Stochastiek voor Informatici

S.A. van de Geer, 2000
E. Belitser, 2001, 2002
nieuwe versie F.M. Spieksma, 2003

16 april 2003

Inhoudsopgave

1 Inleiding: Uniforme verdeling, transformaties, wet van de grote aantallen.

Voorbeeld 1.1 Persoon A kiest een geheel getal X uit de getallen 1 t/m 10.

$$X \in \{1, \ldots, 10\}.$$

Persoon B heeft geen idee welk getal A gekozen heeft. Voor B is X een *stochastische grootheid* (*kansvariabele*). De kans dat B het goede getal raadt, is

$$\frac{1}{10}.$$

Algemeen: we spreken van de *uitkomst* X van een *experiment*. We zeggen dat X een *aselecte trekking* is als iedere mogelijke uitkomst dezelfde kans heeft. Bij een aselecte trekking uit de getallen $\{1, \ldots, m\}$, is de kans op getal x dus gelijk aan $1/m$, voor alle $x \in \{1, \ldots, m\}$. We schrijven dit als

$$\mathsf{P}\{X = x\} = \frac{1}{m}, \quad x = 1, \ldots, m.$$

Hier staat P voor P*robability*.

Voorbeeld 1.2 We gooien met een zuivere dobbelsteen. Laat X het aantal ogen zijn. Dan is

$$\mathsf{P}\{X = x\} = \frac{1}{6}, \ x = 1, \ldots, 6.$$

Voorbeeld 1.3 (college Evolutionaire Algoritmen) Gegeven twee willekeurige binaire getallen a_1, $a_2 \in \{0,1\}^n$ ter lengte n, en de gebruikelijke mutatieoperator $m : \{0,1\}^n \to \{0,1\}^n$, met kans $p \in (0,1)$ dat een gegeven bit muteert. Wat is de kans dat a_1 muteert in a_2, d.w.z. de kans $\mathsf{P}\{m(a_1) = a_2\}$?

Analoog wordt met

$$\mathsf{P}\{X \in A\}$$

de kans dat X in de verzameling A valt aangegeven.

Voorbeeld 1.4 Stel X is het aantal ogen bij het gooien met een dobbelsteen. Dan is

$$\mathsf{P}\big\{X \in \{2,4,6\}\big\}$$

de kans op een even aantal ogen. Hoe groot is deze kans?

[1] dit laatste behoort niet tot de tentamenstof voor zover het continue verdelingen betreft
[2] convolutieformules behoren niet tot tentamenstof

1.1 Discrete uniforme verdeling.

Als X een aselecte trekking uit $\{1,\ldots,m\}$ is, dan zeggen we dat X *uniform verdeeld* is over de getallen $\{1,\ldots,m\}$.

Aan één getal kun je niet zien of het de uitkomst is van een aselecte trekking. Als het experiment een aantal keren herhaald wordt, dan zal, in het geval de trekkingen *onderling onafhankelijke* en *aselect* zijn, iedere mogelijke uitkomst ongeveer even vaak voorkomen.

Laat $X_1,\ldots,X_n$ de uitkomsten zijn van n *onderling onafhankelijke* (o.o.) trekkingen uit de getallen $\{1,\ldots,m\}$. Met onderling onafhankelijk bedoelen we dat de uitkomst van het ene experiment geen informatie bevat over de uitkomst van een ander experiment. Dan geldt:

$$\lim_{n\to\infty} \frac{\{\text{aantal } X_i \text{ gelijk aan } x,\ i\le n\}}{n} = \frac{1}{m},\ x=1,\ldots,m,$$

d.w.z. voor n groot (veel herhalingen van het experiment), is de *frequentie* van een uitkomst ongeveer gelijk aan de *kans* op die uitkomst.

Opmerking 1.1 Dit resultaat noemt men de **wet van de grote aantallen**. Het volgt (wiskundig) uit de zogenaamde kansaxioma's. Volgens de frequentisten is het per definitie zo, d.w.z. zij definiëren een kans als de limiet van herhaalde experimenten. Zou je zelf voorbeelden kunnen verzinnen waarbij dit misschien niet een realistische benadering zou kunnen zijn?

We gebruiken nu een software pakket om wat "feeling"voor toevalsgetallen aan te kweken. De volgende simulaties zijn gedaan met Matlab. Je kunt natuurlijk ook Maple, Splus, SAS, of je eigen programma gebruiken. De computer genereert *deterministische* getallen, d.m.v. een programma dat *random number generator* wordt genoemd (*random* = stochastisch). De manier waarop dat gebeurt is zó, dat ze haast niet van toevalsgetallen te onderscheiden zijn. Er zijn diverse statistische tests om na te gaan of bepaalde getallen zich gedragen als toevalsgetallen. Een voorbeeld van zo'n test is boven al genoemd: bij onderling onafhankelijke aselecte trekkingen komt iedere mogelijke uitkomst ongeveer even vaak voor.

```
>>n=10;
>>unidrnd(6,1,n;)
 % dit levert n o.o. aselecte trekkingen uit 1...6
>>x=unidrnd(6,1,n)
 % dit vult de 1xn vector x met aselecte trekkingen uit 1...6
   x=
   3 6 5 4 5 4 3 2 3 4
>>x(3)
   ans=
     5
 % dit geeft het 3-de element van de trekking
 % N.B. een ; aan het eind van een commando onderdrukt output op het scherm van
   het uitgevoerde commando.
```

1.2 Realisaties.

Stel we doen 10 aselecte trekkingen uit de getallen $\{1,\ldots,6\}$ die we in Matlab in de vector x opbergen. We vinden dan 10 *getallen* $\{x_1,\ldots,x_{10}\}$. Dit noemt men wel een *realisatie* van de stochastische grootheden $X_1,\ldots,X_{10}$. De rij $3,6,5,4,5,3,2,3,4$ is zo'n realisatie.

1.3 Histogram.

Men kan de verdeling van n getallen $x_1,\ldots,x_n$ weergeven d.m.v. een *histogram*. Hierbij wordt het waardebereik van de getallen onderverdeeld in een aantal intervallen, en geteld hoeveel van de x_i in een bepaald interval liggen.

```
>> hist(x)
 % maakt histogramplotje van de 10 random getallen in een apart venster Figure no 1
>>hist(x,20)
 % verdeelt de waarnemingen over 20 bins van gelijke breedte, en plot deze
>>title('Tien aselecte trekkingen uit 1,...,6')
 %geeft titel 'Tien ...' mee aan het plaatje
>>[u,t]=hist(x)
   u =
        1     0     3     0     3     0     0     2     0     1
   t =
     Columns 1 through 7
       2.2000    2.6000    3.0000    3.4000    3.8000    4.2000    4.6000
     Columns 8 through 10
       5.0000    5.4000    5.8000
 % bergt in u op hoeveel waarnemingen er in elke bin zitten, en in t de bin-centra
```

Kun je deze uitkomst verklaren? Met het commando `help histc` kun je lezen hoe het histogram meer naar eigen wens aan te passen. Bijv.

```
>>hist(x,1:6)
 % 1:6 creeert een vector met elementen 1,2,3,4,5,6 (stapgrootte 1!),
 % dus we krijgen een histogram met bin-centra de elementen van deze vector.
```

1.4 Transformaties.

Laat X een aselecte trekking uit de getallen $\{1, \ldots, m\}$ zijn, en Y een transformatie van X:

$$Y = g(X).$$

Dan is Y in het algemeen niet meer uniform verdeeld.

Voorbeeld 1.5 Laat X een aselecte trekking uit $\{1, \ldots, 10\}$ zijn, en g de functie

$$g(x) = \begin{cases} x+1, & \text{als } x \text{ een priemgetal is, of als } x = 1\ , \\ x, & \text{anders.} \end{cases}$$

De transformatie ziet er dus als volgt uit:

$$1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10 \quad \xrightarrow{g} \quad 2\ 3\ 4\ 4\ 6\ 6\ 8\ 8\ 9\ 10.$$

Noem $Y = g(X)$. Er zijn nu twee waarden van X ($X = 3$ en $X = 4$) die allebei de waarde $Y = 4$ opleveren. De kans op $Y = 4$ is daarom $2 \times \frac{1}{10} = \frac{1}{5}$. We vinden de verdeling

$$\begin{aligned} \mathsf{P}\{Y = 2\} = \mathsf{P}\{Y = 3\} &= \frac{1}{10}, \\ \mathsf{P}\{Y = 4\} = \mathsf{P}\{Y = 6\} = \mathsf{P}\{Y = 8\} &= \frac{1}{5}, \\ \mathsf{P}\{Y = 9\} = \mathsf{P}\{Y = 10\} &= \frac{1}{10}. \end{aligned}$$

Is de stochastische grootheid Y uniform verdeeld?

Neen, want de mogelijke waarden voor Y zijn $\{2, 3, 4, 6, 8, 9, 10\}$, maar deze waarden hebben niet alle dezelfde kans.

Hoe kun je nu trekkingen uit Y simuleren? Onderstaand voorbeeld illustreert dat.

```
>>x=unidrnd(10,1,20);
 % x bevat een realisatie van 20 o.o. trekkingen uit de getallen 1 t/m 10
>>y=x+isprime(x)+(x==1);
 % isprime is een boolean vector die de waarde 1 geeft als het
     corresponderende element van x een priemgetal (dus ongelijk 1) is.
 % (x==1) is een boolean vector die 1 is als het corresponderende
     element van x gelijk is aan 1.
```

1.5 Discrete stochastische grootheden.

We geven stochastische grootheden aan met hoofdletters (X, Y, etc.). Een stochastische grootheid X heet *discreet*, als de mogelijke waarden deel zijn van een aftelbare deelverzameling van de reële getallen $\mathbf{R}$. De verdeling van X kunnen we dan beschrijven door de kansen op de mogelijke waarden op te sommen. Als $\{w_1, w_2, \ldots\}$ de mogelijke waarden van X zijn, dan geldt altijd dat

$$\sum_j \mathsf{P}\{X = w_j\} = 1.$$

1.6 De verdelingsfunctie.

De (cumulatieve) verdelingsfunctie F van een stochastische grootheid X is

$$\mathsf{F}(x) = \mathsf{P}\{X \leq x\}, \quad x \in \mathbf{R}.$$

Als X een discrete stochastische grootheid is, met mogelijke waarden $\{w_1, w_2, \ldots\}$, dan geldt dus

$$\mathsf{F}(x) = \sum_{w_j \leq x} \mathsf{P}\{X = w_j\}, \quad x \in \mathbf{R}.$$

Merk op dat F een stijgende trapfunctie is, met sprongen in de punten w_j. Deze is rechts-continu **(waarom?)**.

Laten we veronderstellen dat de waarden in oplopende volgorde genummerd zijn: $w_1 < w_2 < \ldots$. Dan

$$\mathsf{P}\{X = w_j\} = \mathsf{F}(w_j) - \mathsf{F}(w_{j-1}),\ j = 1, 2, \ldots.$$

(Hierbij nemen we voor w_0 (het geval $j = 1$) een willekeurig getal kleiner dan de kleinste waarde w_1.) M.a.w., gegeven de verdelingsfunctie F, dan kunnen we de verdeling van X (de opsomming van de kansen) weer terugvinden.

De verdelingsfunctie F geeft dus een complete beschrijving van de verdeling van X. Soms wordt F dan ook kortweg de *verdeling* genoemd. (In het geval van discrete stochastische grootheden is de beschrijving d.m.v. F misschien niet zo interessant. In het geval van continue stochastische grootheden (zie verderop) speelt de verdelingsfunctie een grotere rol.)

N.B. Om verwarring te voorkomen, hangen we de stochast zelf soms als sub-index aan de F en P-symbolen.

Voorbeeld 1.6 Kies de stochast Y gedefinieerd in Voorbeeld 1.5. Dan

$$\mathsf{F}_Y(x) = \begin{cases} 0, & x < 2, \\ 1/10, & 2 \leq x < 3, \\ 1/5, & 3 \leq x < 4\ , \\ 2/5, & 4 \leq x < 6, \\ 3/5, & 6 \leq x < 8, \\ 4/5, & 8 \leq x < 9, \\ 9/10, & 9 \leq x < 10, \\ 1, & x \geq 10. \end{cases}$$

Als de mogelijke waarden gegeven zijn, is het wat overzichtelijker om F alleen aan te geven in deze waarden:

$$\mathsf{F}_Y(2) = \frac{1}{10},\ \mathsf{F}_Y(3) = \frac{1}{5},\ \mathsf{F}_Y(4) = \frac{2}{5},\ \mathsf{F}_Y(6) = \frac{3}{5},\ \mathsf{F}_Y(8) = \frac{4}{5},\ \mathsf{F}_Y(9) = \frac{9}{10},\ \mathsf{F}_Y(10) = 1.$$

Dit is dus een cumulatieve weergave van de kansen.

1.7 Steekproef.

Stel $X_1, \ldots, X_n$ zijn onderling onafhankelijke stochastische grootheden, die alle dezelfde verdeling hebben. Ze hebben dan alle dezelfde verdelingsfunctie F:

$$\mathsf{P}\{X_i \leq x\} = \mathsf{F}(x),\ x \in \mathbf{R},\ \text{voor alle } i = 1, \ldots, n.$$

We noemen $X_1, \ldots, X_n$ een *steekproef* (uit (de verdeling) F). We zeggen ook wel dat $X_1, \ldots, X_n$ een steekproef is uit X, waarbij X verdeling F heeft (n o.o. *copieën* van een *populatiegrootheid* X).

1.8 Wet van de grote aantallen.

Laat $X_1, \ldots, X_n$ een steekproef zijn uit X, $n \geq 1$. Dan geldt voor iedere verzameling A:

$$\lim_{n\to\infty} \frac{\{\text{aantal } X_i \text{ in } A,\ i \leq n\}}{n} = \mathsf{P}\{X \in A\}.$$

In woorden: de fractie waarnemingen die in de verzameling A terecht komt is ongeveer gelijk aan de kans op die verzameling. We illustreren dit met de volgende simulaties, waarbij we weer de stochast Y uit voorbeeld 1.5 gebruiken. De eerste simulatie 'schat' geeft de fractie van het aantal waarnemingen i voor $i = 1, \ldots, 10$.

```
>>n=1000;
 % doe 1000 aselecte trekkingen
>>x=unidrnd(10,1,n);
>>y=x+isprime(x)+(x==1);
>>h=hist(y,1:10)/n;
 % het i-de element van h bevat de fractie waarnemingen i, voor i=1,...,10
>>idx=find(h==0);
>>h(idx)=nan;
 % de waarnemingen die niet voorkomen, worden niet meegeplot
>>bar(h,'y')
 % maakt staafdiagram van de gevonden fractie in geel
>>hold on
 % we willen er nog wat bij plotten
>> f=[0 0.1 0.1 0.2 0 0.2 0 0.2 0.1 0.1];
 % f bevat de kansen oftewel de limiet fracties
>>idx=find(f==0);
>>f(idx)=nan;
>>plot(f,'ob')
 % plot nu de f-waarden met blauwe rondjes.
```

De volgende simulatie geeft de voortschrijdende fractie van het aantal waarnemingen van één van de getallen 2,4,6. Hierbij is y al bepaald.

```
>>hits=cumsum(y==2|y==4|y==6);
 % dit telt het aantal waarnemingen 2,4,6 van de eerste k trekkingen, k=1,...,n
>>hits=hits./[1:n];
 % de punt operator deelt het k-de element door k.
```

Waar moet dit op gaan lijken en klopt dat?

1.9 Gemiddelde.

Het gemiddelde van een rij getallen $x_1, \ldots, x_n$ is gegeven door

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Het gemiddelde van n stochastische grootheden $X_1, \ldots, X_n$ is de stochastische grootheid

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Zelfs als de stochasten X_i o.o. zijn, dan nog is het berekenen van de verdeling van $\bar{X}$ vaak niet eenvoudig. In de praktijk poogt men een idee te krijgen van deze verdeling door simulatie.

We nemen in het onderstaande 10 000 simulaties van het gemiddelde van 2, 10, en 40 aselecte trekkingen uit $\{1, \ldots, 10\}$.

```
>>x=unidrnd(10,40,10000);
>>size(x)
   ans=
       40 10000
 % controleert of we inderdaad een matrix met 40 rijen en 10000 kolommen hebben gevuld
>>s2=sum(x([1 2],:))/2;
 % middelt de eerste twee elementen van elke kolom, dit geeft een vector met 10000 elementen,
   nl. de 10000 simulaties van het gemiddelde van 2 aselecte trekkingen.
>>hist(s2)
>>s10=sum(x([1 10],:))/10;
>>s40=mean(x);
 % hier middelen we over hele kolommen
```

Voor steeds groter wordende waarden van n convergeert het gemiddelde naar een getal, volgens de wet van de grote aantallen (bijv. mits de X_i o.o. zijn en eindige verwachting hebben- zie Hoofdstuk 4), namelijk de verwachting oftewel gemiddelde waarneming.

De onderstaande simulatie bepaalt het gemiddelde van n aselecte trekkingen uit $\{1, \ldots, 10\}$, voor $n = 1, ..100000$. Waar moet dit op gaan lijken en lijkt het er inderdaad redelijk op?

```
>>x=unidrnd(10,1,100000);
>>x=cumsum(x);
 % we nemen de cumulatieve sommen, d.w.z. het k-de element is de som van de eerste
   k getallen
>>x=x./[1:100000];
 % middelen doen je met de .operator
>>plot(x,'r')
 % plot een rode lijn
```

1.10 De centrale limiet stelling.

Volgens de wet van de grote aantallen convergeert het gemiddelde van een steekproef van grootte n naar een *getal*, voor $n \to \infty$. De *centrale limietstelling* specifeert de afwijking van dit getal, door de uitspraak dat het gemiddelde van de steekproef ongeveer *normaal* verdeeld is, als n groot is. Nu hebben we nog niet gedefinieerd wat we bedoelen met de normale verdeling (zie Hoofdstuk 3). Het komt er ongeveer op neer dat de verdeling van $\bar{X}$ altijd dezelfde *klokvorm* krijgt. Het doet er niet toe uit welke verdeling de steekproef is getrokken (als deze maar eindige *variantie* (zie Hoofdstuk 4) heeft), de klokvorm komt altijd weer terug. Dit is een van de redenen waarom de normale verdeling zo'n belangrijke rol speelt. Laten we dit eens bekijken met een voorbeeld. Eerst nemen we een steekproef uit X, met X uniform verdeeld op $\{1, \ldots, 10\}$.

```
function g = gauss(x,m,s)
% g=gauss(x,m,s) calculates the prob density of X in N(m,s^2)
g=exp(-0.5*((x-m)/s).^2)/(s*sqrt(2*pi));

% eerst definieren we de functie gauss die de normale verdeling
%  (zit ook standaard in de Stats-toolbox van Matlab)
% deze moet je in de Matlab Editor maken en opslaan in de Matlab work directory.
```

```
>>n=10000;
>>x=unidrnd(10,100,10000);
>>m=sum(x)/100;
 % m is een vector met 10 000 gemiddelden van 100 trekkingen uit x
>>hist(m)
>>hold on
 % nu de goede normale verdeling erover heen plotten, met m=5.5 en
   s=sigma(X)/sqrt(100)=0.28723
 % de variabele r doorloopt met stapjes ter grootte 0,01 het interval [4,7]
>>[u,t]=hist(m);
>>r=4:0.01:7;
>>plot(r,10000*(t(2)-t(1))*gauss(r,5.5,0.28723))
 of: >>plot(r,10000*(t(2)-t(1))*normpdf(r,5.5,0.28723))
```

Voorbeeld 1.7 Stel er zijn 1 miljoen lotto-getallen. Persoon A koopt één lot. De kans dat A de hoofdprijs wint is dan één op miljoen: laat X het nummer van de hoofdprijs zijn, en x het nummer dat A getrokken heeft, dan

$$\mathsf{P}\{X = x\} = \frac{1}{1000000}, \quad x \in \{1, \ldots, 1000000\}.$$

De verdelingsfunctie is

$$\mathsf{F}(x) = \frac{x}{1000000}, \quad x \in \{1, \ldots, 1000000\}.$$

Bij de uniforme verdeling op de getallen $\{1, \ldots, m\}$, wordt de kans op een getal $(= 1/m)$ steeds kleiner als het aantal mogelijkheden $(= m)$ groter wordt. Bij grote m is een herschaling handig.

Voorbeeld 1.8 Beschouw bovenstaand voorbeeld, maar deel alle getallen door $1000000 = 10^6$, oftewel $X \mapsto X \times 10^{-6}$. Dan bezit deze herschaalde X de uniforme verdeling op $\{1 \times 10^{-6}, 2 \times 10^{-6}, \ldots, 1\}$, zodat

$$\mathsf{P}\{X = x\} = 10^{-6}, \quad x \in \{1 \times 10^{-6}, 2 \times 10^{-6}, \ldots, 1\}.$$

De verdelingsfunctie is nu

$$F(x) = x, \ x \in \{1 \times 10^{-6}, 2 \times 10^{-6}, \ldots, 1\}.$$

1.11 De uniforme verdeling op $[0, 1]$.

Stel dat we blindelings een getal tussen 0 en 1 kiezen. Noem het resultaat X. Dan bezit X de (continue) uniforme verdeling op het interval [0,1]. De verdelingsfunctie van X is

$$F(x) = x, \text{ voor alle } x \in [0, 1].$$

Dit is de limiet van de (discrete) uniforme verdeling op $\{1/m, 2/m, \ldots, 1\}$, met $m \to \infty$. Als X uniform verdeeld is op $[0, 1]$ kan X **alle** waarden in het interval $[0, 1]$ aannemen, d.w.z. er zijn oneindig veel (zelfs overaftelbaar veel) mogelijke waarden. Alle mogelijke waarden hebben bovendien dezelfde kans, namelijk kans nul! Het is daarom vaak meer zinvol om in plaats van over de *kans* op een waarde, te spreken over de *aannemelijkheid* van een waarde. Bij de uniforme verdeling op $[0, 1]$ is de aannemelijkheid van alle $x \in [0, 1]$ gelijk, en wel gelijk aan één. We definiëren de dichtheid $f(x)$ van X als de afgeleide van de verdelingsfunctie

$$f(x) = 1, \text{ voor alle } x \in [0, 1].$$

De dichtheid $f(x)$ in het punt x wordt dan ook wel de aannemelijkheid van de waarde x genoemd. Er geldt voor $0 \leq s < t \leq 1$

$$P(s \leq X \leq t) = P(X \leq t) - P(X \leq s) = F(t) - F(s) (= \int_s^t f(x)dx) = t - s.$$

M.a.w., de kans op het interval $[s, t]$ is gelijk aan de lengte $t - s$.

1.12 Afronden.

Stel X is uniform verdeeld op $[0,1]$. We ronden een meting van X nu naar boven af, tot op 6 cijfers achter de komma, en wel als volgt: we nemen het kleinste gehele getal dat groter of gelijk is aan $X \times 10^6$. Laten we dit getal Y noemen. Dan bezit Y de (discrete) uniforme verdeling op $\{1, \ldots, 10^6\}$. In Matlab hebben we discrete uniforme verdelingen geconstrueerd door uit te gaan van de continue uniforme verdeling (de laatste zit standaard in de Stats Toolbox van Matlab):

```
>>x=ceil(m*rand(1,n));
 % dit levert n o.o. trekkingen uit de getallen 1,...,m
```

Afronding in Matlab kan op verschillende manieren: `floor` rond naar beneden af, `fix` gooit het gedeelte achter de komma weg, `round` rond af naar het dischtsbijzijnde gehele getal (probeer zelf wat hij met 0,5 doet).

1.13 Lineaire transformaties.

Stel U is uniform verdeeld op $[0,1]$, en noem $X = a + bU$, met $b > 0$. Dan is X uniform verdeeld op het interval $[a, a+b]$. De verdelingsfunctie van X is

$$\mathsf{F}(x) = \frac{x-a}{b}, \ x \in [a, a+b],$$

met dichtheid

$$f(x) = \frac{1}{b}, \ x \in [a, a+b].$$

Verder geldt voor $a \leq s < t \leq a+b$,

$$\mathsf{P}\{s \leq X \leq t\} = \frac{t-s}{b} = \frac{\text{lengte subinterval}}{\text{lengte totale interval}}.$$

Hoe kun je dus nu een trekking uit de verdeling van X simuleren?

1.14 Andere transformaties.

Stel U is uniform verdeeld op $[0,1]$, en zij $X = g(U)$ met g een gegeven niet-lineaire functie. Dan is X niet meer uniform verdeeld.

Voorbeeld 1.9 Neem $g(u) = u^2$, d.w.z. $X = U^2$. De verdelingsfunctie van X wordt nu

$$\begin{aligned} \mathsf{F}_X(x) &= \mathsf{P}\{X \leq x\} = \mathsf{P}\{g(U) \leq x\} \\ &= \mathsf{P}\{U^2 \leq x\} = \mathsf{P}\{U \leq \sqrt{x}\} = \sqrt{x}, \quad x \in [0,1]. \end{aligned}$$

De dichtheid van X is

$$f_X(x) = \frac{dF_X(x)}{dx} = \frac{d\sqrt{x}}{dx} = \frac{1}{2\sqrt{x}}, \ x \in (0,1].$$

(In $x = 0$ is de dichtheid niet gedefinieerd, want daar bestaat de afgeleide van de verdelingsfunctie niet.)

We kunnen weer m.b.v. een histogram kijken of de dichtheid er inderdaad zo uit ziet. We nemen $m = 10000$ simulaties.

```
>>u=rand(1,10000);
 % 10000 trekkingen uit de homogene verdeling
>>x=u.^2;
>>hist(x,30)
 % plot een histogram met 30 bins, dus binbreedte is 1/30
>>r=0.001:0.001:1;
 % maak een vector r, met elementen 0,001, 0,002,...,1;
   je kunt niet bij 0 beginnen omdat dichtheid niet bestaat in dat punt
>>f=2*sqrt(r);
```

```
 % eerst worteltrekken en met 2 vermenigvuldigen
>>f=f.^{-1};
 % omgekeerde nemen
>>plot(r,(1000/3)*f,'r')
 % f opblazen met aantal waarnemingen vermenigvuldigd met binbreedte.
```

1.15 De empirische verdelingsfunctie.

Laat $X_1, \ldots, X_n$ een steekproef uit de verdeling F zijn. We noemen

$$\mathsf{F}_n(x) = \frac{\{\#X_i \leq x,\ i \leq n\}}{n}, \quad x \in \mathbf{R}$$

de *empirische verdelingsfunctie*. Volgens de wet van de grote aantallen geldt voor alle x

$$\lim_{n\to\infty} \mathsf{F}_n(x) = \mathsf{P}\{X_1 \leq x\} = \mathsf{F}(x).$$

1.16 Opgaven Hoofdstuk 1

Opgave 1.1 Men gooit tweemaal met een dobbelsteen. Bereken de kans op de volgende gebeurtenissen:

a) de hoogste worp levert 5 ogen;

b) de laagste worp levert 5 ogen,

c) de aantallen ogen zijn gelijk.

Opgave 1.2 Hoe vaak moet men gemiddeld met een dobbelsteen gooien, totdat men alle aantallen ogen gehad heeft? (Gebruik de computer en simuleer!)

Opgave 1.3 Laat X een aselecte trekking zijn uit de getallen $\{1, \ldots, 7\}$, en zij $Y = (X-4)^2$. Bepaal $\mathsf{P}\{Y = y\}$ voor alle mogelijke waarden van y. Bepaal de verdelingsfunctie van Y.

Opgave 1.4 Laat X en Y o.o. aselecte trekkingen zijn uit $\{1, \ldots, 7\}$. Hoe ziet de verdeling van $Z = X + Y$ er uit? Doe een simulatie, d.w.z. neem een steekproef $Z_1, \ldots, Z_m$ uit Z en maak een histogram.

Opgave 1.5 Stel U is uniform verdeeld op $[0, 1]$ en zij $X = -\log U$, met log de natuurlijke logaritme. Bepaal de verdelingsfunctie van X. (Dit noemt men de standaard exponentiële verdeling.)

Doe vervolgens een simulatie van 10 000 aselecte trekkingen uit deze verdeling en plot het histogram. Plot in hetzelfde plaatje ook de dichtheid f van X (na geschikt 'opblazen').

Ga experimenteel na wat de verwachting van X zal zijn door gebruik van de wet van de grote aantallen.

Opgave 1.6 Herhaal de simulatie in paragraaf 1.10 van het gemiddelde van $n = 2$, 3, 12 en 40 aselecte trekkingen uit $\{1, 2, \ldots, 10\}$. Voor welke waarden van n heeft al enigszins een klokvorm? Hierbij moet je het getal $0,28723$ vervangen door $2,8723/\sqrt{n}$.

Opgave 1.7 Voor X uniform verdeeld op $\{1, \ldots, 10\}$, definiëren we nu $Y = X^2 - X$. In Matlab kun je dat doen door de zogenaamde punt-operator te gebruiken:

```
>>y=x.^2-x;
```

Dan heeft het gemiddelde (verwachting) waarde $\mu = 33$ en de standaarddeviatie $\sigma = 29,627$ Doe de simulatie analoog aan de simulatie in Opgave 1.6 voor het gemiddelde van aselecte trekkingen uit de verdeling van Y.

Opgave 1.8 Neem een steekproef $X_1, \ldots, X_n$ van grootte $n = 12$ uit $X = -\log U$ met U uniform verdeeld. Doe dit 10000 keer en maak een histogram van de 10000 zo verkregen gemiddelden $\bar{X} = \sum_{i=1}^{n} X_i/n$.

Heeft dit al enigszins een klokvorm? Dit kun je bekijken door de normale verdeling met $m = 1$ en $s = 1/\sqrt{12}$ op de juiste manier op te blazen en in het histogram te plotten.

2 Kansruimtes

2.1 Wat is een kans?

In het dagelijks taalgebruik komt men het begrip *kans* regelmatig tegen. Dit heeft te maken met onzekerheid over, oftewel gebrek aan informatie (dit kan ook onstaan door de complexiteit van het te beschouwen systeem); een zekere mate aan onvoorspelbaarheid wat betreft de gevolgen van een handeling; het feit dat handelingen (waarnemingen, metingen) niet identiek herhaalbaar zijn (onzekerheid is inherent aan de natuur oftewel God dobbelt wél!), etc. Typische voorbeelden zijn de volgende.

Voorbeeld 2.1 Het gebruik van veiligheidsgordels doet de kans op een ongeluk met dodelijke afloop afnemen.

Voorbeeld 2.2 De kans van slagen van een experiment is groter als de proef door deskundigen wordt verricht.

In de dagelijkse praktijk komt het begrip kans overeen met *fractie*, *frequentie*, of *percentage*.

Voorbeeld 2.3 Bij herhaald gooien met een zuivere munt zal de *fractie* van het aantal keren dat munt valt, uiteindelijk ongeveer $1/2$ zijn (ten gevolge van de wet van de grote aantallen). Met andere woorden: de empirische kans op munt by n keer gooien is ongeveer $1/2$, voor n groot:

$$\frac{\{\text{aantal keren munt in } n \text{ worpen}\}}{n} \approx \frac{1}{2}.$$

Voorbeeld 2.4 De uitspraak dat in Nederland de kans op een baan gelijk is aan x, betekent niets anders, dan dat $x\%$ van de mensen in Nederland tussen de 18 en 65 momenteel een baan heeft.

Voorbeeld 2.5 Gegeven is dat van een partij van 1000 lampen, er 265 defect zijn, maar er is niet bekend welke dat zijn. De uitspraak dat de kans op een defecte lamp $0,265$ is, kun je op verschillende manieren interpreteren. Enerzijds betekent dat een willekeurige partij van m lampen *gemiddeld* $0,265 \times m$ kapotte zal bevatten. Anderzijds kun je de uitspraak zien als een verhouding van het aantal lampen met een bepaalde (in dit geval ongewenste) eigenschap en het totaal aantal lampen.

De laatste twee voorbeelden zijn typisch gevalleen waarbij het niet gaat om *herhaling* van een experiment. Dit vraagt om een abstracte omschrijving van het begrip kans. Het idee is (zoals bij de meeste wiskundige theorieën) om een aantal z.g. axioma's op te stellen waaraan een kans moet voldoen, en wel zodanig dat de eigenschappen die uit de axioma's volgen ongeveer voldoen aan een intuïtief idee van kans en consistent met de 'praktijk van alledag'.

2.2 Uitkomstenruimte en kansmaat

Voor een beschrijving van een kansexperiment, moeten we de verzameling van mogelijke uitkomsten karakteriseren, waarbij de op dat moment beschikbare informatie wordt gebruikt. Deze verzameling noemen we de *uitkomstenruimte* Ω. Als werkhypothese gaan we er vanuit dat de mogelijke uitkomsten voor de waarnemer identificeerbaar zijn (dit hoeft formeel niet, maar dan wordt de verdere constructie gecompliceerder).

Herhaalde experimenten zijn onafhankelijke uitvoeringen van hetzelfde experiment. Deze tezamen vormen weer een experiment met een gecompliceerdere uitkomstenruimte. In dit geval kan men spreken over de frequentie van een gebeurtenis (de empirische kans op):

$$\frac{\{n(A) = \text{aantal keren dat gebeurtenis } A \text{ optreedt bij } n \text{ herhalingen}\}}{n}.$$

Voorbeeld 2.6 Experiment: eenmalig gooien met een dobbelsteen. Uitkomstenruimte $\Omega = \{1, 2, 3, 4, 5, 6\}$. Herhaald experiment: n keer gooien met een dobbelsteen. Empirisch is vastgesteld dat: $n(\{4, 6\})/n \approx 1/3$ voor n groot (wet van de grote aantallen). Welke uitkomstenruimte hoort bij dit experiment? Welke bij het oneindig vaak gooien?

Nu plakken we de zijden van de dobbelsteen af: met rood voor de zijden met 1,2,3 of 4 ogen en de overige zijden met blauw. We doen hetzelfde experiment. Uitkomstenruimte $\Omega' = \{\text{rood, blauw}\}$. De kans op elke uitkomst kun je afleiden uit de kansen voor de onafgeplakte dobbelsteen.

Voorbeeld 2.7 (College Evolutionaire Algoritmen) Bij Voorbeeld 1.3 is het experiment: de nieuwe mutatie gegeven de oude. De uitkomstenruimte is $\Omega = \{0,1\}^n$, de collectie binaire rijtjes van lengte n.

Voorbeeld 2.6 suggereert dat Ω niet altijd eindig of zelfs maar aftelbaar hoeft te zijn.

Voorbeeld 2.8 Een binair getal ω tussen 0 en 1 kan men schrijven als $X = 0.\omega_1\omega_2\omega_3\ldots$ met $\omega_i \in \{0,1\}$. Wat is dus Ω? Bekijk nu

$$n(\{1\})/n = \text{ de fractie énen in de eerste } n \text{ digits.}$$

Het blijkt dat als X een *willekeurig* gekozen getal tussen 0 en 1 is (d.w.z. als X uniform verdeeld is op $[0,1]$), dan $\lim_{n\to\infty} n(\{1\})/n = 1/2$.

Hier kun je ook anders tegenaan kijken. Stel je springt op de natuurlijke getallen: een sprongetje naar rechts (+1) met kans 1/2, en een sprongetje op de pas (0) met kans 1/2 (soort half-dronkemanswandeling). Dan is de positie na n sprongen precies $n(\{1\})$. Je bent op den duur halverwege van waar je had kunnen zijn! Waar zou je op den duur zijn als je met kans 1/3 een sprongetje naar rechts had gemaakt?

Uitkomst en gebeurtenis. Elementen van Ω stellen (enkelvoudige) *uitkomsten* voor. Vaak zijn we in samengestelde uitkomsten geïnteresseerd (vaak hebben die een bepaalde eigenschap gemeen): een *gebeurtemis* (eventualiteit, Engels: event) is een deelverzameling van Ω.

De collectie eventualiteiten leggen in feite een structuur op de uitkomstenverzameling Ω, maar daar zullen hier niet op in gaan.

Voorbeeld 2.9 Experiment: het werpen van een dobbelsteen. Gebeurtenis A is gegeven door

$$A = \text{het werpen van een oneven getal} = \{1,3,5\}.$$

In Voorbeeld 2.6 heb je een relatie tussen de twee uitkomstenruimtes Ω en Ω': gebeurtenis $\{1,2,3,4\}$ correspondeert met uitkomst $\{\text{rood}\}$, en gebeurtenis $\{5,6\}$ met uitkomst $\{\text{blauw}\}$.

De volgende voorbeelden gaan over rijtjes of woorden van n verschillende symbolen of letters.

Voorbeeld 2.10 Trekken met teruglegging oftewel herhaald experiment Gegeven n verschillende symbolen $\ell_1,\ldots,\ell_n$ oftewel een alfabet met n letters.

i) Experiment: kies willekeurig een één-letterwoord. Uitkomstenruimte $\Omega = \{\ell_1,\ldots,\ell_n\}$.

ii) Experiment: herhaal experiment (i) m keer. Dit komt neer op het kiezen een willekeurig woord van m letters uit het n-alfabet $\{\ell_1,\ldots,\ell_n\}$.

De uitkomstenruimte is nu de collectie Ω^m van alle m-letterwoorden uit het n-alfabet.

iii) Experiment: het totaal aantal letters ℓ_1 in een willekeurig gekozen m-woord, het aantal letters symbolen ℓ_2, etc. (komt neer op de frequentie).
De uitkomstenruimte is dan

$$\tilde{\Omega} = \{(m_1,\ldots,m_n) \,|\, 0 \le m_i \le m, i = 1,\ldots,n, m_1 + \cdots m_n = m\}$$

Je kunt nu symbolen of letters een label geven en gaan kijken naar hoe vaak welke labels voorkomen. Label bijvoorbeeld de eerste r letter $b_1,\ldots,b_r$ rood en de rest blauw.

iv) Experiment: welke kleur label tref je bij één trekking? Het eerste experiment heeft nu uitkomstenruimte $\Omega' = \{\text{rood},\ \text{blauw}\}$.

v) Experiment: herhaal experiment (iv) m keer. Dit is hetzelfde als de reeks van m successievelijke labels van een willekeurig m-letterwoord waarnemen. Uitkomstenruimte is nu $(\Omega')^m$: de rijtjes van m rode en blauwe labels, d.w.z. je maakt zo m-woorden uit een 2-alfabet. Het verschil is echter dat hetzelfde m-letterwoord uit het 2-alfabet afkomstig kan zijn van *verschillende* m-letterwoorden uit het n-alfabet! Dit heeft een gevolg voor de waarde van de kansen!

vi) Experiment: het aantal rode en blauwe labels in de m trekkingen. Uitkomstenruimte is nu $\tilde{\Omega}' = \{(i,j) \mid 0 \leq i,j \leq m, i+j=m\}$, waarbij i het aantal rode en j het aantal blauwe labels is.

De ruimte Ω^m van alle m-letterwoorden uit het n-alfabet is de onderliggende ruimte, die de prettige eigenschap heeft dat alle woorden een even grote kans hebben. Deze bepaalt de kansverdelingen op de 'afgeleide' uitkomstenruimtes.

We geven een getallenvoorbeeld. We kiezen het alfabet $\{a,b,c\}$, d.w.z. $n=3$. We doen 3 herhalingen van het experiment waarbij we één letter trekken, d.w.z. $m=3$ en de mogelijke uitslagen zijn alle 3-letterwoorden.

$$
\begin{array}{lcl}
aaa & \leftrightarrow & (3,0,0) \\
bbb & \leftrightarrow & (0,3,0) \\
ccc & \leftrightarrow & (0,0,3)
\end{array}
\qquad
\left.\begin{array}{l} aab \\ aba \\ baa \end{array}\right\} \leftrightarrow (2,1,0)
\qquad
\left.\begin{array}{l} aac \\ aca \\ caa \end{array}\right\} \leftrightarrow (2,0,1)
$$

$$
\left.\begin{array}{l} bba \\ bab \\ abb \end{array}\right\} \leftrightarrow (1,2,0)
\qquad
\left.\begin{array}{l} bbc \\ bcb \\ cbb \end{array}\right\} \leftrightarrow (0,2,1)
\qquad
\left.\begin{array}{l} caa \\ cac \\ acc \end{array}\right\} \leftrightarrow (1,0,2)
\qquad
\left.\begin{array}{l} ccb \\ cbc \\ bcc \end{array}\right\} \leftrightarrow (0,1,2)
$$

$$
\left.\begin{array}{l} abc \\ acb \\ bac \\ bca \\ cab \\ cba \end{array}\right\} \leftrightarrow (1,1,1)
$$

Labellen we nu a en b met blauw (B) en c met rood (R), dan krijgen we de 3-letterwoorden uit het $\{R,B\}$-alfabet:

$$
\left.\begin{array}{l} aaa \\ bbb \\ aab \\ aba \\ baa \\ bba \end{array}\right\} \leftrightarrow RRR
\qquad
\left.\begin{array}{l} bab \\ baa \\ aac \\ abc \\ bbc \\ bac \end{array}\right\} \leftrightarrow RRB
$$

$$
\left.\begin{array}{l} aca \\ acb \\ bcb \\ bca \end{array}\right\} \leftrightarrow RBR
\qquad
\left.\begin{array}{l} caa \\ cab \\ cbb \\ cba \end{array}\right\} \leftrightarrow BRR
$$

$$
\left.\begin{array}{l} cca \\ ccb \end{array}\right\} \leftrightarrow BBR
\qquad
\left.\begin{array}{l} cac \\ cbc \end{array}\right\} \leftrightarrow BRB
$$

$$
\left.\begin{array}{l} caa \\ cab \\ cba \\ cbb \end{array}\right\} \leftrightarrow RBB
\qquad
\left.\begin{array}{l} ccc \end{array}\right\} \leftrightarrow RRR
$$

Voorbeeld 2.11 Trekken zonder teruglegging oftewel woorden bestaande uit *verschillende* letters
Gegeven n verschillende symbolen $\ell_1, \ldots, \ell_n$ oftewel een alfabet met n letters.

i) Experiment: kies willekeurig een één-letterwoord. Uitkomstenruimte $\Omega = \{\ell_1, \ldots, \ell_n\}$.

ii) Experiment: kies een willekeurig woord van m *verschillende* letters uit het n-alfabet. De uitkomstenruimte is nu de collectie Ω^m van alle m-letterwoorden van verschillende letters uit het n-alfabet. $ell_1 \ldots, ell_n$.

iii) Experiment: het totaal aantal letters ℓ_1 in een willekeurig gekozen m-woord, het aantal letters symbolen ℓ_2, etc. (komt neer op de frequentie).
De uitkomstenruimte is dan

$$\tilde{\Omega} = \{(m_1, \ldots, m_n) \,|\, 0 \leq m_i \leq 1, i = 1, \ldots, n, m_1 + \cdots m_n = m\}.$$

Je kunt nu weer symbolen of letters een label geven en gaan kijken naar hoe vaak welke labels voorkomen. Evenala in Voorbeeld 2.10 is de ruimte Ω^m weer bepalend voor de afgeleide uitslagenruimtes, omdat in Ω^m alle uitkomsten even waarschijnlijk zijn bij aselecte trekking.

Bovenstaand getallenvoorbeeld met $n = m = 3$ wordt in het geval van trekken zonder teruglegging, oftewel woorden met allemaal verschillende letters, heel veel gereduceerd: $\Omega^3 = \{abc, acb, bac, bca, cba, cab\}$!

Voorbeeld 2.12 Verdelen van n verschillende symbolen over m dozen
Dit is terug te voeren op Voorbeeld 2.10.

Gegeven n symbolen $\ell_1, \ldots, \ell_n$. Experiment: verdeel de symbolen over m genummerde dozen. Uitkomstenruimte Ω bestaat uit alle mogelijke toekenningen van doosnummers $d_1, \ldots, d_m$ aan symbolen. Elke toekenning correspondeert dus met een n-letterwoord uit een m-alfabet $\{d_1, \ldots, d_m\}$.

Merk op dat Voorbeeld 2.10 (ii), zogenaamd *trekken met teruglegging uit een vaas met n knikkers voorstelt* waarbij de *volgorde van belang is*; (iii) stelt *trekken met teruglegging voor* waarbij de *volgorde* **niet** *van belang is.* Voorbeeld 2.11 (i) stelt *trekken zonder teruglegging uit een vaas met n knikkers* voor, waarbij de *volgorde van belang is*; en geval (iii) stelt *trekken zonder teruglegging* van k knikkers voor, waarbij de *volgorde* **niet** *van belang is.*

Verzamelingenleer. De gebruikelijke operaties van de verzamelingenleer (doorsnede, vereniging, complement) zijn ook van toepassing op gebeurtenissen:

- $A \cap B$: A door(sneden met) B. Dit is de verzameling van alle elementen die zowel in A als in B zitten. We zeggen ook wel dat gebeurtenissen A en B allebei optreden.
- $A \cup B$: A verenigd met B. Dit is de verzameling van elementen die in A of in B zitten, of in beide. We zeggen ook wel dat gebeurtenis A of B optreedt.
- $\bar{A}$: het complement van A. Dit zijn alle elementen die niet in A zitten. We zeggen ook wel dat de gebeurtenis A niet optreedt.
- Als $B \subset A$, d.w.z. B is een deelverzameling van A. Gebeurtenis B impliceert gebeurtenis A.

Als $A \cap B = \emptyset$, de lege verzameling, dan hebben A en B geen elementen gemeen. We zeggen ook wel dat de gebeurtenissen A en B niet tegelijk kunnen optreden en we noemen A en B disjunct. We noemen $A_1, \ldots, A_m$ disjuncti als A_i en A_j disjunct zijn voor elk paar indices i, j, $i \neq j$. Tenslotte schrijf je:

$$A \setminus B = A \cap \bar{B}.$$

Verder noemt men Ω de zekere gebeurtenis en $\emptyset$ de onmogelijke gebeurtenis.

Het aantal 'gebeurtenissen' kan zeer groot zijn: als Ω N elementen bevat, dan zijn er 2^N veschillende 'gebeurtenissen'. Is Ω niet eindig, maar wel aftelbaar, dan zijn er reeds *overaftelbaar* veel deelverzamelingen. Als bijv. $\Omega = \mathbf{R}^p$, dan is het aantal deelverzamelingen nog veel groter. In dit geval beperkt men zich tot een klasse gebeurtenissen $\mathcal{A}$, die enigszins nette eigenschappen hebben, en die consistent zijn met de informatie die wij over een model hebben.

Uitgaande van de collectie Ω van mogelijke en onderscheidbare uitkomsten, kun je een geloof uitspreken over de kansen op verschillende gebeurtenissen uit de relevante collectie gebeurtenissen $\mathcal{A}$. Dat geloof is o.a. gebaseerd op nadere informatie omtrent je model, of op experimentele gegevens over de frequentie van gebeurtenissen bij een herhaald experiment, zonder dat interne consistentie verloren gaat.

Kansaxiomata We noemen P een *kans(maat)* op de gebeurtenissen in Ω als

i) $\mathsf{P}\{A\} \geq 0$ voor alle gebeurtenissen $A \in \mathcal{A}$;

ii) $\mathsf{P}\{\Omega) = 1$;

iii) $\mathsf{P}\{\cup_{i=1}^{\infty} A_i\} = \sum_{i=1}^{\infty} \mathsf{P}\{A_i\}$, mits de $A_i \in \mathcal{A}$ disjunct zijn.

We zeggen nu dat $\mathsf{P}\{A\}$ de *kans* op gebeurtenis A is. Uit deze axiomata zijn de volgende eigenschappen gemakkelijk af te leiden.

Eigenschappen van een kansmaat

1. $\mathsf{P}\{\emptyset\} = 0$;
2. $\mathsf{P}\{A \cup B\} = \mathsf{P}\{A\} + \mathsf{P}\{B\}$ als A en B disjunct zijn.
3. $\mathsf{P}\{\bar{A}\} = 1 - \mathsf{P}\{A\}$;
4. $\mathsf{P}\{B \setminus A\} = \mathsf{P}\{B\} - \mathsf{P}\{A \cap B\}$;
5. $\mathsf{P}\{A \cup B\} = \mathsf{P}\{A\} + \mathsf{P}\{B\} - \mathsf{P}\{A \cup B\}$;
6. $\mathsf{P}\{A \cup B\} \leq \mathsf{P}\{A\} + \mathsf{P}\{B\}$;
7. als $A \subseteq B$ dan is $\mathsf{P}\{B \setminus A\} = \mathsf{P}\{B\} - \mathsf{P}\{A\}$ en $\mathsf{P}\{A\} \leq \mathsf{P}\{B\}$;
8. $0 \leq \mathsf{P}\{A\} \leq 1$.

Bewijs.

- (1) in axioma (iii) neem je alle $A_i = \emptyset$. Dan zijn de A_i disjunct en $\cup_i A_i = \emptyset$. We krijgen: $\mathsf{P}\{\emptyset\} = \sum_i \mathsf{P}\{\emptyset\}$. Dit kan alleen als er overal 0 staat, d.w.z. $\mathsf{P}\{\emptyset\} = 0$.
- (2) pas axioma (iii) toe met $A_1 = A$, $A_2 = B$ en $A_i = \emptyset$, $i \geq 3$.
- (3) $A \cup \bar{A} = \Omega$ en $A \cup \bar{A} = \emptyset$. We kunnen dus (2) en (ii) toepassen, zodat $1 = \mathsf{P}\{\Omega\} = \mathsf{P}\{A \cup \bar{A}\} = \mathsf{P}\{A\} + \mathsf{P}\{\bar{A}\}$.
- (4) ten eerste geldt $B = (B \cap A) \cup (B \setminus A)$; ten tweede zijn $B \cap A$ en $B \setminus A$ disjunct. Het gestelde volgt nu uit (2).
- (5) er geldt: $A \cup B = (B \setminus A) \cup (A \setminus B) \cup (A \cap B)$, en deze drie gebeurtenissen zijn disjunct. Het volgt nu uit (4).
- (6) volgt uit axioma (i) en uit (5).
- (7) omdat $B = A \cup (B \setminus A)$ en omdat A en $B \setminus A$ disjunct zijn, geldt door toepassing van axioma (i) $\mathsf{P}\{B\} = \mathsf{P}\{A\} + \mathsf{P}\{B \setminus A\} \geq \mathsf{P}\{A\}$.
- (8) omdat $A \subseteq \Omega$ volgt het gestelde uit (7) en axiomata (i) en (ii).

QED

2.3 Combinatoriek en enige discreet-uniforme kansruimten

De vraag is nog steeds *hoe* je in praktische problemen kansen aan gebeurtenissen moet toekennen. Uitganspunt is gewoonlijk de relatieve frequentie waarmee gebeurtenissen zich naar verwachting voordoen. Deze *modelbouw* speelt zich derhalve af op het grensgebied van statistiek en het toepassingsterrein. De meest overzichtelijke situatie is wanneer Ω aftelbaar is. In dat geval is het voldoende aan elk punt van Ω een kans toe te kennen, die aan de kansaxiomata voldoet.

De eenvoudigste situatie is wanneer Ω zelfs eindig is, zeg n elementen bevat, en alle uitkomsten even waarschijnlijk zijn, d.w.z. dat we een aselecte trekking uit Ω doen. Elke *uitkomst* heeft dus kans $1/n$. De kans op een gebeurtenis A is dan

$$\mathsf{P}\{A\} = \text{aantal elementen van } A \cdot \frac{1}{n},$$

en moet je derhalve het aantal elementen van A bepalen om deze kans te weten te komen.

Een vuistregel bij zo'n model is dat: *de kans op gebeurtenis A=aantal corresponderende (gunstige) uitkomsten gedeeld door het totaal aantal uitkomsten.* Dit principe is van toepassing op de voorbeelden uit de vorige paragraaf, mits de juiste 'basis'uitkomstenruimte wordt gebruikt (namelijk diegene, waaruit de trekkingen aselect zijn, d.w.z. elke uitkomst is even waarschijnlijk).

We geven een aantal combinatorische regels, die ook in Discrete Wiskunde zijn behandeld.

(A) Het aantal rijtjes ter lengte k bestaande uit hooguit n verschillende symbolen is n^m.

Dit is dus equivalent aan het aantal woorden van m letters uit een alfabet van n letters.

(B) Het aantal manieren om n symbolen te rangschikken (m.a.w. het aantal permutaties van n symbolen) is $n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2(\times 1) = n!$ (spreek uit: *n faculteit*).

Dit is op te vatten als het aantal woorden bestaande uit n verschillende letters.

(C) (*Zonder herhaling, lettend op de volgorde*) Het aantal geordende m-rijtjes uit n symbolen is gegeven door $A_{n,m} = n \cdot (n-1) \cdot \ldots \cdot (n-m+1) = n!/(n-m)!$. Voor $m > n$ is dit aantal logischerwijze gelijk aan 0. We definiëren $0! = 1$ en dus zijn we voor het geval $m = n$ weer terug in situatie **(B)**.

Dit is op te vatten als het aantal woorden van m *verschillende* letters uit een alfabet van n letters.

(D) (*Zonder herhaling, niet lettend op de volgorde*) Het aantal manieren om een groepje van m symbolen uit n symbolen te kiezen is

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

(spreek uit: *nboven m*). Men noemt $\binom{n}{m}$ een *binomiaal coëfficiënt*.

Dit is op te vatten als het aantal alfabetten van m symbolen dat je uit een alfabet van n symbolen kunt maken. Elk m-alfabet legt een $(n-m)$-alfabet van *niet* gekozen letter vast: dus het aantal m-alfabetten is gelijk aan het aantal $(n-m)$-alfabetten oftewel $\binom{n}{m} = \binom{n}{n-m}$.

Dit krijgen we als volgt uit **(C)**: met een gegeven alfabet van k symbolen kun je volgens **(B)** $m!$ woorden van m verschillende letters maken. Dus elk m-letter alfabet tel je in $A_{n,m}$ $m!$ keer mee; d.w.z. het aantal m-letter alfabetten is $A_{n,m}/m!$.

(E) Het aantal manieren om een n-letterwoord te maken uit een m-alfabet, met n_i letters ℓ_i, $i = 1, 2, \ldots, m$, $n_1 + \cdots + n_m = n$, is gelijk aan

$$\binom{n}{n_1, \ldots, n_m} = \frac{n!}{n_1! \cdot n_2! \cdots n_m!}, \quad n_1 + \cdots + n_m = n.$$

Dit kun je als volgt inzien: de plaatsen in het woord zijn de symbolen $1, \ldots, n$. Ga plaatsen in het woord aan de verschillende letters $1, \ldots, m$ toekennen: voor letter ℓ_1 heb je $\binom{n}{n_1}$ mogelijkheden. Voor letter ℓ_2 heb je nog $n - n_1$ plekken (symbolen) over waaruit je kunt kiezen: hieruit kies je een groepje van n_2 en dat kan op $\binom{n-n_1}{n_2}$ manieren. Zo doorgaand krijg je in het totaal

$$\binom{n}{n_1, n_2, \ldots, n_m} = \binom{n}{n_1} \cdot \binom{n-n_1}{n_2} \cdot \binom{n-n_1-n_2}{n_3} \cdots \binom{n-n_1-n_2-\cdots-n_{m-1}}{n_m} = \frac{n!}{n_1! \cdot n_2! \cdots n_m!}.$$

Men noemt $\binom{n}{n_1,n_2,\cdots,n_m}$ *multinomiaalcoëfficiënten*. Het totaal aantal mogelijke n-letterwoorden uit een m alfabet is dus m^n: voor elke plek binnen het woord heb je m mogelijkheden.

(F) Het aantal mogelijkheden om frequenties toe te kennen aan de m verschillende letters binnen een n-letterwoord is

$$\binom{n+m-1}{m-1} = \binom{n+m-1}{n}.$$

Dit is equivalent aan het aantal mogelijkheden om een n-letterwoord uit een 1-alfabet in m lettergrepen op te splitsen (het aantal letters in de eerste lettergreep correspondeert met het aantal letters ℓ_1 in het woord, etc.)

Een overgang tussen twee lettergrepen geven we aan met |; hiervan zijn er $m-1$. Een opsplitsing van een n-letterwoord uit een 1-alfabet in m lettergrepen correspondeert dan met een $n+m-1$-letterwoord van $m-1$ symbolen | en n symbolen 1 (bijv.), oftewel een toekenning van $m-1$ plekken in het woord aan een doos met label | en n plekken in het woord aan de doos met label 1. Dit kan volgens **(E)** op bovenstaand aantal manieren.

Eigenschappen van binomiaal coëfficiënten.
De *driehoek van Pascal* is

$$\begin{array}{ccccccccc} & & & & 1 & & & & \\ & & & 1 & & 1 & & & \\ & & 1 & & 2 & & 1 & & \\ & 1 & & 3 & & 3 & & 1 & \\ 1 & & 4 & & 6 & & 4 & & 1 \\ & & & & \cdots & & & & \end{array}$$

Op de $(n+1)$-ste rij van de driehoek vindt men de binomiaal coëfficiënten

$$\binom{n}{0}, \binom{n}{1}, \ldots, \binom{n}{k}, \ldots, \binom{n}{n-1}, \binom{n}{n}.$$

Er geldt:

$$\binom{n}{0} = \binom{n}{n} = 1, \qquad \binom{n}{1} = \binom{n}{n-1} = n,$$

en de symmetrie $\binom{n}{k} = \binom{n}{n-k}$. Verder ziet men aan de driehoek van Pascal dat

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}.$$

Het *binomium van Newton* is de formule

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

en het *multinomium van Newton* is

$$(a_1 + a_2 + \cdots a_m)^n = \sum_{\substack{0 \le n_i \le n \\ n_1+\cdots+n_m=n}} \binom{n}{n_1, n_2, \ldots, n_m} a_1^{n_1} a_2^{n_2} \cdots a_m^{n_m}.$$

Dit kun je handig gebruiken voor het afleiden van relaties tussen binomiaal (multinomiaal) coëfficiënten. Bijv. $a = -1$ en $b = 1$ in vullen in het *binomium* geeft:

$$0 = (-1+1)^n = \sum_{k=0}^{n} \binom{n}{k} (-1)^k.$$

Invullen van $a = 1 = b$ geeft:

$$2^n = (1+1)^n = \sum_{k=0}^{n} \binom{n}{k}.$$

Verder relaties zijn:

$$\begin{aligned}
\sum_{k=0}^{n} \binom{n_1}{k}\binom{n_2}{n-k} &= \binom{n_1+n_2}{n} \\
\sum_{k=n}^{N} \binom{k}{n} &= \binom{N+1}{n+1} \\
\sum_{k=0}^{n} \binom{N-k}{r} &= \binom{N+1}{r+1} - \binom{N-n}{r+1}
\end{aligned}$$

Voorbeeld 2.13 Gooi 5 keer met een dobbelsteen (dit is een 5-letterwoord uit een 6-alfabet). Het aantal mogelijke uitkomsten is 6^5 volgens **(A)**. Het aantal mogelijke uitkomsten waarbij elke worp een ander getal oplevert, is 5! volgens **(B)**.

Voorbeeld 2.14 We gooien n keer met een munt. Laat X het aantal keren kruis (K) zijn. Dan is $n - X$ het aantal keren munt (M). De stochast X heeft dan een *binomiale verdeling*. Deze ziet er als volgt uit.

We willen de kans weten op $X = x$ keer K. Een rijtje van n worpen met x keer K correspondeert met een n-letterwoord van x keer K en $(n - x)$ keer M. Daarvan zijn er volgens **(E)** $\binom{n}{x,n-x} = \binom{n}{x}$ van. Noem p de kans op kruis bij één keer gooien ($p = 1/2$ bij een zuivere munt). Dan is de kans op een gegeven n-letterwoord met x K en $(n - x)$ M gelijk aan $p^x(1-p)^{n-x}$. We vinden zo dat de kans op x keer kruis gelijk is aan

$$\mathsf{P}\{X = x\} = \binom{n}{x} p^x (1-p)^{n-x},\ x = 0, 1, \ldots, n.$$

Voorbeeld 2.15 Gevraagd: de kans dat van 4 bridgespelers N,O,Z en W speler N alle azen heeft.

Deze kans wordt gegeven door 'het aantal gewenste mogelijkheden gedeeld door het totaal aantal mogelijkheden'.

Voor het totaal aantal mogelijkheden moet je het aantal mogelijkheden berekenen waarop 52 (verschillende) kaarten over 4 spelers verdeeld kunnen worden. Iedere speler krijgt dus 13 kaarten. Dit komt neer op het bepalen van het aantal 52-letterwoorden met 13 N, 13 O, 13 W, en 13 Z. Volgens **(E)** zijn dit er

$$\binom{52}{13,13,13,13} = \frac{52!}{13!13!13!13!}.$$

Hoeveel verdelingen zijn zo dat de azen bij speler N terecht komt? Er zijn 4 azen. Er zijn dus nog 52-4=48 kaarten te vergeven aan de spelers, waarbij N er 9 krijgt en de overigen 13. Dat komt neer op het berekenen van het aantal 48-letterwoorden, met 9 N, 13 O, 13 W, 13 Z. Volgens **(E)** is dit

$$\binom{48}{9,13,13,13} = \frac{48!}{9!13!13!13!}.$$

De gevraagde kans is dan: $(48!13!)/(52!9!)$.

Wat is nu de kans dat een willekeurige speler alle azen heeft?

Voorbeeld 2.16 Als voorbeeld bekijken we een partij van N chips, waarvan een onbekend aantal, zeg R, kapot is. Definieer $p = R/N$. Dus p is de fractie kapotte chips in de partij. We willen nu iets te weten komen over p, maar het is teveel werk om alle chips in de partij te controleren. We nemen daarom slechts een steekproef van n chips.

Bij al dit modellen uit macro-wereld veronderstel je dat het type objecten dat je bekijkt, onderling onderscheidbaar (identificeerbaar zijn). In ons geval betekent dat dat de chips in feite een nummertje hebben $(c_1, \ldots, c_N)$, waardoor ze herkend kunnen woorden. Op micro-niveau hoeft dat niet per sé zo te zijn, zoals je bij de opgaven zult zien.

Het nemen van een steekproef kan op twee manieren:

(a) Steekproef *met* teruglegging. Trek n keer aselect een chip, noteer of deze chip functioneert, en leg de getrokken chip vervolgens weer terug in de partij. De kans op een kapotte chip bij één keer aselect trekken is dan p. Dus bij n keer trekken is het aantal kapotte chips in de steekproef binomiaal verdeeld:

$$\mathsf{P}\{x \text{ kapotte chips in de steekproef }\} = \binom{n}{x} p^x (1-p)^{n-x},\ x = 0, \ldots, n.$$

Hoe beredeneer je dit via het tellen van rijtjes?

Eerst kijken we naar het resultaat: x kapotte (K) en $n-x$ hele (H) chips correspondeert met een n-letterwoord uit het alfabet $\{K, H\}$, met precies x letters K. Volgens (**E**) zijn er $\binom{n}{x}$ zulke woorden. Elke K ontstond uit R chips $(c_1, \ldots, c_R)$ en elke H uit één der $N-R$ overige. Dus elk dezer woorden correspondeer met $R^x \cdot (N-R)^{n-x}$ trekkingen van chips. Totaal is dus

$$\binom{n}{x} R^x \cdot (N-R)^{n-x}.$$

Deel door het aantal mogelijkheden N^n, dan krijgen we

$$\binom{n}{x} \frac{R^x (N-R)^{n-x}}{N^n} = \binom{n}{x} \Big(\frac{R}{N}\Big)^x \Big(\frac{N-R}{N}\Big)^{n-x} = \binom{n}{x} p^x (1-p)^{n-x}.$$

b) Steekproef *zonder* teruglegging. Trek n keer aselect een chip, en houdt deze apart (we veronderstellen hier dat $n \leq N$).

Wat is nu de kans op x kapotte chips? Dat is weer het 'aantal goede mogelijkheden gedeeld door totaal het aantal mogelijkheden'. Het totaal aantal mogelijkheden om n chips te trekken is in dit geval het aantal woorden van n verschillende letters uit het alfabet $c_1, \ldots, c_N$: volgens **(C)** is dit $N!/(N-n)!$.

Zeg de kapotte chips hebben de nummers $c_1, \ldots, c_R$. Op hoeveel manieren kun je een groepje van x chips uit de nummers $c_1, \ldots, c_R$ trekken en (dus) $n-x$ uit de nummers $c_{R+1}, \ldots, c_N$? **(B)** levert dat je de x chips op $\binom{R}{x}$ manieren trekken en de $n-x$ op $\binom{N-R}{n-x}$ manieren. Van deze gekozen n chips kun je $n!$ woorden maken (op $n!$ manieren rangschikken). In totaal levert dit

$$\binom{R}{x} \cdot \binom{N-R}{n-x} \cdot n!$$

manieren. Derhalve is

$$\mathsf{P}\{x \text{ kapotte chips in de steekproef}\} = \frac{\binom{R}{x}\binom{N-R}{n-x} \cdot n!}{\frac{N!}{(N-n)!}} = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}}.$$

Dit geldt voor $0 \leq x \leq \min(n, R)$, en $0 \leq n-x \leq \min(n, N-R)$. We noemen dit de *hypergeometrische verdeling*.

Volgens de wet van de grote aantallen geldt zowel in geval **a)** als in geval **b)** (met N groot), dat als n groot is de fractie kapotte chips in de steekproef ongeveer gelijk zal zijn aan de fractie kapotte chips in de partij. In die zin geeft de steekproef dus informatie over de onbekende fractie p.

Het lijkt intuïtief aannemelijk, dat met of zonder teruglegging trekken vrijwel geen effect heeft op de grootte van de kans om x kapotte chips aan te treffen in een trekking van n, wanneer de partij maar heel erg groot is. Dit kun je ook bewijzen.

Lemma 2.1 *Gegeven een partij van N chips, waaronder R kapotte. Dan is*

$$\lim_{\substack{N-R\to\infty \\ R\to\infty}} \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}} \frac{1}{\binom{n}{x}\big(\frac{R}{N}\big)^x \big(\frac{N-R}{N}\big)^{n-x}} = 1,$$

oftewel de kans dat x chips kapot zijn van een n-tal chips is vrijwel gelijk, of de trekking nu met of zonder teruglegging is, zolang $N-R$ en R maar groot genoeg zijn.

Bewijs. We schrijven bovenstaand quotiënt uit: dit is gelijk aan

$$\begin{aligned}\frac{R!(N-R)!n!(N-n)!}{x!(R-x)!(n-x)!(N-R-n+x)!N!}\cdot\frac{x!(n-x)!N^n}{n!R^x(N-R)^{n-x}}\\ = \frac{R!}{(R-x)!R^x}\frac{(N-R)!}{(N-R-n+x)!(N-R)^{n-x}}\frac{N^n(N-n)!}{N!}.\end{aligned}$$

Merk op dat

$$\frac{R!}{(R-x)!R^x}=\frac{R(R-1)(R-2)\cdots(R-x+1)}{R\cdot R\cdots R}=\big(1-\frac{1}{R}\big)\cdot\big(1-\frac{2}{R}\big)\cdots\big(1-\frac{x-1}{R}\big)\to 1,\quad R\to\infty.$$

Analoog geldt dit voor de andere termen, zodat het gestelde volgt. QED

2.4 Opgaven Hoofdstuk 2

Opgave 2.1 Een aap zit achter de computer willekeurig letters te typen. Op die manier onstaan afentoe toevallig woorden. Wat duurt langer: het wachten op het woord `informatica` of het wachten op het woord `abracadabra`?

Opgave 2.2 Laat zien dat

$$\begin{aligned}\mathsf{P}\{A\cup B\cup C\} &= \mathsf{P}\{A\}+\mathsf{P}\{B\}+\mathsf{P}\{C\}\\ &\quad -\mathsf{P}\{A\cap B\}-\mathsf{P}\{A\cap C\}-\mathsf{P}\{B\cap C\}\\ &\quad +\mathsf{P}\{A\cap B\cap C\}.\end{aligned}$$

Doe ditzelfde voor $\mathsf{P}\{A\cup B\cup C\cup D\}$.

Opgave 2.3 Welke kans is groter: met vier dobbelstenen in één worp minstes één zes, of met twee dobbelstenen in 24 worpen minstens één dubbelzes? (Probleem van Chevalier de Méré)

Opgave 2.4 Een bericht wordt versleuteld verstuurd, met n mogelijke decodeer-sleutels. De ontvanger weet niet welke decodeer-sleutel en probeert ze één voor één. Bereken de kans dat de achtste poging het bericht juist decodeert.

Opgave 2.5 In een natuurgebied met N zeldzame vogels vangt men er 5 en voorziet deze van een merkteken. Vervolgens laat men deze vogels weer los. Later worden er opnieuw 5 gevangen. Wat is de kans dat 2 van de 5 gemerkt zijn als $N = 12$?

Opgave 2.6 Uit een populatie van N elementen worden er $n(\leq N)$ met teruglegging getrokken. Bereken de kans dat alle n elementen verschillend zijn.

Opgave 2.7 De getallen 1 tot en met n worden in een willekeurige volgorde geplaatst. Bereken de kans dat de getallen $1,\ldots,k$ op volgorde naast elkaar staan.

Opgave 2.8 Een eenvoudig model voor *wachtrijen*. Deze wachtrij wordt op diescrete tijdstippen $t = 0, 1, 2, \ldots$ geobserveerd en per tijdseenheid wordt de eerste klant in de rij bediend met kans p, en arriveert er een nieuwe klant met kans q. Op tijdstip 0 is er één klant in de wachtrij. Bepaal de kansen op $0, 1, 2, 3$ klanten in de wachtrij op tijdstip 2.

Opgave 2.9 Een eenvoudig *genetisch* model. Stel dat de genen in een organisme in tweetallen voorkomen, en dat elk lid van zo'n tweetal ofwel type a ofwel A is. De mogelijke genotypes van een organisme zijn dan AA, Aa en aa (aA en Aa zijn equivalent). Als twee organismen paren, draagt elk van hen onafhankelijk één van zijn genen bij: ieder van de twee genen wordt overgedragen met kans $1/2$.

a) Stel dat de genotypes van de ouders AA en Aa zijn. Bepaal de mogelijke genotypes van de kinderen en de kansen daarop.

b) Stel dat de kansen op genotypes AA, Aa en aa in de eerste generatie resp. p $2q$ en r zijn. Bereken de kansen voor de tweede en derde generaties en laat zien dat deze dezelfde zijn. Dit heet de Wet van Hardy-Weinberg. Hoe zullen deze kansen voor de komende generaties zijn?

c) Bereken de kansen voor de tweede en derde generatie als in b), maar nu onder de extra aanname dat de kans dat een individu van type AA, Aa of aa de paringsleeftijd bereikt, gelijk zijn aan u, v en w.

Opgave 2.10 Bose-Einstein statistiek voor bosonen
Gegeven n identieke deeltjes. Deze willen we over m cellen verdelen. Op hoeveel manieren kan dat? Experimenteel is vastgesteld dat elk van deze mogelijkheden even waarschijnlijk is.

Laat nu $n = 4$ en $m = 3$. Wat is dus de kans dat er 2 deeltjes in cel 1 terecht komen, 1 in cel 2 en 1 in cel 3?

Opgave 2.11 Fermi-Dirac statistiek voor electronen
Gegeven n identieke deeltjes. Op hoeveel manieren kunnen deze over $m \geq n$ cellen verdeeld worden, waarbij hooguit 1 deeltje in elke cel komt? Ook hier is experimenteel vastgesteld dat elke mogelijkheid even waarschijnlijk is.

Laat nu $n = 4$ en $m = 6$. Wat is de kans dat precies 1 deeltje in de cellen 1,2,5,7 terecht komt?

Opgave 2.12 Bereken de kans dat de spelers N en Z in Voorbeeld 2.15 tezamen alle azen hebben. Wat is de kans dat twee spelers alle azen hebben?

Opgave 2.13 Een systeembeheerder zit met het probleem om dusdanige restricties aan toegestane paswoorden op te leggen, dat een random trial and error methode een willekeurig paswoord met voldoend kleine kans vindt. Tegelijk is het zo dat een systeembeheerder bij het uitgeven van accounts random paswoorden moet genereren, en ter wille van de efficiëntie wil hij dat dat niet al te veel niet-toegestane paswoorden oplevert.

Hoe is een paswoord samengesteld? Er zijn 95 mogelijke symbolen die je met een toetsaanslag op je scherm kunt afdrukken. Deze zijn opgedeeld in 5 zogenaamde 'character classes': 26 kleine letters, 26 hoofdletters, 10 cijfers, 10 leestekens

```
! ‘ ’ " : ; , . ? !
```

en de 'rest', 23 in getal

```
~ @ * $ % ^ & # ( ) -  _  + = { } [ ] | \ < > /
```

Het minimale aantal symbolen in een paswoord is gesteld op 7. De systeembeheerder twijfelt nu of beter is het aantal verschillende character classes waaruit een paswoord moet bestaan op 3 of op 4 te stellen. Experiment wijst uit dat random genereren van paswoorden met 7 symbolen te weinig paswoorden oplevert met 4 verschillende character classes, maar dat 3 verschillende character classes voldoend vaak voorkomt.

We willen dit theoretisch verifiëren. Bereken de kans dat random genereren van een paswoord van 7 symbolen een paswoord oplevert met 3 dan wel 4 verschillende character classes.

Doe in Matlab een aantal simulaties van 1000 paswoorden van 7 symbolen en vergelijk de theoretische kansen met de gevonden frequenties.

Opgave 2.14 Intuïtief is het duidelijk dat trekken met of zonder teruglegging uit een hele grote populatie geen invloed heeft, waar het gaat om de grootte van de kans op het aantal objecten met een gegeven kenmerk. Reken dit in Matlab voor Voorbeeld 2.16 eens na: d.w.z. kies x en n. Vergelijk de kansen voor grote waarden van R en N, zó dat die een vast verhouding hebben (bijv. $R/N = 0,25$).

3 Voorwaardelijke kansen en onafhankelijkheid

3.1 Voorwaardelijke kansen

Als we gedeeltelijke informatie over de uitkomst van een experiment hebben, dan kan dit de kans op een bepaalde uitkomst beïnvloeden.

Voorbeeld 3.1 Stel dat er met evenveel kans een jongen als een meisje wordt geboren. Dan heeft een willekeurig gezin van twee kinderen met kans 1/2 één meisje en één jongen, met kans 1/4 twee meisjes en met kans 1/4 twee jongens. Stel dat het oudste kind uit een gegeven gezin een jongen is, wat is de kans dat het andere kind ook een jongen is?

De uitkomst 'twee meisjes' kan al niet meer voorkomen, dus de extra informatie verandert de uitkomstruimte. Wat is deze?

Voorwaardelijke kans. Laat $\mathsf{P}\{B\} \neq 0$. De *voorwaardelijke kans* op A gegeven B is gedefineerd als:

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

Voorwaardelijke kansen voldoen aan de kansaxiomata. In feite zorgt de 'extra informatie' ervoor, dat je je uitkomstenruimte Ω kunt beperken tot een nieuwe, namelijk uitkomstenruimte B.

Voorbeeld 3.2 Vervolg Voorbeeld 3.1.
De uitkomstenruimte $\Omega = \{JJ, JM, MJ, MM\}$. De gebeurtenis $B = \{\text{oudste kind is J}\}$ correspondeert met de gebeurtenis $\{JJ, JM\}$.

Laat nu $A = \{\text{beide kinderen zijn J}\}$, d.w.z. $A = \{JJ\}$. Dan is de kans dat beide kinderen een jongen zijn gegeven dat de oudste het is:

$$\mathsf{P}\{A|B\} = \frac{\mathsf{P}\{A \cap B\}}{\mathsf{P}\{B\}} = \frac{\mathsf{P}\{JJ\}}{\mathsf{P}\{JJ, JM\}} = \frac{1/4}{1/2} = \frac{1}{2}.$$

Stel dat gegeven is dat één van de twee kinderen een jongen is, d.w.z. $B' = \{JJ, JM, MJ\}$. Dit gegeven zijnde, wat is nu de kans dat beide kinderen een jongen zijn? Dat is

$$\mathsf{P}\{A|B'\} = \frac{\mathsf{P}\{A \cap B'\}}{\mathsf{P}\{B'\}} = \frac{\mathsf{P}\{JJ\}}{\mathsf{P}\{JJ, JM, MJ\}} = \frac{1/4}{3/4} = \frac{1}{3}.$$

Voorbeeld 3.3 Men gooit drie keer met een zuivere munt. Wat is nu de kans op minstens 1 keer kruis (K) gegeven minstens 2 keer munt (M)? Noem X het aantal keren K. Dan is de kans op minstens 2 keer M gelijk aan de kans op hooguit 1 keer K, oftewel

$$\mathsf{P}\{X \leq 1\} = \mathsf{P}\{X = 0\} + \mathsf{P}\{X = 1\} = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}.$$

Minstens 1 keer K èn minstens 2 keer M kan alleen, als je *precies* 1 *keer* K vindt. De kans hierop is

$$\mathsf{P}\{X = 1\} = \frac{3}{8}.$$

Dus het antwoord is

$$\mathsf{P}\{X \geq 1|X \leq 1\} = \frac{\mathsf{P}\{X \geq 1 \cap X \leq 1\}}{\mathsf{P}\{X \leq 1\}} = \frac{\mathsf{P}\{X = 1\}}{\mathsf{P}\{X \leq 1\}} = \frac{3/8}{1/2} = \frac{3}{4}.$$

In de praktijk worden vaak voorwaardelijke kansen gegeven (in situatie x is de kans op gebeurtenis y ...). De volgende eigenschap is handig bij modelbouw om kansen uit voorwaardelijke kansen af te leiden.

Behoud van totale kans
Stel $B_1, B_2, \ldots$ is een *partitie* van Ω. Dat wil zeggen de collectie $B_1, B_2, \ldots$ is *disjunct* en $\cup_{i \geq 1} B_i = \Omega$.

Dan is

$$\mathsf{P}\{A\} = \sum_{i:\mathsf{P}\{B_i \neq 0\}} \mathsf{P}\{A|B_i\}\mathsf{P}\{B_i\}.$$

Bewijs. Er geldt

$$A = A \cap \Omega = A \cap \big(\cup_i B_i \big) = \cup_i \big(A \cap B_i \big).$$

De collectie $A \cap B_i$ zijn disjunct om de B_i dat zijn. Verder impliceert $\mathsf{P}\{B_i\} = 0$ dat $\mathsf{P}\{A \cap B_i\} = 0$. Dus krijgen we

$$\mathsf{P}\{A\} = \sum_{i \geq 1} \mathsf{P}\{A \cap B_i\} = \sum_{i : \mathsf{P}\{B_i\} \neq 0} \mathsf{P}\{A \cap B_i\} = \sum_{i : \mathsf{P}\{B_i\} \neq 0} \mathsf{P}\{A|B_i\}\mathsf{P}\{B_i\}.$$

QED

In het bijzonder geldt als $\mathsf{P}\{B\} \neq 0$ dat $\mathsf{P}\{A\} = \mathsf{P}\{A|B\}\mathsf{P}\{B\} + \mathsf{P}\{A|\bar{B}\}\mathsf{P}\{\bar{B}\}$.

Voorbeeld 3.4 Bij een vervoersbedrijf is de kans op vertraging 1/5 bij regenweer en 1/15 bij droog weer. Als de kans op regen 1/10 is, wat is dan de kans op vertraging op een willekeurige dag?

$$\mathsf{P}\{\text{vertraging}\} = \mathsf{P}\{\text{vertraging}|\text{regen}\}\mathsf{P}\{\text{regen}\} + \mathsf{P}\{\text{vertraging}|\text{droog}\}\mathsf{P}\{\text{droog}\} = \frac{1}{5}\frac{1}{10} + \frac{1}{15}\frac{9}{10} = \frac{4}{50}.$$

Voorbeeld 3.5 Twee vrienden J en K worden gedwongen te kiezen uit 3 chocolaatjes, waarvan er één vergiftigd is. Het gekozen chocolaatje dient meteen genuttigd te worden. We schrijven $J = 1$ als J overleeft, en $J = 0$ anders, en analoog voor K. Stel dat J eerst kiest. De kans dat hij het overleeft is dan

$$\mathsf{P}\{J = 1\} = \frac{2}{3}.$$

Als J het overleeft, zijn er twee chocolaatjes over, waarvan er één vergiftigd is. Nu moet K kiezen. De kans dat hij het vergiftigde chocolaatje kiest is $\frac{1}{2}$:

$$\mathsf{P}\{K = 1|J = 1\} = \frac{1}{2}.$$

Mocht J het vergiftigde chocolaatje gekozen hebben, dan hoeft K nergens meer voor te vrezen:

$$\mathsf{P}\{K = 1|J = 0)\} = 1.$$

Behoud van totale kans levert dat

$$\begin{aligned} \mathsf{P}\{K = 1\} &= \mathsf{P}\{K = 1|J = 1\}\mathsf{P}\{J = 1\} + \mathsf{P}\{K = 1|J = 0\}\mathsf{P}\{J = 0\} \\ &= \frac{1}{2} \times \frac{2}{3} + 1 \times \frac{1}{3} = \frac{2}{3}. \end{aligned}$$

M.a.w. K heeft dezelfde kans om te overleven als J. Het maakt dus niet uit wie de eerste keus heeft. Waar zou je toch uit kiezen: de eerste of de tweede keus te hebben?

Ook voorwaardelijke kansen kun je soms bepalen door de voorwaarde om te draaien. Dit is een belangrijke methode in de zogenaamde Bayesiaanse statistiek.

De regel van Bayes.
Stel $B_1, B_2, \ldots$ is een partitie van de uitkomstenruimte Ω. Als $\mathsf{P}\{A\}$, $\mathsf{P}\{B_i\} > 0$ voor gegeven index i, dan geldt

$$\mathsf{P}\{B_i|A\} = \frac{\mathsf{P}\{A|B_i\}\mathsf{P}\{B_i\}}{\sum_{k : \mathsf{P}\{B_k\} \neq 0} \mathsf{P}\{A|B_k\}\mathsf{P}\{B_k\}}.$$

Bewijs. Volgens de definitie van voorwaardelijke kans geldt:

$$\mathsf{P}\{B_i|A\} = \frac{\mathsf{P}\{A \cap B_i\}}{\mathsf{P}\{A\}} = \frac{\mathsf{P}\{A|B_i\}\mathsf{P}\{B_i\}}{\mathsf{P}\{A\}}.$$

Behoud van totale kans geeft

$$\mathsf{P}\{A\} = \sum_{k : \mathsf{P}\{B_k\} \neq 0} \mathsf{P}\{A|B_k\}\mathsf{P}\{B_k\}.$$

Deze twee combineren levert het gewenste. QED

Voorbeeld 3.6 In de studierichtingen Informatica, Natuurkunde en Wiskunde zijn achtereenvolgens 15%, 10% dan wel 30% van de studenten vrouw. Van alle studenten van deze drie studierichtingen studeert 40% Informatica, 35% Natuurkunde en 25% Wiskunde. Wat is de kans dat een willekeurige studente uit één dezer drie studierichtingen Informatica blijkt te studeren? Dat is

$$\begin{aligned}\mathsf{P}\{\mathsf{I}|\text{vrouw}\} &= \frac{\mathsf{P}\{\mathsf{I}\cap\text{vrouw}\}}{\mathsf{P}\{\text{vrouw}\}}\\ &= \frac{\mathsf{P}\{\text{vrouw}|\mathsf{I}\}\mathsf{P}\{\mathsf{I}\}}{\mathsf{P}\{\text{vrouw}|\mathsf{I}\}\mathsf{P}\{\mathsf{I}\}+\mathsf{P}\{\text{vrouw}|\mathsf{N}\}\mathsf{P}\{\mathsf{N}\}+\mathsf{P}\{\text{vrouw}|\mathsf{W}\}\mathsf{P}\{\mathsf{W}\}}\\ &= \frac{0,15\cdot 0,40}{0,15\cdot 0,40+0,1\cdot 0,35+0,3\cdot 0,25}=\frac{6}{17}.\end{aligned}$$

Merk op dat in dit voorbeeld $\mathsf{P}\{\mathsf{I}\}=0,4\neq\mathsf{P}\{\mathsf{I}|\text{vrouw}\}=6/17$: d.w.z. dat de gebeurtenissen 'vrouw' en 'Informatica' geen onafhankelijke gebeurtenissen zijn. Het optreden van de ene gebeurtenis heeft invloed op het optreden van de ander.

3.2 Onderling onafhankelijke gebeurtenissen

Twee gebeurtenissen A en B heten *onderling onafhankelijk* (afgekort: *o.o.*) als

$$\mathsf{P}\{A\cap B\}=\mathsf{P}\{A\}\mathsf{P}\{B\}.$$

Een collectie gebeurtenissen $A_1, A_2, \ldots$ heten onderling onafhankelijk als

$$\mathsf{P}\{A_{i(1)}\cap A_{i(2)}\cap\cdots\cap A_{i(n)}\}=\mathsf{P}\{A_{i(1)}\cdot\mathsf{P}\{A_{i(2)}\}\cdots\mathsf{P}\{A_{i(n)}\},$$

voor elke *eindige greep* indices $i(1),\ldots,i(n)$.

Dit begrip wint aan helderheid, wanneer je je realiseert dat onafhankelijkheid van A en B hetzelfde is als de eis $\mathsf{P}\{A|B\}=\mathsf{P}\{A\}$ (mits $\mathsf{P}\{B\}>0$!). Dat wil zeggen: de frequentie waarmee A optreedt wordt niet beïnvloed door het gegeven dat B (al dan niet) optreedt. Oftewel: het optreden van B verschaft geen enkele informatie over het al dan niet optreden van A.

Voorbeeld 3.7 In Zeeland is een bouwwerk gemaakt van 60 pijlers, die bij storm neergelaten kunnen worden zodat ze een dam vormen. De kans dat één zo'n pijler functioneert op het moment, dat de dam in werking wordt gezet, is vrij groot: ongeveer 95%. Het functioneren van een pijler is onafhankelijk vn het functioneren van de ander pijlers. Echter, als één pijler niet goed is, dan zullen er overstromingen zijn. Het is dus van belang dat *alle* 60 pijlers goed functioneren. De kans hierop is ongeveer $(0,95)^{60}<0,05$!

Voorbeeld 3.8 Om de veiligheid van een kerncentrale te vergroten, bouwt men diverse veiligheidsmechanismen in. Slechts als àl deze mechanismen haperen, kan er een kernramp gebeuren. Men zegt nu, dat de kans op een kernramp erg klein is, omdat het wel erg toevallig zou zijn als alle veiligheidsmechanismen het tegelijk laten afweten. Een impliciete veronderstelling in deze redenering is vaak dat de veiligheidsmechanismen o.o. zijn. In dat geval is de kans dat geen van allen werkt, gelijk aan het prodoct van de afzonderlijke kansen. Deze kans is evenzoveel kleiner naarmate er meer veiligheidsmechanismen zijn. Bij een risico-analyse is het derhalve van groot belang om na te gaan of de veronderstelling van onderlinge onafhankelijkheid wel klopt.

Voorbeeld 3.9 Stel we hebben dezelfde situatie als in Voorbeeld 2.16: een partij van N chips genummerd $c_1,\ldots,c_N$, waarvan er R (de nummers $c_1,\ldots,c_R$) kapot zijn. We doen n keer een aselecte trekking. Noem $A_i=\{i\text{-de trekking levert een kapotte chip}\}$.

Laten we eerst *met teruglegging* trekken. Gebeurtenis A_i correspondeert met alle n-letterwoorden uit de symbolen $c_1,\ldots,c_N$, met één van de symbolen $c_1,\ldots,c_R$ (kapotte chips) op de i-de plaats. Daarvan zijn er $R\cdot N^{n-1}$. Het totaal aantal n-letterwoorden van N symbolen is N^n (**A**). Aangezien trekking uit de uitkomstenruimte Ω van de n-letterwoorden *aselect* is, volgt dat

$$\mathsf{P}\{A_i\}=\frac{\text{aantal 'goede' gebeurtenissen'}}{\text{totaal aantal gebeurtenissen}}=\frac{R\cdot N^{n-1}}{N^n}=\frac{R}{N}.$$

De gebeurtenis dat zowel i-de als j-de ($j \neq i$) trekking een kapotte chip opleveren heeft kans $\mathsf{P}\{A_i \cap A_j\} = R^2N^{n-2}/N^n = (R/N)^2$. Hieruit volgt dat $\mathsf{P}\{A_i \cap A_j\} = \mathsf{P}\{A_i\}\mathsf{P}\{A_j\}$, en dus zijn deze gebeurtenissen onderling onafhankelijk. Had je dit kunnen verwachten?

Bij trekking *zonder teruglegging* ligt dit anders. De uitkomstenruimte Ω bestaat nu alle woorden van n *verschillende* symbolen uit $c_1, \ldots, c_N$. Gebeurtenis A_i correspondeert nu met alle woorden van n verschillende symbolen met éń van de symbolen $c_1, \ldots, c_R$ (kapotte chips) op de i-de plaats. Hoeveel zijn dat er?

Kies eerst één van de kapotte chips $c_1, \ldots, c_R$, en zet die op plaats i: hiervoor zijn R mogelijkheden. Dan zijn er nog $N-1$ symbolen over, waaruit we $n-1$-letterwoord van moeten maken (we knippen dit tussen plaatsen $i-1$ en i doormidden en zetten daar dat eerst gekozen symbool tussen). Dit kan op $(N-1)!/(N-1-(n-1))! = (N-1)!(N-n)!$ manieren (**C**). In het totaal zijn dit $R\cdot(N-1)!/(N-n)!$ woorden (manieren). Evenzo is het aantal woorden van n verschillende letters gelijk aan $N!/(N-n)!$, weer volgens (**C**). Ook hier trekking uit de uitkomstenruimte van 'woorden' van n verschillende letters uit de N chips $c_1, \ldots, c_N$ weer *aselect*. Dus

$$\mathsf{P}\{A_i\} = \frac{\text{aantal 'goede' gebeurtenissen'}}{\text{totaal aantal gebeurtenissen}} = \frac{R \cdot \frac{(N-1)!}{(N-n)!}}{\frac{N!}{(N-n)!}} = \frac{R}{N},$$

dus gelijk aan de kans 'met teruglegging'! De kans dat we een kapotte chip trekken in zowel de i-de als de j-de ($j \neq i$) trekking, kunnen we op dezelfde manier beredeneren:

$$\mathsf{P}\{A_i \cap A_j\} = \frac{R^2 \frac{(N-2)!}{(N-n)!}}{\frac{N!}{(N-n)!}} = \frac{R^2}{N(N-1)} \neq \mathsf{P}\{A_i\} \cdot \mathsf{P}\{A_j\} \quad (= \frac{R^2}{N^2}),$$

d.w.z. de gebeurtenissen A_i en A_j zijn nu *niet* onafhankelijk!

Nu hoeft onafhankelijkheid van een stel gebeurtenissen niet hetzelfde te zijn als 'paarsgewijze onafhankelijkheid' (d.w.z. elk tweetal van het stel is onderling onafhankelijk), ook al lijkt dat op het eerste gezicht wel te gelden. Het begrip 'onafhankelijkheid' is dus een gecompliceerd ding! We geven een voorbeeld.

Voorbeeld 3.10 Laat $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ en doe hieruit aselect een trekking. De kans op elk der punten is derhalve $1/4$. Laat

$$\begin{aligned} A &= \{\omega_1, \omega_4\} \\ B &= \{\omega_2, \omega_4\} \\ C &= \{\omega_3, \omega_4\} \end{aligned}$$

We beweren dat A, B en C paarsgewijs onafhankelijke gebeurtenissen zijn, d.w.z. A en B zijn o.o., A en C zijn o.o. en B en C zijn o.o. Controle: $\mathsf{P}\{A \cap B\} = \mathsf{P}\{A \cap C\} = \mathsf{P}\{B \cap C\} = \mathsf{P}\{\omega_4\} = 1/4$, $\mathsf{P}\{A\} = \mathsf{P}\{B\} = \mathsf{P}\{C\} = 1/2$, dus we hebben inderdaad $\mathsf{P}\{A \cap B\} = \mathsf{P}\{A\} \cdot \mathsf{P}\{B\}$; $\mathsf{P}\{A \cap C\} = \mathsf{P}\{A\} \cdot \mathsf{P}\{C\}$ en $\mathsf{P}\{B \cap C\} = \mathsf{P}\{B\} \cdot \mathsf{P}\{C\}$. Merk dus op dat bijvoorbeeld het optreden van B geen enkele extra informatie verschaft of A ook optreedt (wat is die voorwaardelijke kans?)!

Voor onderlinge onafhankelijkheid van A, B en C hebben we ook nodig dat $\mathsf{P}\{A \cap B \cap C\} = \mathsf{P}\{A\} \cdot \mathsf{P}\{B\} \cdot \mathsf{P}\{C\}$. Maar dan moet gelden: $\mathsf{P}\{A \cap B \cap C\} = \mathsf{P}\{\omega_4\} = 1/4 = (1/2)^3$! Dus A, B en C zijn wel paarsgewijs onafhankelijk maar niet onderling onafhankelijk!

3.3 Opgaven Hoofdstuk 3

Opgave 3.1 Stel n componenten zijn in serie verbonden. Voor iedere unit is er een back-up, en het systeem faalt zodra minstens één unit plus bijbehorende back-up falen. Veronderstel onafhankelijkheid van de units en backups. Noem p de kans op falen van een unit of een back-up. Wat is de kans dat het systeem werkt?

Opgave 3.2 Uit een goed geschud spel van 52 kaarten trek je achtereeenvolgens twee kaarten.

i) Geef de uitkomstenruimte Ω voor dit experiment.

ii) Hoe groot is de kans op twee ruitenkaarten?

iii) Zijn de gebeurtenissen 'eerste kaart is ruiten' en 'tweede kaart is ruiten' onderling onafhankelijk?

iv) Zijn de gebeurtenissen 'eerste kaart is ruiten' en 'tweede kaart is boer' onderling onafhankelijk?

v) Zijn de gebeurtenissen 'eerste kaart is ruiten' en 'eerste kaart is boer' onderling onafhankelijk?

vi) Bereken de kans dat beide kaarten dezelfde 'kleur' hebben.

Opgave 3.3 Een spel van 52 kaarten wordt geschud. Vervolgens worden de kaarten één voor één omgedraaid en getoond. Bepaal de uitkomstenruimte en kansmaat voor dit experiment. Wat is de kans dat de 10-de kaart een aas is? Laat ook zien dat de gebeurtenissen $A = \{\text{de eerste kaart is een aas}\}$ en $B = \{\text{de 10-de kaart is een aas}\}$ afhankelijk zijn. Hoe zit het met (on)afhankelijkheid van A en B als $A = \{\text{eerste kaart is schoppen-10}\}$?

Opgave 3.4 In een Amerikaans onderzoek naar de invloed van ras op veroordeling tot de doodstraf wegens moord werden in 1981 de volgende gegevens gepubliceerd:

	blanke beschuldigde	
	wel doodstraf	niet doodstraf
blank slachtoffer	19	132
zwart slachtoffer	0	9

	zwarte beschuldigde	
	wel doodstraf	niet doodstraf
blank slachtoffer	11	52
zwart slachtoffer	6	97

i) Leidt uit deze cijfers af, dat de kans tot de doodstraf veroordeeld te worden voor een blanke groter is dan voor een zwarte.

ii) Laat tevens zien, dat de kans om tot de doodstraf veroordeeld te worden voor een blanke kleiner is dan voor een zwarte bij uitsplitsing naar blanke en zwarte slachtoffers. Verklaar deze paradox!

Opgave 3.5 Er wordt blindelings een kaart getrokken uit een spel van 52 kaarten. Laat zien dat de gebeurtenissen $A = \{\text{kaart is schoppen}\}$ en $B = \{\text{kaart is een aas}\}$ onderling onafhankelijk zijn.

Opgave 3.6 In elk van drie kastjes zitten twee laden: in elke la zit één munt. Het eerste kastjes bevat twee zilveren munten, het tweede een zilveren en een gouden munt; het derde kastje bevat twee gouden munten. Stel dat je bij aselecte trekking van een la een gouden munt vindt. Wat is de kans dat de andere la van dit kastje een zilveren munt bevat? Evenzo: wat is de kans dat de andere la van dit kastje een gouden munt bevat?

Opgave 3.7 Stel dat de kans om ouder te worden dan 70 jaar gelijk aan 0,6, en de kans om ouder te worden dan 80 jaar gelijk is aan 0,2. Als iemand nu haar 70-ste verjaardag heeft bereikt, wat is dan de kans dat zij ook haar 80-ste verjaardag zal vieren?

Opgave 3.8 Beschouw alle gezinnen met twee kinderen. De vier mogelijkheden JJ, JM, MJ, MM treden ieder met kans 1/4 op. Bereken de kans dat beide kinderen jongens zijn, indien gegeven is dat één van de kinderen een jongen is. Kies vervolgens aselect een jongen uit de verzameling van alle jongens afkomstig uit een gezin met twee kinderen. Wat is de kans dat deze jongen een broertje heeft?

Opgave 3.9 Aan een universiteit zijn over een reeks van jaren de volgende gemiddelde instroomcijfers en rendementen per faculteit bekend:

	Letteren	W&N	Rechten
rendement	30%	50%	40&
instroom mannen	100	150	200
instroom vrouwen	250	50	200

Als we ons tot dit drietal faculteiten beperken, hoe groot is dan de kans dat een man resp. vrouw afstudeert? Hoe groot is de kans dat een willekeurige aankomende student bij één dezer drie faculteiten afstudeert?

Opgave 3.10 We stoppen achtereenvolgens 4 ballen in 4 dozen $d_1, \ldots, d_4$. De kans voor elke bal om in een willekeurige doos terecht te komen is 1/4. Als we weten dat de eerste 2 ballen in verschillende dozen terecht zijn gekomen, wat is dan de kans dat een doos 3 ballen bevat?

Opgave 3.11 Bij transport van flessen vindt vervoer plaats over goede wegen (80% van de afstand) en over slechte wegen (20% van de afstand). Bij vervoer over slechte wegen is de breuk $0,5\%$, bij vervoer over goede wegen $0,1\%$. Hoe groot is de kans dat een willekeurige fles onderweg breekt?

Opgave 3.12 In een showroom staan 8 auto's. Daarvan heeft er één rembekrachtiging (R), twee hebben automatische transmissie (A), twee hebben stuurbekrachtiging (S), één heeft zowel S als A, één heeft alle drie de opties en één heeft geen van drieën. Kies een willekeurige auto. Zij aanwezigheid van A, S en R bij de gekozen auto paarsgewijs onafhankelijke gebeurtenissen? Zijn het onafhankelijke gebeurtenissen?

Opgave 3.13 Voer een reeks onafhankelijke experimenten uit, steeds met kans p op succes. Hoe groot is de kans dat

i) bij de eerste 5 experimenten 2 maal een succes optreedt;

ii) het eerste succes pas bij het 5-de experiment optreedt;

iii) het tweede succes bij het 5-de deelexperiment optreedt?

Als je de experimenten net zolang blijft herhalen totdat het eerste succes optreedt, hoe ziet de uitkomstenruimte Ω er dan uit? Welke kansen ken je aan de uitkomsten toe? Verifieer dat deze kansen tot 1 optellen.

Opgave 3.14 Het aantal eieren dat (een bepaalde soort) vogels legt, kan als volgt worden beschreven: 1 ei met kans $1/10$, 2 eieren met kans $2/10$, 3 eieren met kans $3/10$, 4 eieren met kans $3/10$, en 5 eieren met kans $1/10$. Als de gelegde eieren onafhankelijk van elkaar met kans $1/2$ uitkomen, wat is dan de kans op 3 jonge vogels in een nest?

Opgave 3.15 Voer 10 onafhankelijk experimenten uit, steeds met kans $1/2$ op succes. Als in het totaal 8 successen optreden, hoe groot is dan de (voorwaardelijke!) kans, dat het eerste experiment geen succes was?

Opgave 3.16 Bij een spelshow wordt de kandidaat verzocht te kiezen uit drie deuren. Achter één dezer staat de hoofdprijs. Nadat de kandidaat een deur heeft gekozen, loopt de spelleider naar één van de twee overige deuren en doet deze open. De prijs staat niet achter de deur die de spelleider heeft gekozen. De kandidaat wordt gevraagd of hij alsnog wil wisselen. Wat zou je in dat geval doen?

Opgave 3.17 Van een bepaalde populatie (bestaande uit evenveel mannen als vrouwen) is bekend dat 5% van de mannen en $0,25\%$ van de vrouwen kleurenblind is. Men kiest aselect een persoon, die kleurenblind blijkt te zijn. Bereken de kans dat deze persoon een vrouw is.

Opgave 3.18 Gegeven zijn 5 vazen V_0, V_1, V_2, V_3 en V_4. In elke vaas zitten 4 knikkers. In vaas V_k zitten k rode en $4-k$ witte knikkers. Men kiest aselect een vaas.

a) Uit deze vaas wordt aselect één knikker gepakt. Bereken de kans dat deze rood is.

b) Uit deze vaas worden aselect met teruglegging drie knikkers gepakt. Bereken de kans dat ze alle drie rood zijn.

c) Vraag (b) maar nu zonder teruglegging.

Opgave 3.19 Deze opgave behandelt een eenvoudig voorbeeld van zogenaamde *vertakkingsprocessen*. Een populatie begint met éé individu (bijv een ééncellig beestje); op tijdstip $t = 1$ zal deze zich ofwel delen met kans p, ofwel stervan met kans $(1-p)$. Als het zich deelt, dan gedragen beide 'kinderen' zich onafhankelijk van elkaar en hun 'ouder', volgens hetzelfde principe, maar dan op tijdstip 2.

Wat is de kans dat de populatie voor de derde generatie is uitgestorven? Voor welke waarde van p is deze kans gelijk aan $1/2$?

Opgave 3.20 Bereken de kans dat een gezin met 6 kinderen bestaat uit 4 jongens en 2 meisjes. Neem aan dat de kans op een jongen gelijk is aan $1/2$.

Opgave 3.21 Een verzekeringsmaatschappij onderscheidt high-risk, medium-risk en low-risk cliënten met kansen $0,02$, $0,01$ en $0,0025$ dat een dergelijke cliënt een claim indient. De percentages cliënten van de verschillende categorieën zijn resp. 10%, 20% en 70%. Welk percentage van de claims komt van high-risk cliënten?

Opgave 3.22 Iemand heeft 5 backups, ieder op een andere flop, maar hij is vergeten op welke. Hij heeft 25 flops, en bekijkt hiervan achtereenvolgens 4 flops. Bereken de kans dat deze 4 flops twee van de gezochte backups bevatten.

Opgave 3.23 Een systeem bestaat uit n onafhankelijke units, die elk kans p hebben om te falen. Het systeem gaat down als minstens k units falen (k is hierbij een gegeven maar ons onbekend getal). Wat is de kans dat het systeem down gaat?

Opgave 3.24 Gegeven de volgende binaire string van 7 bits: $01**10*$ oftewel een 'schema' in de terminologie van Genetische Algoritmen. Het symbool $*$ betekent dat het betreffende bitje niet gespecificeerd is. Laten we aannemen dat dat met even grote kans een 0 dan wel een 1 is. Elk bitje heeft voorts een kans p om te muteren. Wat is de kans dat we na mutatie de string 1111111 waarnemen? Reken deze kans uit voor $p = 1/2$. Wat is de p die de kans op 1111111 maximaliseert, oftewel wat is een 'meest aannemelijke' schatter voor p bij de gegeven waarneming dat $01**10*$ tot 1111111 muteert?

4 Stochastische grootheden

4.1 Algemene definities

Een experiment voer je in de regel uit om iets te weten te komen: bijvoorbeeld waarnemingen in de vorm van metingen of ondervraging van personen. De *uitkomst* dient dan als basis voor het trekken van conclusies.

Waarnemingen hoeven niet per sé bij herhaling van het experiment hetzelfde resultaat op te leveren. We hebben dit gezien bij het herhaald gooien van een (zuivere) dobbelsteen. Andere oorzaken kunnen zijn: oncontroleerbare meetfouten, veranderende omstandigheden, andere waarnemers, etc.

Uitkomsten die onderhevig zijn aan dergelijke (toevals)fluctuaties, worden in kansmodellen gerepresenteerd door *stochastische grootheden* (oftewel *stochasten*; afkorting s.g.). Stochastische grootheden worden meestal met *hoofdletters* weergegeven, de waarden die deze s.g. kan aannemen worden met een *kleine letter* aangeduid.

Bijvoorbeeld stelt

$$\{X = x\}$$

de gebeurtenis voor, dat de stochast X de waarde x heeft.

We spreken af, dat $X \in \mathbf{R}$, d.w.z. X neemt waarden in de reële getallen aan. Soms is dat natuurlijk: als X de executietijd van een programma is. Soms is het een codering: antwoorden *ja, nee* kun je met $\{0, 1\}$ coderen.

Definitie 4.1 Een stochastische grootheid is een meetbare functie

$$X : \Omega \to \mathbf{R}.$$

Wij zullen ons hier niet bekommeren om het begrip *meetbaarheid*, omdat we ons verderop zullen beperken tot twee typen stochastische grootheden: discrete en continue.

Voorbeeld 4.1 Experiment: één keer gooien met een zuivere munt, d.w.z. $\Omega = \{\text{kruis}, \text{munt }\}$. Kies $X\{\text{kruis}\} = 1$, $X\{\text{munt}\} = 0$. Dan stelt X een codering van de uitkomsten voor.

Voorbeeld 4.2 Experiment: een worp met een zuivere dobbelsteen, d.w.z. $\Omega = \{1, 2, 3, 4, 5, 6\}$. Definieer $X(i) = i$, $i = 1, \ldots, 6$.

Voorbeeld 4.3 Experiment: twee worpen met een zuivere dobbelsteen, d.w.z. $\Omega = \{(i, j) \,|\, i, j = 1, \ldots, 6\}$. Je kunt bijv. geïnteresseerd zijn in de stochast $X\big((i, j)\big) = i + j$, of $Y\big((i, j)\big) = i$, of $Z\big((i, j)\big) = i \cdot j$. Wat stellen deze voor?

Als een kansverdeling P op de uitkomstenruimte Ω is gegeven, dan induceert de stochast X een kansverdeling P_X op de reële rechte $\mathbf{R}$. Deze geeft aan hoe de totale kansmassa wordt 'uitgesmeerd' over de waarden van X. Met andere woorden: X brengt de kansverdeling op Ω over naar een kansverdeling op $\mathbf{R}$. Hiermee kun je X ook zien als een transformatie van Ω naar $\mathbf{R}$.

Definitie 4.2 Zij X een stochastische grootheid gedefinieerd op de kansruimte $(\Omega, \mathcal{A}, \mathsf{P})$. Dan is de kansverdeling P_X van X op de volgende wijze bepaald:

$$\mathsf{P}_X\{B\} = \mathsf{P}\{\omega : X(\omega) \in B\}, \quad \text{voor alle "nette" } B \subset \mathbf{R}.$$

Notatie: $\mathsf{P}_X\{B\} = \mathsf{P}\{\omega : X(\omega) \in B\} = \mathsf{P}\{X \in B\}$.

Voorbeeld 4.4 (**vervolg Voorbeeld 4.3**) De kansverdeling op Ω wordt gegeven door $\mathsf{P}\{A\} = \frac{\#A}{36}$. De stochast X neemt de waarden $\{2, \ldots, 12\}$ aan. Dan krijgen we bijv. $\mathsf{P}_X\{3\} = \mathsf{P}\{X = 3\} = \mathsf{P}\{(1, 2) \cup (2, 1)\} = 2/36$.

Kansverdelingen van stochastische grootheden worden ook bepaald door verdelingsfuncties.

Definitie 4.3 De *verdelingsfunctie* van een stochastische grootheid X is gedefinieerd als

$$\mathsf{F}_X(x) = \mathsf{P}\{X \leq x\}.$$

We laten index X weg, als er geen verwarring kan ontstaan.

Duidelijk is dat de kansverdeling P_X de verdelingsfunctie F_X bepaalt. Dat dat andersom ook geldt, volgt met behulp van 'maattheorie'. Dit is gebaseerd op het feit dat de kans dat $a < X \leq b$, $a < b$ direct te halen valt uit de verdelingsfunctie:

$$\mathsf{P}\{a < X \leq b\} = \mathsf{P}\{X \leq b\} - \mathsf{P}\{X \leq a\} = \mathsf{F}_X(b) - \mathsf{F}_X(a).$$

Verdelingsfuncties F zijn reëelwaardige functies met de volgende eigenschappen:

i) $0 \leq \mathsf{F}(x) \leq 1$;

ii) $\mathsf{F}(x)$ is niet-dalend, d.w.z. $a < b$ impliceert $\mathsf{F}(a) \leq \mathsf{F}(b)$;

iii) $\mathsf{F}(-\infty) = \lim_{x \to -\infty} \mathsf{F}(x) = 0$; $\mathsf{F}(\infty) = \lim_{x \to \infty} \mathsf{F}(x) = 1$;

iv) F is rechts-continu, oftewel $\lim_{x \downarrow a} \mathsf{F}(x) = \mathsf{F}(a)$; verder $\lim_{x \uparrow a} \mathsf{F}(x) = \mathsf{P}\{X < a\}$.

Voorbeeld 4.5 Kies blindelings 2 verschillende getallen uit $\{1, \ldots, 10\}$ (in één greep). Wat is de verdelingsfunctie van het grootste getal X? Bepaal $\mathsf{P}\{X \geq 5\}$ en $\mathsf{P}\{X = 9\}$.

Dit experiment kunnen we modelleren als een trekking uit de (discreet-homogene -waarom?) uitkomstenruimte $\Omega = \{(i,j) \,|\, 1 \leq i, j \leq 10, i < j\}$ (we trekken in feite een 2-letter alfabet uit een 10-letter alfabet). De stochast X die hier aan de orde is, wordt gegeven door $X\big((i,j)\big) = \max\{i,j\}$. De waarden die X kan aannemen, zijn $\{2, \ldots, 10\}$. We hoeven alleen voor deze waarden van X de verdelingsfunctie te bepalen:

$$\mathsf{F}(x) = \mathsf{P}\{X \leq x\} = \frac{|\{X \leq x\}|}{|\Omega|} = \frac{\binom{x}{2}}{\binom{10}{2}}.$$

Nu geldt: $\mathsf{P}\{X \geq 5\} = 1 - \mathsf{P}\{X < 5\} = 1 - \mathsf{P}\{X \leq 4\} = 1 - \mathsf{F}(4)$ en $\mathsf{P}\{X = 9\} = \mathsf{P}\{X \leq 9\} - \mathsf{P}\{X \leq 8\} = \mathsf{F}(9) - \mathsf{F}(8)$. Leid zelf direct een formule voor $\mathsf{P}\{X = x\}$ af!

Voor we nu specifieke stochasten gaan bestuderen, is het handig om het begrip *onafhankelijkheid van gebeurtenissen* uit te breiden naar *onafhankelijkheid van stochasten*. Deze eigenschap speelt een rol bij de constructie van een aantal bekende stochasten.

Definitie 4.4 De stochasten $X_1, \ldots, X_k$ heten *onderling onafhankelijk* (o.o.) als voor alle 'nette' deelverzamelingen $B_1, \ldots, B_k \subset \mathbf{R}$ geldt dat

$$\mathsf{P}\{X_1 \in B_1, \ldots, X_k \in B_k\} = \mathsf{P}\{X_1 \in B_1\} \cdot \mathsf{P}\{X_2 \in B_2\} \cdot \mathsf{P}\{X_k \in B_k\}.$$

Met andere woorden: zij zijn o.o. dan en slechts dan wanneer alle gebeurtenissen van de vorm $\{X_1 \in B_1\}, \ldots \{X_k \in B_k\}$ o.o. zijn.

Zoals gezegd, gaan we twee typen stochasten bestuderen: discrete en continue.

Voorbeeld 4.6

- Het aantal functionerende verbindingen in een electriciteitscircuit is een discrete s.g.
- Het aantal ogen bij het gooien van een dobbelsteen is een discrete s.g.
- De executietijd van een programma is in principe een continue s.g. (zou je over kunnen twisten- waarom?)
- Analoge signalen zijn continue s.g., digitale signalen zijn discrete s.g.

Van deze twee typen zullen we een aantal voorbeelden behandelen.

4.2 Discrete en continue verdelingen: definities

Discrete verdelingen

X heet een *discrete* stochast als X maar eindig of aftelbaar veel (reële) waarden uit $\Omega^* = \{x_1, x_2 \ldots\}$ kan aannemen. Definieer

$$p_j = \mathsf{P}\{X = x_j\},$$

d.w.z. de stochast X legt kansmassa p_j op het punt x_j.

Het is duidelijk dat $p_j \geq 0$ en $\sum_{j:x_j \in \Omega^*} p_j = 1$. Verder is duidelijk dat de getallen $p_1, p_2, \ldots$ de kansverdeling en verdelingsfunctie van X volledig bepalen. Immers voor elke "nette" $B \subset \mathbf{R}$ geldt:

$$\mathsf{P}\{X \in B\} = \sum_{j:x_j \in B} p_j.$$

Fysisch kun je je een discrete kansverdeling voorstellen als de verdeling van een eenheidsmassa over de reële rechte: hier leg je een massa p_j in het punt x_j.

Verder geldt

$$\mathsf{F}(x) = \sum_{j:x_j \leq x} p_j.$$

Dus maakt F een sprong ter hoogte p_j in het punt x_j, en blijft F constant *tussen* deze sprongpunten in. Het is eenvoudig jezelf ervan te overtuigen dat de verdelingsfunctie de kansverdeling volledig beschrijft: als de verdelingsfunctie F is gegeven voor alle reële x, dan zijn de p_j's en x_j's terug te vinden.

Continue verdelingen

Ofschoon praktische metingen en waarnemingen een eindige nauwkeurigheid hebben, werkt men toch veel met modellen waarin de uitkomstenruimte Ω de reële rechte is (of een deel daarvan).

We noemen X een *continue* stochast als X alle X alle waarden in een zeker interval kan aannemen. In de regel is de verdelingsfunctie $\mathsf{F}(x) = \mathsf{P}\{X \leq x\}$ continu en is er een functie f is, die de (kans)dichtheid van X wordt genoemd, zó dat

$$\mathsf{P}\{a < X \leq b\} = \int_a^b \mathsf{f}(x)dx.$$

(Dit geldt niet altijd: met behulp van de Cantorverzameling kun je tegenvoorbeelden construeren!) M.a.w. de kans op een interval is gelijk aan het *oppervlak onder de grafiek van* f bij dat interval. Fysisch kun je je voorstellen dat de eenheidsmassa is uitgesmeerd over de reële rechte met een dikte die evenredig is met f.

Als f zelf continu is, geldt

$$\mathsf{P}\{x < X < x + \Delta x\} = \int_x^{x+\Delta x} \mathsf{f}(y)dy \approx \Delta x \cdot \mathsf{f}(x),$$

d.w.z. hoe groter de dichtheid, des te groter de kans oftewel des te aannemelijker om een waarde dichtbij x aan te treffen. De waarde $\mathsf{f}(x)$ in het punt x is de *aannemelijkheid* waarmee waarde x voorkomt.

Voorbeeld 4.7 De lichaamstemperatuur X van een mens is een continue stochast, die in principe alle waarden tussen 375^0 en 42^0 kan aannemen (lager of hoger is wellicht mogelijk, maar dan hebben we in elk geval niet met een erg gezond mens te maken). Bij gezonde mensen schommelt X rond de 37^0: de dichtheid f zal dus een maximum hebben in de buurt van 37^0 (waarom?).

Voorbeeld 4.8 Stel X is een aselect gekozen digitaal getal tussen 0 en 1. Aselect wil zeggen dat iedere waarde even *aannemelijk* wordt geacht. Een binair getal in het interval $(0, 1]$ is te schrijven als $\omega = 0.\omega_1\omega_2\ldots$, met $\omega_i \in \{0, 1\}$, $i = 1, 2, \ldots$. Laat Ω de uitkomstenruimte van alle binaire getallen in $(0, 1]$ zijn (wat is Ω dus?). Noteer nu

$$\Omega_n(\omega) = \{ \text{ alle binaire getallen waarvoor de eerste } n \text{ digits overeenkomen met die van } \omega\}.$$

Stel dat ω en ω' in minstens één digit ω_i, $i < n$, verschillen. Dan zijn $\Omega_n(\omega)$ en $\Omega_n(\omega')$ disjunct. Bovendien zijn de twee gebeurtenissen even aannemelijk. Ga na dat je 2^n van dit soort disjuncte gebeurtenissen hebt, wier vereniging de hele Ω is. Daaruit volgt dat $\mathsf{P}\{X \in \Omega_n(\omega)\} = 2^{-n}$.

We vinden dat

$$\mathsf{P}\{X = \omega\} \leq \mathsf{P}\{X \in \Omega_n(\omega)\} = 2^{-n}, \quad \text{voor alle } n,$$

en dus is $\mathsf{P}\{X = \omega\} = 0$: iedere gegeven uitkomst heeft kans 0, maar toch heb je altijd een uitkomst. Je weet alleen niet in eindige tijd welke je te pakken hebt, als je hem zou moeten genereren: na n iteraties weet je hooguit in welke $\Omega_n(\omega)$ je uitkomst zit.

Heeft X een dichtheid? De enige kandidaat is $\mathsf{f}(x) \equiv 1$, omdat we alle waarden even aannemelijk achten. Voor elke tweetal getallen x en x' moet dan gelden

$$\mathsf{P}\{x < X \leq x'\} = \int_x^{x'} \mathsf{f}(y)dy = \int_x^{x'} 1dy = x' - x.$$

Ga na dat dit klopt voor getallen x en x' met *eindige* binaire ontwikkeling! Zo hebben we een beschrijving van de verdeling van X.

Kansen moeten altijd niet-negatief zijn. Derhalve *moet* de dichtheid f zelf ook niet-negatief zijn. Verder neemt X met kans 1 een waarde uit $\mathbf{R}$ (per definitie), d.w.z. $\mathsf{P}\{X \in \mathbf{R}\} = 1$. De dichtheid f voldoet altijd aan:

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

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

Men kan bewijzen dat bij iedere functie f op $\mathbf{R}$ met deze twee eigenschappen een stochast X kan worden gemaakt, waarvan f de dichtheid is.

Wat is nu de verdelingsfunctie F van een stochast X met dichtheid f? Per definitie

$$\mathsf{F}(x) = \mathsf{P}\{X \leq x\} = \int_{-\infty}^{x} \mathsf{f}(t)dt.$$

Als de dichtheid f zelf continu is, dan (Hoofdstelling der Integraalrekening) is deze is de afgeleide van de verdelingsfunctie F:

$$\mathsf{f}(x) = \frac{d}{dx}\mathsf{F}(x).$$

Verder volgt uit de veronderstelde continuïteit van de verdelingsfunctie F dat $\mathsf{P}\{X = x\} = 0$, immers

$$\mathsf{P}\{X = x\} \leq \mathsf{P}\{x - \epsilon < X \leq x + \epsilon\} = \mathsf{F}(x + \epsilon) - \mathsf{F}(x - \epsilon) \to 0, \quad \text{als } \epsilon \to 0.$$

Symmetrische verdelingen

Een bijzonder type kansverdelingen zijn *symmetrische verdelingen*: X heeft een symmetrische kansverdeling om 0, als X en $-X$ dezelfde kansverdeling hebben. In het algemeen geldt dat de kansverdeling van X symmetrisch is om het punt a als $X - a$ en $-(X - a) = a - X$ dezelfde kansverdeling hebben.

Voor het *discrete* geval betekent symmetrie om a niets anders dan dat

$$\mathsf{P}\{X = a + x\} = \mathsf{P}\{X = a - x\}.$$

In het *continue* geval houdt dit in dat de kansdichtheid f symmetrisch is om a oftewel $\mathsf{f}(a + x) = \mathsf{f}(a - x)$. Immers als dit geldt, dan hebben we

$$\begin{aligned} \mathsf{P}\{X - a \leq x\} &= \int_{-\infty}^{x+a} \mathsf{f}(t)dt \overset{u=t-a}{=} \int_{-\infty}^{x} \mathsf{f}(a + u)du = \int_{-\infty}^{x} \mathsf{f}(a - u)du \overset{t=a-u}{=} \int_{a-x}^{\infty} \mathsf{f}(t)dt \\ &= \mathsf{P}\{X \geq a - x\} = \mathsf{P}\{a - X \leq x\} \end{aligned}$$

In de volgende paragraaf bespreken we een aantal belangrijke kansverdelingen. Feitelijk zijn dit steeds families van kansverdelingen, waarvan de leden door één of meerdere parameters worden geïdentificeerd.

Transformaties

Stel we hebben een discrete of continue stochast X gegeven, en we willen de kansverdeling van een functie $g(X)$ van X bepalen. I.h.a. is dit lastig. Voorbeelden van transformaties van een discrete en een continue stochast zijn in Hoofdstuk 1 gegeven. Simulatie van trekkingen uit $g(X)$ is daarentegen betrekkelijk eenvoudig als we een simulatie van trekkingen uit X kunnen doen (waarom?).

4.3 Voorbeelden van discrete kansverdelingen

1. Ontaarde verdeling (gedegenereerde verdeling).
X bezit een *ontaarde* verdeling als X maar één waarde kan aannemen, d.w.z. als voor zeker getal x_0 geldt $P(X = x_0) = 1$. De verdelingsfunctie $F(x)$ is dan constant gelijk aan nul voor $x < x_0$ en constant gelijk aan één voor $x \geq x_0$.

Het komt erop neer dat X deterministisch is, d.w.z. de uitkomst ligt van tevoren vast!

2. Alternatieve verdeling met parameter p.
X bezit een *alternatieve* verdeling $\mathsf{alt}(\mathrm{p})$ als X slechts 2 waarden kan aannemen. Zonder verlies van algemeenheid noemen we deze waarden 1 en 0 en dus

$$X = \begin{cases} 1, & \text{met kans } p \\ 0, & \text{met kans } 1-p. \end{cases}$$

Een stochast X met een dergelijke verdeling noemt men wel een *Bernouilli* grootheid. Dit type stochast komt voor wanneer een experiment maar twee mogelijke uitkomsten heeft, aangeduid met "succes"$(X = 1)$ en "mislukking"$(X = 1)$, waarbij p de *succeskans* is.

3. Binomiale verdeling met parameters n en p.
X is binomiaal $\mathsf{bin}(n, p)$ verdeeld, X alleen de waarden $0, 1, \ldots, n$ kan aannemen en

$$\mathsf{P}\{X = x\} = \binom{n}{x} p^x (1-p)^{n-x}, \text{ voor } x = 0, 1, \ldots, n.$$

Hierbij zijn n en p parameters met $n \in \{1, 2, \ldots\}$ en $0 < p < 1$. Merk op dat de $\mathsf{bin}(n, 1/2)$ verdeling symmetrisch is om $n/2$!

De alternatieve verdeling is een speciaal geval van de binomiale verdeling, nl. $n = 1$. Er is echter een veel nauwere relatie met Bernouilli grootheden.

Beschouw weer experimenten met kans p op succes. Worden n van deze experimenten onafhankelijk uitgevoerd, dan is het aantal successen X $\mathsf{bin}(n, p)$ verdeeld: zie Voorbeeld 2.16 Steekproef met teruglegging. Hierbij wordt n keer een steekproef met teruglegging gedaan uit een populatie van N chips, waaronder R kapotte. Als we "succes" interpreteren als "het trekken van een kapoote chip", dan is $p = R/N$.

Anderzijds representeert de Bernouilli grootheid X_i of het i-de experiment een succes oplevert, voor $i = 1, \ldots, n$ (met succeskans p). Dan is X ook te schrijven als

$$X = X_1 + X_2 + \cdots + X_n.$$

Onafhankelijkheid van de experimenten impliceert onderlinge onafhankelijkheid van de Bernouilli grootheden $X_1, \ldots, X_n$. Kennelijk is de som van n onderling onafhankelijke Bernouilli grootheden (met dezelfde succeskans p) $\mathsf{bin}(n, p)$ verdeeld!

4. Hypergeometrische verdeling.
X bezit een hypergeometrische verdeling als

$$\mathsf{P}\{X = x\} = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}}, \quad x = \max(0, n-(N-R)), \ldots, \min(n, R).$$

Deze verdeling krijg je bij het doen van n trekkingen zonder teruglegging uit een populatie van N objecten, waaronder R objecten met een "gewenste" eigenschap (zie Voorbeeld 2.16 Steekproef zonder teruglegging).

N.B. de uitkomst van de i-de trekking is een Bernouilli grootheid X_i met succeskans $p = R/N$ (zie Voorbeeld 3.9); er geldt wel dat $X = X_1 + \cdots + X_n$, maar de $X_1, \ldots, X_n$ zijn nu *niet* onderling onafhankelijk!

N.B. in de praktijk is R, of, equivalent hieraan, de fractie $p = R/N$ onbekend. Door het nemen van steekproeven wil men nagaan hoe groot p (dus R) (ongeveer) is. Dit is een eenvoudig geval van kwaliteitscontrole en daar zullen we later nog op ingaan.

5. Negatief binomiale verdeling met parameters k en p.
Stel X is het aantal keren dat een computerprogramma gedraaid heeft totdat het voor de eerste keer fout liep. Noem

$$Y_i = \begin{cases} 1, & \text{als het bij de } i\text{-de keer draaien fout loopt;} \\ 0, & \text{als het bij de } i\text{-de keer draaien goed gaat,} \end{cases}$$

en zij $p = \mathsf{P}\{Y_i = 1\}$. Neem nu aan dat het al of niet fout lopen van de individuele executies o.o. gebeurtenissen zijn. Dan geldt

$$\mathsf{P}\{X = x\} = \mathsf{P}\{Y_1 = Y_2 = \ldots = Y_{x-1} = 0, Y_x = 1\} = (1-p)^{x-1}p, \quad x = 1, 2, \ldots.$$

Dit noemt men de *geometrische verdeling*. De geometrische verdeling is een speciaal geval van de *negatief binomiale* verdeling. De laatste krijgt men, als men naar de verdeling kijkt van de s.g. $X^{(k)}$, de wachttijd tot het voor de k-de keer fout loopt. Dan geldt

$$\begin{aligned} \mathsf{P}\{X^{(k)} = x\} &= \mathsf{P}\{\sum_{i=1}^{x-1} Y_i = k-1, Y_x = 1\} \\ &= \binom{x-1}{k-1} p^k (1-p)^{x-k}, \quad x = k, k+1, \ldots. \end{aligned}$$

Voor $k = 1$ is dit de geometrische verdeling.

6. Poissonverdeling met parameter μ.
X heeft een *Poisson*(μ) verdeling, $\mu > 0$, als

$$\mathsf{P}\{X = x\} = \frac{\mu^x}{x!}\mathrm{e}^{-\mu}, \quad x = 0, 1, \ldots.$$

De interpretatie zullen we aan de hand van een voorbeeld proberen duidelijk te maken. Stel dat X het aantal logins gedurende tijdsperiode $[0, T]$ is. We willen de verdeling van X weten. Daartoe verdelen we het interval $[0, T]$ in n kleine deelintervalletjes van lengte T/n. Definieer X_i =het aantal logins in intervalletje i. Stel

a) De kans op één login in interval i is ongeveer evenredig met de lengte van dat interval: $\mathsf{P}\{X_i = 1\} \approx \lambda T/n$. Hier is λ de evenredigheidsconstante.

b) De kans op meer dan één login in een klein interval is ongeveer nul (d.w.z. verwaarloosbaar t.o.v. de kans op één login wanneer het interval maar klein genoeg is): $\mathsf{P}\{X_i > 1\} \approx 0$.

c) Het aantal logins in een klein interval is onafhankelijk van het aantal logins in een ander interval.

Nu is $X = \sum_{i=1}^n X_i$. De bovenstaande veronderstellingen zeggen dat X_i ongeveer alternatief verdeeld is met parameter $p = \lambda T/n$. De onafhankelijkheidsveronderstelling (c) geeft dan

$$\mathsf{P}\{X = x\} \approx \binom{n}{x}\Big(\frac{\lambda T}{n}\Big)^x\Big(1 - \frac{\lambda T}{n}\Big)^{n-x}, \ x = 1, \ldots, n.$$

Herschrijven geeft

$$\mathsf{P}\{X = x\} \approx \frac{n!}{(n-x)!n^x}\frac{(\lambda T)^x}{x!}\Big(1 - \frac{\lambda T}{n}\Big)^{n-x}.$$

Er geldt

$$\lim_{n\to\infty} \frac{n!}{(n-x)!n^x} = 1$$

en

$$\lim_{n\to\infty}\Big(1 - \frac{\lambda T}{n}\Big)^{n-x} = \lim_{n\to\infty}\Big(1 - \frac{\lambda T}{n}\Big)^{n} = e^{-\lambda T}.$$

Dus

$$\mathsf{P}\{X = x\} = \lim_{n\to\infty} \frac{n!}{(n-x)!n^x}\frac{(\lambda T)^x}{x!}(1 - \frac{\lambda T}{n})^{n-x} = \frac{(\lambda T)^x}{x!}e^{-\lambda T}, x = 0, 1, \ldots.$$

M.a.w. X heeft een Poisson(μ) verdeling met $\mu = \lambda T$. We noemen μ de intensiteit. Als μ groot is, is de kans op een login gedurende een klein tijdsinterval ook groot. Dat betekent dat het druk is.

Laten nu X en Y twee onafhankelijke stochasten zijn, zó dat X een Poisson(λ) en Y een Poisson(μ) verdeling heeft. Dan kun je met soortgelijke argumenten aantonen dat $X + Y$ een Poisson($\lambda + \mu$) verdeling heeft. Het bewijs zal in een opgave gevraagd worden!

4.4 Voorbeelden van continue kansverdelingen

1. Uniforme verdeling op $[a,b]$.
De stochast X bezit een *uniforme* of *homogene* verdeling $\mathsf{hom}(a,b)$ op $[a,b]$ als de dichtheid f van X gelijk is aan

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b, \\ 0, & \text{anders} \end{cases} .$$

De dichtheid is constant, zeg gelijk aan c, op $[a,b]$ en we hebben c zó gekozen dat f tot 1 integreert. De verdelingsfunctie wordt

$$\mathsf{F}(x) = \begin{cases} 0, & x \le a \ , \\ \int_{-\infty}^{x} \frac{1}{b-a} dx = \frac{x-a}{b-a}, & x \in [a,b] \ , \\ 1, & x \ge b \ . \end{cases}$$

en

$$\mathsf{P}\{s \le X \le t\} = \frac{t-s}{b-a} = \frac{\text{lengte subinterval}}{\text{lengte hele interval}},$$

voor alle $a \le s < t \le b$.

Deze kansverdeling modelleert volstrekt willekeurige trekkingen uit het interval $[a,b]$: de aannemelijkheid van elk getal is even groot. Trekkingen uit de $\mathsf{hom}(0,1)$ verdeling vormen de basis van vele simulatie studies: dit komen we o.a. in de volgende paragraaf tegen. Men spreekt dan van *aselecte getallen* ("random numbers"). Aselecte getallen die door computers worden geproduceerd, zijn geen echte aselecte getallen, maar deterministische rijen getallen (tussen 0 en 1), die niet te onderscheiden zijn van een aselecte rij getallen: men spreekt dan van *pseudo-aselecte getallen* ("pseudo-random numbers")

2. Exponentiële verdeling met parameter λ.
X bezit een *exponentiële* verdeling met paramter $\lambda > 0$ als de dichtheid is:

$$\mathsf{f}(x) = \lambda \mathrm{e}^{-\lambda x}, \ x \ge 0.$$

De verdelingsfunctie is nu

$$\mathsf{F}(x) = 1 - \mathrm{e}^{-\lambda x}.$$

We zullen een interpretatie geven aan de hand van een voorbeeld. Stel dat de tijd X tussen twee opeenvolgende logins (in seconden) een exponentiële verdeling heeft met parameter λ.

Stel dat de laatste login op tijdstip 0 was. Na t seconden neem je waar dat er sindsdien nog steeds geen login heeft plaatsgevonden. Wat is de kans dat het nog minstens x seconden duurt tot de eerstvolgende login? Deze wordt gegeven door

$$\mathsf{P}\{X > t + x \,|\, X > t\}.$$

Laten we deze berekenen:

$$\mathsf{P}\{X > t + x \,|\, X > t\} = \frac{\mathsf{P}\{X > t + x, X > t\}}{\mathsf{P}\{X > t\}} = \frac{\mathsf{P}\{X > t + x\}}{\mathsf{P}\{X > t\}} = \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} = e^{-\lambda x} = \mathsf{P}\{X > x\}.$$

Dit betekent dat het vanaf *elk* waarneemtijdstip een exponentieel verdeelde tijd met parameter λ duurt tot de volgende login. Deze eigenschap noemt men wel de Markov-eigenschap oftewel 'geheugenloosheid' van de exponentiële verdeling.

Wat impliceert dat o.a. in de praktijk? Stel dat je onderstelt dat de levensduur van lampen exp(2) met verwachtingswaarde van 1/2 jaar (bij continue branden. Dan betekent de geheugenloosheidseigenschap dat de lampen 'niet verouderen'. Modellering van levensduren met exponentiële lijkt dus een niet erg realistische aanname en i.h.a. zal men verdelingen daartoe kiezen die *wel* een verouderingseigenschap hebben.

Wat is nu de kans op een login in het eerstvolgende tijdsinterval ter lengte h? Deze is

$$\mathsf{P}\{X \le h\} = 1 - e^{-\lambda h} \approx \lambda \cdot h.$$

Wat is de kans op meer dan één login in datzelfde tijdsinterval? Dat blijkt verwaarloosbaar klein te zijn t.o.v. de kans op precies één login.

Hieruit volgt dan, dat aan de aannames a), b) en c) van het Poissonproces zijn voldaan. D.w.z. dat het *aantal logins* in een gegeven tijdsinterval ter lengte T een Poisson(λT) verdeling heeft! Gegeven dat n logins

in een bepaald tijdsinterval heeft plaatsgevonden, zijn bovendien de logintijdstippen homogeen verdeeld over dat interval!

3. Gammaverdeling met parameters r en λ
X bezit een $\Gamma(r,\lambda)$ verdeling als de dichtheid gegeven is door

$$\mathsf{f}(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)}, \quad x \geq 0,$$

met

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx$$

de gamma-functie. N.B. $\Gamma(n) = (n-1)!$.

Het is direct in te zien dat dat $\Gamma(1,\lambda)$ verdeling gelijk is aan de exponentiële verdeling met parameter λ. Het blijkt dat de som van n onafhankelijke exponentiële verdelingen met parameter λ een $\Gamma(n,\lambda)$-verdeling heeft In het voorbeeld met logins betekent dit, dat de tijd die het duurt tot de n-de login, een $\Gamma(n,\lambda)$-verdeling heeft.

4. Normale verdeling met parameters μ en σ^2. X bezit een *normale* verdeling met parameters μ en σ^2 als de dichtheid de klokvorm is gegeven door

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

Hierbij is σ de positieve wortel uit σ^2. De parameter μ geeft het maximum van $\mathsf{f}(x)$ aan, en $\mu \pm \sigma$ zijn de buigpunten. De breedte van de grafiek wordt bepaald door σ.

De notatie voor de normale verdeling is: $\mathsf{N}(\mu,\sigma^2)$-verdeling. We schrijven soms $X \sim \mathsf{N}(\mu,\sigma^2)$, waarmee dan bedoeld wordt dat X normaal verdeeld is met parameters μ en σ^2.

De *standaard* normale verdeling ($\mathsf{N}(0,1)$-verdeling) betreft het geval $\mu = 0$, $\sigma^2 = 1$. De dichtheid is dan

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

en de standaard normale verdelingsfunctie is

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

(de zogenaamde *Laplace*-functie). Deze kan verder niet expliciet worden uitgerekend, maar er bestaan wel tabellen van. Uit de symmetrie in 0 volgt dat $\Phi(-x) = 1 - \Phi(x)$. Tabellen bevatten dus alleen de waarden voor $x \geq 0$!

"Normale" $\mathsf{N}(\mu,\sigma^2)$-verdelingen zijn gerelateerd aan standaard normale. Als $X \sim \mathsf{N}(\mu,\sigma^2)$, dan is de "gestandaardiseerde grootheid $Y := (X-\mu)/\sigma$ standaard normaal verdeeld. Dit kun je als volgt inzien:

$$\begin{aligned}
\mathsf{F}_Y(y) = \mathsf{P}\{Y \leq y\} &= \mathsf{P}\Big\{\frac{X-\mu}{\sigma} \leq y\Big\} = \mathsf{P}\{X \leq \mu + \sigma y\} \\
&= \int_{-\infty}^{\mu+\sigma y} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx \\
&\overset{u=(x-\mu)/\sigma}{=} \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = \Phi(y).
\end{aligned}$$

Andersom geldt ook: als Y standaard normaal is verdeeld, dan heeft $X = \mu + \sigma Y$ een $\mathsf{N}(\mu,\sigma^2)$ verdeling:

$$\begin{aligned}
\mathsf{F}_X(x) = \mathsf{P}\{X \leq x\} &= \mathsf{P}\{\mu + \sigma Y \leq x\} = \mathsf{P}\{Y \leq (x-\mu)/\sigma\} \\
&= \int_{-\infty}^{(x-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\
&\overset{u=\mu+\sigma z}{=} \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-(u-\mu)^2/2\sigma^2} dx.
\end{aligned}$$

Ten gevolge geldt voor X $\mathsf{N}(\mu,\sigma^2)$ verdeeld dat

$$\mathsf{F}(x) = \Phi(\frac{x-\mu}{\sigma}).$$

Zo kan men m.b.v. de tabel voor de standaard normale verdeling, de verdelingsfunctie voor iedere andere normale verdeling berekenen.

Vooruitlopend melden we nog één eigenschap van de normale verdeling: als $X \sim \mathsf{N}(\mu,\sigma^2)$, $Y \sim \mathsf{N}(\nu,\tau^2)$ en als X en Y onafhankelijk zijn, dan geldt

$$aX + bY + c \sim \mathsf{N}(a\mu + b\nu + c, a^2\sigma^2 + b^2\tau^2), \quad a,b,c \in \mathbf{R}.$$

Deze bewering kan pas worden bewezen, wanneer we het begrip 'simultane verdeling' hebben behandeld (zie Hoofdstuk 6).

Het belang van de normale verdeling is, dat deze op veel plaatsen opduikt: zij is de basis voor de zogenaamde Brownse beweging. Dit is gebaseerd op de studie van Brown naar de verspreiding van stuifmeel. Later heeft Einstein hiervoor een kansmodel ontworpen, dat ook de beweging van deeltjes (in bijv. een vloeistof) modelleert. De Centrale Limietstelling ondersteunt het belang van de normale verdeling.

Voorbeeld 4.9 Zij $X \sim \mathsf{N}(3,9)$. Dan is $U = (X-\mu)/\sigma \sim \mathsf{N}(0,1)$ verdeeld.

$$\begin{aligned}
\mathsf{P}\{2 < X < 5\} &= \mathsf{P}\{\frac{2-3}{3} < \frac{X-3}{3} < \frac{5-3}{3}\} = \mathsf{P}\{-\frac{1}{3} < U < \frac{2}{3}\} \\
&= \mathsf{P}\{U < \frac{2}{3}\} - \mathsf{P}\{U < -\frac{1}{3}\} \\
&= \Phi\big(\frac{2}{3}\big) - \Phi\big(-\frac{1}{3}\big) = \Phi\big(\frac{2}{3}\big) - \big(1 - \Phi\big(\frac{1}{3}\big)\big) \\
&\approx 0,7475 - 0,3694 \approx 0.378.
\end{aligned}$$

Voorbeeld 4.10 Bij het versturen van een binaire boodschap (0 of 1) van A naar B via een kabel is sprake van ruis, die het signaal vervormt. Het in B ontvangen signaal Y heeft dan de vorm $0+X$ of $1+X$, waarin X de stochastische ruis voorstelt. Neem aan $X \sim \mathsf{N}(0,1/4)$. In B interpreteert men de ontvangen boodschap Y als 0, indien $Y \le 1/2$ en als 1, indien $Y > 1/2$. Hoe groot is de kans op een foute interpretatie?

$$\begin{aligned}
\mathsf{P}\{\text{fout bij verzenden 0}\} &= \mathsf{P}\{0+X > 1/2\} = \mathsf{P}\big\{\frac{X}{1/2} > 1\big\} = 1 - \Phi(1) \approx 0,16 \\
\mathsf{P}\{\text{fout bij verzenden 1}\} &= \mathsf{P}\{1+X < 1/2\} = \mathsf{P}\big\{\frac{X}{1/2} < -1\big\} = \Phi(-1) = 1 - \Phi(1) \approx 0,16
\end{aligned}$$

5. Cauchy-verdeling
De stochast X bezit een Cauchy-verdeling wanneer de dichtheid gegeven is door

$$\mathsf{f}(x) = \frac{1}{\pi(1+x^2)}.$$

Deze verdeling krijg je als quotiënt van 2 onafhankelijke standaard normale verdelingen.

6. Chi-kwadraat verdeling met r vrijheidsgraden
X heeft een chikwadraat verdeling met r vrijheidsgraden, als de dichtheid gegeven is door

$$\mathsf{f}(x) = c_r x^{(r-2)/2} e^{-x/2}, \quad x > 0,$$

waarin c_r een ongespecificeerde constante is, zó dat $\int_0^\infty \mathsf{f}(x)dx = 1$ (de kans op de hele ruimte is 1). Notatie: χ^2_r. Voor $r = 2$ onstaat een exponentiële verdeling met parameter $1/2$!

Deze verdeling is gerelateerd aan de normale verdeling in de volgende zin. Als $X_1,\ldots,X_r$ onafhankelijke standaard normaal verdeelde stochasten zijn, dan heeft $X_1^2 + \cdots + X_r^2$ een χ^2_r-verdeling. Voor $r = 1$ laten we dit zien, als voorbeeld hoe de verdeling van een transformatie van een stochast met een bekende verdeling te bepalen. Er geldt voor $x > 0$ dat

$$\mathsf{P}\{X_1^2 \le x\} = \mathsf{P}\{-\sqrt{x} \le X_1 \le \sqrt{x}\} = 2\mathsf{P}\{0 \le X_1 \le \sqrt{x}\} = 2\big(\Phi(\sqrt{x}) - \tfrac{1}{2}\big).$$

Derhalve geldt voor de kansdichtheid van X_1^2 dat

$$\mathsf{f}(x) = \frac{d}{dx} 2\big(\Phi(\sqrt{x}) - \tfrac{1}{2}\big) = \frac{1}{\sqrt{x}}\phi(\sqrt{x}) = \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2}.$$

Ten gevolge is $c_1 = 1/\sqrt{2\pi}$.

7. Beta-verdeling met parameters α en β
X heeft een Beta-verdeling met parameters $\alpha > 0$ en $\beta > 0$, wanneer de dichtheid gegeven is door

$$\mathsf{f}(x) = \begin{cases} c_{\alpha,\beta} x^{\alpha-1}(1-x)^{\beta-1}, & 0 < x < 1, \\ 0, & \text{anders.} \end{cases}$$

Hierin is $c_{\alpha,\beta}$ de constante die ervoor zorgt dat $\mathsf{f}(x)$ over het interval (0,1) tot 1 uitintegreert. De algemene formule is

$$c_{\alpha,\beta}^{-1} = \text{Beta}(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx,$$

dat is de inverse van de Beta-functie in het punt (α, β). Is α een geheel getal, dan volgt door herhaald partieel integreren

$$\begin{aligned} \text{Beta}(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx &= \frac{\alpha-1}{\beta} \int_0^1 x^{\alpha-2}(1-x)^{\beta} dx \\ &= \frac{\alpha!}{\beta(\beta+1)\cdots(\beta+\alpha-1)}. \end{aligned}$$

De beta-verdelingen spelen een rol in de Bayesiaanse statistiek. Als $\alpha = \beta$, dan zijn de kansverdelingen symmetrisch om 1/2.

4.5 Simulatie van kansverdelingen

Met behulp van de Stats Toolbox van Matlab kun je waarnemingen doen uit een aantal standaard kansverdelingen, zoals de verdelingen die boven besproken zijn. Hierbij moet je nog wel opletten hoe de in te stellen parameters van de kansverdeling in Matlab zijn gedefinieerd. Bijv. het commando

```
>>x=exprnd(mu,10,1000)
```

genereert een 10×1000 array (matrix) van waarnemingen uit de (negatief) exponentiële verdeling met parameter $1/mu$!

In de praktijk kan het nodig zijn waarnemingen te doen uit andere dan het lijstje standaardkansverdelingen. Hieronder volgen enkele methodes.

Discrete verdelingen
Stel X een discrete stochast, d.w.z. X neemt waarden aan uit $\{\omega_1, \omega_2, \ldots\}$. Definieer

$$p_j = \mathsf{P}\{X = \omega_j\}.$$

We gaan nu een aselecte trekking u uit het interval $(0, 1)$ transformeren tot een trekking uit X als volgt: de trekking uit X krijgt waarde ω_j als

$$\sum_{k \le j-1} p_k < u \le \sum_{k \le j} p_k.$$

In Matlab kun je dit op verschillende manieren maken. We geven een manier voor het geval X *eindig veel* waarden aanneemt. Het leven wordt gemakkelijker, wanneer er formules voor de ω_j's en p_j's zijn, maar daar gaan we niet van uit. We geven een voorbeeld.

```
>>w=[ 1 3.5 6 7 9 ];
>>p=[.2 .2 .3 .15 .15 ];
 % geef de waarden w_j en p_j op
>>sum(p)
```

```
 % berekent de som van de kansen: daar moet 1 uit komen.
 % je kunt ook testen op niet-negativiteit.
>>p=[0 cumsum(p)];
 % je moet cumulatieve sommen van kansen vergelijken
 % voeg 0 als eerste element toe om ook waarneming w_1 te kunnen doen
>>u=rand(3,10);
>>x=zeros(3,10);
 % u is een 3x10 array van waarnemingen van aselecte trekkingen uit (0,1)
 % na transformatie bevat x de waarnemingen uit X,
 % eerst initialisatie
>>for k=1:5
>>  x=x+w(k)*((u>p(k))&(u<=p(k+1)));
>>end
 % geeft x(i,j) waarde w(k) als p(k)<u(i,j)<= p(k+1)
```

Continue verdelingen

De eerste methode voor het doen van waarnemingen uit een continue kansverdeling berust op inversie.

Stel een verdelingsfunctie F is gegeven en we willen waarnemingen hieruit simuleren. Je weet dat F niet-dalend is, en waarden tussen 0 en 1 aanneemt. Als je nu bijv. weet dat de kans $\mathsf{F}(x) = \mathsf{P}\{X \leq x\} = 1/2$ voor één of andere x, dan kun je x hieruit oplossen. In het geval F strict stijgend, is $x = \mathsf{F}^{-1}(1/2)$, met F^{-1} de inverse van F. Echter, F kan horizontale stukken hebben, en dan is de inverse niet eenduidig gedefinieerd. Dat probleem kun je oplossen door de kleinste x te kiezen waarvoor $\mathsf{F}(x) = 1/2$ (deze bestaat vanwege de rechts-continuïteit). Deze keuze noeme je dan 'gegeneraliseerde inverse'.

Dit betoog is de grondslag voor het simuleren van een trekking uit F.

Laat nu voor het gemak F strict stijgend zijn. Laat U een uniforme $\mathsf{hom}(0,1)$ verdeling hebben. Dan is $X = \mathsf{F}^{-1}(U)$ een stochast met verdelingsfunctie F. Immers,

$$\mathsf{P}\{X \leq x\} = \mathsf{P}\{\mathsf{F}^{-1}(U) \leq x\} = \mathsf{P}\{U \leq \mathsf{F}(x)\} = \mathsf{F}(x).$$

In het geval de inverse van F gemakkelijk te bepalen is, kunnen we hiermee waarnemingen uit F eenvoudig simuleren. Laten we dit aan de hand van het volgende voorbeeld doen.

Voorbeeld 4.11 Stel we willen 100 waarnemingen doen uit de volgende verdeling (ga na dat dit inderdaad een verdelingsfunctie is, en dat deze strict stijgend is)

$$\mathsf{F}(x) = 1 - \frac{1}{1+x}, \quad x \geq 0.$$

De inverse is (ga na)

$$\mathsf{F}^{-1}(u) = \frac{u}{1-u}, \quad 0 \leq u < 1.$$

In Matlab kun je dit als volgt doen.

```
>>n=100
>>u=rand(1,n);
 % simuleert een 1xn array van trekkingen uit de homogene hom(0,1) verdeling.
>>g=u./(1-u);
 % dit is een trekking uit F
```

Vergelijk nu de empirische verdelingsfunctie (zie Hoofdstuk 1.15) F_{100} met F.

```
>>x=2*(1:n)/n;
 % genereert n punten op gelijke afstand in het interval (0,2].
>>F=zeros(1,n);
 %genereert een 1xn array gevuld met 0-en
>>for k=1:n
>> F(k)=length(find(g<=x(k)));
>>end
```

```
 % F is de cumulatieve verdeling!
>>stairs(x,F)
 % plot de empirische verdelingsfunctie F_100 als een trappetje
>>r=0:0.01:2;
>>plot(r,1-(1/(1+r)),'r')
 % plot de echte verdelingsfunctie
```

In veel gevallen is het lastig de verdelingsfunctie te inverteren, terwijl er wel een nette uitdrukking voor de dichtheid f is. Laat X een stochast zijn met dichtheid f.

Stel eerst dat f een dichtheid op een eindig interval $[a, b]$ is, en dat $\mathsf{f}(x) \leq c$ voor een constante c. Laten nu Z en Y twee onderling onafhankelijke stochasten zijn, met Z een homogene $\mathsf{hom}(a, b)$ verdeling en Y een homogene $\mathsf{hom}(0, 1)$ verdeling.

We doen onafhankelijke trekkingen $z_1, z_2, \ldots$ uit Z en $y_1, y_2, \ldots$ uit Y, en we vergelijken de waarden y_k met $\mathsf{f}(z_k)/c$. Als $y_k \leq \mathsf{f}(z_k)/c$ dan *accepteren* we de trekking z_k en stellen $X = z_k$; anders doen we nieuwe (onafhankelijke) trekkingen z_{k+1} uit Z, en y_{k+1} uit Y. De wiskundige verantwoording voor deze procedure is dat de *voorwaardelijke verdeling* van Z gegeven $Y \leq \mathsf{f}(Z)$ de gezochte verdeling met dichtheid f is (zie Hoofdstuk 6 Simultane verdelingen).

Voorbeeld 4.12 Stel

$$\mathsf{f}(x) = 6x \cdot (1 - x), \quad x \in [0, 1].$$

Dan geldt

$$\mathsf{F}(x) = 3x^2 - 2x^3, \quad x \in [0, 1].$$

Deze verdelingsfunctie is lastig te inverteren, dus passen we bovenstaande procedure toe in Matlab. We gegeneren een groot aantal trekkingen uit Z en Y; vervolgens maken we een histogramplotje van de geaccepteerde trekkingen. Als de geaccepteerde trekkingen inderdaad trekkingen uit de gegeven verdeling met dichtheid f zijn, dan moet het histogram behoorlijk gaan lijken op de dichtheid zelf, na correct schalen.

```
>>n=1000;
>>z=rand(1,n);
>>y=rand(1,n);
 % gegenereer 1000 trekkingen uit Y en Z.
>>r=0:0.01:1;
>>f=6*r.*(1-r);
 % berekent de dichtheid in een groot aantal punten van [0,1].
>>c=max(f);
 % bepaalt de kleinste constante c met f(x)<=c
>>hits=find(y<=6*z.*(1-z)/c);
 % bepaal de indices van de geaccepteerde trekkingen
>>hist(z(hits));
 % genereert het histogram van de geaccepteerde waarden
>>hold on
>>m=length(hits);
 % m is aantal geaccepteerde trekkingen, dus beoogt het aantal waarnemingen uit
 % de gewenste X te zijn
>>[u,t]=hist(z(hits));
 % we moeten ook de binbreedte van het histogram weten om de juiste schalingsfactor
 % te bepalen
>>plot(r,m*(t(2)-t(1))*f,'r')
 % plot de geschaalde dichtheid in rood
```

Stel dat f een dichtheid op een onbegrensd interval is. Dan valt er niet homogeen uit dat interval te trekken en moet een andere methode van stal gehaald worden. We nemen nu aan, dat er een kansdichtheid r en een constante c bestaan, zó dat

$$\mathsf{f}(x) \leq c \cdot \mathsf{r}(x), \quad \text{voor alle } x.$$

De stochast Z is nu verondersteld de kansdichtheid r te hebben. De k-de trekking wordt geaccepteerd, wanneer

$$y_k \leq \frac{\mathsf{f}(z_k)}{c \cdot \mathsf{r}(z_k)}.$$

4.6 Opgaven Hoofdstuk 4

Opgave 4.1 Laat X het aantal ogen bij één worp van een eerlijke dobbelsteen zijn. Bepaal de mogelijke waarden en de kansverdeling van de stochasten $Y = X^2$ en $Y = -X$.

Opgave 4.2 Werp 3 keer met een zuivere munt en noteert met X het aantal keren dat Kruis bovenop ligt. Bepaal mogelijke waarden en kansverdeling van X.

Opgave 4.3 Trek 2 keer zonder teruglegging uit een vaas met 5 genummerde ballen (de nummers zijn 1 t/m 5) en noteer met X het kleinst getrokken nummer. Bepaal mogelijke waarden en kansverdeling van X.

Opgave 4.4 Een zuivere munt wordt gegooid tot voor het eerst Munt verschijnt, maar niet meer dan 10 keer. Bepaal waardebereik en verdeling van X, het aantal keren Kruis.

Opgave 4.5 Drie identieke en zuivere munten worden herhaald tegelijk geworpen, net zolang tot ze alle drie dezelfde kant boven hebben. Wat is de kans dat je meer dan drie keer moet werpen?

Opgave 4.6 Van het aantal e-mails dat per uur binnenkomt, is vastgesteld dat dat een Poisson(2) verdeling heeft.

i) Bereken de kans dat er een e-mail binnenkomt tijdens de koffiepauze van 10 minuten.

ii) Hoe lang kans men pauze nemen, als men eist dat de kans dat er *geen* e-mail tijdens de pauze binnekomt, tenminste 1/2 is?

Opgave 4.7 Een zeldzame ziekte heeft een incidentie van één op duizend. Stel dat individuen in een populatie onafhankelijk van elkaar al of niet geïnfecteerd raken. Bepaal de kans op k gevallen in een populatie van 10 000 individuen, voor $k = 0, 1, \ldots$.

Opgave 4.8 Stel dat X en Y onderling onafhankelijk, binomiaal verdeeld zijn: X heeft een $\mathsf{bin}(n, p)$-verdeling en Y een $\mathsf{bin}(m, p)$.

i) Beredeneer wat de verdeling van $X + Y$ is.

ii) Bereken $\mathsf{P}\{X = k \,|\, X + Y = N\}$.

Opgave 4.9 Laten X en Y onderling onafhankelijke Poisson verdeelde stochasten zijn met parameters μ resperctievelijk ν.

i) Beredeneer wat de verdeling van $X + Y$ is.

ii) Bereken $\mathsf{P}\{X = k \,|\, X + Y = N\}$.

Opgave 4.10 De stochastische grootheid X bezit de volgende kansdichtheid

$$\mathsf{f}(x) = \begin{cases} cx(3-x), & 0 < x < 3 \\ 0, & \text{anders.} \end{cases}$$

i) Voor welke waarde van c is f een kansdichtheid?

ii) Bereken $\mathsf{P}\{1 < X < 3\}$ en $\mathsf{P}\{X < 2\}$.

Opgave 4.11 Veronderstel dat de gespreksduur van telefoongesprekken een $\mathsf{N}(\mu, \sigma^2)$ verdeling heeft met $\mu = 2$ minuten en $\sigma = 30$ seconden. Bereken de kans dat een gesprek

i) langer duurt dan 3 minuten;

ii) korter duurt dan 30 seconden;

iii) tussen de 30 seconden en $2\frac{1}{2}$ minuut duurt.;

iv) een negatieve duur heeft.

De aanname van normaliteit is wat vreemd omdat negatieve gespreksduren dan voor kunnen komen. Een meer realistische aanname is *lognormaliteit*: X is lognormaal verdeeld als $\log X$ normaal verdeeld is.

Opgave 4.12 Stel $X_1, \ldots, X_4$ zijn onderling onafhankelijk en standaard normaal verdeeld. Noem $\bar{X} = (X_1 + \cdots + X_4)/4$. Bereken $\mathsf{P}\{\bar{X} > 1/2\}$. Sommigen hebben de neiging om te denken dat $\mathsf{P}\{\bar{X} > 1/2\} = \mathsf{P}\{4X_1/4 > 1/2\}$. Immers, de X_i's hebben alle dezelfde verdeling. Is dat waar en waarom?

Opgave 4.13 Stel X en Y zijn onderling onafhankelijk en $X \sim \mathsf{N}(0, 25)$ en $Y \sim \mathsf{N}(-1, 9)$. Bereken $\mathsf{P}\{X - Y > 2\}$ en $\mathsf{P}\{2X + 3Y > 5\}$.

Opgave 4.14 Stel X en Y zijn onderling onafhankelijk exponentieel verdeeld met parameter λ. Bepaal de verdeling van $\min(X, Y)$.

Opgave 4.15 Stel X heeft verdelingsfunctie F. Bereken de dichtheid f in de volgende gevallen

i) $\mathsf{F}(x) = x^2$, $0 \leq x \leq 1$.

ii) $\mathsf{F}(x) = 1 - (1 + x)^{-4}$, $x \geq 0$.

iii) $\mathsf{F}(x) = \sin(x)$, $0 \leq x \leq \pi/2$.

Opgave 4.16 Genereer een steekproef ter grootte n uit F met F gegeven in opgave 4.15. Teken de empirische verdelingsfunctie F_n en de theoretische verdelingsfunctie F in één plaatje.

Opgave 4.17 Stel X heeft dichtheid f. Bereken de verdelingsfunctie F in de volgende gevallen:

i) $\mathsf{f}(x) = 12x^2(1 - x)$, $0 \leq x \leq 1$.

ii) $\mathsf{f}(x) = (x + 1)/2$, $-1 \leq x \leq 1$.

iii) $\mathsf{f}(x) = \sin(x)$, $0 \leq x \leq \pi/2$.

Opgave 4.18 Iemand zet een baan uit voor een hardloopwedstrijd van 100 meter door 100 stappen te nemen. De lengte van de stappen zijn onderling onafhankelijk en $\mathsf{N}(\mu, 0,01)$ verdeeld.

i) Als $\mu = 0,97$ meter, bereken de kans dat de lengte van de baan minder dan 5 meter van 100 meter verschilt.

ii) Hoe groot moet μ tenminste zijn, opdat de kans dat de lengte van de baan groter is dan 100 meter, groter is dan 0,95?

Opgave 4.19 Een bedrijf is slecht bereikbaar: bijna iedere keer dat men belt, is de lijn bezet. Laat X het aantal keren zijn dat men moet bellen om uiteindelijk iemand aan de lijn te krijgen. Stel dat

$$\mathsf{P}\{X = x\} = (1 - p)p^{x-1}, \quad x = 1, 2, \ldots$$

Wat stelt p voor? Bepaal de verdelingsfunctie $\mathsf{F}(x)$ van X in $x = 1, 2, \ldots$

Opgave 4.20 Een multiple-choice test bestaat uit 20 vragen, met bij elke vraag de keuze uit 4 mogelijke antwoorden. Een zekere student beheerst de stof niet al te best, maar kan wel bij iedere vraag één van de antwoorden (correct) elimineren. Van de overige 3 mogelijke antwoorden kiest de student er één *op goed geluk*. De eis is dat tenminste 12 vragen goed zijn beantwoord.

i) Wat is de kans dat de student slaagt?

ii) Bepaal deze kans nog eens, maar nu onder de aanname dat de student 2 van de 4 antwoorden kan elimineren.

Opgave 4.21 Door drie extra bits aan een vier-bit woord toe te voegen op een bepaalde manier (Hamming code), kan men één foute bit detecteren en corrigeren. Als elke bit kans 0,05 heeft om gedurende de communicatie te zijn veranderd, en de bits onafhankelijk van elkaar al of niet veranderen, wat is dan de kans dat een woord correct wordt ontvangen (d.w.z. 0 of één bit fout)? Wat is de kans dat het woord correct wordt ontvangen als er geen check bits zijn?

Opgave 4.22 Stel X is de reparatietijd (in uren) van een apparaat. Onderstel dat X exponentieel verdeeld is met parameter $\lambda = 1/2$.

i) Bereken de kans dat X kleiner is dan 2 (uur).

ii) Als Y de reparatietijd van hetzelfde apparaat voorstelt in minuten, wat is dan de kansverdeling van Y?

Opgave 4.23 Als een meting naar behoren wordt uitgevoerd. is de uitkomst $\mathsf{N}(3, 4)$ verdeeld. Zo nu en dan worden er bij de meting grove fouten gemaakt, in welk geval de uitkomst $\mathsf{N}(8, 16)$ verdeeld is. Als grove fouten worden gemaakt met kans $1/20$, en X de uitkomst van het experiment voorstelt, wat zijn dan verdelingsfunctie en kansdichtheid van X? Test je antwoord door simulatie.

Opgave 4.24 Zij het aantal coli bacteriën in een literfles water Poisson(2) verdeeld. Als de aantallen coli bacteriën in verschillende flessen onafhankelijk zijn, bereken dan de kans dat zich in 5 flessen geen enkele coli bacterie bevindt.

5 Momenten, verwachting en variantie

Een aantal simpele karakteristieken van kansverdelingen van stochasten spelen in de praktijk een grote rol. Vooral locatie (plaats) en spreiding (variatie) zijn van belang. Deze intuïtieve concepten worden gekarakteriseerd met behulp van zogenaamde *momenten* van kansverdelingen, een wellicht niet altijd bevredigende procedure.

Definitie 5.1 Stel X is een discrete stochastische grootheid met waarden $\{x_1, x_2, \ldots\}$. Dan is de *verwachting* van een functie $g(X)$ van X het getal

$$\mathsf{E}g(X) = \sum_j g(x_j)\mathsf{P}\{X = x_j\}$$

(mits de reeks absoluut convergeert). Als X continu verdeeld is met dichtheid f_X, dan is de verwachting van een functie $g(X)$ van X

$$\mathsf{E}g(X) = \int_{-\infty}^{\infty} g(x)\mathsf{f}_X(x)dx$$

(mits deze integraal absoluut convergeert). Hier staat E voor *expectation*.
N.B. Het getal $\mathsf{E}X$ is de *verwachting* van X.

In het discrete geval is de verwachting $\mathsf{E}X$ een gewogen gemiddelde van de mogelijke waarden van X, met de kansen als gewichten. Als X zelfs een aselecte trekking is uit een eindige populatie elementen $x_1, \ldots, x_N$ (en de kansen dus gelijk zijn aan de relatieve frequenties van deze elementen binnen de populatie), dan is $\mathsf{E}X$ het *populatiegemiddelde*.

Het fysisch analogon van verwachting is *zwaartepunt*. De kans $p_j = \mathsf{P}\{X = x_j\}$ heeft de interpretatie van de hoeveelheid massa die in punt x_j wordt gelegd. Het zwaartepunt wordt dan gegeven door $(\sum_i x_j p_j)/(\sum_j p_j) = \sum_j x_j p_j = \mathsf{E}X$.

Ook in het continue geval gelden deze interpretaties. In relatie to de kansverdeling van X stelt het getal $\mathsf{E}X$ dus het "midden" van de kansverdeling voor. Merk op dat de verwachtingswaarde $\mathsf{E}X$ niet van de betekenis van X afhangt, maar alleen van zijn kansverdeling. Kun je $\mathsf{E}X$ bij een bepaalde interpretatie van X eenvoudig bepalen, dan geldt het resultaat voor alle stochasten met dezelfde kansverdeling. Deze eigenschap is een invariantie-eigenschap, waarvan soms handig gebruik van kan worden gemaakt.

Voorbeeld 5.1 Laat X het aantal ogen zijn bij één keer gooien met een dobbelsteen. De gemiddelde waarde (populatiegemiddelde) is dan $(1+2+3+4+5+6)/6 = 3,5$. Formele berekening van de verwachting geeft

$$\begin{aligned}\mathsf{E}X &= 1\cdot\mathsf{P}\{X=1\} + 2\cdot\mathsf{P}\{X=2\} + 3\cdot\mathsf{P}\{X=3\} + 4\cdot\mathsf{P}\{X=4\} + 5\cdot\mathsf{P}\{X=5\} + 6\cdot\mathsf{P}\{X=6\}\\ &= 1\times\frac{1}{6} + 2\times\frac{1}{6} + 3\times\frac{1}{6} + 4\times\frac{1}{6} + 5\times\frac{1}{6} + 6\times\frac{1}{6} = 3,5.\end{aligned}$$

Willen we de verwachting van $g(X) = X^2$ berekenen, dan krijgen we:

$$\begin{aligned}\mathsf{E}X^2 &= 1^2\cdot\mathsf{P}\{X=1\} + 2^2\cdot\mathsf{P}\{X=2\} + 3^2\cdot\mathsf{P}\{X=3\} + 4^2\cdot\mathsf{P}\{X=4\} + 5^2\cdot\mathsf{P}\{X=5\} + 6^2\cdot\mathsf{P}\{X=6\}\\ &= 1\times\frac{1}{6} + 4\times\frac{1}{6} + 9\times\frac{1}{6} + 16\times\frac{1}{6} + 25\times\frac{1}{6} + 36\times\frac{1}{6} = \frac{91}{6}.\end{aligned}$$

Voorbeeld 5.2 Laat X nu een aselecte trekking uit het interval $(0,1)$ voorstellen. Dat wil zeggen dat X homogeen is verdeeld op $(0,1)$. Dan is

$$\mathsf{E}X = \int_0^1 x\cdot\mathsf{f}_X(x)\,dx = \int_0^1 x\cdot 1\,dx = \big[\tfrac{1}{2}x^2\big]_{x=0}^{x=1} = \frac{1}{2}.$$

Willen we $\mathsf{E}\sqrt{1-X^2)}$ weten, dan krijgen we

$$\begin{aligned}\mathsf{E}\sqrt{1-X^2} &= \int_0^1 \sqrt{1-x^2}\cdot\mathsf{f}_X(x)\,dx = \int_0^1 \sqrt{1-x^2}\cdot 1\,dx\\ &\overset{x=\cos u}{=} \int_{u=\pi/2}^{0} \sin u\,d(\cos u) = \int_{u=0}^{\pi/2}\sin^2(u)\,du = \int_{u=0}^{\pi/2}\big(\frac{1}{2} - \frac{1}{2}\cos(2u)\big)\,du\\ &= \big[\tfrac{1}{2}u - \frac{1}{4}\sin(2u)\big]_{u=0}^{u=\pi/2} = \pi/4.\end{aligned}$$

Merk nu op, dat de tweede integraal niets anders voorstelt dan de oppervlakte onder de grafiek van de functie $\sqrt{1-x^2}$ tussen 0 en 1, d.w.z. de oppervlakte van de kwartcirkel.

Dit betekent dat de verwachting van een functie van de homogene verdeling niets anders is dan de oppervlakte is onder de grafiek van deze functie. Dit geeft een gereedschap om oppervlakten te berekenen via simulatie van trekkingen uit de homogene verdeling. Dat is de basis voor zogenaamde Monte-Carlo integratie, een methode die in de Statistische Mechanica nogal van belang is en die we verderop nog zullen behandelen.

We zullen later ook zien dat kansen en verwachting op dezelfde wijze kunnen worden gesimuleerd. Dat heeft te maken met het feit dat een kans *ook* als verwachting geschreven kan worden.

Hoe gaat dit in zijn werk? Stel we willen $\mathsf{P}\{X \in B\}$, $B \subset \mathbf{R}$, als een verwachting schrijven. Daartoe definiëren we de functie g als de zogenaamde indicatorfunctie op B:

$$g(x) = \mathbf{1}_{\{x \in B\}} = \begin{cases} 1, & x \in B \\ 0, & \text{anders.} \end{cases}$$

Dan geldt

$$\mathsf{E}g(X) = \mathsf{E}\mathbf{1}_{\{X \in B\}} = 1 \cdot \mathsf{P}\{X \in B\} + 0 \cdot \mathsf{P}\{X \notin B\} = \mathsf{P}\{X \in B\}.$$

Enkele eigenschappen van verwachtingen
Uit de definitie van verwachting volgt onmiddellijk de eigenschap dat $X \geq 0$ impliceert dat $\mathsf{E}X \geq 0$. Voort geldt vrij direct dat

$$\mathsf{E}(af(X) + bg(X) + c) = a\mathsf{E}f(X) + b\mathsf{E}g(X) + c, \tag{5.1}$$

waarbij $f(X)$ en $g(X)$ functies van de stochast X zijn en a, b en c constanten (mits de betreffende integralen/sommen absoluut convergeren). Dat impliceert dat: de verwachting van een som is de som van de verwachtingen. Waarom?

Symmetrische verdelingen
Stel X heeft een symmetrische verdeling met symmetrie-punt a. Als $\mathsf{E}X$ bestaat, dan geldt $\mathsf{E}X =$ symmetriepunt $= a$. Waarom?

Vanwege de symmetrie hebben $X - a$ en $a - X$ dezelfde verdeling en dus ook dezelfde verwachting:

$$\mathsf{E}(X - a) = \mathsf{E}(a - X), \text{ oftewel } \mathsf{E}X - a = a - \mathsf{E}X \quad \text{oftewel } \mathsf{E}X = a.$$

In geval van symmetrie geeft de verwachting ook het 'midden', de 'locatie' van de verdeling weer. Wanneer symmetrie ontbreekt, is $\mathsf{E}X$ misschien een minder geschikte maat voor de 'locatie', vooral wanneer er extreme waarden zijn die met kleine kans optreden.

Voorbeeld 5.3 Laat X gedefinieerd zijn door $\mathsf{P}\{X = -1\} = \mathsf{P}\{X = 1\} = 0,499$ en $\mathsf{P}\{X = 10000\} = 0,001$. Dan is

$$\mathsf{E}X = (-1) \cdot 0,499 + 1 \cdot 0,499 + 10000 \cdot 0,001 = 10,$$

hoewel de verdeling bijna geheel rond 0 is gecentreerd.

De verwachting is het zogenaamde *eerste moment* van de stochast X.

Definitie 5.2 Voor een stochast X is

$$\mathsf{E}(X^k)$$

(mits dit bestaat) het *k-de moment.*, $k = 1, 2, \ldots$
De *variantie* van X is gedefinieerd als

$$\text{var}(X) = \mathsf{E}\big(X - \mathsf{E}X\big)^2 = \mathsf{E}X^2 - (\mathsf{E}X)^2$$

(mits de verwachtingen bestaan).

De laatste gelijkheid volgt door uitschrijven en gebruikmaken van het feit dat de verwachting van een som de som van de verwachtingen is (5.1)

$$\mathsf{E}\big(X - \mathsf{E}X\big)^2 = \mathsf{E}\big(X^2 - 2X\mathsf{E}X + \big(\mathsf{E}X\big)^2\big) = \mathsf{E}X^2 - 2 \cdot \big(\mathsf{E}X\big)^2 + \big(\mathsf{E}X\big)^2 = \mathsf{E}X^2 - \big(\mathsf{E}X\big)^2.$$

De variantie is de verwachte (gemiddelde) kwadratische afstand van X tot de centrale waarde $\mathsf{E}X$, en beschrijft daarmee de variatie van X om zijn verwachting. Door het kwadrateren is het effect van extreme waarden van X op de variantie nog groter dan bij verwachting. Verder is de eenheid waarin $\mathrm{var}(X)$ wordt uitgedrukt het kwadraat van de eenheid waarin X wordt gemeten. Dit is één van de redenen om de standaarddeviatie van X in te voeren:

$$\sigma(X) = \sqrt{\mathrm{var}(X)} = \sqrt{\sigma^2(X)}.$$

Naar analogie met deze notatie, schrijven we i.h.a. $\sigma^2(X)$ voor de variantie van X!

Voorbeeld 5.4 Bij een roulettespel zetten we één euro in op oneven. De kans om een euro te winnen is dan $18/37$ en de kans om een gulden te verliezen is $19/37$ (nul is even!). Dus als X de winst is, dan is $\mathsf{E}X = 1 \cdot 18/37 + (-1) \cdot 19/37 = -1/37$ en $\mathsf{E}X^2 = 1$. Dus is $\sigma^2(X) = 1 - (-1/37)^2 \approx 0,9993$.

We kunnen ook een andere strategie kiezen. Stel we zetten één euro in op 23. De winst is dan $Y = 36-1 = 35$ met kans $1/37$ en $Y = -1$ met kans $36/37$. Dus is $\mathsf{E}Y = -1/37$ en de tweede strategie heeft dezelfde winstverwachting als de eerste. Echter voor de tweede strategie geldt $ExpY^2 = (35)^2/37 + 36/37 = 1261/37$, zodat $\sigma^2(Y) = 1261/37 - (1/37)^2 \approx 34,0803$. De spreiding is bij de tweede strategie groter, d.w.z. de gemiddelde fluctuaties rond het gemiddelde zijn groter. Dit betekent dat je meer risico neemt, maar daar staat tegenover dat je ook meer kunt winnen.

Aan welke strategie zou je de voorkeur geven?

Voorbeeld 5.5 Laat X een aselecte trekking uit het interval $(0,1)$ zijn, d.w.z. X is homogeen verdeeld op $(0,1)$. In voorbeeld 5.2 hadden we berekend dat $\mathsf{E}X = 1/2$. Voorts geldt

$$\mathsf{E}X^2 = \int_0^1 x^2 \cdot \mathsf{f}_X(x)\,dx = \int_0^1 x^2\,dx = \big[\frac{1}{3}x^3\big]_{x=0}^{x=1} = \frac{1}{3},$$

zodat $\sigma^2(X) = 1/3 - (1/2)^2 = 1/12$.

Enkele eigenschappen van variantie

i) Uit de definitie volgt dat

$$\sigma^2(X) \geq 0.$$

Immers, $\sigma^2(X) = \mathsf{E}\big(X - \mathsf{E}X\big)^2$, en $(X - \mathsf{E}X)^2 \geq 0$!

ii) Door herschrijving weten we $\sigma^2(X) = \mathsf{E}X^2 - \big(\mathsf{E}X\big)^2$ en dus geldt ook

$$\mathsf{E}X^2 \geq \big(\mathsf{E}X\big)^2.$$

iii) $\sigma^2(X) = 0$ dan en slechts dan als X een ontaarde verdeling heeft. Dat wil zeggen dat $\mathsf{P}\{X = x_0\} = 1$ voor een zeker getal x_0. Voor dit getal geldt $\mathsf{E}X = x_0$ (waarom?).

Voor discrete stochasten is dit niet moeilijk in te zien. Laten we aannemen dat X de waarde x_j aanneemt met kans p_j. Er geldt

$$\sigma^2(X) = \mathsf{E}\big(X - \mathsf{E}X\big)^2 = \sum_j (x_j - \mathsf{E}X)^2 p_j.$$

Als voor een zekere waarde x_j geldt dat $x_j \neq \mathsf{E}X$, dan volgt $(x_j - \mathsf{E}X)^2 > 0$, en dus

$$\sigma^2(X) \geq (x_j - \mathsf{E}X)^2 p_j > 0.$$

Wil $\sigma^2(X) = 0$, dan moet $x_j = \mathsf{E}X$ voor alle j, d.w.z. X heeft een ontaarde verdeling en en neemt waarde $\mathsf{E}X$ met kans 1 aan.

iv) Voor elk tweetal getallen a en b geldt

$$\sigma^2(aX + b) = a^2 \cdot \sigma(X).$$

Immers,

$$\sigma^2(aX + b) = \mathsf{E}\big(aX + b - \mathsf{E}(aX + b)\big)^2 = \mathsf{E}\big(aX - a\mathsf{E}X)\big)^2 = a^2\mathsf{E}\big(X - \mathsf{E}X\big)^2 = a^2 \cdot \sigma^2(X).$$

I.h.b. geldt $\sigma^2(X) = \sigma^2(-X)$ (waarom?)!

5.1 Verwachting en variantie van bekende kansverdelingen

Discrete stochasten

1. Ontaarde verdeling.
Stel $\mathsf{P}\{X = x_0\} = 1$. Dan geldt $\mathsf{E}X = x_0\mathsf{P}\{X = x_0\} = x_0$. Voorts hebben we al gezien dat $\sigma^2(X) = 0$.

2. Alternatieve verdeling met parameters p.
Stel X heeft een $\mathsf{alt}(p)$ verdeling. Dan is

$$\mathsf{E}X = 1 \cdot \mathsf{P}\{X = 1\} + 0 \cdot \mathsf{P}\{X = 0\} = 1 \cdot p = p.$$

Omdat $X = X^2$ ($1^2 = 1$ en $0^2 = 0$) geldt verder

$$\sigma^2(X) = \mathsf{E}X^2 - \big(\mathsf{E}X\big)^2 = p - p^2 = p(1-p).$$

3. Binomiale verdeling met parameters n en p.
Stel X heeft een $\mathsf{bin}(n,p)$ verdeling. Dan is

$$\begin{aligned}
\mathsf{E}X &= \sum_{x=0}^{n} x \cdot \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x \cdot \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \\
&\overset{x\binom{n}{x}=n\binom{n-1}{x-1}}{=} n\sum_{x=1}^{n} \binom{n-1}{x-1} \cdot p^x \cdot (1-p)^{n-x} \\
&\overset{n-x=n-1-(x-1)}{=} np \cdot \sum_{x=1}^{n} \binom{n-1}{x-1} \cdot p^{x-1} \cdot (1-p)^{n-1-(x-1)} \\
&\overset{u=x-1}{=} np \cdot \sum_{u=0}^{n-1} \binom{n-1}{u} \cdot p^u \cdot (1-p)^{n-1-u} \\
&= np \cdot 1 = np.
\end{aligned}$$

De éénnalaatste gelijkheid volgt uit het feit dat we de kansen van een $\mathsf{bin}(n-1,p)$ verdeling bij elkaar optellen en daar komt (per definitie) 1 uit. Voor het berekenen van $\mathsf{E}X^2$ moet je gebruiken dat

$$x^2 \cdot \binom{n}{x} = x \cdot n\binom{n-1}{x-1} = n\big((x-1)\binom{n-1}{x-1} + \binom{n-1}{x-1}\big) = n\big((n-1)\binom{n-2}{x-2} + \binom{n-1}{x-1}\big),$$

en dit op dezelfde manier doorrekenen als boven. Dan krijg je $\mathsf{E}X^2 = n(n-1)p^2 + np$, zodat

$$\sigma^2(X) = np(1-p).$$

Een eenvoudiger bewijs komt in het volgende hoofdstuk. Hierbij zullen we gebruik maken van het feit dat X de som is van n onafhankelijke $\mathsf{alt}(p)$ verdeelde stochasten.

4. Hypergeometrische verdeling
X is het aantal objecten met een zekere eigenschap in een trekking van n objecten zonder teruglegging uit een populatie van N. Het aantal objecten met de gewenste eigenschap is R. In dit geval is X te beschrijven als een som van n $\mathsf{alt}(p)$ verdeelde stochasten die *afhankelijk* zijn (zie Hoofdstuk 3 voorbeeld 3.9).

In het volgende hoofdstuk zullen we zien, dat de afhankelijkheid voor het berekenen van de verwachting geen problemen geeft, maar de variantie wel. Daar zullen we aantonen dat

$$\mathsf{E}X = np, \quad \sigma^2(X) = np(1-p) \cdot \frac{N-n}{N-1}.$$

5. Negatieve binomiale verdeling met parameters k en p
Laat X een negatief binomiale verdeling hebben met parameters 1 en p (oftewel een geometrsiche verdeling).

Dan geldt

$$\begin{aligned}
\mathsf{E}X &= \sum_{x\geq 1} x\cdot p\cdot (1-p)^{x-1}\\
&= p\cdot \sum_{x\geq 1} x\cdot (1-p)^{x-1}\\
&= p\cdot \frac{d}{dp}\sum_{x\geq 0} -(1-p)^{x}\\
&= p\cdot \frac{d}{dp}\frac{-1}{1-(1-p)} = p\cdot \frac{d}{dp}\frac{-1}{p}\\
&= \frac{p}{p^2} = \frac{1}{p},
\end{aligned}$$

oftewel de verwachte wachttijd 'tot de eerste gebeurtenis' is $1/p$. Verder geldt

$$\begin{aligned}
\mathsf{E}X^2 &= \sum_{x\geq 1} x^2\cdot p\cdot (1-p)^{x-1}\\
&= p\cdot \sum_{x\geq 1} \big((x+1)x - x\big)\cdot (1-p)^{x-1}\\
&= p\cdot \frac{d^2}{dp^2}\sum_{x\geq 0}(1-p)^{x} - \mathsf{E}X\\
&= p\cdot \frac{d^2}{dp^2}\frac{1}{p} - \frac{1}{p}\\
&= \frac{2p}{p^3} - \frac{1}{p} = \frac{2}{p^2} - \frac{1}{p}.
\end{aligned}$$

Dus volgt: $\sigma^2(X) = 2/p^2 - 1/p - (1/p)^2 = 1/p^2 - 1/p = (1-p)/p^2$.

Het algemene geval van de wachttijd $X^{(k)}$ tot de k-de gebeurtenis gaat als volgt: je kunt $X^{(k)}$ schrijven als som van k onafhankelijk stochasten die ieder de wachttijd tot de eerste volgende gebeurtenis voorstellen. Daarmee kun je na het volgende hoofdstuk gemakkelijk nagaan dat

$$\mathsf{E}X^{(k)} = \frac{k}{p}, \quad \sigma^2(X^{(k)}) = \frac{k(1-p)}{p^2}.$$

6. Poissonverdeling met parameter μ.
Nu geldt dat

$$\begin{aligned}
\mathsf{E}X &= \sum_{x=0}^{\infty} x\cdot \mathsf{P}\{X=x\}\\
&= \sum_{x=0}^{\infty} x\frac{\mu^x}{x!}e^{-\mu}\\
&= \mu\cdot \sum_{x=1}^{\infty} \frac{\mu^{x-1}}{(x-1)!}e^{-\mu}\\
&\overset{u=x-1}{=} \mu\cdot \sum_{u=0}^{\infty} \frac{\mu^{u}}{u!}e^{-\mu}\\
&= \mu,
\end{aligned}$$

want die laatste som is weer een som van Poissonkansen en dus gelijk aan 1. Voor de variantie moeten we $\mathsf{E}X^2$ berekenen. In dit geval is het gemakkelijker om $\mathsf{E}X(X-1)$ te berekenen:

$$\mathsf{E}X(X-1) = \sum_{x=0}^{\infty} x(x-1)\cdot \mathsf{P}\{X=x\}$$

$$
\begin{aligned}
&= \sum_{x=2}^{\infty} x(x-1)\frac{\mu^x}{x!}e^{-\mu} \\
&= \mu^2 \cdot \sum_{x=2}^{\infty} \frac{\mu^{x-2}}{(x-2)!}e^{-\mu} \\
&\overset{u=x-2}{=} \mu^2 \cdot \sum_{u=0}^{\infty} \frac{\mu^u}{u!}e^{-\mu} \\
&= \mu^2.
\end{aligned}
$$

Omdat $\mathsf{E}X^2 = \mathsf{E}X(X-1) + \mathsf{E}X$, geldt $\mathsf{E}X^2 = \mu^2 + \mu$. Hieruit volgt $\sigma^2(X) = \mu^2 + \mu - \mu^2 = \mu$!

Continue verdelingen

1. Uniforme verdeling op (a, b)
Als X een aselecte trekking uit (a, b) is, dan is $Y = (X-a)/(b-a)$ een aselecte trekking uit $(0, 1)$. Oftewel $X = a + (b-a)Y$. Voor Y hadden we verwachting en variantie reeds uitgerekend, nl. $1/2$ en $1/12$. Dus m.b.v. (5.1) volgt

$$\mathsf{E}X = \mathsf{E}\big(a + (b-a)Y\big) = a + (b-a)\frac{1}{2} = \frac{a+b}{2}.$$

M.b.v. eigenschap (iv) van varianties volgt

$$\sigma^2(X) = \sigma^2\big(a + (b-a)Y\big) = \frac{1}{12(b-a)^2}.$$

2. Exponentiële verdeling met parameters λ
Laat X een exponentieel verdeelde stochast zijn met parameter λ. Met behulp van partiële integratie krijgen we

$$
\begin{aligned}
\mathsf{E}X &= \int_0^{\infty} x \cdot \mathsf{f}_X(x)\,dx \\
&= \int_0^{\infty} x \cdot \lambda e^{-\lambda x}\,dx \\
&= \big[-x \cdot e^{-\lambda x}\big]_{x=0}^{x=\infty} - \int_0^{\infty} -e^{-\lambda x}\,dx \\
&= 0 - \big[\frac{1}{\lambda}e^{-\lambda x}\big]_{x=0}^{x=\infty} = \frac{1}{\lambda}.
\end{aligned}
$$

Eveneens geldt na twee maal partieel integreren

$$
\begin{aligned}
\mathsf{E}X^2 &= \int_0^{\infty} x^2 \cdot \lambda e^{-\lambda x}\,dx \\
&= \big[-x^2 \cdot e^{-\lambda x}\big]_{x=0}^{x=\infty} - \int_0^{\infty} -2x \cdot e^{-\lambda x}\,dx \\
&= 0 - \Big(\big[2x \cdot \frac{1}{\lambda}e^{-\lambda x}\big]_{x=0}^{x=\infty} - \int_0^{\infty} 2 \cdot \frac{1}{\lambda}e^{-\lambda x}\,dx\Big) \\
&= 0 - \Big(0 - \big[-2 \cdot \frac{1}{\lambda^2}e^{-\lambda x}\big]_{x=0}^{x=\infty}\Big) = \frac{2}{\lambda^2}.
\end{aligned}
$$

Bijgevolg $\sigma^2(X) = 2/\lambda^2 - (1/\lambda)^2 = 1/\lambda^2$.

3. Gammaverdeling met parameters r **en** λ
Als de parameterwaarde r een geheel getal is, is de Gamma-verdeling $\Gamma(r, \lambda)$ hetzelfde als de verdeling van de som van r exponentiële verdelingen met parameter λ. Wanneer r een niet-geheel getal is, dan is dat natuurlijk niet zo. Verwachting en variantie zijn desalniettemin betrekkelijk gemakkelijk uit te rekenen door de gewenste

integraal weer 'terug te formuleren' tot een integraal over een dichtheid. Deze levert per definitie 1 op. Laat X een $\Gamma(r,\lambda)$-verdeling bezitten. Dan:

$$\begin{aligned}\mathsf{E}X &= \int_0^\infty x\cdot\frac{\lambda^r x^{r-1}e^{-\lambda x}}{\Gamma(r)}\,dx\\ &= \int_0^\infty \frac{\lambda^r x^r e^{-\lambda x}}{\Gamma(r)}\,dx\\ &= \frac{\Gamma(r+1)}{\lambda\Gamma(r)}\int_0^\infty \frac{\lambda^{r+1}x^r e^{-\lambda x}}{\Gamma(r+1)}\,dx\\ &= \frac{\Gamma(r+1)}{\lambda\Gamma(r)},\end{aligned}$$

waarbij we de integraal hebben teruggebracht tot een integraal over de kansdichtheid van de $\Gamma(r+1,\lambda)$-verdeling. Op dezelfde wijze volgt:

$$\mathsf{E}X^2 = \frac{\Gamma(r+2)}{\lambda^2\Gamma(r)}.$$

Bijgevolg is

$$\sigma^2(X) = \frac{1}{\lambda^2}\Big(\frac{\Gamma(r+2)}{\Gamma(r)} - \Big(\frac{\Gamma(r+1)}{\Gamma(r)}\Big)^2\Big).$$

4. Normale verdeling met parameters μ en σ^2.
Laat X een $\mathsf{N}(\mu,\sigma^2)$ verdeelde stochast zijn. Dan is X symmetrisch verdeeld om μ. Derhalve geldt $\mathsf{E}X=\mu$.

Laten we nu eerst de variantie berekenen voor een standaard normaal verdeelde stochast Y. Partieel integreren geeft:

$$\begin{aligned}\mathsf{E}Y^2 &= \int_{-\infty}^\infty y^2\cdot\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\,dy\\ &= \big[-y\cdot\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\big]_{y=-\infty}^{y=\infty} - \int_{-\infty}^\infty \frac{-1}{\sqrt{2\pi}}e^{-y^2/2}\,dy\\ &= 0--1=1.\end{aligned}$$

Dus is $\sigma^2(Y)=\mathsf{E}Y^2-0=1$.

We weten dat $Z=(X-\mu)/\sigma$ standaard normaal verdeeld is. Vanwege eigenschap (iv) van varianties en de relatie $X=\sigma Z+\mu$ krijgen we dat $\sigma^2(X)=\sigma^2$. De parameters van de normale verdeling zijn dus precies verwachting en variantie!

5. Cauchyverdeling
Stel de stochast X heeft een Cauchy-verdeling. Dan heeft X geen eindige verwachting. Immers:

$$\int_{-M}^N x\cdot\mathsf{f}_X(x)\,dx = \int_{-M}^N x\cdot\frac{1}{\pi(1+x^2)}\,dx = \big[\frac{1}{2\pi}\ln(1+x^2)\big]_{x=-M}^{x=N} = \frac{1}{2\pi}\ln\frac{1+N^2}{1+M^2}$$

en dit convergeert niet voor alle rijen $M,N\to\infty$ naar hetzelfde getal! De variantie kan dus ook niet bestaan!

6. Chi-kwadraat verdeling met r vrijheidsgraden
Stel X heeft een χ-kwadraatverdeling met r vrijheidsgraden. Dan is X te verkrijgen als de som van de kwadraten van r onafhankelijk kwadraten van standaarnormale verdelingen.

Na het volgende hoofdstuk kun je uitrekenen dat $\mathsf{E}X=r$. Als daarbij ook nog een aantal malen partieel integreert krijg je dat $\sigma^2(X)=2r$.

7. Betaverdeling met parameters α en β
Stel X heeft een Betaverdeling met parameters $\alpha>0$ en $\beta>0$. Laten we aannemen dat α een geheel getal is. Dan krijgen we met dezelfde soort truc als voor de Gammaverdeling dat

$$\mathsf{E}X=\frac{\alpha}{\alpha+\beta},\quad \sigma^2(X)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Het rekenen aan verwachtingen en varianties
In het bovenstaande zijn er enkele algemene truccendozen gebruikt om verwachtingen en varianties uit de rekenen:

i) Wanneer de verdeling symmetrisch is, volgt de verwachting onmiddellijk zonder verder rekenen.

ii) Transformatie naar een stochast met een verdeling uit dezelfde parameterfamilie, maar dan met eenvoudigere parameters, liefst 0 of 1 of iets dergelijks.

iii) De te berekenen integraal zo herschrijven dat er een aantal constante termen te isoleren zijn, en het overgeblevene de integraal over een kansdichtheid is (uit dezelfde parameterfamilie, maar misschien met andere parameters dan die waarmee je begon). Ook al is deze integraal lastig, je weet op grond van het feit dat het om een kansdichtheid gaat, dat er 1 uit moet komen.

iv) Soms kun je de stochast beschrijven als een som van bijv. n stochasten met bekende verwachting μ en variantie σ^2. Met technieken uit het volgende hoofdstuk volgt dat de gevraagde verwachting gelijk is aan $n\mu$; bovendien, als de stochasten onderling onafhankelijk zijn, dan volgt dat de variantie gelijk is aan $n\sigma^2$.

Er is nog een andere methode die in de praktijk vaak werkt. We hadden gezien dat het handig kan zijn te conditioneren op gebeurtenissen, waardoor bepaalde kansen gemakkelijker uit te rekenen zijn. Hierbij maakten we gebruik van het *behoud van volledige kans* (zie Hoofdstuk 3).

We kunnen eenzelfde procedure ook volgen voor het uitrekenen van verwachtingen. Dit geldt voor algemene stochasten, maar zullen het hier alleen laten zien voor *discrete stochasten*.

Voorwaardelijke verwachtingen van discrete stochasten
Stel $B_1, B_2, \ldots$ is een partitie van de uitkomstenruimte Ω. Dan is $\mathsf{P}\{X = x_j \mid B_k\} = \mathsf{P}\{X = x_j \cap B_k\}/\mathsf{P}\{B_k\}$, de voorwaardelijke kans dat $X = x_j$ gegeven B_k.

We kunnen ook de *voorwaardelijke verwachting* van X gegeven B_k definiëren:

$$\mathsf{E}\{X \mid B_k\} = \sum_j x_j \mathsf{P}\{X = x_j \mid B_k\}.$$

Uit de *behoudswet van totale kans* volgt nu dat

$$\begin{aligned}
\mathsf{E}X &= \sum_j x_j \mathsf{P}\{X = x_j\} \\
&= \sum_j x_j \big(\sum_k \mathsf{P}\{X = x_j \mid B_k\} \mathsf{P}\{B_k\} \big) \\
&= \sum_k \Big(\sum_j x_j \mathsf{P}\{X = x_j \mid B_k\} \Big) \mathsf{P}\{B_k\} \\
&= \sum_k \mathsf{E}\{X \mid B_k\} \mathsf{P}\{B_k\}
\end{aligned}$$

(mits de sommen absoluut convergeren).

Laten we aan de hand van een voorbeeldje zien wat je hiermee op schiet.

Voorbeeld 5.6 In opgave 3.14 definiëren we X als het aantal jonge vogels in het nest. Wat is $\mathsf{E}X$?

Gegeven dat je weet hoeveel eieren een vogel heeft gelegd, kun je het verwachte aantal jonge vogels berekenen, want een ei komt uit met kans 1/2, onafhankelijk van de andere eieren. Zeg $B_k = \{\text{nest van k eieren}\}$. Dan heeft het aantal uitgekomen eieren een $\mathsf{bin}(k, 1/2)$ verdeling. Deze heeft verwachting $k/2$, d.w.z.

$$\mathsf{E}\{X \mid B_k\} = \frac{k}{2}.$$

De kansen op B_k zijn gegeven in de opgave. Die invullen geeft

$$\mathsf{E}X = \sum_k \mathsf{E}\{X \mid B_k\} \mathsf{P}\{B_k\} = \frac{1}{2} \cdot \frac{1}{10} + \frac{2}{2} \cdot \frac{2}{10} + \frac{3}{2} \cdot \frac{3}{10} + \frac{4}{2} \cdot \frac{3}{10} + \frac{5}{2} \cdot \frac{1}{10} = \frac{31}{20}.$$

In alle andere gevallen zul je toch echt moeten gaan rekenen!

5.2 Simulatie van verwachting en variantie, Monte Carlo integratie

Laat een stochast $Y = g(X)$ gegeven zijn en we willen het populatiegemiddelde (theoretisch gemiddelde) $\mathsf{E}g(X)$ weten. Als dat analytisch niet te doen is, dan kun je proberen het gewenste getal via simulatie te schatten. Daartoe doe je een flink aantal (n) trekkingen uit X, zeg $X_1, \ldots, X_n$. Hiermee krijg je trekkingen $Y_1 = g(X_1), \ldots, Y_n = g(X_n)$ uit Y.

Met behulp van de wet der grote aantallen volgt (onder redelijke aannamen) dat

$$\bar{Y}(n) = \frac{Y_1 + \cdots Y_n}{n} \to \mathsf{E}Y,$$

met kans 1, voor $n \to \infty$. Dat wil zeggen dat het steekproefgemiddelde $\bar{Y}(n)$ convergeert naar het populatiegemiddelde $\mathsf{E}Y$. We noemen $\bar{Y}(n)$ dan ook een *schatter* van $\mathsf{E}Y$.

Merk wel op dat het steekproefgemiddelde een *stochast* is (elke simulatie levert een ander getal!), terwijl het populatiegemiddelde een *getal* is. Wel zullen we in Hoofdstuk 6 zien, dat $\mathsf{E}\bar{Y}(n) = \mathsf{E}Y$!

Met dit gegeven hebben we feitelijk al eerder schattingen gemaakt van verwachtingen: zie §1.9. Een ander voorbeeldje is een schatting van π, gebaseerd op voorbeeld 5.2.

Voorbeeld 5.7 Laat X een aselecte trekking uit $(0,1)$ zijn. Dan is $\mathsf{E}4\sqrt{1-X^2} = \pi$. We simuleren $n = 10000$ trekkingen uit X. Dit geeft 10000 trekkingen uit $4\sqrt{1-X^2}$ en het steekproefgemiddelde wordt een schatter van π.

```
>>n=10000;
>>x=rand(1,n);
>>y=4*sqrt(1-x.^2);
>>mean(y)-pi
 %berekent verschil tussen steekproefgemiddelde y en pi
```

Het antwoord lijkt overigens nog nergens op!

Zoals gezegd, kun je deze methode gebruiken om analytisch niet te berekenen integralen te schatten: stel gevraagd $I = \int_a^b g(x)dx$ voor een gecompliceerde functie g. Dan is

$$I = \mathsf{E}g(X_{a,b}) \cdot (b-a),$$

waarbij $X_{a,b}$ homogeen verdeeld is op (a,b) (ga dit na!). Dus kunnen we de gevraagde integraal schatten door trekkingen uit $g(X_{a,b})$ te doen via trekkingen uit $X_{a,b}$ en vervolgens het steekproefgemiddelde te schalen met een factor $(b-a)$. Deze methode heet *Monte Carlo integratie* en wordt veel in de fysica toegepast. De methode is vooral efficiënt bij hoog-dimensionale problemen (veel-deeltjes systemen, het web etc). Voor laagdimensionale problemen zijn veelal numerieke methoden efficiënter.

Het probleem is dat we nog geen kwalitatieve uitspraken hebben over de schatter. Hierbij speelt de Centrale Limietstelling een rol en dat komt later aan de orde.

Een andere mogelijkheid om een te schatten, beschrijven we hieronder voor het probleem van het schatten van π. Dit is een feite een acceptatie-rejectie methode zoals we in het vorige hoofdstuk al zijn tegengekomen (je kunt het ook zien als Monte-Carlo integratie maar dan voor een functie van twee variabelen).

Voorbeeld 5.8 Op het interval $(0,1)$ is het bereik van de functie $\sqrt{1-x^2}$ eveneens het interval $(0,1)$. Laten X en Y beide homogeen verdeeld zijn op $(0,1)$. We doen n onafhankelijke trekkingen uit het paar (X,Y). Feitelijk doen we aselecte trekkingen uit het eenheidsvierkant $(0,1) \times (0,1)$.

Merk op dat $y \leq \sqrt{1-x^2}$ dan en slechts dan als $x^2 + y^2 \leq 1$. Dus een trekking (x,y) ligt op het gevraagde oppervlak wanneer $x^2 + y^2 \leq 1$ en dan accepteren we de trekking. Het aantal geaccepteerde trekkingen gedeeld door n benadert steeds beter het getal

$$\frac{\pi}{4 \cdot \text{oppervlakte eenheidsvierkant}} = \frac{\pi}{4}.$$

Dit simuleren we als volgt.

```
>>n=10000;
>>x=rand(1,n);
>>y=rand(1,n);
>>f=x.^2+y.^2;
>>hits=sum(f<=1);
>>v=4*length(hits)/n;
```

Dit kun je als volgt op een aardige manier plotten. We definiëren een functie die dat doet, waarbij we de lengte van de simulatierun n als parameter meegeven.

```
>>function v=plot_shots(n);
>>%v=plot_shots(n) computes an approximation to Pi, using n pairs
>>%of samples from the uniform distribution on 0..1 to cover the unit square
>>%in the first quadrant, and counting hits inside the quarter circle.
>>x=rand(1,n);
>>y=rand(1,n);
>>f=x.^2+y.^2;
>>hit=find(f<=1);
>>miss=find(f>1);
>>v=4*length(hit)/n;
>>xc=0:.01:1:
>>yc=sqrt(1-xc.^2);
>>hold off
>>plot(xc,yc,'k',x(hit),y(hit),'+g',x(miss),y(miss),'+r','linewidth',2)
>>axis square
```

Een analoge methode is overigens in vorige eeuwen lijfelijk uitgevoerd door duizenden malen op een vierkant papier met ingeschreven cirkel en een fijn grid naalden te gooien. Vervolgens werd het gemiddeld aantal keren geteld dat de naald het grid raakte. Dit zette duidelijk geen zoden aan de dijk!

Hoe kun je tenslotte een schatting maken van de variantie?

Laat $X_1, \ldots, X_n$ een steekproef zijn uit de stochast X, met (populatie)variantie $\sigma^2 = \mathrm{var}(X)$. We noemen dan

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

de *steekproefvariantie*. Volgens de wet van de grote aantallen geldt dat S^2 ongeveer gelijk is aan σ^2 als n groot is. We noemen $S = \sqrt{S^2}$ de steekproefstandaarddeviatie. Deze ligt in de buurt van σ, voor n groot.

Merk op dat we bij de berekening van S^2 door $n-1$ delen en niet door n. Hier zijn theoretische gronden voor (delen door n betekent vaak een onderschatting van de theoretische variantie σ^2). Voor grote n maakt het natuurlijk niet zoveel uit, en kun je dus ook

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}(n))^2}{n}$$

als benadering gebruiken. We noemen S^2 en $\hat{\sigma}^2$ *schatters* van σ^2 (en S en $\hat{\sigma}$ schatters van σ).

5.3 Opgaven Hoofdstuk 5

Opgave 5.1 Toon aan dat $\lambda \cdot X$ een exponentiële verdeling heeft met parameter 1 als X een exponentiële verdeling heeft met parameter λ.

Opgave 5.2 Een stochast X heeft de volgende dichtheid

$$\mathsf{f}(x) = \begin{cases} 1+x, & -1 \le x \le 0, \\ 1-x, & 0 \le x \le 1 \\ 0, & x > 1 \text{ of } x < 0. \end{cases}$$

Bereken

i) $\mathsf{P}\{|X| > 1/2\}$, $\mathsf{P}\{X < -2\}$;

ii) $\mathsf{E}X$;

iii) $\sigma^2(X)$.

Opgave 5.3 Stel X is $\mathsf{N}(10, 36)$ verdeeld, maar je hebt geen rekentuig bij de hand en slechts een tabel van de standaard-normale verdeling (d.w.z. de tabel van de verdelingsdfunctie in positieve getallen). Hoe moet je dan de volgende kansen berekenen: $\mathsf{P}\{X < 20\}$, $\mathsf{P}\{X > 16\}$ en $\mathsf{P}\{|X - 10| > 6\}$?

Opgave 5.4 Stel X heeft verdeling $\mathsf{P}\{X = 1\} = 1/6$, $\mathsf{P}\{X = 2\} = 2/6$, $\mathsf{P}\{X = 3\} = 1/6$ en $\mathsf{P}\{X = 6\} = 2/6$. Bereken $\mathsf{E}X$ en $\sigma^2(X)$.

Opgave 5.5 Een vaas bevat 2 groene, 4 blauwe en 4 rode (genummerde) knikkers. Men trekt aselect knikkers uit de vaas. Hoe lang duurt het gemiddeld totdat men voor het eerst een blauwe knikker heeft getrokken bij

i) trekkingen met terugleggen,

ii) trekkingen zonder terugleggen?

Opgave 5.6 Gooi met twee dobbelstenen. Bereken verwachting en variantie van het totaal aantal ogen. Stel nu: X is het aantal gegooide ogen op dobbelsteen 1 en Y het aantal op dobbelsteen 2. Wat is $\mathsf{E}(X+Y)$, $\mathsf{E}XY$ en $\sigma^2(X+Y)$? Valt je hierbij iets op?

Opgave 5.7 Stel X bezit de Poissonverdeling met parameter μ. Laat voor $k = 0, 1, 2, \ldots$ zien dat

$$\mathsf{E}\big(X(X-1)(X-2)\cdots(X-k)\big) = \mu^{k+1}.$$

Opgave 5.8

Laat X het aantal keren gooien met een munt zijn, tot men voor het eerst n keer achter elkaar kruis heeft gegooid. Bereken $\mathsf{E}X$.

Opgave 5.9 Laat X een stochastische grootheid zijn met waarden in de natuurlijke getallen $0, 1, 2, \ldots$ Laat zien dat

$$\mathsf{E}X = \sum_{k=0}^{\infty} \mathsf{P}\{X > k\}.$$

Opgave 5.10 Stel X is standaard normaal verdeeld. Bepaal $\mathsf{E}e^{2X}$.

Opgave 5.11 Laat X een stochastische grootheid zijn met dichtheid $\mathsf{f}(x) = (1/2)e^{-|x|}$, $x \in \mathbf{R}$. Bepaal verwachting en variantie van X.

Opgave 5.12 Stel X is een aselecte trekking uit (a, b). Bereken verwachting en variantie van X door uit te rekenen:

$$\int_a^b x \cdot \mathsf{f}_X(x)\,dx, \quad \int_a^b x^2 \cdot \mathsf{f}_X(x)\,dx.$$

Opgave 5.13 Reistijden bij de NS zijn afhankelijk van het weer. Bij veel gevallen blaadjes in de herfst nemen deze dramatisch toe.

Laten X en Y en Z onafhankelijke een homogene verdelingen zijn op $(0, 1)$. Na een herfststorm is de vertraging op het traject Amsterdam-Leiden een random getal tussen 0 en 3 uur, en is dus verdeeld als $3X$. Bij anderzins zwaar weer is de vertraging verdeeld als $1,5Y$, en bij 'normaal' weer, is de vertraging verdeeld als $Z/3$. De fractie van de tijd per jaar dat er herfststormen zijn, is $1/24$. De fractie van de tijd per jaar dat er anderszins zwaar weer is, is $1/2$ en de fractie van de tijd dat het weer normaal is, is $1 - 1/24 - 1/2 = 11/24$. Wat is de gemiddelde vertraging per jaar?

Opgave 5.14 Een gokker bedenkt een nieuw gokspelletje: hij trekt aselect uit de getallen $3/10, \cdots, 9/10$ een getal, zeg p. Dan gooit hij een onzuivere munt op met kans p op Kop. Als Kop bovenkomt krijgt hij a euro's, als Munt bovenkomt moet hij 2a euro's betalen. Bereken zijn verwachte opbrengst. Kan zijn verwachte winst positief zijn voor een of andere waarde van a? Zo ja, voor welke?

Opgave 5.15 Stel X heeft een Cauchyverdeling. Dan kun je X maken als quotiënt van twee onafhankelijke standaardnormaal $\mathsf{N}(0,1)$ verdeelde stochasten Y en Z: $X = Y/Z$. Deze bewering hebben we niet bewezen, dus die willen we controleren via simulatie. Simuleer een flink aantal trekkingen Y/Z m.b.v. trekkingen uit Y en Z. Plot deze in een histogram. Plot hierover de dichtheid f_X na deze op correcte manier te hebben geschaald. Lijkt dit enigszins?

Ga via simulatie na dat $\mathsf{E}X$ niet bestaat.

Opgave 5.16 Laat X een $\Gamma(r,\lambda)$-verdeling hebben. Bereken $\sigma^2(X)$.

Opgave 5.17 Stel X heeft een Beta-verdeling met parameters α en β, waarbij α een geheel getal is. Ga na dat $\mathsf{E}X = c_{\alpha,\beta}/c_{\alpha+1,\beta} = \alpha/(\alpha+\beta)$ en dat $\mathsf{E}X^2 = c_{\alpha,\beta}/c_{\alpha+2,\beta} = \alpha(\alpha+1)/(\alpha+\beta)(\alpha+\beta+1)$ (zie Hoofdstuk 4). Toon aan, dat $\sigma^2(X) = \alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$.

Ga met Matlab na dat deze uitdrukkingen kloppen voor $\alpha = 2.5$ en $\beta = 4$:

```
>>[mu,s2]=betastat(2.5,4)
```

Ga dit ook na m.b.v. simulatie van een flink aantal trekkingen uit de beta verdeling met de gewenste parameters.

Opgave 5.18 We willen een beetje gevoel krijgen hoe snel het steekproefgemiddelde al 'redelijk' in de buurt ligt van de te schatten verwachting. Volgens de Centrale Limietstelling (komt later) heeft een verschil van meer dan $2\sigma/\sqrt{n}$ minder dan 5% kans. Voor het schatten van π in Voorbeeld 5.7, betekent dit dat het aantal goede decimalen, in de orde ${}_{10}\log(\sqrt{n}/2\sigma) - 1 \approx {}_{10}\log(\sqrt{n}) - 1$ ligt. Ga dit na voor beide schattingsmethoden van π uit voornoemd voorbeeld. Hoe groot moet de steekproef zijn om 10 decimalen goed te hebben?

6 Stochastische vectoren

Bij een toevalsexperiment neemt men vaak meer dan één variabele tegelijk in beschouwing. Zo zou een effectenhandelaar naast de maximale waarde ook de minimale waarde van een bepaald aandeel gedurende een gegeven periode kunnen willen weten. Men kan daarbij ook geïnteresseerd zijn in samenhang tussen de variabelen.

We beperken ons tot twee stochastische variabelen X en Y. Het paar (X, Y) kan gezien worden als een $\mathbf{R}^2$-waardige functie op de onderliggende uitkomstenruimte Ω en wordt een (in dit geval twee-dimensionale) *stochastische vector* genoemd. Generalisatie naar meer dimensies is analoog.

Twee stochasten X en Y hebben een *simultane verdeling* als alle kansen van de vorm

$$\mathsf{P}\{X \in A, Y \in B\},$$

gedefinieerd zijn. Als X en Y op dezelfde onderliggende kansruimte $(\Omega, \mathcal{A}, \mathsf{P}\}$ (de collectie deelverzamelingen $\mathcal{A}$ zijn de 'basis' verzamelingen die de kansverdeling P definiëren) zijn gedefinieerd, dan is dit automatisch het geval.

Een simultane kansverdeling van X en Y is dus eigenlijk een kansverdeling op $\mathbf{R}^2$, die voor te stellen is als een fysische massaverdeling op het platte vlak.

De *marginale verdelingen* van (X, Y) zijn de verdelingen van X en Y. De marginale verdelingen zijn uit de simultane verdeling af te leiden door uitsommeren en/of uitintegreren. Omgekeerd, kun je uit de marginale verdelingen van X en Y de simultane verdeling niet 'terughalen' zolang je de relatie tussen X en Y niet kent!

We kunnen ook een *simultane verdelingsfunctie* $\mathsf{F}_{X,Y}$ van de vector (X, Y) definiëren door

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

De marginale verdelingsfuncties van de coördinaten X en Y zijn dan:

$$\mathsf{F}_X(x) = \mathsf{F}_{X,Y}(x, \infty) = \lim_{y \to \infty} \mathsf{F}_{X,Y}(x, y), \quad \mathsf{F}_Y(y) = \mathsf{F}_{X,Y}(\infty, y).$$

Discrete simultane kansverdelingen
In het *discrete geval* betekent dit dat de *simultane kansverdeling*

$$\mathsf{P}\{X = x, Y = y\}$$

voor een eindige of aftelbare collectie waarden van x en y bestaat, zó dat

$$\sum_{x,y} \mathsf{P}\{X = x, Y = y\} = \sum_x \sum_y \mathsf{P}\{X = x, Y = y\} = 1.$$

De kansen op gebeurtenissen $(X, Y) \in B$ voor $B \subset \mathbf{R}^2$ worden berekend door te sommeren over alle punten $(x, y) \in B$ die in het waardebereik van X en Y liggen

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

De marginale kansverdelingen van X en Y krijg je door uitsommeren:

$$\mathsf{P}\{X = x\} = \sum_y \mathsf{P}\{X = x, Y = y\}, \quad \mathsf{P}\{Y = y\} = \sum_x \mathsf{P}\{X = x, Y = y\}.$$

De verdelingsfunctie $\mathsf{F}_{(X,Y)}$ heeft in het discrete geval de gedaante van een terras-functie, met oplopende terrassen wanneer x en/of y toenemen.

Voorbeeld 6.1 Laat X het aantal jongens en Y het aantal meisjes zijn in een aselect gekozen Nederlands huishouden. De kansen waarmee simultane waarden van X en Y worden aangenomen (corresponderend met

de relatieve frequenties) zijn (afgerond op 3 decimalen)

$y\backslash x$	0	1	2	3	4		
0	0,16	0,12	0,08	0,032	0,008	0,40	marginale
1	0,12	0,09	0,06	0,024	0,006	0,30	kansen
2	0,08	0,06	0,04	0,016	0,004	0,20	$\mathsf{P}\{Y=y\}$
3	0,032	0,024	0,016	0,0006	0,002	0,08	
4	0,008	0,006	0,004	0,002	0,000	0,02	
	0,40	0,30	0,20	0,08	0,02		

marginale kansen $\mathsf{P}\{X=x\}$

Uit het kanstableau valt direct af te lezen dat

$$\begin{aligned}\mathsf{P}\{\text{3 kinderen}\} &= \mathsf{P}\{X+Y=3\}\\ &\quad 0,032+0,06+0,06+0,032=0,184\end{aligned}$$

Wat is de kans op minstens één meisje?

Het aantal kinderen $Z=X+Y$ heeft als verwachte waarde:

$$\begin{aligned}\mathsf{E}Z &= 1\cdot(0,12+0,12)+2\cdot(0,08+0,09+0,08)+3\cdot(0,184)\\ &\quad +4\cdot(0,008+0,024+0,04+0,024+0,008)+5\cdot(0,006+0,016+0,016+0,006)\\ &\quad +6\cdot(0,004+0,006+0,004)+7\cdot(0,002+0,002)\\ &= 0,24+0,50+0,552+0,416+0,22+0,084+0,028=2,04\end{aligned}$$

Merk op dat dit gelijk is aan $\mathsf{E}X+\mathsf{E}Y$ (ietsje prettiger om te berekenen!)

Voorbeeld 6.2 Men gooit 5 maal met een zuivere dobbelsteen. Zij X het aantal gegooide enen en Y het aantal gegooide tweeën. Dan is $\mathsf{P}\{X=0,Y=0\}=4^5/6^5$ (aantal 'goede rijtjes'/totaal aantal rijtjes). Verder, $\mathsf{P}\{X=0,Y=1\}=\binom{5}{1}\cdot 4^4/6^5$ (voor het aantal 'goede rijtjes' trek je eerst de plaats waarop je een 1 zet, en vervolgens kun je elke andere plaats vullen met een keuze uit 4 (nl.3,4,5,6)). In het algemeen geldt

$$\mathsf{P}\{X=x,Y=y\}=\frac{\binom{5}{x}\binom{5-x}{y}4^{5-x-y}}{6^5}=\frac{5!}{x!y!(5-x-y)!}\Big(\frac{1}{6}\Big)^{x+y}\Big(\frac{4}{6}\Big)^{5-x-y}.$$

Het aantal goede rijtjes krijgen we als volgt: kies eerst uit 5 plaatsen de x plaatsen waarop een 1 wordt gezet; uit de $5-x$ overige kiezen we er y waarop een 2 wordt gezet; voor de overige 4 plaatsen hebben we 4^4 mogelijkheden om getallen uit de verzameling $3,4,5,6$ te zetten.

Wat zijn de marginale verdelingen van X en Y?

Voorbeeld 6.3 (Multinomiale verdeling) In plaats van alleen het aantal énen en tweeën bij het gooien van een zuivere dobbelsteen, willen we de frequenties van alle ogen-aantallen bijhouden.

De algemene opzet van dit probleem is als volgt: stel gegeven een m-zijdige dobbelsteen met kans p_i op uitkomst i, $\sum_i p_i=1$. Het aantal keren dat uitkomst i voorkomt in n 'worpen' geven we aan met X_i, $i=1,\ldots,m$. Automatisch geldt dan $X_1+\cdots X_m=n$ (waarom?). Gevraagd: $\mathsf{P}\{X_1=x_1,\ldots,X_m=x_m\}$ $(x_1+\cdots+x_m=n)$. Elke rijtje met x_1 enen, x_2 tweeën, etc, heeft evenveel kans om voor te komen, namelijk kans $p_1^{x_1}\cdots p_m^{x_m}$. De gevraagde kans is dit product vermenigvuldigd met het aantal 'goede rijtjes'. Het aantal 'goede rijtjes' is volgens formule **E** uit Hoofdstuk 2 de multinomiaalcoëfficiënt

$$\binom{n}{x_1,\ldots,x_m}=\frac{n!}{x_1!\cdots x_m!}.$$

Derhalve hebben we gevonden

$$\mathsf{P}\{X_1=x_1,\ldots,X_m=x_m\}=\binom{n}{x_1,\ldots,x_m}p_1^{x_1}\cdots p_m^{x_m},$$

en we zeggen dat $(X_1,\ldots,X_m)$ de *multinomiale verdeling* heeft met parameters n en $(p_1,\ldots,p_m)$.

Wat is de marginale verdeling van X_i? Het aantal keren dat we uitkomst i in n worpen hebben, moet $\mathsf{bin}(n, p_i)$ verdeeld zijn (waarom?). Krijgen we dit ook door uitsommeren? Controle voor X_1 levert:

$$\begin{aligned}
\mathsf{P}\{X_1 = x_1\} &= \sum_{x_2,\ldots,x_m} \mathsf{P}\{X_1 = x_1, \ldots, X_m = x_m\} \\
&= \sum_{x_2,\ldots,x_m} \frac{n!}{x_1!x_2!\cdots x_m!} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m} \\
&= \frac{n!}{x_1!(n-x_1)!} p_1^{x_1} \sum_{x_2,\ldots,x_m} \frac{(n-x_1)!}{x_2!\cdots x_m!} p_2^{x_2} \cdots p_m^{x_m} \\
&= \binom{n}{x_1} p_1^{x_1} \sum_{x_2,\ldots,x_m} \binom{n-x_1}{x_2,\ldots,x_m} p_2^{x_2} \cdots p_m^{x_m} \\
&= \binom{n}{x_1} p_1^{x_1} (p_2 + \cdots + p_m)^{n-x_1} = \binom{n}{x_1} p_1^{x_1} (1-p_1)^{n-x_1}.
\end{aligned}$$

Inderdaad, maar met een handige redenering kom je sneller tot dit antwoord!

Continue simultane verdelingen

In het *continue geval* bestaat de *simultane kansdichtheid* $\mathsf{f}_{X,Y}$ z'odat voor elke 'nette' (meetbare) deelverzameling $B \subset \mathbf{R}^2$

$$\mathsf{P}\{(X,Y) \in B\} = \iint_{(x,y)\in B} \mathsf{f}_{X,Y} d(x,y).$$

Voor een rechthoekje $B = (x_1, x_2) \times (y_1, y_2)$ geldt

$$\mathsf{P}\{(X,Y) \in B\} = \mathsf{P}\{x_1 \le X \le x_2, y_1 \le Y \le y_2\} = \int_{x_1}^{x_2} \int_{y_1}^{y_2} \mathsf{f}_{X,Y}(x,y) dy dx.$$

Het maakt niet uit of je de zijden van het rechthoekje in de integraal meeneemt of niet: de kans op een 'zijde' is toch 0 (waarom?)! Het berekenen van deze dubbele integraal komt neer op het twee maal uitrekenen van een enkele integraal, waarbij het niet uitmaakt of je eerst naar x of eerst naar y integreert.

De simultane dichtheid voldoet ook aan

$$\mathsf{f}_{X,Y}(x,y) \ge 0,\ (x,y) \in \mathbf{R}^2, \quad \iint_{\substack{-\infty<x<\infty \\ -\infty<y<\infty}} \mathsf{f}_{X,Y}(x,y) d(x,y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathsf{f}_{X,Y}(x,y) dx dy = 1.$$

De *marginale dichtheden* van X en Y krijg je door uit te integreren over de andere variabele. Omdat geldt (Y moet 'ergens' zijn)

$$\mathsf{P}\{x_1 \le X \le x_2\} = \mathsf{P}\{x_1 \le X \le x_2, -\infty < Y < \infty\} = \int_{x_1}^{x_2} \Big(\int_{-\infty}^{\infty} \mathsf{f}_{X,Y}(x,y) dy \Big) dx,$$

moet gelden dat $\mathsf{f}_X(x) = \int_{-\infty}^{\infty} \mathsf{f}_{X,Y}(x,y) dy$. Evenzo geldt $\mathsf{f}_Y(y) = \int_{-\infty}^{\infty} \mathsf{f}_{X,Y}(x,y) dx$.

Uit de simultane kansdichtheid, kun je de simultane kansverdeling bepalen:

$$\mathsf{F}_{X,Y}(x,y) = \mathsf{P}\{X \le x, Y \le y\} = \int_{-\infty}^{x} \int_{-\infty}^{y} \mathsf{f}_{X,Y}(u,v) dv du.$$

Omgekeerd is i.h.a. de simultane kansdichtheid de afgeleide naar beide variabelen van de kansverdeling:

$$\mathsf{f}_{X,Y}(x,y) = \frac{\delta}{\delta x} \frac{\delta}{\delta y} \mathsf{F}_{X,Y}(x,y) = \frac{\delta^2}{\delta x \delta y} \mathsf{F}_{X,Y}(x,y).$$

Voorbeeld 6.4 Stel (X,Y) is continu verdeeld met dichtheid $\mathsf{f}_{X,Y}(x,y) = x + y$ voor $x, y \in [0,1]$ en 0 elders. Is dit inderdaad een dichtheid? $\mathsf{f}_{X,Y} \ge 0$ en

$$\int_{x=0}^{1} \int_{y=0}^{1} (x+y) dy dx = \int_{x=0}^{1} \big[xy + \tfrac{1}{2}y^2\big]_{y=0}^{y=1} dx = \int_{x=0}^{1} (x + \tfrac{1}{2}) dx = \Big[\tfrac{1}{2}x^2 + \tfrac{1}{2}x\Big]_{x=0}^{x=1} = 1.$$

De marginale dichtheden van X en Y zijn gegeven door: $\mathsf{f}_X(x) = \int_0^1 \mathsf{f}_{X,Y}(x,y) dy = \int_0^1 (x+y) dy = x + 1/2$, $0 \le x \le 1$; evenzo geldt $\mathsf{f}_Y(y) = y + 1/2$, $0 \le y \le 1$.

Voorbeeld 6.5 Stel we willen een aslecte trekking doen uit de rechthoek $[z,b]\times[c,d]$. Een trekking heeft twee componenten, dus we hebben hier te maken een twee-dimensionale stochastische vector (X,Y). Aangezien elke trekking even aannemelijk is, is de dichtheid constant op de rechthoek en deze moet dus gelijk zijn aan

$$\mathsf{f}_{X,Y}(x,y) = \begin{cases} \frac{1}{(b-a)(d-c)}, & a \le x \le b, c \le y \le d \\ 0, & \text{anders.} \end{cases}$$

Dan geldt voor $a \le x_1 \le x_2 \le b$ en $c \le y_1 \le y_2 \le d$

$$\begin{aligned} \mathsf{P}\{x_1 \le X \le x_2, y_1 \le Y \le y_2\} &= \int_{x_1}^{x_2}\int_{y_1}^{y_2} \frac{1}{(b-a)(d-c)} dydx \\ &= \frac{x_2-x_1)(y_2-y_1)}{(b-a)(d-c)} = \frac{\text{oppervlakte } [x_1,x_2]\times[y_1,y_2]}{\text{oppervlakte } [a,b]\times[c,d]}. \end{aligned}$$

Ga na dat de marginale verdeling van X de homogene verdeling op $[a,b]$ is en die van Y de homogene verdeling op $[c,d]$.

Stel nu $a=0=c$ en $b=1=d$. Gevraagd: $\mathsf{P}\{X^2+Y^2 \le 1\}$. Je kunt nu de integraal analytisch uitrekenen. Gezien de relatie tussen kans en oppervlakte, volgt echter rechtstreeks

$$\begin{aligned} \mathsf{P}\{X^2+Y^2 \le 1\} &= \frac{\text{oppervlakte van de punten } (x,y) \text{ met } x^2+y^2 \le 1, 0 \le x,y \le 1}{(1-0)(1-0)} \\ &= \frac{\text{oppervlakte kwartcirkel}}{1^2} = \frac{\pi}{4}. \end{aligned}$$

N.B. Kansen voor een drie-dimensionale homogeen verdeelde stochastische vector zijn dus *inhouden*!

6.1 Onafhankelijkheid en voorwaardelijke verwachting [1]

De definitie voor onafhankelijkheid van stochasten hebben we al eerder genoemd, maar wordt hier herhaald.

De stochasten $X_1,\ldots,X_n$ heten onderling onafhankelijk wanneer voor elke collectie 'nette' deelverzamelingen $B_1,\ldots,B_n \subset \mathbf{R}$ geldt dat

$$\mathsf{P}\{X_1 \in B_1,\ldots,X_n \in B_n\} = \mathsf{P}\{X_1 \in B_1\}\cdot\mathsf{P}\{X_2 \in B_2\}\cdots\mathsf{P}\{X_n \in B_n\},$$

in woorden: hun simultane verdeling factoriseert.

Onafhankelijke stochasten komen met name voor wanneer we n onafhankelijk experimenten doen. Aan de simultane verdeling is gemakkelijk te zien of de componenten onafhankelijk zijn. Dat blijkt uit de volgende caracteriseringsstelling.

Theorem 6.1 *Stochastische grootheden zijn dan en slechts dan onafhankelijk wanneer hen simultane verdelingsfunctie factoriseert:*

$$\begin{aligned} \mathsf{F}_{X_1,\ldots,X_n}(x_1,\ldots,x_n) &= \mathsf{P}\{X_1 \le x_1,\ldots,X_n \le x_n\} \\ &= \mathsf{P}\{X_1 \le x_1\}\cdot\mathsf{P}\{X_2 \le x_2\}\cdots\mathsf{P}\{X_n \le x_n\} = \mathsf{F}_{X_1}(x_1)\mathsf{F}_{X_2}(x_2)\cdots\mathsf{F}_{X_n}(x_n). \end{aligned}$$

In het bijzonder, zijn discrete stochasten $X_1,\ldots,X_n$ onafhankelijk dan en slechts dan wanneer hun simultane kansverdeling factoriseert:

$$\mathsf{P}\{X_1 = x_1,\ldots,X_n = x_n\} = \mathsf{P}\{X_1 = x_1\}\cdots\mathsf{P}\{X_n = x_n\},$$

voor alle waarden $x_1,\ldots,x_n$.

Continue stochasten $X_1,\ldots,X_n$ zijn onafhankelijk dan en slechts dan wanneer hun simultane kansdichtheid factoriseert:

$$\mathsf{f}_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \mathsf{f}_{X_1}(x_1)\mathsf{f}_{X_2}(x_2)\cdots\mathsf{f}_{X_n}(x_n),$$

voor alle $x_1,\ldots,x_n$.

[1] dit laatste behoort niet tot de tentamenstof voor zover het continue verdelingen betreft

Voorbeeld 6.6 Ga na dat de stochasten in voorbeelden 6.1 en 6.5 wel onafhankelijk zijn, maar in de overige voorbeelden niet.

Voorbeeld 6.7 (Ordestatistieken) Stel $X_1, \ldots, X_n$ zijn onafhankelijke trekkingen uit eenzelfde verdeling met verdelingfunctie F. Dat wil zeggen dat zij onafhankelijke en identiek verdeelde stochasten zijn. Wanneer je de n trekkingen ordent in opklimmende volgorde $X_{(1)}, \ldots, X_{(n)}$, noem je het resultaat de *ordestatistiek* (*order statistics*). De verdelingsfunctie van de grootste waarneming $X_{(n)}$ (wanneer deze bestaat) is gemakkelijk af te leiden:

$$\begin{aligned}\mathsf{F}_{X_{(n)}}(x) &= \mathsf{P}\{X_1 \le x, \ldots, X_n \le x\} = \mathsf{P}\{X_1 \le x\}\mathsf{P}\{X_2 \le x\}\cdots\mathsf{P}\{X_n \le x\} \\ &= \Big(\mathsf{P}\{X_1 \le x\}\Big)^n \\ &= \big(\mathsf{F}(x)\big)^n.\end{aligned}$$

Heeft X_1 een kansdichtheid, dan heeft $X_{(n)}$ kansdichtheid

$$\mathsf{f}_{X_{(n)}}(x) = \frac{d}{dx}\Big(\big(\mathsf{F}(x)\big)^n\Big) = n\big(\mathsf{F}(x)\big)^{n-1}\mathsf{f}(x).$$

De verdelingsfunctie van de kleinste orde statistiek $X_{(1)}$ is

$$\begin{aligned}\mathsf{F}_{X_{(1)}}(x) &= \mathsf{P}\{X_{(1)} \le x\} = 1 - \mathsf{P}\{X_{(x)} > x\} \\ &= 1 - \mathsf{P}\{\text{alle } X_i > x\} = 1 - \mathsf{P}\{X_1 > x\}\cdots\mathsf{P}\{X_n > x\} \\ &= 1 - \big(1 - \mathsf{F}(x)\big)^n.\end{aligned}$$

We hadden eerder al onafhankelijkheid gecaracteriseerd met behulp van voorwaardelijke kansen: bij onafhankelijkheid van stochasten X en Y zal informatie over de waarde van X de uitkomst van Y niet beïnvloeden, d.w.z. $\mathsf{P}\{Y \in B\} = \mathsf{P}\{Y \in B \,|\, X \in A\}$, mits $\mathsf{P}\{X \in A\} \neq 0$.

Voorwaardelijke kansen en verwachting van discrete stochasten
Stel gegeven twee discrete stochasten met een simultane kansverdeling.

De voorwaardelijke kansverdeling van Y gegeven $X = x$ krijgen we op precies dezelfde wijze als in het vorige hoofdstuk: we kiezen nu $B_x = \{X = x\}$, dan is de collectie B_x, voor alle waarden x van X met positieve kans, een partitie van de uitkomstenruimte.

De voorwaardelijke kansen worden nu gedefinieerd als $\mathsf{P}\{Y = y \,|\, X = x\} = \mathsf{P}\{Y = y, X = x\}/\mathsf{P}\{X = x\}$. De voorwaardelijke verwachting van $h(Y)$ gegeven $X = x$ (h een functie van Y) is dan $\mathsf{E}\{h(Y) \,|\, X = x\} = \sum_y y\mathsf{P}\{Y = y \,|\, X = x\}$. De verwachting $\mathsf{E}h(Y)$ kunnen we uit de voorwaardelijke verwachtingen terugrekenen door $\mathsf{E}h(Y) = \sum_x \mathsf{E}\{h(Y) \,|\, X = x\}\mathsf{P}\{X = x\}$.

Onafhankelijkheid van discrete stochasten X en Y kan als volgt worden gecaracteriseerd: X en Y zijn onafhankelijk dan en slechts dan wanneer $\mathsf{P}\{Y = y\} = \mathsf{P}\{Y = y \,|\, X = x\}$, voor alle y en alle x waarvoor $\mathsf{P}\{X = x\} \neq 0$.

Met continue stochasten hebben we hier een probleem omdat $\mathsf{P}\{X = x\} = 0$ voor *alle waarden* van X! Dit wordt als volgt opgelost: de uitspraken over kansen worden omgezet in uitspraken over aannemelijkheden.

Voorwaardelijke kansverdeling en verwachting van continue stochasten
Stel X en Y zijn twee continue stochasten met een simultane kansverdeling $\mathsf{f}_{X,Y}$. De *voorwaardelijke dichtheid* van Y gegeven $X = x$ wordt gegeven door

$$\mathsf{f}_{Y \,|\, X=x}(y) = \begin{cases} \frac{\mathsf{f}_{X,Y}(x,y)}{\mathsf{f}_X(x)}, & \mathsf{f}_X(x) > 0 \\ 0, & \text{anders.} \end{cases}$$

Nu geldt

$$\mathsf{P}\{Y \in B \,|\, X = x\} = \int_{y \in B} \mathsf{f}_{Y \,|\, X=x}(y)dy$$

en voor de voorwaardelijke verwachting van een functie h van Y:

$$\mathsf{E}\{h(Y) \,|\, X = x\} = \int_y h(y)\mathsf{f}_{Y \,|\, X=x}(y)dy.$$

De verwachting $\mathsf{E}h(Y)$ krijgen we terug uit de voorwaardelijke verwachtingen door uit te integreren over x:

$$\mathsf{E}h(Y) = \int_x \mathsf{E}\{h(Y) \mid X = x\}\mathsf{f}_X(x)dx.$$

Onafhankelijkheid van X en Y wordt in dit geval gecaracteriseerd door: X en Y zijn onafhankelijk d.e.s.d.a. $\mathsf{f}_{Y \mid X=x}(y) = \mathsf{f}_Y(y)$ wanneer $\mathsf{f}_X(x) > 0$. Dit is niets anders dan de uitspraak dat informatie over de waarde van X niets zegt over de waarde van Y.

6.2 Som van onafhankelijke stochasten [2]

Laten we eerste twee *onafhankelijke discrete* stochasten X en Y beschouwen. De vraag is: wat is de kansverdeling van $Z = X + Y$? Laten we deze afleiden.

$$\begin{aligned} \mathsf{P}\{Z = z\} &= \mathsf{P}\{X + Y = z\} = \sum_x \mathsf{P}\{X = x, Y = z - x\} \\ &\overset{\text{onafhankelijkheid}}{=} \sum_x \mathsf{P}\{X = x\}\mathsf{P}\{Y = z - x\}. \end{aligned}$$

Deze formule is een convolutieformule (voor de som van onafhankelijke discrete stochasten).

In een aantal gevallen is de verdeling van de som gemakkelijker te berekenen. Dit komt voor wanneer bijvoorbeeld X en Y tot eenzelfde parameterfamilie van kansverdelingen behoren.

Voorbeeld 6.8 Stel $X_1, \ldots, X_n$ zijn onafhankelijke $\mathsf{alt}(p)$ verdeelde stochasten. Dan is $X = X_1 + \cdots + X_n$ het aantal enen in de n experimenten, d.w.z. X heeft een $\mathsf{bin}(n, p)$ verdeling.

Voorbeeld 6.9 Stel X en Y zijn onafhankelijk en beide binomiaal verdeeld met parameters (n, p) respectievelijk (m, p). Dan heeft $X + Y$ een $\mathsf{bin}(n + m, p)$ verdeling. Dit kun je als volgt inzien.

X is te schrijven als som van n onafhankelijke $\mathsf{alt}(p)$ verdeelde stochasten $X_1, \ldots, X_n$; Y is te schrijven als som van m onafhankelijke $\mathsf{alt}(p)$ verdeelde stochasten, zeg $X_{n+1}, \ldots, X_{n+m}$. De som $X + Y$ is dus een som van $(n + m)$ onafhankelijke $\mathsf{alt}(p)$ verdeelde stochasten $X_1 + \cdots X_{n+m}$ en is derhalve $\mathsf{bin}(n + m, p)$ verdeeld.

Voorbeeld 6.10 Stel X en Y zijn onafhankelijk en beide Poisson verdeeld met parameters μ en ν respectievelijk. De interpretatie van het Poissonproces kun je gebruiken om af te leiden dat $X + Y$ een Poisson verdeling heeft met parameter $\mu + \nu$.

Je kunt dit ook met de convolutieformule berekenen:

$$\begin{aligned} \mathsf{P}\{X + Y = z\} &= \sum_x \mathsf{P}\{X = x\}\mathsf{P}\{Y = z - x\} \\ &= \sum_{x \geq 0, z-x \geq 0} \frac{\mu^x}{x!} \exp^{-\mu} \frac{\nu^{z-x}}{(z-x)!} \exp^{-\nu} \\ &= \exp^{-(\mu+\nu)} \sum_{x=0}^{z} \frac{1}{x!(z-x)!} \mu^x \nu^{z-x} \\ &= \exp^{-(\mu+\nu)} \frac{1}{z!} \sum_{x=0}^{z} \frac{z!}{x!(z-x)!} \mu^x \nu^{z-x} = \exp^{-(\mu+\nu)} \frac{1}{z!} (\mu + \nu)^z, \end{aligned}$$

waarbij we het binomium van Newton hebben gebruikt.

Stel nu dat X en Y *onafhankelijke continue* stochasten zijn met We berekenen de verdelingsfunctie van $X+Y$:

$$\begin{aligned} \mathsf{F}_{X+Y}(z) &= \mathsf{P}\{X + Y \leq z\} = \iint_{x+y \leq z} \mathsf{f}_{X,Y}(x, y)d(x, y) \\ &\overset{\text{onafhankelijkheid}}{=} \iint_{x+y \leq z} \mathsf{f}_X(x)\mathsf{f}_y(y)d(x, y) \end{aligned}$$

[2] convolutieformules behoren niet tot tentamenstof

$$
\begin{aligned}
&= && \int_{x=-\infty}^{\infty} \Big(\int_{y:x+y\le z} \mathsf{f}_X(x)\mathsf{f}_Y(y)dy \Big) dx \\
&\overset{t=x+y}{=} && \int_{x=-\infty}^{\infty} \Big(\int_{t\le z} \mathsf{f}_X(x)\mathsf{f}_Y(t-x)dt \Big) dx \\
&\overset{\text{wissel om}}{=} && \int_{t=-\infty}^{z} \Big(\int_{-\infty}^{\infty} \mathsf{f}_X(x)\mathsf{f}_Y(t-x)dx \Big) dt.
\end{aligned}
$$

Tussen haakjes moet dan de dichtheid van $X+Y$ staan in het punt t, oftewel:

$$\mathsf{f}_{X+Y}(t) = \int_{-\infty}^{\infty} \mathsf{f}_X(x)\mathsf{f}_Y(t-x)dx.$$

Dit heet de *convolutieformule* voor de som van twee onafhankelijke continue stochasten X en Y.

Voorbeeld 6.11 Stel dat X en Y onafhankelijke normaal verdeelde stochasten met parameters (μ,σ^2) resp. (ν,τ^2). Toepassing van de convolutieformule en een hoop rekenwerk levert dat $aX+bY+c$ $\mathsf{N}(a\mu+b\nu+c, a^2\sigma^2+b^2\tau^2)$ verdeeld is.

Ook hier kun je de verdeling van de som soms op eenvoudiger wijze afleiden door gebruik te maken van handige interpretaties van de stochasten die je moet sommeren.

Voorbeeld 6.12 Stel X en Y zijn onafhankelijke chi-kwadraat verdeelde stochasten met parameters r en s. Dat wil zeggen dat X te maken is als de som van het kwadraat van r onafhankelijke standaardnormale stochasten $X_1,\ldots,X_r$ en Y als de som van het kwadraat van s onafhankelijke standaardnormale stochasten $X_{r+1},\ldots,X_{r+s}$. M.a.w. $X = X_1^2+\cdots+X_r^2$ en $Y = X_{r+1}^2+\cdots+X_{r+s}^2$. Hieruit volgt dat $X+Y = X_1^2+\cdots+X_{r+s}^2$ een χ^2_{r+s} verdeling heeft.

6.3 Verwachtingen en covarianties

Als de stochasten $X_1,\ldots,X_n$ niet onafhankelijk zijn, noemt men ze *afhankelijk*. Afhankelijkheden worden vaak gemeten met behulp van *covarianties* en *correlaties*.

Laten we eerst de verwachting van een functie van een stochastische twee-dimensionale vector (X,Y) definiëren.

Definitie 6.1 Stel $g:\mathbf{R}^2\to\mathbf{R}$ is een (meetbare) functie, en X en Y zijn stochasten. Stel X en Y zijn discreet en hebben een simultane kansverdeling, dan is de verwachting $\mathsf{E}g(X,Y)$ van $g(X,Y)$

$$\mathsf{E}g(X,Y) = \sum_{x,y} g(x,y)\mathsf{P}\{X=x, Y=y\}.$$

Anderzijds, wanneer X en Y een continue verdeling hebben met simultane dichtheid $\mathsf{f}_{X,Y}$, dan is de verwachting van $g(X,Y)$

$$\mathsf{E}g(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\mathsf{f}_{X,Y}(x,y)dxdy.$$

De volgende cruciale eigenschap volgt direct uit de definitie.

Lineariteit van de verwachting
Stel X en Y zijn twee stochasten, dan is $\mathsf{E}(aX+bY+c) = a\mathsf{E}X+b\mathsf{E}Y+c$. N.B. zelfs bij afhankelijkheid van X en Y geldt deze formule!

Dit heeft zeer aangename consequenties. Ten eerste geldt de eigenschap dat $\mathsf{E}X \le \mathsf{E}Y$ als $X\le Y$. Je kunt de verwachting van een aantal standaardverdelingen gemakkelijk berekenen, zoals van de chi-kwadraat verdeling, Γ-verdeling met geheelwaardige parameter r.

Voorbeeld 6.13 Stel $X_1,\ldots,X_n$ zijn trekkingen uit dezelfde verdeling. Zelfs als deze trekkingen niet onafhankelijk zijn geldt voor de verwachting van het gemiddelde

$$\mathsf{E}\{\frac{X_1+\cdots+X_n}{n}\} = \mathsf{E}X_1.$$

Voorbeeld 6.14 Stel X heeft een $\mathsf{bin}(n,p)$ verdeling, d.w.z. $X = X_1 + \cdots + X_n$, waarbij $X_1, \ldots, X_n$ onafhankelijke $\mathsf{alt}(p)$ verdeelde stochasten zijn. Dan is $\mathsf{E}X = \sum_{i=1}^{n} \mathsf{E}X_i = n\mathsf{E}X_1 = np$.

Voorbeeld 6.15 We trekken n objecten zonder teruglegging uit een populatie van N objecten, waarvan R met een gewenste eigenschap. X is het aantal objecten in de trekking met eigenschap R. Dan is $\mathsf{E}X = nR/N$.

Hie kun je dit inzien? Definieer $X_i = 1$ als de i-trekking een object met eigenschap R oplevert, en $X_i = 0$ anders. In Hoofdstuk 3 voorbeeld 3.9 hebben we gezien dat $\mathsf{P}\{X_i = 1\} = R/N$. Dat wil zeggen dat de X_i alle een $\mathsf{alt}(R/N)$ verdeling hebben, alleen zijn ze niet onafhankelijk! Dat maakt voor de verwachting niet uit en dus krijgen we $\mathsf{E}X = \mathsf{E}(X_1 + \cdots + X_n) = \mathsf{E}X_1 + \cdots + \mathsf{E}X_n = nR/N$, want $\mathsf{E}X_i = 1 \cdot R/N + 0(1 - R/N) = R/N$.

Voorbeeld 6.16 (vervolg voorbeeld 6.4) We hebben

$$\mathsf{E}X = \int_0^1 x(x + \tfrac{1}{2})dx = \left[\tfrac{1}{3}x^3 + \tfrac{1}{4}x^2\right]_{x=0}^{x=1} = \tfrac{7}{12} = \mathsf{E}Y.$$

Dus $\mathsf{E}\{X + Y\} = 14/12$. Verder is bijvoorbeeld

$$\mathsf{E}\{XY\} = \int_{x=0}^1 \int_{y=0}^1 xy(x+y)dydx = \int_{x=0}^1 \left[\tfrac{1}{2}x^2y^2 + \tfrac{1}{3}xy^3\right]_{y=0}^{y=1} dx = \int_{x=0}^1 \tfrac{1}{2}x^2 + \tfrac{1}{3}xdx = \left[\tfrac{1}{6}(x^3 + x^2)\right]_{x=0}^{x=1} = \tfrac{1}{3}.$$

In geval van onafhankelijkheid van X en Y geldt de volgende belangrijke eigenschap.

Theorem 6.2 *Als de stochasten X en Y onafhankelijk zijn, dan geldt $\mathsf{E}\{XY\} = \mathsf{E}X \cdot \mathsf{E}Y$. In het bijzonder geldt voor meetbare functies $g, h : \mathbf{R} \to \mathbf{R}$ dat $\mathsf{E}\{g(X)h(Y)\} = \mathsf{E}g(X) \cdot \mathsf{E}h(Y)$.*

Bewijs. We bekijken alleen de gevallen dat X en Y beide discreet dan wel continu zijn. Stel X en Y zijn discreet. Dan

$$\begin{aligned}
\mathsf{E}\{XY\} &= \sum_x \sum_y xy\mathsf{P}\{X = x, Y = y\} \\
&= \sum_x \sum_y \mathsf{P}\{X = x\}\mathsf{P}\{Y = y\} \\
&= \sum_x x\mathsf{P}\{X = x\} \sum_y y\mathsf{P}\{Y = y\} = \mathsf{E}X \cdot \mathsf{E}Y.
\end{aligned}$$

Stel X en Y beide continu verdeeld. Dan

$$\begin{aligned}
\mathsf{E}\{XY\} &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} xy\mathsf{f}_{X,Y}(x,y)dydx \\
&= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} xy\mathsf{f}_X(x)\mathsf{f}_Y(y)dydx \\
&= \int_{x=-\infty}^{\infty} x\mathsf{f}_X(x)dx \int_{y=-\infty}^{\infty} y\mathsf{f}_Y(y)dy = \mathsf{E}X \cdot \mathsf{E}Y.
\end{aligned}$$

QED

Voorbeeld 6.17 (vervolg voorbeeld 6.4) We zien hieruit ook dat X en Y niet onafhankelijk kunnen zijn. Want $\mathsf{E}\{XY\} = 1/3$ en $\mathsf{E}X \cdot \mathsf{E}Y = (7/12)^2 \neq 1/3$!

We willen nu een grootheid hebben die de mate van afhankelijkheid van twee stochasten meet. Een redelijke maat hiervoor is de *covariantie*.

Definitie 6.2 De *covariantie* tussen twee stochasten X en Y is gedefinieerd als

$$\mathrm{cov}(X,Y) = \mathsf{E}\{(x - \mathsf{E}X) \cdot (Y - \mathsf{E}Y)\}.$$

De *correlatiecoëfficiënt* van X en Y is de genormeerde covariantie

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma(X) \cdot \sigma(Y)}.$$

Eigenschappen van covariantie en relatie met variantie

i) $\mathrm{cov}(X,X) = \sigma^2(X)$.
Dit volgt uit de definitie van de twee begrippen.

ii) $\mathrm{cov}(X,Y) = \mathsf{E}\{XY\} - \mathsf{E}X \cdot \mathsf{E}Y$.
Dit volgt uit: $\mathrm{cov}(X,Y) = \mathsf{E}\{(XY - X \cdot \mathsf{E}Y - \mathsf{E}X \cdot Y + \mathsf{E}X \cdot \mathsf{E}Y)\} = \mathsf{E}\{XY\} - \mathsf{E}X \cdot \mathsf{E}Y$.

iii) $\mathrm{cov}(aX+b, cY+d) = ac \cdot \mathrm{cov}(X,Y)$.
Dit volgt uit:

$$\begin{aligned}\mathrm{cov}(aX+b, cY+d) &= \mathsf{E}\{(aX+b)(cY+d)\} - \mathsf{E}\{aX+b\}\mathsf{E}\{cY+d\} \\ &= \mathsf{E}\{acXY\} + ad\mathsf{E}X + bc\mathsf{E}Y + bd - (a\mathsf{E}X + b)\cdot(c\mathsf{E}Y + d) \\ &= ac\big(\mathsf{E}\{XY\} - \mathsf{E}X \cdot \mathsf{E}Y\big) = ac \cdot \mathrm{cov}(X,Y).\end{aligned}$$

iv) $\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y) + 2 \cdot \mathrm{cov}(X,Y)$.
Dit volgt uit:

$$\begin{aligned}\sigma^2(X+Y) &= \mathsf{E}(X+Y)^2 - \big(\mathsf{E}\{X+Y\}\big)^2 \\ &= \mathsf{E}\{X^2 + 2XY + Y^2\} - \big(\mathsf{E}X + \mathsf{E}Y\big)2 \\ &= \mathsf{E}X^2 + 2\mathsf{E}\{XY\} + \mathsf{E}Y^2 - \big(\mathsf{E}X\big)^2 - 2\mathsf{E}X \cdot \mathsf{E}Y - \big(\mathsf{E}Y\big)^2 \\ &= \sigma^2(X) + \sigma^2(Y) + 2 \cdot \mathrm{cov}(X,Y).\end{aligned}$$

v) $-1 \le \rho(X,Y) \le 1$.
Volgt uit de zogenaamde ongelijkheid van Cauchy-Schwarz.

Gevolg
Als X en Y onafhankelijk dan is $\mathrm{cov}(X,Y) = 0$ en $\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y)$.

Voorbeeld 6.18 (vervolg voorbeeld 6.4) $\mathrm{cov}(X,Y) = 1/3 - (7/12)^2 = -1/144$.

Voorbeeld 6.19 (vervolg voorbeeld 6.14) Voor X een $\mathsf{bin}(n,p)$ verdeelde stochast geldt $\sigma^2(X) = n\sigma^2(X_1) = np(1-p)$.

Voorbeeld 6.20 (vervolg voorbeeld 6.15) De $\mathsf{alt}(R/N)$ verdeelde stochasten $X_1, \ldots, X_n$ zijn afhankelijk bij het trekken zonder teruglegging uit de populatie van N objecten, waaronder R met een gewenste eigenschap. We willen nu de variantie van $X = X_1 + \cdots + X_n$, het aantal objecten met eigenschap R in de trekking van N berekenen. We weten dat $\sigma^2(X_i) = (R/N)(1-(R/N))$. Daartoe berekenen we eerst

$$\mathrm{cov}(X_iX_j) = \mathsf{E}\{X_iX_j\} - \mathsf{E}X_i\mathsf{E}X_j = 1\mathsf{P}\{X_i = 1, X_j = 1\} - \big(\frac{R}{N}\big)^2 = \frac{R(R-1)}{N(N-1)} - \big(\frac{R}{N}\big)^2 = -\frac{R}{N}\frac{N-R}{N(N-1)}.$$

We krijgen nu

$$\begin{aligned}\sigma^2(X) &= \sigma^2(X_1 + \cdots + X_n) \\ &= \sum_{i=1}^{n} \sigma^2(X_i) + 2\sum_{i<j} \mathrm{cov}(X_i, X_j) \\ &= n\frac{R}{N}\big(1 - \frac{R}{N}\big) - 2\sum_{i<j} \frac{R}{N}\frac{N-R}{N(N-1)} \\ &= n\frac{R}{N}\frac{N-R}{N} - 2\cdot\frac{n(n-1)}{2}\frac{R}{N}\frac{N-R}{N(N-1)} \\ &= n\frac{R}{N}\frac{N-R}{N}\frac{N-n}{N-1}.\end{aligned}$$

Wanneer we trekken *met* teruglegging, heeft het aantal objecten met de gewenste eigenschap een $\mathsf{bin}(n, R/N)$ verdeling, en dus variantie $n(R/N)(N-R)/N$. Deze is *groter* dan bij trekken zonder teruglegging! D.w.z. de fluctuaties rond het gemiddelde zijn groter wanneer we met teruglegging trekken.

Voorbeeld 6.21 (vervolg voorbeeld 6.13) Als de X_i ook onafhankelijk zijn, dan geldt

$$\sigma^2\Big(\frac{X_1+\cdots+X_n}{n}\Big) = \frac{1}{n^2}\Big(\sigma^2(X_1)+\cdots+\sigma^2(X_n)\Big) = \frac{1}{n}\sigma^2(X_1).$$

Voor de standaarddeviatie van het gemiddelde krijgen we:

$$\sigma\Big(\frac{X_1+\cdots+X_n}{n}\Big) = \sqrt{\frac{1}{n}\sigma^2(X_1)} = \frac{1}{\sqrt{n}}\sigma(X_1).$$

Je ziet dat het gemiddelde een steeds kleinere variantie krijgt, en zich voor grote waarden van n steeds meer om het gemiddelde gaat concentreren.

Voorbeeld 6.22 Voor de som van n onafhankelijke en identiek verdeelde stochasten $X_1,\ldots,X_n$ geldt dus $\mathsf{E}(X_1+\cdots+X_n = n\mathsf{E}X_1$, $\sigma^2(X_1+\cdots+X_n) = n\sigma^2(X_1)$ en $\sigma(X_1+\cdots+X_n) = \sqrt{n}\sigma(X_1)$. Het gemiddelde groeit linear met n, maar de deviatie (die dezelfde maat heeft als de verwachting) groeit slechts als $\sqrt{n}$.

De covariantie is helaas niet helemaal een goede maat voor afhankelijkheid, want er zijn afhankelijke stochasten te vinden die covariantie 0 hebben.

Voorbeeld 6.23 Stel X is homogeen verdeeld op $[-1/2,1/2]$ en $Y = X^2$. X en Y zijn dan afhankelijk, want de waarde X legt de waarde Y vast. Maar $\mathsf{E}X = 0$, $\mathsf{E}\{XY\} = \mathsf{E}X^3 = 0$ (ga na) en dus $\mathrm{cov}(X,Y) = \mathsf{E}\{XY\} - \mathsf{E}X\cdot\mathsf{E}Y = 0$.

De X en Y noemen we *ongecorreleerd*.

Definitie 6.3 X en Y heten *ongecorreleerd* als $\mathrm{cov}(X,Y) = 0$ oftewel $\mathsf{E}\{XY\} = \mathsf{E}X\cdot\mathsf{E}Y$.

Eigenschap van ongecorreleerde stochasten
Als X en Y ongecorreleerde stochasten zijn, dan is $\sigma^2(X+Y) = \sigma^2(X)+\sigma^2(Y)$.

Wat meet covariantie dan wel, als het niet afhankelijkheid is?

Covariantie en lineair verband
De covariantie is een maat voor een *lineair verband* tussen stochasten. We zeggen dat er een *exact lineair verband* tussen X en Y is als voor zekere a en b geldt dat $Y = aX+b$. In het algemeen is er geen exact lineair verband, maar we verwachten wel vaak een relatie in de trant van "hoe groter X, des te groter Y" of juist "hoe groter X des te kleiner Y".

Voorbeeld 6.24 Stel $Y = aX+b+V$ waarbij X en V onafhankelijk zijn. Men kan V interpreteren als een verstoring van het lineaire verband, vaak in de vorm van ruis. Nu is $\mathsf{E}\{XY\} = \mathsf{E}\{X(aX+b+V)\} = a\mathsf{E}X^2 + b\mathsf{E}X + \mathsf{E}X\mathsf{E}V$. Verder, $\mathsf{E}X\cdot\mathsf{E}Y = a\big(\mathsf{E}X\big)^2 + b\mathsf{E}X + \mathsf{E}X\mathsf{E}V$, en dus is $\mathrm{cov}(X,Y) = a\sigma^2(X)$. We zien dat de covariantie positief is als $a>0$ en negatief als $a<0$.

In het algemeen noemen we het geval $\mathrm{cov}(X,Y) > 0$ een positief verband en $\mathrm{cov}(X,Y) < 0$ een negatief verband. Als $\mathrm{cov}(X,Y) = 0$ dan is er geen lineair verband maar er kan wel degelijk een relatie tussen de twee stochasten zijn! (zie voorbeeld 6.23)

Correlatiecoëfficiënt
De correlatiecoëfficiënt is een dimensieloos begrip, d.w.z. het is onafhankelijk van de eenheid waarin X en Y worden gemeten. Stel we standaardiseren X en Y, d.w.z. we beschouwen

$$\tilde{X} = \frac{X-\mathsf{E}X}{\sigma(X)}, \quad \tilde{Y} = \frac{Y-\mathsf{E}Y}{\sigma Y}.$$

Dan geldt $\rho(X,Y) = \rho(\tilde{X},\tilde{Y})$.

Meer-dimensionale normale verdeling
Er is een verdeling waar ongecorreleerdheid en onafhankelijk wel dezelfde begrippen zijn, en dat is de meer-dimensionale normale verdeling.

X en Y hebben een twee-dimensionale normale verdeling, als ze een simultane dichtheid hebben van de vorm

$$\mathsf{f}_{X,Y}(x,y) = \frac{1}{2\pi\sigma\tau\sqrt{1-\rho^2}} \exp\Big\{ -\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2(1-\rho^2)} - \frac{1}{2}\frac{(y-\nu)^2}{\tau^2(1-\rho^2)} + \rho\frac{(x-\mu)(y-\nu)}{\sigma\tau(1-\rho^2)}\Big\}.$$

De betekenis van de verschillende parameters is: $\mu = \mathsf{E}X$, $\nu = \mathsf{E}Y$, $\sigma = \sigma(X)$, $\tau = \sigma(Y)$, $\rho = \rho(X,Y)$. Als $\rho = 0$, dan factoriseert de simultane dichtheid in het product van de twee dichtheden van de $\mathsf{N}(\mu,\sigma^2)$ en de $\mathsf{N}(\nu,\tau^2)$ verdelingen, en dus zijn X en Y in dat geval onafhankelijk! Ook als $\rho \neq 0$ zijn de marginale dichtheden van X en Y die van de $\mathsf{N}(\mu,\sigma^2)$ en $\mathsf{N}(\nu,\tau^2)$ verdelingen, maar X en Y zijn dan niet onafhankelijk.

Voor het beschrijven van hoger dimensionale normale verdelingen is matrix notatie handig, hetgeen momenteel niet aan de orde is.

6.4 Simulatie van covariantie

Laat $(X_1,Y_1),\ldots,(X_n,Y_n)$ een steekproef zijn uit een (bivariate) verdeling, d.w.z. n onafhankelijke copiën van (X,Y). Laat $\mathrm{cov}(X,Y)$ de covariantie zijn tussen X en Y. De steekproefcovariantie is

$$S_{XY} = \frac{\sum_{i=1}^n \big(X_i - \bar{X}(n)\big)\big(Y_i - \bar{Y}(n)\big)}{n-1}.$$

Volgens de wet van de grote aantallen geldt weer dat S_{XY} ongeveer gelijk is aan $\mathrm{cov}(X,Y)$ voor grote waarden van n.

Evenzo, kun je de correlatie schatten met de steekproefcorrelatie

$$\hat{\rho}_{XY} = \frac{S_{XY}}{S_X S_Y}$$

waarbij de S_X en S_Y de steekproefvarianties van X en Y zijn (zie Hoofdstuk 5).

6.5 Opgaven Hoofdstuk 6

Opgave 6.1 Stel dat de aantallen fouten in elke 20 regels van een computerprogramma (van 2000 regels) onafhankelijk Poisson verdeeld zijn met parameter 1/4, wat is dan de verdeling van het aantal fouten in het hele programma?

Opgave 6.2 Gegeven de stochastische vector (X,Y) met simultane kansverdeling $\mathsf{P}\{Y=y, X=x\}$ gegeven in het volgende tableau

$y\backslash x$	1	2
-1	0,25	0,20
1	0,35	0,20

i) Bepaal de marginale verdelingen van X en Y.

ii) Zijn X en Y onafhankelijk?

iii) Bereken $\mathsf{E}X$, $\mathsf{E}Y$, $\mathsf{E}\{X+Y\}$, $\sigma^2(X)$ en $\sigma^2(Y)$.

iv) Bereken $\mathsf{E}\{XY\}$, $\mathrm{cov}(X,Y)$ en $\sigma^2(X+Y)$.

Opgave 6.3 Laat X het aantal worpen zijn totdat we voor het eerst 6 gooien met een dobbelsteen en Y het aantal worpen totdat we voor het eerst 6 gooien gerekend *vanaf* X (startend bij $X+1$).

i) Bepaal de simultane verdeling van $(X, X+Y)$.

ii) Bepaal de verdeling van $X+Y$, interpreteer het resultaat.

iii) Bepaal de voorwaardelijke verdeling van X gegeven $X+Y=z$, voor elke mogelijke waarde van z.

Opgave 6.4 We werpen n maal een onzuivere munt met kans p op Kop. Laat X het aantal keren Kop zijn in de n worpen, en Y het aantal keren Munt. Wat voor verdeling heeft (X,Y)? Zijn X en Y onafhankelijk? Bereken $\mathsf{E}(X+Y)$ en $\sigma^2(X+Y)$.

Opgave 6.5 Laat X de levensduur van een instrument voorstellen en Y de corrosiegevoeligheid. De simultane dichtheid van X en Y wordt gegeven door ($\lambda > 0$)

$$\mathsf{f}_{X,Y}(x,y) = \begin{cases} \lambda x \exp^{-x(y+\lambda)}, & x,y > 0 \\ 0, & \text{anders.} \end{cases}$$

Bepaal de marginale dichtheden. Ga na of X en Y onafhankelijk zijn. Probeer de covariantie van X en Y te bepalen.

Opgave 6.6 Laten $X_1, \ldots, X_{10}$ onafhankelijke stochasten zijn met dezelfde kansdichtheid $\mathsf{f}(x) = 2x$ voor $0 \leq x \leq 1$, en 0 elders. Bereken $\mathsf{E}\{X_1 + \cdots + X_{10}\}$ en $\sigma^2(X_1 + \cdots + X_{10})$.

Opgave 6.7 Zij (X,Y) continu verdeeld met simultane kansdichtheid

$$\mathsf{f}_{X,Y}(x,y) = \begin{cases} 1/(x+y)^3, & x,y > c \\ 0, & \text{anders.} \end{cases}$$

Bepaal de constante c zó dat $\mathsf{f}_{X,Y}$ een kansdichtheid is. Bepaal marginale kansdichtheden van X en Y. Ga na of X en Y onafhankelijk zijn.

Opgave 6.8 Stel dat X een Γ-verdeling heeft met parameters r en λ. Gegeven is dat r een geheel getal is. Bereken $\mathsf{E}X$ en $\sigma^2(X)$.

Opgave 6.9 Bereken $\mathsf{E}Y^2$ voor Y een standaard normaal verdeelde stochast. Bereken hieruit $\mathsf{E}X$ voor X een stochast met een χ^2_r-verdeling.

Opgave 6.10 Zij (X,Y) continu verdeeld met simultane kansdichtheid

$$\mathsf{f}_{X,Y}(x,y) = \begin{cases} c \cdot e^{-(x-y)}, & 0 < y < 1, x > y \\ 0, & \text{anders} \end{cases}$$

Bepaal de constante c zó dat $\mathsf{f}_{X,Y}$ een kansdichtheid is. Bepaal marginale kansdichtheden van X en Y en $\mathsf{E}\{X+Y\}$. Ga na of X en Y onafhankelijk zijn.

Opgave 6.11 Zij (X,Y) continue verdeeld met simultane kansdichtheid

$$\mathsf{f}_{X,Y}(x,y) = \begin{cases} 1/\pi, & x^2 + y^2 \leq 1 \\ 0, & \text{anders} \end{cases}$$

Bepaal de marginale verdelingen van X en Y. Ga na of X en Y onafhankelijk zijn. Zijn X en Y ongecorreleerd? Bepaal verdelingsfunctie en kansdichtheid van $Z = X^2 + Y^2$.

Opgave 6.12 Laten X en Y onafhankelijk stochasten zijn die beide homogeen verdeeld zijn op $[0,1]$. Bepaal de dichtheid van de stochast $Z = X + Y$, en zijn verwachting en variantie. Bepaal $\mathrm{cov}(Z,X)$.

Opgave 6.13 De stochastische vector (X,Y) heeft simultane verdeling

$$\mathsf{P}\{X = x, Y = y\} = \frac{e^{-\mu}\mu^{x+y}}{(x+y+1)!}, \quad \mu > 0,\ x,y = 0,1,2,\ldots$$

i) Bereken $\mathsf{P}\{X = 0\}$ en $\mathsf{P}\{Y = 0\}$.

ii) Bereken $\mathsf{P}\{X + Y = k\}$, $k = 0,1,2,\ldots$.

iii) Bereken $\mathsf{P}\{X = x \,|\, X + Y = k\}$.

iv) Zijn X en Y onafhankelijk?

Opgave 6.14 Laten a en b positieve getallen zijn, en X en Y twee stochasten. Ga na dat de correlatie tussen aX en bY gelijk is aan de correlatie tussen X en Y.

Opgave 6.15 Toon aan dat $-1 \leq \rho(X, Y) \leq 1$. Hint: gebruik eerst opgave 6.14 om te beargumenteren dat je mag aannemen dat $\sigma^2(X) = \sigma^2(Y) = 1$. Geef vervolgens een uitdrukking voor $\sigma^2(X + Y)$ en $\sigma^2(X - Y)$.

Opgave 6.16 Laten X en Y twee stochasten zijn met dezelfde varianties. Wat is $\mathrm{cov}(X - Y, X + Y)$?

Opgave 6.17 Laat $X_1, \ldots, X_n$ een steekproef uit een stochast X zijn. Laat zien dat

$$\mathsf{E}\Big\{\frac{\bar{X}(m)}{\bar{X}(n)}\Big\} = 1, \quad m < n.$$

7 Limietstellingen

In Hoofdstuk 1 en in de opgaven zijn al een aantal malen twee belangrijke limietstellingen aan de orde geweest: de wet der grote aantallen en de centrale limietstelling.

7.1 Wet der grote aantallen

Zoals eerder aan de orde is geweest, is het wiskundige kansbegrip gefundeerd op het idee dat vaak herhaalde series experimenten gemiddelde uitkomsten opleveren die ongeveer gelijk zijn. Dat zou je een experimentele wet der grote aantallen kunnen noemen. Het theoretisch analogon (sterke vorm) zegt dat het gemiddelde van de uitkomsten steeds beter de verwachte uitkomst van één experiment benadert.

Er zijn vele varianten van deze wet. We geven er twee.

Theorem 7.1 (Sterke wet van de grote aantallen) *Stel $X_1, X_2, \ldots$ zijn onderling onafhankelijk identiek verdeelde stochasten met eindige verwachting $\mathsf{E}X_i = \mu$. Dan geldt met kans 1 dat*

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \lim_{n\to\infty} \bar{X}(n) = \mu.$$

We zullen een "zwakkere" versie van deze wet bewijzen.

Theorem 7.2 (Zwakke wet der grote aantallen) *Stel dat $X_1, X_2, \ldots$ onafhankelijke stochasten zijn met dezelfde (eindige) verwachting $\mathsf{E}X_i = \mu$ en dezelfde (eindige) variantie $\sigma^2(X_i) = \sigma^2$. Dan geldt voor alle $c > 0$ dat*

$$\lim_{n\to\infty} \mathsf{P}\{|\bar{X}(n) - \mu| > c\} = 0.$$

Men zegt dat $\bar{X}(n)$ in kans convergeert naar *μ, als $n \to \infty$.*

Let wel: de zwakke wet staat toe dat de X_i's verschillende verdelingen hebben, als ze maar dezelfde μ en σ^2 hebben.

Is het theoretisch aannemelijk dat dit resultaat moet gelden, afgezien van het feit dat dit experimenteel geconstateerd is?

We weten dat $\sigma^2(\bar{X}(n)) = \sigma^2/n$, d.w.z. de variantie van $\bar{X}(n)$ wordt kleiner naarmate n toeneemt. In de limiet $n \to \infty$ wordt de variantie 0, en we hebben gezien dat een stochast met variantie 0 maar één waarde kan aannemen, namelijk zijn verwachting!

Interpretatie
We kunnen $X_1, \ldots, X_n$ zien als n metingen van een getal μ met meetfout (error) $\epsilon_i = X_i - \mu$, $i = 1, \ldots, n$. Dat wil zeggen dat

$$X_i = \mu + \epsilon_i, \quad i = 1, \ldots, n.$$

De metingen bevatten geen systematische fout, in de zin dat we gemiddeld genomen geen fout maken: $\mathsf{E}\epsilon_i = 0$, $i = 1, \ldots, n$. De variantie is een maat voor nauwkeurigheid van de meting. Als de variantie groot is, hebben we tamelijk onnauwkeurige metingen (fluctueren veel). Nemen we het gemiddelde, dan geldt $\bar{X}(n) = \mu + \bar{\epsilon}(n)$, met $\bar{\epsilon}(n) = (\epsilon_1 + \cdots + \epsilon_n)/n$. Dus $\bar{X}(n)$ meet μ met meetfout $\bar{\epsilon}(n)$. De onnauwkeurigheid is kleiner geworden dan die van de individuele metingen want $\sigma^2\big(\bar{\epsilon}(n)\big) = \sigma^2/n$. Hoe meer metingen we verrichten, des te kleiner de onnauwkeurigheid.

Chebyshev ongelijkheid
Voor het bewijs van de zwakke wet hebben we een simpel lemma nodig, dat ondanks zijn eenvoud een cruciale rol speelt in vele afleidingen.

Lemma 7.1 *Gegeven zij een (niet-negatieve) stochast $Z \geq 0$. Dan geldt voor alle $c > 0$*

$$\mathsf{P}\{Z \geq c\} \leq \frac{\mathsf{E}Z}{c}.$$

Bewijs. We hebben gezien dat $\mathsf{P}\{Z \geq c\} = \mathsf{E}\mathbf{1}_{\{Z \geq c\}}$. Verder geldt dat

$$0 \leq c \cdot \mathbf{1}_{\{Z \geq c\}} \leq Z,$$

want als $Z < c$ dan is $\mathbf{1}_{\{Z \geq c\}} = 0$ dus ook $c \cdot \mathbf{1}_{\{Z \geq c\}}$. Nu verwachting nemen

$$\mathsf{E}\{c \cdot \mathbf{1}_{\{Z \geq c\}}\} \leq \mathsf{E}Z.$$

Maar $\mathsf{E}\{c \cdot \mathbf{1}_{\{Z \geq c\}} = c \cdot \mathsf{E}\mathbf{1}_{\{Z \geq c\}} = c \cdot \mathsf{P}\{Z \geq c\}$. Delen door c geeft het gewenste resultaat. QED

Gevolg. Gegeven stochast X. Kies $Z = X^2$. Dan krijgen we

$$\mathsf{P}\{|X| \geq c\} = \mathsf{P}\{X^2 \geq c^2\} = \mathsf{P}\{Z \geq c^2\} \leq \frac{\mathsf{E}Z}{c^2} = \frac{\mathsf{E}X^2}{c^2}$$

vooropgesteld dat $\mathsf{E}X^2$ eindig is. Willen we een afschatting hebben voor $\mathsf{P}\{|X - \mu| \geq c\}$, dan kunnen we $Z = (X - \mu)^2$ kiezen. Dan is $\mathsf{E}Z = \mathsf{E}(X - \mu)^2 = \sigma^2$, zodat

$$\mathsf{P}\{|X - \mu| \geq c\} = \mathsf{P}\{(X - \mu)^2 \geq c^2\} = \mathsf{P}\{Z \geq c^2\} \leq \frac{\mathsf{E}Z}{c^2} = \frac{\sigma^2}{c^2}.$$

Tenslotte vervangen we X door $\bar{X}(n)$, μ blijft μ en σ^2 wordt σ^2/n:

$$\mathsf{P}\{|\bar{X}(n) - \mu| \geq c\} = \mathsf{P}\{(\bar{X}(n) - \mu)^2 \geq c^2\} = \mathsf{P}\{Z \geq c^2\} \leq \frac{\mathsf{E}Z}{c^2} = \frac{\sigma^2}{n \cdot c^2} \to 0,$$

als $n \to \infty$. Hiermee is de zwakke wet bewezen!

Speciaal geval: de alternatieve verdeling
Bekijk n onafhankelijke experimenten, waarbij we geïnteresseerd in het al dan niet optreden van een bepaald kenmerk of een bepaalde eigenschap. Noem $X_i = 1$ al dit kenmerk of deze eigenschap in het i-de experiment wordt gevonden. Dan is $X_i = 0$, wanneer dit in het i-de experiment *niet* het geval is.

Laat $p = \mathsf{P}\{X_i = 1\}$. Dan zijn $X_1, \ldots, X_n$ voor elke n onafhankelijk en alternatief verdeeld met succeskans p. Het aantal keren dat het te bestuderen kenmerk voorkomt in n experimenten is $X(n) = \sum_{i=1}^{n} X_i$, en dit heeft een $\mathsf{bin}(n, p)$ verdeling!

Omdat $\mathsf{E}X_i = p$, geldt volgende de wet van de grote aantallen dat

$$\frac{\text{aantal keren dat kenmerk voorkomt in } n \text{ experimenten}}{n\text{=aantal experimenten}} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}(n) \to p, \quad n \to \infty.$$

Als voorbeeld van onafhankelijke experimenten, kun je trekkingen $Y_1, \ldots, Y_n$ uit een stochast Y doen. Het kenmerk kan bijvoorbeeld zijn of de uitkomst van een trekking in een bepaalde deelverzameling A ligt: $X_i = 1$ d.e.s.d.a. $Y_i \in A$, oftewel $X_i = \mathbf{1}_{\{Y_i \in A\}}$. Dan is $p = \mathsf{E}X_i = \mathsf{P}\{X_i = 1\} = \mathsf{P}\{Y_i \in A\} = \mathsf{P}\{Y \in A\}$. Dat impliceert dat

$$\frac{\text{aantal keren in } n \text{ epxerimenten dat een uitkomst in } A \text{ voorkomt}}{n} \to \mathsf{P}\{Y \in A\}.$$

7.2 De centrale limietstelling

Stel dat we n onafhankelijke trekkingen $X_1, \ldots, X_n$ doen uit een stochast X met verwachting $\mathsf{E}X = \mu$ en variantie $\sigma^2 = \sigma^2(X)$. Volgens de wet der grote aantallen is de $\bar{X}(n)/n \approx \mu$. Maar hoe groot is nu de afwijking? Zijn daar kwantitatieve uitspraken over te doen?

Omdat $\bar{X}(n)$ een stochast is, is de afwijking $|\bar{X}(n) - \mu|$ *ook* een stochast. Men kan dus nooit *exact* zeggen, hoe groot de afwijking is, maar daar hooguit een *kansuitspraak* over doen. De centrale limietstelling geeft een benadering voor kansen van de vorm $\mathsf{P}\{|\bar{X}(n) - \mu| > c\}$, door de uitspraak dat $\bar{X}(n)$ ongeveeer normaal verdeeld is.

Voor de formulering van de stelling, moet we eerst standaardiseren, zodat we over stochasten praten met dezelfde verwachting en dezelfde variantie.

Standaardiseren
Het standaardiseren van een stochast X betekent dat we verwachting $\mathsf{E}X = \mu$ aftrekken en delen door de standaarddeviatie $\sigma = \sigma(X)$: $(X - \mu)/\sigma$. De gestandaardiseerde heeft verwachting 0 en variantie en standaarddeviatie 1.

Voor de gemiddelden $\bar{X}(n)$, is de gestandaardiseerde

$$\frac{\bar{X}(n) - \mu}{\sigma/\sqrt{n}}.$$

Theorem 7.3 (Centrale limietstelling) *Laten $X_1, \ldots, X_n$ onafhankelijke stochasten zijn die alle dezelfde verdeling hebben, met verwachting $\mu = \mathsf{E}X_1$ en $\sigma^2 = \sigma^2(X_1)$. Dan geldt voor elke $x \in \mathbf{R}$*

$$\lim_{n\to\infty} \mathsf{P}\Big\{ \frac{\bar{X}(n) - \mu}{\sigma/\sqrt{n}} \le x \Big\} = \Phi(x).$$

De stelling impliceert dat $\bar{X}(n)$ ongeveer $\mathsf{N}(\mu, \sigma^2/n)$ verdeeld is, oftewel $\bar{X}(n)$ is *asymptotisch normaal verdeeld*. Op de lange duur *vergeet* het steekproefgemiddelde uit welke verdeling er getrokken was.

Feitelijk geeft de centrale limietstelling ook een uitspraak over de verdeling van de som $X_+(n) = \sum_{i=1}^n X_i$. Immers: $(\bar{X}(n) - \mu)/(\sigma/\sqrt{n})$ is de gestandaardiseerde van zowel $\bar{X}(n)$ als $\sum_{i=1}^n X_i$:

$$\frac{\bar{X}(n) - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}\big(\bar{X}(n) - \mu\big)}{\sigma} = \frac{\sqrt{n}}{\sqrt{n}} \frac{\sqrt{n}\big(\bar{X}(n) - \mu\big)}{\sigma} = \frac{X_+(n) - n\mu}{\sqrt{n}\sigma}.$$

Deze laatste uitdrukking is de gestandaardiseerde van $X_+(n)$ omdat $\mathsf{E}X_+(n) = n\mu$ en $\sigma^2\big(X_+(n)\big) = n\sigma^2(X_1) = n\sigma^2$, d.w.z. $\sigma\big(X_+(n)\big) = \sqrt{n\sigma^2(X_1)} = \sqrt{n}\sigma$. In woorden zegt de centrale limietstelling dus ook, dat $X_+(n)$ ongeveer $\mathsf{N}(n\mu, \sqrt{n}\sigma)$ is verdeeld.

Willen we nu $\mathsf{P}\{(\bar{X}(n) - \mu)/(\sigma/\sqrt{n}) \in B\}$ schatten voor grote waarden van n, dan is

$$\lim_{n\to\infty} \mathsf{P}\big\{ \frac{\bar{X}(n) - \mu}{\sigma/\sqrt{n}} \in B \big\} = \mathsf{P}\{Z \in B\},$$

met Z standaard normaal verdeeld. Wanneer $B = [x_1, x_2]$ een interval is, krijgen we

$$\lim_{n\to\infty} \mathsf{P}\big\{ x_1 \le \frac{\bar{X}(n) - \mu}{\sigma/\sqrt{n}} \le x_2 \big\} = \Phi(x_2) - \Phi(x_1).$$

Willen we nu een uitspraak over de kansverdeling van $\bar{X}(n)$ of $X_+(n)$ voor grote waarden van n hebben, dan krijgen we bijvvoorbeeld

$$\mathsf{P}\{\bar{X}(n) \le x\} = \mathsf{P}\{\frac{\bar{X}(n) - \mu}{\sigma/\sqrt{n}} \le \frac{x - \mu}{\sigma/\sqrt{n}} \approx \Phi\big(\frac{x - \mu}{\sigma/\sqrt{n}}\big),$$

en

$$\mathsf{P}\{X_+(n) \le x\} = \mathsf{P}\big\{\frac{X_+(n) - n\mu}{\sqrt{n}\sigma} \le \frac{x - n\mu}{\sqrt{n}\sigma}\big\} \approx \Phi\big(\frac{x - n\mu}{\sqrt{n}\sigma}\big).$$

Speciaal geval: sommen van onafhankelijke $\mathsf{alt}(p)$ verdeelde stochasten oftewel binomiale verdeling.
In de vorige paragraaf hadden we het als speciaal geval al over onafhankelijke stochasten $X_1, \ldots, X_n$ met een $\mathsf{alt}(p)$ verdeling gehad. Daar was besproken dat $\bar{X}(n) = \sum_{i=1}^n X_i/n \approx p$ oftewel $X_+(n) = \sum_{i=1}^n X_i \approx np$. Willen we iets weten over het verschil $|X_+(n) - np|$, dan kunnen we de centrale limietstelling aanroepen. Omdat $\sigma^2\big(X_+(n)\big) = n\sigma^2(X_1) = np(1-p)$, geldt dat $X_+(n)$ ongeveer $\mathsf{N}(np, np(1-p))$ verdeeld is, oftewel

$$\frac{X_+(n) - np}{\sqrt{np(1-p)}}$$

is ongeveer standaardnormaal verdeeld. De exacte verdeling van $X_+(n)$ weten we echter ook: dit is de binomiale verdeling met parameters n en p, d.w.z.

$$\mathsf{P}\{X_+ = x\} = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, \ldots, n.$$

Hieruit volgt een relatie tussen binomiale kansen en standaardnormale:

$$\sum_{k\le x}\binom{n}{k}p^k(1-p)^{n-k} = \mathsf{P}\{X_+(n)\le x\} \approx \Phi\big(\frac{x-np}{\sqrt{np(1-p)}}\big).$$

Analytisch is deze relatie overigens niet simpel te bewijzen!

Laten we nu een getallenvoorbeeldje bekijken: $n = 20$, $p = 0,40$ en $x = 5$. Voor $n = 20$ heeft de verdeling van $X_+(n)$ al redelijk de klokvorm. Echter

$$\mathsf{P}\{X_+(20)\} = \sum_{k\le 5}\binom{20}{k}(0,40)^k(0,60)^{20-k} = 0,1256$$

en

$$\Phi\big(\frac{x-np}{\sqrt{np(1-p)}}\big) = \Phi\big(\frac{5-20\cdot 0,40}{\sqrt{20\cdot 0,40\cdot 0,60}}\big) = \Phi(-1,37) = 1-\Phi(1,37) = 1-0,9147 = 0,0853.$$

Dit lijkt niet zo erg! Dat heeft te maken met het feit dat we een discrete stochast met mogelijke waarden $0, 1, 2, \ldots, 20$ benaderen met een continue verdeling. Het is beter om in zo'n geval continuïteitscorrectie toe te passen.

Continuïteitscorrectie
Als een discrete stochast Y te benaderen is met een normaal verdeelde stochast, kan het verstandig zijn continuïteitscorrectie toe te passen. Gebruik dat

$$\mathsf{P}\{Y = y\} = \mathsf{P}\{Y \le y + \tfrac{1}{2}\} - \mathsf{P}\{Y \le y - \tfrac{1}{2}\}.$$

Dan

$$\mathsf{P}\{Y = y\} \approx \Phi\big(\frac{y+1/2-\mathsf{E}Y}{\sigma(Y)}\big) - \Phi\big(\frac{y-1/2-\mathsf{E}Y}{\sigma(Y)}\big)$$

en dus

$$\mathsf{P}\{Y \le y\} \approx \Phi\big(\frac{y+1/2-\mathsf{E}Y}{\sigma(Y)}\big).$$

In bovenstaand voorbeeld krijgen we na continuïteitscorrectie:

$$\mathsf{P}\{X_+(20)\le 5\} \approx \Phi\big(\frac{5+\frac{1}{2}-20\cdot 0,40}{\sqrt{20\cdot 0,40\cdot 0,60}}\big) = \Phi(-1,14) = 1-\Phi(1,14) = 1-0,8729 = 0,1271.$$

Dit lijkt al beter! De volgende vuistregel blijkt in de praktijk redelijk te werken.

Vuistregel De binomiale $\mathsf{bin}(n,p)$ verdeling kan goed met een normale verdeling (met juiste verwachting en variantie) benaderd worden voor waarden van n en p met $np(1-p) > 5$ mits continuïteitscorrectie wordt toegepast!.

Waarom speelt dit probleem van nodige continuïteitscorrectie geen rol wanneer we naar de steekproefgemiddelden kijken? Deze kunnen tenslotte ook discreet verdeeld zijn!

Speciaal geval: Poissonverdeling
Stel $X_1, \ldots, X_n$ zijn onafhankelijk Poisson verdeelde stochasten met parameter μ. Dan is $X_+(n)$ Poisson verdeeld met parameter $n\mu$. Dat wil zeggen dat $\mathsf{E}X_+(n) = \sigma^2\big(X_+(n)\big) = n\mu$. Volgens de centrale limietstelling is $X_+(n)$ ongeveer $\mathsf{N}(n\mu, n\mu)$ verdeeld: de Poisson verdeling kan dus benaderd worden met een normale verdeling. Hierbij leidt continuïteitscorrectie in het algemeen weer tot een betere benadering!

Speciaal geval: Normale verdeling
Als $X_1, \ldots, X_n$ een $\mathsf{N}(\mu, \sigma^2)$ verdeling hebben, dan heeft de som $X_+(n)$ een $\mathsf{N}(n\mu, n\sigma^2)$ verdeling en het steekproefgemiddelde $\bar{X}(n)$ een $\mathsf{N}(\mu, \sigma^2/n)$ verdeling.

7.3 Betrouwbaarheidsinterval

Stel we hebben een steekproef $X_1, \ldots, X_n$ uit een stochast X met eindige verwachting $\mu = \mathsf{E}X$ en eindige variantie $\sigma^2(X)$. Volgens de wet der grote aantallen is $\bar{X}(n) \approx \mu$.

Laten we een kans α vastleggen. Daarbij is een interval $-c \leq \bar{X}(n) - \mu \leq c$ te vinden, zó dat

$$\mathsf{P}\{-c \leq \bar{X}(n) - \mu \leq c\} = \mathsf{P}\{|\bar{X}(n) - \mu| \leq c\} = 1 - \alpha.$$

Dit betekent dat $\alpha\%$ van de series van n waarnemingen is een afwijking $|\bar{X}(n) - \mu| > c$ oplevert.

De verdeling van $\bar{X}(n)$ is vaak niet bekend, en dus kunnen we c niet exact berekenen. Voor voldoend grote n is de centrale limietstelling van toepassing en kunnen we hiermee een benadering voor de gevraagde c bepalen:

$$\begin{aligned} 1 - \alpha = \mathsf{P}\{|\bar{X}(n) - \mu| \leq c\} &= \mathsf{P}\{\bar{X}(n) - \mu \leq c\} - \mathsf{P}\{\bar{X}(n) - \mu \leq -c\} \\ &= \mathsf{P}\Big\{\frac{\bar{X}(n) - \mu}{\sigma/\sqrt{n}} \leq \frac{c}{\sigma/\sqrt{n}}\Big\} - \mathsf{P}\Big\{\frac{\bar{X}(n) - \mu}{\sigma/\sqrt{n}} \leq \frac{-c}{\sigma/\sqrt{n}}\Big\} \\ &\approx \Phi\big(\frac{c}{\sigma/\sqrt{n}}\big) - \Phi\big(\frac{-c}{\sigma/\sqrt{n}}\big) \\ &= 2\Phi\big(\frac{c}{\sigma/\sqrt{n}}\big) - 1. \end{aligned}$$

Hoe moet je nu c kiezen? Als moet gelden

$$\Phi\big(\frac{c}{\sigma/\sqrt{n}}\big) - \Phi\big(\frac{-c}{\sigma/\sqrt{n}}\big) = 1 - \alpha$$

dan geldt vanwege de symmetrie dat

$$\Phi\big(\frac{-c}{\sigma/\sqrt{n}}\big) = \frac{\alpha}{2}$$

en dus is

$$\Phi\big(\frac{c}{\sigma/\sqrt{n}}\big) = 1 - \alpha + \frac{\alpha}{2} = 1 - \frac{\alpha}{2}.$$

Met $\xi_\beta > 0$ duiden we het getal aan, waarvoor

$$\Phi(\xi_\beta) = \beta.$$

Dan is $c/(\sigma/\sqrt{n}) = \xi_{1-\alpha/2}$ en dus is

$$c = \frac{\xi_{1-\alpha/2}\sigma}{\sqrt{n}}.$$

Dit kun je gebruiken in volgende situatie: stel we hebben de kans α vastgelegd en derhalve het interval $|\bar{X}(n) - \mu| \leq \xi_{1-\alpha/2}\sigma/\sqrt{n}$. Dat wil zeggen: in $\alpha \times 100\%$ van series van n waarnemingen constateer je een verschil tussen $\bar{X}(n)$ en μ van meer dan $c = \xi_{1-\alpha/2}\sigma/\sqrt{n}$.

Stel nu dat de verwachtingswaarde μ *onbekend* is. Volgens de wet der grote aantallen is $\bar{X}(n)$ een goede schatter voor μ.

Hoe goed? In $\alpha \times 100\%$ van series van n waarnemingen verschilt μ meer dan $\xi_{1-\alpha/2}\sigma/\sqrt{n}$ van $\bar{X}(n)$, oftewel in $(1-\alpha) \times 100\%$ van series van n waarnemingen zit μ in een interval

$$[\bar{X}(n) - \frac{\xi_{1-\alpha/2}\sigma}{\sqrt{n}}, \bar{X}(n) + \frac{\xi_{1-\alpha/2}\sigma}{\sqrt{n}}].$$

Dit interval noem je *het asymptotisch* $(1-\alpha) \times 100\%$ *betrouwbaarheidsinterval voor* μ.
Let wel: dit interval hangt van de waarneming $\bar{X}(n)$ af, en varieert dus per waarneming!

Kies bijvoorbeeld $\alpha = 0,05$. Dan is $\xi_{1-\alpha/2} = \xi_{0,975} \approx 1,96$ en dus is

$$[\bar{X}(n) - \frac{1,96\sigma}{\sqrt{n}}, \bar{X}(n) + \frac{1,96\sigma}{\sqrt{n}}]$$

het asymptotisch 95%-betrouwbaarheidsinterval voor μ.

Onbekende variantie
In het betrouwbaarheidsinterval komt σ voor en dus nemen we aan dat we de variantie en dus standaarddeviatie *wel* kunnen bepalen (maar de verwachting niet). Dat moge niet altijd een realistische aanname zijn. Wat te doen als de standaarddeviatie evenmin bekend is? Dan vul je voor σ een *schatter* van de standaarddeviatie in:

$$S(n) = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\big(X_k - \bar{X}(n)\big)^2},$$

want je weet dat $S(n) \to \sigma$ voor $n \to \infty$ (met kans 1), onder redelijke aannamen.

Het asymptotische $(1-\alpha)\times 100\%$ betrouwbaarheidsinterval voor μ wordt dan

$$[\bar{X}(n) - \frac{\xi_{1-\alpha/2}S(n)}{\sqrt{n}}, \bar{X}(n) + \frac{\xi_{1-\alpha/2}S(n)}{\sqrt{n}}]$$

Marge
Het getal $\xi_{1-\alpha/2}\sigma/\sqrt{n}$ of de schatter $\xi_{1-\alpha/2}S/\sqrt{n}$ wordt ook wel de *marge* genoemd. Deze marge is gebaseerd op een benadering (m.b.v. de centrale limietstelling). Als de steekproefgrootte n klein is, kans men aan de conservatieve kant gaan zitten door de marge groter te kiezen. Met name kan men dan de normale verdeling vervangen door de zogenaamde Student verdeling met $(n-1)$ vrijehidsgraden. Daar gaan we hier niet verder op in.

Vuistregel
Als vuistregel kan men (in woorden) hanteren: schatter en geschatte waarde verschillen niet meer dan $2\times$ de (geschatte) standaarddeviatie van de schatter. Hierin is 2 een afronding van de *beroemde* 1,96, die volgt uit de eis van 95% betrouwbaarheid in combinatie met de normale verdeling. De standaarddeviatie van de schatter $\bar{X}(n)$ is tenslotte $\sigma/\sqrt{n}$.

Voorbeeld 7.1 (Binomiale verdeling) Stel dat X een $\mathsf{bin}(n,p)$ verdeling heeft. Dan is

$$\frac{X-p}{\sqrt{np(1-p)}}$$

voor grote waarden van n bij benadering standaard normaal verdeeld, wegens de centrale limietstelling. Dit betekent dat de kans op de gebeurtenis

$$\Big\{\frac{|X-np|}{\sqrt{np(1-p)}} \le \xi_{1-\alpha/2}\Big\}$$

ongeveer $1-\alpha$ is. Maar deze verzameling is in feite het betrouwbaarheidsinterval voor p: je krijgt een kwadratische vergelijking in p die je kunt oplossen.

Je kunt het ook iets eenvoudiger doen: de wet der grote aantallen zegt dat $X/n \approx p$ voor grote waarden van n. Men kan dan bewijzen dat

$$\frac{X-np}{\sqrt{n(X/n)\big(1-(X/n)\big)}}$$

ook bij benadering standaardnormaal verdeeld is voor grote waarden van n. Ten gevolge is

$$\mathsf{P}\Big\{\frac{|X-np|}{\sqrt{n(X/n)\big(1-(X/n)\big)}} \le \xi_{1-\alpha/2}\Big\} \approx 1-\alpha.$$

Omschrijven levert dat in $(1-\alpha)\times 100\%$ van de waarnemingen de waarde p in het interval

$$[\frac{X}{n} - \frac{1}{\sqrt{n}}\sqrt{(X/n)\big(1-(X/n)\big)}, \frac{X}{n} + \frac{1}{\sqrt{n}}\sqrt{(X/n)\big(1-(X/n)\big)}]$$

ligt, oftewel dit interval is het $(1-\alpha)\times 100\%$ betrouwbaarheidsinterval voor p.

7.4 Opgaven Hoofdstuk 7

Opgave 7.1 Zijn $X_1, \ldots, X_{10}$ onafhankelijke stochasten met dezelfde kansdichtheid

$$\mathsf{f}(x) = \begin{cases} 2x, & 0 \leq x \leq 1 \\ 0, & \text{anders} \end{cases}$$

i) Bereken $\mathsf{E}\{X_1 + \cdots + X_{10}\}$ en $\sigma^2\big(X_1 + \cdots + X_{10}\big)$.

ii) Benader de verdeling van $X = X_1 + \cdots + X_{10}$ met een normale verdeling. Bereken hiermee een benadering voor de kans $\mathsf{P}\{X \leq 5\}$.

Opgave 7.2 Laat X een $\mathsf{bin}(n = 35, p = 1/5)$ verdeling hebben. Bereken de waarde $np(1-p)$. Zal een normale benadering voor de kansverdeling van X met continuïteitscorrectie redelijke schattingen geven? Geef een benadering voor de kansen $\mathsf{P}\{X = 3\}$ en $\mathsf{P}\{X \geq 7\}$.

Opgave 7.3 Stel $X_1, \ldots, X_{12}$ zij onafhankelijke Poisson verdeelde stochasten met parameter 2. Wat is de verdeling van de som $X = X_1 + \cdots + X_{12}$. Bereken (op je rekenmachine) $\mathsf{P}\{X = 10\}$. Vergelijk deze met een normale benadering voor deze kans (met en zonder continuïteitscorrectie). Benader de kans $\mathsf{P}\{x \geq 10\}$.

Opgave 7.4 Stel $X(n)$ is een Poisson verdeelde stochast met parameter n. Bereken $\lim_{n\to\infty} \mathsf{P}\{n - \sqrt{n} \leq X(n) < n + 2\sqrt{n}\}$. Bereken $\lim_{n\to\infty} \mathsf{P}\{X(n) \leq 2003\}$ en $\lim_{n\to\infty} \mathsf{P}\{n - 1905 < X(n) < n + 2003\}$.

Opgave 7.5 Stel $(X_1, Y_1), \ldots, (X_{100}, Y_{100})$ zijn onafhankelijk en identiek verdeeld met simultane verdeling

$X_1 \backslash Y_1$	1	2	4
1	0	0	1/4
2	0	1/2	0
4	1/4	0	0

i) Bepaal de marginale verdelingen van X_1 en Y_1.

ii) Bereken $\mathsf{E}\{X_1 - Y_1\}$ en $\sigma^2\big(X_1 - Y_1\big)$.

iii) Benader de kans $\mathsf{P}\{\sum_{k=1}^{100} X_k \leq 10\sqrt{2} + \sum_{k=1}^{100} Y_k\}$.

Opgave 7.6 Stel dat X een binomiale verdeling heeft met parameters $n = 100$ en $p = 1/3$. Kan men de kansverdeling van X redelijk met een normale verdeling benaderen? Zo ja, bepaal de waarde c z'o dat het verschil tussen X en $\mathsf{E}X$ hoogstens c is met kans

i) ongeveer 95%;

ii) ongeveer 99%.

Opgave 7.7 Beschouw $n = 4$ onafhankelijke waarnemingen $X_1, \ldots, X_4$ uit een stochast X. De gevonden waarden zijn

$$x_1 = 0,26, \quad x_2 = 5,12, \quad x_3 = -0,16, \quad x_4 = 3,91.$$

Bepaal een (asymptotisch) 95% betrouwbaarheidsinterval voor $\mathsf{E}X$. **N.B.** We nemen hier weinig waarnemingen omdat het rekenwerk dan doenlijk is zonder computer. In het algemeen geeft de normale verdeling natuurlijk geen goede benadering van het betrouwbaarheidsinterval, wanneer dit slechts gebaseerd is op 4 waarnemingen.

Opgave 7.8 Neem een steekproef $X_1, \ldots, X_100$ uit de uniforme verdeling op [0,1]. Bepaal een 95% (asymptotisch) betrouwbaarheidsinterval voor de verwachting 1/2.

8 Tabel standaardnormale verdeling

We tabelleren de tabel voor de standaardnormale verdeling; de laatste kolom bevat de dichtheid in de punten van de eerste kolom.

$u\backslash\Phi(u)$	+.00	+.01	+.02	+.03	+.04	+.05	+.06	+.07	+.08	+.09		$\phi(u)$
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359		.3989
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753		.3970
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141		.3910
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517		.3814
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879		.3683
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224		.3521
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549		.3332
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852		.3123
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133		.2897
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389		.2661
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621		.2420
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830		.2179
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015		.1942
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177		.1714
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319		.1497
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441		.1295
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545		.1109
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633		.0940
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706		.0790
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767		.0656
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817		.0540
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857		.0440
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890		.0355
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916		.0283
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936		.0224
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952		.0175
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964		.0136
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974		.0104
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981		.0079
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986		.0060
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990		.0044
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993		.0033
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995		.0024
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997		.0017
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998		.0012
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998		.0009
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999		.0006