeg is, zal zo'n deelinterval hoogstens één van de duizend randomwaarden bevatten (want bij de ideale randomgenerator is de kans op twee gelijke waarden in een serie van duizend nul), maar in de meeste deelintervallen zal geen enkel random getal voorkomen. Ook de derde kolom zal voornamelijk uit nullen bestaan, en is het geen 0, dan zal er 0.001 staan.
De laatste kolom van Tabel 2.1 behoudt echter zijn structuur: een rij getallen die geleidelijk van 0 naar 1 stijgt. Die structuur wordt nog duidelijker als we het bij die kolom behorende staafdiagram opvatten als de grafiek van een 'trapfunctie'. Op den duur, als er nog hoogstens één randomwaarde per interval optreedt, stijgen alle traptreden met een bedrag van precies $\frac{1}{1000}$, en de x-waarden waarbij de treden omhoog gaan, zullen steeds meer naderen tot de gekozen randomwaarden. Als R de verzameling is die bestaat uit de duizend randomwaarden, dan nadert de trapfunctie steeds dichter tot de trapfunctie $F_R(x)$ met treden van hoogte $\frac{1}{1000}$ die precies bij de duizend gekozen randomwaarden verspringen. We noemen deze functie de *cumulatieve frequentiefunctie* bij R. In Figuur 2.1 ziet u de grafiek van $F_R(x)$ bij zo'n keuze van duizend randomwaarden. Door het grote aantal zijn de afzonderlijke trapjes nauwelijks te onderscheiden. Overigens, het computeralgebrapakket dat deze en de meeste volgende grafieken getekend heeft, trekt verticale verbindingslijntjes

cumulatieve frequentiefunctie

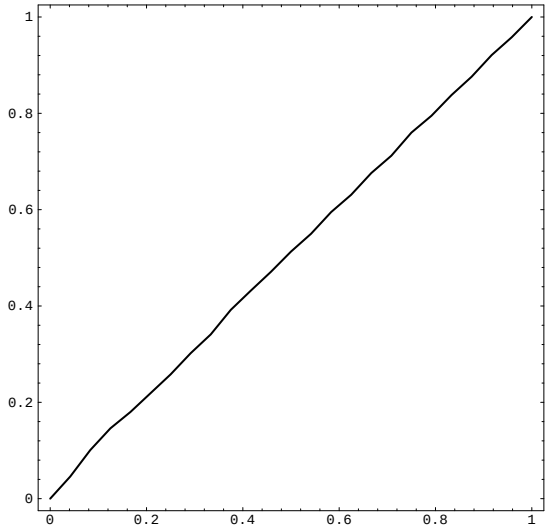
Tabel 2.1: Duizend randomgetallen, verdeeld in 15 klassen

deelinterval	aantal	rel. freq.	cum. rel. freq.
$[0,\ 1/15\rangle$	64	0.064	0.064
$[1/15,\ 2/15\rangle$	66	0.066	0.130
$[2/15,\ 3/15\rangle$	62	0.062	0.192
$[3/15,\ 4/15\rangle$	76	0.076	0.268
$[4/15,\ 5/15\rangle$	50	0.050	0.318
$[5/15,\ 6/15\rangle$	78	0.078	0.396
$[6/15,\ 7/15\rangle$	69	0.069	0.465
$[7/15,\ 8/15\rangle$	64	0.064	0.529
$[8/15,\ 9/15\rangle$	67	0.067	0.596
$[9/15,\ 10/15\rangle$	75	0.075	0.671
$[10/15,\ 11/15\rangle$	72	0.072	0.743
$[11/15,\ 12/15\rangle$	66	0.066	0.809
$[12/15,\ 13/15\rangle$	68	0.068	0.877
$[13/15,\ 14/15\rangle$	64	0.064	0.941
$[14/15,\ 1]$	59	0.059	1.000



Staafdiagram van de cumulatieve relatieve frequenties

waar een grafiek sprongdiscontinuïteiten heeft. Bij een trapfunctie worden dus ook de verticale lijntjes van de afzonderlijke trapjes getekend.
We kunnen nu ook het aantal randomwaarden, dat we tot nu toe op duizend gefixeerd hadden, steeds groter laten worden. Bij discrete modellen hebben we gezien dat de relatieve frequenties dan steeds meer naar kansen gaan tenderen (dit was de zogenaamde experimentele wet van de grote aantallen). In het continue geval hebben we nog geen kansen gedefinieerd, maar het wordt nu wel duidelijk wat we daarvoor moeten doen. Bij een steeds groter wordend aantal randomwaarden zal de cumulatieve frequentiefunctie $F_R(x)$ op het interval $[0, 1]$ steeds meer gaan lijken op de functie $F(x) = x$. Dat betekent dat de cumulatieve relatieve frequentie van het aantal randomwaarden in het interval $[0, x]$ op den duur de limietwaarde $F(x) = x$ krijgt, en dat zal dus ook de kans moeten zijn die we aan het interval $[0, x]$ toekennen. Diezelfde kans kennen we natuurlijk toe aan de intervallen $[0, x\rangle$, $\langle 0, x]$ en $\langle 0, x\rangle$, want losse punten hebben kans 0. Zo krijgen we bijvoorbeeld voor $x = 0.7$ een kans van 0.7 voor elk van de intervallen $[0, 0.7]$, $[0, 0.7\rangle$, $\langle 0, 0.7]$ en $\langle 0, 0.7\rangle$.
Maar nu kunnen we ook kansen definiëren voor willekeurige deelintervallen van $[0, 1]$. Zo zal de kans op het interval $[0.3, 0.7]$ natuurlijk $0.7 - 0.3 = 0.4$ moeten zijn, want



Figuur 2.1: Cumulatieve frequentiefunctie bij duizend randomwaarden

$[0, 0.3\rangle \cup [0.3, 0.7] = [0, 0.7]$, en dus moet $P([0, 0.3\rangle) + P([0.3, 0.7]) = P([0, 0.7])$ zijn. Maar de kansen $P([0, 0.3\rangle) = 0.3$ en $P([0, 0.7]) = 0.7$ hadden we al gedefinieerd, en dus moeten we aan $P([0.3, 0.7])$ inderdaad de waarde 0.4 toekennen.
In het algemeen definiëren we voor een willekeurig deelinterval $[a, b]$ van $[0, 1]$ de kans $P([a, b])$ door

$$P([a, b]) = F(b) - F(a).$$

cumulatieve verdelingsfunctie

In dit geval is $F(x) = x$, en dus is hier $P([a, b]) = b - a$. Diezelfde kans kennen we natuurlijk toe aan de open en halfopen intervallen met dezelfde eindpunten. De functie $F(x)$ noemen we in dit verband de *cumulatieve verdelingsfunctie*. De kans op een interval is dus gedefinieerd als de toename van de cumulatieve verdelingsfunctie over dat interval.
In Figuur 2.2 ziet u een grafiek van deze cumulatieve verdelingsfunctie $F(x) = x$. U ziet ook dat deze grafiek als het ware een 'gladgestreken' versie is van de grafiek van Figuur 2.1, de cumulatieve frequentiefunctie bij duizend randomwaarden.
Het continue kansmodel dat we hiermee hebben beschreven, heet de *uniforme verdeling* op het interval $[0, 1]$. De 'ideale randomgenerator' heeft deze verdeling als continu kansmodel.

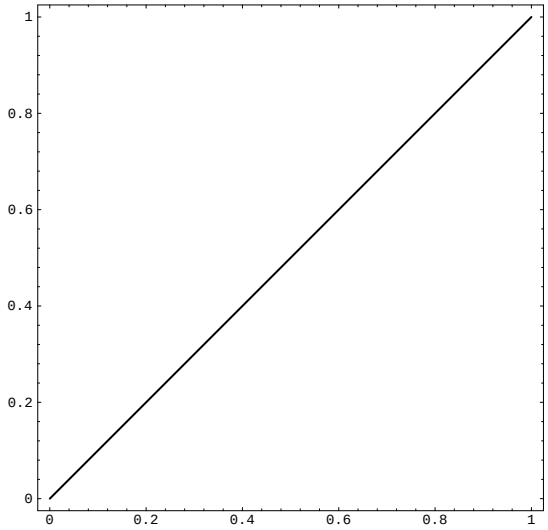
uniforme verdeling

Opgave 2.1 *Op een soortgelijke wijze kan men de uniforme verdeling op een ander interval definiëren, bijvoorbeeld op het interval* $[1, 2]$*. Denk maar aan een randomgenerator die willekeurige getallen tussen 1 en 2 produceert, waarbij elk getal uit dit interval 'even waarschijnlijk' is.*

a. Wat is in dit geval de cumulatieve verdelingsfunctie $F(x)$*?*
b. Wat is de kans op het interval $[\sqrt{2}, \sqrt{3}]$*?*

Opgave 2.2 *In deze opgave beschouwen we de uniforme verdeling op het interval* $[2, 8]$*.*

a. Bepaal de cumulatieve verdelingsfunctie $F(x)$*.*
b. Bepaal $P([3, \pi])$*.*
c. Bepaal in het algemeen een formule voor de kans $P([a, b])$*, waarbij* $[a, b]$ *een willekeurig deelinterval is van* $[2, 8]$*.*



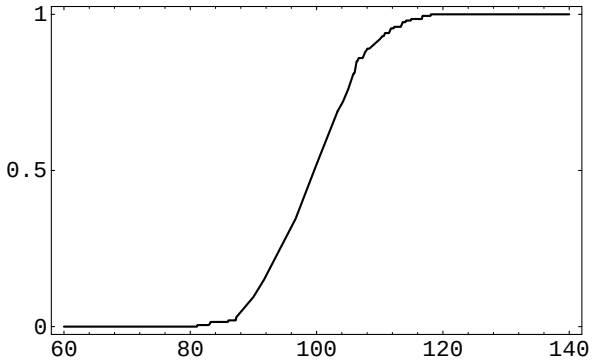
Figuur 2.2: Cumulatieve verdelingsfunctie van de uniforme verdeling op $[0, 1]$

Uit het bovenstaande begint duidelijk te worden hoe we in het algemeen bij een continu kansmodel op een zinvolle wijze kansen kunnen definiëren: niet via kansen op individuele uitkomsten (want die kunnen alleen maar nul zijn), maar via kansen op deelintervallen van de uitkomstenruimte. Alvorens echter een algemene formele definitie te geven van continue kansmodellen en cumulatieve verdelingsfuncties, behandelen we nog twee andere voorbeelden.

Voorbeeld 2.3 *De brandtijd van een gloeilamp*

We keren weer terug naar het kansexperiment van het bepalen van de brandtijd van een gloeilamp. We zouden dat experiment bijvoorbeeld tweehonderd keer uit kunnen voeren, en net als boven hiervoor een cumulatieve frequentiefunctie kunnen tekenen. We zullen dan ook weer een stijgende trapfunctie te zien krijgen: in dit geval een trapfunctie die met stapjes van $\frac{1}{200}$ van 0 naar 1 stijgt. Het interval waarop die stijging plaatsvindt, loopt van de kleinste brandtijd in de serie tot aan de grootste.

In plaats van het uitvoeren van zo'n kostbaar en tijdrovend experiment (waarvan de details ons op dit moment toch niet echt interesseren), voeren we een computersimulatie uit. We gebruiken daarvoor een aangepaste vorm van de randomgenerator, een vorm die zogenaamde 'trekkingen uit een normale verdeling' kan simuleren. In de loop van dit blok zult u leren wat er hier allemaal voor theorie achter zit, maar voor dit moment is het voldoende als u accepteert dat dit een 'redelijke' simulatie oplevert voor zo'n kansexperiment. Bij een simulatie met behulp van de normale verdeling moeten we nog twee parameters kiezen: de verwachting μ, die we voor dit moment kunnen interpreteren als een soort 'gemiddelde brandtijd', en de standaardafwijking σ, die een maat is voor de 'spreiding' van de resultaten. Een grote waarde van σ resulteert bij zo'n simulatie in brandtijden die ver uiteen liggen, en een kleine σ levert brandtijden die zich rond de verwachting μ concentreren. In onze simu-

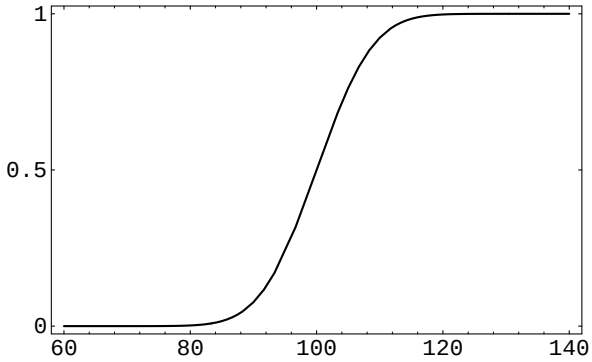


Figuur 2.3: Cumulatieve frequentiefunctie van een simulatie van de brandtijd van tweehonderd lampen

latie hebben we, min of meer willekeurig, $\mu = 100$ en $\sigma = 7$ genomen. Figuur 2.3 toont het resultaat.

In Figuur 2.4 ziet u de cumulatieve verdelingsfunctie $F(x)$ die aan het betreffende continue kansmodel ten grondslag ligt. Het is weer een 'gladgestreken' versie van Figuur 2.3. Ook in dit model geeft een functiewaardenverschil, zoals bijvoorbeeld $F(110) - F(100)$ de kans op een deelinterval weer, in dit geval de kans op het deelinterval $[100, 110]$. Het verschil $F(110) - F(100)$ stelt dus de kans voor dat de brandtijd van een gloeilamp tussen de 100 en 110 uur ligt.

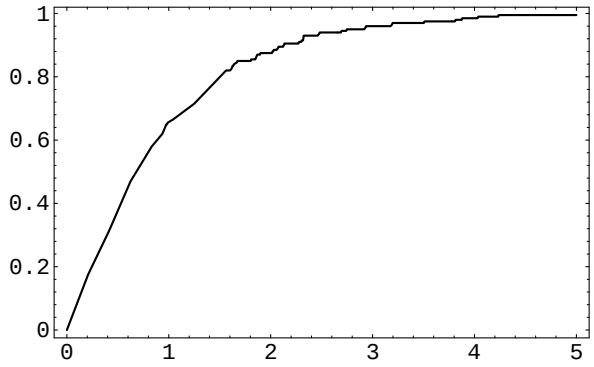
De functie $F(x)$ is de cumulatieve verdelingsfunctie van de zogenaamde normale verdeling met parameters $\mu = 100$ en $\sigma = 7$. Aan het einde van dit hoofdstuk zullen we uitgebreid op de normale verdeling terugkomen.



Figuur 2.4: Cumulatieve verdelingsfunctie van de normale verdeling met parameters $\mu = 100$ en $\sigma = 7$

Voorbeeld 2.4 *Aankomsttijden in een wachtrij*

Dit voorbeeld sluit aan bij Voorbeeld 2.2. Als kansexperiment registreren we de tijd die verloopt tussen de aankomst van twee opvolgende klanten in een postkantoor. Ook hier doen we het niet echt, maar voeren we een computersimulatie uit aan de hand van een continu kansmodel



Figuur 2.5: Cumulatieve frequentiefunctie van een simulatie van tweehonderd tussenaankomsttijden

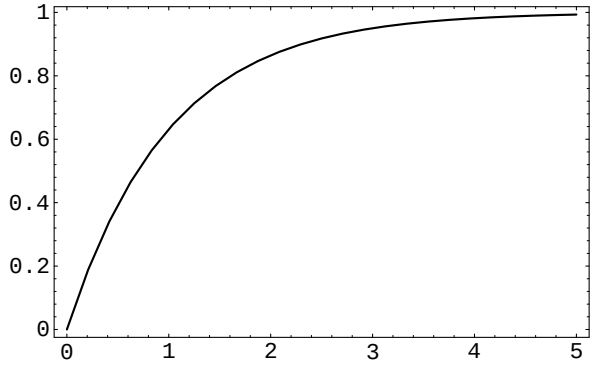
dat bij dit soort situaties vaak gebruikt wordt. Het is de zogenaamde negatief-exponentiële verdeling. Die verdeling, die gedefinieerd is op het interval $[0, \infty\rangle$, heeft één parameter die meestal λ genoemd wordt. Wij hebben $\lambda = 1$ gekozen en met een aangepaste randomgenerator tweehonderd waarden bepaald. In Figuur 2.5 ziet u de bijbehorende cumulatieve frequentiefunctie.

Figuur 2.6 toont de gladgestreken versie, de cumulatieve verdelingsfunctie van de bijbehorende negatief-exponentiële verdeling. De formule van de verdelingsfunctie van de negatief-exponentiële verdeling met parameter λ luidt:

$$F(x) = 1 - e^{-\lambda x} \quad (x \geq 0).$$

negatief-exponentiële verdeling

We zien dat dit een stijgende functie is die bij 0 begint. De limiet voor $x \to \infty$ is 1, maar de waarde 1 wordt nooit bereikt; er is sprake van een horizontale asymptoot. Bij dit alles nemen we aan dat $\lambda > 0$ is; in Figuur 2.6 hebben we $\lambda = 1$ genomen.



Figuur 2.6: Cumulatieve verdelingsfunctie van de negatief-exponentiële verdeling met parameter $\lambda = 1$

U heeft nu een aantal voorbeelden gezien van continue kansmodellen. Die modellen bleken steeds afhankelijk te zijn van bepaalde parameters: bij de uniforme

verdeling het interval $[a, b]$ waarop de verdeling gedefinieerd is, bij de normale verdeling de verwachting μ en de standaardafwijking σ en bij de negatief-exponentiële verdeling de parameter λ. In veel praktijksituaties, met name bij normale en de negatief-exponentiële verdelingen, kent men op theoretische gronden wel de aard van de verdeling, maar niet de precieze waarde van de parameters. Door kansexperimenten uit te voeren probeert men dan vaak verantwoorde schattingen voor die parameterwaarden te krijgen, of hypothesen omtrent de parameterwaarden te toetsen. Wiskundige technieken hiervoor worden ontwikkeld in de *schattingstheorie* en de *toetsingstheorie*. U zult daarmee in Hoofdstuk 6 kennismaken.

2.2 Verdelingsfuncties

We zijn nu in staat een formele definitie te geven van een continu kansmodel. Essentieel daarvoor blijkt de cumulatieve verdelingsfunctie te zijn: een continue functie $F(x)$ die stijgt van 0 naar 1. Kortheidshalve zullen we in het vervolg het bijvoeglijk naamwoord 'cumulatieve' meestal weglaten, en gewoon spreken over een 'verdelingsfunctie'.
Zonder beperking van de algemeenheid kunnen we veronderstellen dat zo'n verdelingsfunctie $F(x)$ op de gehele $\mathbb{R}$ gedefinieerd is, want als de uitkomstenruimte U slechts een deelinterval van $\mathbb{R}$ is, definiëren we eenvoudig $F(x) = 0$ voor alle x links van U, en $F(x) = 1$ voor alle x rechts van U. Met behulp van de verdelingsfunctie definiëren we daarna een kansfunctie op de verzameling van alle deelintervallen van $\mathbb{R}$.

Definitie 2.1 *Een continu kansmodel wordt gedefinieerd door een* continue verdelingsfunctie $F(x)$ *op $\mathbb{R}$ en een met behulp van die verdelingsfunctie gedefinieerde* kansfunctie *op de deelintervallen van $\mathbb{R}$. De verdelingsfunctie $F(x)$ moet voldoen aan de volgende eigenschappen:*

eigenschappen verdelingsfunctie

1. $\lim_{x \to -\infty} F(x) = 0$,
2. $\lim_{x \to +\infty} F(x) = 1$,
3. *Als* $x_1 < x_2$ *dan* $F(x_1) \leq F(x_2)$.

Zo'n verdelingsfunctie $F(x)$ definieert voor elk deelinterval van $\mathbb{R}$ als volgt een kans:

$$P([a,b]) = P([a,b\rangle) = P(\langle a,b]) = P(\langle a,b\rangle) = F(b) - F(a).$$

kans op interval is toename verdelingsfunctie

De kans op een interval is dus de toename van de verdelingsfunctie op dat interval. Ook als zo'n interval onbegrensd is, wordt de kans erop gedefinieerd als de toename van de verdelingsfunctie:
$P(\langle -\infty, b]) = P(\langle -\infty, b\rangle) = F(b) - F(-\infty) = F(b)$,
$P([a, \infty\rangle) = P(\langle a, \infty\rangle) = F(\infty) - F(a) = 1 - F(a)$,
$P(\langle -\infty, \infty\rangle) = F(\infty) - F(-\infty) = 1$.

Voorbeelden van verdelingsfuncties heeft u gezien in Figuur 2.1 (de uniforme verdeling op $[0, 1]$), Figuur 2.3 (de normale verdeling met $\mu = 100$ en $\sigma = 7$) en Figuur 2.5 (de negatief-exponentiële verdeling met $\lambda = 1$).

Opgave 2.3 *Bepaal bij de uniforme verdeling op* $[2, 8]$ *de volgende kansen:* $P(\langle 0, 7\rangle)$, $P(\langle -\infty, 0])$ *en* $P([0, \infty\rangle)$.

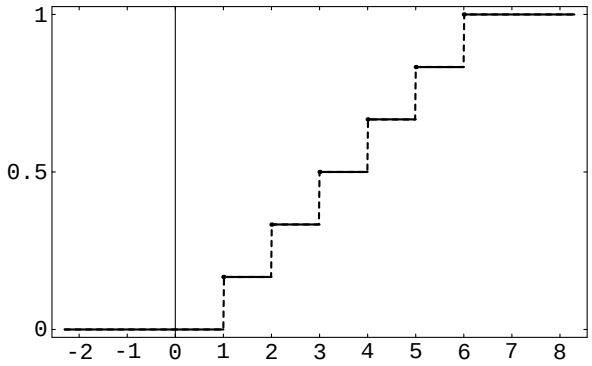
Opgave 2.4 *Bepaal bij de negatief-exponentiële verdeling met* $\lambda = 2$ *de kansen* $P([-1, 1])$ *en* $P([1, \infty\rangle)$.

verdelingsfunctie bij een discreet kansmodel

We kunnen ook bij discrete kansmodellen waarvan de uitkomstenruimte U een (discrete) deelverzameling van $\mathbb{R}$ is, een verdelingsfunctie definiëren. Dat gaat als volgt:

$$F(x) = P(\{u \in U \mid u \leq x\}).$$

In woorden: $F(x)$ is de kans op de gebeurtenis dat er een uitkomst kleiner dan of gelijk aan x optreedt. Men verifieert gemakkelijk dat $F(x)$ dan altijd voldoet aan de voorwaarden (1), (2) en (3) uit Definitie 2.1. Alleen is $F(x)$ nu geen continue functie meer maar een trapfunctie: in elke $u \in U$ treedt een sprong op ter grootte $P(\{u\})$ en tussen de sprongpunten is $F(x)$ constant.



Figuur 2.7: Cumulatieve verdelingsfunctie van het discrete kansmodel van de zuivere dobbelsteen

Als voorbeeld nemen we het model van de zuivere dobbelsteen met uitkomstenruimte $U = \{1, 2, 3, 4, 5, 6\} \subset \mathbb{R}$, waarvoor dus $P(\{u\}) = 1/6$ geldt voor iedere $u \in U$. In Figuur 2.7 ziet u de bijbehorende cumulatieve verdelingsfunctie. Voor deze functie geldt

$$F(x) = \begin{cases} 0 & (x < 1) \\ 1/6 & (1 \leq x < 2) \\ 2/6 & (2 \leq x < 3) \\ 3/6 & (3 \leq x < 4) \\ 4/6 & (4 \leq x < 5) \\ 5/6 & (5 \leq x < 6) \\ 1 & (x \geq 6) \end{cases}$$

Merk op dat we in zulke discrete kansmodellen nu ook voor ieder deelinterval van $\mathbb{R}$ een kans kunnen definiëren. In tegenstelling tot het geval van continue kansmodellen maakt het bij discrete modellen dan wèl uit of de randpunten van het interval meedoen of niet. Zo geldt in het dobbelsteenvoorbeeld natuurlijk dat $P([1, 3]) = 1/2$, want het interval bevat drie uitkomsten die als resultaat van een worp voor kunnen komen, namelijk 1, 2 en 3, maar $P(\langle 1, 3\rangle) = 1/6$, want in dit interval zit alleen de dobbelsteenwaarde 2.
In termen van de verdelingsfunctie is de regel dus dat de kans van een interval I gelijk is aan de som van alle sprongen die de verdelingsfunctie op I maakt. Eventuele sprongen in de randpunten van I tellen daarbij alleen mee als het betreffende randpunt in I bevat is, dat wil zeggen dat I aan die kant gesloten is.

Opgave 2.5 *Bereken in het zojuist geschetste dobbelsteenvoorbeeld de kans van de intervallen* $\langle 0, \infty\rangle$, $\langle 1, \infty\rangle$, $[1, \infty\rangle$ *en* $\langle 1, 2\rangle$.

Opgave 2.6 *In een discreet kansmodel geldt $U = \{1, 2, 3, 4, \ldots\}$ en $P(n) = 2^{-n}$ voor alle $n \in U$. Bepaal de bijbehorende verdelingsfunctie, laat zien dat aan eis (2) uit Definitie 2.1 voldaan is en schets de grafiek van $F(x)$.*

Opgave 2.7 *Laat $P(n)$ de kans zijn dat bij het tossen met een zuivere munt de eerste maal 'kop' optreedt bij de n-de worp. Bepaal $P(n)$. Hoe groot is de kans dat er helemaal geen 'kop' verschijnt?*

Volledigheidshalve merken we nog op dat men op de bovenbeschreven wijze naast discrete en continue kansmodellen ook 'gemengde' modellen zou kunnen definiëren uitgaande van een verdelingsfunctie $F(x)$ die monotoon niet-dalend is, sprongpunten bezit, maar ook intervallen waarop sprake is van een continue stijging. We laten die mogelijkheid in deze cursus terzijde.

2.3 Kansdichtheidsfuncties

We keren terug naar Voorbeeld 2.4, waarin een continu kansmodel gegeven wordt voor de tijd die verloopt tussen de aankomst van twee opvolgende klanten in een postkantoor. De bijbehorende verdelingsfunctie was $F(x) = 1 - e^{-\lambda x}$ voor $x \geq 0$ en $F(x) = 0$ voor $x < 0$. In Figuur 2.6 was de grafiek geschetst bij de parameterkeuze $\lambda = 1$. De grafiek vertoont een gladde stijging: de functie is differentieerbaar (behalve voor $x = 0$). Hoe kunnen we de afgeleide $f(x) = F'(x) = \lambda e^{-\lambda x}$ in termen van ons kansmodel interpreteren? Aangezien

$$F'(x) = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x}$$

ligt het voor de hand om het quotient

$$\frac{F(x + \Delta x) - F(x)}{\Delta x}$$

nader te beschouwen. De teller is de kans van het interval $[x, x + \Delta x]$, de noemer is de lengte ervan. Hun quotient is in eerste benadering gelijk aan de afgeleide $f(x) = F'(x)$, dus

$$P([x, x + \Delta x]) = F(x + \Delta x) - F(x) \approx f(x)\Delta x.$$

We zien dat de kans op een klein interval evenredig is met de lengte ervan, en dat de evenredigheidsfactor in eerste benadering gelijk is aan de afgeleide $f(x)$ van de verdelingsfunctie $F(x)$. $f(x)$ geeft als het ware het quotiënt van de 'hoeveelheid kans' die er op zo'n klein interval aanwezig is, en de lengte van dat interval. Dit verklaart de naam *kansdichtheidsfunctie* die men voor $f(x)$ gebruikt. In het geval van de negatief-exponentiële verdeling geldt voor $x > 0$ dat $F(x) = 1 - e^{-\lambda x}$ en de kansdichtheidsfunctie is dan $f(x) = \lambda e^{-\lambda x}$. In Figuur 2.8 ziet u onder elkaar de grafieken van de kansdichtheidsfunctie $f(x)$ en de verdelingsfunctie $F(x)$ voor een aantal waarden van λ. Merk op dat al die functies nul zijn voor $x < 0$, en dat $f(0)$ niet gedefinieerd is.

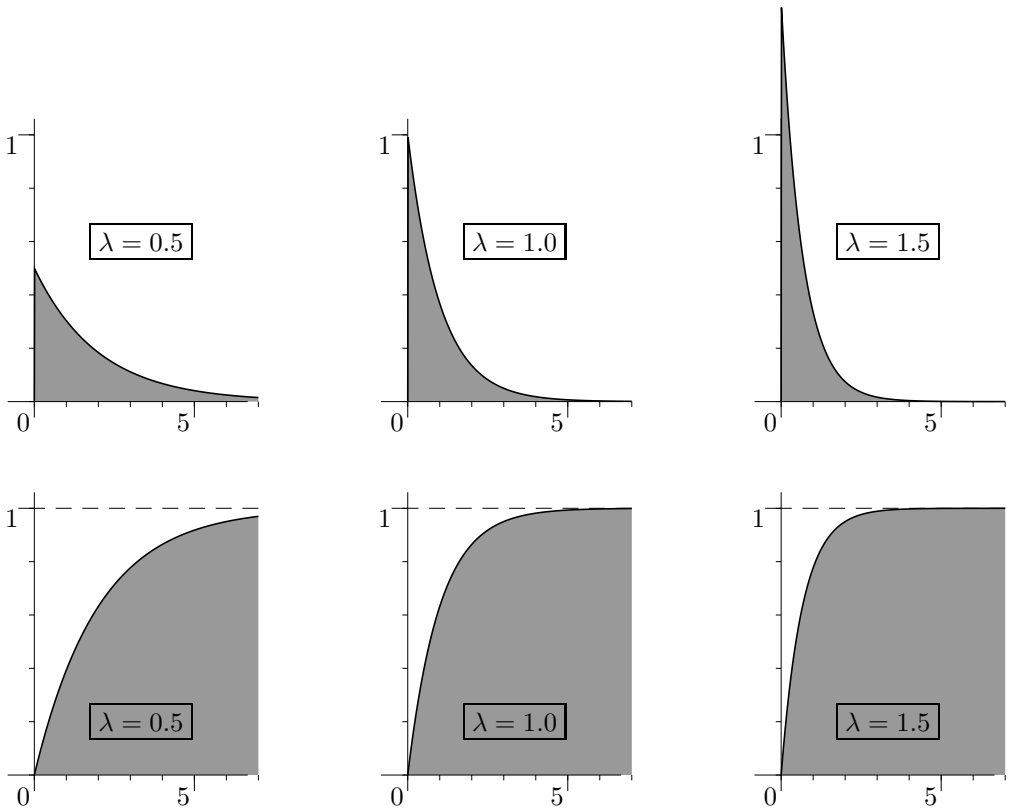
kansdichtheids-functie

Ten aanzien van deze kansdichtheidsfuncties merken we op:

1. $f(x) \geq 0$
2. $f(x)$ is een integreerbare functie op $\mathbb{R}$
3. $\int_{-\infty}^{\infty} f(x)\, dx = 1$

Eigenschap (3) kunt u door berekening verifiëren:

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_{0}^{\infty} \lambda e^{-\lambda x}\, dx = \left[-e^{-\lambda x}\right]_0^{\infty} = 1$$



Figuur 2.8: Kansdichtheidsfuncties $f(x)$ (boven) en verdelingsfuncties $F(x)$ (onder) voor de negatief-exponentiële verdeling bij drie waarden van λ.

maar u kunt natuurlijk ook opmerken dat $f(x)$ de afgeleide is van $F(x)$, dat de integraal over $\mathbb{R}$ van $f(x)$ dus gelijk is aan de toename van $F(x)$ op $\mathbb{R}$, en dat die toename 1 is omdat $F(x)$ een verdelingsfunctie is.
In het algemeen noemt men elke functie $f(x)$ die aan de voorwaarden (1) tot en met (3) voldoet een kansdichtheidsfunctie. Uit (1), (2) en (3) volgt dan dat de functie

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

een verdelingsfunctie is (zie Definitie 2.1), zodat hiermee ook een continu kansmodel vastligt. Voor elk interval interval $[a, b]$ geldt dan

$$P([a, b]) = F(b) - F(a) = \int_{a}^{b} f(t)\, dt.$$

Merk op dat zo'n kansdichtheidsfunctie $f(x)$ niet in alle punten continu behoeft te zijn: slechts de integreerbaarheid van $f(x)$ is van belang. Men kan echter wel bewijzen dat $F(x)$ in elk punt continu is en dat $F'(x) = f(x)$ in elk punt geldt waar $f(x)$ continu is.
Voor de volledigheid formuleren we de definitie van het begrip kansdichtheidsfunctie nog eens expliciet:

Definitie 2.2 *Een functie $f(x)$ van $\mathbb{R}$ naar $\mathbb{R}$ heet een kansdichtheidsfunctie op $\mathbb{R}$ als $f(x)$ voldoet aan de volgende eigenschappen:*

1. $f(x) \geq 0$ *voor alle* x,
2. $f(x)$ *is een integreerbare functie op* $\mathbb{R}$,
3. $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

Opgave 2.8 *Bepaal de kansdichtheidsfunctie bij de uniforme verdeling op het interval* $[a, b]$.

Opgave 2.9 *Ga na dat*

$$f(x) = \frac{1}{\pi(1+x^2)} \qquad (x \in \mathbb{R})$$

een kansdichtheidsfunctie is en bepaal de bijbehorende verdelingsfunctie (het bijbehorende continue kansmodel wordt de Cauchy-verdeling *genoemd).*

2.4 De normale verdeling

De normale verdeling speelt in de toepassingen van de kansrekening en de statistiek een zeer belangrijke rol: veel kansexperimenten blijken met deze verdeling goed gemodelleerd te kunnen worden. Dit is met name het geval bij kansexperimenten waarbij de uitkomsten opgevat kunnen worden als 'toevallige fluctuaties' rond een zekere gemiddelde waarde. In Hoofdstuk 5 zullen we hierop nog terugkomen; hier leggen we slechts de wiskundige basis door de formules en enige elementaire eigenschappen te geven van de verdelingsfunctie en de kansdichtheidsfunctie.
Men kan bewijzen dat

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}.$$

Uit dit resultaat volgt direct:

Stelling 2.1 *De functie*

$$n_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

is voor elke μ *en elke* $\sigma > 0$ *een kansdichtheidsfunctie.*

BEWIJS: Alleen eigenschap (3) van Definitie 2.2 is niet vanzelfsprekend. Noem

$$y = \frac{x-\mu}{\sigma\sqrt{2}}$$

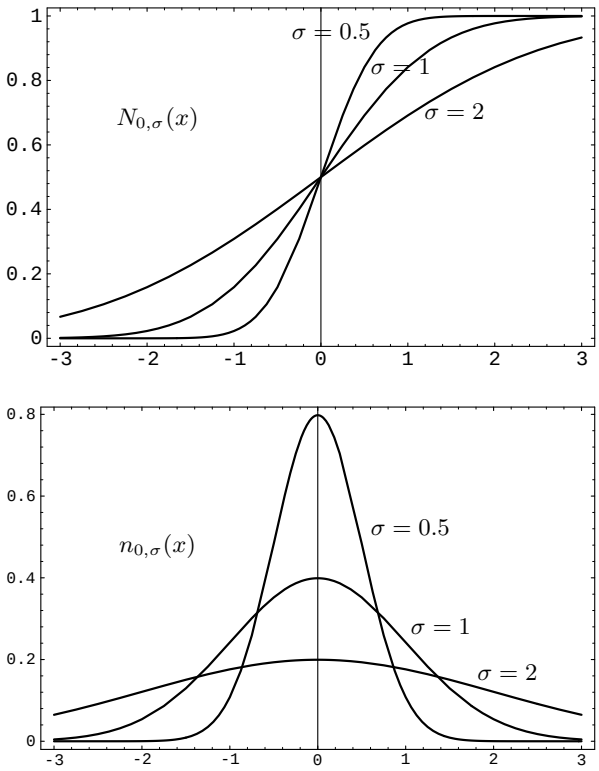
dan geldt

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2} dy = 1.$$

□

Het bijbehorende continue kansmodel heet de *normale verdeling* met parameters μ en σ. In hoofdstuk 4 zult u zien dat μ en σ een duidelijke kanstheoretische betekenis hebben: μ zal de *verwachting* blijken te zijn, en σ de *standaardafwijking*, maar de definitie en de interpretatie van die termen zullen we daar pas behandelen. Hier merken we op dat de bijbehorende verdelingsfunctie

$$N_{\mu,\sigma}(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$



Figuur 2.9: Kansdichtheidsfuncties $n_{\mu,\sigma}(x)$ en de bijbehorende verdelingsfuncties $N_{\mu,\sigma}(x)$ voor $\mu = 0$ en $\sigma = 0.5$, 1 en 2

niet in termen van bekende elementaire functies kan worden uitgedrukt.
In Figuur 2.9 ziet u grafieken van de kansdichtheidsfuncties en de bijbehorende verdelingsfuncties voor $\mu = 0$ en $\sigma = 0.5$, 1 en 2.

standaardnormale verdeling

Voor $\mu = 0$ en $\sigma = 1$ spreekt men over de *standaardnormale verdeling*. De bijbehorende kansdichtheidsfunctie en verdelingsfunctie geeft men meestal aan met $\varphi(x)$, resp. $\Phi(x)$. Er geldt dus

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}$$

en

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}t^2}\, dt$$

Van de functiewaarden van $\Phi(x)$ bestaan er ten behoeve van statistische berekeningen uitgebreide tabellen. Via de eenvoudige transformatie

$$N_{\mu,\sigma}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

kan men met deze tabellen ook waarden van de algemene normale verdelingsfunctie berekenen. Moderne grafische rekenmachines en wiskundige software-pakketten beschikken echter over de mogelijkheid om de waarden van $N_{\mu,\sigma}$ voor elke μ en σ direct te berekenen.
We geven volledigheidshalve nog een bewijs van de transformatieformule. Het is

vrijwel gelijk aan dat van Stelling 2.1. Noem $u = (t-\mu)/\sigma$, dan is

$$N_{\mu,\sigma}(x) = \int_{-\infty}^{t=x} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt =$$

$$= \int_{-\infty}^{u=\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}u^2} du = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

□

Opgave 2.10 *Beschrijf via welke meetkundige transformaties de grafiek van $n_{\mu,\sigma}(x)$ ontstaat uit de grafiek van de standaardnormale kansdichtheidsfunctie $\varphi(x)$. Beantwoord dezelfde vraag voor de grafieken van de verdelingsfuncties $N_{\mu,\sigma}(x)$ en $\Phi(x)$.*

Opgave 2.11 *Toon aan dat $\Phi(0) = \frac{1}{2}$.*

Opgave 2.12 *Toon aan dat $\Phi(-x) = 1 - \Phi(x)$ voor elke x.*

Opgave 2.13 *Aan een tabel voor $\Phi(x)$ ontlenen we:*
$\Phi(1) = 0.8413, \quad \Phi(2) = 0.9772, \quad \Phi(3) = 0.9987.$
Bereken met behulp van deze gegevens in drie decimalen nauwkeurig de onderstaande kansen voor de normale verdeling met verdelingsfunctie $N_{\mu,\sigma}(x)$.

a. $P([-\infty, \mu - 2\sigma])$
b. $P([\mu - \sigma, \mu + \sigma])$
c. $P([\mu - 2\sigma, \mu + 2\sigma])$
d. $P([\mu - 3\sigma, \mu + 3\sigma])$.

Opgave 2.14 *In de literatuur en in veel computeralgebrapakketten wordt gebruik gemaakt van de zogenaamde* error function $\operatorname{erf}(x)$ *die gedefinieerd wordt door*

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

Deze functie is nauw verwant aan de verdelingsfunctie $\Phi(x)$ van de standaardnormale verdeling. In deze opgave laten we u deze samenhang afleiden.

a. Bereken $\lim_{x\to\infty} \operatorname{erf}(x)$ *en* $\lim_{x\to-\infty} \operatorname{erf}(x)$.
b. Druk $\Phi(x)$ uit in termen van de error function.
c. Druk $N_{\mu,\sigma}(x)$ uit in termen van de error function.

Hoofdstuk 3

Stochastische variabelen

3.1 Inleiding

In het vorige hoofdstuk heeft u gezien hoe men discrete en continue kansmodellen kan beschrijven door een uitkomstenruimte te geven en kansen te definiëren op zekere deelverzamelingen daarvan. Bij discrete modellen kenden we daarbij uiteindelijk een kans toe aan elke deelverzameling van de uitkomstenruimte, maar bij continue modellen, waarbij we als uitkomstenruimte steeds de gehele $\mathbb{R}$ namen, definieerden we slechts kansen op intervallen. We deden dit via een (cumulatieve) verdelingsfunctie: een continue functie op $\mathbb{R}$ die monotoon stijgt van 0 naar 1. De kans die hoort bij een interval is dan de toename van de verdelingsfunctie over dat interval. Ook bij discrete kansmodellen met een deelverzameling van $\mathbb{R}$ als uitkomstenruimte konden we kansen definiëren via een verdelingsfunctie; in dat geval bleek zo'n functie een niet-dalende trapfunctie op $\mathbb{R}$ te zijn met sprongdiscontinuïteiten in de punten van $\mathbb{R}$ waaraan een positieve kans wordt toegekend.
Het bleek daarnaast ook mogelijk te zijn om continue kansmodellen te definiëren via een kansdichtheidsfunctie, dat wil zeggen een niet-negatieve integreerbare functie $f(x)$ op $\mathbb{R}$ met totale integraal 1. De bijbehorende verdelingsfunctie is dan $F(x) = \int_{-\infty}^{x} f(t)dt$, en de kans die hoort bij een willekeurig interval I is de integraal over I van de kansdichtheidsfunctie.
Bij deze benadering is het *kansexperiment* een beetje naar de achtergrond verdwenen. Kansen worden via een verdelingsfunctie of een kansdichtheidsfunctie gedefinieerd, en het uitvoeren van het experiment – het gooien van een munt of een dobbelsteen, het bepalen van de brandtijd van een gloeilamp, het bepalen van de tussentijd tussen de aankomst van twee opvolgende klanten – is in het model eigenlijk niet meer terug te vinden. In dit hoofdstuk zullen we dat element nu ook in ons wiskundige model onderbrengen.
We voeren daartoe in ons wiskundige kansmodel het begrip *stochastische variabele* in. Intuïtief kunt u zo'n variabele opvatten als een functie die aan elke uitkomst van het kansexperiment een reëel getal toevoegt. Soms is dat getal de uitkomst zelf: denk bijvoorbeeld aan het aantal ogen dat boven ligt na het werpen met een dobbelsteen, of aan de brandtijd van een lamp. Maar soms moet er na het uitvoeren van het experiment eerst een berekening worden uitgevoerd: denk bijvoorbeeld aan het werpen met vier dobbelstenen, waarbij men als stochastische variabele de som kiest van de geworpen ogenaantallen.
Vaak associeert men uitspraken over kansen met de mogelijke waarden van zo'n stochastische variabele. Men spreekt bijvoorbeeld over 'de kans dat bij het werpen met vier dobbelstenen het totale ogenaantal 12 is', of over 'de kans dat de brandtijd van een gloeilamp meer dan 100 uur is'. In dit hoofdstuk zullen we uitleggen hoe

men in een wiskundig model dit soort uitspraken en begrippen nader kan preciseren.

3.2 Discrete stochastische variabelen

Wanneer we een kansexperiment uitvoeren, bepalen we een uitkomst u uit de uitkomstenruimte U. Bij discrete kansmodellen kunnen we dan direct over de kans $P(u)$ op die uitkomst spreken. Bij continue modellen kan dat ook, maar dat is weinig zinvol, want die kans is altijd nul. Met de uitkomst u kunnen we in het continue geval echter ook de kans $P(\langle -\infty, u])$ associëren, met andere woorden, de functiewaarde $F(u)$ van de verdelingsfunctie. Door de verdelingsfunctie ligt het kansmodel volledig vast.
In veel gevallen zijn we echter niet zozeer geïnteresseerd in de uitkomst u van het kansexperiment, als wel in een zekere reële functiewaarde $\underline{x}(u)$ die we met behulp van die uitkomst kunnen berekenen. Het gaat dan dus om een *functie* $\underline{x}$ met domein U en waarden in $\mathbb{R}$. Zo'n functie noemen we een *stochastische variabele*, of ook wel een *toevalsvariabele*, soms kortweg een *stochast*. We geven hiervan een aantal voorbeelden.

stochastische variabele

Voorbeeld 3.1 *Het werpen met twee dobbelstenen.*

Stel dat we met twee dobbelstenen werpen, bijvoorbeeld een rode en een groene. Een uitkomst van zo'n kansexperiment bestaat uit een getallenpaar (r, g), waarbij r het aantal ogen van de rode dobbelsteen, en g dat van de groene dobbelsteen is. De uitkomstenruimte U bestaat dan uit 36 elementen, en een plausibel discreet kansmodel is dat men aan elk van die uitkomsten dezelfde kans toekent: elk paar krijgt kans $1/36$. Als stochastische variabele kunnen we nu de totale score, dat wil zeggen het totale aantal ogen bij zo'n worp nemen. We noemen die stochastische variabele $\underline{s}$. Voor elke $u \in U$ geldt dan dat $\underline{s}(u) \in \{2, 3, \ldots, 11, 12\}$.

Aan elk van die scores zullen we nu een kans gaan toekennen. Dit sluit aan bij de dagelijkse praktijk, waarin we het ook hebben over 'de kans op een score van 12 bij het gooien met twee dobbelstenen'. Iedereen 'weet' dat de kans op een score van 12 kleiner is dan de kans op een score van 7. Dit kunnen we in het model ook direct zien. In Tabel 3.1 ziet u bij elke uitkomst (r, g) de score $s = r + g$.

Tabel 3.1: Totaal aantal ogen bij het werpen met twee dobbelstenen.

r \ g	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Hieruit blijkt dat er slechts één uitkomst is met score 12, en zes uitkomsten met score 7. Aangezien we in ons model alle uitkomsten dezelfde kans $1/36$ hebben gegeven, ligt het voor de hand om aan de waarde 12 van de stochastische variabele $\underline{s}$ de kans $1/36$ toe te kennen, en aan de

waarde 7 de kans $6/36 = 1/6$. We noteren die kansen als resp. $P(\underline{s} = 12)$ en $P(\underline{s} = 7)$.

In het algemeen kan men zo voor elke waarde s van de stochastische variabele $\underline{s}$ de kans $P(\underline{s} = s)$ definiëren. In formule:

$$P(\underline{s} = s) = P(\{u \in U \mid \underline{s}(u) = s\}).$$

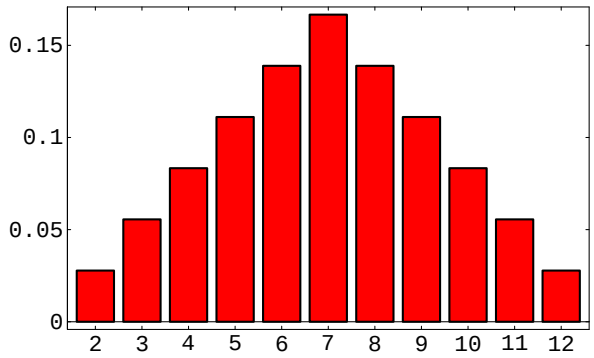
Let op het gebruik van onderstreepte letters en niet-onderstreepte letters: we zullen een stochastische variabele altijd aangeven met een onderstreepte letter, en de waarde van een stochastische variabele met een niet-onderstreepte letter. In Tabel 3.2 ziet u een overzicht van de kansen bij elke score. In tabelvorm staat hier de zogenaamde de *kansfunctie* van de stochastische variabele $\underline{s}$, dat wil zeggen de functie die bij elke waarde van $\underline{s}$ de bijbehorende kans geeft. Met behulp van een staafdiagram kunnen we die kansfunctie in beeld brengen. In Figuur 3.1 ziet u het staafdiagram van de kansfunctie van de stochastische variabele $\underline{s}$.

kansfunctie

We kunnen het zojuist beschreven kansexperiment natuurlijk ook beschrijven als het twee maal achter elkaar werpen met dezelfde dobbelsteen, waarbij we voor elke worp als kansmodel dat van de zuivere dobbelsteen gebruiken. In zulke gevallen zeggen we voortaan eenvoudig dat we 'twee maal met een zuivere dobbelsteen werpen'. De stochastische variabele $\underline{s}$ kan men dan beschrijven als 'het totaal aantal ogen bij het twee maal werpen met een zuivere dobbelsteen'.

Tabel 3.2: Kansfunctie van de totale score bij het twee maal werpen met een zuivere dobbelsteen.

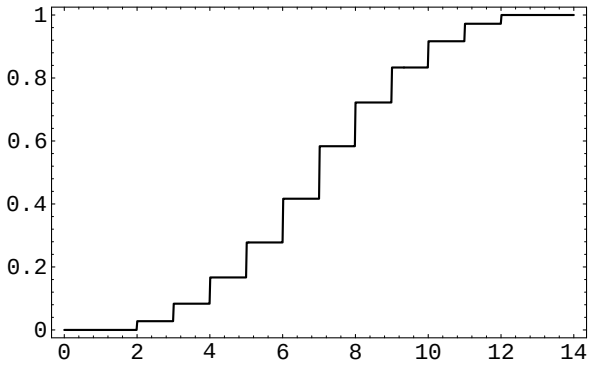
s	2	3	4	5	6	7	8	9	10	11	12
$P(\underline{s} = s)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



Figuur 3.1: Staafdiagram van de kansfunctie van de stochastische variabele $\underline{s}$, de totale score bij het twee maal werpen met een zuivere dobbelsteen.

verdelingsfunctie van een discrete stochastische variabele

Omdat de waarden van een stochastische variabele $\underline{x}$ altijd reële getallen zijn, kunnen we bij $\underline{x}$ in het discrete geval een (cumulatieve) verdelingsfunctie $F_{\underline{x}}(x)$ definiëren via de formule



Figuur 3.2: Verdelingsfunctie van de stochastische variabele $\underline{s}$, de totale score bij het twee maal werpen met een zuivere dobbelsteen.

$$F_{\underline{x}}(x) = P(\underline{x} \leq x) = \sum_{u \leq x} P(\underline{x} = u).$$

In Figuur 3.2 ziet u de grafiek van de verdelingsfunctie $F_{\underline{s}}(\underline{x})$ van de stochastische variabele $\underline{s}$, de totale score bij het twee maal werpen met een zuivere dobbelsteen. Het is weer een trapfunctie die van 0 naar 1 stijgt. In elk sprongpunt is de sprong-grootte gelijk aan de kans op de betreffende waarde. Zo maakt de verdelingsfunctie bij $x = 2$ een sprong ter grootte $1/36$ want $P(\underline{s} = 2) = 1/36$, bij $x = 3$ een sprong ter grootte $2/36$ want $P(\underline{s} = 3) = 2/36$, enzovoort.

Opgave 3.1 *Geef bij elk van de volgende stochastische variabelen de kansfunctie en de verdelingsfunctie.*

a. Het totale aantal malen munt *bij het drie maal werpen met een zuivere munt.*

b. De absolute waarde van het verschil van de uitkomsten bij het twee maal werpen met een zuivere dobbelsteen.

c. Het totale aantal ogen bij het werpen met twee zuivere dobbelstenen waarbij op de zijvlakken van de ene steen 2, 3, 4, 5, 6, 6, ogen staan, en op de zijvlakken van de andere steen 1, 1, 2, 3, 4, 5 ogen.

Opgave 3.2 *Ik speel met Michael het volgende spel. Ieder van ons zet 1 gulden in. We gooien om beurten met een zuivere munt. Wie het eerst 'kop' gooit, wint de inzet; daarna is het spel afgelopen. Michael laat mij beginnen. Onder de stochastische variabele $\underline{x}$ versta ik mijn bezit (0 gulden of 2 gulden) na afloop van het spel. Bepaal de kansfunctie $P(\underline{x} = x)$ en de verdelingsfunctie $F_{\underline{x}}(x)$ van $\underline{x}$.*

Opgave 3.3 *Ik speel tegen Gabriel het volgende spel. We zetten bij iedere ronde beide hetzelfde bedrag in. Daarna gooi ik met een zuivere munt. Is de uitkomst 'kop' dan krijg ik de totale inzet, is het 'munt', dan krijgt Gabriel alles. Ik hanteer de volgende strategie: zodra ik win, stop ik met het spel, maar als ik verlies, verdubbel ik mijn inzet. Mijn eerste inzet is 1 gulden. Gabriel is altijd bereid mee te spelen. Onder $\underline{x}$ versta ik mijn winst na afloop van het spel. Bepaal de kansverdeling en de verdelingsfunctie van $\underline{x}$.*

Opgave 3.4 *Ik speel tegen Lucifer hetzelfde spel als tegen Gabriel. Ik hanteer dezelfde strategie. Lucifer is echter niet bereid meer dan 10 ronden te spelen.*
Onder $\underline{y}$ versta ik mijn winst na afloop van het spel. Bepaal de kansverdeling en de verdelingsfunctie van $\underline{y}$.

3.2.1 De binomiale verdeling

Voorbeeld 3.2 *Lampen in een doos.*

We keren terug naar Voorbeeld 1.2, de massaproductie van gloeilampen, waarbij naast goede exemplaren ook defecte lampen geproduceerd kunnen worden. We hebben gezien dat er een bruikbaar kansmodel ontstaat wanneer men aanneemt dat de productie een kansexperiment is, waarbij defecte exemplaren optreden met een zekere vaste kans p (met $0 < p < 1$). Stel nu dat de lampen met 12 tegelijk in een doos verpakt worden. Het aantal defecte exemplaren in zo'n doos kan dan 0, 1, ..., 11 of 12 zijn. We kunnen dat aantal opvatten als een stochastische variabele $\underline{k}$. Wat is de bijbehorende kansfunctie, met andere woorden, wat is $P(\underline{k} = k)$ voor $k = 0, 1, \ldots, 12$?

Elke doosinhoud kunnen we voorstellen door een rijtje van twaalf letters G of D, voor 'goed' of 'defect'. De kans op het rijtje $GGGGGGGGGGGG$ is dan $(1-p)^{12}$, want de kans op één goed exemplaar is $(1-p)$ en elke lamp in de doos heeft dezelfde kans om 'goed' of 'defect' te zijn. Er geldt dus

$$P(\underline{k} = 0) = (1-p)^{12}.$$

Wat is de kans dat er één defect exemplaar in de doos zit? In het rijtje van twaalf kan de D op twaalf verschillende plaatsen staan, dus er zijn twaalf verschillende rijtjes met één D en elf G's. De kans op elk rijtje afzonderlijk is $p(1-p)^{11}$, dus

$$P(\underline{k} = 1) = 12p(1-p)^{11}.$$

Bij meer defecte exemplaren moeten we binomiaalcoëfficiënten gebruiken:

$$P(\underline{k} = 2) = \binom{12}{2} p^2 (1-p)^{10}$$

want er zijn $\binom{12}{2}$ rijtjes met twee D's en tien G's. (Ter herinnering: het aantal manieren om k exemplaren uit n dingen te kiezen is gelijk aan $\binom{n}{k} = \dfrac{n!}{(n-k)!k!}$.)
In het algemeen geldt dus

$$P(\underline{k} = k) = \binom{12}{k} p^k (1-p)^{12-k}.$$

De bijbehorende verdelingsfunctie is

$$F_{\underline{k}}(x) = \sum_{k \le x} P(\underline{k} = k) = \sum_{k \le x} \binom{12}{k} p^k (1-p)^{12-k}$$

In Voorbeeld 33.2 hadden we $p = 0.03$ genomen. In Tabel 34.3 staan hierbij de waarden $F_{\underline{k}}(x)$ van de verdelingsfunctie voor $x = 0, \ldots, 12$ op vier decimalen afgerond. Omdat p zo klein is, stijgt $F_{\underline{k}}$ erg snel naar 1. Toch is de kans op een doos met één defect exemplaar nog $P(\underline{k} = 1) = 0.2575$, dat is ruim 25% !

We kunnen Voorbeeld 3.2 direct generaliseren tot de situatie waarin sprake is van n herhalingen van een kansexperiment met twee mogelijke uitslagen, die we nu even 'succes' en 'mislukking' zullen noemen. Laat p de kans op 'succes' zijn, en $(1-p)$

Tabel 3.3: Waarden van de kansfunctie en de (cumulatieve) verdelingsfunctie van het aantal defecte exemplaren in een doos van 12 bij een kans $p = 0.03$, afgerond op vier decimalen

k	$P(\underline{k} = k)$	$F_{\underline{k}}(k)$
0	0.6938	0.6938
1	0.2575	0.9513
2	0.0438	0.9951
3	0.0045	0.9996
4	0.0000	1.0000
5	0.0000	1.0000
...	...	
12	0.0000	1.0000

de kans op 'mislukking'. Als stochastische variabele $\underline{x}$ nemen we het totale aantal successen in zo'n serie van n uitvoeringen van het kansexperiment. De kans op k successen is dan

$$P(\underline{x} = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

binomiale verdeling

Vanwege de overeenkomst van deze uitdrukking met termen van de binomiaalformule van Newton noemt men deze kansverdeling de *binomiale verdeling* met *parameters* n en p. De bijbehorende verdelingsfunctie noteren we als $B_{n,p}(x)$. De formule ervan luidt:

$$B_{n,p}(x) = \sum_{k \le x} \binom{n}{k} p^k (1-p)^{n-k}.$$

Dit is een trapfunctie met $n+1$ discontinuïteiten. Er geldt $B_{n,p}(x) = 0$ voor $x < 0$ en $B_{n,p}(x) = 1$ voor $x \ge n$.
De geldigheid van die laatste gelijkheid, $B_{n,p}(x) = 1$ voor $x \ge n$, volgt overigens ook direct uit de formule voor het binomium van Newton

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

wanneer u $a = p$ en $b = 1 - p$ substitueert. Dit is opnieuw een teken van de verbondenheid van deze kansverdeling met de binomiaalformule.
In Figuur 3.3 ziet u de grafiek van de verdelingsfunctie $B_{n,p}(x)$ voor $n = 20$ en resp. $p = 0.3$ en $p = 0.7$, terwijl in Figuur 34.4 de staafdiagrammen getekend zijn van de bijbehorende kansfuncties.

Opgave 3.5 *Ik wed dat ik met twee zuivere dobbelstenen in 24 worpen minstens één maal dubbelzes kan gooien. Hoe groot is de kans dat ik de weddenschap win?*

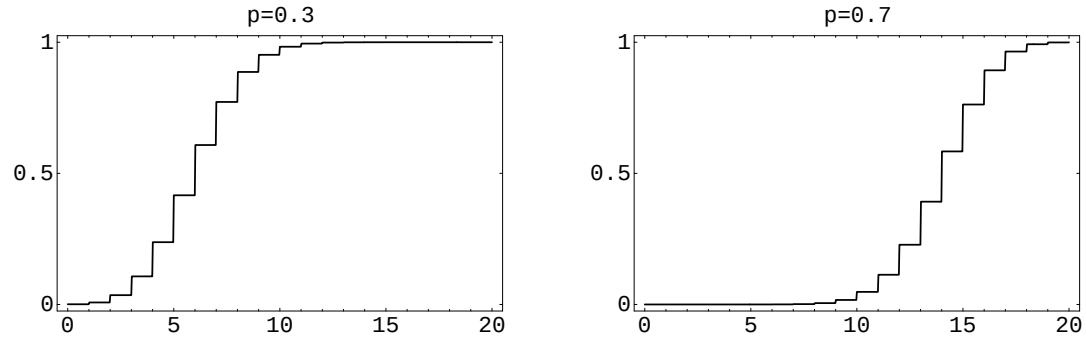
Opgave 3.6 *Een vaas bevat 3 witte en 7 zwarte knikkers die alleen in kleur van elkaar verschillen. Men trekt 20 maal blindelings een knikker uit de vaas, noteert de kleur, doet de knikker weer terug in de vaas en schudt de vaas goed alvorens opnieuw te trekken.*

a. Bereken de kans dat men precies 6 witte knikkers trekt.

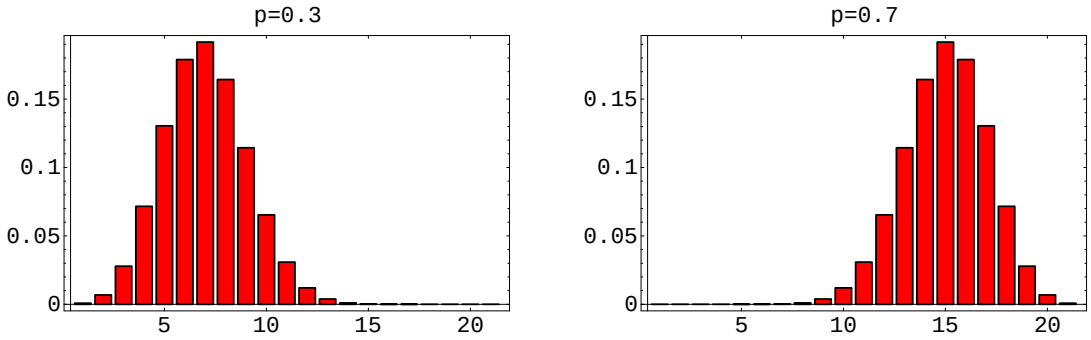
b. Bereken de kans dat men hoogstens 2 witte knikkers trekt.

Geef uw antwoorden zowel in formulevorm als ook in vier decimalen nauwkeurig.

Opgave 3.7 *Bereken de kansverdeling en de verdelingsfunctie van de binomiale verdeling met parameters $n = 4$ en $p = \frac{1}{2}$.*



Figuur 3.3: Grafieken van de binomiale verdelingsfuncties $B_{n,p}(x)$ voor $n = 20$ en $p = 0.3$ resp. $p = 0.7$.



Figuur 3.4: Staafdiagrammen van de bijbehorende binomiale kansfuncties

3.3 Continue stochastische variabelen

Voorbeeld 3.3 *Gemiddelde brandtijd van 12 gloeilampen*

Ter introductie van het begrip stochastische variabele bij continue kansmodellen, nemen we weer het voorbeeld van de massaproductie van gloeilampen. Nu nemen we aan dat we dozen van 12 stuks bekijken waarin geen defecte exemplaren zitten. Als stochastische variabele $\underline{m}$ nemen we het gemiddelde van de brandtijden van de 12 lampen.

Als we in ons kansmodel zijn uitgegaan van een normale verdeling van de brandtijden van de lampen met (bijvoorbeeld) $\mu = 100$ en $\sigma = 7$, ligt daarmee in principe ook de kansverdeling van deze nieuwe stochastische variabele $\underline{m}$ vast: er zal een verdelingsfunctie $F_{\underline{m}}(x)$ bestaan die op de een of andere manier samenhangt met de verdelingsfunctie $N_{\mu,\sigma}(x)$ van de brandtijd per lamp in ons model, en die aan een nieuwe kansverdeling op $\mathbb{R}$ gekoppeld is volgens de formule

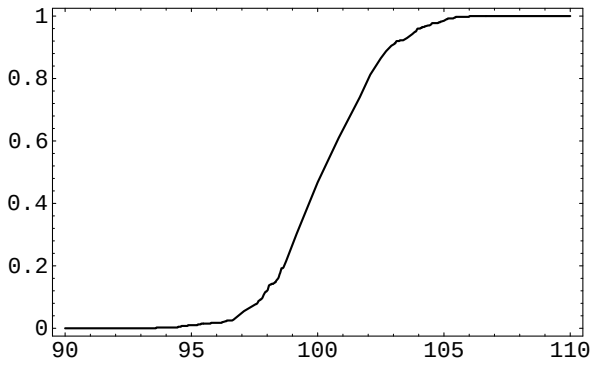
$$F_{\underline{m}}(x) = P(\underline{m} \leq x).$$

Kennen we die verdelingsfunctie, dan hebben we daarmee ook direct de nieuwe kansverdeling te pakken: de kans op een interval I, of, wat realistischer geformuleerd, de kans dat de gemiddelde brandtijd $\underline{m}$ in het interval I valt, is dan immers gelijk aan de toename van de verdelingsfunctie $F_{\underline{m}}(x)$ over I.

Het is echter niet duidelijk of die verdelingsfunctie ook door een eenvoudige formule gegeven kan worden. Om enig idee te krijgen van de aard van de verdelingsfunctie, kunnen we een simulatie uitvoeren. We

bepalen met behulp van een aangepaste randomgenerator (bijvoorbeeld) 4800 trekkingen uit de normale verdeling met parameters $\mu = 100$ en $\sigma = 7$, nemen ze telkens met twaalf stuks samen, bepalen daarvan het gemiddelde, en maken net als in Hoofdstuk 1 een grafiek van de bijbehorende cumulatieve frequentiefunctie. We verwachten dat die sterk zal lijken op die van de verdelingsfunctie van $\underline{m}$ waarnaar we op zoek zijn.

In Figuur 3.5 ziet u het resultaat. Deze figuur lijkt sterk op Figuur 2.3, de cumulatieve frequentiefunctie bij een simulatie van de brandtijd van 200 lampen. Figuur 3.5 wekt dus het vermoeden dat ook het gemiddelde $\underline{m}$ normaal verdeeld is, en wel met een 'verwachting' μ die weer 100 is, maar met een veel kleinere 'spreiding'. Die laatste twee observaties – zelfde verwachting, kleinere spreiding – hadden we op intuïtieve gronden misschien ook wel verwacht bij 'gemiddelden nemen'. Maar dat de vorm van de grafiek bij benadering weer die van een normale verdeling zou zijn, is niet vanzelfsprekend. Het is echter wel waar: men kan het bewijzen, maar het bewijs ervan is niet triviaal! En men kan ook bewijzen dat de μ inderdaad ongewijzigd blijft, en dat de nieuwe σ verkregen wordt door de oude te delen door de wortel uit het aantal lampen per doos. In dit geval is de nieuwe σ dus $7/\sqrt{12} \approx 2.02$.

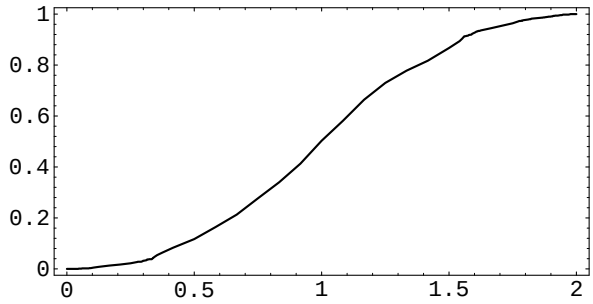


Figuur 3.5: Cumulatieve frequentiefunctie bij 400 gemiddelde brandtijden in dozen van 12 lampen.

Voorbeeld 3.4 *De som van twee uniform verdeelde stochastische variabelen.*

Dat de som van twee stochastische variabelen niet altijd een stochastische variabele hoeft te zijn met dezelfde soort kansverdeling, hebben we eigenlijk al gezien in Voorbeeld 3.1 bij het werpen met twee 'zuivere' dobbelstenen. Bij iedere worp apart is elke uitkomst even waarschijnlijk, maar bij de som van de ogenaantallen bij twee maal werpen is dat niet meer het geval: de uitkomst 7 heeft bijvoorbeeld een veel grotere kans dan de uitkomsten 2 of 12.

Ook bij de uniforme verdeling in de continue situatie doet zich dat verschijnsel voor. We illustreren dit met een simulatie waarin we met de randomgenerator tweeduizend trekkingen uit de uniforme verdeling op $[0, 1]$ doen. We nemen die trekkingen met twee tegelijk samen en bepalen bij elk paar de som. In Figuur 3.6 ziet u de cumulatieve frequentiefunctie van die duizend sommen. Het is een verdeling op $[0, 2]$ die zeker niet uniform is!



Figuur 3.6: Cumulatieve frequentiefunctie bij duizend sommen van twee uniform verdeelde stochastische variabelen op $[0, 1]$

Bij een continue stochastische variabele speelt de verdelingsfunctie een belangrijke rol. Vaak wordt die vastgelegd door een kansdichtheidsfunctie, dat wil zeggen een niet-negatieve integreerbare functie met een totale integraal die gelijk is aan 1. Als $\underline{x}$ de stochastische variabele is, en $f_{\underline{x}}(x)$ is de bijbehorende kansdichtheidsfunctie, dan wordt de verdelingsfunctie $F_{\underline{x}}(x)$ van $\underline{x}$ gegeven door *kansdichtheidsfunctie*

$$F_{\underline{x}}(x) = \int_{-\infty}^{x} f_{\underline{x}}(t)dt.$$

Verder geldt

$$F_{\underline{x}}(x) = P(\underline{x} \leq x)$$

In het algemeen kunnen we voor elk interval I de kans dat een waarde van $\underline{x}$ in I ligt, schrijven als de integraal over I van de kansdichtheidsfunctie: *kans als integraal*

$$P(\underline{x} \in I) = \int_{I} f_{\underline{x}}(t)dt.$$

Zoals uit de hierboven gegeven voorbeelden blijkt, is het in de praktijk echter niet altijd eenvoudig om bij een gegeven stochastische variabele expliciete formules te bepalen voor de kansdichtheidsfunctie of de verdelingsfunctie.

Opgave 3.8 *Definieer*

$$f(x) = \begin{cases} \dfrac{4x}{\pi(1+x^4)} & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

a. Laat zien dat $f(x)$ een kansdichtheidsfunctie is.

Laat $\underline{x}$ de bijbehorende stochastische variabele zijn.

b. Bereken de bijbehorende verdelingsfunctie.
c. Bereken $P(\underline{x} \leq 1)$.

Hoofdstuk 4

Verwachting en variantie

4.1 De verwachting van een discrete stochastische variabele

We komen nog even terug op Tabel 3.2 op bladzijde 29, die de kansfunctie geeft van de stochastische variabele $\underline{s}$, de som van het aantal ogen bij het twee maal werpen met een zuivere dobbelsteen.
Laten we aannemen dat we dit kansexperiment een zeer groot aantal malen, bijvoorbeeld een miljoen maal, uitvoeren, daarbij telkens de waarde van $\underline{s}$ berekenen, en van al die waarden het gemiddelde nemen. Welke uitkomst verwachten we dan te krijgen? Overeenkomstig de experimentele wet van de grote aantallen verwachten we dat de relatieve frequentie waarmee elke waarde van $\underline{s}$ optreedt, vrijwel niet te onderscheiden zal zijn van de bijbehorende kans. Omdat de kans op 2 gelijk is aan 1/36, verwachten we dus dat ongeveer 1/36e deel van de uitkomsten van het kansexperiment de waarde 2 oplevert, enzovoort. Als we de gemiddelde waarde van alle uitkomsten bepalen, verwachten we dus ongeveer de uitkomst

$$\mu = 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + 4 \times \frac{3}{36} + \ldots + 11 \times \frac{2}{36} + 12 \times \frac{1}{36} = \frac{252}{36} = 7$$

te krijgen. We noemen deze theoretische waarde de *verwachting* van $\underline{s}$. In het algemeen definieert men het begrip verwachting van een discrete stochastische variabele als volgt.

> **Definitie 4.1** *Laat $\underline{x}$ een discrete stochastische variabele zijn die de waarden x_1, x_2, ... aanneemt met kansen resp. p_1, p_2, ... De verwachting van $\underline{x}$ wordt dan gedefinieerd door*
>
> $$E(\underline{x}) = \sum_i x_i p_i.$$

(De letter E komt van het Engelse woord 'expectation'.) In sommige gevallen zullen we ook de Griekse letter μ voor de verwachting gebruiken, of $\mu_{\underline{x}}$ wanneer we willen benadrukken dat het om de verwachting gaat van de stochastische variabele $\underline{x}$.
Heeft $\underline{x}$ slechts eindig veel waarden, dan staat hier een eindige som. Er zijn echter ook discrete stochastische variabelen met een aftelbare waardenverzameling. In dat geval kan er een probleem optreden: de (oneindige) som zou dan kunnen divergeren, of de uitkomst zou afhankelijk kunnen zijn van de sommatievolgorde. Men eist in dat geval daarom voor het bestaan van de verwachting dat

$$\sum_i |x_i p_i| = \sum_i |x_i| p_i < \infty$$

(bedenk dat de kansen p_i positief zijn). Slechts als aan deze voorwaarde voldaan is, is $E(\underline{x})$ gedefinieerd.

Voorbeeld 4.1 *Werpen met een zuivere dobbelsteen.*

Als $\underline{x}$ de uitkomst is bij het werpen met een zuivere dobbelsteen, dan geldt

$$E(\underline{x}) = \frac{1}{6}\,1 + \frac{1}{6}\,2 + \frac{1}{6}\,3 + \frac{1}{6}\,4 + \frac{1}{6}\,5 + \frac{1}{6}\,6 = \frac{7}{2}$$

Voorbeeld 4.2 *Bernoulli-experiment met parameter p.*

Stel dat we een kansexperiment hebben met twee uitkomsten 'succes' en 'mislukking' met kansen resp. p en $1-p$. Men noemt zo'n experiment vaak een *Bernoulli-experiment* met parameter p, naar Jakob Bernoulli (1654-1705), wiens *Ars Conjectandi* een van de eerste systematische leerboeken over kansrekening genoemd kan worden.

Definieer bij zo'n Bernoulli-experiment de stochastische variabele $\underline{x}$ door $\underline{x} = 1$ bij 'succes', en $\underline{x} = 0$ bij 'mislukking'. Dan is

$$E(\underline{x}) = p \times 1 + (1-p) \times 0 = p.$$

Voorbeeld 4.3 *De binomiale verdeling.*

Laat $\underline{x}$ binomiaal verdeeld zijn met parameters n en p, dat wil zeggen dat

$$P(\underline{x} = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

voor $k = 0, 1, \ldots, n$. De verwachting van $\underline{x}$ is dan

$$\begin{aligned}
E(\underline{x}) &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= n \sum_{k=1}^{n} \binom{n-1}{k-1} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{(n-1)-j} \\
&= np
\end{aligned}$$

Hierbij hebben we gebruikt dat

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} = \frac{n}{k}\,\frac{(n-1)!}{((n-1)-(k-1))!(k-1)!} = \frac{n}{k}\binom{n-1}{k-1}$$

en dat de som op de voorlaatste regel, als som van alle kansen bij de binomiale verdeling met parameters $n-1$ en p, gelijk is aan 1.

Opgave 4.1 *Ik werp 24 maal met twee zuivere dobbelstenen. Onder de stochastische variabele $\underline{x}$ versta ik het aantal malen dubbelzes dat ik werp. Bereken de verwachting van $\underline{x}$.*

Opgave 4.2 *Bereken de verwachting van elk van de stochastische variabelen uit Opgave 3.1.*

Opgave 4.3 *Idem voor de stochastische variabelen $\underline{x}$ en $\underline{y}$ uit Opgave 3.3 en Opgave 3.4.*

Opgave 4.4 *Laat $\underline{x}$ de uitkomst zijn bij het werpen met een zuivere dobbelsteen. Bepaal de verwachting van $\underline{y} = \underline{x}^2$.*

De volgende stelling is nuttig wanneer een kansexperiment met stochastische variabele $\underline{x}$ gebruikt wordt om een andere, van $\underline{x}$ afhangende stochastische variabele $\underline{y}$ te genereren. We zullen bijvoorbeeld zien dat het vaak voorkomt dat men behalve de stochastische variabele $\underline{x}$ ook de stochastische variabele $\underline{y} = \underline{x}^2$ in berekeningen wil gebruiken. Hierbij bedoelen we met $\underline{x}^2$ de stochastische variabele die de waarden van $\underline{x}$ kwadrateert, dat wil zeggen: als $\underline{x}$ de waarde x aanneemt, dan neemt $\underline{y} = \underline{x}^2$ de waarde x^2 aan. Op een soortgelijke wijze definieert men in het algemeen voor een willekeurige functie g van $\mathbb{R}$ naar $\mathbb{R}$ de samengestelde stochastische variabele $\underline{y} = g(\underline{x})$ als de stochastische variabele die de waarde $g(x)$ aanneemt wanneer $\underline{x}$ de waarde x aanneemt.
De verwachting van zulke samengestelde stochastische variabelen kan men op eenvoudige wijze uitdrukken in de kansen van de elementaire gebeurtenissen van het oorspronkelijke discrete kansmodel.

Stelling 4.1 *Laat $\underline{x}$ een discrete stochastische variabele zijn met waarden x_i die aangenomen worden met kansen resp. p_i. Laat de stochastische variabele $\underline{y}$ gedefinieerd zijn door $\underline{y} = g(\underline{x})$ voor zekere functie g. Als dan de verwachting $E(\underline{y})$ bestaat, geldt*

$$E(\underline{y}) = \sum_i g(x_i)p_i.$$

BEWIJS: We schrijven $y_j = g(x_j)$ en merken op dat verschillende x_j-waarden dezelfde y_j kunnen geven. Er geldt nu

$$P(\underline{y} = y_j) \ = P(\underline{y} = g(x_j)) \ = \sum_{i \mid g(x_i)=y_j} P(\underline{x} = x_i) \ = \sum_{i \mid g(x_i)=y_j} p_i,$$

waarbij de sommaties plaatsvinden over alle i waarvoor $g(x_i)$ gelijk is aan y_j. Dit is in de notatie onder het somteken symbolisch weergegeven als $i \mid g(x_i) = y_j$. Uit het bovenstaande volgt

$$E(\underline{y}) = \sum_j y_j P(\underline{y} = y_j) \ = \sum_i g(x_i)p_i.$$

□

Een gevolg van Stelling 4.1 wordt vaak gebruikt:

Stelling 4.2 *Als $\underline{x}$ een discrete stochastische variabele is met verwachting $E(\underline{x})$ en als de stochastische variabele $\underline{y}$ gedefinieerd wordt door $\underline{y} = a\underline{x} + b$ voor zekere constanten a en b, dan geldt*

$$E(\underline{y}) = aE(\underline{x}) + b.$$

BEWIJS: Volgens Stelling 4.1 geldt

$$E(\underline{y}) = \sum_i (ax_i + b)p_i = a \ \sum_i x_i p_i + b \ \sum_i p_i = aE(\underline{x}) + b$$

waarbij de laatste stap volgt uit $\sum_i p_i = 1$. □

4.2 De verwachting van een continue stochastische variabele

Wanneer een continue stochastische variabele $\underline{x}$ gedefinieerd is via een kansdichtheidsfunctie $f_{\underline{x}}(x)$, kunnen we op een analoge wijze de verwachting $E(\underline{x})$ van $\underline{x}$ definiëren; nu echter niet door een som, maar door een integraal. Voorwaarde is dan dat de integraal

$$\int_{-\infty}^{\infty} |x| f_{\underline{x}}(x) dx$$

eindig is.

Definitie 4.2 *Onder de bovengenoemde voorwaarde definieert men de verwachting van $\underline{x}$ door*

$$E(\underline{x}) = \int_{-\infty}^{\infty} x f_{\underline{x}}(x) dx.$$

Intuïtief kan men zich dit ook als volgt voorstellen: verdeel de x-as in kleine deelintervallen I ter lengte Δx. Op zo'n interval heeft x een bij benadering constante waarde, en de kans dat $\underline{x}$ in zo'n interval valt, is $\int_I f_{\underline{x}}(x)dx$. Vermenigvuldig x met die kans, sommeer over alle intervallen en voer een limietovergang uit, dan ontstaat uiteindelijk de bovenstaande integraal. Let wel, dit is geen bewijs! Het is slechts een motivering voor de definitie van de verwachting in het geval van een continue stochastische variabele die aansluit bij de definitie van verwachting in het discrete geval.

Voorbeeld 4.4 *De verwachting van de uniforme verdeling.*

De kansdichtheidsfunctie van de uniforme verdeling op het interval $[a, b]$ wordt gegeven door

$$f(x) = \frac{1}{b-a} \quad \text{als} \quad a \leq x \leq b$$

en $f(x) = 0$ elders. De verwachting is dus gelijk aan

$$\int_a^b \frac{x}{b-a} dx = \frac{1}{2} \frac{b^2 - a^2}{b-a} = \frac{a+b}{2}.$$

Voorbeeld 4.5 *De verwachting van de normale verdeling.*

De verwachting van de normale verdeling met parameters μ en σ is

$$\begin{aligned} E(\underline{x}) &= \int_{-\infty}^{\infty} \frac{x}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\ &= \int_{-\infty}^{\infty} \frac{(x-\mu)+\mu}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\ &= \int_{-\infty}^{\infty} \frac{(x-\mu)}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx + \mu \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\ &= 0 + \mu \;=\; \mu \end{aligned}$$

want de eerste integraal is convergent en de integrand is oneven ten opzichte van het centrum $x = \mu$, en de tweede integraal is, zoals bekend, gelijk aan 1. Inderdaad is de verwachting van de normale verdeling met parameters μ en σ dus gelijk aan μ.

Voorbeeld 4.6 *De negatief-exponentiële verdeling.*

Bij de negatief-exponentiële verdeling met parameter $\lambda > 0$ kan men de verwachting berekenen met behulp van partiële integratie:

$$\begin{aligned} E(\underline{x}) &= \int_0^\infty x\ \lambda e^{-\lambda x} dx \\ &= -\int_0^\infty x\ de^{-\lambda x} \\ &= -\left[x\ e^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x} dx \\ &= 0 + \left[\frac{-1}{\lambda}\ e^{-\lambda x}\right]_0^\infty \\ &= \frac{1}{\lambda}. \end{aligned}$$

Ook de Stellingen 4.1 en 4.2 hebben continue pendanten. Die van Stelling 4.1 luidt:

Stelling 4.3 *Laat $\underline{x}$ een continue stochastische variabele zijn met kansdichtheidsfunctie $f_{\underline{x}}(x)$ en verwachting $E(\underline{x})$, en laat de stochastische variabele $\underline{y}$ gedefinieerd zijn door $\underline{y} = g(\underline{x})$ voor zekere 'nette' functie g van $\mathbb{R}$ naar $\mathbb{R}$. Als ook de verwachting $E(\underline{y})$ bestaat, dan geldt*

$$E(\underline{y}) = \int_{-\infty}^{\infty} g(x) f_{\underline{x}}(x) dx.$$

De geldigheid van deze stelling zal op intuïtieve gronden wel plausibel zijn; een bewijs vergt echter nogal wat techniek. Daarbij moet men bijvoorbeeld precies duidelijk maken wat men in dit verband onder een 'nette' functie verstaat. Dit alles valt ver buiten het bestek van deze cursus. We volstaan met de uitspraak dat bijvoorbeeld alle continue functies aan deze netheidseis voldoen.
In het vervolg zal het kwadraat van een stochastische variabele een grote rol spelen. Hier merken we daarom nu reeds op: als $\underline{x}$ gedefinieerd wordt door een kansdichtheidsfunctie $f_{\underline{x}}(x)$ en als $\underline{y} = \underline{x}^2$, dan geldt op grond van Stelling 4.3 dat

$$E(\underline{y}) = \int_{-\infty}^{\infty} x^2 f_{\underline{x}}(x) dx$$

mits deze integraal eindig is.

Voorbeeld 4.7 *Het kwadraat van een uniform verdeelde variabele.*

Laat $\underline{x}$ uniform verdeeld zijn op $[0, 1]$, en laat $\underline{y} = \underline{x}^2$. We berekenen de verwachting van $\underline{y}$. Omdat de kansdichtheidsfunctie $f_{\underline{x}}(x)$ van $\underline{x}$ gegeven wordt door

$$f_{\underline{x}}(x) = \begin{cases} 1 & (0 \le x \le 1) \\ 0 & (\text{elders}) \end{cases}$$

geldt

$$E(\underline{y}) = \int_0^1 x^2 dx = \frac{1}{3}.$$

Ook alle lineaire functies voldoen natuurlijk aan de 'netheidseis' uit Stelling 4.3, en dus geldt ook de pendant van Stelling 4.2:

Stelling 4.4 *Als $\underline{x}$ een continue stochastische variabele is met verwachting $E(\underline{x})$ en als de stochastische variabele $\underline{y}$ gedefinieerd wordt door $\underline{y} = a\underline{x} + b$ voor zekere constanten a en b, dan geldt*

$$E(\underline{y}) = aE(\underline{x}) + b.$$

BEWIJS: Op grond van Stelling 4.3 geldt

$$\begin{aligned} E(\underline{y}) &= \int_{-\infty}^{\infty} (ax+b) f_{\underline{x}}(x)dx \\ &= a\int_{-\infty}^{\infty} x f_{\underline{x}}(x)dx + b\int_{-\infty}^{\infty} f_{\underline{x}}(x)dx \\ &= aE(\underline{x}) + b. \end{aligned}$$

□

Opgave 4.5 *De stochastische variabele $\underline{x}$ is uniform verdeeld op het interval $[-1, 1]$. Bereken de verwachting van $\underline{y} = \underline{x}^4$.*

Opgave 4.6 *De stochastische variabele $\underline{x}$ is negatief-exponentieel verdeeld met parameter $\lambda = 2$. Bereken de verwachting van $\underline{y} = \underline{x}^2$.*

Opgave 4.7 *Bereken zo mogelijk de verwachting van de Cauchy-verdeling (zie Opgave 2.9)*

Opgave 4.8 *Toon aan dat de stochastische variabele $\underline{x}$ uit Opgave 3.8 een eindige verwachting heeft (u hoeft die verwachting niet te berekenen!).*

4.3 Variantie en standaardafwijking

In de vorige paragrafen hebben we de verwachting van een stochastische variabele $\underline{x}$ gedefinieerd in termen van een som of een integraal. Men kan de verwachting van een stochastische variabele intuïtief opvatten als een soort 'gemiddelde' waarde ervan; bij grote aantallen uitvoeringen van het bijbehorende kansexperiment zal de gemiddelde waarde van de uitkomsten tenderen naar de verwachting. Over de spreiding van de waarden rond dat gemiddelde is daarmee nog niets gezegd. Er zijn variabelen waarbij die spreiding groot is, en variabelen met dezelfde verwachting maar een veel kleinere spreiding. Met de begrippen *variantie* en *standaardafwijking* zullen we maten definiëren voor de spreiding van een stochastische variabele.
In deze paragraaf zal $\underline{x}$ steeds een discrete of continue stochastische variabele zijn waarvoor de verwachting $E(\underline{x})$ bestaat. We zullen die verwachting meestal korter noteren als μ.

> **Definitie 4.3** *De variantie $Var(\underline{x})$ van de stochastische variabele $\underline{x}$ met verwachting μ wordt gedefinieerd door*
>
> $$Var(\underline{x}) = E((\underline{x} - \mu)^2).$$
>
> *De wortel uit de variantie heet de* standaardafwijking*, ook wel* standaarddeviatie*. Hiervoor gebruikt men vaak de Griekse letter σ, dus*
>
> $$\sigma(\underline{x}) = \sqrt{Var(\underline{x})}.$$

Bij de definitie van de variantie vormt men dus de nieuwe stochastische variabele $\underline{x} - \mu$, kwadrateert die, en berekent daarvan de verwachting. In deze vorm is de intuïtieve betekenis van de variantie ook het duidelijkste: het is een maat voor de afwijkingen van de verwachtingswaarde. Het kwadrateren geschiedt vooral om een wiskundig goed hanteerbare uitdrukking te krijgen. Voor praktische berekeningen zijn de volgende stellingen van belang:

Stelling 4.5 *Als a en b constanten zijn dan geldt*

$$Var(a\underline{x}+b) = a^2 Var(\underline{x}) \quad en \quad \sigma(a\underline{x}+b) = |a|\sigma(\underline{x}).$$

BEWIJS: We behoeven slechts de eerste gelijkheid te bewijzen. Op grond van $E(a\underline{x}+b) = aE(\underline{x}) + b = a\mu + b$ geldt

$$\begin{aligned} Var(a\underline{x}+b) &= E(((a\underline{x}+b)-(a\mu+b))^2) \\ &= E((a(\underline{x}-\mu))^2) = a^2 E((\underline{x}-\mu)^2) \\ &= a^2 Var(\underline{x}). \end{aligned}$$

□

Uit Stelling 4.5 volgt in het bijzonder dat de variantie en de standaardafwijking niet veranderen wanneer men bij een stochastische variabele een constante optelt. Omdat het maten zijn voor de spreiding van de waarden, is dit in overeenstemming met onze intuïtie.
De volgende stelling vereenvoudigt de berekening van de variantie (en dus ook die van de standaardafwijking) wanneer men reeds de verwachting $E(\underline{x}) = \mu$ berekend heeft.

Stelling 4.6 *Voor de variantie van een stochastische variabele $\underline{x}$ geldt*

$$Var(\underline{x}) = E(\underline{x}^2) - (E(\underline{x}))^2 = E(\underline{x}^2) - \mu^2.$$

BEWIJS: Omdat $E(\underline{x}) = \mu$ een constante is, geldt

$$\begin{aligned} Var(\underline{x}) &= E((\underline{x}-\mu)^2) = E(\underline{x}^2 - 2\mu\underline{x} + \mu^2) \\ &= E(\underline{x}^2) - 2\mu E(\underline{x}) + \mu^2 = E(\underline{x}^2) - \mu^2. \end{aligned}$$

□

GEVOLG: als $\underline{x}$ een discrete stochastische variabele is die de waarden x_i met kansen resp. p_i aanneemt, dan is

$$Var(\underline{x}) = \sum_i {x_i}^2 p_i - \left(\sum_i x_i p_i\right)^2$$

en als $\underline{x}$ een continue stochastische variabele is met kansdichtheidsfunctie $f_{\underline{x}}(x)$, dan is

$$Var(\underline{x}) = \int_{-\infty}^{\infty} x^2 f_{\underline{x}}(x)dx - \left(\int_{-\infty}^{\infty} x f_{\underline{x}}(x)dx\right)^2.$$

Stelling 4.7 *Voor een willekeurige stochastische variabele $\underline{x}$ met verwachting μ en standaardafwijking σ geldt dat de stochastische variabele*

$$\underline{z} = \frac{\underline{x}-\mu}{\sigma}$$

verwachting 0 en standaardafwijking 1 heeft. Men noemt $\underline{z}$ wel de gestandaardiseerde vorm *van $\underline{x}$.* *gestandaardiseerde vorm*

BEWIJS: Op grond van de Stellingen 4.2 en 4.4 geldt

$$E(\underline{z}) = \frac{1}{\sigma}(E(\underline{x}) - \mu) = 0$$

en op grond van Stelling 35.1 is

$$Var(\underline{z}) = \frac{1}{\sigma^2} Var(\underline{x}) = 1.$$

□

In de volgende voorbeelden berekenen we de varianties van de in de vorige paragrafen behandelde stochastische variabelen.

Voorbeeld 4.8 *Werpen met een zuivere dobbelsteen.*

Als $\underline{x}$ de uitkomst is bij het werpen met een zuivere dobbelsteen, dan geldt $\mu = 7/2$, en dus is

$$Var(\underline{x}) = \frac{1}{6}\sum_{i=1}^{6} i^2 - \left(\frac{7}{2}\right)^2 = \frac{35}{12}.$$

Voorbeeld 4.9 *Bernoulli-experiment.*

Bij een Bernoulli-experiment met kans p op 'succes' en een stochastische variabele $\underline{x}$ die gedefinieerd wordt door $\underline{x} = 1$ bij 'succes', en $\underline{x} = 0$ bij 'mislukking', is de verwachting p (zie Voorbeeld 4.2). Dan geldt dus voor de variantie

$$Var(\underline{x}) = (1^2 \cdot p + 0^2 \cdot (1-p)) - p^2 = p(1-p).$$

Voorbeeld 4.10 *Binomiale verdeling.*

Laat $\underline{x}$ binomiaal verdeeld zijn met parameters n en p, dat wil zeggen dat

$$P(\underline{x} = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

voor $k = 0, 1, \ldots, n$. De verwachting $E(\underline{x}) = \mu$ van $\underline{x}$ is dan np (zie Voorbeeld 4.3). Voor het bepalen van de variantie voeren we als eerste stap net zo'n soort berekening uit als in Voorbeeld 4.3, maar nu voor de verwachting van $\underline{x}(\underline{x}-1)$.

$$\begin{aligned} E(\underline{x}(\underline{x}-1)) &= \sum_{k=0}^{n} k(k-1)\binom{n}{k} p^k (1-p)^{n-k} \\ &= n(n-1)p^2 \sum_{k=2}^{n} \binom{n-2}{k-2} p^{k-2}(1-p)^{(n-2)-(k-2)} \\ &= n(n-1)p^2 \sum_{j=0}^{n-2} \binom{n-2}{j} p^j (1-p)^{(n-2)-j} \\ &= n(n-1)p^2. \end{aligned}$$

Ter toelichting op de bovenstaande afleiding merken we het volgende op. In de som op de eerste regel zijn de eerste twee termen nul, dus die konden we weglaten. We hebben verder gebruikt dat

$$\binom{n}{k} = \frac{n(n-1)}{k(k-1)}\binom{n-2}{k-2} \quad \text{voor } n, k \geq 2.$$

Uit de zojuist afgeleide betrekking volgt nu dat

$$\begin{aligned} Var(\underline{x}) &= E(\underline{x}^2) - \mu^2 = E(\underline{x}(\underline{x}-1)) + E(\underline{x}) - \mu^2 \\ &= n(n-1)p^2 + np - (np)^2 = np(1-p). \end{aligned}$$

Voorbeeld 4.11 *De variantie van de uniforme verdeling.*

De kansdichtheidsfunctie van de uniforme verdeling op het interval $[a, b]$ wordt gegeven door

$$f(x) = \frac{1}{b-a} \quad \text{als} \quad a \leq x \leq b$$

en $f(x) = 0$ elders. Als de stochastische variabele $\underline{x}$ deze verdeling heeft, geldt, zoals we al bewezen hebben in Voorbeeld 4.4, $\mu = (a+b)/2$. Verder is

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

en een korte berekening geeft

$$Var(\underline{x}) = E(\underline{x}^2) - \mu^2 = \frac{(b-a)^2}{12}.$$

Merk op dat de standaardafwijking $\sigma(\underline{x}) = \sqrt{Var(x)}$ van de uniforme verdeling dus evenredig is met de lengte van het interval.

Opgave 4.9 *Bereken de variantie voor de stochastische variabelen uit de Opgaven 3.1 (a) en (b), 3.2, 3.3 en 3.4.*

Opgave 4.10 *De stochastische variabele $\underline{x}$ is uniform verdeeld op het interval $[-1, 1]$. Bereken de variantie en de standaardafwijking van de stochastische variabele $\underline{y} = \underline{x}^4$ (vergelijk ook Opgave 4.5).*

Opgave 4.11 *Toon aan dat de stochastische variabele $\underline{x}$ uit Opgave 3.8 geen eindige variantie heeft.*

Voorbeeld 4.12 *De variantie van de normale verdeling.*

Voor de variantie van de normale verdeling met parameters μ en σ gebruiken we dat we in Voorbeeld 4.5 al hebben afgeleid dat $E(\underline{x}) = \mu$. Met de substitutie $y = (x - \mu)/\sigma$ en met behulp van partiële integratie volgt dan

$$\begin{aligned}
Var(\underline{x}) &= E((\underline{x} - \mu)^2) = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sigma\sqrt{2\pi}} \, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 \, e^{-\frac{1}{2}y^2} dy \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (-y) d(e^{-\frac{1}{2}y^2}) \\
&= \left[\frac{\sigma^2}{\sqrt{2\pi}} (-y) e^{-\frac{1}{2}y^2} \right]_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2} dy \\
&= 0 + \frac{\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-(y/\sqrt{2})^2} d(y/\sqrt{2}) = \sigma^2
\end{aligned}$$

Bijgevolg is de standaardafwijking van de normale verdeling met parameters μ en σ gelijk aan σ.

Voorbeeld 4.13 *De negatief-exponentiële verdeling.*

In Voorbeeld 4.10 is afgeleid dat

$$E(\underline{x}) = \frac{1}{\lambda}$$

als $\underline{x}$ negatief-exponentiëel verdeeld is met parameter λ. We berekenen nu $E(\underline{x}^2)$ (vergelijk ook Opgave 4.6):

$$E(\underline{x}^2) = \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx = -\int_0^{\infty} x^2 de^{-\lambda x}$$

$$\begin{aligned} &= -\left[x^2 e^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x} 2x dx \\ &= 0 - \frac{2}{\lambda}\int_0^\infty x de^{-\lambda x} \\ &= -\frac{2}{\lambda}\left[x e^{-\lambda x}\right]_0^\infty + \frac{2}{\lambda}\int_0^\infty e^{-\lambda x} dx \\ &= 0 - \frac{2}{\lambda^2}\left[e^{-\lambda x}\right]_0^\infty = \frac{2}{\lambda^2} \end{aligned}$$

Voor de variantie geldt dus

$$Var(\underline{x}) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

en voor de standaardafwijking

$$\sigma(\underline{x}) = \frac{1}{\lambda}.$$

Bij de negatief-exponentiële verdeling zijn de verwachting en de standaardafwijking dus aan elkaar gelijk.

Voorbeeld 4.14 *Tussentijden bij een aankomstproces.*

In veel gevallen waarin sprake is van gebeurtenissen die 'toevallig' optreden, kan men de tijdsduur tussen twee opvolgende gebeurtenissen goed modelleren met behulp van de negatief-exponentiële verdeling. Een voorbeeld is de tussentijd tussen de aankomst van twee opvolgende klanten in een postkantoor. De parameter λ kiest men dan zo dat de waargenomen gemiddelde tussenaankomsttijd overeenkomt met de verwachting $1/\lambda$. Stel bijvoorbeeld dat men uit ervaring weet dat er op een bepaald postkantoor 's woensdags tussen 14 uur en 16 uur gemiddeld 90 klanten komen. Gemiddeld arriveren er dan 3 klanten per 4 minuten, zodat de gemiddelde tussentijd tussen de aankomst van twee opvolgende klanten 4/3 minuut bedraagt. Nemen we een minuut als tijdseenheid, dan ligt het voor de hand om de situatie te modelleren met behulp van een stochastische variabele $\underline{t}$ die de tussenaankomsttijd voorstelt, en die negatief-exponentiëel verdeeld is met parameter $\lambda = 3/4$, want de verwachting van $\underline{t}$ is dan $1/\lambda = 4/3$.

Hoe groot is in dit model nu de kans dat er meer dan 2 minuten verloopt tussen de aankomst van twee opvolgende klanten? Als variabele nemen we de tijd t, uitgedrukt in minuten. De verdelingsfunctie van $\underline{t}$ is $F(t) = 1 - e^{-(3/4)t}$, en dus is

$$P(\underline{t} > 2) = 1 - P(\underline{t} \leq 2) = 1 - F(2) = 1 - (1 - e^{-(3/2)}) \approx 0.223$$

Opgave 4.12 *Een Geigerteller registreert bij een licht-radioactief preparaat gemiddeld 10 klikken per minuut. Bereken met behulp van een negatief-exponentieel verdeelde stochastische variabele de kans dat de tijdsduur tussen twee opvolgende klikken minder dan 10, maar meer dan 5 seconden bedraagt.*

Hoofdstuk 5

De Centrale Limietstelling

5.1 Sommen en gemiddelden

In Hoofdstuk 1 hebben we een kansexperiment gedefinieerd als een experiment waarvan de uitkomst van het toeval afhangt, en dat we in principe net zo vaak als we willen onder dezelfde omstandigheden kunnen herhalen. Voorbeelden zijn het werpen met een munt of een dobbelsteen, het bepalen van een randomgetal met behulp van een randomgenerator of het meten van de brandtijd van een in massaproductie vervaardigde gloeilamp. Vaak zijn we in zulke situaties geïnteresseerd in de som of het gemiddelde van de uitkomsten van een serie herhalingen van zo'n experiment. De binomiale verdeling kan men bijvoorbeeld opvatten als de verdeling van de som van het aantal successen in een rij van n herhalingen van een Bernoulli-experiment met succeskans p. Een ander voorbeeld in Hoofdstuk 3 was het bepalen van de gemiddelde brandtijd per lamp in een doos van 12 lampen (Voorbeeld 3.3).
Laat in het algemeen een kansexperiment gegeven zijn met een zekere verdelingsfunctie $F(x)$, een eindige verwachting μ en een eindige standaardafwijking σ. Onder de stochastische variabelen $\underline{x}_1$, $\underline{x}_2$, ..., $\underline{x}_n$ verstaan we de uitkomsten van n herhalingen van het experiment. Elke $\underline{x}_i$ $(i = 1, \ldots, n)$ heeft dus ook $F(x)$ als verdelingsfunctie, μ als verwachting en σ als standaardafwijking. Men zou zich overigens nog af kunnen vragen wat men precies moet verstaan onder het 'herhalen van een experiment onder dezelfde omstandigheden'. Filosofisch gezien kan men de opvatting verdedigen dat zoiets onmogelijk is omdat de omstandigheden altijd enigszins van elkaar verschillen. Maar natuurlijk gaat het hier weer om idealisaties, om wiskundige modellen waarin men aan al deze intuïtieve termen een precieze betekenis kan geven. In wiskundige termen spreekt men dan over stochastische variabelen $\underline{x}_1$, $\underline{x}_2$, ..., $\underline{x}_n$ die *onderling onafhankelijk* en *identiek verdeeld* zijn. Wij volstaan in deze cursus met een intuïtieve benadering waarbij we de finesses van de wiskundige modelvorming buiten beschouing laten.

onderling
onafhankelijk
identiek
verdeeld

De situatie van n onderling onafhankelijke identiek verdeelde stochastische variabelen doet zich ook voor wanneer men een *aselecte steekproef met teruglegging* ter grootte n doet uit een zekere kansverdeling. Daarmee bedoelen we eigenlijk weer dat we een kansexperiment, met daarbij het bepalen van steeds dezelfde stochastische variabele, n maal uitvoeren. De toevoeging 'met teruglegging' bij de steekproef slaat op het veel gebruikte voorbeeld waarbij men zonder te kijken (we zeggen ook wel 'aselect') een bal uit een vaas met verschillend gekleurde, maar verder niet van elkaar te onderscheiden ballen trekt. Doet men dit niet met teruglegging, dan is na een tijdje de vaas leeg, en het is dan ook duidelijk dat de trekkingen niet meer als onderling onafhankelijke stochastische variabelen kunnen worden beschouwd. Bij een aselecte steekproef met teruglegging wordt de getrokken bal telkens in de vaas

teruggelegd, waarna men de ballen goed mengt alvorens opnieuw te trekken. We zullen dit voorbeeld in de Hoofdstuk 6 verder uitwerken.
In andere situaties waarbij men wel over 'aselecte steekproeven' of 'aselecte trekkingen' spreekt, heeft de term 'met teruglegging' geen betekenis; denk bijvoorbeeld aan het werpen met een dobbelsteen of aan het in serie produceren van industriële producten. Essentieel is echter steeds dat we aannemen dat iedere uitvoering van zo'n kansexperiment 'onder dezelfde omstandigheden' plaatsvindt.

5.2 De $\sqrt{n}$-wetten

Uitgaande van de onderling onafhankelijke en identiek verdeelde stochastische variabelen $\underline{x}_1$, $\underline{x}_2$, ..., $\underline{x}_n$ vormen we nu twee nieuwe stochastische variabelen, te weten de som

$$\underline{s}_n = \underline{x}_1 + \ldots + \underline{x}_n$$

en het gemiddelde

$$\underline{m}_n = \frac{\underline{x}_1 + \ldots + \underline{x}_n}{n}.$$

Zoals we al in de vorige eenheid hebben opgemerkt, is het in het algemeen niet eenvoudig om de verdelingsfuncties van $\underline{s}_n$ en $\underline{m}_n$ uit te drukken in termen van de verdelingsfunctie $F(x)$. Wel kan men de volgende stelling afleiden voor de verwachtingen, de varianties en de de standaardafwijkingen van $\underline{s}_n$ en $\underline{m}_n$. We zullen hier overigens geen bewijs geven van deze stelling.

Stelling 5.1 *Met de notaties als boven geldt*

$$E(\underline{s}_n) = n\mu \qquad en \qquad E(\underline{m}_n) = \mu,$$

$$Var(\underline{s}_n) = n\sigma^2 \qquad en \qquad Var(\underline{m}_n) = \frac{\sigma^2}{n}$$

en

$$\sigma(\underline{s}_n) = \sqrt{n}\,\sigma \qquad en \qquad \sigma(\underline{m}_n) = \frac{\sigma}{\sqrt{n}}.$$

$\sqrt{n}$-wetten

De uitspraken over de standaardafwijkingen staan bekend als de $\sqrt{n}$-wetten. We wijzen op de mogelijke verwarring die de notatie voor de standaardafwijkingen kan opleveren: $\sigma(\underline{s}_n)$ en $\sigma(\underline{m}_n)$ gebruiken we voor de standaardafwijkingen van $\underline{s}_n$ en $\underline{m}_n$, terwijl we met een losse σ de standaardafwijking aangeven van iedere $\underline{x}_i$ afzonderlijk. Misschien zou in dit opzicht een notatie als $SA(\underline{s}_n)$ beter zijn dan $\sigma(\underline{s}_n)$, maar die is niet zo gebruikelijk.
Het is interessant om te zien wat er met het gemiddelde $\underline{m}_n$ gebeurt als $n \to \infty$. De verwachting blijft μ maar de variantie, en dus ook de standaardafwijking, gaat naar 0. Dit is in overeenstemming met onze intuïtie dat het gemiddelde van een zeer groot aantal onafhankelijke trekkingen uit een vaste kansverdeling naar de verwachting zal tenderen.

5.3 De Centrale Limietstelling

Nog interessanter is het om de som $\underline{s}_n = \underline{x}_1 + \ldots + \underline{x}_n$ te standaardiseren in de zin van Stelling 4.7. We trekken er dus de verwachting van af, en delen het resultaat door de standaardafwijking. We krijgen dan

$$\underline{z}_n = \frac{\underline{s}_n - n\mu}{\sigma\sqrt{n}} = \frac{\underline{x}_1 + \ldots + \underline{x}_n - n\mu}{\sigma\sqrt{n}}.$$

Precies dezelfde uitdrukking krijgt u wanneer u het gemiddelde $\underline{m}_n = (1/n)(\underline{x}_1 + \ldots + \underline{x}_n)$ standaardiseert: deel eenvoudig teller en noemer door n, dan ziet u het. De stochastische variabele $\underline{z}_n$ heeft volgens Stelling 4.7 verwachting 0 en standaardafwijking 1. Maar veel verrassender, en ook veel moeilijker om te bewijzen, is dat we ook over de verdelingsfunctie $F_{\underline{z}_n}(x)$ van $\underline{z}_n$ een uitspraak kunnen doen. Men kan namelijk bewijzen dat de rij verdelingsfuncties $F_{\underline{z}_n}(x)$ voor $n \to \infty$ convergeert naar de verdelingsfunctie $\Phi(x)$ van de standaardnormale verdeling, *onafhankelijk van de verdeling van de oorspronkelijke stochastische variabelen* $\underline{x}_i$! Dit is de inhoud van de zogenaamde *Centrale Limietstelling*, een stelling die het grote belang van de normale verdeling onderstreept. We formuleren die stelling nu in haar algemene gedaante, maar ook deze stelling zullen we hier niet bewijzen.

Stelling 5.2 (Centrale Limietstelling) *Als gegeven is een rij onderling onafhankelijke identiek verdeelde stochastische variabelen* $\underline{x}_1$, $\underline{x}_2$, ..., *allemaal met eindige verwachting* μ *en eindige standaardafwijking* σ, *dan geldt voor de verdelingsfuncties* $F_{\underline{z}_n}(x)$ *van de stochastische variabelen*

$$\underline{z}_n = \frac{\underline{x}_1 + \ldots + \underline{x}_n - n\mu}{\sigma\sqrt{n}}$$

dat

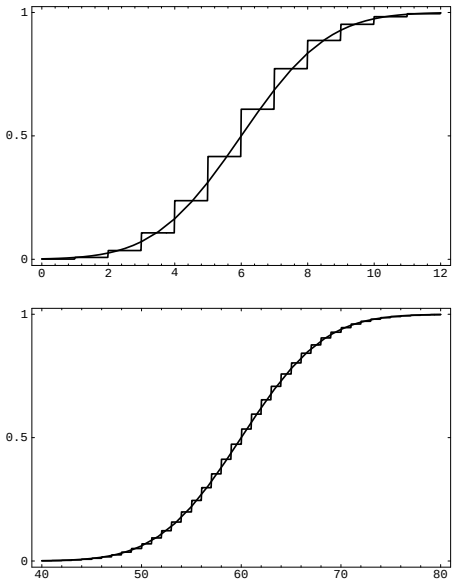
$$\lim_{n\to\infty} F_{\underline{z}_n}(x) = \Phi(x),$$

waarbij $\Phi(x)$ *de verdelingsfunctie is van de standaardnormale verdeling.*

Een gevolg van de Centrale Limietstelling is dat men voor grote waarden van n de verdelingsfuncties van de som $\underline{s}_n$ en het gemiddelde $\underline{m}_n$ 'goed kan benaderen' door de normale verdelingsfuncties $N_{n\mu,\,\sigma\sqrt{n}}(x)$ resp. $N_{\mu,\,\sigma/\sqrt{n}}(x)$. In Figuur 5.1 ziet u hiervan een illustratie; in de volgende paragraaf volgt een uitgebreide toelichting. De Centrale Limietstelling speelt ook een belangrijke rol in de schattings- en toetsingstheorie. Daarvan geven we voorbeelden in Hoofdstuk 6.
We besluiten deze paragraaf met de opmerking dat men kan bewijzen dat in het geval dat de stochastische variabelen $\underline{x}_i$ zelf al normaal verdeeld zijn, hun som $\underline{s}_n$ en hun gemiddelde $\underline{m}_n$ ook normaal verdeeld zijn. In dat geval is de Centrale Limietstelling dus in feite overbodig: de verdelingsfuncties $F_{\underline{z}_n}$ van de gestandaardiseerde vormen van $\underline{s}_n$ en $\underline{m}_n$ zijn dan niet slechts bij benadering, maar volledig identiek gelijk aan de verdelingsfunctie $\Phi(x)$ van de standaardnormale verdeling.

5.4 Normale benaderingen

De Centrale Limietstelling werd voor het eerst bewezen in 1733 door A. de Moivre in het geval van de binomiale verdeling, die immers opgevat kan worden als de som van n onafhankelijke Bernoulli-variabelen die ieder met kans p de waarde 1, en met kans $(1-p)$ de waarde 0 aannemen (zie Voorbeeld 4.2 op pagina 38). In dat geval geldt dus voor iedere $\underline{x}_i$ dat $\mu = p$ en $\sigma = \sqrt{p(1-p)}$. De binomiale verdeling $\underline{x}_1 + \ldots + \underline{x}_n$ heeft dus verwachting np en standaardafwijking $\sqrt{np(1-p)}$, zoals we ook al eerder op een directe wijze afgeleid hebben, en de bijbehorende normale benadering heeft dus de verdelingsfunctie $N_{np,\,\sqrt{np(1-p)}}(x)$.
Ter illustratie ziet u in Figuur 5.1 in één figuur de verdelingsfunctie getekend van de binomiale verdeling en de bijbehorende normale verdeling. Er worden twee gevallen genomen: de bovenste grafieken behoren bij $n = 20$ en $p = 0.3$ en de onderste grafieken bij $n = 200$ en $p = 0.3$. Duidelijk is te zien dat de twee verdelingsfuncties voor grote n vrijwel samenvallen. Merk ook op dat we, om duidelijke figuren te krijgen, niet het gehele domein van de betreffende binomiale verdelingsfuncties in



Figuur 5.1: Verdelingsfuncties $B_{n,p}(x)$ en $N_{np,\sqrt{np(1-p)}}(x)$ voor $n = 20$ en $p = 0.3$ (bovenste grafieken) en voor $n = 200$ en $p = 0.3$ (onderste grafieken).

beeld hebben gebracht: voor $n = 20$ is slechts het interval $[0, 12]$ getekend, en voor $n = 200$ slechts het interval $[40, 80]$. Buiten die intervallen zijn de kansen verwaarloosbaar klein.
In Opgave 2.13 heeft u enige resultaten afgeleid voor de normale verdeling met parameters μ en σ. U heeft toen laten zien dat de kans op een waarde tussen $\mu - \sigma$ en $\mu + \sigma$ ruim 68 procent is, dat de kans op een waarde tussen $\mu - 2\sigma$ en $\mu + 2\sigma$ ruim 95 procent is, en dat de kans op een waarde tussen $\mu - 3\sigma$ en $\mu + 3\sigma$ ruim 99,7 procent is. Deze resultaten gebruikt men vaak als *vuistregels* in gevallen waarbij men een onbekende kansverdeling door een normale verdeling benadert. We geven ze hieronder nogmaals overzichtelijk weer; zie ook Figuur B.2 op bladzijde 68.

vuistregels

Laat $\underline{x}$ een normaal verdeelde stochastische variabele zijn met verwachting μ en standaardafwijking σ, dan geldt:

$$\begin{aligned} P(\underline{x} \in [\mu - \sigma, \mu + \sigma]) &\approx 0.683 \\ P(\underline{x} \in [\mu - 2\sigma, \mu + 2\sigma]) &\approx 0.954 \\ P(\underline{x} \in [\mu - 3\sigma, \mu + 3\sigma]) &\approx 0.997 \end{aligned}$$

De in deze vuistregels genoemde intervallen worden kortweg aangeduid als resp. het σ-, het 2σ- en het 3σ-interval.

Opgave 5.1 *Een gloeilampenfabriek produceert gloeilampen in massaproductie met een kans $p = 0.03$ op defecte exemplaren. Ze worden opgeslagen in containers van 1000 stuks. Het aantal defecte exemplaren in zo'n container vat men op als een stochastische variabele.*

a. Bepaal hiervan de verwachting en de standaardafwijking.

b. De fabriek beweert dat de kans op een container met 48 of meer defecte exemplaren minder dan een half procent is. Gebruik de bovengenoemde vuistregel om te onderzoeken of dat een verantwoorde uitspraak is.

Opgave 5.2 *Een randomgenerator produceert getallen tussen 0 en 1 volgens de uniforme verdeling. Men laat de generator telkens 100 getallen produceren, en neemt daarvan het gemiddelde. Zo verkrijgt men een stochastische variabele* $\underline{x}$.

a. Bepaal van $\underline{x}$ *de verwachting en de standaardafwijking.*

b. Bepaal met behulp van de bovengenoemde vuistregels een interval $[a, b]$ *zo, dat de kans dat* $\underline{x}$ *binnen dat interval valt, ongeveer 95 procent is.*

Opgave 5.3 *Iemand wil* n *maal met een zuivere munt gooien, waarbij* n *zo groot is gekozen, dat het aantal malen* kop*, gedeeld door het totale aantal worpen, met een kans van minstens 95 procent een getal uit het interval* $[0.45, 0.55]$ *zal zijn. Gebruik een van de bovengenoemde vuistregels om te berekenen hoe groot zij* n *dan minstens moet kiezen.*

Opgave 5.4 *Een vulmachine voor jampotten vult potten met een gewicht aan jam dat normaal verdeeld is met gemiddelde van* 501 *gram en een standaardafwijking van 3 gram. De potten worden in dozen van 25 stuks verpakt. Onder* $\underline{G}$ *verstaat men de stochastische variabele die het totale gewicht aan jam in zo'n doos voorstelt. Wat is de kansverdeling van* $\underline{G}$*?*

Hoofdstuk 6

Schatten en Toetsen

6.1 Schattingstheorie

In dit hoofdstuk zult u aan de hand van een aantal voorbeelden zien hoe men met behulp van aselecte steekproeven tot verantwoorde schattingen kan komen van onbekende parameters van een kansverdeling, of hoe men op een verantwoorde wijze hypothesen omtrent onbekende parameters kan toetsen. We beginnen met een inleiding in de schattingstheorie.

6.1.1 Witte en zwarte ballen in een vaas

Stel dat er in een vaas w witte en z zwarte ballen zitten die alleen in kleur van elkaar verschillen. Wanneer men aselect, dat wil zeggen zonder te kijken, een bal trekt, is de kans op een witte bal gelijk aan $p = w/(z + w)$. Noemt men het trekken van een witte bal een 'succes' en het trekken van een zwarte bal een 'mislukking', dan is het duidelijk dat er hier sprake is van een Bernoulli experiment dat men willekeurig vaak kan herhalen, mits men telkens de getrokken bal weer teruglegt en de ballen goed mengt alvorens opnieuw te trekken.
Stel nu dat w en z, en dus ook p, onbekend zijn, maar dat men een zeker aantal aselecte trekkingen met terugleggen uit de vaas mag doen. Het is duidelijk dat men dan nooit enige informatie kan verkrijgen over de precieze aantallen w en z, maar dat men wel enige informatie krijgt over de fractie $p = w/(w+z)$. Echter, *zekerheid* omtrent p is nooit te verkrijgen, hoeveel trekkingen men ook uitvoert! Door het aantal trekkingen op te voeren zal men wel steeds 'betrouwbaarder' schattingen voor p kunnen opstellen.
Laten we nu aannemen dat we 25 aselecte trekkingen mogen uitvoeren. De stochastische variabele die het aantal getrokken witte ballen aangeeft, noemen we $\underline{x}$. Stel dat we 10 maal een witte bal, en 15 maal een zwarte bal trekken. Stel dus dat $\underline{x} = 10$. Wat voor verantwoorde uitspraken over p kunnen we nu doen?
Het ligt voor de hand om de fractie $10/25 = 0.4$ als schatting voor p te hanteren. Maar als we het daarbij laten, gooien we de belangrijke informatie weg dat het hierbij om 25 trekkingen ging. Hoe kunnen we die informatie gebruiken? Waar we in feite over zouden willen beschikken, is een *betrouwbaarheidsinterval*, dat wil zeggen een interval $[p_l, p_r]$ waarvan we kunnen zeggen dat het 'vrijwel zeker' is dat de onbekende p-waarde er in ligt. We verwachten ook dat we zo'n betrouwbaarheidsinterval 'nauwer' kunnen maken naarmate we het aantal trekkingen opvoeren.

betrouwbaarheidsinterval

Dat brengt ons direct op een arbitrair aspect: wat verstaan we onder 'vrijwel zeker'? We zullen proberen dat nader te kwantificeren.
We weten dat we in het onderhavige geval van 25 trekkingen in feite te maken hebben met een binomiaal verdeelde stochastische variabele $\underline{x}$ met als parameters $n = 25$

en de onbekende slagingskans p. Voor iedere p is de grafiek van de verdelingsfunctie $B_{25,p}(x)$ een trapfunctie die van 0 (voor $x = 0$) naar 1 (voor $x = 25$) stijgt. Voor iedere p geeft $B_{25,p}(10)$ de kans weer dat er onder de 25 getrokken ballen *ten hoogste* 10 witte exemplaren zitten. Als p vlak bij 1 zou liggen, zou deze kans heel klein zijn, want naar verwachting zijn de meeste getrokken ballen dan wit. Toch is het niet uitgesloten dat er dan slechts 10 of minder witte ballen getrokken worden. Voor $p = 0.8$ geldt bijvoorbeeld dat

$$P(\underline{x} \leq 10) = B_{25,0.8}(10) \approx 0.00001356$$

In de situatie waarin we op grond van het geconstateerde aantal van 10 witte ballen toch een verantwoorde uitspraak willen doen, zullen we p-waarden waarvoor de kans dermate klein is, als te onwaarschijnlijk verwerpen. Maar waar leggen we precies de grens? Dat is een kwestie van afspraak. Een veel gebruikte grenswaarde is 0.05. We zullen in dat geval alle p-waarden waarvoor

$$B_{25,p}(10) < 0.05$$

geldt, verwerpen. De uiterste waarde hierbij is de p waarvoor

$$B_{25,p}(10) = 0.05 \tag{1}$$

en met de computer kunnen we die vergelijking naar p oplossen. De uitkomst is, op drie decimalen afgerond, $p = 0.583$, en alle grotere p-waarden keuren we dus af.
We hebben hiermee de rechtergrens van ons betrouwbaarheidsinterval gevonden. De linkergrens vinden we op een soortgelijke wijze. We kijken daartoe wat er gebeurt als p heel dicht bij 0 ligt. Neem bijvoorbeeld $p = 0.1$. De kans op 10 of meer witte ballen is dan heel klein, namelijk

$$P(\underline{x} \geq 10) = 1 - P(\underline{x} \leq 9) = 1 - B_{25,0.1}(9) \approx 0.00007898$$

Ook hier gebruiken we weer een drempelwaarde, bijvoorbeeld diezelfde 0.05. We zullen ook alle p-waarden afkeuren waarvoor

$$1 - B_{25,p}(9) < 0.05$$

geldt, dat wil zeggen waarvoor

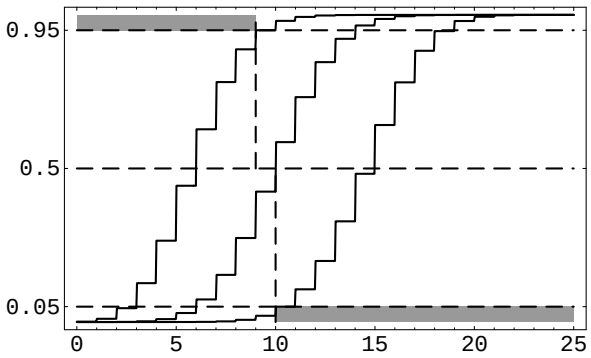
$$B_{25,p}(9) > 0.95.$$

De grenswaarde p waarvoor

$$B_{25,p}(9) = 0.95 \tag{2}$$

vinden we weer met de computer: in drie decimalen nauwkeurig blijkt dit $p = 0.236$ te zijn. Alle kleinere p's worden afgekeurd.
In Figuur 6.1 is de situatie in beeld gebracht. U ziet daarin de grafieken van de verdelingsfunctie $B_{25,p}(x)$ voor drie waarden van p. De middelste grafiek behoort bij $p = 0.4$ (de 'voor de hand liggende' schatting 10/25). Voor de linkergrafiek geldt $B_{25,p}(9) = 0.95$ en voor de rechtergrafiek geldt $B_{25,p}(10) = 0.05$. Slechts de waarden van p waarvoor de gehele grafiek van de verdelingsfunctie buiten de twee grijze gebieden blijft, liggen in het betrouwbaarheidsinterval.
We hebben dus als betrouwbaarheidsinterval het interval $[0.236, 0.583]$ gevonden, en we weten nu dat wanneer p *niet* in dit interval zou liggen, de kans op de gerealiseerde waarde $\underline{x} = 10$ kleiner zou zijn dan 0.05. Dat is de precieze inhoud van de bewering dat p 'waarschijnlijk' binnen het berekende betrouwbaarheidsinterval ligt.

Opgave 6.1 *Ga na dat de vergelijkingen (1) en (2) 25ste-graads vergelijkingen zijn. (In Blok 10 worden technieken behandeld om zulke vergelijkingen numeriek op te lossen met behulp van een computer.)*



Figuur 6.1: Binomiale verdelingsfuncties $B_{25,p}(x)$ voor $p = 0.236$ (links), $p = 0.400$ (midden) en $p = 0.583$ (rechts).

6.1.2 Onbetrouwbaarheidsdrempels

Hoe zijn we hierboven tot de drempelwaarde 0.05 gekomen? Dat was gewoon een arbitraire keuze. Men noemt die waarde meestal een een *onbetrouwbaarheidsdrempel*. Dat is dus een getal α dicht bij 0 met de eigenschap dat de kans op het waargenomen steekproefresultaat, onder de veronderstelling dat de geschatte parameter *niet* in het betrouwbaarheidsinterval ligt, minder is dan α. In het hierboven behandelde voorbeeld hebben we $\alpha = 0.05$ gekozen, maar men werkt ook wel met andere waarden, bijvoorbeeld met $\alpha = 0.01$.

onbetrouwbaarheidsdrempel

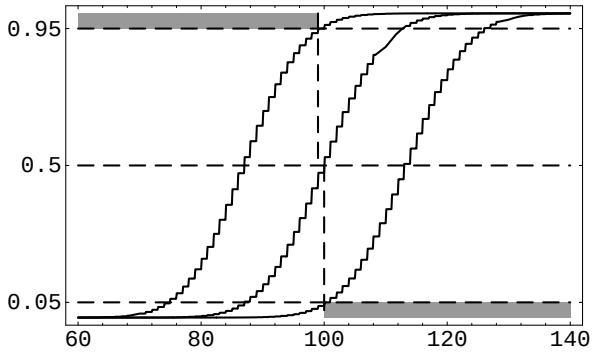
Opgave 6.2 *Waar of niet waar: de keuze $\alpha = 0.01$ geeft een kleiner betrouwbaarheidsinterval.*

De invloed van het aantal trekkingen op de grootte van het betrouwbaarheidsinterval hebben we nog niet onderzocht. Ter illustratie hiervan nemen we nu bij dezelfde $\alpha = 0.05$ een aantal van $n = 250$ trekkingen. Om de vergelijking met het hierboven behandelde voorbeeld goed tot haar recht te laten komen, veronderstellen we dat de fractie van het aantal successen dezelfde is, dus dat er nu 100 witte ballen getrokken zijn. Figuur 6.2 illustreert de resultaten van de bijbehorende berekeningen. Als betrouwbaarheidsinterval krijgen we nu het aanzienlijk kleinere interval [0.348, 0.454].
De mogelijkheid om in het bovenbehandelde voorbeeld goede betrouwbaarheidsintervallen af te leiden, berustte enerzijds op het feit dat we bij een Bernoulli-experiment de kansverdeling van de som van het aantal successen kennen (die som is immers binomiaal verdeeld), en anderzijds op de eenvoudige ligging van de grafieken van de verdelingsfuncties $B_{n,p}(x)$ bij vaste n en variabele p. Het zijn trapfuncties die netjes naast elkaar liggen. Verder konden we de computer inschakelen om de op zichzelf lastige vergelijkingen numeriek op te lossen. We geven nog een voorbeeld van het berekenen van betrouwbaarheidsintervallen.

Voorbeeld 6.1 *Betrouwbaarheidsintervallen bij een munt.*

In Hoofdstuk 1 is in Tabel 1.1 verslag gedaan van een experiment van Buffon die bij 4040 worpen met een munt 2048 maal 'kop' gooide. Bij een betrouwbaarheidsdrempel van $\alpha = 0.05$ kunnen we het bijbehorende betrouwbaarheidsinterval berekenen. We moeten daartoe de volgende twee vergelijkingen oplossen:

$$B_{4040,p}(2048) = 0.05$$



Figuur 6.2: Binomiale verdelingsfuncties $B_{250,p}(x)$ voor $p = 0.348$ (links), $p = 0.400$ (midden) en $p = 0.454$ (rechts).

voor de rechtergrens, en

$$B_{4040,p}(2047) = 0.95$$

voor de linkergrens. Met behulp van een computeralgebrapakket vonden we, afgerond op drie decimalen, het betrouwbaarheidsinterval $[0.494, 0.520]$ met een breedte 0.026.

In dezelfde tabel is ook melding gemaakt van twee experimenten van Pearson. Bij het eerste experiment waren er 12000 worpen met 6019 malen 'kop'. Dit leidt tot het betrouwbaarheidsinterval $[0.4940, 0.5091]$ met een breedte 0.0151.

6.1.3 Het schatten van de verwachting van een normaal verdeelde variabele

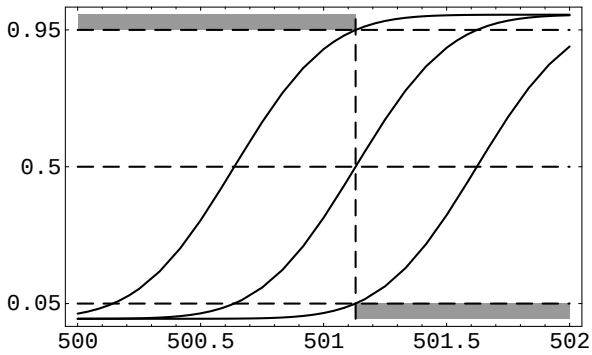
Bij sommige productieprocessen heeft men te maken met een normaal verdeelde stochastische variabele waarvan men wel de variantie kent, maar niet de verwachting. Denk bijvoorbeeld aan een machine die jampotten vult en die na een onderhoudsbeurt of een storing opnieuw moet worden afgeregeld. De variantie in de vulgewichten zal niet of nauwelijks afhangen van de precieze afstelling van het 'gemiddelde gewicht'; dat laatste moet echter wel heel precies worden bijgesteld.
Men zou nu bij die afregeling via een voldoende grote, maar anderzijds niet al te grote steekproef de instelling van de machine willen controleren. Wanneer men ervan uit mag gaan dat het vulgewicht een normaal verdeelde stochastische variabele is met een bekende standaardafwijking σ en een onbekende verwachtingswaarde μ, kan men aan de hand van de uitslag van zo'n steekproef bij een gegeven onbetrouwbaarheidsdrempel een betrouwbaarheidsinterval voor μ opstellen.
Stel bijvoorbeeld dat men weet dat de standaardafwijking σ van het vulgewicht 3 gram is en stel dat men in een aselecte steekproef van $n = 100$ potten als gemiddeld vulgewicht $501,13$ gram vindt. In Leereenheid 35 is vermeld dat het gemiddelde dan $N_{\mu,\,\sigma/\sqrt{n}} = N_{\mu,0.3}$ verdeeld is. Om bij een onbetrouwbaarheidsdrempel $\alpha = 0.05$ een betrouwbaarheidsinterval voor μ te construeren, laten we de computer de beide vergelijkingen

$$N_{\mu,0.3}(501.13) = 0.95$$

en

$$N_{\mu,0.3}(501.13) = 0.05$$

numeriek oplossen naar μ. We vinden resp. $\mu = 500.637$ en $\mu = 501.623$ zodat het betrouwbaarheidsinterval gegeven wordt door [500.637, 501.623]. Figuur 6.3 illustreert de situatie.



Figuur 6.3: Normale verdelingsfuncties $N_{\mu,0.3}(x)$ voor $\mu = 500.637$ (links), $\mu = 501.13$ (midden) en $\mu = 501.623$ (rechts).

We kunnen nog opmerken dat we met behulp van de in Hoofdstuk 1 afgeleide formule

$$N_{\mu,\sigma}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

de op te lossen vergelijkingen terug kunnen brengen tot vergelijkingen voor de standaardnormale verdeling. Vroeger was dit de enige praktisch bruikbare methode omdat men toen uitsluitend beschikte over tabellen voor de standaardnormale verdeling. Als voorbeeld ontlenen we aan zo'n tabel dat

$$\Phi(1.645) = 0.95$$

zodat de linkerrand van het betrouwbaarheidsinterval gevonden wordt uit

$$\frac{501.13-\mu}{0.3} = 1.645$$

hetgeen leidt tot $\mu = 500.6365$, in overeenstemming met de hierboven gevonden waarde (binnen de gebruikte nauwkeurigheid). Op een soortgelijke wijze kunnen we de rechterrand van het betrouwbaarheidsinterval verifiëren.

Opgave 6.3 *Voer deze berekening uit. (N.b.: u hoeft daarvoor geen nieuwe tabelwaarden te hanteren!)*

Opgave 6.4 *Het hier gevonden betrouwbaarheidsinterval ligt symmetrisch ten opzichte van het steekproefgemiddelde* 501.13. *Bij het trekken van ballen uit een vaas in de vorige paragraaf was dat niet het geval. Verklaar dit.*

6.2 Het toetsen van hypothesen

6.2.1 Twee soorten fouten

Nauw verwant met de problematiek van het schatten van parameters is die van het toetsen van hypothesen. In zekere zin is het meer een kwestie van beschouwingswijze dan van een geheel andere wiskundige techniek. Toetsen spelen bijvoorbeeld

een rol bij kwaliteitscontrole: een productieproces wordt periodiek gecontroleerd door middel van steekproeven. Bij welke uitslag van de steekproef moeten de instellingen van de machines worden bijgesteld? Of denk aan een keuringsdienst die partijen consumptieartikelen controleert op de aanwezigheid van schadelijke stoffen. De grootheden die getoetst worden, zijn in zulke gevallen meestal op te vatten als stochastische variabelen waarvan de waarden aan zekere specificaties dienen te voldoen. Kleine variaties zijn onvermijdelijk, maar ze moeten met een grote mate van zekerheid binnen bepaalde toleranties blijven.

Toch is er ook een principieel verschil tussen schatten en toetsen. Bij schatten voeren we *eerst* een steekproef uit, en daarna bepalen we aan de hand van de uitkomst daarvan een schatting van de onbekende parameter en een bijbehorend betrouwbaarheidsinterval. De volgorde is dus: eerst het experiment, dan de berekening. Bij toetsen is het net andersom. Hier gaan we *van tevoren* uit van een zekere hy-

nulhypothese

pothese – in de toetsingstheorie spreekt men altijd over de *nulhypothese*, notatie H_0 – omtrent de waarde van de onbekende parameter die men door middel van een steekproef wil toetsen. Maar ook hier gaat het niet om *zekerheid*, want die is in dit vak onbereikbaar, maar weer om *waarschijnlijkheid*. Het is mogelijk dat de toetsingsprocedure tot een onjuiste beslissing leidt; we kunnen slechts bereiken dat de kans daarop klein is.

Wat is in dit verband een onjuiste beslissing? Er zijn twee soorten onjuiste beslissingen: de eerste is dat de toetsingsprocedure leidt tot het verwerpen van de nulhypothese H_0 terwijl die toch geldig is. Dit noemt men een *fout van de eerste*

fout van de eerste soort

soort. Met andere woorden: *een fout van de eerste soort bestaat uit het ten onrechte verwerpen van de nulhypothese.*

Zoals gezegd: in de toetsingsopzet probeert men de kans op een fout van de eerste soort onder een bepaalde drempelwaarde te houden. Maar er is ook nog een andere ongewenste mogelijkheid, namelijk dat men H_0 niet verwerpt, terwijl die hypothese toch onjuist is. Zo'n fout – men spreekt in dit verband meestal van een *fout van de*

fout van de tweede soort

tweede soort – kan men ook onder bepaalde grenzen proberen te houden. Er zijn echter situaties denkbaar dat dit nauwelijks zinvol is, of dat het zelfs theoretisch is uitgesloten!

6.2.2 Criteria voor het verwerpen van de nulhypothese

Stel bijvoorbeeld dat men bij het werpen met een munt de nulhypothese hanteert dat de kans op 'kop' gelijk is aan $\frac{1}{2}$. Stel verder dat men bij het toetsen van die hypothese de munt een van tevoren vastgesteld aantal malen mag werpen. We zullen zo dadelijk zien dat we dan een interval kunnen opstellen met de eigenschap dat we de nulhypothese zullen verwerpen wanneer het aantal malen 'kop' buiten dat interval ligt. Het is echter niet onredelijk om aan te nemen dat de kans op 'kop' bij een echte munt nooit precies $\frac{1}{2}$ zal zijn, dus dat u in alle gevallen waarin u de nulhypothese niet verwerpt, een fout van de tweede soort maakt!

In de toetsingstheorie wordt uitgebreid op deze problematiek ingegaan; wij zullen ons hier slechts bezighouden met fouten van de eerste soort. Ons probleem is dus: gegeven een nulhypothese H_0, stel dan (vooraf!) criteria op waaronder men H_0 zal verwerpen.

Voorbeeld 6.2 *Kwaliteitscontrole.*

Stel dat we bij een Bernoulli experiment met slagingskans p de nulhypothese $p \geq 0.7$ willen toetsen. Het gaat bijvoorbeeld om een moeilijk industrieel productieproces waarbij, wanneer alle instellingen optimaal geregeld zijn, naar verwachting minstens 70 procent van de geproduceerde exemplaren aan alle specificaties voldoet. Van tijd tot tijd nemen we een steekproef van 25 exemplaren, en bepalen hoeveel exemplaren er

niet in orde zijn. Het aantal goedgekeurde exemplaren in de steekproef vatten we op als een stochastische variabele $\underline{x}$. We weten dat $\underline{x}$ dan binomiaal verdeeld is met $n = 25$ en een onbekende p. De nulhypothese H_0 is $p \geq 0.7$.

Zou de waarde van p inderdaad 0.7 of meer zijn, dan kijken we niet vreemd op wanneer 18 of meer van de 25 exemplaren in orde zijn. Ook een aantal van 17 of 16 goede exemplaren hoeft geen verwondering te wekken, maar zijn er slechts 3 of 4 exemplaren goed, dan zal men waarschijnlijk direct alarm slaan. Tussen 4 en 16 zit dus ergens een grens: een getal c met de eigenschap dat we moeten ingrijpen zodra de waarde van $\underline{x}$ kleiner dan of gelijk aan c is. Opnieuw hebben we dan te maken met een arbitrair element: we moeten weer een drempel kiezen: een getal α dicht bij 0 dat bepaalt waar we die kritische waarde c leggen. We kunnen bijvoorbeeld weer $\alpha = 0.05$ nemen.

Met behulp van de computer berekenen we een aantal waarden van $B_{25,0.7}(x)$. We vinden

$$B_{25,0.7}(13) = 0.0442 \quad \text{en} \quad B_{25,0.7}(14) = 0.0978$$

dus voor alle $x \leq 13$ en alle $p \geq 0.7$ geldt dat

$$B_{25,p}(x) < 0.05.$$

De kritische waarde c is blijkbaar 13, en als toetsingsprocedure spreken we af:

Verwerp H_0 wanneer $\underline{x} \leq 13$ is.

We weten dan dat de kans op een fout van de eerste soort, dat wil zeggen het verwerpen van H_0 in het geval dat $p \geq 0.7$, minder is dan $\alpha = 0.05$. In Figuur 6.4 wordt dit alles geïllustreerd.

Opgave 6.5 *Waar of niet waar: als $\underline{x} > 13$ dan is de kans dat $p \geq 0.7$ is, minstens 95 procent.*

Opgave 6.6 *Waar of niet waar: als $p \geq 0.7$ dan is de kans dat $\underline{x} > 13$ is, minstens 95 procent.*

Opgave 6.7 *Is het waar dat de keuze $\alpha = 0.01$ leidt tot een kleinere waarde van c?*

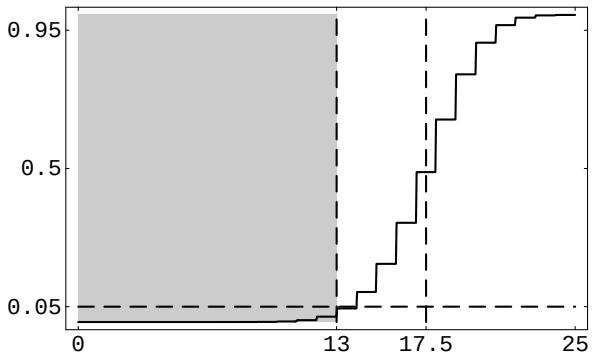
6.2.3 Toelaatbaarheidsintervallen

In Voorbeeld 6.2 is sprake van een zogenaamde *eenzijdige toets*: $\underline{x} \leq c$ leidt tot verwerping van H_0, terwijl $\underline{x} > c$ niet leidt tot verwerping van H_0. Het kan ook voorkomen dat de nulhypothese luidt dat een bepaalde parameter tussen twee grenzen ligt, of zelfs dat de parameter exact een bepaalde waarde heeft. Er zijn dan twee kritische waarden c_1 en c_2 die tezamen een *toelaatbaarheidsinterval* $\langle c_1, c_2 \rangle$ definiëren: we verwerpen H_0 als $\underline{x}$ buiten dit interval valt. We geven twee voorbeelden van zulke tweezijdige toetsen.

toelaatbaarheidsinterval

Voorbeeld 6.3 *Een onzuivere munt?*

Hoe kan men toetsen of een gegeven munt 'onzuiver' is? De nulhypothese H_0 zou dan moeten zijn dat de kans op 'kop' precies gelijk is aan $\frac{1}{2}$. Laten we, met een schuin oog naar een van de experimenten van Pearson die in Tabel 1.1 beschreven zijn, aannemen dat we de munt in kwestie



Figuur 6.4: Grafiek van $B_{25,0.7}(x)$ met kritische waarde $c = 13$. De nulhypothese wordt verworpen als $\underline{x} \leq 13$.

12000 maal mogen gooien. Het aantal malen 'kop' noemen we $\underline{x}$. We kiezen voor α weer 0.05. Welke grenzen moet het aantal malen 'kop' dan overschrijden, willen we de nulhypothese verwerpen?

Blijkbaar hebben we te maken met de verdelingsfunctie $B_{12000,0.5}(x)$. We gaan een toelaatbaarheidsinterval $\langle c_1, c_2\rangle$ construeren met de eigenschap dat H_0 verworpen wordt als $\underline{x} \leq c_1$ of $\underline{x} \geq c_2$. We willen dat de kans op het *ten onrechte* verwerpen van H_0 (dat wil zeggen het maken van een fout van de eerste soort) minder is dan 5 procent. Met andere woorden, we willen dat als $p = \frac{1}{2}$ waar is, de kans dat $\underline{x}$ binnen $\langle c_1, c_2\rangle$ valt, minstens 95 procent is.

Het is gebruikelijk om in zo'n geval de drempelwaarde α in twee gelijke delen te verdelen, en het linkereindpunt c_1 zo te kiezen dat

$$B_{12000,0.5}(c_1) < \frac{1}{2}\alpha = 0.025.$$

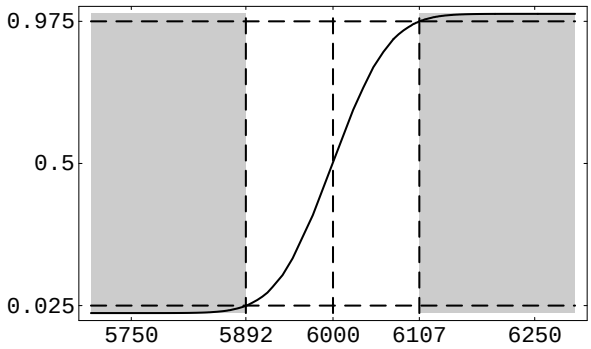
Het rechtereindpunt c_2 moet dan zo gekozen worden dat

$$B_{12000,0.5}(c_2) > 0.975.$$

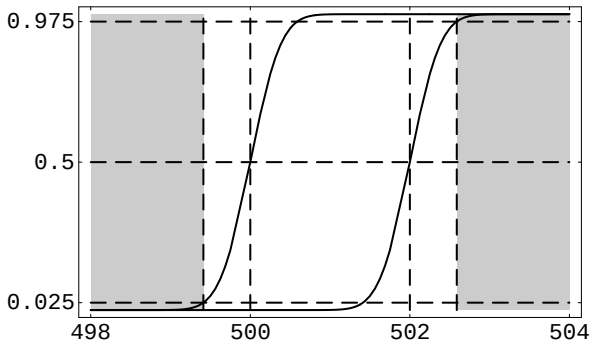
Met de computer vinden we de waarden $c_1 = 5892$ en $c_2 = 6107$. We verwerpen dus de nulhypothese wanneer $\underline{x} \leq 5892$ of $\underline{x} \geq 6107$. De door Pearson gebruikte munt gaf $\underline{x} = 6019$, een waarde die ruim binnen die grenzen valt, en we vinden dus binnen deze opzet geen redenen om aan te nemen dat er met de munt van Pearson iets mis was.

Voorbeeld 6.4 *Nogmaals de vulmachine.*

We keren terug naar de vulmachine van Voorbeeld 6.2, en nemen aan dat die in vol bedrijf is. Van tijd tot tijd nemen een steekproef van 100 potten om te controleren of de instellingen bijgesteld moeten worden. We nemen weer aan dat het vulgewicht normaal verdeeld is met een vaste standaardafwijking σ van 3 gram, en dat de instellingen van de machine alleen invloed hebben op de verwachtingswaarde μ. Als nulhypothese hanteren we $500 \leq \mu \leq 502$, en we vragen ons af wanneer we aan de bel moeten trekken als we, net als in het vorige voorbeeld, $\alpha = 0.05$ als drempelwaarde kiezen. Als toetsingsvariabele $\underline{x}$ kiezen we het gemiddelde vulgewicht in de steekproef. Die variabele is normaal verdeeld



Figuur 6.5: Het toetsen van de nulhypothese $p = \frac{1}{2}$ bij 12000 worpen met een munt en $\alpha = 0.05$. Als $5892 < \underline{x} < 6107$ wordt de nulhypothese niet verworpen.



Figuur 6.6: Toelaatbaarheidsinterval $\langle 499.412, 502.588\rangle$ bij het toetsen van de nulhypothese $500 \leq \mu \leq 502$ voor een steekproef van 100 potten en $\alpha = 0.05$. De toetsingsvariabele is het gemiddelde vulgewicht in de steekproef.

met onbekende verwachting μ en standaardafwijking $\sigma/\sqrt{100} = 0.3$. We zoeken weer een interval $\langle c_1, c_2\rangle$ met de eigenschap dat, onder aanname van de nulhypothese, de kans dat $\underline{x} \leq c_1$ of $\underline{x} \geq c_2$ minder is dan 5 procent. Net als boven delen we die 5 procent in twee gelijke porties van $2\frac{1}{2}$ procent op. We berekenen nu de getallen c_1 en c_2 waarvoor geldt dat

$$N_{500,0.3}(c_1) = 0.025$$

en

$$N_{502,0.3}(c_2) = 0.975.$$

Met behulp van de computer vinden we $c_1 = 499.412$ en $c_2 = 502.588$. Figuur 6 illustreert de situatie. Daarin zijn de normale verdelingsfuncties voor $\mu = 500$ en $\mu = 502$ getekend.

Opgave 6.8 *In de toetsingstheorie wordt bij het bepalen van een toelaatbaarheidsinterval bij een tweezijdige toets de drempelwaarde α in twee (gelijke) porties van $\frac{1}{2}\alpha$ verdeeld; in de schattingstheorie doet men dit niet wanneer men een betrouwbaarheidsinterval bepaalt. Verklaar dit.*

Bijlage A

Gemengde opgaven

GO.1. Bereken de kans dat bij het herhaald werpen met een 'zuivere' dobbelsteen de tweede maal 6 precies optreedt bij de tiende worp.

GO.2. Bij het pokeren werpt men met vijf dobbelstenen.

1. Beschrijf hiervoor een discreet kansmodel, waarbij u ervan uitgaat dat de dobbelstenen 'zuiver' zijn.
2. Bereken in dit model de kans op 'poker', dat wil zeggen de kans dat de vijf dobbelstenen dezelfde uitkomst tonen.
3. Bereken ook de kans op 'carré', dat wil zeggen de kans dat vier stenen dezelfde uitkomst geven, en de vijfde een andere uitkomst.

GO.3. Neem aan dat het aantal minuten dat iemand die een metrostation binnengaat moet wachten op de eerstvolgende metro negatief exponentieel verdeeld is met parameter $\lambda = 2$.

1. Geef van dit model de verdelingsfunctie en de kansdichtheidsfunctie.
2. Bereken in dit model de kans dat iemand na aankomst meer dan 2 minuten moet wachten.
3. Bereken ook de kans dat de wachttijd ligt tussen $\frac{1}{2}$ en $1\frac{1}{2}$ minuut.

GO.4. Voor een positieve constante a is gegeven de functie

$$f(x) = \begin{cases} a \sin x & 0 \leq x \leq \pi \\ 0 & \text{elders} \end{cases}$$

1. Bepaal a zo dat $f(x)$ een kansdichtheidsfunctie is.
2. Bepaal in dat geval de verdelingsfunctie.
3. Bepaal in dat geval ook de kans van het interval $[\pi/4, 3\pi/4]$.

GO.5. Bij de normale verdeling met $\mu = 1$ en $\sigma = 3$ wil men de kans op het interval $[0, \infty\rangle$ berekenen met behulp van een tabel van de standaardnormale verdeling. Druk daartoe deze kans uit in termen van de verdelingsfunctie $\Phi(x)$ van de standaardnormale verdeling.

GO.6. Bij het werpen met twee zuivere dobbelstenen bepaalt men het product van de uitkomsten. Onder de stochastische variabele $\underline{x}$ verstaat men het laatste cijfer van dat product. Bepaal de kansfunctie van $\underline{x}$ en de verwachting van $\underline{x}$.

GO.7. Bij een examen worden 40 multiple-choicevragen gesteld, elk met vijf antwoorden. Telkens is slechts één van de vijf antwoorden juist. Een student gokt elk antwoord naar willekeur.

1. Geef een formule voor de kans dat hij 20 of meer vragen goed gokt.
2. Wat is de verwachting van zijn score?

GO.8. De stochastische variabele $\underline{x}$ is uniform verdeeld op het interval $[0, 5]$. Bereken de verwachting van de stochastische variabele $\underline{x}^2 + \underline{x} + 1$.

GO.9. Een stochastische variabele $\underline{x}$ heeft als kansdichtheidsfunctie

$$f_{\underline{x}}(x) = \begin{cases} xe^{-x} & \text{als } x \geq 0 \\ 0 & \text{als } x < 0 \end{cases}$$

1. Verifieer dat $f_{\underline{x}}(x)$ een kansdichtheidsfunctie is.
2. Bepaal de bijbehorende verdelingsfunctie.
3. Bereken de verwachting van $\underline{x}$.

GO.10. Bij een examen worden 40 multiple-choicevragen gesteld, elk met vijf antwoorden. Telkens is slechts één van de vijf antwoorden juist. Een student gokt elk antwoord naar willekeur. Wat is de standaardafwijking van zijn score?

GO.11. Een stochastische variabele $\underline{x}$ heeft voor zekere positieve a als kansdichtheidsfunctie

$$f_{\underline{x}}(x) = \begin{cases} ax(1-x) & 0 \leq x \leq 1 \\ 0 & \text{elders} \end{cases}$$

1. Bepaal a.
2. Bereken de verwachting en de standaardafwijking van $\underline{x}$.

GO.12. Een stochastische variabele $\underline{x}$ heeft als kansdichtheidsfunctie

$$f_{\underline{x}}(x) = \begin{cases} xe^{-x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Bepaal de variantie van $\underline{x}$.

GO.13. Er worden n onafhankelijke trekkingen gedaan uit een verdeling met verwachting $\mu = 200$ en standaardafwijking $\sigma = 10$. De verdelingsfunctie is verder onbekend. Onder de stochastische variabele $\underline{m}_n$ verstaat men de gemiddelde waarde van deze n trekkingen. Op grond van de Centrale Limietstelling neemt men aan dat $\underline{m}_n$ voor grote n bij benadering normaal verdeeld is.

1. Hoe groot moet men n kiezen opdat de standaardafwijking van $\underline{m}_n$ ongeveer gelijk is aan 0.5?
2. Gebruik $\Phi(2) = 0.9772$ om te verifiëren dat $\underline{m}_n$ dan met een kans van ongeveer 95 procent tussen 199 en 201 ligt.

Bijlage B

Uitwerkingen van de opgaven

1.1. Nummer de feestgangers binair: 0001, 0010, 0011, ..., 1010. Werp vier maal, en laat K corresponderen met 0 en M met 1. Het resultaat is een getal van vier binaire cijfers. Als de uitkomst niet een van de bovengenoemde tien getallen is, werpt u opnieuw vier maal, enzovoort, totdat de taart door iemand is gewonnen.

1.2. a. $5/6$
b. $(1/6)^2$
c. $(5/6)^2$
d. De mogelijkheden om 8 te gooien zijn $(6,2)$, $(5,3)$, $(4,4)$, $(3,5)$, $(2,6)$. Ieder heeft kans $1/36$, dus in totaal is de kans $5/36$.

1.3. In de eerste plaats zou u zeker moeten weten dat er niet met de dobbelsteen geknoeid is. Als het een zuivere dobbelsteen betreft, is de kans op minstens één 6 in vier worpen is gelijk aan 1 minus de kans op geen enkele 6 in vier worpen, en dat is $1 - (5/6)^4 = 671/1296$. Aan u de keuze ...

1.4. Zie de aanwijzing. Gooi paren. Spreek van tevoren af wie wint bij KM en wie bij MK. Bij KK of MM gooit u opnieuw.

1.5. Niet waar. Als $P(A) > 0$ en $P(B) > 0$ is en als A en B disjunct zijn, dan is $P(A \cap B) = P(\phi) = 0 \neq P(A) \cdot P(B)$.

1.6. $P(A \cap B) = 0.15 + 0.33 - 0.45 = 0.03 \neq P(A) \cdot P(B)$. A en B zijn dus noch disjunct, noch onafhankelijk.
$P(B \mid A) = 0.03/0.15 = 0.2$, $P(A \mid B) = 0.03/0.33 = \frac{1}{11}$.

1.7. Als A de gebeurtenis voorstelt dat de kop defect is, en B de gebeurtenis dat de schroefdraad defect is, dan is $P(A \cap B) = 0.008 - 0.00795 = 0.00005$, en dus is $P(A \mid B) = 0.00005/0.005 = 0.01$.

1.8. Nee. Zie de vorige opgave.

1.9. Ja.

1.10. Die kans wordt groter: $P(A \mid B) = 0.164016$

1.11. Nu wordt de kans $P(A \mid B) = 0.227730$

1.12. Er is niets bekend over een 'kansmechanisme', dus de vraagstelling op zichzelf is al twijfelachtig. Als we zelf veronderstellen dat de 'keuze' van de schutter plaatsvindt met gelijke kansen (ieder dus kans $1/4$), dan kunnen we de

voorwaardelijke kansen uitrekenen. Noem A de gebeurtenis dat a geschoten heeft, etc., en noem T de gebeurtenis dat het schot doel treft. Dan is $P(T \mid A) = 0.10$, etc. Een korte berekening leert nu dat $P(A \mid T) = 0.20$, $P(B \mid T) = 0.40$, $P(C \mid T) = 0.30$, $P(D \mid T) = 0.10$.

1.13. (1.) We geven met $P(A)$ de gebeurtenis aan dat de schutter uit groep A afkomstig is, etc. Dan is $P(T \mid A) = 0.10$, etc., en een korte berekening leert dat $P(A \mid T) = 0.210526$, $P(B \mid T) = 0.210526$, $P(C \mid T) = 0.473684$, $P(D \mid T) = 0.105263$.
(2.) Noem M de gebeurtenis dat de schutter drie maal mis schiet. Dan is $P(M \mid A) = 0.9^3$, etc., en een korte berekening leert nu dat $P(A \mid T) = 0.263789$, $P(B \mid T) = 0.0926340$, $P(C \mid T) = 0.333333$, $P(D \mid T) = 0.310242$.

2.1. $F(x) = x - 1$ voor $1 \leq x \leq 2$. $P([\sqrt{2}, \sqrt{3}]) = \sqrt{3} - \sqrt{2}$.

2.2. $F(x)$ is de lineaire functie die op het betreffende interval van 0 naar 1 stijgt. Voor het interval $[2, 8]$ is het dus de functie

$$F(x) = \frac{x-2}{8-2} \qquad \text{als} \qquad 2 \leq x \leq 8$$

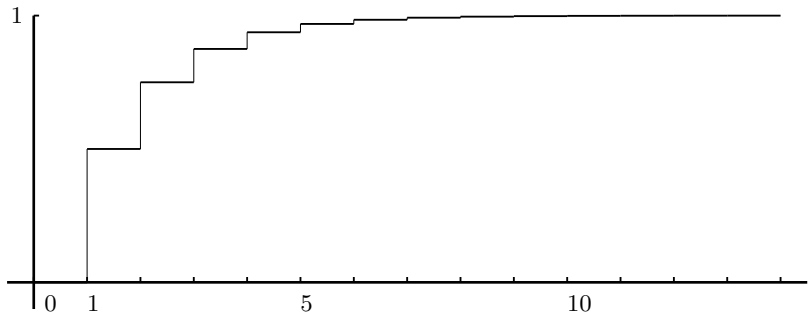
Voor elk deelinterval $[a, b]$ van $[2, 8]$ is dan $P([a, b]) = F(b) - F(a) = (b-a)/6$. In het bijzonder is $P([3, \pi]) = (\pi - 3)/6$.

2.3. $P(\langle 0, 7\rangle) = 5/6$, $P(\langle -\infty, 0]) = 0$ en $P([0, \infty\rangle) = 1$.

2.4 $P([-1, 1]) = F(1) - F(-1) = 1 - e^{-2}$ want $F(-1) = 0$. Evenzo is $P([1, \infty\rangle) = 1 - (1 - e^{-2}) = e^{-2}$.

2.5. Het gaat telkens om de toename van $F(x)$ op het betreffende interval. Maar $F(x)$ is constant behalve in de sprongpunten $x = 1, \ldots, x = 6$, waar telkens een sprong ter grootte $\frac{1}{6}$ optreedt. De kans op een interval is dus gelijk aan het aantal sprongpunten dat zo'n interval bevat, gedeeld door 6. Dit levert de volgende antwoorden:
$P(\langle 0, \infty\rangle) = 1$, $P(\langle 1, \infty\rangle) = \frac{5}{6}$, $P([1, \infty\rangle) = 1$ en $P(\langle 1, 2\rangle) = 0$.

2.6. $F(x) = 0$ voor $x < 1$ en $F(x) = \dfrac{1}{2} + \dfrac{1}{4} + \dfrac{1}{8} + \ldots + \dfrac{1}{2^k} = 1 - \dfrac{1}{2^k}$ als $x \geq 1$, waarbij k het grootste gehele getal is dat kleiner dan of gelijk is aan x.



Figuur B.1: De verdelingsfunctie $F(x)$ van Opgave 1.10

2.7. $P(n) = P(MMM \ldots MK) = \dfrac{1}{2^n}$
($(n-1)$ maal Munt, de laatste maal Kop).
De kans op helemaal geen kop is $1 - (P(1) + P(2) + P(3) + \ldots) = 1 - (\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots) = 0$.

2.8. $f(x) = \begin{cases} \dfrac{1}{b-a} & a \leq x \leq b \\ 0 & \text{elders} \end{cases}$

2.9. Alleen eigenschap (3) is niet helemaal vanzelfsprekend:

$$\int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)} dx = \frac{1}{\pi} \left[\arctan x\right]_{-\infty}^{+\infty} = \frac{1}{\pi} \left(\frac{\pi}{2} + \frac{\pi}{2}\right) = 1$$

De formule voor de verdelingsfunctie luidt:

$$F(x) = \int_{-\infty}^{x} \frac{1}{\pi(1+t^2)} dt = \frac{1}{\pi} \left(\arctan x + \frac{\pi}{2}\right).$$

2.10. De grafiek van $n_{\mu,\sigma}(x)$ ontstaat uit die van $\varphi(x)$ door een horizontale verschuiving van de top naar de lijn $x = \mu$, gevolgd door een horizontale vermenigvuldiging met een factor σ vanuit de lijn $x = \mu$ en een verticale vermenigvuldiging met een factor $1/\sigma$ vanuit de x-as.
De grafiek van $N_{\mu,\sigma}(x)$ ontstaat uit die van $\Phi(x)$ door een horizontale verschuiving van het buigpunt naar de lijn $x = \mu$, gevolgd door een horizontale vermenigvuldiging met een factor σ vanuit de lijn $x = \mu$.

2.11. Omdat de integrand een even functie is, geldt

$$\Phi(0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{1}{2}t^2} dt = \frac{1}{2} \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}t^2} dt = \frac{1}{2}.$$

2.12. Voor iedere x geldt dat

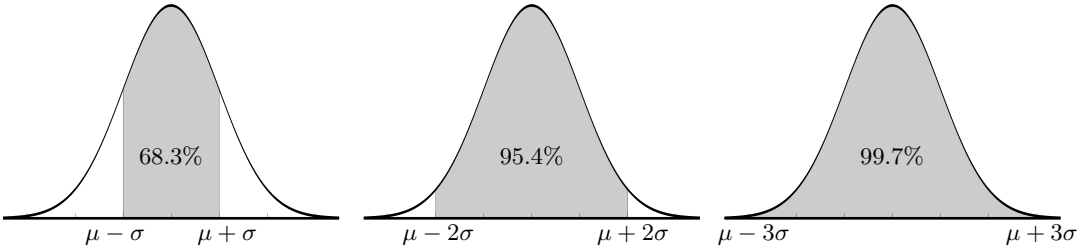
$$\begin{aligned} \Phi(-x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-x} e^{-\frac{1}{2}t^2} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-\frac{1}{2}t^2} dt \\ &= 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt \\ &= 1 - \Phi(x) \end{aligned}$$

2.13. a. Op grond van de omrekenformule en het resultaat van de vorige opgave geldt $P([-\infty, \mu - 2\sigma]) = N_{\mu,\sigma}(\mu - 2\sigma) - N_{\mu,\sigma}(-\infty) = \Phi(-2) - \Phi(-\infty) = (1 - \Phi(2)) - 0 \approx 0.023$

b. Evenzo: $P([\mu - \sigma, \mu + \sigma]) = \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 \approx 0.683$.
Gevolg: bij de kansdichtheidsfunctie van een normale verdeling bevindt zich ruim 68 procent van de oppervlakte onder de 'klokkromme' tussen de waarden $\mu - \sigma$ en $\mu + \sigma$.

c. $P([\mu - 2\sigma, \mu + 2\sigma]) = 2\Phi(2) - 1 \approx 0.954$.
Gevolg: bij de kansdichtheidsfunctie van een normale verdeling bevindt zich ruim 95 procent van de oppervlakte onder de 'klokkromme' tussen de waarden $\mu - 2\sigma$ en $\mu + 2\sigma$.

d. $P([\mu - 3\sigma, \mu + 3\sigma]) \approx 0.997$.
Gevolg: bij de kansdichtheidsfunctie van een normale verdeling bevindt zich ruim 99.7 procent van de oppervlakte onder de 'klokkromme' tussen de waarden $\mu - 3\sigma$ en $\mu + 3\sigma$!



Figuur B.2: Percentages van de oppervlakte onder de klokkromme bij de normale verdeling in *Opgave 1.17 (b), (c)* en *(d).*

2.14. a. $\lim_{x\to\infty} \mathrm{erf}(x) = 1$, $\lim_{x\to-\infty} \mathrm{erf}(x) = -1$.

b. We splitsen de integraal waardoor $\Phi(x)$ gedefinieerd wordt in twee stukken: een integraal over $\langle -\infty, 0]$ en een integraal over $[0, x]$:

$$\begin{aligned}\Phi(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{1}{2}t^2} dt + \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{1}{2}t^2} dt\end{aligned}$$

De eerste integraal is gelijk aan $\frac{1}{2}$, de tweede kunnen we via een substitutie als volgt transformeren in een uitdrukking die te schrijven is in termen van de error function:

$$\begin{aligned}\Phi(x) &= \frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_{t=0}^{t=x} e^{-(t\sqrt{2})^2} d\left(\frac{t}{\sqrt{2}}\right) \\ &= \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\end{aligned}$$

c. Op grond van de omrekenformule geldt nu

$$N_{\mu,\sigma}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)$$

3.1. a. We geven de kansfunctie en de verdelingsfunctie in tabelvorm. Bij de verdelingsfunctie, die eigenlijk op de gehele $\mathbb{R}$ gedefinieerd is, geven we dus alleen de waarden in de sprongpunten. Bedenk hierbij dat $F_{\underline{x}}(x) = P(\underline{x} \leq x) = \sum_{u \leq x} P(\underline{x} = u)$.

x	0	1	2	3
$P(\underline{x} = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$
$F_{\underline{x}}(x) = P(\underline{x} \leq x)$	$\frac{1}{8}$	$\frac{4}{8}$	$\frac{7}{8}$	1

b. In een tabel analoog aan Tabel 3.1 kan men bij elk paar uitkomsten de absolute waarde van het verschil schrijven. Dit leidt tot de volgende tabel voor de kansfunctie en de verdelingsfunctie:

x	0	1	2	3	4	5
$P(\underline{x} = x)$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$
$F_{\underline{x}}(x) = P(\underline{x} \leq x)$	$\frac{6}{36}$	$\frac{16}{36}$	$\frac{24}{36}$	$\frac{30}{36}$	$\frac{34}{36}$	1

c. Op soortgelijke wijze vindt men nu

x	3	4	5	6	7	8	9	10	11
$P(\underline{x} = x)$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{8}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$
$F_{\underline{x}}(x) = P(\underline{x} \leq x)$	$\frac{2}{36}$	$\frac{5}{36}$	$\frac{9}{36}$	$\frac{14}{36}$	$\frac{22}{36}$	$\frac{27}{36}$	$\frac{31}{36}$	$\frac{34}{36}$	1

3.2. Er zijn twee mogelijke uitkomsten, 0 en 2. Bij elke uitkomst berekenen we de kans. De uitkomst is 2 in de volgende gevallen: K, MMK, MMMMK, MMMMMMK, ... Optellen van de betreffende kansen geeft

$$P(\underline{x} = 2) = \frac{1}{2} + \frac{1}{8} + \frac{1}{32} + \ldots = \frac{1/2}{1 - \frac{1}{4}} = \frac{2}{3}$$

waarbij we gebruikt hebben dat hier een meetkundige reeks staat met beginterm $\frac{1}{2}$ en reden $\frac{1}{4}$.
Evenzo krijgen we als uitkomst 0 in de gevallen MK, MMMK, MMMMMK, ..., dus

$$P(\underline{x} = 0) = \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \ldots = \frac{1/4}{1 - \frac{1}{4}} = \frac{1}{3}$$

Voor de verdelingsfunctie $F_{\underline{x}}(x) = P(\underline{x} \leq x)$ geldt dan

$$F_{\underline{x}}(x) = \begin{cases} 0 & \text{als } x < 0 \\ \frac{1}{3} & \text{als } 0 \leq x < 2 \\ 1 & \text{als } x \geq 2 \end{cases}$$

3.3. Win ik na 1 ronde, dan is mijn winst 1 gulden. Win ik na 2 ronden, dan heb ik in totaal $1 + 2 = 3$ gulden ingezet en 4 gulden gekregen, dus is mijn winst weer 1 gulden, etc. Altijd is mijn winst dus 1 gulden, alleen weet ik niet hoe lang het duurt totdat ik die winst kan incasseren.Bovendien moet ik onbeperkt crediet hebben om dit spel te kunnen spelen.
De kansverdeling is dus eenvoudig $P(\underline{x} = 1) = 1$ en de verdelingsfunctie is

$$F_{\underline{x}}(x) = \begin{cases} 0 & \text{als } x < 1 \\ 1 & \text{als } x \geq 1 \end{cases}$$

3.4. Win ik voordat er tien ronden gespeeld zijn, dan is mijn winst weer 1 gulden. Verlies ik echter alle tien de ronden, dan ben ik $1 + 2 + 4 + \ldots + 512 = 1023$ gulden kwijt. De kans hierop is $1/2^{10} = 1/1024$. Er geldt dus

$$P(\underline{y} = -1023) = \frac{1}{1024}$$

$$P(\underline{y} = 1) = \frac{1023}{1024}$$

en

$$F_{\underline{y}}(x) = \begin{cases} 0 & \text{als } x < -1023 \\ \dfrac{1}{1024} & \text{als } -1023 \leq x < 1 \\ 1 & \text{als } x \geq 1 \end{cases}$$

3.5. Ik verlies wanneer ik 24 maal achter elkaar géén dubbelzes gooi. De kans daarop is $(35/36)^{24} \approx 0.5086$ dus de kans dat ik de weddenschap win is $1 - (35/36)^{24} \approx 0.4914$.

3.6. Het gaat hier om een binomiale verdeling met $n = 20$ en $p = 0.3$.
a. $P(\underline{x} = 6) = \binom{20}{6}(0.3)^6(0.7)^{14} \approx 0.1916$.
b. $P(\underline{x} \leq 2) = \sum_{k=0}^{2} \binom{20}{k}(0.3)^k(0.7)^{20-k} \approx 0.0355$.

3.7. We geven het antwoord in dezelfde vorm als in Opgave 3.1.

x	0	1	2	3	4
$P(\underline{x} = x)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$
$F_{\underline{x}}(x)$	$\frac{1}{16}$	$\frac{5}{16}$	$\frac{11}{16}$	$\frac{15}{16}$	1

3.8. a. We behoeven slechts te verifiëren dat de totale integraal 1 is.

$$\int_{-\infty}^{+\infty} f(x)dx = \frac{2}{\pi}\int_0^{\infty} \frac{1}{1+x^4}d(x^2) = \frac{2}{\pi}\left[\arctan x^2\right]_0^{\infty} = 1$$

b. $F_{\underline{x}}(x) = \int_0^x f(t)dt = \begin{cases} \dfrac{2}{\pi}\arctan x^2 & \text{als } x \geq 0 \\ 0 & \text{als } x < 0 \end{cases}$

c. $P(\underline{x} \leq 1) = F_{\underline{x}}(1) = \dfrac{2}{\pi}\arctan 1 = \dfrac{1}{2}$

4.1. Dit is een binomiale verdeling met $n = 24$ en $p = 1/36$ dus $E(\underline{x}) = np = 24/36 = 2/3$.

4.2. (Zie ook de uitwerking van Opgave 3.1)
a. $E(\underline{x}) = 3/2$
b. $E(\underline{x}) = 70/36 = 35/18$
c. $E(\underline{x}) = 252/36 = 7$.

4.3. $E(\underline{x}) = 1$, $E(\underline{y}) = -1023 \cdot \dfrac{1}{1024} + 1 \cdot \dfrac{1023}{1024} = 0$

4.4. De waarden van $\underline{y}$ zijn 1, 4, 9, 16, 25 en 36, die ieder met kans 1/6 worden aangenomen. De verwachting is dus

$$E(\underline{y}) = (1/6)(1 + 4 + 9 + 16 + 25 + 36) = 91/6.$$

4.5. De kansdichtheidsfunctie is 1/2 op het interval $[-1, 1]$ en 0 daarbuiten, dus de verwachting van $\underline{y} = \underline{x}^4$ is

$$E(\underline{y}) = E(\underline{x}^4) = \frac{1}{2}\int_{-1}^{1} x^4 dx = \frac{1}{10}[x^5]_{-1}^{1} = \frac{1}{5}$$

4.6. De kansdichtheidsfunctie is $\lambda e^{-\lambda x}$ op $[0,\infty\rangle$ en 0 daarbuiten, dus de verwachting van $\underline{x}^2$ is

$$\begin{aligned} E(\underline{x}^2) &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx = -\int_0^\infty x^2 de^{-2x} \\ &= -\left[x^2 e^{-2x}\right]_0^\infty + \int_0^\infty e^{-2x} 2x dx = -\int_0^\infty x de^{-2x} \\ &= -\left[x e^{-2x}\right]_0^\infty + \int_0^\infty e^{-2x} dx \\ &= 0 - \frac{1}{2}\left[e^{-2x}\right]_0^\infty = \frac{1}{2} \end{aligned}$$

4.7. Een voorwaarde voor het bestaan van de verwachting is dat de integraal

$$\int_{-\infty}^{\infty} |x| f_{\underline{x}}(x) dx$$

eindig is. Dit is hier niet het geval:

$$\int_{-\infty}^{+\infty} \frac{|x|}{\pi(1+x^2)} dx = 2\int_0^\infty \frac{x}{\pi(1+x^2)} dx = +\infty$$

Voor de Cauchyverdeling is de verwachting dus niet gedefinieerd.

4.8. De verwachting wordt in dit geval gedefinieerd door de integraal:

$$\int_0^{+\infty} \frac{4x^2}{\pi(1+x^4)} dx.$$

De integrand is op het gehele integratieinterval positief, dus de verwachting is, als die bestaat, ook positief. We behoeven blijkbaar alleen maar aan te tonen dat deze integraal eindig is.

Bij 0 is de waarde van de integrand 0, en voor $x \to \infty$ is de x-as een horizontale asymptoot. De orde van grootte van de integrand voor $x \to \infty$ is $1/x^2$, en de integraal daarvan op (bijvoorbeeld) het interval $[1,\infty\rangle$ is eindig. Hetzelfde zal dus ook wel gelden voor de integraal die we willen afschatten.

Om dit vermoeden te bewijzen, splitsen we de integraal in twee stukken die we elk afzonderlijk afschatten:

$$\begin{aligned} E(\underline{x}) &= \int_0^{+\infty} \frac{4x^2}{\pi(1+x^4)} dx \\ &= \int_0^1 \frac{4x^2}{\pi(1+x^4)} dx + \int_1^{+\infty} \frac{4x^2}{\pi(1+x^4)} dx \\ &< \frac{4}{\pi}\int_0^1 1\, dx + \frac{4}{\pi}\int_1^\infty \frac{x^2}{x^4} dx \\ &= \frac{4}{\pi} + \frac{4}{\pi}\left[-\frac{1}{x}\right]_1^\infty = \frac{8}{\pi} \end{aligned}$$

We kennen nog steeds de verwachting niet exact, maar we hebben wel een eindige bovengrens gevonden, en dus hebben we hiermee bewezen dat de integraal waardoor de verwachting wordt gedefinieerd, ook eindig is.

4.9. Opgave 3.1:
a. Dit is een binomiale verdeling met $n = 3$ en $p = 1/2$, dus de variantie is

$np(1-p) = 3/4$.
b. Hier passen we Stelling 4.6 toe: we weten al dat $E(\underline{x}) = 35/18$ (zie Opgave 4.2) en $E(\underline{x}^2) = 210/36 = 35/6$ dus $Var(\underline{x}) = 665/324$.
Opgave 3.2: $E(\underline{x}^2) = 0 \cdot \frac{1}{3} + 4 \cdot \frac{2}{3} = \frac{8}{3}$. Omdat $\mu = \frac{4}{3}$ is $Var(\underline{x}) = \frac{8}{9}$.
Opgave 3.3: $E(\underline{x}^2) = E(\underline{x}) = 1$ dus $Var(\underline{x}) = 0$.
Opgave 3.4: we weten al dat $E(\underline{y}) = 0$. $E(\underline{y}^2) = (1023^2 + 1023)/1024 = 1023$, dus ook $Var(\underline{y}) = 1023$.

4.10. We weten al dat $E(\underline{y}) = E(\underline{x}^4) = 1/5$. We berekenen $E(\underline{y}^2) = E(\underline{x}^8)$

$$E(\underline{x}^8) = \frac{1}{2}\int_{-1}^{1} x^8\, dx = \frac{1}{18}[x^9]_{-1}^{1} = \frac{1}{9}$$

en dus is $Var(\underline{y}) = (1/9) - (1/5)^2 = 16/225$ en $\sigma(\underline{y}) = 4/15$.

4.11. $E(\underline{x}^2) = \displaystyle\int_0^\infty \frac{4x^3}{\pi(1+x^4)}dx = \frac{1}{\pi}\,[\ln(1+x^4)]_0^\infty = +\infty$
en omdat $E(\underline{x})$ eindig is, is $Var(\underline{x})$ dus oneindig.

4.12. We nemen 1 seconde als tijdseenheid. Er is gemiddeld 1 klik per 6 seconden, dus we nemen $\lambda = 1/6$. De gevraagde kans is dan $F(10) - F(5) = e^{-5/6} - e^{-10/6} \approx 0.2457$.

5.1. a. Het gaat hier om de binomiale verdeling met $n = 1000$ en $p = 0.03$. De verwachting is dus $np = 30$, de variantie is $np(1-p) = 29.1$ en de standaardafwijking is $\sqrt{29.1} \approx 5.4$.
b. Benader de verdeling door een normale verdeling met $\mu = 30$ en $\sigma = 5.4$. Dan is $\mu + 3\sigma = 46.2$ en dit is ruim onder de 48. De uitspraak lijkt gezien de vuistregel dus zeker verantwoord.

5.2. a. De uniforme verdeling op $[0, 1]$ heeft verwachting $1/2$ en standaardafwijking $1/\sqrt{12}$. Het gemiddelde X van 100 trekkingen heeft dus verwachting $1/2$ en standaardafwijking $1/10\sqrt{12} \approx 0.029$.
b. Voor het gevraagde interval $[a, b]$ nemen we $[\mu - 2\sigma, u + 2\sigma] = [0.442, 0.558]$.

5.3. Per worp is het een Bernoulli-verdeling met $p = 1/2$, dus $\mu = 1/2$ en $\sigma = 1/2$. Bij n worpen is de standaardafwijking van het gemiddelde dus $\sigma/\sqrt{n} = 1/(2\sqrt{n})$. Op grond van de vuistregel eisen we dat 2 maal dit getal kleiner is dan 0.05, dus $n > 20^2 = 400$.

5.4. $\underline{G}$ is normaal verdeeld met verwachting $\mu = 25 \times 501 = 12525$ gram en standaardafwijking $\sigma = 3 \times \sqrt{25} = 15$ gram.

6.1. Vergelijking (1) luidt

$$\sum_{k=0}^{10} \binom{25}{k} p^k (1-p)^{25-k} = 0.05.$$

Elk van de termen in deze som is een polynoom in p van de graad 25, en hetzelfde geldt dus voor de gehele som. Evenzo voor vergelijking (2).

6.2. Niet waar; het interval wordt juist groter.

6.3. Omdat $\Phi(-1.645) = 1 - \Phi(1.645) = 0.05$ krijgen we de vergelijking

$$\frac{501.13 - \mu}{0.3} = -1.645$$

met als oplossing $\mu = 501.6235$.

6.4. De grafiek van $N_{\mu,\sigma}(x)$ is puntsymmetrisch ten opzichte van het punt $(\mu, \frac{1}{2})$. Dit verklaart de symmetrie van het betrouwbaarheidsinterval rond μ. Bij de binomiale verdeling is grafiek van de verdelingsfunctie echter *niet* puntsymmetrisch ten opzichte van $(\mu, \frac{1}{2}) = (np, \frac{1}{2})$.

6.5. Niet waar! Althans, de uitspraak is in wiskundige zin niet welgedefinieerd. In tegenstelling met de indruk die een oppervlakkige benadering zou wekken, is er namelijk geen kansverdeling gedefinieerd op de parameterruimte. Er is in ons model eenvoudig sprake van een vaste, maar onbekende waarde van p. Men zou wel kunnen nadenken over de vraag of, en zo ja hoe, men zo'n kansverdeling zou kunnen definiëren, maar dat zou hier te ver voeren.

6.6. Deze uitspraak is daarentegen wél waar.

6.7. Ja. Consequentie is dat een kleinere waarde van α leidt tot minder verwerpingen van de nulhypothese.

6.8. De discrepantie zit hem eenvoudig in de volstrekt verschillende definities voor betrouwbaarheidsinterval en toelaatbaarheidsinterval.

Een betrouwbaarheidsinterval is een interval met de eigenschap dat de kans op de in de steekproef gevonden waarde kleiner is dan α wanneer de onbekende parameter *niet* in het betrouwbaarheidsinterval ligt.

Een toelaatbaarheidsinterval is echter een interval met de eigenschap dat onder aanname van de nulhypothese H_0, de kans op een waarde van de toetsingsvariabele die niet in het toelaatbaarheidsinterval ligt, kleiner dan of gelijk is aan α. Met andere woorden, onder aanname van H_0 is de kans op een waarde die wél in het toelaatbaarheidsinterval ligt, groter dan $1 - \alpha$.

In dit verband is het misschien ook verhelderend om op te merken dat een betrouwbaarheidsinterval een deelverzameling is van de parameterruimte, terwijl een toelaatbaarheidsinterval een deelverzameling is van de waardenverzameling van de toetsingsvariabele.

Uitwerking van de gemengde opgaven

GO.1. De tiende worp moet 6 zijn, en van de negen eerdere worpen moet precies één worp 6 zijn. Die ene 6 kan op negen plaatsen zijn voorgekomen. Bij elk van die negen mogelijkheden hoort een kans $(\frac{5}{6})^8 \times \frac{1}{6} \times \frac{1}{6}$. In totaal is de kans dus $9 \times (\frac{5}{6})^8 \times \frac{1}{6} \times \frac{1}{6}$ $(= \frac{390625}{6718464})$.

GO.2. (a) De uitkomstenruimte U bestaat uit alle rijtjes van vijf gehele getallen tussen 1 en 6 (inclusief grenzen); elk rijtje heeft kans 6^{-5}.
(b) Er zijn precies zes mogelijke 'poker'-rijtjes, dus de kans is 6^{-4}.
(c) Bij carré zijn er telkens twee ogenaantallen in het spel: de 'enkele' en de 'viervoudige'. Voor de 'enkele' zijn er zes mogelijke ogenaantallen, en bij elk ogenaantal vervolgens vijf plaatsen in het rijtje waar dat aantal voor kan komen. Daarna kan men voor het 'viervoudige' ogenaantal nog telkens vijf mogelijkheden kiezen. Het aantal rijtjes 'carré' is dus $6 \times 5 \times 5$ en de kans op 'carré' is daarom $6 \times 5 \times 5 \times 6^{-5} = \frac{25}{1296}$.

GO.3. (a)

$$F(x) = \begin{cases} 1 - e^{-2x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$f(x) = \begin{cases} 2e^{-2x} & x > 0 \\ 0 & x < 0 \end{cases}$$

(b) $1 - F(2) = e^{-4} \approx 0.0183$.
(c) $F(\frac{3}{2}) - F(\frac{1}{2}) = e^{-1} - e^{-3} \approx 0.3181$.

GO.4. (a) Uit $\int_0^\pi f(x)dx = 1$ volgt $a = \frac{1}{2}$.
(b)

$$F(x) = \begin{cases} 0 & x < 0 \\ \int_0^x \frac{1}{2} \sin t \, dt = \frac{1}{2}(1 - \cos x) & 0 \leq x \leq \pi \\ 1 & x \geq \pi \end{cases}$$

(c) $F(3\pi/4) - F(\pi/4) = \frac{1}{2}\sqrt{2}$.

GO.5. Die kans is gelijk aan $1 - N_{1,3}(0) = 1 - \Phi(\frac{0-1}{3}) = \Phi(\frac{1}{3})$ (want in het algemeen geldt voor iedere x dat $1 - \Phi(x) = \Phi(-x)$).

GO.6. U kunt de kansfunctie eenvoudig beschrijven door het laatste cijfer van het product van de ogenaantallen in een 6×6 tableau te zetten. Elk van die cijfers heeft kans $\frac{1}{36}$, maar sommige cijfers komen meerdere malen voor:

	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	8	0	2
3	3	6	9	2	5	8
4	4	8	2	6	0	4
5	5	0	5	0	5	0
6	6	2	8	4	0	6

Voor de kansfunctie krijgen we dan

x	0	1	2	3	4	5	6	8	9
$P(\underline{x} = x)$	$\frac{6}{36}$	$\frac{1}{36}$	$\frac{6}{36}$	$\frac{2}{36}$	$\frac{5}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{1}{36}$

en de verwachting blijkt $141/36 = 47/12$ te zijn.

GO.7. (a) De kans op 20 of meer goede antwoorden is gelijk aan

$$\sum_{k=20}^{40} \binom{40}{k} \left(\frac{1}{5}\right)^k \left(\frac{4}{5}\right)^{40-k}.$$

Opmerking: met (bijv.) DERIVE kunt u verifiëren dat deze kans ongeveer gelijk is aan 2.16918×10^{-5}.
(b) De verwachting is $np = 40 \times \frac{1}{5} = 8$.

GO.8.

$$E(\underline{x}^2 + \underline{x} + 1) = \int_0^5 \frac{1}{5}(x^2 + x + 1)\, dx = \frac{71}{6}$$

GO.9. (a) We behoeven slechts te verifiëren dat de totale integraal van $f_{\underline{x}}(x)$ gelijk is aan 1. Dat kan met partiële integratie:

$$\int_0^\infty xe^{-x}dx = -\int_0^\infty xd(e^{-x}) = [-xe^{-x}]_0^\infty + \int_0^\infty e^{-x}dx = 1.$$

(b) Eveneens met partiële integratie (we laten de tussenstappen weg):

$$F(x) = \begin{cases} \int_0^x te^{-t}dt = 1 - (x+1)e^{-x} & \text{als } x \geq 0 \\ 0 & \text{als } x < 0 \end{cases}$$

(c) Evenzo:

$$E(\underline{x}) = \int_0^\infty x^2e^{-x}dx = -\int_0^\infty x^2d(e^{-x}) = [-x^2e^{-x}]_0^\infty + 2\int_0^\infty xe^{-x}dx = 2.$$

GO.10. De standaardafwijking is $\sqrt{np(1-p)} = \sqrt{40 \times \frac{1}{5} \times \frac{4}{5}} \approx 2.53$

GO.11. (a) $\int_0^1 x(1-x)dx = \frac{1}{6}$ dus $a = 6$.
(b) $E(\underline{x}) = 6\int_0^1 x^2(1-x)dx = \frac{1}{2}$,
$E(\underline{x}^2) = 6\int_0^1 x^3(1-x)dx = 0.3$, dus $Var(\underline{x}) = 0.3 - (\frac{1}{2})^2 = 0.05$ en $\sigma(\underline{x}) = \sqrt{0.05} \approx 0.22$.

GO.12. $E(\underline{x}) = 2$ (zie opgave GO.9).
$E(\underline{x}^2) = \int_0^\infty x^3e^{-x}dx = 6$, dus
$Var(\underline{x}) = 6 - 2^2 = 2$.

GO.13. (a) Er moet gelden dat $\sigma/\sqrt{n} \approx 0.5$, dus $n \approx 100/0.25 = 400$.
(b) $\underline{m}_{400}$ is dan bij benadering $N_{200,\,0.5}$ verdeeld. Er geldt dus

$$\begin{aligned} P(199 < \underline{m}_{400} < 201) &\approx N_{200,\,0.5}(201) - N_{200,\,0.5}(199) \\ &= \Phi(\frac{201-200}{0.5}) - \Phi(\frac{199-200}{0.5}) \\ &= \Phi(2) - (1 - \Phi(2)) = 2\,\Phi(2) - 1 \\ &\approx 0.9544. \end{aligned}$$

Bijlage C

Formules

Verwachting en variantie

Als $\underline{x}$ een discrete stochastische variabele is die de waarden x_1, x_2, ... aanneemt met kansen resp. p_1, p_2, ..., dan wordt de verwachting van $\underline{x}$ gedefinieerd door

$$E(\underline{x}) = \sum_i x_i p_i.$$

In het geval $\underline{x}$ oneindig veel waarden kan aannemen, moet men hierbij eisen dat

$$\sum_i |x_i| p_i < \infty.$$

Voor een willekeurige functie $\underline{y} = g(\underline{x})$ waarvoor de verwachting $E(\underline{y})$ bestaat, geldt

$$E(\underline{y}) = \sum_i g(x_i) p_i.$$

Als $\underline{x}$ een continue stochastische variabele is, en

$$\int_{-\infty}^{\infty} |x| f_{\underline{x}}(x) dx$$

is eindig, dan is de verwachting van $\underline{x}$ gedefinieerd door

$$E(\underline{x}) = \int_{-\infty}^{\infty} x f_{\underline{x}}(x) dx.$$

Voor een willekeurige 'nette' functie $\underline{y} = g(\underline{x})$ waarvoor de verwachting $E(\underline{y})$ bestaat, geldt

$$E(\underline{y}) = \int_{-\infty}^{\infty} g(x) f_{\underline{x}}(x) dx.$$

De variantie $Var(\underline{x})$ van de (discrete of continue) stochastische variabele $\underline{x}$ met verwachting μ wordt gedefinieerd door

$$Var(\underline{x}) = E((\underline{x} - \mu)^2).$$

De wortel uit de variantie heet de *standaardafwijking*, ook wel *standaarddeviatie*. Hiervoor gebruikt men vaak de Griekse letter σ, dus

$$\sigma(\underline{x}) = \sqrt{Var(\underline{x})}.$$

Er geldt:

$$Var(\underline{x}) = E(\underline{x}^2) - (E(\underline{x}))^2 = E(\underline{x}^2) - \mu^2.$$

Discrete kansverdelingen

Bernoulli-verdeling met parameter p

kansverdeling: $P(\underline{x}=1)=p \quad P(\underline{x}=0)=1-p$
verwachting: $E(\underline{x})=p$
variantie: $Var(\underline{x})=p(1-p)$
standaardafwijking: $\sigma(\underline{x})=\sqrt{Var(\underline{x})}=\sqrt{p(1-p)}$

Binomiale verdeling met parameters n en p

kansverdeling: $P(\underline{x}=k)=\binom{n}{k}p^k(1-p)^{n-k} \quad (k=0,\ldots,n)$
verwachting: $E(\underline{x})=np$
variantie: $Var(\underline{x})=np(1-p)$
standaardafwijking: $\sigma(\underline{x})=\sqrt{Var(\underline{x})}=\sqrt{np(1-p)}$

Continue kansverdelingen

Uniforme verdeling op $[a,b]$

verdelingsfunctie: $F(x)=\begin{cases}0 & \text{als } x<a\\ (x-a)/(b-a) & \text{als } a\le x\le b\\ 1 & \text{als } x>b\end{cases}$

kansdichtheidsfunctie: $f(x)=\begin{cases}0 & \text{als } x<a\\ 1/(b-a) & \text{als } a\le x\le b\\ 0 & \text{als } x>b\end{cases}$

verwachting: $E(\underline{x})=\frac{1}{2}(a+b)$
variantie: $Var(\underline{x})=\frac{1}{12}(b-a)^2$
standaardafwijking: $\sigma(\underline{x})=\sqrt{Var(\underline{x})}=\frac{1}{\sqrt{12}}(b-a)$

Negatief-exponentiële verdeling met parameter λ

verdelingsfunctie: $F(x)=P(\underline{x}\le x)=1-e^{-\lambda x} \quad (x\ge 0)$
kansdichtheidsfunctie: $f(x)=\lambda e^{-\lambda x} \quad (x\ge 0)$
verwachting: $E(\underline{x})=\frac{1}{\lambda}$
variantie: $Var(\underline{x})=\frac{1}{\lambda^2}$
standaardafwijking: $\sigma(\underline{x})=\sqrt{Var(\underline{x})}=\frac{1}{\lambda}$

Normale verdeling met parameters μ en σ

kansdichtheidsfunctie: $n_{\mu,\sigma}(x)=\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
verdelingsfunctie: $N_{\mu,\sigma}(x)=\int_{-\infty}^{x}\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}\,dt$
verwachting: $E(\underline{x})=\mu$
variantie: $Var(\underline{x})=\sigma^2$
standaardafwijking: $\sigma(\underline{x})=\sqrt{Var(\underline{x})}=\sigma$