Syllabus bij het college

Kansrekening en Statistiek voor Kunstmatige Intelligentie

Bert van Es

Korteweg-de Vries Instituut
Universiteit van Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam

10 oktober 2014

Inhoudsopgave

Hoofdstuk 1

Introductie

Als men bezig is met het ontwerpen van nieuwe methoden in de kunstmatige intelligentie is het onvermijdelijk dat er gewerkt moet worden met metingen en processen die aan toeval onderhevig zijn. De kansrekening is de tak van de wiskunde die het gedrag van dit soort toevalsprocessen beschrijft. De statistiek is de wetenschap van het conclusies trekken uit onzekere gegevens, gebruik makende van kanstheoretische modellen.

Bij de problemen die je tegenkomt zullen natuurlijk veel situaties zijn die je standaard statistische problemen kan noemen. We geven hieronder echter een aantal voorbeelden specifiek uit de praktijk van de kunstmatige intelligentie, met name het evalueren van machine learning methoden.

- *Statistiek voor integratie van informatie.* Stel dat we een systeem willen maken waarmee we kunnen bepalen waar we zijn. Het systeem wordt ergens gedropt en ziet een aantal aanwijzingen. Deze aanwijzingen duiden op een aantal mogelijke plaatsen. We kunnen dit beschrijven als voorwaardelijke kansen: $P(\text{plaats}|\text{ observatie})$. Hoe kunnen we deze informatie combineren tot een soort gecombineerde $P(\text{plaats}|\text{ alle observaties})$? Wat als we van te voren een verwachting hebben over de plaatsen waar we gedropt worden zodat sommige veel waarschijnlijker zijn?

- *Statistiek voor de intelligente pokerspeler.* Een belangrijk element bij pokeren, en ook bij echte onderhandelingen, is het begrijpen van de strategie van de tegenpartij. Als de tegenstander bij pokeren zich random gedraagt dan valt er niet veel te doen. Als een tegenstander een bepaalde strategie volgt dan kunnen we proberen daar gebruik van te maken. Als we bijvoorbeeld de condities kennen waaronder hij bluft, dan kunnen we dat exploiteren. We kunnen nu het gedrag van de tegenstander bij een aantal spelen observeren en daar zijn strategie uit proberen te herkennen. De strategie kunnen we modelleren met kansen, want het zal geen strikte, keiharde strategie zijn. Een methode hiervoor is kijken naar een aantal kenmerken van één spel, of een reeks van spelletjes en te kijken of die samenhangen met bluffen, ofwel: $P(\text{bluft}|\text{ 2 azen})$, $P(\text{bluft}|\text{ in 3 achtereenvolgende spelletjes geen azen})$. Hier weer de vraag hoe we die informatie combineren zodat we ermee kunnen voorspellen. Stel dat we de kans op het voorkomen van strategien kennen. Kunnen we die gebruiken?

Je ziet aan deze voorbeelden dat de begrippen kans en voorwaardelijke kans een belangrijke rol kunnen spelen. Een ander voorbeeld in deze context is taalanalyse waar kanstheoretische modellen voor taal gebruikt worden voor spellingcorrectie en taalherkenning. Met name rekenregels die de verbanden tussen de kansen en voorwaardelijke kansen beschrijven zijn nodig. De bekendste rekenregel is de *Regel van Bayes*.

We vervolgen met twee voorbeelden die van een totaal andere aard zijn.

- *Vergelijken van methoden voor Machine Learning.* We hebben een idee voor een verbetering van een leeralgorithme. Beide algorithmen leren om objekten te classificeren, bijvoorbeeld om patiënten te classificeren naar of ze een ziekte hebben. We implementeren het. Om het te kunnen publiceren moeten we aantonen dat het beter leert dan de oorspronkelijke methode. We proberen beide methodes uit op dezelfde data, N beschrijvingen waarvan we de klasse kennen. De proportie van de data die de oude methode juist herkent is 75%. De nieuwe methode doet er 78% goed. Is dit echt beter of komt het verschil door de toevallige samenstelling van de steekproef?

  We kunnen dezelfde vraag stellen voor een systeem dat een numerieke waarde leert voorspellen, bijvoorbeeld het energieverbruik voor de komende week. We gebruiken beide systemen om een aantal weken het verbruik te voorspellen en meten het echte gebruik. Elke week bepalen we de grootte van de voorspellingsfout. Stel dat 'ons' systeem het beter lijkt te doen en gemiddeld een kleinere fout maakt. Is dat toeval of mogen we aannemen dat het echt beter is?

- *Vergelijken van robot controllers.* In plaats van een leermethode hebben we een systeem gemaakt dat een robot bestuurt. De taak van de robot is om zo snel mogelijk een bepaalde weg af te leggen. Onze claim is dat onze controller beter is dan de standaard controller. We zoeken een aantal routes en laten die door dezelfde robot, maar met verschillende controllers afleggen. We meten hoeveel minuten ze erover doen. Onze robot is gemiddeld wat sneller, maar kom dat door de toevallige selectie van routes?

Dit zijn voorbeelden van het onderzoeken of een nieuwe methode beter is dan een bestaande methode. Dit is natuurlijk het eerste dat je wilt weten als je een nieuwe methode hebt ontwikkeld. De belangrijke vraag is of de methode echt beter is, ook na uitschakeling van de variatie in de metingen. De statistiek levert methoden om daar iets over te zeggen.

Het laatste voorbeeld is een voorbeeld van een classificatieprobleem. Op zich is dat een klassiek probleem uit de statistiek. Echter betrekkelijk recent zijn er voor dit soort problemen, en ook voor regressieproblemen, in de kunstmatige intelligentie zogenaamde neurale netwerken ontwikkeld. Van die moderne methoden kan men zich afvragen of ze beter zijn dan bestaande klassieke statistische methoden.

- *Hoe goed werkt een Machine Learning methode?* Stel dat we ziektes willen leren herkennen. We hebben een AI systeem gebouwd dat data van een aantal patiënten als input neemt en een voorspellingsmodel leert. Het voorspellingsmodel wordt gebruikt om nieuwe patiënten te classificeren naar of ze een bepaalde ziekte hebben. Onze opdrachtgever wil graag weten hoe goed het systeem is. We proberen het uit op een aantal (N) beschrijvingen van patiënten, waarvan we achteraf weten of ze de ziekte hadden. Het systeem blijkt

90% goed te voorspellen. Hebben we nu een goede methode ontworpen of zijn er betere methoden?

In deze syllabus bouwen we het kader van de kansrekening op en geven we een basis van de statistiek, met name schattingstheorie en toetsingstheorie.

Hoofdstuk 2

Kansruimten

2.1 Kansruimten

Als we eigenschappen van kansexperimenten willen onderzoeken moeten we ze eerst kunnen beschrijven met een wiskundig model. De eerste stap daartoe is het invoeren van het begrip **kansruimte** (probability space). Wanneer we experimenten beschouwen, kunnen we twee typen onderscheiden:

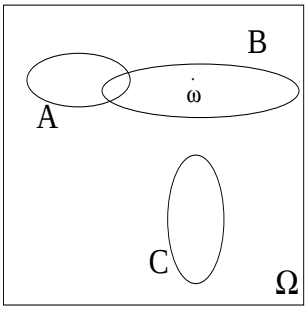
- deterministische experimenten, waarbij de uitkomst van tevoren vastligt,
- toevalsexperimenten, waarbij de uitslag van het 'toeval' afhangt.

Het eerste type experiment, waar ogenschijnlijk geen kansmechanisme aanwezig is, is in feite een speciaal ontaard geval van de het tweede type waar het kansmechanisme een essentiële rol speelt.

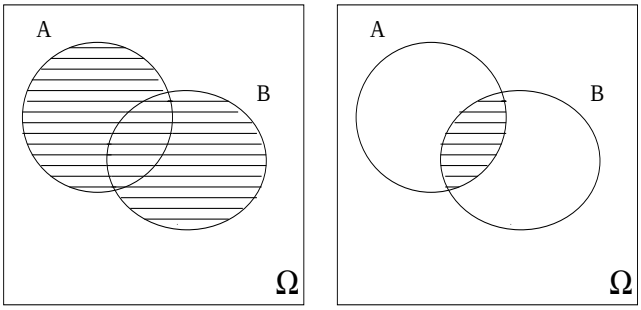
Een **kansruimte** bestaat uit een drietal objecten:

1. Ω: de **uitslagenruimte** of steekproefruimte (sample space). Dit is de verzameling van mogelijke uitkomsten van het experiment. De verzameling wordt doorgaans beschreven door de griekse hoofdletter omega, Ω. De mogelijke uitkomsten zelf, de elementen van de verzameling Ω, geven we aan met de kleine letter omega, ω.
2. $\mathcal{A}$: een collectie deelverzamelingen van Ω. Deze deelverzamelingen worden **eventualiteiten** of gebeurtenissen (events) genoemd. Een deelverzameling A van de uitkomstenruimte Ω beschrijft een gebeurtenis 'ω ligt in A'. Vaak kan die gebeurtenis in gewone woorden omschreven worden. Denk bijvoorbeeld aan gebeurtenissen 'de uitkomst (ω) is een even getal' of 'de uitkomst (ω) is kleiner dan een'.
3. P: een **kansmaat** (probability measure) op $\mathcal{A}$, die aan elke $A \in \mathcal{A}$ een **kans** $P(A)$, een getal, geeft. Voor elke gebeurtenis kunnen we dan spreken van 'de kans op A'.

Als we nu gebeurtenissen A en B hebben dan zijn we eigelijk ook geïnteresseerd in de gebeurtenissen $A \cup B$, de **vereniging** van A en B, en $A \cap B$, de **doorsnede** van A en B. De vereniging is namelijk de gebeurtenis 'A of B treedt op', eventueel alletwee, en de doorsnede is de gebeurtenis 'A en B treden op'.



Figuur 2.1: Schematische voorstelling van een kansruimte waarbij drie gebeurtenissen zijn aangegeven, A, B en C, en een uitkomst ω.



Figuur 2.2: Vereniging van A en B (links) en de doorsnede van A en B (rechts).

Voorbeeld 2.1 (Munt)

Bij het werpen van een munt, met als uitkomsten kruis (k) of munt (m), kunnen we de volgende keuzes maken: $\Omega = \{\text{k,m}\}$, $\mathcal{A}$ bestaat uit alle deelverzamelingen van Ω. Dus $\mathcal{A} = \{\phi, \{k\}, \{m\}, \{k, m\}\}$. Hierbij geeft ϕ de lege verzameling aan. Voor de kansen kiezen we $P(\phi) = 0, P(\{\text{k}\}) = \frac{1}{2}, P(\{\text{m}\}) = \frac{1}{2}$ en $P(\{\text{k,m}\}) = 1$.

Voorbeeld 2.2 (Dobbelsteen)

Bij het werpen van één dobbelsteen, waarbij we geïnteresseerd zijn in het aantal ogen, kunnen we de volgende keuzes maken: $\Omega = \{1, 2, \ldots, 6\}$, $\mathcal{A}$ bestaat uit alle deelverzamelingen van Ω, $P(\{i\}) = \frac{1}{6}$ voor $i = 1, 2, \ldots, 6$. Later zullen we zien dat dan ook alle kansen vastliggen.

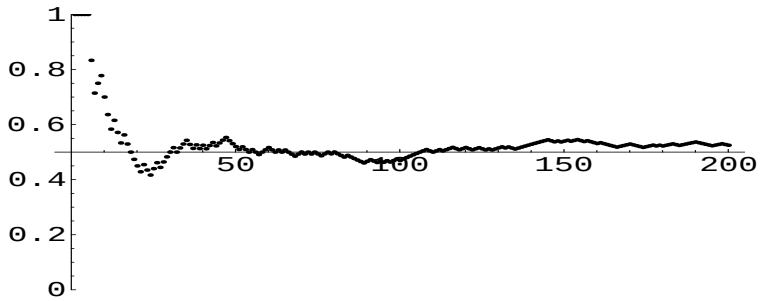
Dat we de kansruimten in deze voorbeelden een redelijke keuze vinden, stoelt op onze intuïtie en ervaring, die we kunnen samenvatten in de volgende **empirische wet van de grote aantallen.** We voeren een gegeven toevalsexperiment n keer uit en geven met $f_n(A)$ de relatieve frequentie aan van het optreden van de eventualiteit A. We definiëren dus

$$f_n(A) = \frac{\text{aantal keren dat } A \text{ optreedt bij de } n \text{ experimenten}}{n}. \tag{2.1}$$

Ons 'gevoel over kansexperimenten' zegt, dat als we dit doen voor $n = 1, 2, \ldots$, dat we dan een rij getallen $f_1(A), f_2(A), \ldots$ krijgen, die allerlei fluctuaties zal vertonen, maar voor groter

wordende n steeds dichter bij een zekere waarde $f(A)$ zal komen te liggen. Hoewel we in werkelijkheid $f(A)$ niet kunnen bepalen, willen we toch graag in ons model $P(A)$ gelijk aan $f(A)$ kiezen. Op deze manier kunnen we dus voor iedere gebeurtenis A de kans op A, genoteerd met $P(A)$ invoeren. Als we ons wiskundig model voor het kansexperiment helemaal opgebouwd hebben kan dat 'gevoel', de convergentie van de relatieve frequenties, wiskundig bewezen worden.

Als voorbeeld kunnen we bijvoorbeeld denken aan worpen met een 'eerlijke' munt. Als we dit een aantal keer doen dan kunnen we de relatieve frequentie van de gebeurtenis 'kruis' berekenen. We kunnen dit experiment op een computer simuleren met een random number generator. In Figuur 2.3 worden de relatieve frequenties weergegeven. We zien dat die frequenties inderdaad dicht bij de waarde 0.5 komen.



Figuur 2.3: De relatieve frequenties $f_n(A), n = 1, \ldots, 200$, van de gebeurtenis A, uitkomst "kruis", bij een rij van 200 gesimuleerde worpen met een eerlijke munt.

Als we kansen op deze manier invoeren dan volgen er ook meteen een aantal eigenschappen uit. Relatieve frequenties $f_n(A)$ en dus ook $f(A)$ hebben de volgende eigenschappen:

$$0 \leq f_n(A) \leq 1, \text{ voor alle gebeurtenissen } A,$$

$$f_n(\Omega) = 1,$$

$$f_n(A_1 \cup \ldots \cup A_m) = f_n(A_1) + \ldots + f_n(A_m),$$

als $A_1, A_2, \ldots, A_m$ disjuncte eventualiteiten zijn. Met disjunctie eventualiteiten bedoelen we eventualiteiten die geen elementen gemeenschappelijk hebben. Dus $A_i A_j = A_i \cap A_j = \phi$ voor $i \neq j, i, j = 1, \ldots m$.

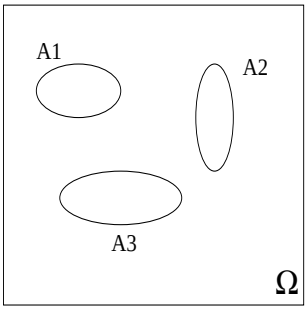
We kunnen nu omschrijven wat we onder een kansruimte verstaan.

Definitie 2.3 *Het drietal $(\Omega, \mathcal{A}, P)$ is een kansruimte als het volgende geldt:*

1. *$\mathcal{A}$ is een collectie deelverzamelingen van Ω met*

   D1 *de uitkomstenruimte Ω zelf is een gebeurtenis,*

   D2 *als A een gebeurtenis is dan is het complement van A, genoteerd met A^c, dat ook,*



Figuur 2.4: Drie disjuncte gebeurtenissen, A_1, A_2 en A_3.

D3 *als* $A_1, \ldots, A_m$ *gebeurtenissen zijn dan is* $A_1 \cup \ldots \cup A_m$*, de vereniging van deze gebeurtenissen, dat ook.*

2. de kansmaat P *kent getallen toe aan gebeurtenissen zódat*

D4 $0 \leq P(A) \leq 1$ *voor alle gebeurtenissen* A*,*

D5 $P(\Omega) = 1$*,*

D6 $P(A_1 \cup \ldots \cup A_m) = P(A_1) + \ldots + P(A_m)$*, als* $A_1, A_2, \ldots A_m$ *disjunct zijn.*

In feite eisen we $D3$ en $D6$ ook voor m gelijk aan oneindig. Voor kansruimten met eindig veel mogelijke uitkomsten maakt dat niet uit want dan zijn er maar eindig veel gebeurtenissen.

Uit deze definitie kunnen we de volgende eigenschappen van $\mathcal{A}$ en P afleiden.

E1: De lege verzameling ϕ is een gebeurtenis en $P(\phi) = 0$,

E2: Als A een gebeurtenis is dan is het complement van A per definitie ook een gebeurtenis en

$$P(A^c) = 1 - P(A), \tag{2.2}$$

E3: Als A en B gebeurtenissen zijn dan is $A \cap B$ ook een gebeurtenis en

$$P(A \cup B) = P(A) + P(B) - P(A \cap B), \tag{2.3}$$

Voorbeeld 2.4 (Alternatief)

Bij een toevalsexperiment met slechts twee mogelijke uitkomsten spreken we van een **alternatief** en noemen we de twee uitkomsten **succes en mislukking** of kruis en munt of 1 en 0. We kiezen $\Omega = \{0, 1\},\ \mathcal{A} = \{\phi, \{0\}, \{1\}, \Omega\}, P(\{1\}) = p$. Met de **succeskans** p ligt de kansmaat P vast.

Voorbeeld 2.5 (Discrete uniforme kansruimte)

Bij een toevalsexperiment met N mogelijke uitkomsten kiezen we $\Omega = \{1, 2, \ldots, N\}$ en $\mathcal{A}$ de collectie van alle deelverzamelingen van Ω. De kansmaat P ligt volledig vast door de $N - 1$ getallen $P(\{1\}), P(\{2\}) \ldots, P(\{N-1\})$. Als alle **simpele eventualiteiten** even waarschijnlijk zijn, geldt $P(\{i\}) = \frac{1}{N},\ i = 1, \ldots, N$, en spreken we van een **aselecte trekking uit** Ω. Dan geldt voor elke gebeurtenis A

$$P(A) = \sum_{i \in A} P(\{i\}) = \frac{\text{aantal elementen van } A}{\text{aantal elementen van } \Omega} = \frac{\text{aantal elementen van } A}{N}.$$

Een dergelijke kansruimte noemen we een **discrete uniforme kansruimte**. Speciale gevallen zijn $N = 2$ (voorbeelden 2.1 en 2.4) en $N = 6$ (voorbeeld 2.2). Het berekenen van kansen bij deze kansruimten is dus een zaak van tellen.

Voorbeeld 2.6 (Taalverwerking)

Een toepassing van kansrekening in de kunstmatige intelligentie vind je bijvoorbeeld bij taalverwerking. In het bijzonder kan je denken aan het verwerken van zinnen. In een **probabilistisch taalmodel** bestaat de uitkomstenruimte Ω uit alle mogelijke zinnen. We gaan er hier even van uit dat dat er eindig veel zijn. Hier zitten dus ook alle mogelijke woorden bij want dat die zien we als zinnen bestaande uit een woord. Aan elke zin wordt een kans toegekend. De som van de kansen op die zinnen is gelijk aan een. Als gebeurtenissen kunnen we hier weer alle deelverzamelingen van Ω nemen. Met dit soort taalmodellen maakt men bijvoorbeeld procedures om de spelling te corrigeren, om de taal te herkennen, of om het volgende woord in een zin te voorspellen op grond van de voorgaande woorden.

Tot nu hebben we als voorbeeld alleen uitkomstenruimten Ω gezien met eindig veel uitkomsten. Het is niet moeilijk een experiment te vinden met oneindig veel mogelijke uitkomsten.

Voorbeeld 2.7

Stel je voor dat we herhaaldelijk eerlijke worpen met een munt doen. De kans op kruis en munt is dan steeds gelijk aan een half. We bekijken het volgende experiment. Je noteert het aantal worpen dat nodig is tot en met de eerste keer dat je munt gooit. Dus de rij k, k, k, k, m levert

als uitkomst vijf op. De uitkomstenruimte Ω is nu gelijk aan de verzameling $\{1, 2, 3, \ldots\}$ en heeft dus oneindig veel elementen.

Laten we de kansen berekenen. De uitkomst 1 betekent dat de eerste worp meteen munt moet zijn. Dus

$$P(\{1\}) = \tfrac{1}{2}.$$

De uitkomst 2 betekent dat de eerste worp kruis moet zijn en de tweede munt. Het totaal aantal mogelijke uitkomsten voor de eerste twee worpen is 4. Deze uitkomsten hebben allemaal gelijke kans. Dus, zie Voorbeeld 2.5,

$$P(\{2\}) = \tfrac{1}{4}.$$

Algemeen, voor een uitkomst k groter dan een geldt net zo dat de kans op de uitkomst k gelijk is aan de kans op eerst $k-1$ keer kruis en dan in de k-de worp munt. Voor de eerste k worpen zijn er nu 2^k mogelijke uitkomsten die allemaal gelijke kans hebben. We vinden nu dus

$$P(\{k\}) = \frac{1}{2^k} = 2^{-k}.$$

Als we nu deze uitkomsten beschrijven met gebeurtenissen $A_k = \{k\}$, voor $k = 1, 2, \ldots$, dan hebben we $P(A_k) = 2^{-k}$ en moet volgens regel D6, met m gelijk aan oneindig, gelden

$$\begin{aligned} 1 = P(\Omega) &= P(\{1, 2, \ldots\}) = P(\{1\} \cup \{2\} \cup \ldots\}) = P(A_1 \cup A_2 \cup \ldots) \\ &\overset{D6}{=} P(A_1) + P(A_2) + \ldots. \end{aligned}$$

Dat de som van deze kansen inderdaad gelijk is aan een volgt uit

$$\tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} + \ldots = \tfrac{1}{2}(1 + \tfrac{1}{2} + (\tfrac{1}{2})^2 + (\tfrac{1}{2})^3 + \ldots) = \frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1.$$

We gebruiken hierbij de meetkundige reeks, met $-1 < r < 1$,

$$1 + r + r^2 + r^3 + r^4 + \ldots = \frac{1}{1-r}. \tag{2.4}$$

Misschien overbodig toch even een bewijsje van (2.4). Schrijf $s_n = 1 + r + r^2 + r^3 + r^4 + \ldots + r^n$. Dan geldt

$$rs_n = r(1 + r + r^2 + r^3 + r^4 + \ldots + r^n) = r + r^2 + r^3 + r^4 + \ldots + r^{n+1}$$

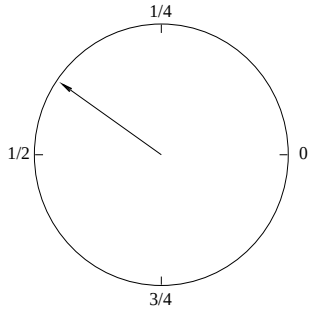
en dus $(1-r)s_n = s_n - rs_n = 1 - r^{n+1}$. We vinden dan

$$s_n = \frac{1 - r^{n+1}}{1-r} \to \frac{1}{1-r},$$

voor n naar oneindig, omdat $r^{n+1} \to 0$ ($|r| < 1$!).

We zien in dit voorbeeld dat oneindige uitkomstenruimten nodig zijn en dat we regel D6 voor de kansmaat, met m gelijk aan oneindig, ook nodig hebben.

Als, zoals in de voorbeelden tot hier toe, Ω een eindige verzameling of aftelbare verzameling (daarmee bedoel je dat je de elementen op een rij kan leggen) is en $\mathcal{A}$ uit alle deelverzamelingen van Ω bestaat, is er sprake van **discrete** kansruimten. Bij het volgende voorbeeld kunnen we de collectie van gebeurtenissen niet gelijk kiezen aan 'alle deelverzamelingen'. In die gevallen zijn er te veel deelverzamelingen en komen we in de wiskundige problemen. Wel kan men $\mathcal{A}$ zo kiezen dat alle in de praktijk voorkomende verzamelingen, zoals bijvoorbeeld intervallen, gebeurtenissen zijn, en dus kansen hebben.



Figuur 2.5: Rad van fortuin.

Voorbeeld 2.8 (Rad van fortuin)

Een wijzer kan om een verticale as draaien, waarbij de punt van de wijzer een cirkel beschrijft waarlangs een lineaire schaalverdeling is aangebracht, die van 0 tot 1 loopt. Men geeft de wijzer een flinke beginsnelheid, wacht tot hij weer stilstaat en leest dan af welk punt op de schaalverdeling wordt aangewezen. We kiezen $\Omega = [0, 1)$. Op grond van de symmetrie kiezen we P zó dat $P([a, b)) = b - a$, waarbij $0 \leq a < b < 1$. Deze kansruimte beschrijft een **aselecte trekking uit** Ω.

2.2 Voorwaardelijke kansen en de regel van Bayes

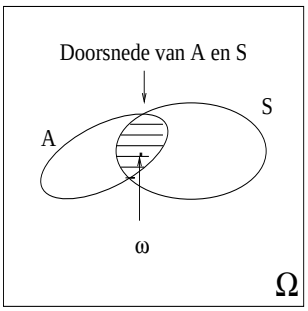
Beschouw een experiment dat beschreven kan worden door een kansruimte $(\Omega, \mathcal{A}, P)$. A en S zijn twee gebeurtenissen. De uitkomst geven we als gebuikelijk aan met ω. We hebben eerder gezien dat we de kans op A, dus $P(A)$, kunnen invoeren als limiet $f(A)$ van de relatieve frequenties $f_n(A)$ van A in een rij herhaalde experimenten. Stel nu eens dat er gegeven is dat de uitkomst ω in S ligt en dat S een gebeurtenis is met positieve kans, niet nul dus. Wat is dan de kans op A? Die kans kunnen we op soortgelijke manier afleiden.

Herhaal het experiment n keer. Haal de uitkomsten die niet in S liggen uit de rij. We houden dan $nf_n(S)$ uitkomsten over. Daarvan liggen er $nf_n(A \cap S)$ ook in A. In de uitgedunde

rij is de relatieve frequentie dan gelijk aan

$$\frac{nf_n(A \cap S)}{nf_n(S)} = \frac{f_n(A \cap S)}{f_n(S)}. \tag{2.5}$$

Volgens ons gevoel zou dit weer moeten convergeren naar $P(A \cap S)/P(S)$. Hier is het dus belangrijk dat $P(S)$ niet gelijk is aan nul! Dit leidt tot de volgende definitie van **voorwaardelijke kans op** A **gegeven** S (conditional probability of A given S).



Figuur 2.6: De gebeurtenissen A en S en de uitkomst ω.

Definitie 2.9 *Als $P(S) > 0$ dan definiëren we de voorwaardelijke kans op A gegeven S, genoteerd als $P(A|S)$, als volgt*

$$P(A|S) = \frac{P(A \cap S)}{P(S)}. \tag{2.6}$$

We kunnen nu drie belangrijke rekenregels afleiden. Eerst doen we de **wet van de totale waarschijnlijkheid**. Laten A en B twee gebeurtenissen zijn en laat $P(B)$ niet nul of een zijn. Er geldt

$$A = (A \cap B) \cup (A \cap B^c),$$

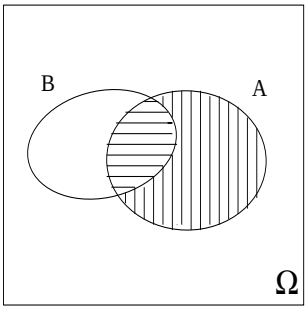
zie Figuur 2.7. A kan dus geschreven worden als vereniging van twee disjuncte gebeurtenissen. Dus volgens regel D6 geldt er dan

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B^c) = P(B)\frac{P(A \cap B)}{P(B)} + P(B^c)\frac{P(A \cap B^c)}{P(B^c)} \\ &= P(B)P(A|B) + P(B^c)P(A|B^c). \end{aligned}$$

We vinden dus

$$\boxed{P(A) = P(B)P(A|B) + P(B^c)P(A|B^c).} \tag{2.7}$$

Voorbeeld 2.10 (Totale waarschijnlijkheid)



Figuur 2.7: A opgesplist als $A \cap B$ verenigd met $A \cap B^c$.

Stel dat we een emmer hebben met 10 ballen waarvan er 4 rood zijn en 6 wit. We trekken willekeurig (aselect) een bal uit de emmer en noteren de kleur. Deze bal leggen we opzij en vervolgens trekken we een tweede bal uit de negen overgebleven ballen en noteren weer de kleur. De uitkomsten van het experiment zijn dus de twee mogelijke kleuren die je getrokken hebt. We bekijken de gebeurtenissen A 'wit in de tweede trekking' en B 'wit in de eerste trekking'. De volgende kansen kunnen we zo uitrekenen

$$P(B) = \tfrac{6}{10}, P(A|B) = \tfrac{5}{9}, P(A|B^c) = \tfrac{6}{9}.$$

Door de wet op de totale waarschijnlijkheid te gebruiken kunnen we deze kansen samenvoegen tot de kans op A. We vinden dan

$$P(A) = P(B)P(A|B) + P(B^c)P(A|B^c) = \tfrac{6}{10}\,\tfrac{5}{9} + \tfrac{4}{10}\,\tfrac{6}{9} = \tfrac{3}{5}.$$

Een hele belangrijke rekenregel is de **regel van Bayes**. Laten A en B weer twee gebeurtenissen zijn. Laat $P(A)$ niet nul zijn en $P(B)$ niet nul of een. Er geldt nu

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

Door (2.7) in te vullen kunnen we dit verder uitwerken tot

$$\boxed{P(B|A) = \tfrac{P(A|B)P(B)}{P(A)} = \tfrac{P(A|B)P(B)}{P(B)P(A|B)+P(B^c)P(A|B^c)}.} \tag{2.8}$$

Voorbeeld 2.11 (Regel van Bayes)

Bekijk weer het experiment van Voorbeeld 2.10. Nu willen we kans op 'wit in de eerste trekking, gegeven wit in de tweede trekking weten', dus $P(B|A)$. Uit de regel van Bayes volgt nu

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{\frac{5}{9}\,\frac{6}{10}}{\frac{3}{5}} = \tfrac{5}{9}.$$

We zien hier dus dat het heel handig is dat deze regel in feite 'A en B verwisselt' in een voorwaardelijke kans.

Voorbeeld 2.12 (Regel van Bayes)

Beschouw de situatie waarin een willekeurige persoon getest wordt op een bepaalde ziekte. Gebeurtenis A is de gebeurtenis dat de test positief is. Gebeurtenis B is de gebeurtenis dat de persoon de ziekte echt heeft. Interessant is dan de voorwaardelijke kans $P(B|A)$, de kans dat de persoon de ziekte echt heeft, gegeven dat de test positief is. Stel dat we voor deze ziekte de volgende kansen weten

$$P(B) = 0.0001, P(A|B) = 0.9, P(A|B^c) = 0.001.$$

We hebben het dus over een zeldzame ziekte. De voorwaardelijke kansen suggereren dat we de test wel kunnen vertrouwen. Volgens de regel van Bayes geldt echter

$$P(B|A) = \frac{P(A|B)P(B)}{P(B)P(A|B) + P(B^c)P(A|B^c)} = \frac{0.9 \times 0.0001}{0.0001 \times 0.9 + 0.9999 \times 0.001} = 0.083.$$

Deze kans is dus onverwacht klein. Dit verschijnsel geldt voor zeldzame ziekten, dus als $P(B)$ heel klein is. Voor meer voorkomende ziekten komt er een redelijke waarde uit.

Als laatste hebben we de **productregel**. Als $A_1, \ldots, A_m$ gebeurtenissen zijn met $P(A_1 \cap \ldots \cap A_m) > 0$ dan geldt er, na uitschrijven van de voorwaardelijke kansen,

$$\boxed{\begin{array}{rcl} P(A_1 \cap \ldots \cap A_m) & = & P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)P(A_4|A_1 \cap A_2 \cap A_3)\times \\ & & \cdots \times P(A_m|A_1 \cap \ldots \cap A_{m-1}). \end{array}} \tag{2.9}$$

We noteren dit product als

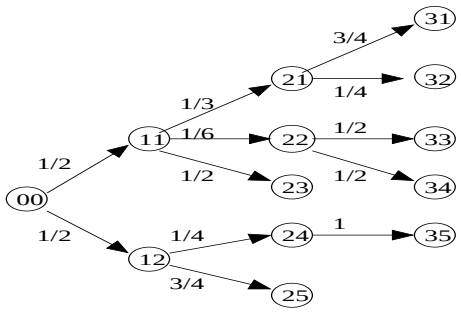
$$P(A_1 \cap \ldots \cap A_m) = P(A_1) \prod_{i=2}^{m} P(A_i|A_1 \cap \ldots \cap A_{i-1}). \tag{2.10}$$

Voorbeeld 2.13 (Productregel)

Bekijk weer het experiment van Voorbeeld 2.10. Laten we nu de kans op $A \cap B$ uitrekenen, de kans op wit in de eerste én de tweede worp. Voor twee gebeurtenissen zegt de productregel

$$P(A \cap B) = P(B)P(A|B) = \tfrac{6}{10}\,\tfrac{5}{9} = \tfrac{1}{3}.$$

Voorbeeld 2.14 (Productregel)



Figuur 2.8: Een boom met overgangskansen.

Beschouw een boom als in Figuur 2.8. Die boom bestaat uit een aantal knopen en gerichte verbindingen. In de boom staan overgangskansen aangegeven. Een overgangskans geeft de voorwaardelijke kans dat je naar een zekere knoop toegaat, gegeven dat je in een gegeven knoop bent. Die kans hangt dan alleen af van de knoop waarin je bent en niet van de eerdere knopen waar je geweest bent. Deze eigenschap van de kansen heet de **Markov eigenschap**.

Laten we de kans uitrekenen dat we na drie stappen in knoop 34 zijn. Voer de volgende gebeurtenissen in: A_1='in stap een ben je in knoop 11', A_2='in stap twee ben je in knoop 22' en A_3='in stap drie ben je in knoop 34'. Uit de productregel volgt nu

$$P(\text{de eindknoop is } 34) = P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) = \tfrac{1}{2}\,\tfrac{1}{6}\,\tfrac{1}{2} = \tfrac{1}{24}.$$

Om de kans op een pad uit te rekenen mag je de overgangskansen dus gewoon vermenigvuldigen.

Voorbeeld 2.15 (Taalverwerking, productregel)

Bij probabilistische taalverwerking, zie Voorbeeld 2.6, krijgen alle mogelijke zinnen een kans. We trekken nu een zin uit Ω. De kansen op de verschillende zinnen zijn gegeven door de kansmaat P. Stel een gegeven zin bestaat uit de n woorden $w_1, \ldots, w_n$. Bekijk de gebeurtenissen W_i='het i-de woord is gelijk aan w_i'. Dan geldt er volgens de productregel

$$P(\text{de getrokken zin is } w_1, \ldots, w_n) = P(W_1) \prod_{i=2}^{n} P(W_i|W_1 \cap \ldots \cap W_{i-1}).$$

Dit model wordt eenvoudiger als we een Markov eigenschap veronderstellen. Men veronderstelt bijvoorbeeld

$$P(W_i|W_1 \cap \ldots \cap W_{i-1}) = P(W_i|W_{i-k} \cap \ldots \cap W_{i-1}).$$

In andere woorden betekent dit dat alleen de laatste k woorden van de zin er toe doen als je de kans op het volgende woord wil weten.

2.3 Onafhankelijkheid van gebeurtenissen

Beschouw twee gebeurtenissen A en B, met $P(B) > 0$, waarvoor geldt dat de voorwaardelijke kans op A gegeven B, dus $P(A|B)$, gelijk is aan de onvoorwaardelijke kans $P(A)$ op A. Kennelijk zegt informatie over het optreden van B dus niets over de kans op het optreden van A. Als we naar de definitie van voorwaardelijke kans (2.9) kijken dan geldt er hier kennelijk

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

Herschrijven we dit dan vinden we

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

In een dergelijke situatie noemen we de gebeurtenissen **onafhankelijk**.

Definitie 2.16 *Twee gebeurtenissen A en B noemen we onafhankelijk als geldt*

$$P(A \cap B) = P(A)P(B). \tag{2.11}$$

Een groep gebeurtenissen $A_1, \ldots, A_m$ noemen we onafhankelijk als

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}), \tag{2.12}$$

voor elk k-tal gebeurtenissen $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$, $k = 2, \ldots, m$ uit $A_1, \ldots, A_m$.

Voorbeeld 2.17

Bij het experiment van twee (eerlijke) worpen met een eerlijke munt geldt met A de gebeurtenis 'kruis in de eerste worp' en B de gebeurtenis 'munt in de tweede worp',

$$\tfrac{1}{4} = P(A \cap B) = P(A)P(B) = \tfrac{1}{2}\,\tfrac{1}{2}.$$

2.4 Geloofswaarden

Laten we enige opmerkingen maken bij de invoering van het kansbegrip. We hebben kansen ingevoerd als limiet van relatieve frequenties van het optreden van een gebeurtenis bij herhaald uitvoeren van het experiment, steeds onder gelijke omstandigheden. Deze benadering wordt daarom de **frequentistische benadering** genoemd. Dit herhalen van het experiment is natuurlijk vaak onmogelijk. Denk bijvoorbeeld aan 'de kans dat het morgen om een uur regent'. We kunnen moeilijk de wereldgeschiedenis een aantal keren opnieuw laten lopen vanaf de oerknal.

In plaats van de objectieve kansen, die uit de frequentistische benadering volgen, wordt er ook veel gebruik gemaakt van **subjectieve kansen**, of **geloofswaarden** (measures of belief). Om zo'n geloofswaarde te koppelen aan een objectieve kans wordt dan de volgende redenering

gevolgd. Stel je een experiment voor waarbij je aselect een bal trekt uit een emmer met witte en zwarte ballen. Dit is een experiment waar een objectieve kans op de gebeurtenis A='de getrokken bal is wit' gedefinieerd kan worden. Je wil nu een geloofswaarde voor bijvoorbeeld de gebeurtenis R='het regent morgen om een uur' invoeren. Er wordt je gevraagd wat je een waarschijnlijker gebeurtenis vind, A of R. Als er geen witte ballen in de emmer zitten kies je natuurlijk voor R en als er alleen maar witte ballen in de emmer zitten dan kies je voor A. Het idee is nu dat er een evenwichtwaarde is, dus dat er aantallen witte en zwarte ballen bestaan waarvoor je geen van beide gebeurtenissen waarschijnlijker vind dan de andere. Die waarde noem je dan de geloofswaarde. Omdat hij zo gekoppeld is aan een objectieve kans gelden er voor deze subjectieve kansen dezelfde rekenregels als voor objectieve kansen.

2.5 Opgaven

1. Bewijs eigenschappen E1, E2 en E3. Om aan te tonen dat $A \cap B$ een gebeurtenis is kan je gebruiken $A \cap B = (A^c \cup B^c)^c$, de zogenaamde regel van De Morgan.

2. Bij een experiment met uitkomstenruimte $\Omega = \{1, 2, \ldots, N\}$ worden de kansen op de simpele gebeurtenissen gegeven door $P(\{1\}), P(\{2\}) \ldots P(\{N\})$. Laat zien dat voor elke gebeurtenis A de kans op A nu vastligt door

   $$P(A) = \sum_{i \in A} P(\{i\}).$$

   Dat betekent dat je de kans op A kan uitrekenen door de kansen op elementen van A op te tellen.

3. Beschouw de gebeurtenissen A en B.

   (a) C is de gebeurtenis dat A optreedt, maar niet B. Druk C uit in termen van A en B.

   (b) D is de gebeurtenis dat A of B optreden maar niet beide tegelijk. Druk D uit in termen van A en B.

4. Als $P(A) = \frac{1}{3}, P(B) = \frac{1}{2}$ en $P(A \cup B) = \frac{3}{4}$, bereken dan

   (a) $P(A \cap B)$,

   (b) $P(A^c \cup B^c)$,

   (c) $P(A^c \cap B)$.

5. Een experiment heeft maar twee mogelijke uitkomsten. Een alternatief dus. De kans op de uitkomsten is p en p^2. Bepaal de waarde van p.

6. In Voorbeeld 2.12 vonden we een kleine waarde voor $P(B|A)$. Bereken deze waarde voor een test met gelijke testeigenschappen en voor ziekten met kans $P(B) = 0.1, 0.01$ en 0.001.

7. Trek willekeurig een kaart uit een standaard pak van 52 kaarten. Elke kaart heeft evenveel kans om getrokken te worden.

   (a) Wat is de kans op de gebeurtenis K=‘de kaart is klaveren’ ?

   (b) Wat is de voorwaardelijke kans op de gebeurtenis A=‘de kaart is een aas’ gegeven dat de kaart klaveren is?

   (c) Geeft het feit dat we weten dat de kaart een klaveren is ons extra informatie over de kans op het optreden van A?

8. Een autobedrijf test auto’s op het teveel uitstoten van vervuilende gassen. Een auto die teveel uitstoot noemen we een “vuile auto”. Stel dat we weten dat 25% van de te testen auto’s vuil is. De test geeft bij een vuile auto in 99% van de gevallen aan dat hij inderdaad vuil is, maar doet dit ook in 17% van de schone auto’s.

(a) Wat is de kans dat de test aangeeft dat een onderzochte auto vuil is ?

(b) Wat is de kans dat een auto vuil is als de test dit aangeeft ?

9. Een willekeurige student moet voor een bepaald vak tentamen doen. De kans dat hij slaagt voor het tentamen als hij het werkcollege gevolgd heeft is 0.8. De kans dat hij slaagt als hij geen werkcollege gevolgd heeft is 0.5. De kans dat hij het werkcollege heeft gevolgd is 0.7.

   (a) Wat is de kans dat de student slaagt ?

   (b) Wat is de kans dat de student het werkcollege heeft gevolgd als gegeven is dat hij voor het tentamen is geslaagd ?

10. We werpen tweemaal met een dobbelsteen. Bij dit experiment bekijken we de gebeurtenissen A='de som van de ogen is vier' en B='tenminste een van de worpen is drie'.

    (a) Zijn A en B onafhankelijk?

    (b) Bereken $P(A|B)$.

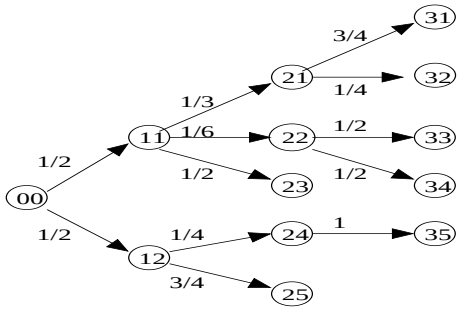
11. Beschouw het volgende experiment. We gooien met een dobbelsteen net zolang tot we voor de eerste keer 6 gooien en noteren het benodigde aantal worpen. Neem aan dat alle zijden van de dobbelsteen even waarschijnlijk zijn om boven te komen en dat alle worpen elkaar niet beïnvloeden.

    (a) Beschrijf een kansruimte voor dit experiment.

    (b) Bereken $P(B_m)$ als B_m de gebeurtenis voorstelt dat men ten hoogste m keer moet gooien.

    (c) Zij C_m de gebeurtenis dat men precies m keer moet gooien. Wat is $P(C_m)$?

12. Beschouw de boom met overgangskansen weergegeven door Figuur 2.9. Dit is de boom van het voorbeeld in Hoofdstuk 2. Stel je beweegt je op een stochastische manier door de boom en je eindigt in een van de eindpunten.



Figuur 2.9: Een boom met overgangskansen.

(a) Bereken de kans dat je in knoop 31 eindigt.

(b) Bereken de voorwaardelijke kans dat je in de eerste stap in knoop 11 bent, gegeven dat je eindigt in knoop 31.

(c) Bereken de voorwaardelijke kans dat je eindigt in knoop 31 , gegeven dat je in de eerste stap in knoop 11 bent.

(d) Zijn de gebeurtenissen 'je bent in de eerste stap in knoop 11' en 'je eindigt in knoop 31' onafhankelijk?

13. Laten de twee gebeurtenissen A en B onafhankelijk zijn. Laat zien dat dan ook de volgende paren gebeurtenissen onafhankelijk zijn:

(a) A en B^c,

(b) A^c en B,

(c) A^c en B^c.

Bij het laatste paar kun je gebruiken $A^c \cap B^c = (A \cup B)^c$.

14. Stel een student doet een multiple choice tentamen. Elke vraag heeft m keuzemogelijkheden. Als hij het goede antwoord weet vult hij dat inderdaad in. Als hij het antwoord niet weet kiest hij, zonder voorkeur, een willekeurig antwoord. Bekijk de volgende gebeurtenissen: A='de student weet het juiste antwoord', B='de student geeft het goede antwoord'. Neem aan dat hij minimaal heeft gestudeerd om te slagen en dat de kans dat hij het juiste antwoord weet gelijk is aan 0.6.

(a) Bereken, bij een gegeven aantal keuzemogelijkheden m, de kans dat hij het goede antwoord geeft.

(b) Bereken de kans dat hij het juiste antwoord wist, gegeven dat hij het goede antwoord heeft gegeven.

(c) Reken deze kansen uit voor $m = 4$ en $m = 10$.

(d) Stel nu eens dat de student lui is, of geen tijd heeft gehad om te leren voor het tentamen. Dit heeft tot gevolg dat de kans dat hij het goede antwoord weet nu gelijk is aan 0.25. Bereken nu weer de bovenstaande kansen voor $m = 4$ en $m = 10$.

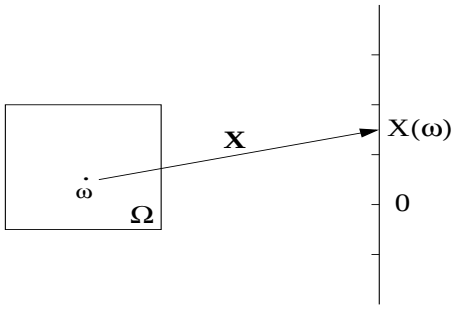
Hoofdstuk 3

Stochastische variabelen, kansverdelingen en verwachtingen

We gebruiken een kansruimte om een kansexperiment wiskundig te beschrijven. De uitkomstenruimte is daarbij de verzameling van mogelijke uitkomsten. Aan de aard van de uitkomsten hebben we geen beperking opgelegd. Met name hoeven ze geen getallen te zijn. Een nadeel daarvan is dat we in het algemeen niet kunnen rekenen met die uitkomsten. Als we het hebben over het experiment van het aselect kiezen van een student uit een bepaalde groep studenten, dan kunnen we twee uitkomsten, twee van die studenten, niet optellen. We kunnen het dan ook niet hebben over 'de gemiddelde student'. Eigenlijk weet je dan wel wat je met 'de gemiddelde student' bedoeld. Als je iets meet aan die student, bijvoorbeeld zijn lengte, dan is zijn lengte gelijk aan de gemiddelde lengte in de groep. Net zo voor zijn gewicht. Je ziet aan dat voorbeeld dus dat je vaak waarden gaat meten aan een uitkomst. Die waarden zijn dan getallen zodat we er mee kunnen rekenen.

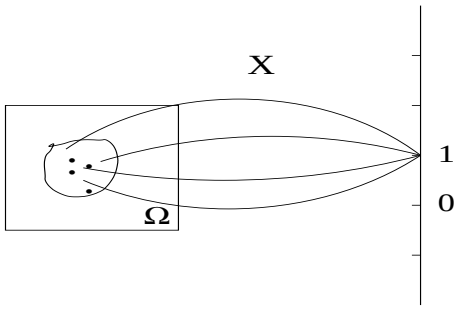
3.1 Stochastische variabelen

Veronderstel dat we een experiment bekijken met als kansruimte $(\Omega, \mathcal{A}, P)$. We beschrijven getalwaardige metingen aan de uitkomsten van dit experiment door middel van zogenaamde **stochastische variabelen** (random variables).



Figuur 3.1: Schematische voorstelling van een stochastische variabele.

Definitie 3.1 *Een stochastische variabele X kent aan elke mogelijke uitkomst ω van het experiment een getal $X(\omega)$ toe. Met andere woorden, een stochastische variabele X is een afbeelding van Ω naar de reële getallen.*



Figuur 3.2: Schematische voorstelling van het berekenen van $P(X = 1)$.

Als we een stochastische variabele hebben, willen we er ook kansen mee kunnen uitrekenen. We kunnen dit als volgt doen. Stel we willen de kans berekenen voor 'X neemt de waarde een aan'. Deze gebeurtenis, die we zullen noteren als $\{X = 1\}$, kunnen we wiskundig als volgt als deelverzameling van Ω beschrijven,

$$\{X = 1\} = \{\omega \in \Omega | X(\omega) = 1\} \subset \Omega.$$

In woorden staat hier 'alle uitkomsten ω waarvoor de X waarde $X(\omega)$ gelijk is aan een'. Als deze verzameling een gebeurtenis van het kansexperiment is, dus als ze in $\mathcal{A}$ zit, dan berekenen we de kans door de kans op de betreffende deelverzameling van Ω te berekenen. Dus

$$P(\{X = 1\}) = P(\{\omega \in \Omega | X(\omega) = 1\}).$$

Net zo kunnen we de kans dat X in een gegeven interval $[a, b]$, met $a < b$, valt, berekenen door middel van

$$P(\{X \in [a, b]\}) = P(\{a \leq X \leq b\}) = P(\{\omega \in \Omega | a \leq X(\omega) \leq b\}).$$

We zien hier dus dat we kansen voor de stochastische variabele X via de afbeelding 'ophalen' uit de kansruimte $(\Omega, \mathcal{A}, P)$.

Voor het gemak van de notatie laten we in het vervolg de accolades in uitdrukkingen als $P(\{X = 1\}), P(\{X \in [a, b]\})$ en $P(\{a \leq X \leq b\})$ weg, en schrijven we $P(X = 1), P(X \in [a, b])$ en $P(a \leq X \leq b)$.

Voorbeeld 3.2 (Tentamencijfers studenten)

Als voorbeeld van een kansruimte met een stochastische variabele bekijken we de deelnemers aan een bepaald tentamen. Er hebben 24 studenten meegedaan aan dit tentamen. Deze studenten noemen we even $\omega_1, \ldots, \omega_{24}$. Het basisexperiment bestaat uit het aselect trekken van een

Student	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	ω_{11}	ω_{12}
Uitslag	5	6.5	5	9.5	8	7	8	6.5	9.5	6.5	8.5	3.5
Student	ω_{13}	ω_{14}	ω_{15}	ω_{16}	ω_{17}	ω_{18}	ω_{19}	ω_{20}	ω_{21}	ω_{22}	ω_{23}	ω_{24}
Uitslag	3.5	9	9.5	3	4.5	8	7.5	6.5	8.5	4.5	9.5	7.5

Tabel 3.1: tentamencijfers.

student uit deze groep. De uitkomstenruimte Ω is dan gelijk aan $\{\omega_1, \ldots, \omega_{24}\}$. Elke individuele student heeft kans 1/24 om geselecteerd te worden. De uitkomsten zijn dus geen getallen.
We zijn misschien geïnteresseerd in de uitslagen van deze studenten, de tentamencijfers. Dit zijn dus getallen die toegekend zijn aan de uitkomsten van het oorspronkelijke experiment. Kortom, aan student ω_i kennen we zijn/haar tentamencijfer $X(\omega_i)$ toe. Dit tentamencijfer is dan een stochastische variabele. Met die stochastische variabele kunnen later gaan rekenen.

3.2 Discrete stochastische variabelen

Als de verzameling van mogelijke uitkomsten van X van de vorm $\{r_1, \ldots, r_k\}$ of $\{r_1, r_2, \ldots\}$ is, in andere woorden, als de uitkomstenruimte eindig of aftelbaar is, dan noemen we X een **discrete stochastische variabele**.

3.2.1 Kansverdeling

In het geval van eindig veel uitkomsten kunnen we de kansen van X in een matrix weergeven. We spreken dan over de **kansverdeling van** X, weergegeven door

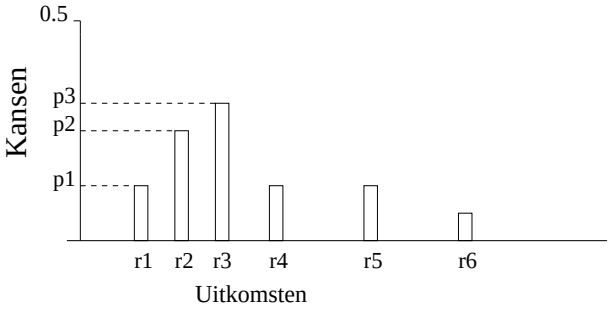
$$X : \begin{pmatrix} r_1 & r_2 & \ldots & r_k \\ p_1 & p_2 & \ldots & p_k \end{pmatrix}, \tag{3.1}$$

waarbij $p_1 = P(X = r_1), \ldots, p_k = P(X = r_k)$, en $p_1 + \ldots + p_k = 1$. Dat de som van deze kansen gelijk is aan een volgt volgens regel D6 van een kansruimte, met hier $\Omega = \{r_1, \ldots, r_k\}$, uit

$$\begin{aligned} 1 &= P(\Omega) = P(\{\omega \in \Omega | X(\omega) = r_1\} \cup \ldots \cup \{\omega \in \Omega | X(\omega) = r_k\}) \\ &\overset{D6}{=} P(\{\omega \in \Omega | X(\omega) = r_1\}) + \ldots + P(\{\omega \in \Omega | X(\omega) = r_k\}) \\ &= P(X = r_1) + \ldots + P(X = r_k) = p_1 + \ldots + p_k. \end{aligned}$$

We kunnen nu alle kansen die beschreven worden in termen van X uitrekenen als we de kansverdeling (3.1) van X kennen. In feite kunnen we dan de achterliggende kansruimte vergeten en alleen met de stochastische variabele X werken. Zo geldt er bijvoorbeeld

$$P(X = r_2 \text{ of } X = r_4) = p_2 + p_4.$$



Figuur 3.3: Schematische voorstelling van de kansverdeling van een stochastische variabele.

Voorbeeld 3.3 (Tentamencijfers studenten, vervolg)

Als we de kans willen uitrekenen dat een aselect gekozen student in Voorbeeld 3.2 is gezakt, dan doen we dat als volgt,

$$P(X \leq 5) = P(\{\omega_1, \omega_3, \omega_{12}, \omega_{13}, \omega_{16}, \omega_{17}, \omega_{22}\}) = \tfrac{7}{24}.$$

De kansverdeling van het tentamencijfer ziet er als volgt uit,

$$X : \begin{pmatrix} 3 & 3.5 & 4.5 & 5 & 6.5 & 7 & 7.5 & 8 & 8.5 & 9 & 9.5 \\ \frac{1}{24} & \frac{2}{24} & \frac{2}{24} & \frac{2}{24} & \frac{4}{24} & \frac{1}{24} & \frac{2}{24} & \frac{3}{24} & \frac{2}{24} & \frac{1}{24} & \frac{4}{24} \end{pmatrix}. \tag{3.2}$$

Berekenen we nu dezelfde kans als boven, maar dan vanuit de kansverdeling van X, dan vinden we

$$P(X \leq 5) = P(X = 3) + P(X = 3.5) + P(X = 4.5) + P(X = 5) = \tfrac{1}{24} + \tfrac{2}{24} + \tfrac{2}{24} + \tfrac{2}{24} = \tfrac{7}{24}.$$

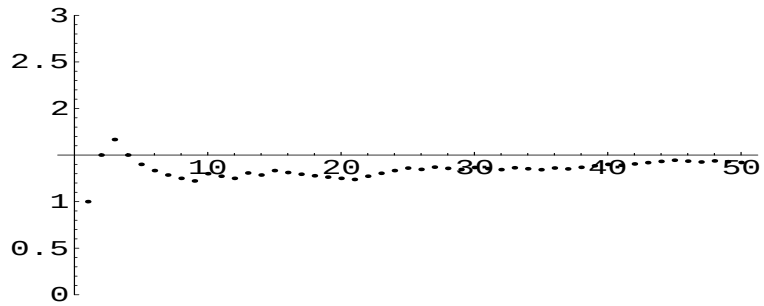
3.2.2 Verwachting en variantie

Als we nu een 'gemiddelde waarde van X' willen invoeren dan kunnen we net zo redeneren als bij het invoeren van de kans. Stel dat we een eindige uitkomstenruimte Ω hebben. Dus $\Omega = \{\omega_1, \ldots, \omega_l\}$. Herhaal het experiment n keer. Dit levert n uitkomsten ω op. Als we de relatieve frequentie gebruiken die ingevoerd is in (2.1), dan kunnen we per uitkomst ω het aantal keren tellen dat die uitkomst voorkomt in de rij. Uitkomst ω komt dan $nf_n(\{\omega\})$ keer voor in de rij. Berekenen we nu het gemiddelde van de waarden van $X(\omega)$ in de rij dan krijgen we

$$\frac{nf_n(\{\omega_1\})X(\omega_1) + \ldots + nf_n(\{\omega_l\})X(\omega_l)}{n} = f_n(\{\omega_1\})X(\omega_1) + \ldots + f_n(\{\omega_l\})X(\omega_l).$$

Deze relatieve frequenties convergeren volgens ons gevoel naar kansen als we het aantal herhalingen groot kiezen. De gemiddelde waarde convergeert dan naar

$$P(\{\omega_1\})X(\omega_1) + \ldots + P(\{\omega_l\})X(\omega_l). \tag{3.3}$$



Figuur 3.4: De gemiddelden $\frac{1}{n}(x_1 + \ldots + x_n), n = 1, \ldots, 50$, bij een rij van 50 gesimuleerde uitkomsten $x_1, \ldots, x_n$ van een stochastische grootheden X met $P(X = 1) = P(X = 2) = 1/2$, voor $i = 1, \ldots, n$.

Dit zullen we de **verwachting van** X noemen. De verwachting van een stochastische variabele is dus in feite zijn gemiddelde over een oneindige rij herhalingen van het experiment. In Figuur 3.5 geven we een gesimuleerd voorbeeld van zo'n rij gemiddelden.

We kunnen echter ook op een andere manier tellen in de rij herhalingen van het experiment. Tel nu per uitkomst r_i het aantal keren dat die uitkomst voorkomt in de rij. Uitkomst r_1 komt dan $nf_n(\{X = r_1\})$ keer voor in de rij, etc. Berekenen we nu het gemiddelde van die waarden dan krijgen we

$$\frac{nf_n(\{X = r_1\})r_1 + \ldots + nf_n(\{X = r_k\})r_k}{n} = f_n(\{X = r_1\})r_1 + \ldots + f_n(\{X = r_k\})r_k. \quad (3.4)$$

Met dezelfde redenering als boven zal dit volgens ons gevoel convergeren naar

$$P(\{X = r_1\})r_1 + \ldots + P(\{X = r_n\})r_k = p_1r_1 + \ldots + p_kr_k.$$

Deze waarde is dan gelijk aan de verwachtingswaarde (3.3). De verwachtingswaarde van een stochastische variabele X geeft de locatie, de plaats, aan van de uitkomsten.

Definitie 3.4 *De verwachting van een stochastische variabele X is gelijk aan*

$$\mathrm{E}\,(X) = \sum_{\omega \in \Omega} X(\omega)P(\{\omega\}). \quad (3.5)$$

Voor een stochastische variabele met kansverdeling (3.1) is deze waarde is gelijk aan $p_1r_1 + \ldots + p_kr_k$.

Als we te maken hebben met een oneindige uitkomstenruimte dan geldt dezelfde definitie. Hierbij zijn de sommen dan oneindige sommen, reeksen. We moeten dan extra veronderstellen dat de som van de absolute waarde van de termen eindig is om problemen met somvolgorden te voorkomen.

Voorbeeld 3.5 (Tentamencijfers studenten, vervolg)

De verwachting van het tentamencijfer X in Voorbeeld 3.2 is gelijk aan

$$\begin{aligned}\mathrm{E}(X) &= 3\tfrac{1}{24} + 3.5\tfrac{2}{24} + 4.5\tfrac{2}{24} + 5\tfrac{2}{24} + 6.5\tfrac{4}{24} + 7\tfrac{1}{24} + 7.5\tfrac{2}{24} + 8\tfrac{3}{24} + 8.5\tfrac{2}{24} \\ &\quad +9\tfrac{1}{24} + 9.5\tfrac{4}{24} = \tfrac{165}{24} = 6.875.\end{aligned}$$

De gemiddelde student heeft dus een voldoende.

Stel nu eens dat we te maken hebben met een functie $Y = g(X)$ van de stochastische variabele X. Van de stochastische variable Y willen we ook de verwachting kunnen uitrekenen. Ook hier is die verwachting gelijk aan de som van de kansen maal de uitkomsten. Voor een stochast X met kansverdeling (3.1) geldt er namelijk.

$$\boxed{\mathrm{E}(Y) = p_1 g(r_1) + \ldots + p_k g(r_k).} \tag{3.6}$$

Hier gebruiken we dus de kansverdeling van X om de verwachting van Y uit te rekenen. Kennelijk is het dus niet nodig eerst de kansverdeling van Y uit te rekenen om de verwachting van Y te bepalen.

Neem nu $g(x) = ax + b$ voor constanten a en b. We bekijken dan in feite een herschaalde en daarna verschoven versie van X. Als we de verwachting van $Y = aX + b$ uitrekenen vinden we

$$\begin{aligned}\mathrm{E}(aX+b) &= \mathrm{E}(Y) = \sum_{i=1}^{k} p_i g(r_i) = \sum_{i=1}^{k} (ar_i + b)p_i = \sum_{i=1}^{k} (ar_i p_i + bp_i) = \sum_{i=1}^{k} ar_i p_i + \sum_{i=1}^{k} bp_i \\ &= a\sum_{i=1}^{k} r_i p_i + b\sum_{i=1}^{k} p_i = a\mathrm{E}(X) + b,\end{aligned}$$

omdat $\sum_{i=1}^{k} p_i$ gelijk is aan een. We hebben nu dus de volgende rekenregel afgeleid,

$$\boxed{\mathrm{E}(aX+b) = a\mathrm{E}(X) + b.} \tag{3.7}$$

Als we vervolgens het kwadraat van X bekijken dan vinden we

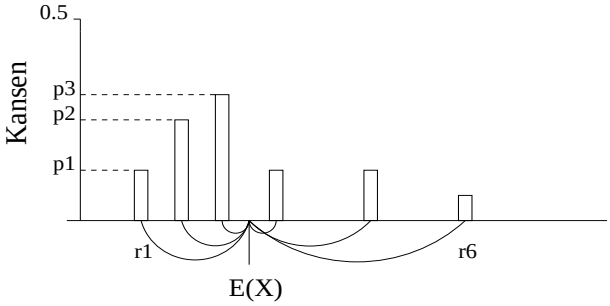
$$\mathrm{E}(X^2) = \mathrm{E}(Y) = p_1 r_1^2 + \ldots + p_k r_k^2.$$

Voor het gemak schrijven we even m voor $\mathrm{E}(X)$. Bekijk nu de stochastische variabele $Y = X - m$. Hierbij is m een constante, een getal. Voor deze Y geldt volgens (3.7), met $a = 1$ en $b = -m$,

$$\mathrm{E}(Y) = \mathrm{E}(X - m) = \mathrm{E}(X) + (-m) = \mathrm{E}(X) - \mathrm{E}(X) = 0.$$

Door van X zijn verwachting af te trekken hebben we de verwachting dus nul gemaakt. We zeggen ook wel dat we X **gecentreerd** hebben. De kansverdeling van de stochast Y is gelijk aan

$$Y : \begin{pmatrix} r_1 - m & r_2 - m & \ldots & r_k - m \\ p_1 & p_2 & \ldots & p_k \end{pmatrix}. \tag{3.8}$$



Figuur 3.5: De kansverdeling van de afstanden van de uitkomsten tot het gemiddelde.

Om een idee te krijgen van de spreiding van de kansverdeling van X berekenen we de verwachting van Y^2. Dat is dan de gewogen som van de gekwadrateerde afstanden van uitkomsten van X tot de verwachting van X. Dus

$$\mathrm{E}\,(Y^2) = p_1(r_1 - m)^2 + \ldots + p_k(r_k - m)^2$$

Dit getal noemen we de **variantie van** X.

Definitie 3.6 *De variantie van X is gelijk aan*

$$\mathrm{Var}\,(X) = \mathrm{E}\,((X - m)^2), \tag{3.9}$$

waarbij $m = \mathrm{E}\,(X)$.

In sommige gevallen is er een handigere manier om de verwachting van een stochastische variabele uit te rekenen. Met wat herschrijven vinden we

$$\begin{aligned}\mathrm{Var}\,(X) &= \mathrm{E}\,((X-m)^2) = \mathrm{E}\,(X^2 - 2Xm + m^2) = \mathrm{E}\,(X^2) + \mathrm{E}\,(-2mX) + \mathrm{E}\,(m^2)\\ &= \mathrm{E}\,(X^2) - 2m\mathrm{E}\,(X) + m^2 = \mathrm{E}\,(X^2) - 2(\mathrm{E}\,(X))^2 + (\mathrm{E}\,(X))^2\\ &= \mathrm{E}\,(X^2) - (\mathrm{E}\,(X))^2.\end{aligned}$$

Dit levert de volgende rekenregel.

$$\boxed{\mathrm{Var}\,(X) = \mathrm{E}\,(X^2) - (\mathrm{E}\,(X))^2.} \tag{3.10}$$

Voor de variantie is het ook interessant om de variantie van een herschaalde en daarna verschoven versie van X te weten. We vinden dan

$$\begin{aligned}\mathrm{Var}\,(aX+b) &= \mathrm{E}\,([aX + b - \mathrm{E}\,(aX+b)]^2) = \mathrm{E}\,([aX + b - (a\mathrm{E}\,(X) + b)]^2)\\ &= \mathrm{E}\,([aX - a\mathrm{E}\,(X)]^2) = \mathrm{E}\,([a(X - \mathrm{E}\,(X))]^2) = \mathrm{E}\,(a^2[X - \mathrm{E}\,(X)]^2)\\ &= a^2\mathrm{E}\,([X - \mathrm{E}\,(X)]^2) = a^2\mathrm{Var}\,(X),\end{aligned}$$

en dus geldt de volgende rekenregel,

$$\boxed{\mathrm{Var}\,(aX+b) = a^2\mathrm{Var}\,(X).} \tag{3.11}$$

Als we nu een maat, een getal, voor de uitgespreidheid van een kansverdeling willen hebben, dan willen we eigenlijk ook dat die maat onder herschalen mee gaat. Dus, als de waarnemingen met een zeker getal vermenigvuldigd worden, dan moet die maat met hetzelfde getal vermenigvuldigd worden. Dat betekent dat we eigenlijk de voorkeur geven aan de wortel van de variantie, de zogenaamde **standaardafwijking van** X.

Definitie 3.7 *De standaardafwijking van X is gelijk aan de wortel van de variantie van X, dus $\sqrt{\mathrm{Var}\,(X)}$.*

Laten we m schrijven voor $\mathrm{E}\,(X)$ en d^2 voor $\mathrm{Var}\,(X)$. De standaardafwijking is dan gelijk aan d. We hebben al gezien dat $Y = X - m$ verwachting nul heeft. De stochastische variabele X is hiermee dan gecentreerd. Nu delen we ook nog door de standaardafwijking. Dus we bekijken $Y = (X - m)/d$. Het is niet moeilijk na te gaan dat de verwachting van Y nog steeds nul is. De variantie en standaardafwijking van Y zijn nu gelijk aan een. We zeggen dat we X hiermee **gestandaardiseerd** hebben.

Voorbeeld 3.8 (Tentamencijfers studenten, vervolg)

We berekenen de verwachting en standaardafwijking van het tentamencijfer X. De verwachting van het tentamencijfer X in Voorbeeld 3.2 is gelijk aan 6.875. Voor de verwachting van X^2 krijgen we

$$\begin{aligned}\mathrm{E}\,(X^2) &= 3^2\,\tfrac{1}{24} + 3.5^2\,\tfrac{2}{24} + 4.5^2\,\tfrac{2}{24} + 5^2\,\tfrac{2}{24} + 6.5^2\,\tfrac{4}{24} + 7^2\,\tfrac{1}{24} + 7.5^2\,\tfrac{2}{24} + 8^2\,\tfrac{3}{24} + 8.5^2\,\tfrac{2}{24} \\ &\quad + 9^2\,\tfrac{1}{24} + 9.5^2\,\tfrac{4}{24} = 51.375.\end{aligned}$$

gebruiken we (3.10) dan krijgen we

$$\mathrm{Var}\,(X) = \mathrm{E}\,(X^2) - (\mathrm{E}\,(X))^2 = 51.375 - 6.875^2 = 4.109375.$$

De standaardafwijking is gelijk aan de wortel hiervan, 2.027.

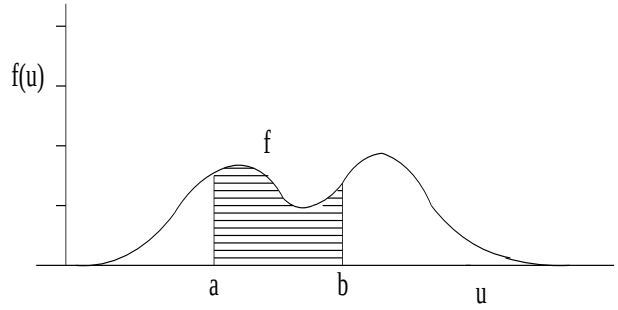
3.3 Continue stochastische variabelen

Discrete stochastische variabele hebben maar eindig veel uitkomsten $r_1, r_2, \ldots, r_k$, of oneindig veel uitkomsten $r_1, r_2, \ldots$ die je op een rij kan leggen. In het laatste geval noem je de verzameling van uitkomsten aftelbaar. De kansverdeling van X kan dan beschreven worden door middel van kansen $p_i = P(X = r_i)$. Zie (3.1). We kunnen kansen op gebeurtenissen die in termen van X beschreven worden dan uitrekenen door de geschikte p_i op te tellen.

Als we nu naar het rad van fortuin, Voorbeeld 2.8, kijken dan hebben we te maken met een uitkomstenruimte Ω die gelijk is aan alle getallen tussen nul en een, het interval $[0, 1)$. Dit zijn er niet eindig veel, en, maar dat is moeilijker te bewijzen, je kan ze ook niet op een rij leggen. Een stochastische variabele bij deze kansruimte zal in het algemeen dan ook niet discreet zijn. Hieronder voeren we zogenaamde continue stochastische variabelen in.

3.3.1 Kansverdeling

Als we een stochastische variabele X hebben waarvoor de uitkomsten alle getallen zijn in een, eventueel oneindig, interval, dan moeten we zijn kansverdeling anders beschrijven. We kunnen immers niet meer dan een rij kansen optellen. In plaats daarvan moeten we dan integreren. De kansverdeling van een **continue stochastische variabele** wordt beschreven door een **kansdichtheidsfunctie** f (probability density function).



Figuur 3.6: Een dichtheidsfunctie f. De oppervlakte van het gearceerde gebied is gelijk aan $P(a \leq X \leq b)$.

Definitie 3.9 *We noemen een functie f een kansdichtheidsfunctie als er aan de volgende twee voorwaarden voldaan is*

1. *$f(u) \geq 0$, voor alle u,*
2. *De oppervlakte onder de functie f is gelijk aan een. Dus*

$$\int_{-\infty}^{\infty} f(u)du = 1. \tag{3.12}$$

We zeggen dat X een continue stochastische variabele is met kansdichtheid f, als

$$P(a \leq X \leq b) = \int_a^b f(u)du, \tag{3.13}$$

voor alle $a < b$.

Als we deze definitie consequent toepassen en de kans op een enkele waarde a uitrekenen dan vinden we

$$P(X = a) = P(a \leq X \leq a) = \int_a f(u)du = 0.$$

De kansen op enkele uitkomsten zijn bij een continue stochastische variabele dus allemaal gelijk aan nul.

Voorbeeld 3.10 (Robot)

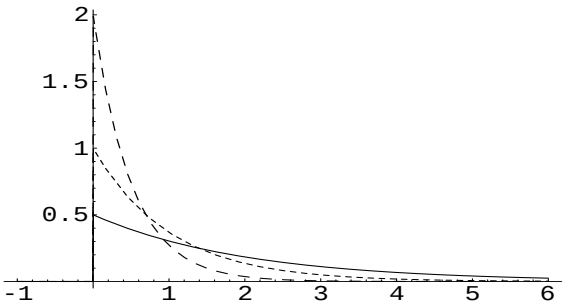
Stel je hebt een robot gebouwd (geprogrammeerd) om een specifieke taak uit te voeren. Je wil dat hij die taak zo snel mogelijk uitvoert. De tijd die hij nodig heeft hangt af beginsituatie. Kies een willekeurige beginsituatie en meet de tijdsduur in minuten. Die tijdsuur is dan een stochastische variabele. De uitkomstenruimte Ω is nu gelijk aan de verzameling van de positieve getallen $[0, \infty)$. Voor de kansdichtheid van X kunnen we een exponentiële functie gebruiken. Voor een zeker getal $\lambda > 0$, dat deze verdeling vastlegt, hebben we dan

$$f(x) = \begin{cases} \frac{1}{\lambda}\, e^{-x/\lambda} & , \text{voor } x \geq 0, \\ 0 & , \text{voor } x < 0. \end{cases}$$

De kans dat de robot minder dan tien minuten nodig heeft is nu gelijk aan

$$\begin{aligned} P(0 \leq X \leq 10) &= \int_0^{10} f(u)du = \int_0^{10} \frac{1}{\lambda}\, e^{-u/\lambda} du = \Big[-e^{-u/\lambda} \Big]_0^{10} \\ &= -e^{-10/\lambda} - (-e^{-0/\lambda}) = 1 - e^{-10/\lambda}. \end{aligned}$$

Dat we deze kans niet kunnen uitrekenen zonder een waarde voor λ in te vullen laat zien dat we een **schatting** van λ, gebaseerd op een aantal gemeten gemeten tijden, nodig hebben. Het getal λ zullen we later een **parameter** noemen. Hiermee zijn we ons eerste statistische probleem tegengekomen.



Figuur 3.7: Een aantal kansdichtheden als in Voorbeeld 3.10.

3.3.2 Verwachting en variantie

De verwachting van een discrete stochastische variabele is gelijk aan de som van zijn mogelijke uitkomsten maal de kansen op die uitkomsten. Het is dus in feite een gewogen gemiddelde van de mogelijke uitkomsten. Bij een continue stochastische variabele wordt de kansverdeling beschreven door zijn kansdichtheidsfunctie f. Hier kunnen we op soortgelijke manier de verwachting invoeren. We doen dit nu als integraal over de mogelijke uitkomsten waarbij we wegen met de kansdichtheidsfunctie. Uitkomsten x met een hogere waarde voor $f(x)$ tellen zwaarder mee dan uitkomsten met een lagere waarde voor $f(x)$. De volgende definities lijken erg op de definities voor discrete stochastische variabelen.

Definitie 3.11 *De verwachting van een continue stochastische variabele X is gelijk aan*

$$\mathrm{E}\,(X) = \int_{-\infty}^{\infty} x f(x) dx. \tag{3.14}$$

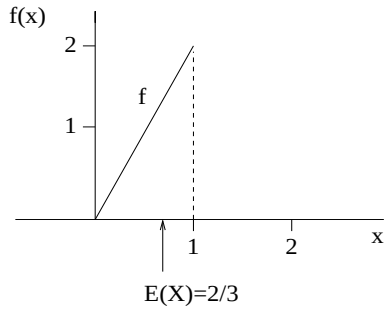
Als $Y = g(X)$, voor een functie g, dan is de verwachting van Y gelijk aan

$$\mathrm{E}\,(Y) = \mathrm{E}\,(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx. \tag{3.15}$$

Schrijven we m voor $\mathrm{E}\,(X)$, dan is de variantie van X gelijk aan

$$\mathrm{Var}\,(X) = \int_{-\infty}^{\infty} (x-m)^2 f(x) dx. \tag{3.16}$$

Voor de verwachtingen en de variantie in deze definitie gelden dezelfde rekenregels als voor discrete stochastische variabelen.



Figuur 3.8: De dichtheidsfunctie f van Voorbeeld 3.12.

Voorbeeld 3.12

Laat X een stochastische variabele zijn met kansdichtheid f gegeven door

$$f(x) = \begin{cases} 2x & \text{, voor } 0 \le x \le 1, \\ 0 & \text{, elders.} \end{cases} \tag{3.17}$$

We berekenen zijn verwachting en variantie. Voor de verwachting van X vinden we

$$\mathrm{E}\,(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x\, 2x dx = 2 \int_0^1 x^2 dx = \tfrac{2}{3}.$$

Voor de variantie berekenen we eerst de verwachting van $Y = X^2$. We vinden dan

$$\mathrm{E}\,(Y) = \mathrm{E}\,(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^1 x^2\, 2x dx = 2 \int_0^1 x^3 dx = \tfrac{1}{2}.$$

De variantie is dan gelijk aan

$$\mathrm{Var}\,(X) = \mathrm{E}\,(X^2) - (\mathrm{E}\,(X))^2 = \tfrac{1}{2} - (\tfrac{2}{3})^2 = \tfrac{1}{18}.$$

Voorbeeld 3.13

Laat X een stochastische variabele zijn met kansdichtheid f gegeven door

$$f(x) = \begin{cases} e^{-x} & \text{, voor } x \geq 0, \\ 0 & \text{, voor } x < 0. \end{cases} \tag{3.18}$$

Dit is een dichtheid als in Voorbeeld 3.10 met de waarde van λ gelijk aan een. We berekenen zijn verwachting en variantie. Voor de berekening van de verwachting passen we de partiële integratie regel (3.19) toe en vinden we

$$\mathrm{E}\,(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_{0}^{\infty} x\, e^{-x} dx = \Big[x(-e^{-x})\Big]_0^{\infty} - \int_{0}^{\infty} -e^{-x} dx = \int_{0}^{\infty} e^{-x} dx = 1,$$

immers de primitieve van e^{-x} is gelijk aan $-e^{-x}$.

Voor de verwachting van $Y = X^2$ vinden we

$$\begin{aligned} \mathrm{E}\,(Y) = \mathrm{E}\,(X^2) &= \int_{-\infty}^{\infty} x^2 f(x) dx = \int_{0}^{\infty} x^2\, e^{-x} dx \\ &= \Big[x^2(-e^{-x})\Big]_0^{\infty} - \int_{0}^{\infty} 2x(-e^{-x}) dx = 2\int_{0}^{\infty} x e^{-x} dx = 2. \end{aligned}$$

De laatste integraal hebben we immers hierboven al uitgerekend. Voor de variantie vinden we dan

$$\mathrm{Var}\,(X) = \mathrm{E}\,(X^2) - (\mathrm{E}\,(X))^2 = 2 - (1)^2 = 1.$$

In de bovenstaande berekeningen van de integralen hebben we twee maal gebruik gemaakt van de partiële integratie regel. Voor de volledigheid noemen we die. Als f en g twee functies zijn met primitieven F en G, dus $f = F'$ en $g = G'$, dan geldt er

$$\boxed{\int_a^b F(x)g(x)dx = \Big[F(x)G(x)\Big]_a^b - \int_a^b f(x)G(x)dx.} \tag{3.19}$$

3.4 Verdelingsfuncties van stochastische variabelen

Als we een algemene stochastische variabele X hebben dan kunnen we zijn kansverdeling ook beschrijven met zijn **verdelingsfunctie** (distribution function) $F(x)$, gedefinieerd door

$$F(x) = P(X \leq x). \tag{3.20}$$

Deze functie geeft de cumulatieve kansen. Voor elke x geeft de verdelingsfunctie de kans dat X ten hoogste de waarde x heeft. Uit de verdelingsfunctie kunnen we alle kansen in termen van X halen. Zo geldt er bijvoorbeeld

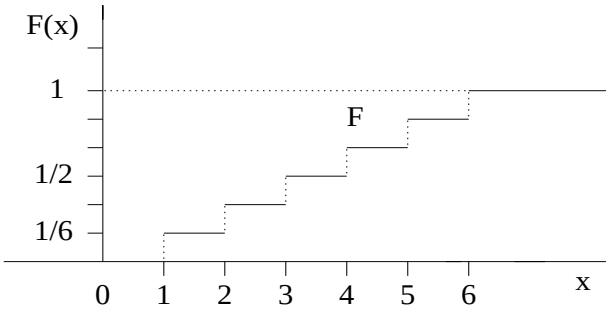
- $P(X \leq a) = F(a)$,

- $P(X > a) = 1 - P(X \leq a) = 1 - F(a)$,
- $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$, voor a en b met $a < b$.

Voor een discrete stochastische variabele X met kansverdeling als (3.1) is F gelijk aan

$$\boxed{F(x) = P(X \leq x) = \sum_{i=1,\ldots,k:r_i \leq x} p_i.} \tag{3.21}$$

In woorden is dat dus : de som van alle kansen p_i die horen bij uitkomsten r_i met $r_i \leq x$. Dit is een stapfunctie die sprongen ter hoogte p_i heeft in de punten $r_1, \ldots, r_k$. Tussen die punten is de functie constant. In Figuur 3.9 wordt de verdelingsfunctie van de uitkomst van een worp met een eerlijke dobbelsteen gegeven.



Figuur 3.9: De verdelingsfunctie van de uitkomst van een worp met een eerlijke dobbelsteen.

Voor een continue stochastische variabele X met kansdichtheid f vinden we

$$\boxed{F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du.} \tag{3.22}$$

In tegenstelling tot de verdelingsfunctie van een discrete stochastische variabele is dit een continue functie. De verdelingsfunctie F is dus een primitieve van f, en omgekeerd is f de afgeleide van F, dus $f(x) = F'(x)$. Figuur 3.10 geeft een illustratie van een verdelingsfunctie van een continue stochastische variabele.

Elke verdelingsfunctie F heeft de volgende eigenschappen:

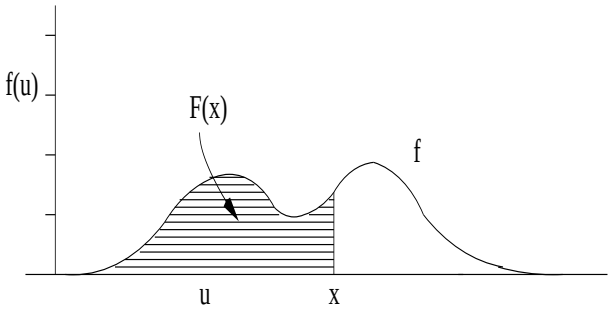
- F is niet dalend. Immers voor twee waarden van x, zeg x_1, x_2, met $x_1 < x_2$ geldt, omdat de gebeurtenis $\{X \leq x_1\}$ een deelverzameling is van de gebeurtenis $\{X \leq x_2\}$,

  $$F(x_1) = P(X \leq x_1) \leq P(X \leq x_2) = F(x_2).$$

- De limiet voor x naar min oneindig van $F(x)$ is gelijk aan nul en de limiet voor x naar oneindig van $F(x)$ is gelijk aan een. Dus

  $$\lim_{x \to -\infty} F(x) = 0 \quad \text{en} \quad \lim_{x \to \infty} F(x) = 1. \tag{3.23}$$

  De eerste bewering volgt uit het feit dat voor een vaste ω, als x steeds kleiner wordt, uiteindelijk geen enkele waarde $X(\omega)$ zal voldoen aan $X(\omega) \leq x$. De kans daarop, $F(x)$ dus, zal dan naar nul gaan. De tweede bewering geldt om een dergelijke reden.



Figuur 3.10: Een dichtheidsfunctie f met de verdelingsfunctie F in het punt x.

Voorbeeld 3.14

In Voorbeeld 3.12 is de verdelingsfunctie van de stochastische variabele X gelijk aan

$$F(x) = \int_{-\infty}^{x} f(u)du = \begin{cases} 0 & \text{, voor } x < 0, \\ \int_0^x 2u du = x^2 & \text{, voor } 0 \leq x \leq 1, \\ 1 & \text{, voor } x \geq 1. \end{cases} \tag{3.24}$$

Voorbeeld 3.15

In Voorbeeld 3.13 is de verdelingsfunctie van de stochastische variabele X gelijk aan

$$F(x) = \int_{-\infty}^{x} f(u)du = \begin{cases} 0 & \text{, voor } x < 0, \\ \int_0^x e^{-u} du = 1 - e^{-x} & \text{, voor } x \geq 0. \end{cases} \tag{3.25}$$

3.5 Functies van stochastische variabelen

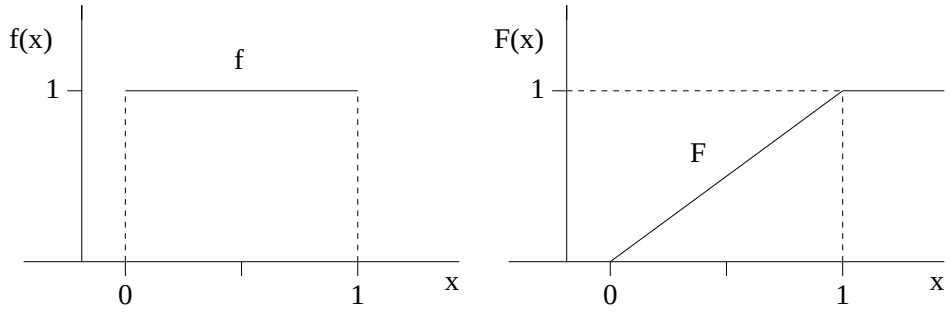
Als we een stochastische variabele X hebben, dan is $Y = g(X)$, voor een zekere functie g, opnieuw een stochastische variabele. We hebben eerder gezien hoe we de verwachting van Y kunnen uitrekenen. We kunnen echter ook de kansverdeling en verdelingsfunctie van Y bepalen. We laten dit zien aan de hand van twee voorbeelden.

Voorbeeld 3.16

Stel X is de stochastische variabele van Voorbeeld 3.12 met kansdichtheid (3.17). We nemen nu $Y = X^2$, dus $g(x) = x^2$. De verdelingsfunctie van Y kunnen we als volgt bepalen. Omdat X alleen waarden tussen nul en een kan hebben, komen negatieve waarden van Y en waarden van Y groter dan een niet voor. Voor $0 \leq y \leq 1$ is de gebeurtenis $\{Y \leq y\} = \{X^2 \leq y\}$ hetzelfde als $\{0 \leq X \leq \sqrt{y}\}$. Dus

$$P(Y \leq y) = P(X^2 \leq y) = P(0 \leq X \leq \sqrt{y}) = F(\sqrt{y}) - F(0) = (\sqrt{y})^2 - 0 = y.$$

Hierbij is F de verdelingsfunctie van X, gegeven door (3.24). De dichtheidsfunctie van Y is gelijk aan de afgeleide van de verdelingsfunctie. Hij is dus constant een tussen nul en een, en elders is hij gelijk aan nul. We zeggen hier dat Y **uniform verdeeld is op het interval** $[0, 1]$.



Figuur 3.11: De dichtheidsfunctie f en de verdelingsfunctie F van een stochastische variabele met een uniforme verdeling.

Voorbeeld 3.17

Laten we nu de stochastische variabele Y uit het vorige voorbeeld verder transformeren. Neem nu $g(x) = -\ln(x)$. We bepalen de kansverdeling van $Z = g(Y) = -\ln(Y)$. Op soortgelijke manier als in het vorige voorbeeld vinden we nu voor $z \geq 0$

$$P(Z \leq z) = P(-\ln(Y) \leq z) = P(Y \geq e^{-z}) = 1 - e^{-z}.$$

Dit is de verdelingsfunctie (3.25) van de stochastische variabele van Voorbeeld 3.13.

Als we de twee transfomaties achter elkaar uitvoeren zien we dat de transformatie $Y = -\ln(X^2) = -2\ln(X)$ de kansverdeling van Voorbeeld 3.12 overvoert in de kansverdeling van Voorbeeld 3.13.

3.6 Computersimulatie van stochastische variabelen

Het kan bij ingewikkelde processen waar meer stochastische variabelen een rol spelen voorkomen dat je de kansverdeling van een stochastiche variabele niet kan uitrekenen. In zo'n geval kan je vaak uitkomsten van die stochastische variabele **simuleren** op de computer. Als je dat maar vaak genoeg doet dan krijg je een goed idee van de kansverdeling van die variabele.

Bij het ontwikkelen van methoden voor het simuleren op een computer onderscheiden we doorgaans twee stappen. De eerste stap is het construeren van een **(pseudo) aselecte getallen generator** ((pseudo) random number generator), die trekkingen uit de uniforme verdeling on $[0, 1]$ simuleert. De tweede stap is dan dat we, gegeven die uniforme trekkingen, een methode construeren om trekkingen uit een bepaalde verdeling te maken. We zullen die twee stappen apart bespreken.

Voor we echter meer in detail treden noemen we een aantal criteria waaraan een pseudo aselecte getallen generator op een computer moet voldoen. We noemen de volgende drie vereisten.

1. De gegenereerde getallen moeten de gewenste *statistische eigenschappen* hebben. Om dit te controleren moeten we op grote aantallen gegenereerde getallen toetsingsgrootheden uitrekenen, die de hypothese dat de onderliggende verdeling van een steekproef gelijk is aan de uniforme verdeling op $[0, 1]$ 'toetsen'.

2. *Snelheid.* Voor simulaties hebben we grote aantallen getallen nodig, dus is het van belang dat deze getallen snel en efficiënt berekend worden, bij voorkeur in machinetaal.

3. Het moet mogelijk zijn de getallen te *reproduceren.* Denk hierbij aan een groot computerprogramma waarin pseudo aselecte getallen worden gegenereerd, die vervolgens worden gebruikt in andere onderdelen van dat programma. Als we echter een fout in het programma gecorrigeerd hebben willen we onderzoeken of dat inderdaad het goede effect op de resultaten heeft gehad. Daarvoor is het nodig dat we exact dezelfde pseudo aselecte getallen kunnen reproduceren.

3.6.1 De lineaire congruentie methode voor het genereren van pseudo aselecte getallen

Van de vele methoden voor het genereren van pseudo aselecte getallen bespreken we er een, de **lineaire congruentie methode**. Bij deze methode maken we eerst getallen volgens het volgende algorithme:

> Kies een natuurlijk getal $x_1 \quad (0 \leq x_1 < m)$
> bepaal $x_2, x_3, ...$ door middel van
>
> $$x_{i+1} = (ax_i + c) \bmod m \quad (i \geq 1),$$
>
> waarbij a, c en m vaste natuurlijke getallen zijn.

Op deze manier verkrijgen we een rij natuurlijke getallen x_i, met $0 \leq x_i < m$. We maken vervolgens een rij getallen u_i, met $0 \leq u_i < 1$, door de x_i te delen door m, dus $u_i = x_i/m$. We kunnen u_{i+1} als volgt uit u_i berekenen

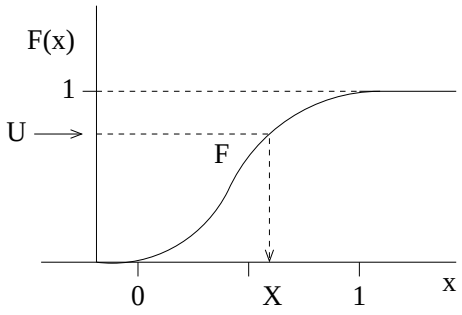
$$u_{i+1} = \frac{1}{m}((a(mu_i) + c) \bmod m), \ i \geq 1.$$

Aangezien het algorithme na hoogstens m stappen weer opnieuw begint, wordt m in de praktijk zo groot mogelijk genomen. Naar de keuze van de constanten a en c is veel onderzoek gedaan. Omdat een aantal theoretische eigenschappen van het algorithme bekend zijn kan men richtlijnen geven voor het kiezen van deze constanten, zie bijvoorbeeld Ripley (1987). De kwaliteit van de generator wordt bepaald door deze keuze en het is ten zeerste af te raden ze zomaar lukraak te kiezen. De meeste computerpakketten hebben tegenwoordig een goede generator. We merken verder nog op dat het belangrijk is dat de berekening van de getallen exact gebeurt in de computer, dus zonder afrondingen, omdat anders de theoretische eigenschappen niet meer hoeven te gelden.

We gaan er nu verder van uit dat we beschikken over een generator van trekkingen uit de uniforme verdeling. In de volgende paragraaf bespreken we methoden voor het genereren van trekkingen uit andere verdelingen.

3.6.2 Het genereren van trekkingen uit algemene verdelingen

We nemen in deze paragraaf aan dat we beschikken over een geschikte generator van trekkingen uit de uniforme verdeling op $[0, 1]$. Stel nu dat we trekkingen willen genereren uit een verdeling met verdelingsfunctie F.



Figuur 3.12: Illustratie van het genereren van een trekking uit een kansverdeling met verdelingsfunctie F.

We geven de trekking uit de uniforme verdeling aan met de stochastische variabele U. We zoeken nu een waarde X zódat $U = F(X)$. Deze waarde X is in feite gelijk aan de inverse van de functie F, genoteerd met F^{-1}, in het punt U. Dus $X = F^{-1}(U)$. De stochastische variabele X heeft dan F als verdelingsfunctie. Dit volgt uit het feit dat de gebeurtenis $\{X \leq x\}$ gelijk is aan de gebeurtenis $\{U \leq F(x)\}$. Dit is geïllustreerd in Figuur 3.12.

Voorbeeld 3.18 (Genereren van een exponentiële trekking)

Een voorbeeld van deze methode vinden we als we een generator van trekkingen met de exponentiële dichtheid (3.18) willen maken. In dat geval is $F(x)$ gelijk aan $(1 - e^{-x})$, voor positieve x, en als we $U = F(X)$ oplossen vinden we

$$U = F(X) \Leftrightarrow U = 1 - e^{-X} \Leftrightarrow e^{-X} = 1 - U \Leftrightarrow -X = \ln(1 - U) \Leftrightarrow X = -\ln(1 - U).$$

Als $u_1, ..., u_n$ een n-tal pseudo aselecte getallen is, dan kunnen we $-\ln(1 - u_1), -\ln(1 - u_2), ..., -\ln(1 - u_n)$ gebruiken als trekkingen uit de standaard exponentiële verdeling. Eigenlijk kunnen we hier ook $-\ln(u_1), -\ln(u_2), ..., -\ln(u_n)$ gebruiken omdat U en $1 - U$ dezelfde kansverdeling hebben (opgave 10).

Het spreekt vanzelf dat deze methode alleen maar bevredigend werkt als de inverse van de verdelingsfunctie F efficiënt berekend kan worden. De hierboven beschreven methode is eigenlijk de meest eenvoudige directe methode. Er bestaan echter nog veel meer methoden. Een aantal daarvan komen later aan de orde.

3.7 Opgaven

1. We werpen tweemaal met een dobbelsteen. Beschouw de twee stochastische variabelen X, het aantal geworpen zessen, en Y, de som van de geworpen ogen.

   (a) Omschrijf de uitkomstenruimte Ω voor dit experiment.

   (b) Bepaal de kansverdelingen van X en Y.

   (c) Zijn de gebeurtenissen $\{X = 1\}$ en $\{Y = 5\}$ onafhankelijk?

2. Laat X een stochastische variabele zijn met kansverdeling

$$X : \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{5} & \frac{2}{5} & \frac{2}{5} \end{pmatrix}. \tag{3.26}$$

   (a) Bereken $\mathrm{E}(X)$.

   (b) Beschouw $Y = X^2$ als nieuwe stochastische variabele. Bepaal de kansverdeling van Y. Bereken daarmee de verwachting van Y.

   (c) Bereken de verwachting van Y met formule (3.6) door gebruik te maken van de kansverdeling van X.

   (d) Bereken de variantie van X.

3. Laten we m schrijven voor $\mathrm{E}(X)$ en d^2 voor $\mathrm{Var}(X)$. De standaardafwijking is dan gelijk aan d. We hebben al gezien dat $Y = X - m$ verwachting nul heeft. De stochastische variabele X is hiermee dan gecentreerd. Nu delen we ook nog door de standaardafwijking. Dus we bekijken $Y = (X - m)/d$. Laat zien dat de verwachting van Y nog steeds gelijk aan nul is. Laat ook zien dat de variantie en standaardafwijking van Y gelijk zijn aan een.

4. Werp één maal met een eerlijke munt. De stochastische variabele X is nul als de uitkomst 'kruis' is, en een als de uitkomst 'munt' is.

   (a) Geef de kansverdeling van X.

   (b) Bereken de verwachting van X en X^2.

   (c) Bereken de variantie van X rechtstreeks met de formule $\mathrm{Var}(X) = \mathrm{E}((X - \mathrm{E}\,X)^2)$ en controleer $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$.

5. Werp één maal met een eerlijke dobbelsteen. Laat X het aantal geworpen ogen zijn.

   (a) Bereken de verwachting en variantie van X.

   (b) Bereken de verwachting van $3X - 1$ en de variantie van $10X + 2$.

6. Laat de kansdichtheidsfunctie f gegeven zijn door

$$f(x) = \begin{cases} 0 & , \text{ als } x < -2, \\ \frac{1}{2} & , \text{ als } -2 \leq x < -1, \\ 0 & , \text{ als } -1 \leq x \leq 1, \\ \frac{1}{2} & , \text{ als } 1 \leq x < 2, \\ 0 & , \text{ als } x \geq 2. \end{cases}$$

Teken de grafiek van deze kansdichtheidsfunctie en bepaal de corresponderende verdelingsfunctie. Teken ook de grafiek van de verdelingsfunctie.

7. Laat X een stochastische variabele zijn met verdelingsfunctie F, gegeven door

$$F(x) = \begin{cases} 0 & , \text{ als } x < 0, \\ x^2 & , \text{ als } 0 \leq x < 1, \\ 1 & , \text{ als } x \geq 1. \end{cases}$$

We gaan $P(\frac{1}{2} < X \leq \frac{3}{4})$ op twee manieren berekenen.

   (a) Gebruik direct de verdelingsfunctie F om de kans uit te rekenen.

   (b) Bepaal de dichtheidsfunctie f van X en bereken de kans door middel van integreren.

8. Stel X is een stochastische variabele met kansdichtheid

$$f(x) = \begin{cases} c(1-x^2) & , \text{ als } -1 \leq x \leq 1, \\ 0 & , \text{ elders.} \end{cases}$$

voor een zekere constante c.

   (a) Bepaal de waarde van c.

   (b) Bereken de verwachting en variantie van X.

   (c) Bereken de verwachting van $|X|$, de absolute waarde van X.

9. Laat X een stochastische variabele zijn met verdelingsfunctie F, gegeven door

$$F(x) = \begin{cases} 0 & , \text{ als } x < 0, \\ \frac{1}{4}x^2 & , \text{ als } 0 \leq x < 2, \\ 1 & , \text{ als } x \geq 2. \end{cases}$$

Neem Y gelijk aan X^2. We gaan de verwachting van Y op twee manieren berekenen.

   (a) Bepaal de kansdichtheidsfunctie van Y en bereken daarmee de verwachting van Y.

   (b) Bereken de verwachting van Y door middel van de regel

$$\mathrm{E}\,(Y) = \int_{-\infty}^{\infty} g(x) f(x) dx,$$

waarbij f de kansdichtheid van X is.

10. Laat de stochastische variabele U als kansdichtheid de uniforme dichtheid op [0.1) hebben. Dus $f(x) = 1$ als $0 \leq x \leq 1$, en $f(x) = 0$, elders.

    (a) Laat zien dat U en $1 - U$ dezelfde kansverdeling hebben.

    (b) Bereken de verwachting en variantie van U.

    (c) Bepaal de kansverdeling van U^2.

    (d) Bereken de verwachting en variantie van U^2.

11. Stel hebt een random getallen generator op je computer die trekkingen uit de uniforme verdeling op $[0, 1)$ levert. Noem een zo'n trekking U. Hoe maak je hiermee een generator die trekkingen genereert met de kansdichtheid (3.17)? Deze dichtheid wordt gegeven door

$$f(x) = \begin{cases} 2x & \text{, voor } 0 \leq x \leq 1, \\ 0 & \text{, elders.} \end{cases}$$

12. Stel je beschikt over een generator voor uniforme getallen uit $[0, 1]$. Bedenk een methode om trekkingen van een discrete stochastische variabele X met kansverdeling

$$X : \begin{pmatrix} r_1 & r_2 & \dots & r_k \\ p_1 & p_2 & \dots & p_k \end{pmatrix},$$

    te genereren.

Hoofdstuk 4

Parametrische families van kansverdelingen

In Voorbeeld 3.10 hebben we de kansverdeling van de tijd die een robot nodig heeft om een taak uit te voeren beschreven met een verdeling die bepaald werd door een getal λ. Dat getal is in het algemeen onbekend. Zo ontstaat er een statistisch probleem om de waarde van λ te benaderen, schatten, als we gegevens hebben, dus als we een aantal tijdsduren hebben gemeten. Zo'n getal λ noemen we een **parameter**. De parameterwaarde wijst in feite één verdeling van een hele familie van verdelingen aan. Een dergelijke familie van kansverdelingen wordt een **parametrische familie** genoemd. De vorm van de kansverdelingen is dan voorgeschreven, maar de parameterwaarde geeft aan welke kansverdeling het precies is. We zullen een aantal veel voorkomende families van kansverdelingen hieronder beschrijven. We geven daarbij ook de verwachtingen en varianties.

4.1 Discrete verdelingen

4.1.1 Bernoulli en Binomiale verdeling (**Bern**(p) **en Bin**(n,p), $0 \leq p \leq 1$)

Stel je doet een alternatief experiment met uitkomst 'succes' of 'mislukking'. De stochastische variabele X heeft waarde een in het geval van 'succes' en waarde nul bij 'mislukking'. Stel nu $p = P(X = 1)$. De kansverdeling van X, gegeven door

$$X : \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}, \tag{4.1}$$

noemen we een **Bernoulli verdeling**. De succeskans p is hier dus de parameter.
De verwachting en variantie worden gegeven door

$$\mathrm{E}\,(X) = p, \tag{4.2}$$
$$\mathrm{Var}\,(X) = p(1-p). \tag{4.3}$$

Als we dit experiment n keer herhalen krijgen we een rij stochastische variabelen $X_1, \ldots, X_n$. Zij $X = X_1 + \ldots + X_n$, het aantal successen in deze n experimenten. We veronderstellen dat de experimenten onafhankelijk zijn. Dat wil zeggen dat de gebeurtenissen in de afzonderlijke experimenten onafhankelijk zijn. We bepalen de kansverdeling van X.

$$\begin{aligned}
P(X=0) &= P(X_1=0,\ldots,X_n=0)\\
&= P(X_1=0)\ldots P(X_n=0)\\
&= (1-p)\ldots(1-p)\\
&= (1-p)^n.\\
P(X=1) &= P(\text{precies een } X_i \text{ is gelijk aan een})\\
&= P(X_1=1,X_2=0,\ldots,X_n=0)+P(X_1=0,X_2=1,\ldots,X_n=0)+\ldots\\
&\quad +P(X_1=0,\ldots,X_{n-1}=0,X_n=1)\\
&= p(1-p)\ldots(1-p)+(1-p)p\ldots(1-p)+\ldots+(1-p)\ldots(1-p)p\\
&= np(1-p)^{n-1}.
\end{aligned}$$

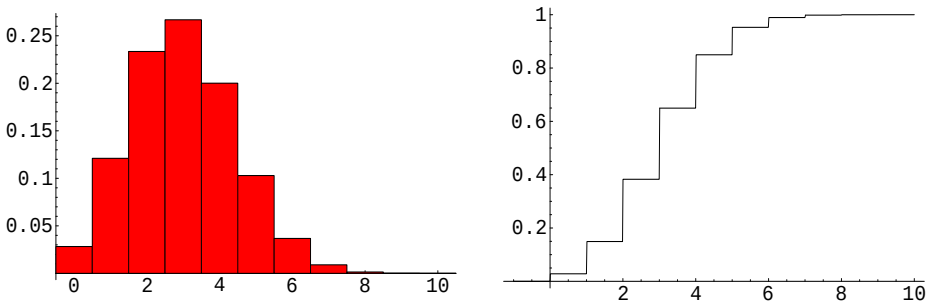
Algemener vinden we, voor $k = 1, \ldots, n$,

$$\begin{aligned}
P(X=k) &= P(\text{precies } k \text{ van de } X_i \text{ zijn gelijk aan een, en } n-k \text{ zijn gelijk aan nul})\\
&= \binom{n}{k} p^k (1-p)^{n-k}.
\end{aligned}$$

Hierbij is de **Binomiaalcoëfficient**,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\ldots(n-k+1)}{k!}, k = 1, 2, \ldots, n, \tag{4.4}$$

gelijk aan het aantal manieren waarop we k eenen kunnen aanwijzen in een rij van n nullen en eenen.



Figuur 4.1: Kansen en verdelingsfunctie van de Bin(10, 0.3) verdeling.

We zeggen nu dat X een **Binomiale verdeling** heeft met parameter p als

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n. \tag{4.5}$$

De verwachting en variantie worden gegeven door

$$\mathrm{E}\,(X) = np, \tag{4.6}$$
$$\mathrm{Var}\,(X) = np(1-p). \tag{4.7}$$

De formule voor de verwachting volgt uit

$$\mathrm{E}\,(X) = \mathrm{E}\,(X_1) + \ldots + \mathrm{E}\,(X_n) = p + \ldots + p = np$$

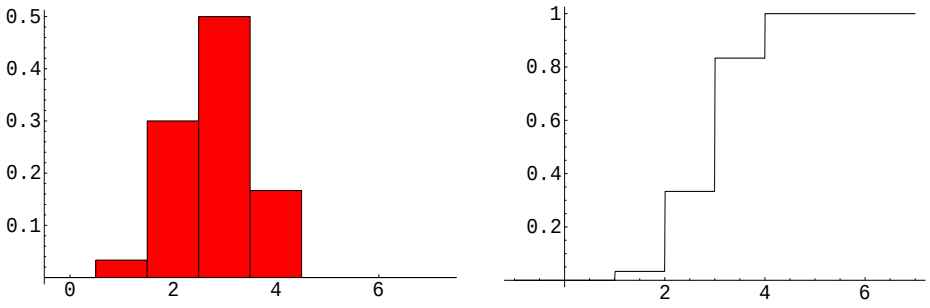
Een bewijs voor de formule van de variantie volgt later.

4.1.2 Hypergeometrische verdeling (Hyp(N, R, n))

Uit een vaas met N knikkers, waarvan er R rood en $N-R$ wit zijn, worden er aselect zonder teruglegging n gekozen. Als X het aantal rode knikkers in de steekproef is, dan geldt voor mogelijke uitkomsten k,

$$P(X=k) = \frac{\binom{R}{k}\binom{N-R}{n-k}}{\binom{N}{n}}. \tag{4.8}$$

Deze kansverdeling noemen we een **Hypergeometrische verdeling**.



Figuur 4.2: Kansen en verdelingsfunctie van de Hyp(10,7,4) verdeling.

Deze kans kan je als volgt afleiden. Je kan op $\binom{R}{k}$ manieren k rode knikkers uit de R rode knikkers in de vaas trekken. Net zo kan je op $\binom{N-R}{n-k}$ manieren $n-k$ witte knikkers uit de $N-R$ witte knikkers in de vaas trekken. In totaal kan je dus op $\binom{R}{k}\binom{N-R}{n-k}$ manieren n knikkers trekken zódat voldaan is aan $\{X=k\}$. Het totaal aantal manieren om n knikkers uit de N knikkers in de vaas te trekken is gelijk aan $\binom{N}{n}$. Als we deze aantallen op elkaar delen vinden we de kans (4.8).

De verwachting en variantie worden gegeven door

$$\mathrm{E}\,(X) = \frac{nR}{N}, \tag{4.9}$$
$$\mathrm{Var}\,(X) = n\frac{R}{N}\frac{N-R}{N}\frac{N-n}{N-1}. \tag{4.10}$$

4.1.3 Poissonverdeling (Poisson(λ), $\lambda > 0$)

We zeggen dat een stochastische variabele X een **Poisson verdeling** heeft met parameter $\lambda > 0$ als

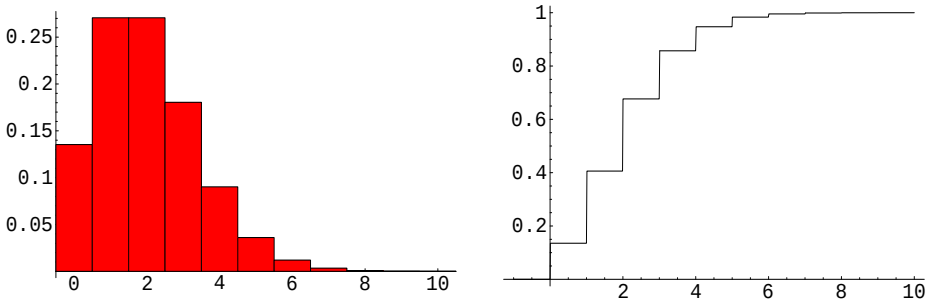
$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots. \tag{4.11}$$

De verwachting en variantie worden gegeven door

$$\mathrm{E}\,(X) = \lambda, \tag{4.12}$$

$$\mathrm{Var}\,(X) = \lambda. \tag{4.13}$$

Deze verdeling kan gebruikt worden als benadering van de Binomiale(n, p) verdeling als n groot is, p klein is, en $np \approx \lambda$.



Figuur 4.3: Kansen en verdelingsfunctie van de Poisson(2) verdeling.

4.2 Continue verdelingen

4.2.1 Uniforme of homogene verdeling (Un$(a, b), a < b$)

Een stochastische variabele X heeft een **uniforme verdeling** op het interval (a, b) als X kansdichtheid

$$f(x) = \begin{cases} \frac{1}{b-a} & , \text{ als } a \leq x \leq b, \\ 0 & , \text{ elders.} \end{cases} \tag{4.14}$$

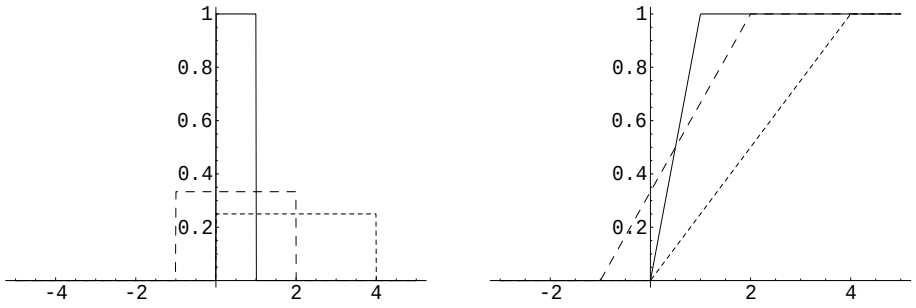
en dus verdelingsfunctie

$$F(x) = \begin{cases} 0 & , \text{ als } x < a, \\ \frac{x-a}{b-a} & , \text{ als } a \leq x < b, \\ 1 & , \text{ als } x \geq b. \end{cases} \tag{4.15}$$

heeft.
De verwachting en variantie worden gegeven door

$$\mathrm{E}\,(X) = \frac{a+b}{2}, \tag{4.16}$$

$$\mathrm{Var}\,(X) = \frac{(b-a)^2}{12}. \tag{4.17}$$



Figuur 4.4: Dichtheden en verdelingsfuncties van homogene verdelingen.

4.2.2 Beta verdeling (**Beta**$(\alpha, \beta),\ \alpha > 0, \beta > 0.$)

Een stochastische variabele X heeft een **beta verdeling** met parameters α en β als hij de volgende kansdichtheid heeft,

$$f(x) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} & , \text{ als } 0 \leq x \leq 1, \\ 0 & , \text{ elders.} \end{cases} , \tag{4.18}$$

waarbij de normaliseringsconstante $B(\alpha, \beta)$ gegeven wordt door

$$B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du. \tag{4.19}$$

De verwachting en variantie worden gegeven door

$$\mathrm{E}\,(X) = \frac{\alpha}{\alpha+\beta}, \tag{4.20}$$

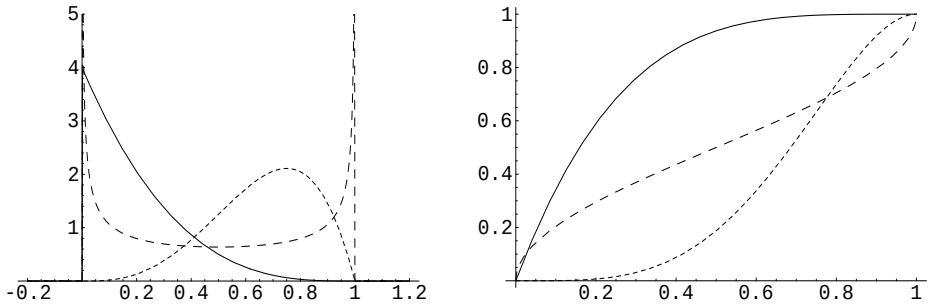
$$\mathrm{Var}\,(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{4.21}$$

Deze verdeling voor een stochastische variabele met waarden tussen nul en een wordt bij de Bayesiaanse aanpak van de statistiek gebruikt om de zogenaamde apriori kennis over een kans te modelleren. Zie hiervoor Sectie 6.5.

4.2.3 Exponentiële verdeling (**Exp**$(\lambda),\ \lambda > 0.$)

Een stochastische variabele X heeft een **exponentiële verdeling** met parameter λ als hij kansdichtheid

$$f(x) = \begin{cases} \frac{1}{\lambda}\, e^{-x/\lambda} & , \text{als } x \geq 0, \\ 0 & , \text{elders.} \end{cases} \tag{4.22}$$



Figuur 4.5: Dichtheden en verdelingsfuncties van Beta verdelingen met parameters (α, β) gelijk aan (1,4), (4,2) en (0.5,0.5).

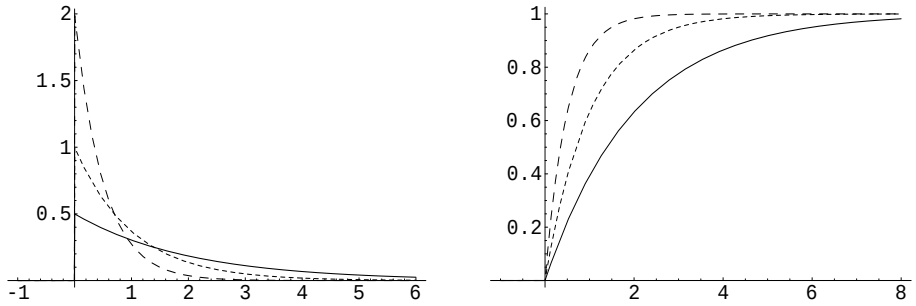
heeft, en verdelingsfunctie

$$F(x) = \begin{cases} 1 - e^{-x/\lambda} & \text{, als } \ x \geq 0, \\ 0 & \text{, elders.} \end{cases} \tag{4.23}$$

De verwachting en variantie worden gegeven door

$$\mathrm{E}\,(X) = \lambda, \tag{4.24}$$
$$\mathrm{Var}\,(X) = \lambda^2. \tag{4.25}$$

Deze verdelingen worden in het algemeen gebruikt om tijdsduren te modelleren. Deze verdelingen hebben de eigenschap dat ze een levensduurverdeling modelleren waarbij geen slijtage optreedt. Zie opgave 4.



Figuur 4.6: Dichtheden en verdelingsfuncties van exponentiële verdelingen.

4.2.4 Gammaverdeling (Gamma(n, λ), $\lambda > 0$.)

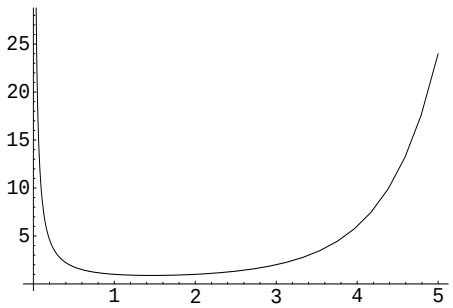
Een stochastische variabele X heeft een Gamma verdeling met parameter (n en) λ als hij kansdichtheid

$$f(x) = \begin{cases} \frac{1}{\lambda^n} \frac{1}{\Gamma(n)} x^{n-1} e^{-x/\lambda} & \text{, als } x \geq 0, \\ 0 & \text{, elders.} \end{cases} \tag{4.26}$$

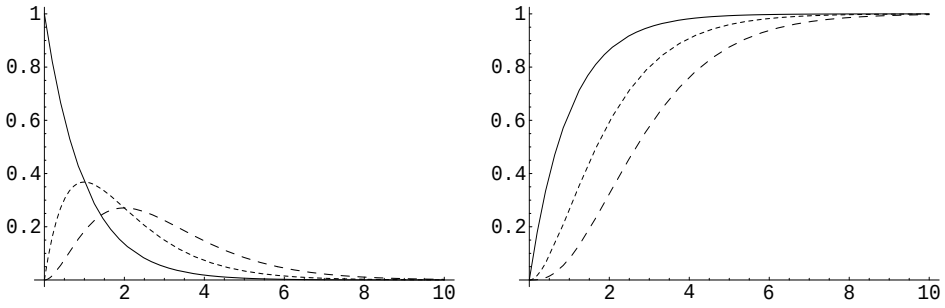
heeft. Hier is $\Gamma(z)$ de zogenaamde Gamma-functie, gedefinieerd door

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt, \qquad z > 0. \tag{4.27}$$

Merk op dat $\Gamma(1) = \int_0^\infty e^{-t} dt = 1$, en dat door middel van partiële integratie voor $z > 1$ geldt



Figuur 4.7: Gamma-functie.



Figuur 4.8: Dichtheden en verdelingsfuncties van gamma verdelingen.

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt = \Big[-t^{z-1} e^{-t} \Big]_0^\infty + (z-1) \int_0^\infty t^{z-2} e^{-t} dt = (z-1)\Gamma(z-1),$$

met als gevolg dat $\Gamma(n) = (n-1)!$ voor $n \in \mathbb{N}$. Merk ook op dat Gamma$(1, \lambda)$ en Exp(λ) gelijk zijn.
De verwachting en variantie worden gegeven door

$$\mathrm{E}\,(X) = n\lambda, \tag{4.28}$$
$$\mathrm{Var}\,(X) = n\lambda^2. \tag{4.29}$$

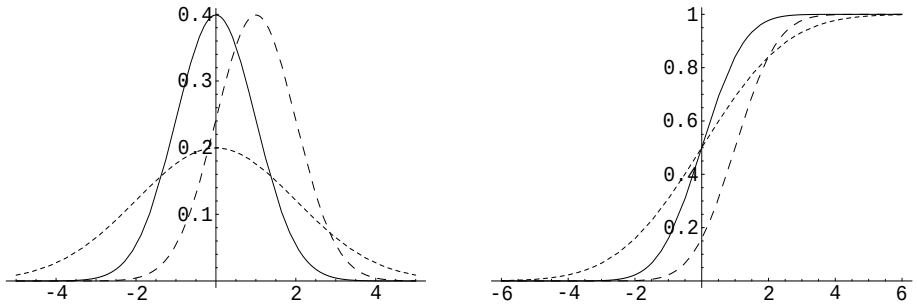
Deze verdeling is gelijk aan de verdeling van de som van n onafhankelijke Exp(λ) stochastische variabelen.

4.2.5 Normale of Gaussische verdeling ($\mathcal{N}(\mu, \sigma^2), -\infty < \mu < \infty, \sigma > 0.$)

Een stochastische variabele X heeft een **normale verdeling** als hij kansdichtheid

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{4.30}$$

heeft. Voor de zogenaamde **standaardnormale verdeling**, $\mathcal{N}(0,1)$, noteren we de dichtheid



Figuur 4.9: Dichtheden en verdelingsfuncties van normale verdelingen.

met

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}. \tag{4.31}$$

en de verdelingsfunctie met

$$\Phi(x) = \int_{-\infty}^{x} \phi(y)dy. \tag{4.32}$$

Als X een $\mathcal{N}(\mu, \sigma^2)$ verdeling heeft, dan geldt

$$f(x) = \frac{1}{\sigma}\, \phi\left(\frac{x-\mu}{\sigma}\right) \tag{4.33}$$

en

$$F(x) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \phi(y)dy = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{4.34}$$

Dit volgt uit de volgende berekening, waarbij we $(u-\mu)/\sigma$ vervangen door v,

$$\begin{aligned} F(x) &= P(X \le x) = \int_{-\infty}^{x} f(u)du = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2} du \\ &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}v^2} \sigma dv = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}v^2} dv \\ &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \phi(v)dv = \Big[\Phi(v)\Big]_{-\infty}^{\frac{x-\mu}{\sigma}} = \Phi\Big(\frac{x-\mu}{\sigma}\Big). \end{aligned}$$

nemen we de afgeleiden naar x dan vinden we (4.33).

De verwachting en variantie van de $\mathcal{N}(\mu, \sigma^2)$ verdeling worden gegeven door

$$\mathrm{E}(X) = \mu, \tag{4.35}$$
$$\mathrm{Var}(X) = \sigma^2. \tag{4.36}$$

De normale verdeling wordt vaak gebruikt om meetfouten te modelleren. Ook bij andere experimenten kan hij vaak gebruikt worden. Dit komt omdat de som van een groot aantal individuele bijdragen, die ongeveer dezelfde kansverdeling hebben, en die elkaar niet beïnvloeden, ongeveer normaal verdeeld is. Dit volgt uit de zogenaamde Centrale Limietstelling van de kansrekening.

4.3 Opgaven

1. Bij de beschrijving van de Bernoulli verdeling worden de verwachting en variantie gegeven in formules (4.2) en (4.3). Bewijs de formules.

2. Bij de beschrijving van de uniforme verdeling worden de verwachting en variantie gegeven in formules (4.16) en (4.17). Bewijs de formules.

3. Bij de beschrijving van de exponentiële verdeling worden de verwachting en variantie gegeven in formules (4.24) en (4.25). Bewijs de formules.
   Hint: Stel $Y = \lambda X$, waarbij X een exponentiële verdeling heeft met parameter gelijk aan een. Dan heeft Y de kansverdeling (4.22).

4. Stel dat X een levensduur is met verdelingsfunctie

   $$F(x) = P(X \leq x) = 1 - e^{-x/\lambda}, x \geq 0,$$

   voor een gegeven getal λ. Laten s en $s + t$ twee tijdstippen zijn met $s > 0$ en $t > 0$. Laat zien dat

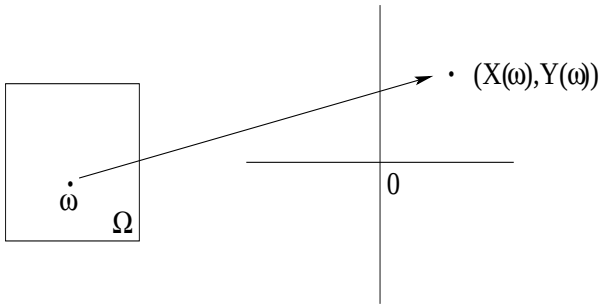
   $$P(X > s + t | X > s) = P(X > t).$$

   Met andere woorden, er treedt geen slijtage op.

Hoofdstuk 5

Meerdere stochastische variabelen tegelijk

In het voorafgaande hebben we, uitgaande van een achterliggende kansruimte $(\Omega, \mathcal{A}, P)$, een stochastische variabele X ingevoerd als een afbeelding van Ω naar de reële getallen. We kennen dus aan elke uitkomst ω in Ω een getal toe. Denk bijvoorbeeld aan een meting aan de uitkomst. Vaak zullen we meerdere metingen aan de uitkomst doen. We krijgen dan bijvoorbeeld in plaats van een getal $X(\omega)$ twee getallen $(X(\omega), Y(\omega))$. We noemen dit een **stochastische vector** (random vector).



Figuur 5.1: Schematische voorstelling van een stochastische vector (X, Y) .

5.1 Discreet verdeelde stochastische vectoren

We voeren begrippen als kansverdeling, verwachting etc., van een discrete stochastische vector in aan de hand van een voorbeeld.

Voorbeeld 5.1 (Tentamencijfers studenten en aantal jaren studie)

Als voorbeeld van een kansruimte waarbij we twee stochastische variabelen tegelijk bekijken gaan we terug naar een voorbeeld in Hoofdstuk 3. In Voorbeeld 3.2 trokken we aselect een student uit 24 studenten die een tentamen hebben gedaan. De stochastische variabele X was

de uitslag van het tentamen. We kunnen echter ook kijken naar het aantal jaren, zeg Y, dat de student studeert. Het gaat hierbij om een tweedejaars vak, dus Y is minimaal gelijk aan twee. We hebben nu dus twee stochastische variablen, gegeven door

$X(\omega) =$ uitslag van het tentamen,
$Y(\omega) =$ aantal jaren studie.

In Tabel 5.1 wordt voor elke student ω het cijfer en het aantal jaren studie gegeven.

Student	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	ω_{11}	ω_{12}
Uitslag	5	6.5	5	9.5	8	7	8	6.5	9.5	6.5	8.5	3.5
Aantal jaren	2	3	2	2	2	3	3	4	2	3	2	2
Student	ω_{13}	ω_{14}	ω_{15}	ω_{16}	ω_{17}	ω_{18}	ω_{19}	ω_{20}	ω_{21}	ω_{22}	ω_{23}	ω_{24}
Uitslag	3.5	9	9.5	3	4.5	8	7.5	6.5	8.5	4.5	9.5	7.5
Aantal jaren	2	2	2	3	2	2	2	4	2	2	2	3

Tabel 5.1: Tentamencijfers en aantal jaren studie.

Stel we willen nu de kans uitrekenen dat de getrokken student een 9.5 heeft gehaald en tweedejaars is. We noteren die gebeurtenis als $\{X = 9.5, Y = 2\}$. Hier mee bedoelen we dus $\{X = 9.5\} \cap \{Y = 2\}$. De kans 'halen we op' uit de oorspronkenkelijke kansruimte. Net als eerder schrijven we de gebeurtenis als volgt

$$\{X = 9.5, Y = 2\} = \{\omega \in \Omega | X(\omega) = 9.5, Y(\omega) = 2\}. \tag{5.1}$$

In Tabel 5.2 staan de tellingen van dergelijke gebeurtenissen in de oorspronkelijke kansruimte

		X										
		3	3.5	4.5	5	6.5	7	7.5	8	8.5	9	9.5
Y	2	0	2	2	2	0	0	1	2	1	1	4
	3	1	0	0	0	2	1	1	1	1	0	0
	4	0	0	0	0	2	0	0	0	0	0	0

Tabel 5.2: Tellingen van tentamencijfers en aantal jaren studie.

gegeven. We zien dat er vier studenten ω zijn die voldoen aan $X(\omega) = 9.5$ en $Y(\omega) = 2$. Elke student heeft kans $1/24$ dus we vinden

$$P(X = 9.5, Y = 2) = \tfrac{4}{24} = \tfrac{1}{6}. \tag{5.2}$$

We noemen dit een **simultane kans** (joint probability). Als we dit voor alle mogelijke uitkomsten van X en Y doen, dan krijgen we een zogenaamde **tabel van simultane kansen**. Tabel 5.3 geeft deze tabel voor het huidige voorbeeld. Voor het overzicht zullen we de breuken in de tabellen hieronder niet vereenvoudigen.

		X										
		3	3.5	4.5	5	6.5	7	7.5	8	8.5	9	9.5
Y	2	0	$\frac{2}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	0	0	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{4}{24}$
	3	$\frac{1}{24}$	0	0	0	$\frac{2}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	0	0
	4	0	0	0	0	$\frac{2}{24}$	0	0	0	0	0	0

Tabel 5.3: Simultane kansen van tentamencijfers en aantal jaren studie.

Deze tabel legt de hele kansverdeling van de stochastische vector vast. We kunnen bijvoorbeeld ook de kans op de gebeurtenis $X = 8$ uit de tabel halen. Immers

$$\{X = 8\} = \{X = 8, Y = 2\} \cup \{X = 8, Y = 3\} \cup \{X = 8, Y = 4\}. \tag{5.3}$$

Bovendien zijn deze gebeurtenissen disjunct. We mogen dus de kansen optellen. We vinden dan

$$\begin{aligned} P(X = 8) &= P(X = 8, Y = 2) + P(X = 8, Y = 3) + P(X = 8, Y = 4) \\ &= \tfrac{2}{24} + \tfrac{1}{24} + 0 = \tfrac{1}{8}. \end{aligned}$$

In de context van een simultane verdeling noemen we de kans $P(Y = 9.5)$ een **marginale kans** (marginal probability). We zien nu dus dat we de kansverdelingen van X en Y alleen, kunnen verkrijgen uit de simultane kanstabel door rijen en kolommen op te tellen. In Tabel 5.4 zijn deze kansverdelingen, de **marginale kanverdelingen**, aan de simultane kanstabel toegevoegd. Ze heten zo omdat ze in feite in de marge, de kantlijn, van de tabel staan. Onderaan de tabel vind je de kanverdeling van X, zoals gegeven in (3.2) terug.

		X											
		3	3.5	4.5	5	6.5	7	7.5	8	8.5	9	9.5	
Y	2	0	$\frac{2}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	0	0	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{4}{24}$	$\frac{15}{24}$
	3	$\frac{1}{24}$	0	0	0	$\frac{2}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	0	0	$\frac{7}{24}$
	4	0	0	0	0	$\frac{2}{24}$	0	0	0	0	0	0	$\frac{2}{24}$
		$\frac{1}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	$\frac{2}{24}$	$\frac{4}{24}$	$\frac{1}{24}$	$\frac{2}{24}$	$\frac{3}{24}$	$\frac{2}{24}$	$\frac{1}{24}$	$\frac{4}{24}$	

Tabel 5.4: Simultane kansen en marginale kansen van tentamencijfers en aantal jaren studie.

We kunnen nu ook voorwaardelijke kansen, zoals de kans dat de getrokken student tweedejaars is, gegeven dat hij een acht heeft gehaald, berekenen. Immers

$$P(Y = 2|X = 8) = \frac{P(Y = 2, X = 8)}{P(X = 8)} = \frac{\frac{2}{24}}{\frac{3}{24}} = \frac{2}{3}.$$

Terugkerende tot de algemene theorie willen we onafhankelijkheid van twee stochastische variabelen definiëren. Voor meerdere stochastische variabelen is de definitie net zo.

Definitie 5.2 *We noemen twee discrete stochastische variabelen X en Y* **onafhankelijk** *als voor alle uitkomsten x en y van X en Y geldt*

$$P(X = x, Y = y) = P(X = x)P(Y = y). \tag{5.4}$$

Hier staat dus dat de gebeurtenissen $\{X = x\}$ en $\{Y = y\}$ voor alle x, y onafhankelijke gebeurtenissen zijn. Ga na dat in het voorgaande voorbeeld de twee stochastische variabelen niet onafhankelijk zijn maar afhankelijk.

Als we geïnteresseerd zijn in een functie van X en Y, zeg $Y = g(X, Y)$ dan zouden we de verwachting van deze nieuwe stochastische variabele willen uitrekenen. Dit doen we weer door het gemiddelde van $g(x, y)$ te berekenen over uitkomsten (x, y), gewogen met de simultane kansen op (x, y).

Definitie 5.3 *De verwachting van $Z = g(X, Y)$ wordt gegeven door*

$$\mathrm{E}\,(Z) = \mathrm{E}\,(g(X, Y)) = \sum_{x,y} g(x, y)P(X = x, Y = y). \tag{5.5}$$

Je rekent zo'n verwachting dus uit door voor elk vakje in de simultane kanstabel van X en Y de waarde van $g(x, y)$ te berekenen, deze te vermenigvuldigen met de kans, en ze allemaal op te tellen. Als je dit doet voor $Y = aX + bY + c$, dus $g(x, y) = ax + by + c$, dan vind je

$$\boxed{\mathrm{E}\,(aX + bY + c) = a\mathrm{E}\,(X) + b\mathrm{E}\,(Y) + c}. \tag{5.6}$$

Bekijken we het product van X en Y, zeg $Z = XY$, dan ligt het voor de hand om te denken dat $\mathrm{E}\,(Z)$ gelijk is aan $\mathrm{E}\,(X)\mathrm{E}\,(Y)$. Dit geldt echter niet in het algemeen maar wel als X en Y onafhankelijk zijn. We kunnen dit als volgt inzien. Er geldt **in het geval van onafhankelijkheid van X en Y**

$$\begin{aligned}\mathrm{E}\,(Z) &= \mathrm{E}\,(XY) = \sum_{x,y} xyP(X = x, Y = y) = \sum_{x,y} xP(X = x)yP(Y = y) \\ &= \sum_{x} xP(X = x) \sum_{y} yP(Y = y) = \mathrm{E}\,(X)\mathrm{E}\,(Y).\end{aligned}$$

Er geldt dus

$$\boxed{\mathrm{E}\,(XY) = \mathrm{E}\,(X)\mathrm{E}\,(Y), \text{ als } X \text{ en} Y \text{ onafhankelijk zijn}}. \tag{5.7}$$

5.2 Continu verdeelde stochastische vectoren

Net als voor continue stochastische variabelen beschrijven we de kansverdeling van een continue stochastische vector door middel van een (simultane) kansdichtheidsfunctie. Waar we bij discrete stochastische vectoren simultane kansen moeten optellen om een kans te berekenen dat de vector in een gebied terecht komt, moeten we in het continue geval deze kansdichtheidsfunctie (dubbel) integeren over dat gebied. In het een dimensionale geval is de integraal van een positieve functie over een gebied gelijk aan de oppervlakte onder die functie en boven dat gebied. Bij een dubbelintegraal is dat de inhoud onder de functie en boven een gebied in het grondvlak.

Definitie 5.4 *We noemen een functie $f(x, y)$ een kansdichtheidsfunctie als er aan de volgende twee voorwaarden voldaan is*

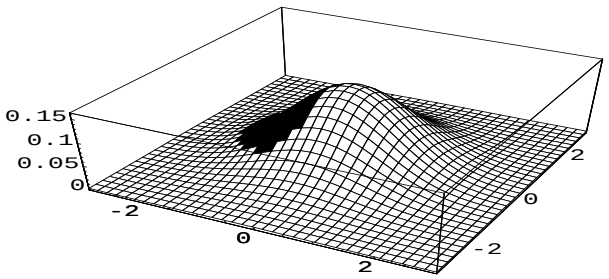
1. *$f(x, y) \geq 0$, voor alle x, y,*
2. *De inhoud van het gebied onder de functie f en boven het grondvlak is gelijk aan een. Dus*

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1. \tag{5.8}$$

We zeggen dat (X, Y) een continue stochastische vector is met kansdichtheid f, als

$$P((X, Y) \in G) = \int \int_{G} f(x, y) dx dy, \tag{5.9}$$

voor alle gebieden G in het vlak. Met andere woorden, de gegeven kans is gelijk aan de inhoud van het gebied onder f en boven het gebied G in het grondvlak.



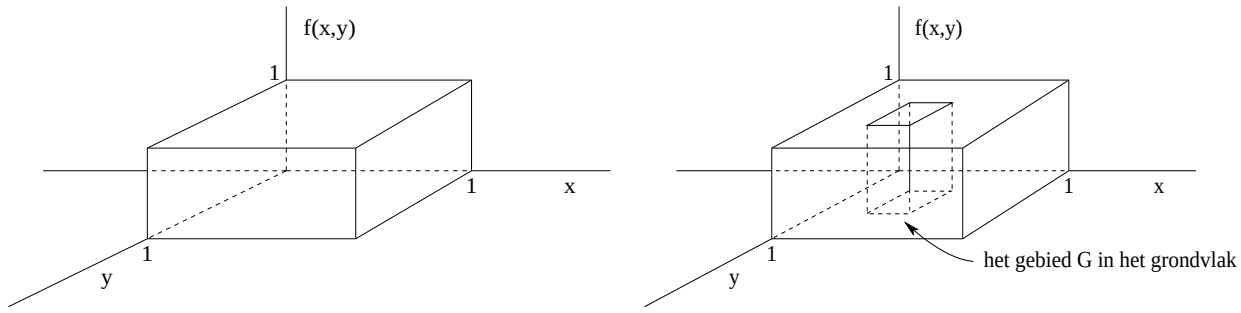
Figuur 5.2: Een kansdichtheidsfunctie $f(x, y)$.

Voorbeeld 5.5 (uniforme verdeling op het eenheidsvierkant)

Laat (X, Y) een stochastisch punt zijn in het eenheidsvierkant $[0, 1] \times [0, 1]$. Als er geen voorkeur voor de plaats is dan ligt het voor de hand om de dichtheid f in dit geval constant een te nemen op het vierkant en nul daarbuiten. Dus

$$f(x, y) = \begin{cases} 1 & \text{, als } 0 \leq x \leq 1, 0 \leq y \leq 1, \\ 0 & \text{, elders} \end{cases} \tag{5.10}$$

Als we de kans willen uitrekenen dat (X, Y) in een gebied G in het eenheidsvierkant terecht



Figuur 5.3: De kansdichtheidsfunctie $f(x, y)$ van de uniforme verdeling op het eenheidsvierkant.

komt dan is dat volgens de definitie gelijk aan de inhoud van het gebied boven G en onder f. Het geval dat G een rechthoek is is getekend in Figuur 5.3. We vinden dan, voor deze uniforme verdeling,

$$P((X, Y) \in G) = \text{ oppervlakte van } G. \tag{5.11}$$

Als we een uniforme verdeling op een gebied S, waarvan de oppervlakte niet gelijk is aan een, willen beschrijven dan wordt dit

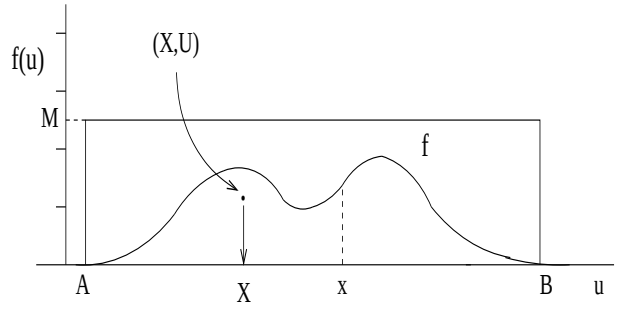
$$P((X, Y) \in G) = \frac{\text{oppervlakte van } G \cap S}{\text{oppervlakte van } S}. \tag{5.12}$$

Voor uniforme verdelingen kunnen we dus gewoon werken met oppervlaktes.

Voorbeeld 5.6 (Fundamentale benadering van simulatie)

We kunnen uniforme twee dimensionale stochastische vectoren mooi gebruiken om een trekking van een stochastische variabele met een dichtheid f te genereren. We gaan dan als volgt te werk. Teken het gebied in het vlak boven de horizontale as en onder de functie f. Noem dit gebied S. Zie Figuur 5.4. De oppervlakte van dit gebied gelijk aan een omdat f een dichtheidsfunctie is.

Stel dat het paar (X, U) een uniforme verdeling heeft op dit gebied S. De gebeurtenis $\{X \leq x\}$ is gelijk aan de gebeurtenis dat het punt (X, U) belandt in het gebied tussen de



Figuur 5.4: Schematische voorstelling van het genereren van een trekking met dichtheidsfunctie f.

horizontale as en de functie f, en links van de vertikale lijn door $(x, 0)$. Dit is het gearceerde gebied in Figuur 3.10. De oppervlakte van dat gebied is gelijk aan $F(x)$. Kortom, als we een uniform punt in S trekken, dan heeft de eerste coördinaat, X, verdelingsfunctie F en dus ook kansdichtheidsfunctie f.

Nu moeten we nog een procedure maken om een uniform punt uit S te trekken. Stel nu dat f nul is buiten een interval $[A, B]$, en dat $f(x) \leq M$, voor alle x. Dan ligt S helemaal in de rechthoek $[A, B] \times [0, M]$. We trekken nu eerst een uniform punt in die rechthoek. Dat is betrekkelijk eenvoudig. Genereer eerst twee uniforme trekkingen, U_1 en U_2, op $[0, 1]$. Dan zijn $V_1 = A + (B - A)U_1$ en $V_2 = MU_2$ respectievelijk uniform verdeeld op $[A, B]$ en $[0, M]$. Het paar (V_1, V_2) is uniform verdeeld op de gegeven rechthoek. Pas nu een **acceptance-rejection** stap toe. Als het punt (V_1, V_2) in S ligt dan accepteer je V_1 als gegenereerde waarde. Als dat niet zo is dan begin je opnieuw.

Het fundamentele aan deze aanpak is dat je een acceptance-rejection stap hebt en dat je een (tweede) hulpvariabele nodig hebt. Dit zijn eigenschappen die een aantal andere simulatie methoden ook hebben.

We hebben in het discrete geval gezien dat we de marginale kansverdelingen van X en Y uit de simultane kanstabel kunnen verkrijgen door over de kolommen en rijen te sommeren. Dit werkt net zo in het continue geval. Alleen moeten we dan integreren. Er geldt dan

$$\boxed{f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy, \quad \text{en} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx.} \tag{5.13}$$

Voor het berekenen van de verwachting van een stochastische variabele Z die een functie is van X en Y, zeg $Z = g(X, Y)$ gebruiken we

$$\boxed{\mathrm{E}\,(Z) = \mathrm{E}\,(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)dxdy.} \tag{5.14}$$

Onafhankelijkheid van twee stochastische variabelen betekent voor discrete variabelen dat alle simultane kansen $P(X = x, Y = y)$ gelijk moeten zijn aan het product van de marginale kansen $P(X = x)$ en $P(Y = y)$. Voor continue stochastische variabelen hebben een soortgelijke definitie in termen van de simultane kansdichtheidsfunctie.

Definitie 5.7 *We noemen twee continue stochastische variabelen X en Y, met simultane kansdichtheidsfunctie f,* **onafhankelijk** *als voor alle uitkomsten x en y van X en Y geldt*

$$f(x,y) = f_X(x) f_Y(y), \tag{5.15}$$

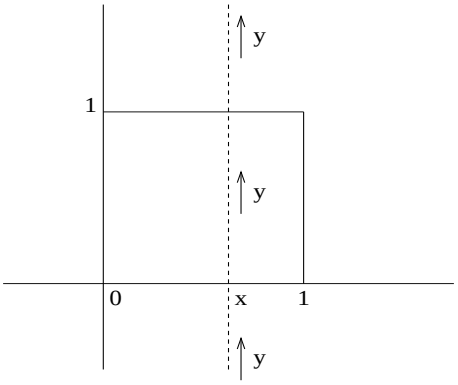
waarbij $f_X(x)$ en $f_Y(y)$ de marginale kansdichtheidsfuncties van X en Y zijn.

Voorbeeld 5.8 (uniforme verdeling op het eenheidsvierkant)

Als het paar (X, Y) uniform verdeeld is op het eenheidsvierkant, zie Voorbeeld 5.5, dan kunnen we (5.13) gebruiken om bijvoorbeeld de kansdichtheid van X uit te rekenen. We vinden dan voor $0 \leq x \leq 1$

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^1 1 dy = 1.$$

Voor andere waarden van x vinden we nul. Om de integraal uit te rekenen kies je een vaste waarde van x en laat je y lopen over de gestippelde lijn die aangegeven is in Figuur 5.5. Onder het integraalteken staan dan de waarden van (x, y) die je tegenkomt. Hier zien we dus dat X



Figuur 5.5: Berekening van de marginale kansdichtheid van X.

uniform verdeeld is op $[0, 1]$. Met een dergelijke berekening kan je laten zien dat Y ook deze verdeling heeft. De stochastische variabelen X en Y in dit voorbeeld zijn onafhankelijk omdat, voor (x, y) in het vierkant,

$$f(x,y) = 1 = 1 \times 1 = f_X(x) f_Y(y). \tag{5.16}$$

Voor (x, y) buiten het vierkant is $f(x, y)$ gelijk aan nul, en minstens een van $f_X(x)$ en $f_Y(y)$ is ook gelijk aan nul. Dan klopt het product dus ook.

5.3 Covariantie en simultane verdelingsfunctie

Men zou kunnen denken dat de variantie van de som van twee stochastische variabelen gelijk is aan de som van hun varianties. Laten we dat eens nader bekijken. Uit de definitie van variantie volgt

$$\begin{aligned}
\mathrm{Var}\,(X+Y) &= \mathrm{E}\left((X+Y-\mathrm{E}\,(X+Y))^2\right) \\
&= \mathrm{E}\left(([X-\mathrm{E}\,(X)]+[Y-\mathrm{E}\,(Y)])^2\right) \\
&= \mathrm{E}\left([X-\mathrm{E}\,(X)]^2+[Y-\mathrm{E}\,(Y)]^2+2[X-\mathrm{E}\,(X)][Y-\mathrm{E}\,(Y)]\right) \\
&= \mathrm{E}\left([X-\mathrm{E}\,(X)]^2)+\mathrm{E}\,([Y-\mathrm{E}\,(Y)]^2)+2\mathrm{E}\,([X-\mathrm{E}\,(X)][Y-\mathrm{E}\,(Y)]\right) \\
&= \mathrm{Var}\,(X)+\mathrm{Var}\,(Y)+2\mathrm{E}\,([X-\mathrm{E}\,(X)][Y-\mathrm{E}\,(Y)]).
\end{aligned}$$

We zien hieruit dat de variantie van de som van twee stochastische variabelen in het algemeen niet gelijk is aan de som van de varianties. Er is een correctieterm die gelijk is aan twee maal $\mathrm{E}\,([X-\mathrm{E}\,(X)][Y-\mathrm{E}\,(Y)]\Big)$. Deze laatste grootheid noemen we de covariantie van X en Y.

Definitie 5.9 *De* **covariantie** *(covariance) van twee stochastische variabelen X en Y wordt gegeven door*

$$\mathrm{Cov}(X,Y)=\mathrm{E}\,((X-\mathrm{E}\,(X))(Y-\mathrm{E}\,(Y))). \tag{5.17}$$

Net als voor de variantie is er een alternatieve manier om de covariantie uit te rekenen. Er geldt namelijk

$$\boxed{\mathrm{Cov}(X,Y)=\mathrm{E}\,(XY)-\mathrm{E}\,(X)\mathrm{E}\,(Y).} \tag{5.18}$$

Als we Y gelijk nemen aan X dan vinden we

$$\boxed{\mathrm{Cov}(X,X)=\mathrm{E}\,([X-\mathrm{E}\,(X)]^2)=\mathrm{Var}\,(X).} \tag{5.19}$$

Echter, als X en Y onafhankelijk zijn, dan hebben we gezien dat de verwachting van hun product gelijk is aan het product van de verwachtingen. We vinden dan dus

$$\boxed{\mathrm{Cov}(X,Y)=0,\ \text{als } X \text{ en Y onafhankelijk zijn.}} \tag{5.20}$$

Uit het laatste volgt dat we inderdaad de varianties mogen optellen als de stochasten maar onafhankelijk zijn.

$$\boxed{\mathrm{Var}\,(X+Y)=\mathrm{Var}\,(X)+\mathrm{Var}\,(Y),\ \text{als } X \text{ en Y onafhankelijk zijn.}} \tag{5.21}$$

Dit geldt ook als je meerdere stochastische variabelen optelt.

Voorbeeld 5.10 (Genereren (bij benadering) van een normale trekking)

Als we een stochastische variabele nemen die uniform verdeeld is op het interval $[0, 1)$, zie Sectie 4.2.1, dan weten we dat zijn verwachting gelijk is aan $\frac{1}{2}$ en zijn variantie gelijk aan $\frac{1}{12}$. Neem nu 12 onafhankelijke stochastische variabelen $U_1, \ldots, U_{12}$ met zo'n uniforme verdeling. Dan geldt voor

$$X = U_1 + \ldots + U_{12} - 6 \tag{5.22}$$

dat zijn verwachting en variantie gegeven worden door

$$\begin{aligned} &\mathrm{E}\,(X) = \mathrm{E}\,(U_1 + \ldots + U_{12} - 6) = 12 \times \tfrac{1}{2} - 6 = 0, \\ &\mathrm{Var}\,(X) = \mathrm{Var}\,(U_1 + \ldots + U_{12} - 6) = \mathrm{Var}\,(U_1 + \ldots + U_{12}) = 12\mathrm{Var}\,(U_1) = 1. \end{aligned}$$

Omdat we hier te maken hebben met een som van onafhankelijke gelijkverdeelde bijdragen is de kansverdeling van X ook vrijwel, maar niet helemaal, normaal. We zien hier dus dat we bij benadering een standaardnormale trekking kunnen genereren uit twaalf uniforme trekkingen. Het is wel een goede benadering.

We hebben de verdelingsfunctie van afzonderlijke stochastische variabelen ingevoerd als de functie die de cumulatieve kansen geeft. Iets dergelijks kunnen we ook doen voor meerdere stochastische variabelen tegelijk.

Definitie 5.11 *De* **simultane verdelingsfunctie** *(joint distribution function) van twee stochastische variabelen X en Y, wordt gegeven door*

$$F(x, y) = P(X \leq x, Y \leq y). \tag{5.23}$$

Als X en Y onafhankelijk zijn dan zijn de gebeurtenissen $X \leq x$ en $Y \leq y$ onafhankelijke gebeurtenissen. We vinden dan

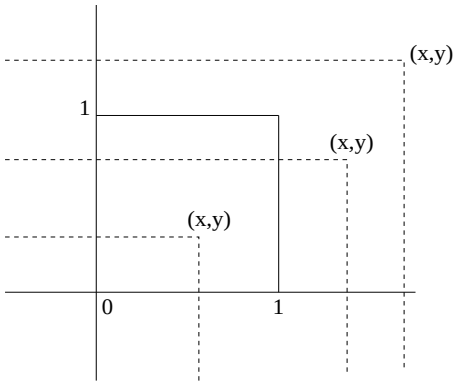
$$\boxed{F(x, y) = F_X(x)F_Y(y), \text{ als } X \text{ en Y onafhankelijk zijn,}} \tag{5.24}$$

waarbij F_X en F_Y de marginale verdelingsfuncties van X en Y zijn.

Voorbeeld 5.12 (uniforme verdeling op het eenheidsvierkant)

Bij Voorbeeld 5.5 hebben we gezien dat we bij een paar (X, Y) dat uniform verdeeld is op het eenheidsvierkant kansen kunnen uitrekenen door middel van oppervlakten. Als we de simultane verdelingsfunctie van X en Y willen bepalen dan moeten we de kans op de gebeurtenis $\{X \leq x, Y \leq y\}$ bepalen voor elke x, y. Dit zijn dus kansen op de linker beneden rechthoeken, kwadranten, met als hoekpunt (x, y). Deze zijn voor drie gevallen aangegeven in Figuur 5.6. Als we de oppervlakten uitrekenen dan vinden we

$$F(x, y) = \begin{cases} 0 & , \text{ als } x < 0 \text{ of } y < 0, \\ x & , \text{ als } 0 \leq x \leq 1 \text{ en } y > 1, \\ y & , \text{ als } 0 \leq y \leq 1 \text{ en } x > 1, \\ xy & , \text{ als } 0 \leq x \leq 1 \text{ en } 0 \leq y \leq 1, \\ 1 & , \text{ als } x \geq 1 \text{ en } y \geq 1. \end{cases}$$



Figuur 5.6: Berekening van de simultane verdelingsfunctie van X en Y.

Deze simultane verdelingsfunctie ziet er wat ingewikkeld uit door het aantal gevallen dat we moeten onderscheiden. In dit geval is de simultane dichtheid een stuk eenvoudiger. We kunnen hier ook controleren dat (5.24) geldt. Dat moet ook want we hebben al eerder gezien dat X en Y hier onafhankelijk zijn.

5.4 Multivariate waarnemingen

Naast het schatten van parameters en het toetsen van hypothesen zijn er andere vraagstellingen die een statistische aanpak vereisen. Als men bijvoorbeeld per object (persoon, plant, etc.) meerdere kenmerken meet, dan kan het bijvoorbeeld van belang zijn een kenmerk uit de andere te verklaren of te voorspellen. Dit soort problemen noemen we *regressieproblemen*. Ook zou het van belang kunnen zijn op grond van de gegevens van verschillende groepen objecten een nieuw gemeten object aan een van die groepen toe te kunnen wijzen. Dit noemen we *classificatie* of *discriminantanalyse*.

Laten we een voorbeeld geven van een classificatieprobleem. De onderstaande tabel bevat metingen aan 150 Irissen. Gemeten zijn de lengte en breedte van het bloemblad en van het kelkblad. In feite bestaat het bestand uit drie soorten Irissen, 50 van elke soort. Van de Irissen in de tabel is de soort bekend. Het probleem is nu een methode te ontwerpen waarmee we van een plant, waarvan alleen de vier metingen bekend zijn, de soort kunnen bepalen. De gegevens staan in een artikel van R.A. Fisher uit 1936. Fisher is een van de grondleggers van de moderne statistiek, maar dit ter zijde.

nr.	bloemblad-lengte	bloemblad-breedte	kelkblad-lengte	kelkblad-breedte
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5.0	3.4	1.5	0.2
⋮	⋮	⋮	⋮	⋮
148	6.5	3.0	5.2	2.0
149	6.2	3.4	5.4	2.3
150	5.9	3.0	5.1	1.8

Tabel 1.1. *Lengte en breedte van bloemblad en kelkblad van 150 Irissen (in cm).*

In het algemeen beschrijven we een dergelijke datamatrix door middel van een matrix met componenten $x_{i,j}$,

$$\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}, \tag{5.25}$$

waarbij i het object aangeeft en j het gemeten kenmerk. In het voorbeeld is p gelijk aan vier. Per Iris zijn vier kenmerken gemeten.

Als we voor deze waarnemingen een statistisch model willen maken dan zullen we veronderstellingen moeten doen. We nemen bijvoorbeeld aan dat metingen aan verschillende Irissen elkaar niet beïnvloeden. We beschouwen de datamatrix dan ook als een realisatie van een matrix met stochastische variabelen als componenten,

$$\begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,p} \end{pmatrix}, \tag{5.26}$$

waarvan de stochasten in verschillende rijen onafhankelijk zijn. De stochasten binnen elke rij mogen echter wel afhankelijk zijn van elkaar. Dit is in feite essentieel als we bijvoorbeeld een van de componenten uit andere componenten willen verklaren.

Laat X_i de staande vector zijn die bestaat uit de elementen van de i-de rij van de datamatrix, dus

$$X_i = \begin{pmatrix} X_{i,1} \\ \vdots \\ X_{i,p} \end{pmatrix}. \tag{5.27}$$

Deze vector bestaat dus uit de metingen aan het i-de object. Het statistische model voor de gegevens bestaat uit een rij $X_1, \ldots, X_n$ van onafhankelijke stochastische p-vectoren. Eigenschappen van deze stochastische vectoren bespreken we in de volgende sectie.

5.5 Stochastische vectoren

De kansverdeling van een stochastische variabele wordt bepaald door zijn verdelingsfunctie $F_X(x) = P(X \leq x)$. Hij wordt ook bepaald door de kansen $P(X = x)$ op de verschillende uitkomsten in het discrete geval, of door zijn kansdichtheidsfunctie $f_X(x)$ in het continue geval. Als we alleen geïnteresseerd zijn in locatie en schaal van de verdeling van X dan berekenen we zijn verwachting en variantie.

Voor een stochastische vector zijn de simultane kansen $P(X_1 = x_1, \ldots, X_p = x_p)$ van zijn componenten, in het discrete geval, en de simultane kansdichtheidsfunctie $f_{X_1,\ldots,X_p}(x_1, \ldots, x_p)$, in het continue geval, bepalend voor de kansverdeling. Ook hier willen we karakteristieken voor locatie en schaal hebben. De voor de hand liggende generalisaties van het eendimensionale geval verkrijgen we als volgt.

Laat X een stochastische p-vector zijn. Dus

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}. \tag{5.28}$$

met $X_1, \ldots, X_p$ stochastische variabelen. De *verwachtingsvector* van X definiëren we als de p-vector van verwachtingen van de componenten, dus

$$\boxed{\mathrm{E}\, X = \begin{pmatrix} \mathrm{E}\, X_1 \\ \vdots \\ \mathrm{E}\, X_p \end{pmatrix}.} \tag{5.29}$$

Vervolgens definiëren we het analogon van de variantie als de *covariantiematrix*

$$\boxed{\mathrm{Cov}(X) = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \cdots & \mathrm{Cov}(X_p, X_p) \end{pmatrix} = \begin{pmatrix} \sigma_{1,1} & \cdots & \sigma_{1,p} \\ \vdots & \ddots & \vdots \\ \sigma_{p,1} & \cdots & \sigma_{p,p} \end{pmatrix},} \tag{5.30}$$

waarbij

$$\sigma_{i,j} = \mathrm{Cov}(X_i, X_j) = \mathrm{E}\,(X_i - \mathrm{E}\,X_i)(X_j - \mathrm{E}\,X_j), \quad \text{voor } i = 1, \ldots, p,\ j = 1, \ldots, p. \tag{5.31}$$

Deze matrix noteren we gebruikelijk met de griekse hoofdletter sigma Σ. Een covariantiematrix is symmetrisch omdat $\mathrm{Cov}(X_i, X_j)$ gelijk is aan $\mathrm{Cov}(X_j, X_i)$. Op de diagonaal staan de varianties van de componenten van de stochastische vector, immers $\sigma_{i,i} = \mathrm{Cov}(X_i, X_i) = \mathrm{Var}\,(X_i)$.

Als de componenten van X onafhankelijk zijn dan weten we dat hun covarianties nul zijn. Dus $\mathrm{Cov}(X_i, X_j) = 0$ voor $i \neq j$. De covariantiematrix van X is dan dus gelijk aan een diagonaal matrix met de varianties van $X_1, \ldots, X_p$ op de diagonaal, immers

$$\Sigma = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \cdots & \mathrm{Cov}(X_p, X_p) \end{pmatrix} = \begin{pmatrix} \mathrm{Var}\,(X_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathrm{Var}\,(X_p) \end{pmatrix}. \tag{5.32}$$

Hoe meer een covariantiematrix lijkt op een diagonaalmatrix, des te onafhankelijker zijn de componenten. Let hierbij wel op dat het gelijk aan nul zijn van een covariantie niet automatisch impliceert dat de stochasten onafhankelijk zijn.

Vervolgens zullen we de verwachtingsvector en covariantiematrix van een lineaire afbeelding van X uitdrukken in de verwachtingsvector en covariantiematrix van X zelf. Zij A een matrix met p kolommen en q rijen, dus

$$A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,p} \\ \vdots & \ddots & \vdots \\ a_{q,1} & \cdots & a_{q,p} \end{pmatrix}, \tag{5.33}$$

dan is AX een stochastische q vector,

$$AX = \begin{pmatrix} a_{1,1} & \cdots & a_{1,p} \\ \vdots & \ddots & \vdots \\ a_{q,1} & \cdots & a_{q,p} \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} a_{1,1}X_1 + \ldots + a_{1,p}X_p \\ \vdots \\ a_{q,1}X_1 + \ldots + a_{q,p}X_p \end{pmatrix}. \tag{5.34}$$

Voor de verwachting van de stochastische vector vinden we

$$\boxed{\mathrm{E}\,(AX) = A(\mathrm{E}\,X),} \tag{5.35}$$

immers

$$\begin{pmatrix} \mathrm{E}\,(a_{1,1}X_1 + \ldots + a_{1,p}X_p) \\ \vdots \\ \mathrm{E}\,(a_{q,1}X_1 + \ldots + a_{q,p}X_p) \end{pmatrix} = \begin{pmatrix} a_{1,1}\mathrm{E}\,X_1 + \ldots + a_{1,p}\mathrm{E}\,X_p \\ \vdots \\ a_{q,1}\mathrm{E}\,X_1 + \ldots + a_{q,p}\mathrm{E}\,X_p \end{pmatrix}. \tag{5.36}$$

Voor de covariantie matrix van de stochastische vector X geldt ook een eenvoudige regel

$$\boxed{\mathrm{Cov}(AX) = A\mathrm{Cov}(X)A^T.} \tag{5.37}$$

Het bewijs berust op uitschrijven van matrixproducten. Zie Paragraaf 5.9.1.

Een betere karakteristiek om de samenhang van de componenten van de stochastische vector X te beschrijven is de *correlatiematrix* die bestaat uit de correlaties tussen de componenten. Deze matrix verandert niet als we de locatie en schaal van de componenten veranderen. Noteren we de correlatie tussen X_i en X_j met $\rho_{i,j}$, dus

$$\boxed{\rho_{i,j} = \mathrm{Corr}(X_i, X_j) = \frac{\mathrm{Cov}(X_i,X_j)}{\sqrt{\mathrm{Var}\,(X_i)\mathrm{Var}\,(X_j)}},} \tag{5.38}$$

dan is de correlatiematrix gelijk aan ($\rho_{i,i} = 1, i = 1, \ldots, p$)

$$\begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,p-1} & \rho_{1,p} \\ \rho_{2,1} & 1 & \cdots & \rho_{2,p-1} & \rho_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{p,1} & \rho_{p,2} & \cdots & \rho_{p,p-1} & 1 \end{pmatrix}. \tag{5.39}$$

Ook hier geldt, des te meer de correlatiematrix lijkt op de eenheidsmatrix, des te onafhankelijker zijn de componenten van X.

5.6 De multivariate normale verdeling

Een belangrijke multivariate verdeling is de multivariate normale verdeling. We voeren de p-variate normale kansdichtheid in als generalisatie naar p dimensies van de univariate normale dichtheid. De univariate normale dichtheid met verwachting μ en variantie σ^2 wordt gegeven door

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}((x-\mu)/\sigma)^2}, \quad -\infty < \mu < \infty,\ \sigma \geq 0. \tag{5.40}$$

Deze dichtheid noemen we de $N(\mu, \sigma^2)$ dichtheid.

De kwadratische term in de exponent is gelijk aan

$$-\tfrac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2 = -\tfrac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu) \tag{5.41}$$

Dit kunnen we generaliseren naar een kansdichtheid van gelijke vorm voor een stochastische p vector met verwachtingsvector μ en covariantiematrix Σ door $(\sigma^2)^{-1}$ in (5.41) te vervangen door de inverse van de covariantiematrix Σ. Als generalisatie krijgen we dan

$$-\tfrac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) \tag{5.42}$$

waarbij $x = (x_1, \ldots, x_p)^T$ en $\mu = (\mu_1, \ldots, \mu_p)^T$. De resulterende kansdichtheid van X is dan gelijk aan

$$f(x) = C\ e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}, \tag{5.43}$$

met C een normeringsconstante die zo gekozen is dat de integraal van f, of, met andere woorden, het volume onder het oppervlak beschreven door f, gelijk is aan een. Deze constante kan worden uitgedrukt in termen van Σ. We vinden dan de volgende formule voor de zogenaamde $N_p(\mu, \Sigma)$ dichtheid

$$f(x) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{p/2}}\ e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}, \tag{5.44}$$

waarbij $|\Sigma|$ de determinant van Σ is.

Zonder bewijs noemen we een aantal eigenschappen van de multivariate normale verdeling:

1. Lineaire combinaties van componenten van X zijn normaal verdeeld.
2. Deelverzamelingen van de componenten van X zijn weer multivariaat normaal verdeeld. In het bijzonder zijn de marginale verdelingen van de componenten normaal. De i-de component X_i is $N(\mu_i, \sigma_{i,i})$ verdeeld.
3. Als de covarianties nul zijn dan zijn de stochasten onafhankelijk (dit geldt niet voor willekeurige stochasten !).

Bewijzen van deze beweringen zijn te vinden in bijvoorbeeld Johnson & Wichern (1998). Bewering (3) kunnen we echter eenvoudig controleren. Als de covarianties tussen de componenten van X nul zijn dan worden de covariantiematrix van X en zijn inverse gegeven door de diagonaalmatrices

$$\Sigma = \begin{pmatrix} \sigma_{1,1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{p,p} \end{pmatrix} \quad \text{en} \quad \Sigma^{-1} = \begin{pmatrix} \sigma_{1,1}^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{p,p}^{-1} \end{pmatrix}.$$

Dus $|\Sigma| = \sigma_{1,1} \ldots \sigma_{p,p}$ en

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = \frac{(x_1-\mu_1)^2}{\sigma_{1,1}} + \cdots + \frac{(x_p-\mu_p)^2}{\sigma_{p,p}}.$$

Als we dit invullen in de algemene formule (5.44) voor een multivariate normale dichtheid, dan zien we dat deze gelijk is aan een product van p univariate normale dichtheden

$$\frac{1}{\sqrt{2\pi\sigma_{1,1}}}\ e^{-\frac{1}{2}((x_1-\mu_1)/\sqrt{\sigma_{1,1}})^2} \cdots \frac{1}{\sqrt{2\pi\sigma_{p,p}}}\ e^{-\frac{1}{2}((x_p-\mu_p)/\sqrt{\sigma_{p,p}}))^2},$$

de simultane dichtheid van p onafhankelijk univariate normale stochasten.

De bivariate normale verdeling.

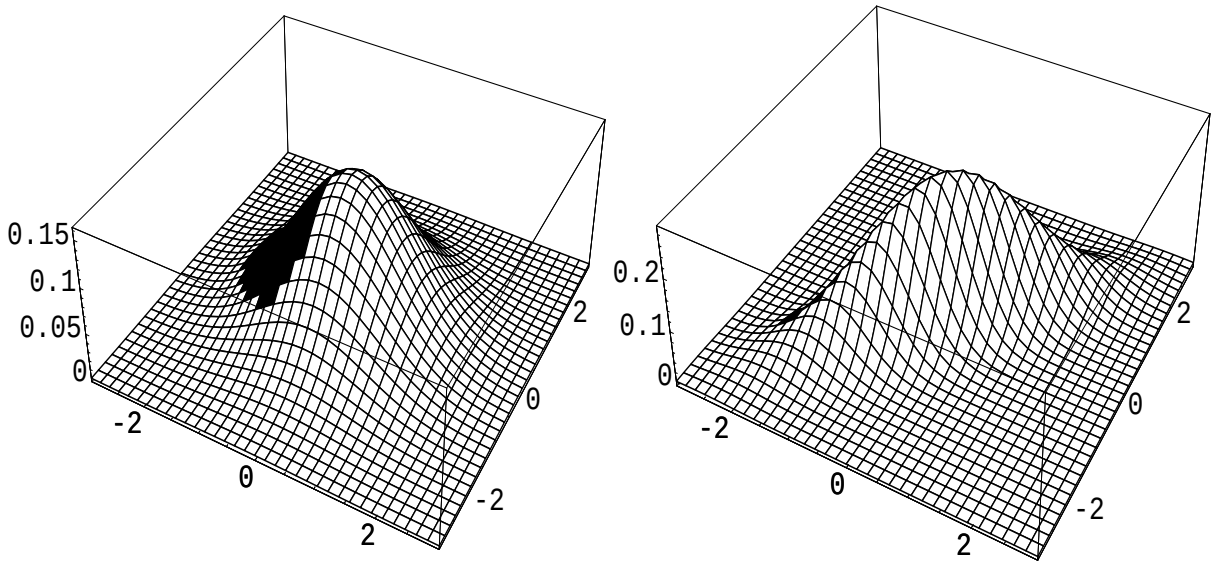
Nemen we p gelijk aan twee dan verkrijgen we de zogenaamde bivariate normale verdeling. Stel dat $X = (X_1, X_2)^T$ een bivariaat normaal verdeelde stochastische vector is met $\mathrm{E}\, X_1 =$

μ_1, $\mathrm{E}\,X_2 = \mu_2$, $\mathrm{Var}\,(X_1) = \sigma_{1,1}$, $\mathrm{Var}\,(X_2) = \sigma_{2,2}$, $\mathrm{Cov}(X_1, X_2) = \sigma_{1,2}$ en $\mathrm{Corr}(X_1, X_2) = \sigma_{1,2}/(\sqrt{\sigma_{1,1}}\sqrt{\sigma_{2,2}}) = \rho_{1,2}$. De covariantiematrix en zijn inverse worden gegeven door

$$\Sigma = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_{2,2} \end{pmatrix} \quad \text{en} \quad \Sigma^{-1} = \frac{1}{\sigma_{1,1}\sigma_{2,2} - \sigma_{1,2}^2} \begin{pmatrix} \sigma_{2,2} & -\sigma_{1,2} \\ -\sigma_{1,2} & \sigma_{1,1} \end{pmatrix}. \tag{5.45}$$

Na wat rekenwerk blijkt, met $x = (x_1, x_2)^T$,

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = \frac{1}{1-\rho_{1,2}^2}\Big[\Big(\frac{x_1-\mu_1}{\sqrt{\sigma_{1,1}}}\Big)^2 + \Big(\frac{x_2-\mu_2}{\sqrt{\sigma_{2,2}}}\Big)^2 - 2\rho_{1,2}\Big(\frac{x_1-\mu_1}{\sqrt{\sigma_{1,1}}}\Big)\Big(\frac{x_2-\mu_2}{\sqrt{\sigma_{2,2}}}\Big)\Big]. \tag{5.46}$$



Figuur 5.7: *Bivariate normale dichtheden. Links: correlatie nul. Rechts: correlatie 0.834.*

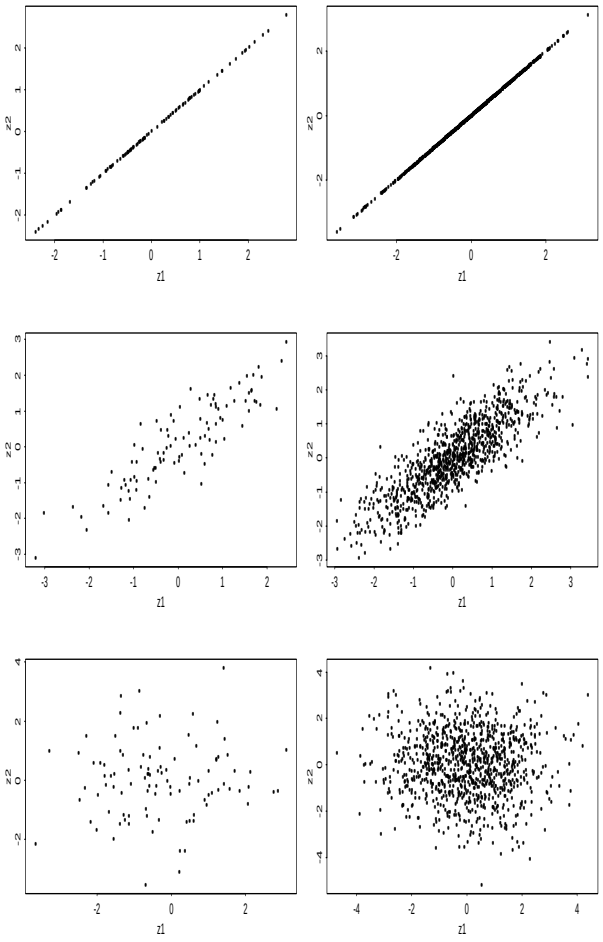
Omdat de determinant van Σ gelijk is aan

$$|\Sigma| = \sigma_{1,1}\sigma_{2,2} - \sigma_{1,2}^2 = \sigma_{1,1}\sigma_{2,2}(1-\rho_{1,2}^2) \tag{5.47}$$

vinden we de volgende uitdrukking voor de bivariate normale dichtheid

$$\begin{aligned} f(x_1, x_2) \;=\; & \frac{1}{2\pi\sqrt{\sigma_{1,1}\sigma_{2,2}(1-\rho_{1,2}^2)}} \\ & \times \exp\Big\{ -\frac{1}{2(1-\rho_{1,2}^2)}\Big[\Big(\frac{x_1-\mu_1}{\sqrt{\sigma_{1,1}}}\Big)^2 + \Big(\frac{x_2-\mu_2}{\sqrt{\sigma_{2,2}}}\Big)^2 \\ & -2\rho_{1,2}\Big(\frac{x_1-\mu_1}{\sqrt{\sigma_{1,1}}}\Big)\Big(\frac{x_2-\mu_2}{\sqrt{\sigma_{2,2}}}\Big)\Big]\Big\} \end{aligned} \tag{5.48}$$

Als $\rho_{1,2}$ gelijk is aan nul, en X_1 en X_2 dus ongecorreleerd zijn, dan is $f(x_1, x_2)$ gelijk aan $f_{X_1}(x_1)f_{X_2}(x_2)$. Volgens de definitie zijn X_1 en X_2 dan dus onafhankelijk.



Figuur 5.8: *Gesimuleerde steekproeven uit de verdeling van Z. Boven: $a = 0$, $n = 100$ (links) en $n = 1000$ (rechts). Midden: $a = 0.3$, $n = 100$ en $n = 1000$. Onder: $a = 1$, $n = 100$ en $n = 1000$.*

5.7 Een voorbeeld

Laten X en Y onafhankelijke standaardnormaal verdeelde stochastische variabelen zijn. Definieer nu de stochastische vector

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} X + aY \\ X - aY \end{pmatrix} = \begin{pmatrix} 1 & a \\ 1 & -a \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}, \tag{5.49}$$

voor een zeker getal a. Uit het voorafgaande volgt dat Z bivariaat normaal verdeeld is met verwachtingsvector nul en covariantiematrix

$$\Sigma_a = \begin{pmatrix} 1 & a \\ 1 & -a \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & a \\ 1 & -a \end{pmatrix}^T = \begin{pmatrix} 1+a^2 & 1-a^2 \\ 1-a^2 & 1+a^2 \end{pmatrix}.$$

De correlatiematrix van Z is gelijk aan

$$R_a = \begin{pmatrix} 1 & (1-a^2)/(1+a^2) \\ (1-a^2)/(1+a^2) & 1 \end{pmatrix}.$$

Voor a gelijk aan een hebben we dus onafhankelijkheid en voor a gelijk aan nul volledige afhankelijkheid. In Figuur 5.8 staan zes gesimuleerde steekproeven uit de verdeling van Z. De steekproefomvang is 100 en 1000. De waarden van a zijn 0, 0.3 en 1. De corresponderende correlaties tussen Z_1 en Z_2 zijn gelijk aan 1, 0.834 en 0.

5.8 Schatters van de verwachtingsvector en covariantiematrix

Stel dat we beschikken over een steekproef die in het kansmodel beschreven wordt door een rij $X_1, \ldots, X_n$ van onafhankelijke stochastische p vectoren met gelijke verwachtingsvectoren μ en covariantiematrices $\Sigma = (\sigma_{i,j})$. Dan is het gemiddelde van de vectoren, $\bar{X}$, een zuivere schatter van μ, immers

$$\boxed{\begin{array}{rcl} \bar{X} & = & \frac{1}{n}(X_1 + \ldots + X_n) = \dfrac{1}{n}\left[\begin{pmatrix} X_{1,1} \\ \vdots \\ X_{1,p} \end{pmatrix} + \ldots + \begin{pmatrix} X_{n,1} \\ \vdots \\ X_{n,p} \end{pmatrix}\right] \\ & = & \begin{pmatrix} \frac{1}{n}(X_{1,1} + \ldots + X_{n,1}) \\ \vdots \\ \frac{1}{n}(X_{1,p} + \ldots + X_{n,p}) \end{pmatrix} = \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{pmatrix}. \end{array}} \tag{5.50}$$

Hierbij hebben we de i-de component van deze gemiddelde vector geschreven als $\bar{X}_i$. In termen van de datamatrix is $\bar{X}$ de vector van de kolomgemiddelden.

De steekproefcovariantiematrix S voeren we in als

$$\boxed{S = \begin{pmatrix} S_{1,1} & \cdots & S_{1,p} \\ \vdots & \ddots & \vdots \\ S_{p,1} & \cdots & S_{p,p} \end{pmatrix},} \tag{5.51}$$

waarbij

$$\boxed{S_{i,k} = \tfrac{1}{n-1}\textstyle\sum_{j=1}^{n}(X_{j,i} - \bar{X}_i)(X_{j,k} - \bar{X}_k).} \tag{5.52}$$

Dit is dus de matrix van de inproducten van de gecentreerde kolommen van de datamatrix, gedeeld door $n-1$. De steekproefcovarianties $S_{i,k}$ zijn zuivere schatters van $\sigma_{i,j}$, zie bijvoorbeeld Johnson & Wichern (1998).

Voorbeeld Beschouw de datamatrix

$$X = \begin{pmatrix} 3 & 12 \\ 4 & 6 \\ 1 & 6 \\ 4 & 8 \end{pmatrix}$$

Bereken eerst de kolomgemiddelden: $\bar{X}_1 = 3, \bar{X}_2 = 8$. Vervolgens centreren we, nemen we de inproducten en delen we door $n - 1$

$$X \overset{centreren}{\rightarrow} \begin{pmatrix} 0 & 4 \\ 1 & -2 \\ -2 & -2 \\ 1 & 0 \end{pmatrix} \overset{inproducten\ nemen\ en\ delen}{\rightarrow} S = \tfrac{1}{3} \begin{pmatrix} 6 & 2 \\ 2 & 24 \end{pmatrix} = \begin{pmatrix} 2 & \frac{2}{3} \\ \frac{2}{3} & 8 \end{pmatrix}$$

5.9 Technische details

5.9.1 Bewijs van (5.37)

De matrix $\mathrm{Cov}(AX)$ is een q bij q matrix. Het i,j-de element is gelijk aan

$$\mathrm{Cov}(a_{i,1}X_1 + \ldots + a_{i,p}X_p, a_{j,1}X_1 + \ldots + a_{j,p}X_p) = \sum_{k=1}^{p}\sum_{l=1}^{p} a_{i,k}a_{j,l}\mathrm{Cov}(X_k, X_l). \tag{5.53}$$

Dit volgt uit de rekenregel voor covarianties van stochastische variabelen in Groeneboom et al. (1997).

Vervolgens laten we zien dat dit gelijk is aan het i,j-de element van $A\Sigma A^T$. Het i,j-de element van $A\Sigma$ is gelijk aan

$$(A\Sigma)_{i,j} = \sum_{k=1}^{p} a_{i,k}\mathrm{Cov}(X_k, X_j).$$

Het i,j-de element van $A\Sigma A^T = (A\Sigma)A^T$ is nu gelijk aan

$$\sum_{l=1}^{p}(A\Sigma)_{i,l}(A^T)_{l,j} = \sum_{l=1}^{p}\Big(\sum_{k=1}^{p} a_{i,k}\mathrm{Cov}(X_k, X_l)\Big)a_{j,l} = \sum_{k=1}^{p}\sum_{l=1}^{p} a_{i,k}a_{j,l}\mathrm{Cov}(X_k, X_l),$$

hetgeen inderdaad gelijk is aan (5.53).

5.10 Opgaven

1. Laten X en Y twee stochastische variabelen zijn met

$$\mathrm{E}(X) = 2, \mathrm{E}(Y) = 3 \text{ en } \mathrm{Var}(X) = 4.$$

   (a) Bereken $\mathrm{E}(3 - 2X)$ en $\mathrm{Var}(3 - 2X)$.

   (b) Bereken $\mathrm{E}(3X - 2Y + 2)$.

   (c) Bereken $\mathrm{E}(2(X + 3)^2)$.

2. Stel dat de simultane kansverdeling van X en Y gegeven wordt door de volgende tabel:

		X		
		1	2	3
Y	1	1/4	0	1/4
	2	0	1/4	0
	3	1/8	0	1/8

   (a) Vul de tabel aan met de marginale verdelingen van X en Y.

   (b) Bereken de verwachtingen van X en Y.

   (c) Bereken de verwachting van XY.

   (d) Zijn X en Y onafhankelijk?

   (e) Bereken de covariantie van X en Y.

3. Stel dat de simultane kansverdeling van X en Y gegeven wordt door de volgende tabel:

		X			
		0	1	2	
	0	1/12	?	?	?
Y	1	1/4	?	1/8	?
	2	1/6	1/12	?	1/3
		1/2	?	1/4	

   (a) Vul de tabel verder in als gegeven is dat de verwachting van Y gelijk is aan 7/6.

   (b) Bepaal de covariantie van X en Y. Zijn X en Y onafhankelijk?

4. Stel dat de simultane kansverdeling van X en Y gegeven wordt door de volgende tabel:

		X				
		-1	0	1	2	
Y	-1	?	0	1/8	?	?
	0	?	?	1/16	1/8	?
	1	1/8	?	?	?	1/2
		1/4	?	1/4	?	

(a) Vul de tabel aan als gegeven is dat $\mathrm{E}(X) = \mathrm{E}(Y) = \frac{1}{4}$.

(b) Bereken $P(X = -1|Y = 1)$.

(c) Zijn X en Y onafhankelijk?.

5. Laten X en Y de uitskomsten zijn van twee onafhankelijke worpen met een munt. We hebben dan, met 0 voor kruis en 1 voor munt,

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

Bekijk nu

$$U = X + Y \text{ en } V = |X - Y|.$$

(a) Bepaal de kansverdelingen van U en V.

(b) Bepaal de simultane kansverdeling van U en V.

(c) Bereken de covariantie van U en V.

(d) Zijn U en V onafhankelijk?

6. Bepaal de marginale kansdichtheden bij de volgende simultane kansdichtheden van X en Y. Controleer ook of X en Y onafhankelijk zijn.

(a)

$$f(x,y) = \begin{cases} \frac{1}{2} & , \text{ als } 0 < x < 2, 0 < y < 1, \\ 0 & , \text{ elders.} \end{cases} \tag{5.54}$$

(b)

$$f(x,y) = \begin{cases} xye^{-(x+y)} & , \text{ als } x > 0, y > 0, \\ 0 & , \text{ elders.} \end{cases}$$

(c)

$$f(x,y) = \begin{cases} \frac{3}{2}\, y^2 & , \text{ als } 0 \leq x \leq 2, 0 < y < 1, \\ 0 & , \text{ elders.} \end{cases}$$

(d)

$$f(x,y) = \begin{cases} 2e^{-(x+y)} & , \text{ als } 0 < x < y, y > 0, \\ 0 & , \text{ elders.} \end{cases}$$

7. Stel dat X en Y de functie (5.54) als simultane kansdichtheid hebben. Bereken de kansen

   (a) $P(X \geq Y)$.

   (b) $P(X < Y)$.

   (c) $P(Y > \frac{1}{2}\, X)$.

   (d) $P(Y \geq (X-1)^2)$.

8. De simultane kansdichtheid $f(x,y)$ van de stochasten X en Y wordt gegeven door

$$f(x,y) = \begin{cases} \frac{9}{16}\, x^2 y^2 & , \text{ als } -1 \leq x \leq 1, 0 \leq y \leq 2 \\ 0 & , \text{ elders.} \end{cases}$$

   (a) Bepaal de marginale kansdichtheden en verdelingsfuncties van X en Y.

   (b) Bereken $\mathrm{E}\,(X+Y)$.

   (c) Zijn X en Y onafhankelijk ?

9. De simultane kansdichtheid $f(x,y)$ van de stochasten X en Y wordt gegeven door

$$f(x,y) = \begin{cases} 3(1-\sqrt{y})\, e^{-x} & , \text{ als } 0 \leq y \leq 1,\ x \geq 0, \\ 0 & , \text{ elders.} \end{cases}$$

   (a) Bepaal de marginale kansdichtheden en verdelingsfuncties van X en Y.

   (b) Zijn X en Y onafhankelijk ?

   (c) Bereken $\mathrm{E}\,(XY)$.

10. Laat $X = (X_1, X_2)^T$ een stochastische 2-vector zijn met verwachtingsvector μ en covariantiematrix Σ gegeven door

$$\mu = \begin{pmatrix} 1 \\ -2 \end{pmatrix} \quad \text{en} \quad \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}.$$

    Bepaal de verwachtingsvector en covariantiematrix van

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix}.$$

    Kunnen we nu concluderen dat Z_1 en Z_2 onafhankelijk zijn ?

11. Laat $X = (X_1, X_2)^T$ een stochastische 2-vector zijn met verwachtingsvector $(0,0)^T$ en covariantiematrix Σ gegeven door

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & \sigma_{2,2} \end{pmatrix},$$

voor zekere variantie $\sigma_{2,2} \geq 0$. Voor welke waarde van $\sigma_{2,2}$ zijn de componenten van

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix}$$

ongecorreleerd. Kunnen we in dat geval ook concluderen dat Z_1 en Z_2 onafhankelijk zijn ?

12. Beschouw de matrix

$$\Sigma = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}.$$

(a) Kan dit een covariantiematrix zijn?
(Hint: bereken de variantie van $X_1 - X_2$.)

(b) Laat zien dat voor elke $p \times p$ covariantiematrix Σ geldt: $a^T \Sigma a \geq 0$ voor alle p-vectoren a , m.a.w. de matrix is *positief definiet*.
(Hint: bereken de variantie van $a_1 X_1 + \cdots + a_p X_p$.)

13. Laat $X = (X_1, X_2)^T$ een stochastische 2-vector zijn met verwachtingsvector μ en covariantiematrix Σ gegeven door

$$\mu = \begin{pmatrix} 1 \\ -2 \end{pmatrix} \quad \text{en} \quad \Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}.$$

Bepaal de verwachtingsvector en covariantiematrix van

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} 1 + 2X_1 - X_2 \\ 2 + X_1 + X_2 \end{pmatrix}.$$

Zijn Z_1 en Z_2 onafhankelijk ?

14. Laat X een $N_3(\mu, \Sigma)$ verdeelde stochastische vector zijn met

$$\mu = \begin{pmatrix} -3 \\ 1 \\ 4 \end{pmatrix} \quad \text{en} \quad \Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Welke van de volgende stochastische variabelen zijn onafhankelijk ?

(a) X_1 en X_2

(b) X_2 en X_3

(c) $\frac{1}{2}(X_1 + X_2)$ en X_3

(d) X_2 en $X_2 - \frac{5}{2}X_1 - X_3$

15. Laat X een $N_3(\mu, \Sigma)$ verdeelde stochastische vector zijn met

$$\mu = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} \quad \text{en} \quad \Sigma = \begin{pmatrix} 4 & 0 & -1 \\ 0 & 5 & 0 \\ -1 & 0 & 2 \end{pmatrix}.$$

Welke van de volgende stochastische variabelen zijn onafhankelijk ?

(a) X_1 en X_2

(b) X_1 en X_3

(c) X_2 en X_3

(d) X_2 en $X_1 + 3X_2 - 2X_3$

16. Laat X een $N_3(\mu, \Sigma)$ verdeelde stochastische vector zijn met

$$\mu = \begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix} \quad \text{en} \quad \Sigma = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{pmatrix}.$$

(a) Bepaal de kansverdeling van $3X_1 - 2X_2 + X_3$.

(b) Bepaal een 2-vector $a = (a_1, a_2)^T$ zodat X_2 en

$$X_2 - (a_1, a_2) \begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$$

onafhankelijk zijn.

Hoofdstuk 6

Schattingstheorie

We hebben ons tot nu toe bezig gehouden met kansrekening, het beschrijven van kansexperimenten. Als je conclusies wil trekken over aspecten van een kansexperiment op grond van uitkomsten van dat experiment, dan begeef je je op het terrein van de statistiek. Die conclusies kunnen van verschillende aard zijn. Zo kan je denken aan schattingstheorie, toetsingstheorie en classificatieproblemen. We beginnen in dit hoofdstuk met schattingstheorie.

Voorbeeld 6.1 (Robot)

In Voorbeeld 3.10 hebben we de kansverdeling van een stochastische variabele X, de tijd die het duurt totdat een robot een bepaalde taak heeft verricht, gemodelleerd. We hebben daar gekozen voor een exponentiële verdeling met een parameter λ. De kansdichtheid van X is dan gelijk aan

$$f(x) = \begin{cases} \frac{1}{\lambda}\, e^{-x/\lambda} & , \text{ als } x > 0, \\ 0 & , \text{ als } x \leq 0. \end{cases} \tag{6.1}$$

Normaliter is die parameter onbekend maar zou je hem wel willen weten om de prestatie van je robot te kunnen beschrijven. Om een idee te krijgen wat de waarde van de parameter λ is kan je de robot een aantal keren, zeg n keer, vanuit willekeurig gekozen startposities, zijn taak laten doen. Laten we zeggen dat dit n tijdsduren $x_1, \ldots, x_n$ oplevert. We zeggen dan dat je een steekproef van omvang n hebt genomen uit de verdeling van X. Later in dit hoofdstuk zullen we zien dat het gemiddelde van deze tijdsduren een goede schatting is van λ.

Om te kunnen beoordelen of deze procedure verstandig is hebben we een kansmodel nodig dat de steekproef en de kansverdeling van het daarop gebaseerde resultaat beschrijft. Bovendien moet je formuleren wat je goede procedures vind.

6.1 Steekproef

Stel dat we iets willen concluderen over een kansverdeling, beschreven door de verdelingsfunctie F, van een stochastische variabele X. We herhalen het experiment n keer op een dusdanige manier dat de experimenten elkaar niet beïnvloeden. Het resultaat is dan n uitkomsten $x_1, \ldots, x_n$.

We kunnen deze uitkomsten zien als een **realisatie**, een uitkomst, van een rij onafhankelijke stochastische variabelen $X_1, \ldots, X_n$, waar bij elke X_i dezelfde kansverdeling heeft als X. We spreken hier van een **steekproef van omvang** n (sample of size n) uit F. We zeggen ook wel dat $X_1, \ldots, X_n$ onafhankelijke, gelijkverdeelde (independent and identically distributed, i.i.d.) stochastische variabelen zijn met verdelingsfunctie F.

In de voorafgaande hoofdstukken hebben we voldoende kansrekening ingevoerd om de kansverdeling van een steekproefuitkomst te kunnen beschrijven. Als de kansverdeling van X discreet is dan kunnen we de kansverdeling van de steekproefuitkomst beschrijven door

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n) = P(X = x_1) \cdots P(X = x_n), \quad (6.2)$$

voor alle mogelijke realisaties $x_1, \ldots, x_n$.

Als de kansverdeling van X continu is, en X een kansdichtheid f heeft, dan kunnen we de simultane kansdichtheid $f(x_1, \ldots, x_n)$ van de steekproefuitkomst beschrijven door

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) = f(x_1) \cdots f(x_n), \quad (6.3)$$

voor alle mogelijke realisaties $x_1, \ldots, x_n$.

We hebben nu dus een wiskundig model dat de variatie van een steekproefuitkomst beschrijft. De volgende stap is het beschrijven van wenselijke eigenschappen van (de kansverdeling van uitkomsten van) procedures die op de steekproef gebaseerd zijn.

6.2 Schatters

We kunnen een (parametrisch) schattingsprobleem als volgt beschrijven. Stel we hebben genoeg inzicht in de aard van de waarnemingen dat we kunnen zeggen dat de kansverdeling van X er een is uit een parametrische familie, maar dat we niet weten welke. Die parametrische familie zullen we beschrijven in termen van een parameter θ. De verdelingsfunctie van de verdelingen in die familie schrijven we als F_θ. We weten dus dat X een van die kansverdelingen heeft, maar we weten niet welke θ daarbij hoort.

Als we, zoals nu, met algemene theorie bezig zijn, noemen we de parameter θ. Bij specifieke schattingsproblemen hebben de parameters meestal een standaard (meestal Griekse) letter om ze te beschrijven. Dat is, zoals in het voorbeeld hierboven, voor de exponentiële verdeling een λ.

Het doel van een schattingsprocedure is om de onbekende waarde van θ, of van een functie van θ, te benaderen, te schatten, op grond van de realisatie van een steekproef van omvang n uit F_θ. Daartoe moeten we eerst beschrijven wat een schatter is.

Definitie 6.2 *Een* **schatter** *(estimator) T is een functie van alleen de waarnemingen $X_1, \ldots, X_n$ van de steekproef. Dus*

$$T = t(X_1, \ldots, X_n). \quad (6.4)$$

Een **schatting** *(estimate) is een waarde, $t(x_1, \ldots, x_n)$, van de schatter gebaseerd op een realisatie $x_1, \ldots, x_n$ van de steekproef.*

Een schatter is dus een stochastische variabele die de uitkomst van de schattingsprocedure beschrijft. Een schatting is de uitkomst van die procedure op grond van een specifieke uitkomst van de steekproef, het getal dat je hebt uitgerekend op grond van jouw steekproef. Wat nadrukkelijk niet mag is dat je de θ gebruikt in je schatter want die ken je immers niet.

Voorbeeld 6.3 (Robot)

Voor onze robot nemen we het steekproefgemiddelde als schatter van λ. Hier hebben we dus

$$T = \frac{1}{n}\left(X_1 + \ldots + X_n\right) = t(X_1, \ldots, X_n), \tag{6.5}$$

waarbij $t(x_1, \ldots, x_n)$ gelijk is aan $\frac{1}{n}\left(x_1 + \ldots + x_n\right)$.

We weten nu dus wat schatters zijn maar op zich schieten we daar niet zoveel mee op. Er zijn slechte en goede schatters. Neem bijvoorbeeld de schatter $T = 1$. Die geeft altijd de waarde een als schatting, ongeacht de steekproefuitkomst. Dat is mooi als de echte waarde gelijk is aan een, maar heel fout als de echte waarde niet gelijk is aan een. In het voorbeeld van het schatten van λ op basis van de gemeten tijdsduren zouden we ook de schatter $T = X_1$ kunnen nemen. Je voelt hier echter wel aan dat je dan informatie weggooit als je de andere tijdsduren niet gebruikt. We moeten dus formuleren wat we goede schatters vinden.

6.3 Zuiverheid en variantie

Stel we willen $g(\theta)$ schatten, voor een functie g. Die functie kan de indentieke functie zijn, $g(\theta) = \theta$, of bijvoorbeeld het kwadraat, $g(\theta) = \theta^2$. Het kan ook een kans zijn. In het robot voorbeeld wil je misschien de kans schatten dat de robot minder dan tien minuten nodig heeft. Die kans is gelijk aan $1 - e^{-10/\lambda}$, zie Voorbeeld 3.10. Als we in dit voorbeeld λ zelf schatten, dan schatten we in feite $\mathrm{E}\left(X\right)$, de verwachting van de tijdsduur, want die is gelijk aan λ, zie (4.24).

Eigenlijk willen we dat de schatter gemiddeld de echte waarde van $g(\theta)$ geeft als we het schattingsexperiment een groot aantal keren herhalen. Dit leidt tot de volgende definitie.

Definitie 6.4 *We noemen een schatter T* **zuiver** *(unbiased) als zijn verwachting gelijk is aan $g(\theta)$. Dus*

$$\mathrm{E}\left(T\right) = g(\theta), \tag{6.6}$$

voor alle mogelijke waarden van θ.

Voor een zuivere schatter geldt dus dat hij, voor alle mogelijke waarden van θ, en dus ook voor de echte waarde, gemiddeld de goede waarde $g(\theta)$ zal geven bij herhalingen van het experiment. Hij heeft dus geen systematische fout.

Voorbeeld 6.5 (Robot)

We kunnen controleren of de schatter in het robot voorbeeld, het steekproefgemiddelde, zuiver is. Berekenen we de verwachting van T dan vinden we

$$\mathrm{E}\,(T) = \mathrm{E}\left(\frac{1}{n}\,(X_1 + \ldots + X_n)\right) = \frac{1}{n}\,(\mathrm{E}\,(X_1) + \ldots + \mathrm{E}\,(X_n)) = \frac{1}{n}\,(\lambda + \ldots + \lambda) = \lambda. \quad (6.7)$$

Deze schatter is dus zuiver.

Opmerking 6.6 *Als je de berekening (6.7) nog eens bekijkt dan zie je het steekproefgemiddelde altijd een zuivere schatter is voor de verwachting van X, dus van de verwachting van de waarnemingen. Er bestaat ook een zuivere schatter voor de variantie van de waarnemingen.*

Stel $X_1, \ldots, X_n$ is een steekproef uit een verdeling met verwachting m en variantie d^2. Dus we hebben $\mathrm{E}\,(X_i) = m$ en $\mathrm{Var}\,(X_i) = d^2$, voor $i = 1, \ldots, n$. Voor het steekproefgemiddelde $\bar{X}_n$ en de steekproefvariantie S_n^2, gegeven door

$$\bar{X}_n = \frac{1}{n}\,(X_1 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (6.8)$$

en

$$S_n^2 = \frac{1}{n-1}\Big((X_1 - \bar{X}_n)^2 + \ldots + (X_n - \bar{X}_n)^2\Big) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad (6.9)$$

geldt

$$\mathrm{E}\,(\bar{X}_n) = m \quad \textit{en} \quad \mathrm{E}\,(S_n^2) = d^2. \quad (6.10)$$

Het bewijs van de zuiverheid van S_n^2 slaan we over. Vaak zie je S_n^2 met een factor $1/n$ in plaats van $1/(n-1)$. Die schatter is dus niet zuiver maar voor grote steekproefomvang n is het verschil minimaal.

In Voorbeeld 6.5 geldt echter ook dat de verwachting van de eerste tijdsduur gelijk is aan λ. Met $T = X_1$ hebben we ook $\mathrm{E}\,(T) = \mathrm{E}\,(X_1) = \lambda$. Deze schatter is dus ook zuiver. Toch geven we naar ons gevoel er de voorkeur aan alle waarnemingen te gebruiken. Dat is omdat we denken dat we dan een nauwkeurigere schatting krijgen. De nauwkeurigheid kunnen we nagaan door de variantie te berekenen van deze twee schatters. We vinden dan

$$\mathrm{Var}\,(X_1) = \lambda^2,$$

en

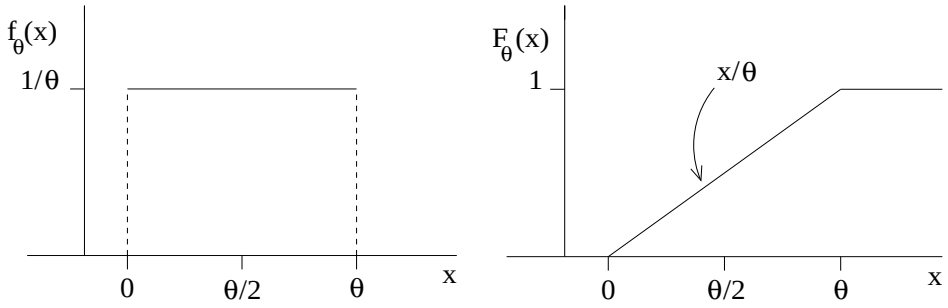
$$\mathrm{Var}\left(\frac{1}{n}\,(X_1 + \ldots + X_n)\right) = \frac{1}{n^2}(\mathrm{Var}\,(X_1) + \ldots + \mathrm{Var}\,(X_n)) = \frac{1}{n^2}(\lambda^2 + \ldots + \lambda^2) = \frac{\lambda^2}{n}.$$

De variantie van het steekproefgemiddelde is dus, als er meer dan een waarneming is, altijd kleiner dan de variantie van een enkele waarneming. Het gemiddelde is dus beter. Bovendien

gaat de variantie naar nul als de steekproefomvang naar oneindig gaat. Als je oneindig veel waarnemingen zou hebben zou je de echte λ dus zeker weten.

We besluiten deze paragraaf met een schattingsprobleem waar je twee voor de hand liggende schatters kan bekijken, die beiden alle waarnemingen gebruiken, waarvan er een duidelijk de voorkeur heeft.

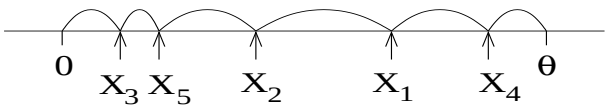
Voorbeeld 6.7 (Schatten van een drempelwaarde)



Figuur 6.1: De kansdichtheid en verdelingsfunctie van de uniforme verdeling op $[0, \theta]$.

Stel je hebt waarnemingen $X_1, \ldots, X_n$ uit de uniforme verdeling op het interval $[0, \theta], \theta > 0$. Je wil θ schatten. Er liggen twee schatters voor de hand. In (4.16) hebben we gezien dat de verwachting van X_i gelijk is aan $\theta/2$. Het steekproefgemiddelde $\bar{X}_n$ heeft dan dus ook verwachting $\theta/2$, en twee maal het gemiddelde heeft verwachting θ. Hieruit volgt dat $T_1 = 2\bar{X}_n$ een zuivere schatter van θ is,

$$\mathrm{E}\,(T_1) = \mathrm{E}\,(2\bar{X}_n) = 2\mathrm{E}\,(\bar{X}_n) = 2\,\frac{\theta}{2} = \theta. \tag{6.11}$$



Figuur 6.2: De tussenafstanden bij vijf waarnemingen.

We kunnen het probleem ook van een andere kant bekijken. Als we kijken naar de tussenafstanden tussen de waarnemingen, inclusief nul en het eindpunt θ, dan zijn dat er $n + 1$. De laatste afstand, tussen de maximale X_i en θ, is dus betrekkelijk klein. Dat kan je het idee geven om een een schatter te maken met die maximale waarneming. Laten we die M_n noemen, dus

$$M_n = \max_{i=1,\ldots,n} X_i. \tag{6.12}$$

Nu kunnen we M_n zelf meteen afvoeren als schatter van θ want M_n is altijd kleiner dan θ. Hij is dus duidelijk niet zuiver. Als we hem een zetje naar rechts zouden kunnen geven dan

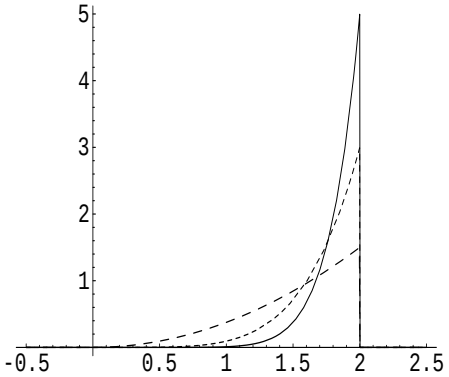
kunnen we hem misschien zuiver maken. Daartoe zullen we eerst zijn kansverdeling bepalen. We noteren de kansdichtheid van M_n door middel van f_{M_n} en zijn verdelingsfunctie door middel van F_{M_n}. Van Paragraaf 4.2.1 weten we dat de verdelingsfunctie F_θ van de X_i gelijk is aan $F_\theta(x) = x/\theta$, voor $0 \leq x \leq \theta$. We gaan nu als volgt te werk. Neem een x met $0 \leq x \leq \theta$. We gebruiken de observatie dat de gebeurtenis 'het maximum van de X_i is kleiner dan x' gelijk is aan de gebeurtenis 'alle X_i zijn kleiner dan x'. We hebben nu, vanwege de onafhankelijkheid van de waarnemingen,

$$\begin{aligned} F_{M_n}(x) &= P(M_n \leq x) = P(\max_{i=1,\ldots,n} X_i \leq x) \\ &= P(X_1 \leq x, \ldots, X_n \leq x) = P(X_1 \leq x) \cdots P(X_n \leq x) \\ &= P(X_1 \leq x) \cdots P(X_n \leq x) \\ &= \frac{x}{\theta} \cdots \frac{x}{\theta} = \frac{1}{\theta^n}\, x^n. \end{aligned}$$

De kansdichtheid van M_n krijgen we nu door de verdelingsfunctie te differentiëren. Dus

$$f_{M_n} = \begin{cases} n\, \frac{1}{\theta^n}\, x^{n-1} & , \text{ als } 0 \leq x \leq \theta, \\ 0 & , \text{ elders.} \end{cases} \tag{6.13}$$

Om de zuiverheid van M_n te onderzoeken moeten we zijn verwachting berekenen. We vinden



Figuur 6.3: Kansdichtheden van M_n voor $\theta = 2$ en $n = 3, 6, 10$.

dan

$$\begin{aligned} \mathrm{E}\,(M_n) &= \int_{-\infty}^{\infty} x f_{M_n}(x) dx = \int_0^\theta x\, n\, \frac{1}{\theta^n}\, x^{n-1} dx = n\, \frac{1}{\theta^n} \int_0^\theta x^n dx \\ &= n\, \frac{1}{\theta^n}\, \frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\, \theta. \end{aligned}$$

De verwachting van M_n zelf is dus altijd kleiner dan θ zoals we al aan voelden komen. Echter als we M_n vermenigvuldigen met een factor $(n+1)/n$ dan komt het goed. Immers, voor

$$T_2 = \frac{n+1}{n} M_n \tag{6.14}$$

vinden we

$$\mathrm{E}\,(T_2) = \mathrm{E}\left(\frac{n+1}{n} M_n\right) = \frac{n+1}{n}\,\mathrm{E}\,(M_n) = \frac{n+1}{n}\,\frac{n}{n+1}\,\theta = \theta. \tag{6.15}$$

Zowel T_1 als T_2 zijn dus zuivere schatters van θ.

Laten we nu eens naar hun nauwkeurigheid kijken en de varianties uitrekenen. Voor de variantie van T_1 vinden we

$$\mathrm{Var}\,(T_1) = \mathrm{Var}\left(2\bar{X}_n\right) = 4\mathrm{Var}\,(\bar{X}_n) = 4\,\frac{1}{n}\,\mathrm{Var}\,(X_1) = 4\,\frac{1}{n}\,\frac{\theta^2}{12} = \frac{1}{3n}\,\theta^2. \tag{6.16}$$

Hierbij hebben we de variantie van de uniforme verdeling gebruikt die in (4.17) gegeven wordt.

Met een berekening als voor de verwachting van M_n kunnen we ook de verwachting van M_n^2 berekenen. We vinden dan

$$\mathrm{E}\,(M_n^2) = \frac{n}{n+2}\,\theta^2. \tag{6.17}$$

De variantie van M_n is dan gelijk aan

$$\mathrm{Var}\,(M_n) = \mathrm{E}\,(M_n^2) - (\mathrm{E}\,(M_n))^2 = \frac{n}{n+2}\,\theta^2 - \left(\frac{n}{n+1}\,\theta\right)^2 = \frac{n}{(n+1)^2(n+2)}\,\theta^2. \tag{6.18}$$

Hieruit volgt dat de variantie van T_2 gelijk is aan

$$\begin{aligned}\mathrm{Var}\,(T_2) &= \mathrm{Var}\left(\frac{n+1}{n} M_n\right)) = \left(\frac{n+1}{n}\right)^2 \mathrm{Var}\,(M_n) = \left(\frac{n+1}{n}\right)^2 \frac{n}{(n+1)^2(n+2)}\,\theta^2 \\ &= \frac{1}{n(n+2)}\,\theta^2.\end{aligned}$$

Als we de varianties van T_1 en T_2 vergelijken dan zien we

$$\mathrm{Var}\,(T_1) = \frac{1}{3n}\,\theta^2 \geq \frac{1}{n(n+2)}\,\theta^2 = \mathrm{Var}\,(T_2),$$

omdat $3n \leq n(n+2)$. Immers $n+2 \geq 3$. De schatter T_2, die gebaseerd is op het maximum, is dus beter dat de schatter T_1, gebaseerd op het gemiddelde. De variantie van T_2 gaat ook nog eens sneller naar nul dan die van T_2.

6.4 Betrouwbaarheidsintervallen voor de verwachting van een normale verdeling

Stel we kunnen een experiment n keer herhalen op zo'n manier dat de uitkomsten elkaar niet beïnvloeden, en dat ze dezelfde verdeling hebben. We kunnen die uitkomsten modelleren door middel van een steekproef $X_1, \ldots, X_n$. We nemen aan dat we kunnen veronderstellen dat de gemeenschappelijke verdeling van die uitkomsten een normale verdeling is met parameters μ, de verwachting, en σ^2, de variantie. Dus de stochastische variabelen $X_1, \ldots, X_n$ zijn onafhankelijk en ze hebben een $\mathcal{N}(\mu, \sigma^2)$ verdeling.

We zijn geïnteresseerd in de verwachting μ in situaties waar σ^2 bekend is, of, wat realistischer is, in situaties waar σ^2 niet bekend is. In beide gevallen is het steekproefgemiddelde

$$\bar{X}_n = \tfrac{1}{n}(X_1 + \ldots + X_n)$$

een goede schatter van μ. Het steekproefgemiddelde is een zuivere schatter van μ. Immers

$$\mathrm{E}\,\bar{X}_n = \mathrm{E}\,\tfrac{1}{n}(X_1 + \ldots + X_n) = \tfrac{1}{n}(\mathrm{E}\,X_1 + \ldots + \mathrm{E}\,X_n) = \tfrac{1}{n} n\mu = \mu.$$

Als we dit gemiddelde van de waarnemingen uitrekenen krijgen we alleen een getal, de schatting. We hebben dan geen idee over de nauwkeurigheid van de schatting. Als het aantal waarnemingen groot is zal de schatting nauwkeuriger zijn dan bij een relatief klein aantal waarnemingen. We willen dus een idee hebben over de nauwkeurigheid van de schatting. Daarvoor gebruiken we een **betrouwbaarheidsinterval** (confidence interval). In deze sectie doen we dat voor het voorbeeld van het schatten van μ bij normaal verdeelde waarnemingen. Bij andere schattingsproblemen kan je in het algemeen ook betrouwbaarheidsintervallen maken.

6.4.1 De variantie is bekend

Laten we eerst veronderstellen dat we de waarde van σ^2, de variantie van de waarnemingen, kennen, en dat μ, de verwachting, de enige onbekende parameter is. We gaan een betrouwbaarheidsinterval voor μ afleiden. We noteren $\bar{x}_n$ voor $\frac{1}{n}(x_1 + \ldots + x_n)$, de realisatie van het steekproefgemiddelde. Bekijk nu de bewering

$$\bar{x}_n - \tfrac{1.96}{\sqrt{n}}\,\sigma \le \mu \le \bar{x}_n + \tfrac{1.96}{\sqrt{n}}\,\sigma. \tag{6.19}$$

Deze bewering is waar of niet waar. Meer kan je er niet over zeggen. Als we de bewering bekijken in de context van ons kansmodel voor de methode van dit schattingsprobleem, en niet het resultaat, dan hebben we

$$\bar{X}_n - \tfrac{1.96}{\sqrt{n}}\,\sigma \le \mu \le \bar{X}_n + \tfrac{1.96}{\sqrt{n}}\,\sigma. \tag{6.20}$$

Omdat het steekproefgemiddelde hier een stochastische variabele is kan deze bewering nu waar of niet waar zijn. We kunnen bovendien de kans uitrekenen dat de bewering waar is. Dan krijgen we

$$\begin{aligned}
&P(\text{de bewering is waar})\\
&= P(\bar{X}_n - \tfrac{1.96}{\sqrt{n}}\,\sigma \le \mu \le \bar{X}_n + \tfrac{1.96}{\sqrt{n}}\,\sigma)\\
&= P(-\tfrac{1.96}{\sqrt{n}}\,\sigma \le \bar{X}_n - \mu \le +\tfrac{1.96}{\sqrt{n}}\,\sigma)\\
&= P(-1.96 \le \tfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le 1.96)\\
&= P(-1.96 \le U \le 1.96) = P(U \le 1.96) - P(U < -1.96) \approx 0.975 - 0.025 = 0.950,
\end{aligned}$$

waarbij U een standaardnormaal verdeelde stochastische variabele is. De kansen in de laatste regel hierboven kan je vinden in tabellen voor de standaardnormale verdeling. Je kan ze ook berekenen met procedures in statistische computerpakketten.

We zien nu dat de bewering (6.20) waar is met kans 0.95. Het (stochastische) interval

$$[\bar{X}_n - \frac{1.96}{\sqrt{n}}\,\sigma\,, \bar{X}_n + \frac{1.96}{\sqrt{n}}\,\sigma] \tag{6.21}$$

wordt een 95% betrouwbaarheidsinterval voor μ genoemd. De bewering (6.20) geeft een idee over de nauwkeurigheid van de schatting.

6.4.2 De variantie is niet bekend

Men kan het betrouwbaarheidsinterval (6.21) natuurlijk niet gebruiken als σ niet bekend is. Dit komt in de praktijk vaak voor. We volgen dan een soortgelijke procedure waarbij we σ door een schatting van σ vervangen. Eerder in dit hoofdstuk hebben we gezien dat de steekproefvariantie

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

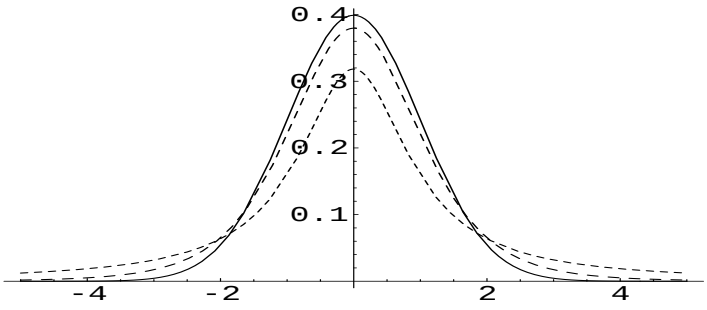
een zuivere schatter is voor $\mathrm{Var}\,(X_1) = \sigma^2$.

William Sealy Gossett, hoofdbrouwer van de Guinness brouwerij, die een opleiding had gehad in scheikunde en wiskunde, heeft in een artikel in 1908 voorgesteld σ in (6.20) te vervangen door

$$S_n = \sqrt{S_n^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2},$$

en om de factor 1.96, die gebasseerd is op de standaardnormale verdeling, in (6.20), te vervangen door een positieve factor c_n die we hieronder zullen bepalen. Dit resulteerde in het volgende betrouwbaarheidsinterval

$$\bar{X}_n - \frac{c_n}{\sqrt{n}}\,S_n \le \mu \le \bar{X}_n + \frac{c_n}{\sqrt{n}}\,S_n. \tag{6.22}$$



Figuur 6.4: *Kansdichtheden van de Student t-verdelingen voor $n = 2$ en $n = 6$ (de gearceerde lijnen), en de standaard normale kansdichtheid (de continue lijn).*

Gossett publiceerde zijn artikelen onder de naam Student. De reden daarvoor was waarschijnlijk dat een andere onderzoeker niet lang daarvoor bedrijfsgeheimen had gepubliceerd samen met zijn wetenschappelijke resultaten. Guiness verboodt daarom dergelijke publicaties.

We willen nu c_n zó kiezen dat de kans dat de bewering (6.22) waar is, gelijk is aan 0.95. In de vorige paragraaf hebben we gebruikt dat de stochastische variabele $U = (\bar{X}_n - \mu)/(\sigma/\sqrt{n})$ een standaardnormale verdeling heeft. Hier, daarentegen, heeft de stochastische variabele

$$T_n = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} = \sqrt{n}\,\frac{\bar{X}_n - \mu}{S_n}, \tag{6.23}$$

hoewel hij een symmetrische verdeling heeft met verwachting nul, geen standaardnormale verdeling. Expliciete uitdrukkingen voor zijn kansdichtheid zijn bekend. De verdeling van T_n wordt Student t-verdeling (met $n-1$ vrijheidsgraden) genoemd. Deze verdeling hangt af van n.

Het feit dat de verdeling verschilt van de standaardnormale verdeling zien we in Figuur 6.4. Voor grote n vervaagt het verschil. Het verschil zien we ook in de variantie van T_n. Voor $n \geq 4$ kan je afleiden

$$\mathrm{Var}\,(T_n) = \frac{n-1}{n-3}.$$

Dit laat ook zien dat deze verdeling meer uitgespreid is dan de standaardnormale verdeling waarvan de variantie gelijk is aan een. Hieruit volgt min of meer dat we waarden van c_n moeten kiezen die groter zijn dan twee. In Tabel 6.1 staan een aantal van die waarden. Als je dergelijke

n	c_n
2	12.7
3	4.3
6	2.6
11	2.23
∞ (standaard normaal)	1.96

Tabel 6.1: *Waarden van c_n voor een aantal steekpoefomvangen n.*

intervallen wil gebruiken voor andere steekproefomvang, of voor andere betrouwbaarheidskans, niet 0.95, dan kan je de waarden van c_n vinden in standaardtabellen. Of je berekend ze met standaardprocedures.

6.5 Bayesiaanse statistiek

In de frequentistische schattingstheorie zoals we die hier beschreven hebben is de parameter θ een onbekend getal, of een onbekend aantal getallen, een vector. We hebben een steekproef van waarnemingen $X_1, \ldots, X_n$ die allemaal dezelfde verdeling uit een parametrische familie van

verdelingen hebben. Namelijk de verdeling die aangegeven wordt door de parameter waarde θ. Het resultaat van de schattingsprocedure, de schatting, is dan een getal, of een vector.

Als je statistiek bedrijft vanuit het gezichtspunt van de Bayesiaanse statistiek dan is dit anders. Je redeneert dan als volgt. Je weet eigelijk altijd wel al iets over de parameter, wellicht uit eerder onderzoek. Die kennis kan je samenvatten door middel van een kansverdeling van de parameter θ, de **a priori verdeling** (prior distribution). Dat is dus heel anders dan we eerder hebben gedaan. Daar is de parameter een vast, onbekend, getal of vector. Vervolgens zegt men dat de waarneming X, gegeven de waarde van de parameter θ, een verdeling uit een parametrische familie heeft die aangegeven wordt door θ. De waarneming kan de hele steekproef zijn, maar bijvoorbeeld ook het aantal successen in een steekproef zoals in het voorbeeld beneden. Stel dat het een continu verdeelde waarneming is en dat de kansdichtheid van de waarneming, gegeven θ, gelijk is aan $f(x|\theta)$. Waar we in geïnteresseerd zijn in de Bayesiaanse benadering is de verdeling van θ, gegeven de waarneming X. Dit is de zogenaamde **a posteriori verdeling** (posterior distribution) van θ.

In de Bayesiaanse filosofie verandert de waarneming dus de a priori kennis in a posteriori kennis. We kunnen de a posteriori verdeling als volgt uitrekenen. Door middel van een continue versie van de regel van Bayes, zie (2.8), kan de a posteriori kansdichtheid geschreven worden als

$$f(\theta|x) = \frac{f(x,\theta)}{f(x)} = \frac{f(x|\theta)p(\theta)}{f(x)}. \tag{6.24}$$

Hierbij is $p(\theta)$ de a priori kansdichtheid van θ, $f(x,\theta)$ de simultane kansdichtheid van X en θ en $f(x)$ de marginale dichtheid van X. Dit is de belangrijke stap in een Bayesiaanse schattingsprocedure. Het verklaart ook de naam. De a priori kennis wordt aangepast aan de waarneming. Dit levert de a posteriori kennis.

Voorbeeld 6.8 (Bayesiaans schatten van een kans)

Stel je moet een nieuwe methode testen. De methode kan werken of niet werken. De methode wordt getest in een aantal, zeg n, onafhankelijke situaties. Dit levert een uitkomst X op van het totaal aantal keren dat de methode heeft gewerkt. Als θ gelijk is aan de kans dat de methode werkt dan is X $\text{Bin}(n,\theta)$ verdeeld.

We vatten onze a priori kennis over θ nu samen in een kansdichtheidsfunctie $p(\theta)$. Dit moet dus een kansdichtheid op $[0,1]$ zijn want θ is een kans. De Beta verdelingen, zie paragraaf 4.2.2, zijn hier heel geschikt voor. De parameter θ wordt dus behandeld als een stochastische variabele met kansdichtheid $p(\theta)$. Gegeven θ heeft de uitkomst X een $\text{Bin}(n,\theta)$ verdeling,

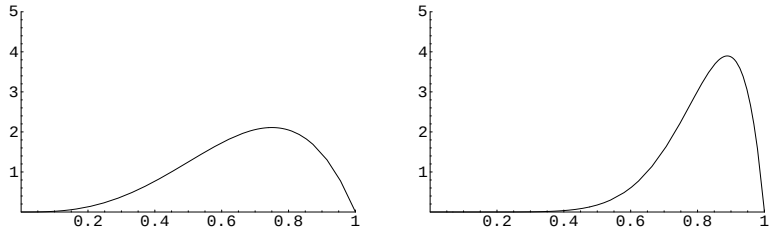
$$P(X = x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x = 0,1,\ldots,n, \tag{6.25}$$

Het blijkt dat de a posteriori verdeling in dit model ook weer een Beta verdeling is, als de a priori verdeling een Beta verdeling is. Alleen de parameters veranderen. Als de a priori verdeling voor θ een $\text{Beta}(a,b)$ verdeling is en de waarneming is $\text{Bin}(n,\theta)$ verdeeld, dan is de

a posteriori verdeling van θ, gegeven de waarneming, Beta$(x+a, n-x+b)$. De a postriori kansdichtheid voor een waarneming x is dus gelijk aan

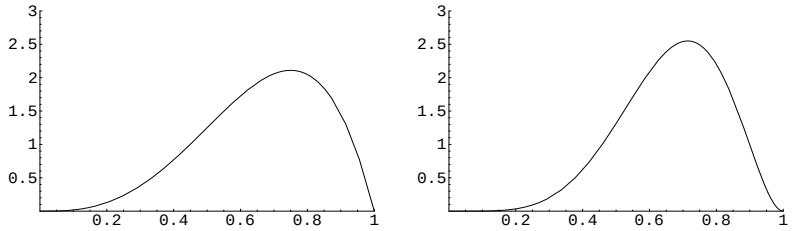
$$f(\theta|x) = \frac{\theta^{x+a-1}(1-\theta)^{n-x+b-1}}{B(x+a, n-x+b)}, \quad 0 < \theta < 1.$$

We illustreren dit met twee voorbeelden.



Figuur 6.5: $n = 5, x = 5$, a priori Beta dichtheid met $a = 4, b = 2$ (links) en a posteriori Beta dichtheid $a = 9, b = 2$ (rechts) .

In Figuur 6.5 hebben we vijf waarnemingen die allemaal een succes zijn. De nieuwe methode heeft in alle gevallen gewerkt. De a priori verdeling schuift dus flink naar rechts.



Figuur 6.6: $n = 5, x = 3$, a priori Beta dichtheid met $a = 4, b = 2$ (links) en a posteriori Beta dichtheid $a = 7, b = 4$ (rechts) .

In Figuur 6.6 hebben we vijf waarnemingen waarvan er drie een succes zijn. De nieuwe methode heeft in drie gevallen gewerkt. De a priori verdeling schuift dan een stuk minder naar rechts.

6.6 Opgaven

1. We hebben het schattingsprobleem als volgt omschreven. Je hebt een steekproef van waarnemingen $X_1, \ldots, X_n$ die allemaal dezelfde verdeling hebben. Die verdeling is er een uit een parametrische familie. De parameter bij deze familie noemen we in het algemeen θ. Voor specifieke families wordt de parameter vaak weergegeven door een andere letter. Je wilt een waarde $g(\theta)$, voor een zekere functie g, schatten. Noem de parameter θ en bepaal in de volgende schattingsproblemen de waarde $g(\theta)$ die je wil schatten.

   (a) Stel $X_i \sim \text{Bern}(p)$, $0 \leq p \leq 1$. Schat $\text{Var}\,(X_i)$.

   (b) Stel $X_i \sim \text{Bin}(10, p)$, $0 \leq p \leq 1$. Schat $P(X_i = 0)$.

   (c) Stel $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $-\infty < \mu < \infty, \sigma > 0$. Schat $\text{E}\,(X_i)$.

2. Stel T_1 en T_2 zijn zuivere schatters van θ. Voor welke waarden van twee constanten a en b is de gecombineerde schatter $T = aT_1 + bT_2$ opnieuw een zuivere schatter van θ. Als je twee zuivere schatters hebt dan heb je er dus meteen ook oneindig veel.

3. Laat X een trekking zijn uit een uniforme verdeling op $[0, \theta]$. Bepaal een zuivere schatter van $g(\theta) = \theta^2$.
   Hint: Bereken $\text{E}\,(X^2)$.

4. Laat $X_1, \ldots, X_n$ een steekproef zijn uit de uniforme verdeling op $[-\theta, \theta]$. Laat zien dat

   $$T = \frac{5}{n}\left(X_1^4 + \ldots + X_n^4\right)$$

   een zuivere schatter is van θ^4.

5. In het voorbeeld van de robot die een taak moet uitvoeren hebben we de tijdsduur gemodelleert door middel van een exponentiële verdeling met parameter $\lambda > 0$. Stel we hebben een steekproef van tijdsduren $X_1, \ldots, X_n$ gemeten. We hebben gezien dat $\bar{X}_n$ een zuivere schatter is van λ. Stel we willen $g(\lambda) = P(X \leq 10) = 1 - e^{-10/\lambda}$, de kans dat de robot hoogstens tien minuten nodig heeft om zijn taak uit te voeren, schatten. Dan ligt het voor de hand om de schatter $T = 1 - e^{-10/\bar{X}_n}$ te gebruiken. Dit is helaas geen zuivere schatter van $g(\lambda)$. Laat zien dat

   $$T = \frac{\text{aantal } X_i \leq 10}{n}$$

   wel een zuivere schatter is van $g(\lambda)$.

6. Men beschikt over een dataset die men opvat als een realisatie van een rij onafhankelijke stochasten $X_1, ..., X_n$. Men veronderstelt dat elke X_i een verdeling heeft met kansdichtheidsfunctie

   $$f(x) = \begin{cases} 0 & , \text{ als } x < 0, \\ \frac{3}{\theta^3}(\theta - x)^2 & , \text{ als } 0 \leq x \leq \theta, \\ 0 & , \text{ als } x > \theta. \end{cases}$$

   waarbij $\theta > 0$.

(a) Bereken de verwachting en de verdelingsfunctie van X_1.

(b) Voor welke a en b is de schatter

$$T_1 = a(X_1 + \cdots + X_n) + b$$

een zuivere schatter van θ.

(c) Maak een zuivere schatter van θ^2.

7. Men beschikt over een dataset die men opvat als een realisatie van een rij onafhankelijke stochasten $X_1, ..., X_n$. Men veronderstelt dat elke X_i een verdeling heeft met kansdichtheidsfunctie

$$f(x) = \begin{cases} 0 & , \text{ als } x < -\theta, \\ \frac{3}{2}\frac{x^2}{\theta^3} & , \text{ als } -\theta \leq x \leq \theta, \\ 0 & , \text{ als } x > \theta, \end{cases}$$

waarbij $\theta > 0$.

(a) Bereken de verdelingsfunctie en de variantie van X_1.

(b) Voor welke a en b is de schatter

$$T = a(X_1^2 + \cdots + X_n^2) + b$$

een zuivere schatter van θ^2 ?

(c) Bereken de verwachting van $|X_1|$ en maak een zuivere schatter van θ.

8. Stel je hebt 16 waarnemingen met een normale verdeling. (Deze waarnemingen zijn gegenereerd uit een $\mathcal{N}(1, 4)$ verdeling). De waarnemingen zijn.

1.21101, 2.60702, 2.26202, 1.20541, 1.00957, -0.190817, 5.12068, 0.0961735, 2.27243, 4.15292, 2.44197, 3.35327, 1.59089, 0.111128, -2.12492, 1.69269.

Het gemiddelde van deze waarnemingen is gelijk aan 1.67572.

(a) Stel je weet dat de variantie van de normale verdeling gelijk is aan $\sigma^2 = 4$. Schat μ en bepaal het 95% betrouwbaarheidsinterval voor μ.

(b) Stel je weet niet wat de variantie is. Schat μ en geef een 95% betrouwbaarheidsinterval voor μ. Hierbij mag je gebruiken dat S_n^2, de steekproefvariantie, voor deze steekproef gelijk is aan 3.11553 en dat de waarde c_{16} die hoort bij de Student verdeling met 15 vrijheidsgraden gelijk is aan 2.131.

(Antwoorden: a) [0.70,2.66], b) [0.74,2.62])

9. Stel je hebt 25 waarnemingen met een normale verdeling. (Deze waarnemingen zijn gegenereerd uit een $\mathcal{N}(1, 4)$ verdeling). De waarnemingen zijn.

2.26671, 1.20015, -2.04985, -1.03673, 4.61856, -0.873123, -3.78571, -0.0470981, 0.295566, 0.138019, -0.655424, 0.539686, 2.25125, 2.27426, 2.3197, 1.24384, 3.57203, -0.419302, 1.34173, -1.40434, 2.13266, 0.164572, 2.06112, 0.191722, 1.36672.

Het gemiddelde van deze waarnemingen is gelijk aan 0.708268.

(a) Stel je weet dat de variantie van de normale verdeling gelijk is aan $\sigma^2 = 4$. Schat μ en bepaal het 95% betrouwbaarheidsinterval voor μ.

(b) Stel je weet niet wat de variantie is. Schat μ en geef een 95% betrouwbaarheidsinterval voor μ. Hierbij mag je gebruiken dat S_n^2, de steekproefvariantie, voor deze steekproef gelijk is aan 3.3822 en dat de waarde c_{25} die hoort bij de Student verdeling met 24 vrijheidsgraden gelijk is aan 2.064.

(Antwoorden: a) [-0.08,1.49], b) [-0.05,1.47])

Hoofdstuk 7

Toetsingstheorie

Bij het schatten van een parameterwaarde, of een functiewaarde daarvan, willen we die waarde zo precies mogelijk benaderen op grond van de waarnemingen. Bij het toetsen hebben we een ander doel voor ogen. Stel eens dat iemand iets beweert over de parameterwaarde. Er wordt bijvoorbeeld beweerd ' de parameterwaarde is gelijk aan tien'. De waarde zelf interesseert je dan in eerste instantie niet zo. Je zou echter op grond van de waarnemingen wel willen controleren of je die bewering gelooft of niet. Dit is een ander probleem dan het schatten van de waarde van de parameter.

We beschrijven eerst de algemene theorie en de de relevante begrippen van de toetsingstheorie. Daarna behandelen we twee belangrijke toetsingsproblemen met de bijbehorende standaard toetsen, de T-toets en de Chi-kwadraat toets.

7.1 Algemene theorie

We beschrijven eerst netjes wat een toestingsprocedure is. Omdat je toetst op grond van waarnemingen die aan variatie onderhevig zijn moet je dan een kansmodel maken voor het toetsingsprobleem en je goed realiseren wat de fouten zijn die je kan maken bij zo'n conclusie. Vervolgens ga je bepalen wat een goede toetsingsprocedure is. Deze onderwerpen, de algemene toetsingstheorie, zullen we hieronder behandelen. We illustreren de theorie aan de hand van het voorbeeld van de robot.

Voorbeeld 7.1 (Robot)

In Voorbeeld 3.10 hebben we de kansverdeling van een stochastische variabele X, de tijd die het duurt totdat een robot een bepaalde taak heeft verricht, gemodelleerd. We hebben daar gekozen voor een exponentiële verdeling met een parameter $\lambda > 0$. De kansdichtheid van X is dan gelijk aan

$$f(x) = \begin{cases} \frac{1}{\lambda}\, e^{-x/\lambda} & , \text{ als } x > 0, \\ 0 & , \text{ als } x \leq 0. \end{cases} \tag{7.1}$$

De parameter λ is onbekend. Bij deze familie van verdelingen is de parameter λ gelijk aan de verwachting van X. Stel dat er beweerd wordt dat de verwachte tijd die de robot nodig heeft

om zijn taak te verrichten gelijk is aan tien minuten. De bewering is dus: $\lambda = 10$. Om een idee te krijgen of je die bewering nu moet geloven of niet, kan je de robot een aantal keren, zeg n keer, vanuit willekeurig gekozen startposities, zijn taak laten doen. Laten we zeggen dat dit n tijdsduren $x_1, \ldots, x_n$ oplevert. Je hebt dan een steekproef van omvang n genomen uit de verdeling van X, de exponentiële verdeling met een parameter $\lambda > 0$.

We zullen nu het toestingsprobleem formeel invoeren. Een **toets** (test) is gebaseerd op een steekproef $X_1, \ldots, X_n$, dus op onafhankelijke waarnemingen, die allemaal dezelfde verdeling hebben. Deze verdeling is er een uit een parametrische familie van verdelingen met een parameter θ die onbekend is. De bewering die we willen toetsen noemen we de **nulhypothese** (null hypothesis). Deze wordt genoteerd met H_0. We willen deze hypothese toetsen tegen een **alternatieve hypothese** (alternative hypothesis) die we noteren met H_1. Het meest eenvoudige voorbeeld is

$$H_0 : \theta = \theta_0 \quad \text{tegen} \quad H_1 : \theta = \theta_1, \tag{7.2}$$

waarbij θ_0 en θ_1 twee mogelijke verschillende waarden voor θ zijn. Het gaat hier dus om het toetsen van de waarde θ_0 voor θ, terwijl er maar een mogelijke andere waarde is, namelijk θ_1. Andere mogelijke hypothesen zijn bijvoorbeeld

$$H_0 : \theta = \theta_0 \quad \text{tegen} \quad H_1 : \theta \neq \theta_0, \tag{7.3}$$
$$H_0 : \theta = \theta_0 \quad \text{tegen} \quad H_1 : \theta > \theta_0, \tag{7.4}$$
$$H_0 : \theta \leq \theta_0 \quad \text{tegen} \quad H_1 : \theta > \theta_0. \tag{7.5}$$

Als de hypothese uit één waarde van θ bestaat heet hij een **enkelvoudige hypothese**. In de andere gevallen spreken we over een **samengestelde hypothese**.

Om de toets uit te voeren moeten we eerst een toetsingsgrootheid kiezen waarmee we gaan werken.

Definitie 7.2 *Een* **toetsingsgrootheid** *(test statistic) T is een functie van alleen de waarnemingen $X_1, \ldots, X_n$ van de steekproef. Dus*

$$T = t(X_1, \ldots, X_n). \tag{7.6}$$

Een toestingsgrootheid is dus eigenlijk hetzelfde als een schatter, alleen gebruiken we hem voor een ander doel. We zien hier dus dat T een stochastische variabele is met een kansverdeling die bepaald wordt door de parameter θ. Hij is immers een functie van de waarnemingen die ook allemaal een kansverdeling hebben die door θ bepaald wordt.

We berekenen eerst de waarde van T op basis van de realisatie $x_1, \ldots, x_n$ van je steekproef. Laten we die uitkomst t noemen. Dus $t = t(x_1, \ldots, x_n)$. We redeneren nu als volgt. Als de nulhypothese (7.2) waar is, dus θ is gelijk aan θ_0, dan kennen we de kansverdeling van T, en weten we dus ongeveer wat we als uitkomsten kunnen verwachten. Als de gevonden uitkomst t 'te onwaarschijnlijk' is dan geloven we H_0 niet. De algemene procedure is dus als volgt.

1. Stel we willen de hypothesen (7.2) toetsen. Bereken de waarde van T op basis van de steekproefuitkomst. Dit levert een waarde t.

2. Als de gevonden waarde t 'te onwaarschijnlijk' is dan **verwerp** je de nulhypothese H_0 ten gunste van de alternatieve hypothese H_1.

3. Als de gevonden waarde t niet 'te onwaarschijnlijk' is, dus als hij best had kunnen optreden, als H_0 waar is, dan **verwerpen we H_0 niet**.

Deze beschrijving is nog wat vaag maar we gaan hem verder uitwerken. Opvallend is ook dat we niet zeggen dat we de nulhypotese aanvaarden, geloven, als we hem niet verwerpen. Dit zit hem in de manier waarop we de toets verder gaan inrichten.

Bij een toetsingsprocedure kan je de volgende fouten maken. Je zou H_0 kunnen verwerpen als H_0 in feite wel waar is. Dit noem je een **fout van de eerste soort** (Type I error). Ook zou je H_0 niet kunnen verwerpen als je dat wel zou moeten doen, dus als H_1 waar is. Dit noem je een **fout van de tweede soort** (Type II error). Andere benamingen hiervoor zijn **false negative** en **false positive**. In Tabel 7.1 wordt een overzicht gegeven. Omdat de conclusies

	H_0 is waar	H_1 is waar
Verwerp H_0	Fout eerste soort	Juiste beslissing
Verwerp H_0 niet	Juiste beslissing	Fout tweede soort

Tabel 7.1: Twee soorten fouten.

gebaseerd zijn op de uitkomst van de steekproef zijn ze aan variatie onderhevig. Bij verschillende uitkomsten zal je wellicht verschillende conclusies trekken. Je kan dus spreken over de kans op een fout van de eerste soort en van een kans op een fout van de tweede soort. Eigenlijk wil je die twee kansen beiden klein hebben. Dit kan in het algemeen niet. We concentreren ons dan op het begrensen van de kans op een fout van de eerste soort. Daarnaast willen we de kans op een fout van de tweede soort zo klein mogelijk hebben. Hier geven we dus in feite prioriteit aan de fout van de eerste soort.

We kiezen van te voren een een grens op de kans op een fout van de eerste soort. Die grens noteren we met α, de **onbetrouwbaarheidsdrempel** (size) van de toets. Gebruikelijk is die kans gelijk te kiezen aan 0.05. Dus eens op de twintig keer mogen we die fout maken. Als de gevolgen van een foute conclusie van de eerste soort zeer ernstig zijn zou je de α wellicht kleiner kiezen. We richten de toets dus zo in dat geldt

$$P(\text{Verwerp } H_0) \leq \alpha \ (0.05), \quad \text{als } H_0 \text{ waar is.} \tag{7.7}$$

De kwaliteit van de toets hangt natuurlijk af van de toetsingsgrootheid die gekozen is. Voor een goede toets is voor θ's, waarvoor H_1 waar is, de kans op verwerpen van H_0 zo groot mogelijk. Voor θ's waarvoor H_1 waar is wil je dat de toets H_0 verwerpt. We noemen we de kans

$$P(\text{Verwerp } H_0) \tag{7.8}$$

voor waarden van θ waarvoor H_1 waar is, het **onderscheidingsvermogen** (power) van de toets in θ. Dit onderscheidingsvermogen willen we dus zo hoog mogelijk hebben.

De waarden van t waarvoor we H_0 gaan verwerpen noemen we het **kritieke gebied**. Meestal verwerpen we H_0 als T te groot uitvalt, dus als $T \geq k$, of als T te klein uitvalt, dus als $T \leq k$. Dit soort toetsen noemen we **eenzijdige toetsen**. Bij sommige toetsingsproblemen verwerp je H_0 als T te groot of te klein uitvalt dus als $T \leq k_1$ of $T \geq k_2$, een **tweezijdige toets**. De getallen k, k_1 en k_2 noemen we de **kritieke waarden** van de toets. De kritieke waarden moeten zo gekozen worden dat aan (7.7) voldaan is. Als een steekproefuitkomst zodanig is dat de nulhypothese H_0 verworpen wordt dan noemen we die uitkomst **significant**.

Voorbeeld 7.3 (Robot)

We illustreren de algemene theorie weer aan de hand van het voorbeeld met de robot. Stel we hebben $n = 20$ tijdsduren gemeten, zeg $x_1, \ldots, x_{20}$. We hebben eerder gezien dat het steekproefgemiddelde een zuivere schatter is van λ. De waarde van het gemiddelde geeft dus informatie over de echte waarde van λ. We besluiten om het steekproefgemiddelde als toetsingsgrootheid te gebuiken. Dus

$$T = \bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n). \tag{7.9}$$

Als hypothesen kiezen we hier

$$H_0 : \lambda = 10 \quad \text{en} \quad H_1 : \lambda \neq 10 \ (\lambda > 0). \tag{7.10}$$

T is een zuivere schatter van λ. We hebben dus reden om aan de waarheid van H_0 te twijfelen als T veel kleiner dan tien of veel groter dan tien uitvalt. De vraag is nu: hoeveel kleiner en hoeveel groter? We kiezen dus voor een tweezijdige toets. Met kritieke waarden k_1 en k_2 zit de toets er als volgt uit

1. Verwerp H_0 als $T \leq k_1$ of $T \geq k_2$,
2. Verwerp H_0 niet als $k_1 < T < k_2$.

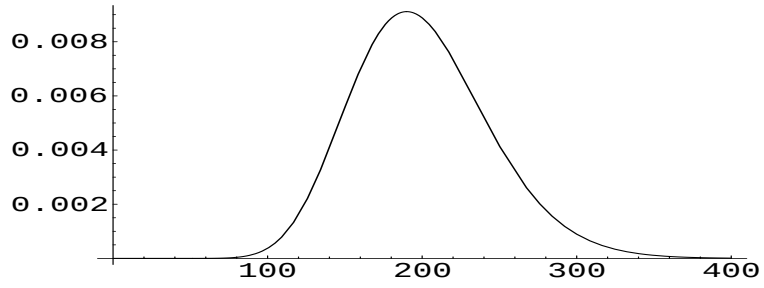
De kritieke waarden moeten we zó kiezen dat aan (7.7) voldaan is. Dus zó dat de kans op een fout van de eerste soort kleiner of gelijk is aan 0.05. Dit is voor elkaar als, gegeven dat H_0 waar is, dus als λ gelijk is aan tien,

$$P(T \leq k_1) = 0.025 \quad \text{en} \quad P(T \geq k_2) = 0.025. \tag{7.11}$$

Samen is dat dan een kans van 0.05. We moeten dus de kansverdeling van T afleiden voor λ gelijk aan tien. De kansverdeling van $X_1 + \ldots + X_n$, voor onafhankelijke Exp(λ) verdeelde stochastische variabelen, is een bekende verdeling, namelijk de Gamma(n, λ) verdeling. In ons geval is n gelijk aan 20 en λ, als H_0 waar is, gelijk aan tien. De kansdichtheid van $X_1 + \ldots + X_{20}$ is gegeven in Figuur 7.1. De verwachting van deze verdeling is overigens gelijk aan

$$\mathrm{E}\,(X_1 + \ldots + X_{20}) = \mathrm{E}\,(X_1) + \ldots + \mathrm{E}\,(X_{20}) = 20\mathrm{E}\,(X_1) = 200.$$

Laten we $X_1 + \ldots + X_{20}$ aanduiden met U. We hebben dan $T = U/20$. Van de stochastische variabele U kennen we dus onder H_0 de kansverdeling. Voor die Gamma$(20, 10)$ verdeling



Figuur 7.1: $n = 20, \lambda = 10$, Gamma(20,10) dichtheid .

bestaan er procedures om zijn verdelingsfunctie uit te rekenen. Daarmee kunnen we dan twee waarden $u_1 < u_2$ uitrekenen waarvoor geldt

$$P(U \leq u_1) = 0.025 \quad \text{en} \quad P(U \geq u_2) = 0.025. \tag{7.12}$$

We vinden dan $u_1 = 122.165$ en $u_2 = 296.709$. De kritieke waarden k_1 en k_2 vinden we vervolgens door de volgende twee vergelijkingen op te lossen

$$P(T \leq k_1) = P(U \leq 20k_1) = 0.025 \quad \text{en} \quad P(T \geq k_2) = P(U \geq 20k_2) = 0.025. \tag{7.13}$$

Dus er moet gelden $20k_1 = u_1$ en $20k_2 = u_2$, met als oplossingen $k_1 = 122.165/20 = 6.11$ en $k_2 = 296.709/20 = 14.83$. De toets die we geconstrueerd hebben is dus nu gelijk aan

1. Verwerp H_0 als $\bar{X}_n \leq 6.11$ of $\bar{X}_n \geq 14.83$.
2. Verwerp H_0 niet als $6.11 < \bar{X}_n < 14.83$.

Deze toets heeft onbetrouwbaarheid 0.05.

Stel we hebben de volgende 20 tijdsduren gemeten:

> 2.11057, 17.3518, 0.783778, 4.82514, 0.612949, 0.320131, 10.9404, 5.72977, 5.02806, 8.79567, 4.97478, 19.6352, 18.895, 19.8331, 1.63212, 36.2795, 29.814, 2.25983, 9.1043, 9.53233.

Het gemiddelde van de metingen is gelijk aan 10.4229. In dit geval zouden we H_0 dus niet verwerpen.

Stel dat we de volgende tijdsduren hebben gemeten:

> 96.0342, 19.3856, 3.96892, 46.0638, 10.5378, 15.4965, 8.59473, 18.6583, 0.00340679, 11.854, 1.73543, 0.605641, 13.604, 0.251315, 24.9287, 23.8209, 44.3242, 21.8183, 22.8898, 15.7644.

Het gemiddelde van de metingen is gelijk aan 20.017. In dit geval zouden we H_0 dus wel verwerpen.

Een alternatieve manier om te conroleren of je met de toets moet verwerpen of niet, is om de zogenaamde **overschrijdingskansen** (p values) uit te rekenen. Bij deze toets bereken je dan bijvoorbeeld voor de eerste dataset, voor λ gelijk aan tien, de kansen $P(T \leq 10.4229)$ en

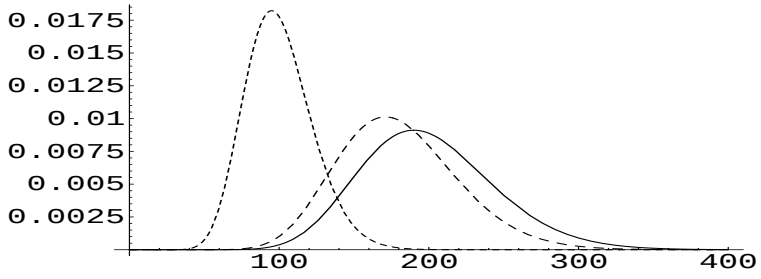
$P(T \geq 10.4229)$. Dit zijn dus de kansen dat je, als H_0 waar is, onder of boven de gevonden waarde 10.4229 van T terecht komt. In dit geval zijn die kansen gelijk aan 0.60 en 0.40. Als een van deze kansen kleiner is dan 0.025 verwerp je H_0. In dit geval is dat niet zo.

Laten we nu voorbeeld met andere hypothesen bekijken. Stel je wil aantonen dat de robot gemiddeld meer dan tien minuten nodig heeft, dus $\lambda > 10$, om de taak te volbrengen. Je kan eigenlijk alleen de nulhypothese veilig verwerpen. De toets is namelijk zo opgebouwd dat de kans op een fout van de eerste soort kleiner of gelijk is aan 0.05. Voor de kans op een fout van de tweede soort heb je een dergelijke garantie niet. Je moet dus de conclusie die je wil trekken als alternatieve hypothese kiezen. De keuze voor de hypothesen is in dit geval

$$H_0 : \lambda \leq 10 \quad \text{en} \quad H_1 : \lambda > 10. \tag{7.14}$$

We gebruiken hier dezelfde toetsingsgrootheid, het steekproefgemiddelde. Je gaat hier de nulhypothese verwerpen als T, het gemiddelde van de metingen, te groot uitvalt. We gebruiken hier dus een eenzijdige toets. De toets ziet er dan als volgt uit

1. Verwerp H_0 als $T \geq k$,
2. Verwerp H_0 niet als $T < k$.



Figuur 7.2: Kansdichtheden van de Gamma(20,λ) verdelingen voor $\lambda = 5, 9$ en 10 . De solide lijn is die van de Gamma(20,10) kansdichtheid.

We moeten nu de kritieke waarde k bepalen. Om de fout van de eerste soort kleiner dan 0.05 te houden moet er dan gelden, voor de waarden van θ waarvoor H_0 waar is,

$$P(\text{Verwerp } H_0) = P(T \geq k) \leq \alpha \ (0.05). \tag{7.15}$$

Dit moet dus gelden voor alle $\lambda \leq 10$. Hier is de nulhypothese dus niet meer enkelvoudig maar is hij samensgesteld. De kans (7.15) moet dus voor alle $0 < \lambda \leq 10$ kleiner of gelijk zijn aan 0.05. In Figuur 7.2 zien we dat dit zo is als we de kritieke waarde k zó kiezen dat voor λ gelijk aan tien de kans op verwerpen van H_0 gelijk is aan 0.05. We moeten dus k hier zó kiezen dat voor λ gelijk aan tien

$$P(\text{Verwerp } H_0) = P(T \geq k) = P(U \geq 20k) \leq \alpha \ (0.05). \tag{7.16}$$

Net als boven bepalen we eerst een waarde u waarvoor $P(U \geq u) = 0.05$. Die waarde van u is gelijk aan 278.792. De kritieke waarde k is dan gelijk aan $u/20 = 278.792 = 13.94$. De toets ziet er dan als volgt uit

1. Verwerp H_0 als $T \geq 13.94$.

2. Verwerp H_0 niet als $T < 13.94$.

Als we kijken naar de twee datasets hierboven verwerpen we dus H_0 in dit geval alleen bij de tweede dataset.

7.2 De T-toets

Een van de meest voorkomende toetsingsproblemen is de toets van een hypothese over de verwachting μ van een $\mathcal{N}(\mu, \sigma^2)$ verdeling waarbij σ^2 onbekend is. De nulhypothese en alternatieve hypothese zijn dan bijvoorbeeld

$$H_0 : \mu = \mu_0 \quad \text{en } H_1 : \mu \neq \mu_0. \tag{7.17}$$

De standaard toets voor dit probleem is de Student T-toets gebaseerd op de T toetsingsgrootheid

$$T_n = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} = \sqrt{n}\,\frac{\bar{X}_n - \mu_0}{S_n}, \tag{7.18}$$

waarbij S_n, de steekproefstandaardafwijking, is gegeven door

$$S_n = \sqrt{S_n^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}.$$

Zie ook paragraaf 6.4.2 waar de stochastische variabele T gebruikt wordt voor het contrueren van een betrouwbaarheidsinterval.

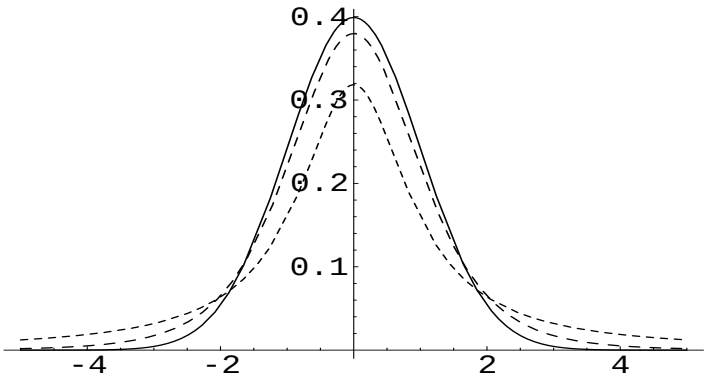
We weten dat het steekproefgemiddelde een goede schatter is van de echte μ, de verwachting van de verdeling van de waarnemingen. Als die μ niet gelijk is aan μ_0 dan zal het steekproefgemiddelde afwijken van μ_0 en zal T een 'te kleine' of 'te grote' waarde aannemen. Voor het toetsen van de nulhypothese in (7.17) ziet de toets er dan als volgt uit

1. Verwerp H_0 als $T \leq -k$ of $T \geq k$,

2. Verwerp H_0 niet als $-k < T < k$,

voor kritieke waarden $-k$ en k.

Als de nulhypothese waar is, dus als de μ die hoort bij de verdeling van de waarnemingen gelijk is aan μ_0, dan heeft T een Student verdeling met $n-1$ vrijheidsgraden. Deze verdeling is symmetrisch om nul. Hierdoor kunnen we de kritieke waarde gelijk nemen aan $-k$ en k, dus ook symmetrisch om nul. Een aantal Student kansdichtheden worden gegeven in Figuur 7.3.



Figuur 7.3: Dichtheden van de Student verdeling voor $n = 1, 5$ en de normale dichtheid.

Om de kans op een fout van de eerste soort in de hand te houden eisen we nu, voor μ gelijk aan μ_0, dus als H_0 waar is,

$$\begin{aligned} 0.05 &= P(\text{Verwerp } H_0) = P(T \leq -k) + P(T \geq k) \\ &= 2P(T \leq -k), \end{aligned}$$

als we een onbetrouwbaarheidsdrempel $\alpha = 0.05$ eisen. De waarde van k kunnen we aflezen uit een tabel van de betreffende Student verdeling of berekenen met een computerprocedure. Voor een tabel zie Sectie 7.4.

De hierboven beschreven toets is een tweezijdige T-toets. De aanpassing voor een eenzijdige toets ligt voor de hand en is net zo als in het voorbeeld van de robot.

Voorbeeld 7.4 (Temperatuurstijging)

Een milieugroepering bestudeert de temperatuurstijging van water dat 50 meter verwijderd is van een kernreactor. Men maakt zich zorgen of de temperatuur meer dan 3 graden Celsius is gestegen. Men heeft 16 metingen $x_1, \ldots, x_{16}$ van temperatuurstijgingen. Hiervan is $\bar{x} = 3.24$ en $s^2 = 1.12$. Men veronderstelt dat de data mogen worden opgevat als een realisatie van een steekproef uit een $\mathcal{N}(\mu, \sigma^2)$ verdeling en besluit om een T-toets voor

$$H_0 : \mu = 3 \quad \text{tegen} \quad H_1 : \mu > 3$$

uit te voeren bij significantieniveau 0.05. Ze willen immers concluderen dat de temperatuur gestegen is. We verwerpen H_0 hier als T groot uitvalt, dus als $T \geq k$. We kunnen k bepalen door

$$P(T \geq k) = 0.05 \tag{7.19}$$

op te lossen in het geval dat H_0 waar is, dus als de echte waarde van μ gelijk is aan 3. In dat geval heeft T een Student verdeling met 15 vrijheidsgraden. Als we de kritieke waarde bij de Student verdeling met 15 vrijheidsgraden, en α gelijk aan 0.05, opzoeken of uitrekenen, dan vinden we k=1.753.

De waarde, zeg t, van T die we op grond van deze waarnemingen kunnen uitrekenen is gelijk aan

$$t = \sqrt{16}\,\frac{3.24 - 3}{\sqrt{1.12}} = 0.91.$$

We kunnen op grond van deze waarnemingen de nulhypothese dus niet verwerpen.

7.3 De Chi-kwadraat toets

Een belangrijke toets voor discreet verdeelde waarnemingen is de zogenaamde Chi-kwadraat toets. De modelveronderstelling hier is dat we waarnemingen hebben die we in het model kunnen omschrijven als een steekproef $X_1, \ldots, X_n$ uit een discrete verdeling gegeven door

$$X_i : \begin{pmatrix} r_1 & r_2 & \ldots & r_m \\ p_1 & p_2 & \ldots & p_m \end{pmatrix}, \tag{7.20}$$

waarbij $p_1 = P(X = r_1), \ldots, p_m = P(X = r_m)$, en $p_1 + \ldots + p_m = 1$. Deze kansen, de parameter vector in dit model, zijn onbekend. De nulhypothese die we hier willen toetsen is dat de p's gelijk zijn aan gegeven kansen $p_1^0, \ldots, p_m^0$. Dus

$$H_0 : (p_1, \ldots, p_m) = (p_1^0, \ldots, p_m^0) \text{ en } H_1 : (p_1, \ldots, p_m) \neq (p_1^0, \ldots, p_m^0). \tag{7.21}$$

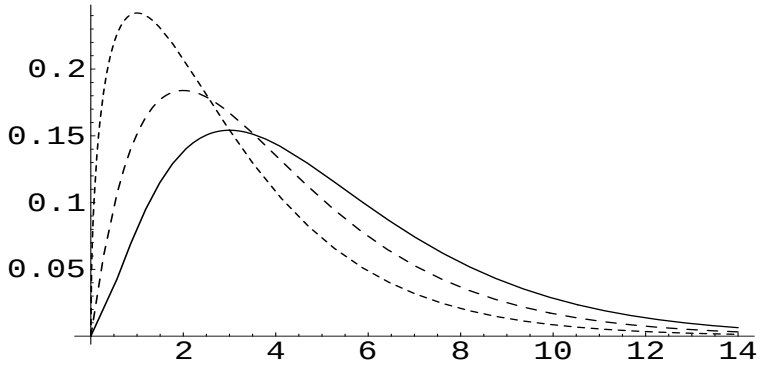
We tellen eerst het aantal keren dat de verschillende uitkomsten $r_1, \ldots, r_m$ voorkomen in de waarnemingen. Laat N_i het aantal keren zijn dat r_i voorkomt. We weten dan dat N_i een $\text{Bin}(n, p_i)$ verdeling heeft. De verwachting van N_i is dus gelijk aan np_i. Als H_0 waar is, dan is de verwachting van N_i dus gelijk aan np_i^0. De volgende toetsingsgrootheid, de zogenaamde Chi-kwadraat toetsingsgrootheid, is een maat voor de afwijking van de gevonden aantallen N_i tot de verwachte aantallen onder de nulhypothese,

$$\chi^2 = \sum_{i=1}^{m} \frac{(N_i - np_i^0)^2}{np_i^0}. \tag{7.22}$$

Als de waarde van χ^2 groot is dan wijken de gevonden aantallen af van de verwachte waarden als de nulhypothese waar is. We verwerpen H_0 dan ook als $\chi^2 \geq k$, voor een zekere kritieke waarde k. Voor grote steekproeven, dus voor grote waarden van n, is de kansverdeling van χ^2, als de nulhypothese waar is, bij benadering gelijk aan een zogenaamde χ^2_{m-1} verdeling, een Chi-kwadraat verdeling met $m - 1$ vrijheidsgraden. In Figuur 7.4 worden de kansdichtheden van een aantal van deze verdelingen gegeven. Deze verdelingen zijn getabelleerd en er bestaan computerprocedures voor. We kunnen dan ook k zó kiezen dat, als H_0 waar is,

$$0.05 = P(\text{Verwerp } H_0) = P(\chi^2 \geq k) \approx P(U \geq k),$$

waarbij U een Chi-kwadraat verdeling heeft met $m - 1$ vrijheidsgraden.



Figuur 7.4: Dichtheden van de Chi-kwadraat verdelingen met vrijheidsgraden $3, 4, 5$.

Voorbeeld 7.5 (Vroege Genetica)

Waarschijnlijk de eerste experimenten in de Genetica zijn uitgevoerd door een Oostenrijkse monnik, Gregor Mendel (1822-1884). Hij had een theorie over net overerven van genetische eigenschappen die hij in de praktijk toetste door het kruisen van erwtenrassen. Bij een van zijn experimenten gebruikte hij erwten met ronde gele zaadjes en erwten met gerimpelde groene zaadjes. Na kruisen van deze planten kreeg hij zaadjes met een van vier mogelijke eigenschappen: rond en geel (1), gerimpeld en geel (2), rond en groen (3) en gerimpeld en groen (4). De getallen 1,2,3,4 geven de uitkomst van een kruisingsexperiment aan. Het aantal verschillende uitkomsten, m, is hier gelijk aan vier.

Volgens Mendels theorie zou de kansverdeling van de uitkomst van elk experiment, X_i, gelijk moeten zijn aan

$$X_i : \begin{pmatrix} 1 & 2 & 3 & 4 \\ \frac{9}{16} & \frac{3}{16} & \frac{3}{16} & \frac{1}{16} \end{pmatrix}. \tag{7.23}$$

Hij heeft 556 van die kruisingen gedaan. Dus de steekproefomvang n is hier gelijk aan 556. Tellingen van de vier soorten zaadjes leverde : $N_1 = 315, N_2 = 101, N_3 = 108, N_4 = 32$. De verwachte aantallen onder de nulhypothese dat de theorie klopt, dus onder

$$H_0 : (p_1, \ldots, p_4) = (\tfrac{9}{16}, \tfrac{3}{16}, \tfrac{3}{16}, \tfrac{1}{16}) \tag{7.24}$$

zijn gelijk aan: 312.75, 104.25, 104.25, 34.75. We berekenen voor deze uitkomsten de waarde van χ^2 en vinden dan $\chi^2 = 0.47$. Om te kijken of deze waarde in het kritieke gebied ligt berekenen we de overschrijdingskans en vinden dan $P(\chi^2 \geq 0.47) = 0.9$. De uitkomst 0.47 ligt dus niet in het kritieke gebied. We kunnen de theorie van Mendel dus ook niet verwerpen.

7.4 Tabellen van kritieke waarden voor de T-toets en de Chi-kwadraat toets

De volgende tabellen zijn overgenomen van
http://www.statsoft.com/textbook/sttable.html .

m/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
inf	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905

Tabel 7.2: Kritieke waarden voor de Student(m) verdeling. Gegeven zijn de waarden van k waarvoor $P(Z_m \geq k) = p$, waarbij Z_m een Student (m) verdeling heeft.

m/p	.500	.250	.100	.050	.025	.010	.005
1	0.45494	1.32330	2.70554	3.84146	5.02389	6.63490	7.87944
2	1.38629	2.77259	4.60517	5.99146	7.37776	9.21034	10.59663
3	2.36597	4.10834	6.25139	7.81473	9.34840	11.34487	12.83816
4	3.35669	5.38527	7.77944	9.48773	11.14329	13.27670	14.86026
5	4.35146	6.62568	9.23636	11.07050	12.83250	15.08627	16.74960
6	5.34812	7.84080	10.64464	12.59159	14.44938	16.81189	18.54758
7	6.34581	9.03715	12.01704	14.06714	16.01276	18.47531	20.27774
8	7.34412	10.21885	13.36157	15.50731	17.53455	20.09024	21.95495
9	8.34283	11.38875	14.68366	16.91898	19.02277	21.66599	23.58935
10	9.34182	12.54886	15.98718	18.30704	20.48318	23.20925	25.18818
11	10.34100	13.70069	17.27501	19.67514	21.92005	24.72497	26.75685
12	11.34032	14.84540	18.54935	21.02607	23.33666	26.21697	28.29952
13	12.33976	15.98391	19.81193	22.36203	24.73560	27.68825	29.81947
14	13.33927	17.11693	21.06414	23.68479	26.11895	29.14124	31.31935
15	14.33886	18.24509	22.30713	24.99579	27.48839	30.57791	32.80132

Tabel 7.3: Kritieke waarden voor de Chi-kwadraat(m) verdeling. Gegeven zijn de waarden van k waarvoor $P(\chi^2_m \geq k) = p$, waarbij χ^2_m een Chi-kwadraat(m) verdeling heeft.

7.5 Opgaven

1. Stel je hebt een dataset die je kan opvatten als een realisatie van een steekproef $X_1, \ldots, X_n$ uit een kansverdeling met onbekende verwachting m. Je wil de hypothese $H_0 : m = 0$ toetsen tegen $H_1 : m \neq 0$ toetsen. Welke toetsingsgrootheid zou je gebruiken? Voor welke uitkomsten van de toetsingsgrootheid zou je de nulhypothese verwerpen? (in termen van 'T groot' of 'T klein', of 'T groot of klein') Is dit een eenzijdige of een tweezijdige toets?

2. Bekijk dezelfde situatie als in opgave 1. Beantwoordt de vragen voor het geval je de volgende hypothesen wil toetsen

   (a) $H_0 : m = 1$ tegen $H_1 : m > 1$.
   (b) $H_0 : m = 5$ tegen $H_1 : m < 5$.
   (c) $H_0 : m \geq 10$ tegen $H_1 : m < 10$.

3. Voor het toetsen van een bepaalde hypothese H_0 met significantieniveau $\alpha = 0.05$, op grond van een dataset, gebruiken we de toetsingsgrootheid T. Die toetsingsgrootheid is zó gekozen dat we de nulhypothese verwerpen als de waarde van T te groot is. Dit is dus een eenzijdige toets.

   (a) Stel de waarde van T is voor de dataset gelijk aan 2.34. Onder H_0 geldt $P(T \geq 2.34) = 0.23$. Moeten we op grond hiervan H_0 verwerpen?
   (b) Stel de waarde van T is voor de dataset gelijk aan 2.34. Onder H_0 geldt $P(T \leq 2.34) = 0.23$. Moeten we op grond hiervan H_0 verwerpen?
   (c) Stel de waarde van T is voor de dataset gelijk aan 0.03. Onder H_0 geldt $P(T \geq 0.03) = 0.968$. Moeten we op grond hiervan H_0 verwerpen?
   (d) Stel de waarde van T is voor de dataset gelijk aan 1.07. Onder H_0 geldt $P(T \leq 1.07) = 0.981$. Moeten we op grond hiervan H_0 verwerpen?
   (e) Stel de waarde van T is voor de dataset gelijk aan 1.07. Onder H_0 geldt $P(T \leq 1.07) = 0.01$. Moeten we op grond hiervan H_0 verwerpen?

4. Beantwoordt de vragen van opgave 3 nu voor de situatie waarin we gekozen hebben voor een toetsingsgrootheid T waarmee je de nulhypothese verwerpt als T te klein is.

5. Beantwoordt de vragen van opgave 3 nu voor de situatie waarin we gekozen hebben voor een toetsingsgrootheid T waarmee je de nulhypothese verwerpt als T te klein of te groot is. Dit is dan dus een tweezijdige toets.

6. Stel je hebt een munt. Iemand beweert dat het een eerlijke munt is, dus dat de kans op kruis gelijk is aan een half. Je hebt reden om daaraan te twijfelen en gaat dit statistisch onderzoeken. Je doet 10 worpen met de munt. De uitkomsten zijn 0, 1, 1, 1, 0, 1 , 0, 1, 1, 1 waarbij 1 staat voor kruis en 0 staat voor munt. Je besluit het totaal aantal keren kruis, zeg X, als toestingsgrootheid te nemen.

(a) Wat zijn de hypothesen?

(b) Bepaal de kritieke waarden voor onbetrouwbaarheid α gelijk aan 0.05 .

(c) Wat is je conclusie?

Een gedeelte van de verdelingsfunctie van de Bin(10,0.5) verdeling wordt in de volgende tabel gegegeven.

x	5	6	7	8	9	10
$P(X \leq x)$	0.6230	0.8281	0.9453	0.9893	0.9990	1.0000

7. Een statisticus gaat op congres naar Kreta. Men beweert dat de temperatuur in de betreffende maand gemiddeld minstens 20 graden is. Van internet heeft hij 20 temperatuur metingen van eerdere jaren gehaald, $x_1, \ldots, x_{20}$. Hiervan is $\bar{x} = 21$ en $s^2 = 4$. Hij veronderstelt dat de data mogen worden opgevat als een realisatie van een steekproef uit een $\mathcal{N}(\mu, \sigma)$ verdeling en besluit om de t-toets voor

$$H_0 : \mu = 20 \quad \text{tegen} \quad H_1 : \mu > 20$$

uit te voeren bij significantieniveau 0.05.

(a) Geef de toetsingsgrootheid, de kritieke waarde en het kritieke gebied van deze toets.

(b) Bepaal de waarde van de toetsingsgrootheid.

(c) Wat kan men concluderen op grond van de 20 waarnemingen.

(d) Waarom toetst hij juist deze nulhypothese ?

Kritieke waarden van Student (m) verdelingen worden gegeven in de tabel in Sectie 7.4. Dit zijn dus de waarden van k waarvoor geldt $P(T_m \geq k) = p$, waarbij T_m een Student (m) verdeling heeft.

8. Een bedrijf heeft een systeem bedacht voor het toekennen van salarisverhogingen aan zijn werknemers. Volgens een bepaald systeem krijgt een werknemer geen salarisverhoging, een standaardverhoging of nog een extra verhoging daarbovenop. Volgens de planning van het managment zouden deze verhogingen moeten voorkomen in respectievelijk in 10%, 65% en 25% van de gevallen.

Men heeft het systeem een jaar lang toegepast en gevonden dat er van de 600 werknemers 42 geen, 365 een standaard, en 193 een extra verhoging kregen.

Pas een Chi-kwadraat toets toe om te toetsen of de gevonden percentages significant afwijken van de nagetreefde percentages. Neem de onbetrouwbaarheid van de toets gelijk aan $\alpha = 0.05$.

Kritieke waarden van Chi-kwadraat (m) verdelingen worden gegeven in de tabel in Sectie 7.4. Dit zijn dus de waarden van k waarvoor geldt $P(\chi^2_m \geq k) = p$, waarbij χ^2_m een Chi-kwadraat (m) verdeling heeft.

Bibliografie

[1] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis (2nd ed.)*. Wiley, New York, 1984.

[2] P. Bickel and K. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, Oakland, Calif., 1977.

[3] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, Inc., New Jersey, 1998.

[4] B.D. Ripley. *Stochastic Simulation*. Wiley, New York, 1987.