Stochastiek voor Informatici

Sara van de Geer

voorjaar 2000

Inhoud hoofdstuk 1 t/m 3

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

Voorbeeld. 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* (*kansvariable*). 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

$$P(X = x) = \frac{1}{m},\ x = 1, \ldots, m.$$

Hier staat P voor *Probability*.

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

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

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 kan je niet zien of het de uitkomst is van een aselecte trekking. Als het experiment een aantal keren herhaald wordt, dan zal bij *onderling onafhankelijke* aselecte trekkingen, 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 uikomst van het ene experiment niets zegt over de uitkomst van een ander experiment. Dan geldt:

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

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. Dit resultaat noemt men de *wet van de grote aantallen*. Het volgt (wiskundig) uit de z.g. kansaxioma's. Volgens de frequentisten is het per definitie zo, d.w.z. zij definiëren een kans als de limiet van herhaalde experimenten.

We gebruiken nu een software pakket om wat "feeling" voor toevalsgetallen aan te kweken. De volgende simulaties zijn gedaan met Splus. U kunt ook Maple, Matlab, SAS, of uw 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 zo 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.

```
$ Splus
> help.start(gui="motif")
> # dit roept het help window op, met "motif" als graphical user interface
> n<-100
> # 100 experimenten
> x<-ceiling(runif(n)*6)
> # dit levert n o.o.  aselecte trekkingen uit 1...6
> x
[1] 4 5 2 1 2 5 4 1 2 3 2 5 2 2 3 5 5 6 6 4 5 4 3 4 6 5 3 3 2 1 1 4 3 2 1 1 4
```

```
[38] 5 4 6 5 5 4 1 4 6 6 3 5 2 6 2 4 3 4 2 3 1 1 6 4 1 2 5 6 6 2 1 1 2 2 3 5 5
[75] 5 1 6 6 2 4 2 3 2 6 5 6 5 3 3 1 6 2 3 6 3 5 2 2 4 2
> motif()
> hist(x,main="n=100")
> n<-1000
> x<-ceiling(runif(n)*6)
> hist(x, breaks=0:6,main="n=1000")
>q()
$
```

1.2. Realisaties. Stel we nemen 100 aselecte trekkingen uit de getallen $\{1, \ldots, 6\}$. We vinden dan 100 *getallen* $\{x_1, \ldots, x_n\}$. Dit noemt met wel de *realisaties* van de stochastische grootheden $X_1, \ldots, X_{100}$.

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.

Met

$$P(X \in A)$$

wordt de kans dat X in de verzameling A valt aangegeven.

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

$$P(X \in \{2, 4, 6\})$$

de kans op een even aantal ogen ($= 1/2$).

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

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

De transformatie ziet er dus als volgt uit:

$$1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10$$

$$\downarrow g$$

$$2\ 3\ 4\ 4\ 6\ 6\ 8\ 8\ 9\ 10$$

(Hierbij beschouwen we 1 als priemgetal.) 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

$$P(Y = 2) = P(Y = 3) = \frac{1}{10},\ P(Y = 4) = P(Y = 6) = P(Y = 8) = \frac{1}{5},\ P(Y = 9) = P(Y = 10) = \frac{1}{10}.$$

De stochastische grootheid Y is niet uniform verdeeld. De mogelijke waarden voor Y zijn $\{2, 3, 4, 6, 8, 9, 10\}$, maar deze waarden hebben niet alle dezelfde kans.

1.5. Discrete stochastische grootheden. We geven stochastische grootheden aan met hoofdletters (X, Y, etc.). De verdeling van een discrete stochastische grootheid, zeg X, kunnen we beschrijven door de mogelijke waarden, en de kans op zo'n waarde, op te sommen. Als $\{w_1, w_2, \ldots\}$ de mogelijke waarden van X zijn, dan geldt altijd dat

$$\sum_j P(X = w_j) = 1.$$

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

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

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

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

Merk op dat F een stijgende trapfunctie is, met sprongen in de punten w_j. Laten we veronderstellen dat de waarden in oplopende volgorde staan: $w_1 < w_2 < \ldots$. Dan

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

(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.)

Voorbeeld. Stel

$$P(X = 2) = P(X = 3) = \frac{1}{10},\ P(X = 4) = P(X = 6) = P(X = 8) = \frac{1}{5},\ P(X = 9) = P(X = 10) = \frac{1}{10}.$$

Dan

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

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

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

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:

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

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. *kopietjes* 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:

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

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 simulatie.

```
>par(mfrow=c(2,2))
```

```
>ns<-c(10,100,1000,10000)
>for (n in ns) {
+x<-ceiling(runif(n)*10)
+for (i in 1 :  n) {
+if (x[i] < 4 ) { x[i]<-x[i]+1 }
else
+if (x[i] == 5) { x[i]<-x[i]+1 }
else
+if (x[i] == 7) { x[i]<-x[i]+1 }}
+titel<-paste("n = ", format (n))
+hist (x,nclass=7,breaks=c(1,2,3,4,6,8,9,10), main=titel)}
```

1.9. De verdeling van de som van twee o.o. stochastische grootheden. Stel X en Y zijn twee discrete o.o. stochastische grootheden (s.g.[n]). Noem

$$Z = X + Y.$$

Dan geldt voor alle z,

$$P(Z = z) = \sum_x P(X = x)P(Y = z - x),$$

waarbij gesommeerd wordt over de mogelijke waarden van x (de rol van X en Y mogen verwisseld worden). De verdeling van Z wordt de *convolutie* van de verdeling van X en Y genoemd.

Voorbeeld. Stel X en Y zijn twee o.o. aselecte trekkingen uit $\{1, \ldots, 10\}$. Dan zijn $\{2, \ldots, 20\}$ de mogelijke waarden van $Z = X + Y$, met kansen

$$P(Z = 2) = P(X = 1, Y = 1) = \frac{1}{100},$$

$$P(Z = 3) = P(X = 1, Y = 2) + P(X = 2, Y = 1) = \frac{2}{100},$$

$$P(Z = 4) = P(X = 1, Y = 3) + P(X = 2, Y = 2) + P(X = 3, Y = 1) = \frac{3}{100},$$

enz.:

z	$P(Z = z)$
2	0.01
3	0.02
4	0.03
5	0.04
6	0.05
7	0.06
8	0.07
9	0.08
10	0.09
11	0.10
12	0.09
13	0.08
14	0.07
15	0.06
16	0.05
17	0.04
18	0.03
19	0.02
20	0.01

1.10. Gemiddelde. Het gemiddelde van een rij getallen $x_1, \ldots, x_n$ is

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

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

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

Merk op dat, in het geval van o.o. X_i, de verdeling van $\sum_{i=1}^{n} X_i$ een n-voudige convolutie is. Het exact berekenen van dergelijke convoluties kan best lastig zijn! Door simulaties kan men echter ook een idee krijgen van de verdeling. We nemen in het onderstaande 10 000 simulaties van het gemiddelde van respectievelijk 2,3 10 en 40 aselecte trekkingen uit $\{1, \ldots, 10\}$.

```
>m<-10000
># het aantal simulaties is 10000
>ns<-c(2,3,10,40)
># we bekijken de som van 2,3 10 en 40 stochastische grootheden
>for (n in ns) {
+x<-ceiling(runif(m*n)*10)
+dim(x)<-c(m,n)
+# we hebben nu m steekproeven ter grootte n uit de uniforme
verdeling op 1 t/m 10, samengebracht in een mxn matrix
u<-rep(1,n)
+# dit levert een n-vector van 1'en
+som<-x%*%u
+gemiddelde<-som/n
+titel<-paste("n = ", format(n))
+hist(gemiddelde,nclass=40,main=titel)}
```

1.11. De centrale limiet stelling. De *centrale limietstelling* zegt dat het gemiddelde van een steekproef van grootte n ongeveer *normaal* verdeeld is, als n groot is. Nu hebben we nog niet gedefinieerd wat we bedoelen met de normale verdeling (zie 3.8(2)). 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 voorbeeldje. In plaats van een steekproef uit X, met X uniform verdeeld op $\{1, \ldots, 10\}$, nemen we een steekproef uit $Y = X^2 - X$.

```
>m<-10000
>ns<-c(2,3,10,40)
>for (n in ns) {
+x<-ceiling(runif(m*n)*10)
+y<-x**2-x
+dim(y)<-c(m,n)
+u<-rep(1,n)
+som<-y%*%u
+gemiddelde<-som/n
+titel<-paste("n = ", format(n))
+hist(gemiddelde,nclass=40,main=titel)}
```

Voorbeeld. 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

$$P(X = x) = \frac{1}{1000000}, \ x \in \{1, \ldots, 1000000\}.$$

De verdelingsfunctie is

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

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

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

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

De verdelingsfunctie is nu

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

1.12. 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 $t - s$.

1.13. 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 Splus hebben we discrete uniforme verdelingen geconstrueerd door uit te gaan van de continue uniforme verdeling (de laatste zit standaard in Splus):

```
>x <- runif(n)
># dit levert n o.o.  trekkingen uit de uniforme verdeling op [0,1]
>x<-ceiling(x*m)
># dit levert n o.o.  trekkingen uit de getallen 1 ..  m
```

1.14. 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

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

met dichtheid

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

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

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

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

$$F(x) = P(X \leq x) = P(g(U) \leq x)$$

$$= P(U^2 \leq x) = P(U \leq \sqrt{x}) = \sqrt{x},\ x \in [0,1].$$

De dichtheid van X is

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

(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.

```
> m<-10000
> u<-runif(m)
> u<-sort(u)
> hist(u,nclass=30,probability=T,main="f(u)=1")
> x<-u**2
> hist(x,nclass=30,probability=T,main="f(x)=1/(2sqrt(x))")
> Fm<-1:m/m
> plot(u,Fm,main="F(u)=u")
> plot(x,Fm,main="F(x)=sqrt(x)")
```

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

$$F_n(x) = \frac{\{\#X_i \leq x,\ i \leq n\}}{n},\ x \in \mathbf{R}$$

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

$$\lim_{n\to\infty} F_n(x) = F(x).$$

2. Axioma's, voorwaardelijke kans en combinatoriek

Wat is een kans? In het dagelijks taalgebruik komt men het begrip *kans* regelmatig tegen.

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

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

Het begrip kans komt gedeeltelijk overeen met *mogelijkheid*, en wordt soms geoperationaliseerd door *fractie*, *frequentie*, of *percentage*.

Voorbeeld.. Van de mensen in Nederland tussen de 18 en 65 jaar heeft x % een baan.

Een interpretatie van het laatste voorbeeld is dat in Nederland de kans op een baan x % is. In de wiskunde wordt echter een veel abstracter begrip kans gehanteerd. 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 volgen uit de axioma's ongeveer voldoen aan een intuïtief idee van kans.

2.1. Stochastiek. We kunnen een onderscheid maken tussen deterministische modellen en stochastische modellen. Deterministisch zijn b.v. de wetten van Newton (b.v. $F = m \cdot a$). Stochastische modellen hebben een bepaalde mate van onzekerheid ingebouwd. De reden kan gebrek aan gegevens zijn, maar vaak ziet men onzekerheid als inherent aan de natuur.

2.2.Terminologie. We spreken over een *experiment*, en de verzameling van alle mogelijke uitkomsten noemen we de *uitkomstenruimte* Γ. *Herhaalde* experimenten zijn verscheidene uitvoeringen van hetzelfde experiment. De herhaalde experimenten vormen tezamen weer een experiment met gecompliceerdere uitkomstenruimte. Bij herhaalde experimenten kan men spreken van de frequentie van een gebeurtenis. Dit is het

aantal keren dat de gebeurtenis optreedt gedeeld door het aantal experimenten. Bij n experimenten waarbij de gebeurtenis A $n(A)$ keer optreedt is dus

$$f_q(A) = \frac{n(A)}{n}$$

de frequentie van gebeurtenis S.

Voorbeeld.
Experiment: gooien met een dobbelsteen
Uitkomstenruimte: $\{1, 2, 3, 4, 5, 6\}$
Herhaalde experimenten: $n\times$ gooien met een dobbelsteen
$n(\{6\})$ = het aantal keren dat 6 is gegooid
$f_q(\{6\}) = n(\{6\})/n$ = de frequentie van 6
Empirisch vastgesteld: $f_q(\{6\}) \approx 1/6$ als n groot (wet van de grote aantallen).

Voorbeeld: Binaire getallen.
Een binair getal ω tussen 0 en 1 kan men schrijven als $X = 0.\omega_1\omega_2\omega_3\ldots$ met $\omega_i \in \{0, 1\}$. Bekijk nu $f_q(\{1\}) :=$ de fractie énen in de eerste n 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 $\lim_{n\to\infty} f_q(\{1\}) = 1/2$.

2.3. Gebeurtenissen. We bekijken de verzameling van alle mogelijke uitkomsten van een experiment (notatie: Γ): de uitkomstenruimte. Een *gebeurtenis* is een deelverzameling van Γ. We noemen een gebeurtenis ook wel een *eventualiteit* (Engels: *event*).

Voorbeeld. Het werpen met een dobbelsteen.
$\Gamma = \{1, 2, 3, 4, 5, 6\}$ en een gebeurtenis is b.v. $A = \{1, 3, 5\}$, een oneven getal.

2.4. Uitkomst. De uitkomst van een experiment is formeel gesproken een deelverzameling van de uitkomstenruimte Γ bestaande uit maar één element.

2.5. Verzamelingenleer. Op verzamelingen A en B kan men de volgende operaties uitvoeren:
$A \cap B$: 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, dan zitten alle elementen van B ook in 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.

2.6. Axioma's. We noemen P een *kans* op de gebeurtenissen in Γ als
(1) $0 \le P(A) \le 1$ voor alle gebeurtenissen $A \subset \Gamma$,
(2) $P(\emptyset) = 0$,
(3) $P(\Gamma) = 1$,
(4) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ voor alle gebeurtenissen $A, B \subset \Gamma$,
(5) Als $A_1, A_2, \ldots$ disjuncte gebeurtenissen zijn (d.w.z. de doorsnede van ieder tweetal is leeg), dan $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$.

2.7. Voorwaardelijke kans. Als $P(B) \neq 0$, dan is de *voorwaardelijke* kans op A *gegeven* B:

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

Voorwaardelijke kansen voldoen ook aan axioma's (1) t/m (5).

Voorbeeld. Men gooit drie keer met een zuivere munt. Wat is nu de kans op minstens $1\times$ kruis gegeven minstens $2\times$ munt? Noem X het aantal keren kruis. Dan is de kans op minstens $2\times$ munt gelijk aan

$$P(X \le 1) = P(X = 0) + P(X = 1) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}.$$

Minstens 1× kruis èn minstens 2× munt kan alleen als je precies 1× kruis vindt. De kans hierop is

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

Dus het antwoord is

$$P(X \geq 1 | X \leq 1) = \frac{\frac{3}{8}}{\frac{1}{2}} = \frac{3}{4}.$$

2.8. De regel van Bayes. Soms zijn alleen voorwaardelijke kansen gegeven. De onvoorwaardelijke kansen kan men dan terugvinden m.b.v. de eigenschap hieronder. Laat $B_1, \ldots B_k$ heet een *partitie* van Γ zijn, d.w.z. $B_1, \ldots, B_k$ zijn disjunct en $B_1 \cup \ldots \cup B_k = \Gamma$. Dan geldt voor iedere gebeurtenis A,

$$P(A) = \sum_{i=1}^{k} P(A \cap B_i) = \sum_{P(B_i) \neq 0} \frac{P(A \cap B_i)}{P(B_i)} P(B_i) = \sum_{P(B_i) \neq 0} P(A|B_i) P(B_i).$$

Voorbeeld. 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

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

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 $\frac{1}{2}$:

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

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

$$P(K = 1 | J = 0) = 1.$$

Hieruit volgt volgens de regel van Bayes dat

$$P(K = 1) = P(K = 1|J = 1)P(J = 1) + P(K = 1|J = 0)P(J = 0)$$

$$= \frac{1}{2} \times \frac{2}{3} + 1 \times \frac{1}{3} = \frac{2}{3}.$$

M.a.w. K heeft dezelfde kans om te overleven als J. Het maakt dus niet uit wie de eerste keus heeft.

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

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

De interpretatie: gebeurtenis B zegt niets over het al of niet optreden van gebeurtenis A (en andersom). Dus (als $P(B) \neq 0$) A en B dan en slechts dan o.o. als $P(A|B) = P(A)$.

2.10. Onderling onafhankelijke stochastische grootheden. Twee stochastische grootheden X en Y heten *onderling onafhankelijk* (*o.o.*) als voor iedere A en B

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

We noemen $X_1, \ldots, X_n$ onderling onafhankelijk als voor alle $A_1, \ldots, A_n$

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_1 \in A_1) \ldots P(X_n \in A_n).$$

Voorbeelden.

(1) In Zeeland is een bouwwerk gemaakt bestaande uit 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 %. De pijlers functioneren op onderling onafhankelijke wijze. Als één pijler het niet doet zullen er overstromingen zijn. Het is dus van belang dat àlle 60 pijlers goed functioneren. De kans hierop is echter ongeveer $(0.95)^{60} < 0.05$!

(2) 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 toevallig zou zijn als alle veiligheidsvoorzorgen tegelijkertijd het laten afweten. Vaak is impliciet in deze redenering, de veronderstelling dat de veiligheidsmechanismen o.o. zijn. Immers, dan is de kans dat àlle veiligheidsmechanismen niet werken gelijk aan het product van de kansen dat één veiligheidsmechanisme niet werkt. Deze kans is dan kleiner naarmate er meer veiligheidsmechanismen zijn. Bij een risico-analyse is het daarom van groot belang om na te gaan of de veronderstelling van onderlinge onafhankelijkheid wel klopt.

2.11. Combinatoriek. Stel dat men de kans op een gebeurtenis A wil weten, bijvoorbeeld bij het aselect kiezen uit een verzameling van m elementen. Dan is het van belang te weten hoeveel uitkomsten er in A zitten. Daarbij is enige kennis van de combinatoriek goed bruikbaar. Hieronder volgen de belangrijkste regels.

(A) Het aantal rijtjes van lengte k van n symbolen is n^k.

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

(C) Het aantal rijtjes van lengte k van n symbolen zodanig dat niet twee keer dezelfde optreedt is

$$\frac{n!}{(n-k)!}.$$

We definiëren $0! = 1$ (dus voor het geval $n = k$ zijn we terug in situatie (B)).

(D) Het aantal manieren om uit n symbolen er k te kiezen is

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

(spreek uit: *n boven k*). Men noemt $\binom{n}{k}$ een *binomiaal coëfficiënt*. Het verschil met (C) is dat we niet op de ordening letten. Merk op dat het aantal manieren om er k te kiezen gelijk is aan het aantal manieren om er $(n-k)$ (niet) te kiezen, d.w.z.

$$\binom{n}{k} = \binom{n}{n-k}.$$

2.12. Eigenschappen van binomiaal coëfficiënten.

(1) De *driehoek van Pascal* is

$$\begin{array}{ccccccc} & & & 1 & & & \\ & & 1 & & 1 & & \\ & 1 & & 2 & & 1 & \\ 1 & & 3 & & 3 & & 1 \\ 1 & 4 & & 6 & & 4 & 1 \\ & & & \ldots & & & \end{array}$$

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

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

(2) Er geldt:

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

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

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

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

(3) Het *binomium van Newton* is de formule

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

(4) We gooien n keer met een munt. Laat X het aantal keren kruis zijn. Dan bezit X een *binomiale verdeling*. Deze ziet er als volgt uit. Noem p de kans op kruis bij één keer gooien ($p = 1/2$ bij een zuivere munt). Bij n keer gooien is de kans op een geordend rijtje met (precies) x keer kruis gelijk aan $p^x(1-p)^{n-x}$ Het aantal rijtjes met x keer kruis is $\binom{n}{x}$. We vinden zo dat de kans op x keer kruis gelijk is aan

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

(5) *Steekproefcontrole*. 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. Dit kan op twee manieren:

(5a) 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:

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

(5b) Steekproef *zonder* teruglegging. Trek n keer aselect een chip, en leg deze apart (we veronderstellen hier dat $n \leq N$). Het aantal manieren waarop men n elementen uit N kan kiezen is $\binom{N}{n}$. Het aantal manieren om x elementen te kiezen uit R, en $n-x$ uit de overige $N-R$ is $\binom{R}{x}\binom{N-R}{n-x}$. Dus

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

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 (5a) als in geval (5b) (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.

3. Voorbeelden van kansverdelingen. Een *stochastische grootheid* beschrijft de uitkomst van een experiment. We gebruiken de afkorting *s.g.*. Stochastische grootheden worden meestal met hoofdletters (X, Y, etc.) aangegeven. We spreken af dat $X \in \mathbf{R}$, de reële getallen. Soms is dat natuurlijk, b.v. als X de executietijd van een programma is, soms is het echter een codering. Antwoorden $\{ja, nee\}$ op een vraag

kan men b.v. met $\{1, 0\}$ coderen. Gebeurtenissen zijn nu van de vorm $\{X \in A\}$ met $A \subset \mathbf{R}$. We maken onderscheid tussen discrete s.g.$\underline{n}$ en continue s.g.$\underline{n}$.

Voorbeelden.

a) Het aantal functionerende verbindingen in een electriciteitscircuit is een discrete s.g..

b) Het aantal ogen bij het gooien van een dobbelsteen is een discrete s.g..

c) De executietijd van een programma is een continue s.g..

d) Analoge signalen zijn continue s.g.$\underline{n}$, digitale signalen zijn discrete s.g.$\underline{n}$.

3.1. Discrete stochastische grootheden. X is een *discrete* s.g. als X maar eindig of aftelbaar veel waarden kan aannemen

3.2. Discrete verdeling. Stel X is een discrete s.g. met mogelijke waarden waarden $\{w_1, w_2, \ldots\}$. Definieer

$$p_j = P(X = w_j),\ j = 1, 2, \ldots.$$

Dus p_j is de kans op uitkomst w_j. We noemen $p_1, p_2, \ldots$ de *verdeling* van X. De verdelings*functie* van X is

$$F(x) = \sum_{w_j \leq x} p_j = P(X \leq x).$$

3.3. Eigenschappen van discrete verdelingen.

(i) $0 \leq p_j \leq 1$ en $\sum_j p_j = 1$,

(ii) $0 \leq F(x) \leq 1$ en $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$,

(iii) $F(x)$ is een stijgende (d.w.z. niet-dalende) functie,

(iv) $F(x)$ springt p_j omhoog bij w_j, $j = 1, 2, \ldots$ en is constant tussen de sprongpunten.

(v) $P(s < X \leq t) = F(t) - F(s)$.

3.4. Continue stochastische grootheden. X is een *continue* s.g. als X alle waarden in een zeker interval kan aannemen.

3.5. Dichtheid. Met een continue s.g. kunnen we vaak een *dichtheid* $f(x)$ associëren, die de *aannemelijkheid* van de waarde x aangeeft. De dichtheid f is zo gedefinieerd dat voor alle $s < t$,

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

M.a.w., de kans op een interval is gelijk aan de oppervlakte onder de grafiek van f, bij dat interval.

De verdelingsfunctie $F(x)$ van een continue s.g. is net zo gedefinieerd als bij discrete stochastische grootheden, n.l.

$$F(x) = P(X \leq x).$$

Bij een continue s.g. betekent dit dat

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

zodat F een primitieve van f is, ofwel $f(x) = dF(x)/dx$. Bij een continue s.g. is de verdelingsfunctie ook continu (terwijl de verdelingsfunctie van een discrete s.g. een trapfunctie is).

3.6. Eigenschappen van continue verdelingen.

(i) $f(x) \geq 0$, $\int_{-\infty}^{\infty} f(x)dx = 1$,

(ii) $0 \leq F(x) \leq 1$, $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$,

(iii) $F(x)$ is een stijgende (d.w.z. niet-dalende) functie,

(iv) $F(x)$ is continu,

(v) $P(s < X \leq t) = F(t) - F(s)$.

(vi) $P(s \leq X \leq t) = P(s < X < t) = P(s \leq X < t) = P(s < X \leq t)$ en $P(X = x) = 0$.

3.7. Voorbeelden van discrete verdelingen.

(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 $x \geq x_0$.

(2) **Alternatieve verdeling met parameter p.** X bezit een *alternatieve* verdeling als X slechts 2 waarden kan aannemen. Zonder verlies van algemeenheid noemen we deze waarden 1 en 0. Er geldt dus $P(X=1) = 1 - P(X=0) = p$ (zeg). We noemen p wel de *succeskans*.

(3) **Binomiale verdeling met parameters n en p.** X bezit een binomiale verdeling als

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

waarbij n en p parameters zijn. Als $X_1, \ldots, X_n$ o.o. en alternatief verdeeld zijn, met $P(X_i = 1) = 1 - P(X_i = 0) = p$, dan is $X = \sum_{i=1}^n X_i$ binomiaal verdeeld met parameters n en p.

(4) **Hypergeometrische verdeling.** X bezit een hypergeometrische verdeling als

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

(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 $p = P(Y_i = 1)$. Dan, onder de aanname dat het al of niet fout lopen voor de individuele executies o.o. zijn,

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

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. $\tilde{X}$, de wachttijd tot het voor de k-de keer fout loopt. Dan

$$P(\tilde{X} = x) = P(\sum_{i=1}^{x-1} Y_i = k-1, Y_x = 1)$$

$$= \binom{x-1}{k-1} p^k (1-p)^{x-k},\ x = k, k+1, \ldots.$$

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

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

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

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 intervallletje i is ongeveer evenredig met de lengte van dat intervalletje: $P(X_i = 1) \approx \lambda T/n$. Hier is λ de evenredigheidsconstante.

(b) De kans op meer dan één login in een klein intervalletje is ongeveer nul: $P(X_i > 1) \approx 0$.

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

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

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

(zie ook voorbeeld (3)). Herschrijven geeft

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

Er geldt

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

en

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

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

Dus

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

M.a.w. X is Poisson verdeeld met parameter $\mu=\lambda T$. We noemen μ de intensiteit. Als μ groot is betekent dat dat het druk is.

3.8. Voorbeelden van continue verdelingen.

(1) **Uniforme verdeling op** $[a,b]$**.** De s.g. X bezit een *uniforme* (of homogene) verdeling 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

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

en

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

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

(2) **Normale verdeling met parameters** μ **en** σ^2**.** X bezit een *normale* verdeling als de dichtheid is:

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

Hier zijn $\mu\in\mathbf{R}$ en $\sigma^2>0$ parameters, en σ is de positieve wortel uit σ^2. De parameter μ geeft het maximum van $f(x)$ aan, en $\mu\pm\sigma$ zijn de buigpunten. De breedte van de grafiek wordt bepaald door σ. De notatie voor de normale verdeling is: $N(\mu,\sigma^2)$-verdeling. We schrijven soms $X\sim N(\mu,\sigma^2)$, waarmee dan bedoeld wordt dat X normaal verdeeld is met parameters μ en σ^2.

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

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

en de standaard normale verdelingsfunctie is

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

Deze kan verder niet expliciet worden uitgerekend, maar er bestaan wel tabellen van.

Stel nu dat $X \sim N(\mu, \sigma^2)$. Dan $Y := (X - \mu)/\sigma \sim N(0, 1)$. Andersom geldt ook: als $Y \sim N(0, 1)$ dan $X := \sigma Y + \mu \sim N(\mu, \sigma^2)$. Dus als $F(x)$ de verdelingsfunctie van X is, dan

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

Zo kan men m.b.v. de tabel voor de standaard normale verdeling, de verdelingsfunctie voor iedere andere normale verdeling berekenen. Nu is $\Phi(x)$ meestal alleen getabelleerd voor $x \geq 0$. Omdat $\phi(x)$ symmetrisch rond $x = 0$ is, geldt echter

$$\Phi(x) = 1 - \Phi(-x)$$

zodat $\Phi(x)$ voor negatieve waarden van x ook uit de tabel af te lezen is.

(3) **Exponentiële verdeling met parameter λ.** X bezit een *exponentiële* verdeling als de dichtheid is:

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

Hier is $\lambda > 0$ weer een parameter. De verdelingsfunctie is nu

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

We zullen een interpretatie geven aan de hand van een voorbeeld. Laat X het tijdsinterval zijn, dat verloopt tussen twee opeenvolgende auto's dat langs een vast punt langs de snelweg raast. Noem Y_T het aantal auto's dat langs dit punt komt gedurende een tijdsinterval van lengte T. Stel dat Y_T Poisson verdeeld is met parameter λT (zie voorbeeld (5) van de discrete verdelingen). Dan

$$P(Y_T = k) = \frac{(\lambda T)^k}{k!} \mathrm{e}^{-\lambda T},$$

voor $k \in \{0, 1, \ldots\}$. Zo vinden we

$$P(X \leq x) = P(\text{minstens 1 auto in tijdsinterval met lengte } x)$$

$$= 1 - P(\text{geen auto}'\text{s in tijdsinterval met lengte } x) = 1 - P(Y_x = 0) = 1 - \mathrm{e}^{-\lambda x}.$$

We zien dat X exponentieel verdeeld is met parameter λ. Als de intensiteit λ groot is, komen er veel auto's langs, en zal men over het algemeen niet lang op de volgende auto hoeven te wachten.

3.9. Onderling onafhankelijke stochastische grootheden. Laat $X_1, \ldots, X_n$ een rij van stochastische grootheden zijn. We definiëren dan de n-dimensionele (*simultane*) verdelingsfunctie als

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n), \ (x_1, \ldots, x_n) \in \mathbf{R}^n.$$

Noem F_i de verdelingsfunctie van X_i, $i = 1, \ldots, n$. Dan zijn $X_1, \ldots, X_n$ o.o. dan en slechts dan als

$$F(x_1, \ldots, x_n) = F_1(x_1) \ldots F_n(x_n), \text{ voor alle } (x_1, \ldots, x_n) \in \mathbf{R}^n.$$

Als $X_1, \ldots, X_n$ discrete s.g.$^{\underline{n}}$ zijn, dan zijn ze o.o., dan en slechts dan als

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \ldots P(X_n = x_n), \text{ voor alle } (x_1, \ldots, x_n) \in \mathbf{R}^n.$$

Als $X_1, \ldots, X_n$ continue s.g.$^{\underline{n}}$ zijn, dan zijn ze o.o., dan en slechts dan als

$$f(x_1, \ldots, x_n) = f_1(x_1) \ldots f_n(x_n), \text{ voor alle } (x_1, \ldots, x_n) \in \mathbf{R}^n.$$

Hier is

$$f(x_1, \ldots, x_n) = \frac{\partial F(x_1, \ldots, x_n)}{\partial x_1 \ldots \partial x_n},$$

de n-dimensionele (*simultane*) dichtheid van $X_1, \ldots, X_n$, en f_i de dichtheid van X_i, $i = 1, \ldots, n$ (waarbij we veronderstellen dat deze bestaan).

3.10. De verdeling van de som. Stel X en Y zijn twee o.o. stochastische grootheden. De verdeling van $X + Y$ is dan de *convolutie* van de verdeling van X en de verdeling van Y. Zo'n convolutie kan in het algemeen lastig uit te rekenen zijn. Verder is het zo dat als X en Y een verdeling van een bepaald type hebben, dan bezit $X + Y$ in het algemeen *niet* een verdeling van dat type. Bijvoorbeeld, als X en Y exponentieel verdeeld zijn, dan is $X + Y$ niet exponentieel verdeeld. Enkele uitzonderingen op dit negatieve verschijnsel zijn:

(a) Als X binomiaal verdeeld is met parameters n en p, en Y binomiaal verdeeld is met parameters m en p, dan is $X + Y$ binomiaal verdeeld met parameters $n + m$ en p (zie Opgave 29).

(b) Als X en Y Poisson verdeeld zijn, dan bezit $X + Y$ ook een Poisson verdeling (zie Opgave 30).

(c) Als $X \sim N(\mu, \sigma^2)$ en $Y \sim N(\nu, \tau^2)$, dan $X + Y \asymp N(\mu + \nu, \sigma^2 + \tau^2)$. Meer algemeen: $aX + bY + c \sim N(a\mu + b\nu + c, a^2\sigma^2 + b^2\tau^2)$. (Geen bewijs.)

3.11. Construeren van continue verdelingen. Stel U is uniform verdeeld op $[0, 1]$. Laat F een continue, strict stijgende verdelingsfunctie zijn. Dan is $X = F^{-1}(U)$ een s.g. met verdelingsfunctie F. Immers

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

We kunnen dus uitgaande van de uniforme verdeling de meeste andere continue verdelingen construeren.

Voorbeeld. Stel we willen 100 waarnemingen uit de volgende verdeling:

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

De inverse is nu

$$F^{-1}(u) = \frac{u}{1-u}, \ u \in [0, 1).$$

In Splus

```
> n <-100
> u<-runif(100)
> x<-u/(1-u)
> x<-sort(x)
>Fn<-1:n/n
>plot(x,Fn,type="s",main="F(x)=1-1/(1+x)")
```

In veel gevallen is het lastig de verdelingsfunctie te inverteren, terwijl er wel een nette uitdrukking voor de dichtheid f is. Stel nu dat f een dichtheid is op een eindig interval $[a, b]$, en dat $f(x) \leq c$. Neem dan twee o.o. s.g.^{n} X en Y, met X uniform verdeeld op $[a, b]$ en Y uniform verdeeld op $[0, c]$. Dan is de verdeling van X gegeven $Y \leq f(X)$ precies de gezochte verdeling met dichtheid f.

Voorbeeld. Stel

$$f(x) = 6x(1 - x), \ x \in [0, 1].$$

Dan geldt

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

Deze is lastig te inverteren. We gaan nu aselect trekken uit het gebied onder de grafiek van f, om waarnemingen uit bovenstaande F te genereren.

```
> w<-1:1000/1000
> fw<-6*w*(1-w)
> plot(w,fw,main="dichtheid f")
> Fw<-3*w**2-2*w**3
> plot(w,Fw,main="verdelingsfunctie F")
> n<-200
> x<-runif(n)
> y<-runif(n)*max(fw)
```

```
> fx<-6*x*(1-x)
> for (i in 1:n)
> if (y[i]>fx[i]) x[i]<-1
> x<-sort(x)
> x
[1] 0.05739141 0.09510574 0.09632206 0.10894097 0.12426849 0.12480332
[7] 0.13716313 0.14451770 0.14831981 0.15509735 0.16497428 0.16572703
[13] 0.20748320 0.20839370 0.21451466 0.21921856 0.23708377 0.23745860
[19] 0.24447818 0.25774763 0.26021103 0.26325308 0.26886361 0.26962035
[25] 0.27080283 0.27102546 0.28127016 0.28230138 0.28421747 0.29197379
[31] 0.29888199 0.30210348 0.31154216 0.33701534 0.33745031 0.33754155
[37] 0.34385675 0.35114416 0.35121903 0.35667167 0.35912167 0.36168343
[43] 0.36448175 0.36491687 0.36647552 0.37344540 0.37724002 0.38033625
[49] 0.40155399 0.41175656 0.41222125 0.41490265 0.41508581 0.41769230
[55] 0.42000086 0.42191959 0.42279283 0.42341976 0.42422251 0.42635215
[61] 0.43446630 0.43896168 0.45263497 0.46878322 0.47181784 0.47940891
[67] 0.48442668 0.48492021 0.48626929 0.48908238 0.49078221 0.50747610
[73] 0.50900902 0.52205908 0.52400099 0.53425469 0.54202298 0.54399321
[79] 0.54481964 0.54511086 0.55874832 0.55927327 0.56825010 0.56924045
[85] 0.57223703 0.57646318 0.57738428 0.58787771 0.59719462 0.60065769
[91] 0.60957093 0.61363221 0.61808333 0.62462990 0.62654482 0.62833651
[97] 0.62922727 0.63206508 0.63619266 0.64347180 0.65025440 0.65140896
[103] 0.65486187 0.66049886 0.67262233 0.67496820 0.67619768 0.67703501
[109] 0.67750035 0.68150075 0.68813619 0.69505422 0.69518493 0.70060586
[115] 0.70273611 0.71021159 0.71130019 0.73087059 0.73092217 0.73887201
[121] 0.74731262 0.76181274 0.77536268 0.79292942 0.79350472 0.79644512
[127] 0.82457222 0.83709437 0.84603504 0.84932118 0.87545489 0.88974498
[133] 0.90880412 0.91836347 0.95308807 0.96793078 1.00000000 1.00000000
[139] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
[145] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
[151] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
[157] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
[163] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
[169] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
[175] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
[181] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
[187] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
[193] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
[199] 1.00000000 1.00000000
> nn<-136
> xx<-x[1:nn]
> Fnn<-1:nn/nn
> plot(xx,Fnn,type="s",main="trekking uit opp.  onder grafiek f")
```

3.12. QQ-plots. Stel nu dat X een continue, strict stijgende verdelingsfunctie F heeft. Dan is $F(X)$ uniform verdeeld op $[0,1]$. Dit kan men als volgt toepassen. Laat $X_1,\ldots,X_n$ een steekproef zijn uit een onbekende verdeling. We willen nagaan of het een steekproef uit F is. Als dit het geval is, dan zal de empirische verdelingsfunctie van $F(X_1),\ldots,F(X_n)$ ongeveer op de rechte lijn $y=x$ liggen. We noemen het plaatje een *QQ-plot* (Q = *Quantile*).

```
> n<-100
> x<-rnorm(n)
> # dit levert n waarnemingen uit de N(0,1)-verdeling
> x<-sort(x)
```

```
> Fn<-1:n/n
> plot(x,Fn,type="s",main="PP-plot voor N(0,1)-verdeling")
> z<-1:7000/1000 - 3.5
> lines(z,pnorm(z))
> # pnorm is de standaard normale verdelingsfunctie
> u<-pnorm(x)
> plot(u,Fn,type="s",main="QQ-plot voor N(0,1)-verdeling")
```

4. Verwachting en variantie.

4.1. De verwachting van een discrete stochastische grootheid. Stel X is een discrete stochastische grootheid met waarden $\{w_1, w_2, \ldots\}$. Dan heet

$$EX = \sum_j w_j P(X = w_j)$$

de *verwachting* van X, en

$$Eg(X) = \sum_j g(w_j) P(X = w_j)$$

de verwachting van de functie $g(X)$ van X. Hier staat E voor *expectation*.

Voorbeeld. Laat X het aantal ogen zijn bij één keer gooien met een dobbelsteen. Dan

$$EX = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5,$$

en bijvoorbeeld

$$EX^2 = 1 \times \frac{1}{6} + 4 \times \frac{1}{6} + 9 \times \frac{1}{6} + 16 \times \frac{1}{6} + 25 \times \frac{1}{6} + 36 \times \frac{1}{6} = \frac{91}{6}.$$

4.2. Zwaartepunt. Het fysisch analogon van verwachting is *zwaartepunt*. Stel we leggen massa's $\pi_1, \pi_2, \ldots$ op de punten $w_1, w_2, \ldots$. Dan is het zwaartepunt $(\sum_i w_j \pi_j)/(\sum_j \pi_j) = \sum_j w_j p_j$ met $p_j = \pi_j/(\sum_j \pi_j)$, $j = 1, 2, \ldots$.

4.3. De verwachting van een continue stochastische grootheid. Stel X is een continue s.g. met dichtheid $f(x)$. Dan heet

$$EX = \int_{-\infty}^{\infty} x f(x) dx$$

de verwachting van X, en

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

de verwachting van de functie $g(X)$ van X.

Voorbeeld. Stel X is homogeen verdeeld op het interval $[0, 1]$. Dan

$$EX = \int_0^1 x dx = 1/2$$

en bijvoorbeeld

$$E\cos(X) = \int_0^1 \cos(x) dx = \sin(1).$$

4.4. Gemiddelde. Men noemt de verwachting van een stochastische grootheid ook wel het *gemiddelde*. Dit kan helaas verwarring geven, want in 1.10 hadden we het al over het gemiddelde van een rij getallen, en zelfs over het gemiddelde van een rij stochastische grootheden. Stel nu dat $X_1, \ldots, X_n$ een steekproef is uit X (zie 1.7). Dan noemen we $\bar{X} = \sum_{i=1}^n X_i/n$ ook wel het *steekproefgemiddelde*, en EX het *populatiegemiddelde* (ook wel *theoretisch gemiddelde*). Merk op dat $\bar{X}$ een stochastische grootheid is, terwijl EX een getal (niet-stochastisch) is. We zullen laten zien (zie 4.5) dat $E\bar{X} = EX$. Bovendien zijn $\bar{X}$ en EX voor grote steekproeven ongeveer gelijk aan elkaar (wet van de grote aantallen). We noemen $\bar{X}$ ook wel een *schatter* van EX.

Voorbeeld. Stel we gooien n keer met een dobbelsteen. Laat X_i het aantal ogen zijn bij de i-de worp. Dan $\bar{X} \approx 3.5$.

```
> # we gooien 50 keer met een dobbelsteen
> n<-50
> x<-ceiling(runif(n)*6)
> x
```

```
[1] 3 1 1 4 3 2 1 2 2 6 2 2 6 2 1 6 1 4 6 6 2 4 5 5 4 1 3 4 3 4 6 6 4 1 5 1 6 1
[39] 2 4 4 2 3 1 3 3 1 1 4 3
> # het steekproefgemiddelde is
> mean(x)
[1] 3.14
```

Voorbeeld. Stel we trekken n keer uit de uniforme verdeling op $[0,1]$. Dan is $\bar{X} \approx \frac{1}{2}$ en $\bar{Y} \approx \sin(1)$.

```
> n<-50
> x<-runif(n)
> mean(x)
[1] 0.507087
> y<-cos(x)
> mean(y)
[1] 0.8326261
> sin(1)
[1] 0.841471
```

4.5. De verwachting van de som. Zij X en Y twee stochastische grootheden (discreet of continu) en a, b, c *getallen*. Dan

$$E(aX + bY + c) = aEX + bEY + c.$$

Als $X_1, \ldots, X_n$ een rij van stochastische grootheden is, dan vinden we door het bovenstaande herhaald toe te passen:

$$E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} EX_i.$$

In woorden: de verwachting van de som is de som van de verwachtingen. Als $X_1, \ldots, X_n$ een steekproef is uit X vinden we

$$E\bar{X} = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}E(\sum_{i=1}^{n} X_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n}\sum_{i=1}^{n} EX = \frac{1}{n}nEX = EX.$$

Dus de verwachting van het steekproefgemiddelde is het theoretisch gemiddelde.

4.6. De verwachting van enkele discrete verdelingen.

(1) **Ontaarde verdeling.** Als $P(X = x_0) = 1$, dan $EX = x_0$.

(2) **Alternatieve verdeling met parameter p.** Stel $P(X = 1) = 1 - P(X = 0) = p$. Dan

$$EX = 1 \times p + 0 \times (1 - p) = p.$$

Dus de verwachting van X is gelijk aan de succeskans.

(3) **Binomiale verdeling met parameters n en p.** Als $X_1, \ldots, X_n$ o.o. en alternatief verdeeld zijn, met $P(X_i = 1) = 1 - P(X_i = 0) = p$, dan is $X = \sum_{i=1}^{n} X_i$ binomiaal verdeeld met parameters n en p. Dus door gebruik te maken van het resultaat in (2), en toepassing van 4.5, vinden we

$$EX = \sum_{i=1}^{n} EX_i = np.$$

(4) **Hypergeometrische verdeling.** X bezit een hypergeometrische verdeling als

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

We kunnen weer schrijven $X = \sum_{i=1}^{n} X_i$ met $P(X_i = 1) = 1 - P(X_i = 0) = p$, $i = 1, \ldots, n$, en met $p = R/N$ de succeskans. Er geldt daarom ook hier

$$EX = np.$$

(5) **Negatief binomiale verdeling met parameters k en p.** Stel X bezit de negatief binomiale verdeling met parameters k en p. Dan

$$EX = \frac{k}{p}.$$

(6) **Poissonverdeling met parameter μ.** X bezit een Poissonverdeling met parameter $\mu > 0$ als

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

Dan

$$EX = \sum_{x=0}^{\infty} x\frac{\mu^x}{x!}\mathrm{e}^{-\mu}$$

$$= \mu \sum_{x=1}^{\infty} \frac{\mu^{x-1}}{(x-1)!}\mathrm{e}^{-\mu} = \mu,$$

4.7. De verwachting van enkele continue verdelingen.

(1) **Uniforme verdeling op $[a, b]$.** De s.g. X bezit een uniforme verdeling op $[a, b]$ als de dichtheid f van X gelijk is aan

$$f(x) = \frac{1}{b-a},\ a \le x \le b.$$

Dus

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

(2) **Normale verdeling met parameters μ en σ^2.** Stel Y is $N(0, 1)$-verdeeld. Dan

$$EY = \int_{-\infty}^{\infty} y\phi(y)dy,$$

met $\phi(y)$ de dichtheid van de standaard-normale verdeling (zie 3.8 (2)). Omdat $\phi(y)$ symmetrisch is rond $y = 0$ is $EY = 0$. Als nu X $N(\mu, \sigma^2)$-verdeeld is dan is $Z := (X - \mu)/\sigma$ $N(0, 1)$-verdeeld. We zien dat $EX = E(\sigma Z + \mu) = \mu$.

(3) **Exponentiële verdeling met parameter λ.** X bezit een *exponentiële* verdeling als de dichtheid is:

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

M.b.v. partiële integratie vinden we

$$EX = \int_0^{\infty} x\lambda\mathrm{e}^{-\lambda x}dx = \frac{1}{\lambda}.$$

4.8. De variantie. De *variantie* van een s.g. X is de verwachte kwadratische afwijking van het gemiddelde:

$$\mathrm{var}(X) = E(X - EX)^2.$$

De *standaardafwijking* van X is

$$\sigma_X = \sqrt{\mathrm{var}(X)}.$$

We schrijven ook wel σ_X^2 voor de variantie van X. De standaardafwijking is een maat voor de *spreiding*.

4.9. Andere schrijfwijze. Er geldt

$$\mathrm{var}(X) = EX^2 - (EX)^2,$$

want, als we $EX = \mu$ noemen, dan

$$\text{var}(X) = E(X-\mu)^2 = EX^2 - 2\mu EX + \mu^2$$

$$= EX^2 - 2\mu^2 + \mu^2 = EX^2 - \mu^2.$$

Voorbeeld. Bij het roulettespel zetten we één gulden in op oneven. De kans om een gulden te winnen is 18/37 en de kans om een gulden te verliezen is 19/37 (nul is hier een even getal). Dus als X onze winst is, dan $EX = -1/37$ en $EX^2 = 1$, dus $\text{var}(X) = 1 - (1/37)^2 = 0.9993$. We kunnen ook een andere strategie kiezen. Stel we zetten één gulden in op 23. De winst is dan $X = 35$ met kans $1/37$ en $X = -1$ met kans $36/37$. Dus $EX = -1/37$, net als bij de vorige strategie. Maar $EX^2 = (35)^2/37 + 36/37 = 1261/37$ zodat $\text{var}(X) = 1261/37 - (1/37)^2 = 34.0803$. De spreiding van deze strategie is dus veel groter, wat betekent dat je meer risico neemt, maar ook meer kan winnen.

4.10. Eigenschappen van variantie.

(i) $\text{var}(X) = E(X - EX)^2 \geq 0$. Hier volgt ook uit dat $EX^2 \geq (EX)^2$, want (zie 4.9) $\text{var}(X) = EX^2 - (EX)^2$.

(ii) $\text{var}(X) = 0$ dan en slechts dan als X een ontaarde verdeling bezit, d.w.z. voor zekere constante x_0 is $P(X = x_0) = 1$. Deze constante is dan $x_0 = EX$ (een s.g. die alleen de waarde x_0 kan aannemen heeft natuurlijk ook verwachting x_0). We zeggen ook wel dat X volledig *geconcentreerd* is in x_0. In het algemeen geeft ook de variantie de mate van concentratie van X rond EX aan.

(iii) Als a en b getallen zijn, dan

$$\text{var}(aX + b) = a^2\text{var}(X).$$

Immers, noem $EX = \mu$. Dan $E(aX+b) = a\mu+b$ en $\text{var}(aX+b) = E((aX+b)-(a\mu+b))^2 = E(aX-a\mu)^2 = E(a^2(X-\mu)^2) = a^2E(X-\mu)^2 = a^2\text{var}(X)$.

4.11. Steekproefvariantie. Laat $X_1, \ldots, X_n$ een steekproef zijn uit de populatie s.g. X, met (populatie)variantie $\sigma^2 = \text{var}(X)$. We noemen dan

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

de *steekproefvariantie*. Hier is $\bar{X}$ weer het steekproefgemiddelde (zie 1.7 en 4.4). 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 kan je dus ook

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

als benadering gebruiken. We noemen S^2 en $\hat{\sigma}^2$ *schatters* van σ^2 (en S en $\hat{\sigma}$ schatters van σ).

Voorbeeld.

```
> # we gooien 50 keer met een dobbelsteen
> n<-50
> x<-ceiling(runif(n)*6)
> x
[1] 3 2 4 4 5 2 6 6 3 3 5 3 2 3 4 4 3 5 2 3 1 6 2 2 2 1 3 4 4 5 6 5 6 6 3 1 5 2
[39] 4 5 6 2 1 2 2 2 4 1 6 6
> # het steekproefgemiddelde is
> mean(x)
[1] 3.54
> # de steekproefvariantie is
> var(x)
[1] 2.743265
```

```
> # even controleren of Splus dat goed gedaan heeft ;)
> (sum((x-mean(x))**2))/(n-1)
[1] 2.743265
> # de theoretische variantie is
> (91/6)-(49/4)
[1] 2.916667
> # nu voor de uniforme verdeling op [0,1]
> x<-runif(n)
> mean(x)
[1] 0.4417616
> var(x)
[1] 0.09253191
> # de theoretische variantie is 1/12 (zie 4.13 (1))
> 1/12
[1] 0.08333333
> # laten we n wat groter kiezen
> x<-runif(1000)
> mean(x)
[1] 0.5021352
> var(x)
[1] 0.08574171
> 1/12
[1] 0.08333333
> # nu voor y=cos(x)
> y<-cos(x)
> mean(y)
[1] 0.8394455
> # het theoretisch gemiddelde is
> sin(1)
[1] 0.841471
> var(y)
[1] 0.01970543
> # wat is de theoretische variantie??
```

4.12. De variantie van de som. Stel X en Y zijn twee o.o. stochastische grootheden. Dan

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y).$$

M.a.w. voor *onafhankelijke* stochastische grootheden is de variantie van de som gelijk aan de som van de varianties. Voor *afhankelijke* stochastische grootheden is dit i.h.a. niet het geval. We gaan hier in 4.21 wat meer op in.

4.13. De (verwachting en) variantie van enkele discrete verdelingen.

(1) **Ontaarde verdeling.** Als $P(X = x_0) = 1$, dan $EX = x_0$ en $\mathrm{var}(X) = 0$.

(2) **Alternatieve verdeling met parameter p.** Stel $P(X = 1) = 1 - P(X = 0) = p$. Dan $X^2 = X$, dus

$$EX^2 = EX = p.$$

Volgens 4.9 geldt nu

$$\mathrm{var}(X) = EX^2 - (EX)^2 = p - p^2 = p(1-p).$$

(3) **Binomiale verdeling met parameters n en p.** Als $X_1, \ldots, X_n$ o.o. en alternatief verdeeld zijn, met $P(X_i = 1) = 1 - P(X_i = 0) = p$, dan is $X = \sum_{i=1}^{n} X_i$ binomiaal verdeeld met parameters n en p. Dus door gebruik te maken van het resultaat in (2), en 4.12 toe te passen, vinden we

$$\mathrm{var}(X) = \sum_{i=1}^{n} \mathrm{var}(X_i) = np(1-p).$$

(4) **Hypergeometrische verdeling.** X bezit een hypergeometrische verdeling als

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

Deze verdeling komt naar voren bij een steekproef *zonder* terugleggen. We kunnen weer schrijven $X=\sum_{i=1}^{n}X_i$ met de trekkingen X_i alternatief verdeeld, $P(X_i=1)=1-P(X_i=0)=p$, $i=1,\ldots,n$, waarbij $p=R/N$. Maar $X_1,\ldots,X_n$ zijn *niet* o.o., want als een element eenmaal getrokken is kan deze niet nogmaals getrokken worden. We kunnen daarom 4.12 niet toepassen. Zonder bewijs (liefhebbers: zie 4.22) vermelden we

$$\mathrm{var}(X)=np(1-p)\frac{N-n}{N-1}.$$

De variantie is dus kleiner dan in (3) (als $n>1$).

(5) **Negatief binomiale verdeling met parameters k en p.** Stel X bezit de negatief binomiale verdeling met parameters k en p. Dan $EX=\frac{k}{p}$, en

$$\mathrm{var}(X)=\frac{k(1-p)}{p^2}.$$

(6) **Poissonverdeling met parameter μ.** Er geldt $EX=\mu$, en $EX^2=\mu+\mu^2$, dus

$$\mathrm{var}(X)=\mu.$$

Voor de Poissonverdeling zijn verwachting en variantie dus gelijk.

4.14. De (verwachting en) variantie van enkele continue verdelingen.

(1) **Uniforme verdeling op $[a,b]$.** Dan $EX=\frac{a+b}{2}$, en

$$\mathrm{var}(X)=\frac{b-a}{12}.$$

(2) **Normale verdeling met parameters μ en σ^2.** Stel Y is $N(0,1)$-verdeeld. Dan volgt d.m.v. partiële integratie

$$EY^2=1.$$

Dus ook (omdat $EY=0$), $\mathrm{var}(Y)=1$. Als nu X $N(\mu,\sigma^2)$-verdeeld is dan is $Z:=(X-\mu)/\sigma$ $N(0,1)$-verdeeld. Door 4.10 (iii) toe te passen vinden we

$$\mathrm{var}(X)=\mathrm{var}(\sigma Z+\mu)=\sigma^2\mathrm{var}(Z)=\sigma^2.$$

Samenvattend: bij de $N(\mu,\sigma^2)$-verdeling stelt de lokatieparameter μ de verwachting voor, en en de (gekwadrateerde) schaalparameter σ^2 de variantie.

(3) **Exponentiële verdeling met parameter λ.** We vonden al $EX=1/\lambda$. M.b.v. twee keer partiële integratie vinden we

$$EX^2=\int_0^\infty x^2\lambda\mathrm{e}^{-\lambda x}dx=\frac{2}{\lambda^2}.$$

Dus

$$\mathrm{var}(X)=\frac{1}{\lambda^2}.$$

Bij de exponentiële verdeling is de variantie gelijk aan het kwadraat van de verwachting.

4.15. De covariantie tussen twee stochastische grootheden. Laat X en Y twee stochastische grootheden zijn. De covariantie tussen X en Y is

$$\mathrm{cov}(X,Y)=EXY-EXEY.$$

4.16. Onafhankelijke stochastische grootheden. Stel Xen Y zijn o.o.. Dan $\text{cov}(X,Y) = 0$. Als X en Y discreet verdeeld zijn is dit als volgt in te zien. Er geldt

$$EXY = \sum_i \sum_j x_i y_j P(X = x_i, Y = y_j)$$

De onafhankelijkheidsveronderstelling geeft

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$

voor alle i en j. Dus

$$EXY = \sum_i \sum_j x_i y_j P(X = x_i)P(Y = y_j)$$

$$= \sum_i x_i P(X = x_i) \sum_j y_j P(Y = y_j) = EXEY.$$

4.17. Eigenschappen van covariantie.
(i) $\text{cov}(X,X) = \text{var}(X)$,
(ii) $\text{cov}(X,Y) = E(X-EX)(Y-EY)$, want als we schrijven $EX = \mu$ en $EY = \nu$, dan $E(X-\mu)(Y-\nu) = EXY - \mu EY - \nu EX + \mu\nu = EXY - \mu\nu$.

4.18. Lineair verband. De covariantie is een maat voor een *lineaire* verband tussen stochastische grootheden. We zeggen dat er een *exact* lineair verband is tussen X en Y als voor zekere α en β ($\neq 0$) geldt $Y = \alpha + \beta X$. In het algemeen is er natuurlijk geen exact lineair verband, maar we verwachten wel vaak een relatie in de trant van:"hoe groter X, des te groter Y" (bv. bij lichaamslengte en lichaamsgewicht) of juist: "hoe groter X des te kleiner Y". Merk nu op dat als X en Y o.o. zijn, dan $\text{cov}(X,Y) = 0$, maar dat het omgekeerde niet waar hoeft te zijn.

Voorbeeld. Stel X is homogeen verdeeld op $[-1/2, 1/2]$ en $Y = X^2$. Dan zijn X en Y duidelijk niet o.o., want als men X weet, weet men Y ook. Maar $EX = 0$, $EXY = EX^3 = 0$, dus $EXY = EXEY = 0$, d.w.z. $\text{cov}(X,Y) = 0$.

Voorbeeld. Stel $Y = \alpha + \beta X + V$, waarbij V en X onafhankelijk zijn. Men kan V interpreteren als een verstoring van het lineaire verband. Nu is $EXY = EX(\alpha + \beta X + V) = \alpha EX + \beta EX^2 + EXV$, en $EXEY = EX(\alpha + \beta EX + EV) = \alpha EX + \beta(EX)^2 + EXEV$. Dus $\text{cov}(X,Y) = \beta\text{var}(X)$. We zien dat de covariantie positief is als $\beta > 0$, en anders is de covariantie negatief (of nul).

4.19. Positief of negatief verband. In het algemeen noemen we het geval $\text{cov}(X,Y) > 0$ een positief verband en $\text{cov}(X,Y) < 0$ een negatief verband. Als $\text{cov}(X,Y) = 0$ kan er nog best een zeker verband zijn, er is alleen geen *lineair* verband.

4.20. Steekproefcovariantie. Laat $(X_1,Y_1),\ldots,(X_n,Y_n)$ een steekproef zijn uit een (bivariate) verdeling, dwz n o.o. copietjes van stochastische grootheden (X,Y). Laat σ_{XY} de covariantie zijn tussen X en Y. De steekproefcovariantie is

$$S_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}.$$

Volgens de wet van de grote aantallen geldt weer dat S_{XY} ongeveer gelijk is aan σ_{XY} voor n groot.

Voorbeeld. We trekken de X_i uit een uniforme verdeelde X de Y_i uit een eveneens uniform verdeelde Y, onafhankelijk van de X. Dan is de covariantie tussen X en Y is dus nul. Verder nemen we een steekproef $Z_1,\ldots,Z_n$ uit $Z = Y + X^2$. De stochastische grootheden X en Z zijn dus niet o.o..

```
> n<-100
> x<-runif(n)
> y<-runif(n)
> sxy<-sum((x-mean(x))*(y-mean(y)))/(n-1)
> sxy
[1] -0.004884402
> z<-y+x**2
```

```
> sxz<-sum((x-mean(x))*(z-mean(z)))/(n-1)
> sxz
[1] 0.08622173
```

4.21. De variantie van de som. Er geldt

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X,Y).$$

Immers, noem $EX = \mu$ en $EY = \nu$. Dan

$$\mathrm{var}(X+Y) = E[X+Y-(\mu+\nu)]^2 = E[(X-\mu)^2 + (Y-\nu)^2 + 2(X-\mu)(Y-\nu)]$$

$$= E(X-\mu)^2 + E(Y-\nu)^2 + 2E(X-\mu)(Y-\nu) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X,Y).$$

Dus als X en Y o.o. zijn, dan

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y).$$

Voor een rij $X_1, \ldots, X_n$ vindt men

$$\mathrm{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{var}(X_i) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} \mathrm{cov}(X_i, X_j).$$

4.22. Steekproef met of zonder terugleggen. (Voor liefhebbers.) Bekijk n aselecte trekkingen, uit een populatie van N elementen waarvan er R kenmerk S bezitten. Noem

$$X_i = \begin{cases} 1, & \text{als } S \text{ wordt gevonden in de } i\text{-de trekking ,} \\ 0, & \text{anders.} \end{cases}$$

Bij een steekproef met of zonder terugleggen geldt

$$P(X_i = 1) = \frac{R}{N},\ i = 1, \ldots, n.$$

Dit impliceert dat $EX_i = R/N$, $EX_i^2 = R/N$ en $\mathrm{var}(X_i) = R/N - (R/N)^2 = R/N(1-R/N)$. Zij nu weer $X = \sum_{i=1}^{n} X_i$ het aantal elementen in de steekproef met kenmerk S.

(a) Met terugleggen. $X_1, \ldots, X_n$ zijn o.o., waaruit volgt dat

$$\mathrm{var}(X) = \sum_{i=1}^{n} \mathrm{var}(X_i) = \sum_{i=1}^{n} \frac{R}{N}(1-\frac{R}{N}) = n\frac{R}{N}(1-\frac{R}{N}).$$

(b) Zonder terugleggen.

Er geldt voor $j \neq i$,

$$EX_iX_j = P(X_i = 1, X_j = 1) = \frac{R}{N}\frac{R-1}{N-1}.$$

Dus

$$\mathrm{cov}(X_i, X_j) = \frac{R}{N}\frac{R-1}{N-1} - (\frac{R}{N})^2 = \frac{R}{N}\frac{N-R}{N}\frac{1}{N-1}.$$

We vinden zo

$$\mathrm{var}(X) = \sum_{i=1}^{n} \mathrm{var}(X_i) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} \mathrm{cov}(X_i, X_j)$$

$$= \sum_{i=1}^{n} \frac{R}{N}\frac{N-R}{N} + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1} -\frac{R}{N}\frac{N-R}{N}\frac{1}{N-1}$$

$$= n\frac{R}{N}\frac{N-R}{N} + 2(\frac{n(n-1)}{2})(-\frac{R}{N}\frac{N-R}{N}\frac{1}{N-1})$$
$$= n\frac{R}{N}\frac{N-R}{N}\frac{N-n}{N-1}.$$

De variantie bij een steekproef zonder terugleggen is kleiner dan bij een steekproef met terugleggen.

4.23. Gestandaardiseerde stochastische grootheden. Stel X is een stochastische grootheid met verwachting $EX = \mu$ en variantie $\text{var}(X) = \sigma^2$. Dan heet

$$X^* = (X - \mu)/\sigma$$

de *gestandaardiseerde* van X. Merk op dat $EX^* = 0$ en $\text{var}(X^*) = 1$.

4.24. Correlatie. Laat X en Y twee stochastische grootheden zijn met covariantie $\text{cov}(X, Y) = \sigma_{XY}$, en met varianties $\text{var}(X) = \sigma_X^2$, en $\text{var}(Y) = \sigma_Y^2$. De *correlatie* tussen X en Y is

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

De correlatie is dus de covariantie tussen de gestandaardiseerde stochastische grootheden.

4.25. Eigenschappen van correlatie.

(i) De correlatie is, in tegenstelling tot de covariantie, een dimensielooos begrip, d.w.z. het is onafhankelijk van de meeteenheid. Of X en/of Y b.v. in centimeters of in meters wordt gemeten heeft geen invloed op de correlatiecoëfficiënt.

(ii) Er geldt

$$-1 \leq \rho_{XY} \leq 1.$$

4.26. Steekproefcorrelatie. Beschouw weer een steekproef $(X_1, Y_1), \ldots, (X_n, Y_n)$ uit (X, Y). De steekproefcorrelatie is dan

$$\hat{\rho}_{XY} = \frac{S_{XY}}{S_X S_Y},$$

met S_X^2 en S_Y^2 de steekproefvarianties van X resp. Y. Voor n groot ligt $\hat{\rho}_{XY}$ in de buurt van ρ_{XY}.

Voorbeeld. We gaan verder met het voorbeeld in 4.20.

```
> sx<-sqrt(var(x))
> sy<-sqrt(var(y))
> # de steekproefcorrelatie tussen X en Y is nu
> sxy/(sx*sy)
[1] -0.0546592
> sz<-sqrt(var(z))
> # de steekproefcorrelatie tussen X en Z is
> sxz/(sx*sz)
[1] 0.6818945
```

5. Wet van de grote aantallen en centrale limiet stelling. In dit hoofdstuk is $X_1, \ldots, X_n$ een steekproef zijn uit (de verdeling van) X, d.w.z. $X_1, \ldots, X_n$ zijn o.o. en hebben alle dezelfde verdeling als de populatievariabele X. (Als X verdelingsfunctie F heeft noemen we $X_1, \ldots, X_n$ ook wel een steekproef uit F.)

5.1. De (populatie-)verwachting en variantie. We noteren de verwachting van X met

$$\mu = EX,$$

en de variantie met

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

waarbij we er van uit gaan dat deze bestaan. Dit is niet altijd het geval! Bijvoorbeeld, bij de z.g. Cauchy-verdeling bestaat de variantie niet.

5.2. Steekproefgemiddelde. Het steekproefgemiddelde is

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

We noemen $\bar{X}$ ook wel het empirisch gemiddelde, en μ het theoretisch gemiddelde (zie ook 4.4). Merk op dat $\bar{X}$ een stochastische grootheid is (en μ een getal). Verder hangt $\bar{X}$ af van de steekproefgrootte n. We geven dit soms aan door:

$$\bar{X} = \bar{X}_n.$$

5.3. Verwachting en variantie van het steekproefgemiddelde. Er geldt

$$E\bar{X} = \mu$$

(zie 4.3). De variantie van $\bar{X}$ is

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Om dit in te zien passen we eerst 4.10 (iii) toe:

$$\mathrm{var}(\bar{X}) = \frac{1}{n^2}\mathrm{var}(\sum_{i=1}^{n} X_i).$$

Vervolgens volgt uit herhaald toepassen van 4.12 dat de variantie van de som gelijk is aan de som van de varianties (omdat $X_1, \ldots, X_n$ o.o. zijn):

$$\mathrm{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{var}(X_i)$$

$$= \sum_{i=1}^{n} \sigma^2 = n\sigma^2.$$

Dus

$$\mathrm{var}(\bar{X}) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

5.4. Wet van de grote aantallen. De variantie van $\bar{X}$ is klein als n groot is, d.w.z. $\bar{X}$ concentreert zich dan rond μ. In de limiet wordt de variantie nul, en we hebben gezien dat een s.g. met variantie gelijk aan nul maar één waarde kan aannemen. n.l. zijn verwachting (zie 4.10 (ii)). De wet van de grote aantallen zegt:

$$\lim_{n\to\infty} \bar{X}_n = \mu.$$

5.5. Interpretatie. We kunnen $X_1, \ldots, X_n$ zien als n metingen van de constante μ, met meetfout (error) $\epsilon_i = X_i - \mu$, $i = 1, \ldots, n$. D.w.z.

$$X_i = \mu + \epsilon_i, \ i = 1, \ldots, n.$$

Er is geen systematische fout in de meting, in die zin dat $E\epsilon_i = 0$ voor alle i. De nauwkeurigheid van de meting wordt weergegeven door de variantie van de meetfout σ^2. Als σ^2 groot is hebben we tamelijk onnauwkeurige metingen. Merk nu op dat

$$\bar{X} = \mu + \bar{\epsilon},$$

waarbij

$$\bar{\epsilon} = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i.$$

Dus $\bar{X}$ meet μ met meetfout $\bar{\epsilon}$. De onnauwkeurigheid is kleiner geworden dan die van de individuele metingen X_i, want $\text{var}(\bar{\epsilon}) = \sigma^2/n$. De onnauwkeurigheid gaat naar nul als we steeds meer metingen verrichten.

5.6. Een Chebyshev ongelijkheid. (Voor liefhebbers!) De volgende ongelijkheid geeft aan dat een stochastische grootheid niet veel van zijn verwachting kan afwijken als de variantie klein is.

Lemma. *Laat Z een stochastische grootheid zijn, dan geldt voor alle $c > 0$,*

$$P(|Z - EZ| > c) \leq \frac{\text{var}(Z)}{c^2}.$$

Bewijs.. We tonen het alleen aan voor een discrete s.g. Z. Als Z continu verdeeld is verloopt het bewijs analoog. Per definitie

$$\text{var}(Z) = \sum_j (z_j - EZ)^2 P(Z = z_j),$$

waarbij $\{z_j\}$ de mogelijke uitkomsten van Z zijn. We kunnen dit opsplitsen in twee delen:

$$\text{var}(Z) = \sum_{|z_j - EZ| \leq c} (z_j - EZ)^2 P(Z = z_j) + \sum_{|z_j - EZ| > c} (z_j - EZ)^2 P(Z = z_j).$$

Als we hier de eerste term weglaten wordt het resultaat hoogstens kleiner (want de termen zijn ≥ 0). Wat betreft de tweede term merken we op dat als $|z_j - EZ| > c$, dan $(z_j - EZ)^2 > c^2$, dus

$$\sum_{|z_j - EZ| > c} (z_j - EZ)^2 P(Z = z_j) \geq c^2 \sum_{|z_j - EZ| > c} P(Z = z_j).$$

Nu is $\sum_{|z_j - EZ| > c} P(Z = z_j)$ precies de kans dat $|Z - EZ| > c$. Zo vinden we

$$\text{var}(Z) \geq c^2 P(|Z - EZ| > c),$$

ofwel

$$P(|Z - EZ| > c) \leq \frac{\text{var}(Z)}{c^2}.$$

□

5.7. Speciaal geval: alternatieve verdeling. Veronderstel dat $X_1, \ldots, X_n$ een steekproef is uit de alternatieve verdeling met succeskans p: $P(X_i = 1) = 1 - P(X_i = 0) = p$, $i = 1, \ldots, n$. Merk op dat $X_+ = \sum_{i=1}^{n} X_i$ de binomiale verdeling bezit met parameters n en p. (In 3.7(3) en verder noemden we $\sum_{i=1}^{n} X_i = X$. Dat doen we hier niet, want X is in dit hoofdstuk de populatievariabele.) Het populatiegemiddelde is nu p (zie 4.6 (2)). Volgens de wet van de grote aantallen geldt daarom:

$$\lim_{n \to \infty} \bar{X}_n = p.$$

Dit zegt dat bij onafhankelijke, herhaalde experimenten de frequentie van een gebeurtenis met grote kans in de buurt van de kans op die gebeurtenis ligt, als tenminste het aantal experimenten maar groot genoeg is. Immers, beschouw een gebeurtenis A. Noem $n(A)$ het aantal keren dat die gebeurtenis optreedt, en $f_q(A) = n(A)/n$ de frequentie van gebeurtenis A. Dan $n(A) = X_+$, met $X_i = 1$ als in het i-de experiment gebeurtenis A optreedt, en $X_i = 0$ anders, $i = 1, \ldots, n$. De succeskans is dan $p = P(A)$. De wet van de grote aantallen zegt dus

$$f_q(A) \to P(A).$$

5.8. Hoe groot is de afwijking? Volgens de wet van de grote aantallen is $\bar{X} \approx \mu$, en de afwijking $|\bar{X} - \mu|$ is i.h.a. klein als $\bar{X}$ op veel experimenten gebaseerd is, d.w.z., als de steekproefgrootte n groot is. Men kan niet zeggen hoeveel de fout $|\bar{X} - \mu|$ precies bedraagt, doordat $\bar{X}$ een stochastische grootheid is, zodat de uitkomst onzeker is. Maar men kan wel een kansuitspraak over de afwijking doen. De centrale limietstelling geeft een benadering voor kansen van de vorm $P(|\bar{X} - \mu| > c)$. De stelling zegt dat $\bar{X}$ ongeveer normaal verdeeld is.

5.9. Standaardiseren We zullen $\bar{X}$ eerst standaardiseren (zie 4.23: we trekken de verwachting ervan af en delen door de standaarddeviatie) zodat het resultaat verwachting 0 en variantie 1 heeft. De centrale limietstelling beweert dat de gestandaardiseerde $\bar{X}$ ongeveer standaard normaal verdeeld is. Het steekproefgemiddelde $\bar{X}$ heeft verwachting μ en variantie σ^2/n (zie 5.3). De standaarddeviatie van $\bar{X}$ is dan $\sigma/\sqrt{n}$. Hieruit volgt dat $\sqrt{n}(\frac{\bar{X}-\mu}{\sigma})$ verwachting 0 en variantie 1 heeft.

5.10. De centrale limiet stelling. De centrale limiet stelling beweert dat de gestandaardiseerde $\bar{X}$ ongeveer standaard normaal verdeeld is: Voor alle x

$$\lim_{n\to\infty} P(\sqrt{n}(\frac{\bar{X}_n - \mu}{\sigma}) \leq x) = \Phi(x).$$

Hier is $\Phi(x)$ de standaard normale verdelingsfunctie (zie 3.8 (2)).

5.11. Andere schrijfwijze. Merk op dat

$$\sqrt{n}(\frac{\bar{X} - \mu}{\sigma}) = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}.$$

Nu heeft $\sum_{i=1}^n X_i$ verwachting $n\mu$ en variantie $n\sigma^2$, zodat de laatste uitdrukking gezien kan worden als de gestandaardiseerde van $\sum_{i=1}^n X_i$. Het maakt natuurlijk niet uit of men eerst het gemiddelde neemt en dan deze standaardiseert of dat men $\sum_{i=1}^n X_i$ rechtstreeks standaardiseert. Het er op neer dat $\bar{X}$ ongeveer $N(\mu, \sigma^2/n)$-verdeeld is, ofwel dat $\sum_{i=1}^n X_i$ ongeveer $N(n\mu, n\sigma^2)$-verdeeld is. We schrijven

$$P(\bar{X} \leq x) \approx \Phi\left(\sqrt{n}(\frac{x-\mu}{\sigma})\right),$$

en

$$P(\sum_{i=1}^n X_i \leq x) \approx \Phi\left(\frac{x - n\mu}{\sqrt{n}\sigma}\right).$$

5.12. Speciale gevallen. De verdeling van het steekproefgemiddelde kan dus altijd worden benaderd met de normale verdeling (als gemiddelde en variantie bestaan). In die zin *vergeet* het steekproefgemiddelde uit welke verdeling de oorspronkelijke steekproef is getrokken. We zeggen ook wel dat $\bar{X}$ *asymptotisch* normaal verdeeld is. Enkele speciale gevallen zijn:

(a) **Binomiale verdeling.** Laat $X_1, \ldots, X_n$ o.o. zijn met $p = P(X_i = 1) = 1 - P(X_i = 0)$. Dit is meestal een codering van het al of niet optreden van een bepaalde gebeurtenis A, en $f_q(A) = \bar{X}$ is dan de frequentie van A. Het theoretisch gemiddelde is p. De variantie is

$$\sigma^2 = p(1-p)$$

(zie 4.13 (2)). Dus $E(\bar{X}) = p$ en $\text{var}(\bar{X}) = p(1-p)/n$. Volgens de centrale limiet stelling is nu $\bar{X}$ ongeveer $N(p, p(1-p)/n)$-verdeeld. Merk weer op dat $X_+ = \sum_{i=1}^n X_i$. het aantal keren is dat gebeurtenis A optreedt. De s.g. X_+ is ongeveer $N(np, np(1-p))$-verdeeld, ofwel

$$\frac{X_+ - np}{\sqrt{np(1-p)}}$$

is ongeveer standaard normaal verdeeld. De exacte verdeling van X_+ is de binomiale verdeling met parameters n en p:

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

De binomiale verdeling kan dus worden benaderd door de normale verdeling. De centrale limietstelling zegt in dit geval dat

$$\sum_{k \leq x} \binom{n}{k} p^k (1-p)^{n-k} \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right).$$

Men zou kunnen proberen zoiets analytisch te bewijzen, maar dat ziet er niet eenvoudig uit.

(b) **Poisson verdeling.** Stel $X_1, \ldots, X_n$ zijn o.o. Poisson verdeeld met parameter μ. Dan is $X_+ = \sum_{i=1}^n X_i$ Poisson verdeeld met parameter $n\mu$. (zie 3.10 (b)). De variantie van X_+ is ook $n\mu$ (zie 4.13 (6)). Volgens de centrale limiet stelling is X_+ ongeveer $N(n\mu, n\mu)$-verdeeld. De Poisson verdeling kan dus worden benaderd door de normale verdeling (en de Poisson verdeling is ook weer een benadering van de binomiale verdeling, zie 3.7 (6)).

(c) **Normale verdeling.** In het geval van normaal verdeelde stochastische grootheden is $\bar{X}$ exact normaal verdeeld (zie 3.10 (c)).

5.13 Getallenvoorbeeld voor de binomiale verdeling. We nemen $n = 20$. Voor deze waarde van het aantal experimenten is de binomiale verdeling nog getabelleerd, omdat de benadering met de normale verdeling niet zo goed is. Laten we eens zien wat het verschil is voor $p = 0.40$ en $x = 5$. Uit een tabel halen we dat

$$P(X_+ \leq 5) = 0.1256.$$

Men kan dit narekenen:

$$\sum_{k=0}^{5} \binom{20}{k} (0.40)^k (0.60)^{20-k} = 0.1256.$$

Verder

$$\Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{5 - (20)(0.40)}{\sqrt{(20)(0.40)(0.60)}}\right)$$
$$= \Phi(-1.37) = 1 - \Phi(1.37) = 1 - 0.9147 = 0.0853.$$

Vergelijk deze uitkomst met het exacte resultaat 0.1256. De benadering is dus niet zo best.

5.14. Continuïteitscorrectie.. Als X_+ een binomiale verdeling met parameters n en p bezit, dan kan X_+ alleen de waarden $0, 1, \ldots, n$ aannemen. Het is een beter om bij de benadering van zo'n discrete s.g. met de continue normale verdeling een continuïteitscorrectie toe te passen, m.n. als n klein is. Deze correctie is:

$$P(X = x) \approx \Phi\left(\frac{x + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{x - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right),$$

voor $x \in \{0, 1, \ldots, n\}$. In woorden: $P(X_+ = x)$ benaderen we met de kans dat een $N(np, np(1-p))$-verdeelde s.g. in het interval $[x - \frac{1}{2}, x + \frac{1}{2}]$ ligt. De kans $P(X_+ \leq x)$ benaderen we dan met de kans dat een $N(np, np(1-p))$-verdeelde s.g. in het interval $(-\infty, x + \frac{1}{2}]$ ligt:

$$P(X_+ \leq x) \approx \Phi\left(\frac{x + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right),\ x \in \{0, 1, \ldots n\}.$$

5.15. Getallenvoorbeeld met continuïteitscorrectie. Neem weer $n = 20$ en $p = 0.40$. Dan

$$P(X_+ \leq 5) = 0.1256.$$

Gebruiken we de benadering met continuïteitscorrectie, dan vinden we

$$\Phi\left(\frac{5 + \frac{1}{2} - (20)(0.40)}{\sqrt{(20)(0.40)(0.60)}}\right) = \Phi(-1.14) = 1 - \Phi(1.14) = 1 - 0.8729 = 0.1271.$$

Dit is inderdaad een verbetering. Bekijk ook $P(X = 8) = 0.1797$. Ga na dat de benadering is $\Phi(0.22) - \Phi(-0.22) = 0.1820$.

5.16. Betrouwbaarheidsinterval. Stel we kiezen nu de waarde $c > 0$ zó dat $|\bar{X} - \mu|$ hoogstens c is met grote kans, zeg met 95 % kans. Om c exact te berekenen hebben moeten we de exacte verdeling van $\bar{X}$ weten. We kunnen ook een benadering gebruiken, als n voldoende groot is. Er geldt volgens de centrale limiet stelling

$$P(|\bar{X} - \mu| \leq c) = P(\bar{X} - \mu \leq c) - P(\bar{X} - \mu \leq -c)$$
$$= P\left(\sqrt{n}(\frac{\bar{X} - \mu}{\sigma}) \leq \sqrt{n}\frac{c}{\sigma}\right) - P\left(\sqrt{n}(\frac{\bar{X} - \mu}{\sigma}) \leq -\sqrt{n}\frac{c}{\sigma}\right)$$
$$\approx \Phi(\sqrt{n}\frac{c}{\sigma}) - \Phi(-\sqrt{n}\frac{c}{\sigma})$$
$$= 2\Phi(\sqrt{n}\frac{c}{\sigma}) - 1.$$

Nu is

$$\Phi(1.96) = 0.975,$$

dus

$$2\Phi(1.96) - 1 = 0.95.$$

We nemen daarom

$$\sqrt{n}\frac{c}{\sigma} = 1.96,$$

ofwel

$$c = (1.96)\frac{\sigma}{\sqrt{n}}.$$

We noemen nu

$$[\bar{X} - (1.96)\frac{\sigma}{\sqrt{n}}, \bar{X} + (1.96)\frac{\sigma}{\sqrt{n}}]$$

een asymptotisch 95 % betrouwbaarheidsinterval voor μ. D.w.z. met ongeveer 95 % kans is de afwijking tussen $\bar{X}$ en μ niet meer dan

$$(1.96)\frac{\sigma}{\sqrt{n}}.$$

5.17. Het schatten van de variantie. Bij de meeste statistische problemen is de verdeling waaruit de steekproef getrokken is onbekend. Dit betekent dat zowel μ als σ^2 onbekend zijn. We hebben nu een schatter van μ, n.l. het steekproefgemiddelde $\bar{X}$. We hebben ook een schatter van σ^2, namelijk de steekproefvariantie

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

(zie 4.11), en deze kunnen we gebruiken om te schatten wat de afwijking $|\bar{X} - \mu|$ ongeveer is. We vinden dan dat met ongeveer 95 % kans de afwijking tussen $\bar{X}$ en μ niet meer is dan

$$(1.96)\frac{S}{\sqrt{n}}.$$

5.18. De marge. We noemen $(1.96)\sigma/\sqrt{n}$ of de geschatte variant $(1.96)S/\sqrt{n}$ ook wel de *marge*. Merk op dat het getal 1.96 volgt uit het feit dat we een kans van 5 % dat de afwijking tussen $\bar{X}$ en μ toch buiten deze marge valt nog acceptabel achten. Als men deze kans wil verlagen tot bijvoorbeeld 1 % wordt de marge groter. Verder is de marge gebaseerd op een benadering (de centrale limiet stelling). Als de steekproefgrootte n klein is, kan men aan de conservatieve kant gaan zitten door de marge groter te kiezen. Met name kan men dan de normale verdeling vervangen door de z.g. Student verdeling met $n - 1$ vrijheidsgraden. (Hier gaan we in dit college niet verder op in.)

5.19. Vuistregel. Als vuistregel kan men (in woorden) hanteren: schatter en geschatte waarde verschillen niet meer dan 2× de (geschatte) standaarddeviatie van de schatter. Hier 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 $\bar{X}$ is $\sigma/\sqrt{n}$.

6. Schattingstheorie.

6.1. Steekproef. We beschouwen een rij $X_1, \ldots, X_n$ van o.o. s.g.$^{\underline{n}}$ met dezelfde verdeling:

$$P(X_1 \leq x) = P(X_2 \leq x) = \ldots = P(X_n \leq x) = F(x), \text{ voor alle } x.$$

We zeggen dan dat $X_1, \ldots, X_n$ o.o. en *identiek* verdeeld zijn, en we noemen $X_1, \ldots, X_n$ een steekproef uit de verdeling F, ofwel n o.o. copietjes van de populatiegrootheid X. Een realisatie van $(X_1, \ldots, X_n)$ noteren we met $(x_1, \ldots, x_n)$. Dit zijn de getallen die we hebben waargenomen nadat de steekproef daadwerkelijk is uitgevoerd.

6.2. Onbekende parameters. De verdelingsfunctie $F(x)$ is geheel of gedeeltelijk onbekend. We nemen vaak iets aan over de vorm van $F(x)$. Dit is soms voor het wiskundig gemak, maar het kan ook b.v. zijn dat we het waardebereik van X kennen, of iets anders over de verdeling van de X. Als b.v. $X \in \{0, 1\}$, dan bezit X een alternatieve verdeling. De succeskans $p = P(X = 1)$ zullen we in het algemeen niet kennen, en we zeggen dan dat X een alternatieve verdeling met *onbekende* parameter p bezit. Een ander voorbeeld is dat we op grond van een redenatie als in 3.8 (3) veronderstellen dat X een exponentiële verdeling bezit met onbekende parameter λ. Ook wordt vaak aangenomen dat X normaal verdeeld is met onbekende parameters μ en σ^2.

6.3. Schatter. Een *schatter* $T = t(X_1, \ldots, X_n)$ is een functie van de waarnemingen $X_1, \ldots, X_n$ die niet afhangt van onbekende parameters. De reden waarom we eisen dat de functie T niet van onbekende grootheden mag afhangen, is dat we T in praktijk moeten kunnen uitrekenen. D.w.z. als we waarnemingen $X_1, \ldots, X_n$ hebben, dan is T ook bekend.

6.4. Schatting. Bij realisaties $x_1, \ldots, x_n$ noemen we de realisatie $t = t(x_1, \ldots, x_n)$ een *schatting*. Een schatting is dus een realisatie van een schatter.

6.5. Verwachting en variantie. De (theoretische) verwachting van X geven we aan met

$$\mu = EX$$

en de (theoretische) variantie met

$$\sigma^2 = \text{var}(X)$$

(waarbij we er van uit gaan dat deze bestaan).

6.6. Schatter van de verwachting. Het steekproefgemiddelde

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

kan men opvatten als schatter van μ. Volgens de wet van de grote aantallen is $\bar{X}$ ongeveer gelijk aan μ voor n groot. Er is geen systematische fout in deze schatter, in die zin dat

$$E\bar{X} = \mu$$

(zie 4.5). We noemen daarom $\bar{X}$ ook wel een *zuivere* schatter van μ.

6.7. Schatter van de variantie. De steekproefvariantie

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

kan men opvatten als schatter van σ^2, want, alweer volgens de wet van de grote aantallen, S^2 is ongeveer gelijk aan σ^2 voor n groot. Immers, omdat

$$\bar{X} \approx \mu$$

is

$$S^2 \approx \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \mu)^2$$

$$\approx \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2.$$

Maar nu staat er het steekproefgemiddelde van de variabele $Y=(X-\mu)^2$, en Y heeft theoretisch gemiddelde

$$EY = E(X-\mu)^2 = \sigma^2.$$

Nu is de vraag: waarom deelt men door $n-1$ i.p.v. door n? De reden is dat je er zo voor zorgt dat S^2 geen systematische fout heeft, d.w.z.

$$ES^2 = \sigma^2$$

(zie 6.8 hieronder). We noemen S^2 dan ook een *zuivere* schatter van σ^2.

6.8. Het bewijs dat S^2 een zuivere schatter is van σ^2. Er geldt (vergelijk met 4.9)

$$\sum_{i=1}^{n}(X_i-\bar{X})^2 = \sum_{i=1}^{n}(X_i^2-2\bar{X}X_i+\bar{X}^2) = \sum_{i=1}^{n}X_i^2 - 2\bar{X}\sum_{i=1}^{n}X_i + \sum_{i=1}^{n}\bar{X}^2$$

$$= \sum_{i=1}^{n}X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 = \sum_{i=1}^{n}X_i^2 - n\bar{X}^2.$$

Nu is $\sigma^2 = EX_i^2 - \mu^2$ (zie 4.9), dus $EX_i^2 = \sigma^2+\mu^2$, $i=1,\ldots,n$. Analoog,

$$\frac{\sigma^2}{n} = \mathrm{var}(\bar{X}) = E\bar{X}^2 - \mu^2$$

dus $E\bar{X}^2 = \sigma^2/n + \mu^2$. We vinden zo

$$ES^2 = \frac{1}{n-1}(\sum_{i=1}^{n}EX_i^2 - nE\bar{X}^2) = \frac{1}{n-1}(\sum_{i=1}^{n}(\sigma^2+\mu^2) - n(\sigma^2/n+\mu^2)) = \sigma^2.$$

Bij dit soort berekeningen is de volgende truuk ook handig:

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

Nu staan de s.g.$\underline{n}$ in afwijking van de verwachting. Daarom mag je hier zonder verlies van algemeenheid veronderstellen dat ze verwachting nul hebben. Dit verklaart ook waarom in bovenstaand bewijs dat S^2 zuiver is, de μ's tegen elkaar wegvallen.

6.9. Schatter van de verdelingsfunctie. De *empirische* verdelingsfunctie

$$F_n(x) = \frac{1}{n}\{\text{aantal } X_i \le x,\ 1\le i\le n\}$$

kan men opvatten als schatter van de (theoretische) verdelingsfunctie $F(x)$. Volgens de wet van de grote aantallen is $F_n(x)$ ongeveer gelijk aan $F(x)$ voor n groot. Bovendien is $F_n(x)$ een zuivere schatter:

$$EF_n(x) = F(x).$$

Immers, $F_n(x)$ is het steekproefgemiddelde van de alternatief verdeelde variabele

$$Y = \begin{cases} 1, & \text{als } X\le x \\ 0, & \text{als } X > x\ . \end{cases}$$

De succeskans is in dit geval is

$$P(Y=1) = P(X\le x) = F(x),$$

dus $EY = F(x)$.

6.10. Notatie. Laat nu $\theta \in \mathbf{R}$ een onbekende parameter zijn (bijvoorbeeld de verwachting μ in de normale verdeling, de parameter λ in de exponentiële verdeling, de succeskans p in de alternatieve verdeling). Een schatter van θ hangt als functie van de waarnemingen niet van onbekende grootheden af, maar de *verdeling* van een schatter hangt meestal wel van onbekende grootheden af. Dit komt doordat de verdeling van de waarnemingen zelf van onbekende grootheden afhangt, m.n. van de onbekende parameter θ. In het bijzonder hangt de verwachting en variantie van een schatter T af van θ. dit te benadrukken schrijven we soms $E_\theta T$ voor de verwachting van T als de parameterwaarde θ is, en analoog: $\mathrm{var}_\theta(T)$, $P_\theta(T > 2)$, etc..

6.11. Zuivere schatters. De *onzuiverheid* van een schatter T is

$$\mathrm{bias}_\theta(T) = E_\theta(T) - \theta$$

(bias is het engelse woord voor onzuiverheid). We noemen een schatter T van θ *zuiver* (Engels: *unbiased*) als

$$E_\theta T = \theta,$$

voor alle mogelijke waarden van θ.

Voorbeeld. We hebben gezien dat S^2 een zuivere schatter is van σ^2. De variant

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

die door n deelt i.p.v. door $n-1$ is géén zuivere schatter. Als schatter voor de standaarddeviatie σ gebruikt men meestal $S = \sqrt{S^2}$. Deze is echter niet zuiver, want aangezien $\mathrm{var}(S) = ES^2 - (ES)^2 > 0$, is $(ES)^2 < ES^2 = \sigma^2$ ofwel $ES < \sigma$.

6.12. Verwachte kwadratische fout. We zoeken schatters die volgens een bepaald criterium *goed* zijn. Bijvoorbeeld, het is prettig als een schatter zuiver is, want dan bezit deze geen systematische fout. Een ander criterium is de verwachte kwadratische fout,

$$\mathrm{MSE}_\theta(T) = E_\theta(T - \theta)^2$$

(MSE komt van het engelse begrip Mean Square Error).

6.13. Relatie tussen verwachte kwadratische fout en onzuiverheid. Een schatter moet bij voorkeur een kleine MSE hebben, en we willen ook graag dat een schatter zuiver is, of op z'n minst kleine onzuiverheid heeft. De relatie tussen de MSE en de bias is gegeven in het volgende

Lemma. $\mathrm{MSE}_\theta(T) = \mathrm{var}_\theta(T) + \mathrm{bias}_\theta^2(T)$.

Bewijs.

$$\begin{aligned}
\mathrm{MSE}_\theta(T) &= E_\theta(T-\theta)^2 \\
&= E_\theta\left((T - E_\theta T) + (E_\theta T - \theta)\right)^2 \\
&= E_\theta\left(T - E_\theta T\right)^2 + \left(E_\theta T - \theta\right)^2 + 2E_\theta\left(T - E_\theta T\right)\left(E_\theta T - \theta\right) \\
&= \mathrm{var}_\theta(T) + \mathrm{bias}_\theta^2(T) + 0.
\end{aligned}$$

□

Voor een zuivere schatter T is $\mathrm{MSE}_\theta(T) = \mathrm{var}_\theta(T)$. M.a.w. een zuivere schatter is *goed* als deze kleine variantie heeft. Een onzuivere schatter met kleine variantie kan onbruikbaar zijn, want zo'n schatter concentreert zich rond het verkeerde punt. Soms moet men een afweging maken: aan de éne kant wil men graag een zuivere schatter hebben en aan de andere kant wil men ook de variantie klein houden. Dit kunnen strijdige belangen zijn. Vaak houdt men vast aan de eis dat een schatter zuiver moet zijn, en zoekt men onder alle zuivere schatters diegene met de kleinste variantie. Dit kan problemen geven, bijvoorbeeld een zuivere schatter bestaat niet altijd.

6.14. Een schatter van een dichtheid. Stel dat X continu verdeeld met dichtheid $f(x)$. Per definitie is

$$f(x) = \lim_{h\to 0}\frac{F(x+h) - F(x)}{h},$$

waarbij F weer de verdelingsfunctie is. Voor het schatten van $f(x)$ heeft het weinig zin om hier F door de empirische verdelingsfunctie F_n te vervangen, want F_n is niet differentieerbaar. Wat men wel kan doen is een vaste h kiezen en $f(x)$ schatten met

$$\hat{f}(x) = \frac{F_n(x+h) - F_n(x)}{h}.$$

We noemen h de *bandbreedte*. Er geldt

$$E\hat{f}(x) = \frac{F(x+h) - F(x)}{h} \approx f(x)$$

voor h klein. Dus $\hat{f}(x)$ is niet zuiver, maar de bias is klein als h klein is. Merk op dat

$$F_n(x+h) - F_n(x) = \frac{1}{n}\{\text{aantal der } X_i \text{ met } x < X_i \leq x+h,\ 1 \leq i \leq n\}.$$

Dit is het steekproefgemiddelde van een steekproef uit een alternatieve verdeling met parameter $F(x+h) - F(x)$. Hieruit volgt dat

$$\text{var}(\hat{f}(x)) = \frac{(F(x+h) - F(x))(1 - (F(x+h) - F(x))}{h^2 n} \approx \frac{f(x)}{hn}.$$

De variantie is klein als h groot is. Om een redelijke MSE te krijgen moet men h daarom niet al te groot kiezen om de bias in de hand te houden, en niet al te klein om de variantie in de hand te houden. Voor de keuze van de bandbreedte h is een algemene theorie opgebouwd, waar we in dit college niet op ingaan.

Bovenstaand idee kan worden uitgebreid zodat je een schatter van $f(x)$ voor *alle* waarden van x krijgt, namelijk het *histogram*. Het waardebereik van de waarnemingen wordt verdeeld in intervalletjes van lengte h en met eindpunten $a_0, a_1, \ldots, a_T$ (dus $a_i = a_{i-1} + h$). Voor $x \in (a_{i-1}, a_i]$ schat men $f(x)$ met

$$\hat{f}(x) = \frac{F_n(a_i) - F_n(a_{i-1})}{h}.$$

6.15. Meest-aannemelijke schatters: voorbeeld. Het idee in deze paragraaf is die waarde als schatter van θ te kiezen, waarvoor de gevonden waarnemingen het meest aannemelijk zijn.

Voorbeeld.

$$X_i = \begin{cases} 1, & \text{als het computerprogramma bij gegevensinvoer } i \text{ goed werkt,} \\ 0, & \text{anders .} \end{cases}$$

Stel dat $p = P(X_i = 1)$ de onbekende succeskans is, waarbij we veronderstellen dat p voor alle n soorten gegevensinvoer hetzelfde is. We hebben de realisatie

$$(x_1, \ldots, x_{10}) = (1, 1, 1, 0, 1, 1, 1, 1, 1, 0)$$

waargenomen. Dus 8 van de 10 keer heeft het programma succesvol gedraaid. Op grond van deze waarneming is het aannemelijk dat p niet al te klein is, want anders zouden we i.h.a. wel meer mislukkingen hebben gevonden. De kans op $(1, 1, 1, 0, 1, 1, 1, 1, 1, 0)$ is

$$L(p) := p^8(1-p)^2$$

We noemen $L(p)$ de *aannemelijkheid* van de waarneming $(1, 1, 1, 0, 1, 1, 1, 1, 1, 0)$. Voor welke waarde van p is de aannemelijheid nu het grootst? We zoeken dan het maximum van $L(p)$. Noem die waarde $\hat{p}$ die $L(p)$ maximaliseert de *meest aannemelijke schatting*.

Maximaliseren van $L(p)$ kan hier d.m.v. de afgeleide nemen en die gelijk aan nul te stellen:

$$\frac{d}{dp}L(p) = \frac{d}{dp}(p^8 - 2p^9 + p^{10}) = 8p^7 - 18p^8 + 10p^9.$$

$$\frac{d}{dp}L(p) = 0 \;\Leftrightarrow\; 8p^7 - 18p^8 + 10p^9 = 0$$

$$\Leftrightarrow\; p = 0 \;\vee\; 8 - 18p + 10p^2 = 0$$

$$\Leftrightarrow\; p = 0 \;\vee\; p = 0.8 \;\vee\; p = 1,$$

waarbij we in de laatste stap gebruik maakten van de abc-formule. Nu zijn $p = 0 \;\vee\; p = 1$ minima van $L(p)$ en $p = 0.8$ is het maximum. De meest aannemelijke schatting in dit geval is daarom

$$\hat{p} = 0.8.$$

Voorbeeld.(Algemener) Stel $X_1, \ldots, X_n$ zij o.o. s.g.$\underline{n}$ met $P(X_i = 1) = 1 - P(X_i = 0) = p$ onbekend. Als $(x_1, \ldots, x_n)$ een realisatie is van $(X_1, \ldots, X_n)$, dan is

$$L(p) = P(X_1 = x_x, \ldots, X_n = x_n) = p^{(\sum_{i=1}^n x_i)}(1-p)^{(n-\sum_{i=1}^n x_i)}$$

de aannemelijkheidsfunctie. Het maximum van $L(p)$ kan gevonden worden door de afgeleide gelijk aan nul te stellen en van de oplossingen na te gaan welke het maximum is. Maar het is handiger om $\log L(p)$ te maximaliseren. Dit is hetzelfde als $L(p)$ maximaliseren omdat het nemen van de logaritme een strict stijgende transformatie is. Met "log" bedoelen we de natuurlijke logaritme (men mag ook de logaritme met een ander grondgetal kiezen, maar de natuurlijke logaritme blijkt vaak het gemakkelijkst te zijn). We hebben

$$\log L(p) = \sum_{i=1}^n x_i \log(p) + (n - \sum_{i=1}^n x_i) \log(1-p).$$

Zoek het maximum van $\log L(p)$:

$$\frac{d}{dp} \log L(p)|_{p=\hat{p}} = (\frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p})|_{p=\hat{p}} = 0$$

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

De meest aannemelijke schatting is dus $\hat{p} = \bar{x}$.

Merk op dat $\hat{p}$ van de uitkomsten $x_1, \ldots, x_n$ afhangt. Men kan dit aangeven met $\hat{p} = \hat{p}(x_1, \ldots, x_n)$. We gebruiken dezelfde notatie $\hat{p} = \hat{p}(X_1, \ldots, X_n)$ en noemen de laatste de meest aannemelijke *schatter*. De meest aannemelijke *schatting* is m.a.w. een realisatie van de meest aannemelijke schatter.

6.16. Definitie meest-aannemelijke schatter. Laat $X_1, \ldots, X_n$ een steekproef zijn uit X, waarbij de verdeling van X bekend is op een parameter θ na. Laat $x_1, \ldots, x_n$ de waargenomen waarden zijn. Voor de definitie van de *aannemelijkheidsfunctie* (Engels: *likelihood function*) bekijken we twee gevallen. Als X een discrete verdeling bezit, is de aannemelijkheidsfunctie

$$L(\theta) = P_\theta(X_1 = x_1) \ldots P_\theta(X_n = x_n),$$

Als X een continue verdeling bezit met dichtheid f_θ, dan is de aannemelijkheidsfunctie

$$L(\theta) = f_\theta(x_1) \ldots f_\theta(x_n).$$

De *meest aannemelijke schatting* $\hat{\theta}$ is gedefinieerd door:

$$L(\hat{\theta}) = \max_\theta L(\theta),$$

waarbij gemaximaliseerd wordt over alle mogelijke waarden van θ. De meest-aannemelijke schatting $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ is een realisatie van de meest-aanemelijke schatter $\hat{\theta}(X_1, \ldots, X_n)$. We gebruiken voor schatter en schatting dezelfde notatie $\hat{\theta}$.

6.17. Aanwijzingen voor het berekenen van de meest-aanemelijke schatter.

(i) Het maximum van $L(\theta)$ kan vaak (maar niet altijd) gevonden worden door de afgeleiden naar θ gelijk aan nul te stellen.

(ii) Het is handig om $\log L(\theta)$ te maximaliseren, i.p.v. $L(\theta)$. Bijvoorbeeld, in het discrete geval

$$\log L(\theta) = \sum_{i=1}^{n} \log P_\theta(X_i = x_i).$$

Het is makkelijker om de *som* van een aantal termen te differentiëren, i.p.v. het *produkt* van een aantal termen.

6.18. Meest-aanemelijke schatters in enkele voorbeelden.

Voorbeeld (i). Stel X bezit de geometrsche verdeling met succeskans θ:

$$P_\theta(X = x) = (1-\theta)\theta^x,\ x = 0, 1, 2, \ldots,\ 0 < \theta < 1.$$

Dan

$$\log P_\theta(X_i = x_i) = \log(1-\theta) + x_i \log\theta,$$

en dus

$$\log L(\theta) = n\log(1-\theta) + \sum_{i=1}^{n} x_i \log(\theta).$$

De afgeleide naar θ is nu

$$\frac{d}{d\theta}\log L(\theta)|_{\theta=\hat\theta} = (-\frac{n}{1-\theta} + \frac{\sum_{i=1}^n x_i}{\theta})|_{\theta=\hat\theta} = 0$$

$$\Rightarrow (1-\hat\theta)\sum_{i=1}^{n} x_i - n\hat\theta = 0 \ \Rightarrow\ \sum_{i=1}^{n} x_i - (\sum_{i=1}^{n} x_i + n)\hat\theta = 0$$

$$\Rightarrow \hat\theta = \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i + n} = \frac{\bar x}{\bar x + 1}.$$

De meest aannemelijke schatter is daarom

$$\hat\theta = \frac{\bar X}{\bar X + 1}.$$

Voorbeeld (ii). Stel X is homogeen verdeeld op $[0, \theta]$:

$$f_\theta(x) = \frac{1}{\theta},\ 0 \le x \le \theta.$$

Dan

$$L(\theta) = (\frac{1}{\theta})^n,\ 0 \le \min_{1\le i\le n} x_i \le \max_{1\le i\le n} x_i \le \theta.$$

Deze is niet differentieerbaar, maar we zien dat het maximum ligt bij $\hat\theta = \max(x_1, \ldots, x_n)$. Dus $\hat\theta = \max(X_1, \ldots, X_n)$ is de meest aannemelijke schatter.

Voorbeeld (iii). Laat X $N(\mu, \sigma^2)$-verdeeld zijn. De twee parameters μ en σ^2 veronderstellen we beide onbekend, dus $\theta = (\mu, \sigma^2)$ is nu een twee-dimensionale onbekende parameter. De dichtheid is

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp[-(1/2)(\frac{x-\mu}{\sigma})^2].$$

De log-aanemelijkheidsfunctie is

$$\log L(\mu, \sigma^2) = -n\log(\sqrt{2\pi}) - (n/2)\log(\sigma^2) - (1/2)\frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma^2}.$$

Door de afgeleide naar μ gelijk aan nul te stellen, vind je

$$\frac{d}{d\mu} logL(\mu, \sigma^2)|_{\mu=\hat{\mu}} = \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2} = 0,$$

dus $\hat{\mu} = \bar{x}$. Verder:

$$\frac{d}{d\sigma^2} \log L(\mu, \sigma^2)|_{\mu=\hat{\mu},\ \sigma^2=\hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^2} + (1/2)\frac{\sum_{i=1}^n (x_i - \hat{\mu})^2}{\hat{\sigma}^4} = 0$$

$$\Rightarrow\ \hat{\sigma}^2 = (1/n)\sum_{i=1}^n (x_i - \hat{\mu})^2$$

De meest aannemelijke schatters zijn dus $\hat{\mu} = \bar{X}$ en $\hat{\sigma}^2 = (1/n)\sum_{i=1}^n (X_i - \bar{X})^2$.

6.19. Kleinste-kwadratenschatters. In voorbeeld 6.19. (iii) hebben we gezien dat het steekproefgemiddelde de kwadratensom

$$\sum_{i=1}^n (X_i - \mu)^2$$

minimaliseert naar μ. We noemen het steekproefgemiddelde daarom ook wel de *kleinste-kwadratenschatter* van de verwachting. We gaan nu een algemenere situatie bekijken. Laat $Y_1, \ldots, Y_n$ o.o. stochastische grootheden zijn, maar nu niet noodzakelijk met allemaal dezelfde verdeling. Laat $x_1, \ldots, x_n$ nu gegeven getallen zijn, niet noodezakelijk een realisatie van een steekproef $X_1, \ldots, X_n$. We nemen aan dat

$$EY_i = g_\theta(x_i),$$

met g_θ een functie die afhangt van een onbekende parameter θ. De kleinste-kwadratenschatter $\hat{\theta}$ minimaliseert de kwadratensom

$$\sum_{i=1}^n (Y_i - g_\theta(x_i))^2.$$

Voorbeeld. We beschikken over n datasets met omvang respectievelijk $x_1, \ldots, x_n$. De datasets worden d.m.v. een computerprogramma gecontroleerd op coderingsfouten. Laat y_i de executietijd van het controleprogramma zijn, bij dataset i van omvang x_i, $i = 1, \ldots, n$. We willen nu het verband tussen de omvang van een dataset en de executietijd onderzoeken. Het idee is dat we gegeven een dataset van omvang x_i, de executietijd niet precies kunnen voorspellen. D.w.z. y_i is een realisatie van een stochastische grootheid Y_i. We veronderstellen, dat gegeven x_i, we iets over de verwachte waarde van Y_i kunnen zeggen:

$$EY_i = g_\theta(x_i),$$

waarbij $g_\theta(x)$ één of andere functie is, die afhangt vcan een onbekende parameter θ.

6.20. Definitie lineaire regressiemodel. Laat $Y_1, \ldots, Y_n$ o.o. waarnemingen zijn, met

$$EY_i = \alpha + \beta x_i,\ i = 1, \ldots, n,$$

waarbij $x_1, \ldots, x_n$ gegeven getallen, en α en β onbekende parameters.

Voorbeeld. Laat x_i de druk zijn waaraan plastic buis i bloot staat, en Y_i de levensduur van deze buis. Stel men weet dat de verwachte levensduur van een plastic buis, op een constante na, omgekeerd evenredig is met de druk. In formule:

$$EY_i = \alpha + \frac{\beta}{x_i},\ i = 1, \ldots, n.$$

Door over te gaan op een nieuwe x-variabele kan je dit in de vorm van een lineair model gieten. Noem n.l. $\tilde{x}_i = 1/x_i$. Dan $EY_i = \alpha + \beta\tilde{x}_i$, $i = 1, \ldots, n$.

6.21. Meetfout. De *meetfouten* in het (lineaire) regressiemodel zijn de variabelen

$$\epsilon_i = Y_i - EY_i,\ i = 1, \ldots, n.$$

Er geldt dus

$$Y_i = \alpha + \beta x_i + \epsilon_i,\ i = 1, \ldots, n,$$

met $\epsilon_1, \ldots, \epsilon_n$ o.o. meetfouten met verwachting nul. Meestal neemt men ook aan dat $\text{var}(\epsilon_i)$ constant is voor alle i, zeg $\text{var}(\epsilon_i) = \sigma^2$ (met σ^2 i.h.a. onbekend). Dit zegt dat de nauwkeurigheid van de meting niet van i afhangt. Verder geldt dan ook $\text{var}(Y_i) = \sigma^2$ voor alle i.

6.22. De verklarende variabele. De variabele x_i in het regressiemodel noemt men wel de verklarende variabele, en Y_i de te verklaren variabele. Het kan zijn dat x_i niet instelbaar is maar daarentegen een realisatie van een stochastische grootheid X_i. Dit maakt in principe niets uit voor het regressiemodel, zolang men maar aanneemt dat X_i onafhankelijk van de meetfout ϵ_i is. De modelaannamen moeten zodanig zijn, dat het regressiemodel geldt, *voorwaardelijk* op de waargenomen waarden $x_1, \ldots, x_n$.

6.23. De kleinste kwadratenschatters in het lineaire model. Het idee is nu om die lijn $l(x)$ te zoeken die "het best past" bij de waarnemingen (puntenwolk) (x_i, Y_i), $i = 1, \ldots, n$. We hanteren daarbij het criterium $\sum_{i=1}^n (Y_i - l(x_i))^2$: dit moet voor zekere lijn $l(x)$ zo klein mogelijk zijn. Noem

$$\mathbf{S}(\alpha, \beta) = \sum_{i=1}^n (Y_i - \alpha - \beta x_i)^2.$$

De *kleinste-kwadratenschatters* (KK-schatters) $\hat{\alpha}$ en $\hat{\beta}$ zijn gedefinieerd door

$$\mathbf{S}(\hat{\alpha}, \hat{\beta}) = \min_{\alpha,\beta} \mathbf{S}(\alpha, \beta).$$

Een realisatie van $(\hat{\alpha}, \hat{\beta})$ noemt men een kleinste-kwadratenschatting.

6.24. Uitdrukking voor de kleinste-kwadratenschatters. We schrijven weer $\bar{x} = \sum_{i=1}^n x_i/n$ en $\bar{Y} = \sum_{i=1}^n Y_i/n$ ($\bar{y} = \sum_{i=1}^n y_i/n$).

Lemma.

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x},$$

en

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Bewijs.

$$\frac{d}{d\alpha}\mathbf{S}(\alpha, \beta) = -2\sum_{i=1}^n (Y_i - \alpha - \beta x_i)$$

$$\Rightarrow\ \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

$$\Rightarrow\ \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}.$$

$$\frac{d}{d\beta}\mathbf{S}(\alpha, \beta) = -2\sum_{i=1}^n (Y_i - \alpha - \beta x_i) x_i = -2(\sum_{i=1}^n Y_i x_i - \alpha \sum_{i=1}^n x_i - \beta \sum_{i=1}^n x_i^2)$$

$$\Rightarrow\ \sum_{i=1}^n Y_i x_i - \hat{\alpha} \sum_{i=1}^n x_i - \hat{\beta} \sum_{i=1}^n x_i^2 = 0$$

$$\Rightarrow\ \sum_{i=1}^n Y_i x_i - (\bar{Y} - \hat{\beta}\bar{x}) \sum_{i=1}^n x_i - \hat{\beta} \sum_{i=1}^n x_i^2 = 0$$

$$\Rightarrow\ \hat{\beta}\sum_{i=1}^{n}(x_i^2 - x_i\bar{x}) = \sum_{i=1}^{n}(Y_i x_i - \bar{Y}x_i)$$

$$\Rightarrow\ \hat{\beta} = \frac{\sum_{i=1}^{n} x_i(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})x_i} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

□

6.25. Klein getallenvoorbeeldje. Stel $n = 3$, $(x_1, x_2, x_3) = (1, 2, 3)$ en $(y_1, y_2, y_3) = (2, 1, 3)$. Dan $\bar{x} = 2$, $\bar{y} = 2$, $\sum_{i=1}^{3}(x_i - \bar{x})^2 = 2$ en $\sum_{i=1}^{3}(x_i - \bar{x})y_i = 1$. Dus $\hat{\beta} = 1/2$ en $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 1$.

6.26. Simulatie. Om het gedrag van de kleinste-kwadratenschatters in een simulatiestudie te onderzoeken, kiezen we een steekproefgrootte n, een zekere α en β, de waarden voor $x_1, \ldots, x_n$ en een verdeling voor de meetfouten $\epsilon_1, \ldots, \epsilon_n$. Omdat het nu een simulatie betreft, zijn α en β wèl bekend, en kunnen we dus controleren of $\hat{\alpha}$ en $\hat{\beta}$ in de buurt van α en β liggen.

```
> n<-100
> x<-1:n/n
>#  we nemen alpha = 2 en beta = 3
> lx<-2+3*x
> e<-rnorm(n)
> y<-lx+e
> plot(x,y)
> sxy<-sum((x-mean(x))*(y-mean(y)))/(n-1)
> s2x<-var(x)
> hatbeta<-sxy/s2x
> hatbeta
[1] 2.904354
> hatalpha<-mean(y) - hatbeta*mean(x)
> hatalpha
[1] 2.11699
> hatlx<-hatalpha + hatbeta*x
> lines(x,hatlx)
> lines(x,lx)
```

6.27. Steekproefequivalenten. De kleinste-kwadratenschatters $\hat{\alpha}$ en $\hat{\beta}$ zijn in feite *steekproefequivalenten* van hun theoretische tegenhangers α en β. Om dit te verduidelijken nemen we aan dat $x_1, \ldots, x_n$ realisaties zijn van een steekproef $X_1, \ldots, X_n$ uit een stochastische grootheid X. Veronderstel, als in 4.18, het model

$$Y = \alpha + \beta X + \epsilon,$$

met ϵ onafhankelijk van X, en $E\epsilon = 0$. Dan is

$$\alpha = EY - \beta EX,$$

en (zie 4.18)

$$\beta = \mathrm{cov}(X, Y)/\mathrm{var}(X).$$

De kleinste-kwadratenschatters zijn

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X},$$

en

$$\hat{\beta} = S_{XY}/S_X^2,$$

met S_{XY} de steekproefcovariantie, en S_X^2 de steekproefvariantie van X.

6.28. Zuiverheid. De schatters $\hat{\alpha}$ *en* $\hat{\beta}$ zijn zuivere schatters van α resp. β. Immers

$$E(\hat{\beta}) = E\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) = \frac{E(\sum_{i=1}^{n}(x_i - \bar{x})Y_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})E(Y_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(\alpha + \beta x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= \frac{\alpha \sum_{i=1}^n (x_i - \bar{x}) + \beta \sum_{i=1}^n (x_i - \bar{x})x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta,$$

en

$$E(\hat{\alpha}) = E(\bar{Y} - \hat{\beta}\bar{x}) = E(\bar{Y}) - E(\hat{\beta})\bar{x}$$

$$= \alpha + \beta\bar{x} - \beta\bar{x} = \alpha.$$

6.29. Meest-aannemelijke schatters. Stel dat $\epsilon_1, \ldots, \epsilon_n$ o.o. $N(0, \sigma^2)$-verdeeld zijn, dan zijn de KK-schatters $\hat{\alpha}$ en $\hat{\beta}$ ook de meest aannemelijke schatters. Hierbij gebruiken we een uitbreiding van de definitie van meest-aannemelijke schatters, naar het geval van o.o. maar niet identiek verdeelde stochastische grootheden (de Y_i hebben immers niet alle dezelfde verwachting). Als de meetfouten alle normaal verdeeld zijn met verwachting nul en variantie σ^2, dan zijn de waarnemingen Y_i $N(\alpha+\beta x_i, \sigma^2)$-verdeeld. De dichtheid van Y_i is dus

$$f(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2].$$

De aannemelijkheidsfunctie van $Y_1, \ldots, Y_n$ is nu

$$L(\theta) = (\frac{1}{\sqrt{2\pi}\sigma})^n \exp[-\frac{1}{2\sigma^2}(y_1 - \alpha - \beta x_1)^2] \ldots \exp[-\frac{1}{2\sigma^2}(y_n - \alpha - \beta x_n)^2]$$

$$= (\frac{1}{\sqrt{2\pi}\sigma})^n \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2] = (\frac{1}{\sqrt{2\pi}\sigma})^n \exp[-\frac{1}{2\sigma^2}\mathbf{S}(\alpha, \beta)].$$

De meest aannemelijke schattingen voor α en β vind je door deze uitdrukking te maximaliseren. Dit is hetzelfde als $\mathbf{S}(\alpha, \beta)$ minimaliseren.

Opgaven Stochastiek voor Informatici

Voorjaar 2000

(Tussen haakjes staat een verwijzing naar een paragraaf.)

1. (1.1, 1.4) 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.

2. (1.10) Hoevaak moet men gemiddeld met een dobbelsteen gooien, totdat men alle aantallen ogen gehad heeft? (Gebruik de computer en simuleer!)

3. (2.14(4)) Welke kans is groter: met vier dobbelstenen in één worp minstens één zes, of met twee dobbelstenen in 24 worpen minstens één dubbelzes? (Probleem van Chevalier de Méré.)

4. (2.7) In ieder van drie kastjes zitten twee laden; in elke la zit één munt. Het eerste kastje bevat twee zilveren munten, het tweede een zilveren en een gouden munt; het derde kasje bevat twee gouden munten. Men trekt aselect een la open en vindt een gouden munt. Met welke kans vindt men in de andere la van dit kasje een zilveren munt?

5. Bij een spelshow wordt de kandidaat verzocht te kiezen uit drie deurtjes. Achter één van de deurtjes staat de hoofdprijs. Nadat de kandidaat een deurtje gekozen heeft, loopt de spelleider naar een van de andere deurtjes en doet deze open. De prijs staat niet achter het deurtje dat de spelleider geopend heeft. De kandidaat wordt nu gevraagd of hij/zij wil wisselen, d.w.z. of hij/zij toch niet liever het andere nog dichte deurtje kiest. Wat zou u in zo'n geval doen?

6. (2.8) Een bericht wordt versleuteld verstuurd, met n mogelijke decodeersleutels. De ontvanger weet niet welke sleutel de goede is, en probeert ze één voor één. Bereken de kans dat de achste poging het bericht juist decodeert.

7. (2.12(4)) Bereken de kans dat een gezin met 6 kinderen bestaat uit 3 jongens en 3 meisjes. Neem aan dat de kans op een jongen gelijk is aan $p = \frac{1}{2}$.

8. (2.12(4)) Iemand heeft 5 backups gemaakt op 5 verschillende flops, maar is vergeten op welke flops. Hij is in het bezit van 25 flops, en bekijkt hiervan achtereenvolgens 4 flops. Bereken de kasn dat er bij deze 4 flops 2 van de gezochte backups zitten.

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

10. (2.9) Beschouw het volgende schakelingssysteem. De kans dat een schakel dicht is is p, en de schakels zijn o.o.. Bereken de kans op verbinding tussen A en B.

11. (1.4, 1.6) Laat X een aselecte trekking zijn uit de getallen $\{1, \ldots, 7\}$, en zij $Y = (X-4)^2$. Bepaal $P(Y = y)$ voor alle mogelijke waarden van y. Bepaal de verdelingsfunctie van Y.

12. (1.9) Laat X en Y o.o. aselecte trekkigen zijn uit $\{1, \ldots, 7\}$. Hoe ziet de verdeling van $Z = X + Y$ er uit? (Maak eventueel een simulatie, d.w.z. neem een steekproef $Z_1, \ldots, Z_m$ uit Z en maak een histogram.)

13. (1.15) 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.)

14. (1.10) Neem een steekproef $X_1, \ldots, X_n$ van grootte n uit een willekeurige verdeling (bijvoorbeeld uit $X = -\log U$ met U uniform verdeeld). Doe dit m keer en maak een histogram van de m zo verkregen gemiddelden $\bar{X} = \sum_{i=1}^{n} X_i/n$.

15. (2.6) Laat zien dat

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$
$$-P(A \cap B) - P(A \cap C) - P(B \cap C)$$
$$+P(A \cap B \cap C).$$

Geef een soortgelijke uitdrukking voor $P(A \cup B \cup C \cup D)$.

16. (2.5, 2.7) Van de gebeurtenissen A en B is gegeven:

$$P(A) = 0.30,\ P(B) = 0.78,\ P(A \cap B) = 0.16.$$

Bereken:

$$P(A \cup B),\ P(\bar{A} \cup \bar{B}),\ P(A|B),\ P(B|\bar{A}).$$

17. (2.8) Een verzekeringsmaatschappij onderscheidt high-risk, medium-risk en low-risk cliënten, met kansen resp. 0.02, 0.01 en 0.0025 dat een dergelijk clënt een claim indient. De percentages cliënten van de verschillende categorieën zijn resp. 10 %, 20 % en 70 %. Welk percentage van de claims komen van high-risk cliënten?

18. (2.7) Stel dat de kans om ouder te worden dan 70 jaar gelijk is aan 0.6, en de kans om ouder te worden dan 80 jaar gelijk is aan 0.2. Als iemand nu haar 70ste verjaardag heeft bereikt, wat is dan de kans dat zij ook haar 80ste zal mogen vieren?

19. Stel n componenten zijn in serie verbonden. Voor iedere unit is er een backup, en het syteem faalt dan en slechts dan als minstens één unit plus de bijbehorende backup falen. Veronderstel onafhankelijkheid van de units/backups, en dat de kans op falen voor een unit/backup gelijk is aan p. Wat is de kans dat het systeem werkt?

20. (3.7(3)) Een systeem bestaat uit n onafhankelijke units, die elk kans p hebben om te falen. Het systeem faalt als er minstens k units falen. Wat is de kans dat het systeem faalt?

21. Deze opgave behandelt een eenvoudig voorbeeld van z.g. *vertakkingsprocessen*. Een populatie begin met één individu; op tijdstip $t = 1$ zal deze zich ofwel delen met kans p, ofwel sterven met kans $1-p$. Als het zich deelt, dan gedragen beide kinderen zich onafhankelijk, met dezelfde twee mogelijkheden op tijdstip $t = 2$. Wat is de kans dat er geen individuen zijn in de derde generatie? Voor welke waarde van p is deze kans gelijk aan $\frac{1}{2}$?

22. Hier is een eenvoudig model voor *wachtrijen*. Deze wachtrij verloopt in discrete tijd ($t = 0, 1, 2, \ldots$), en per eenheid van tijd wordt de eerste persoon in de rij bediend met kans p, en arriveert een nieuw persoon met kans q. Op tijdstip $t = 0$ is er één persoon in de wachtrij. Bepaal de kansen op 0,1,2,3 mensen in de wachtrij op tijdstip $t = 2$.

23. Deze opgave introduceert 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 of 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/haar genen bij; één van het tweetal wordt overgedragen met kans $\frac{1}{2}$.

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

b) Stel dat de kansen op genotypes AA, Aa en aa zijn p, $2q$ en r, resp. in de eerste generatie. Bepaal de kansen voor de tweede en derde generatie, en laat zien dat ze dezelfde zijn. Dit heet de wet van Hardy-Weinberg.

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

24. (3.7(5)) Een bedrijf is slecht bereikbaar: bijna iedere keer als men belt is de lijn bezet. Laat X het aantal keren zijn dat men moet bellen om uiteindelijk iemand aan de lijn te krijgen, en stel dat

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

Hier is p de kans dat de lijn bezet is. Bepaal de verdelingsfunctie $F(x)$ (in $x = 1, 2, \ldots$) van X.

25. (3.7(3)) Een multiple-choice test bestaat uit 20 vragen, elk met bij elke vraag de keuze uit 4 mogelijke antwoorden. Een zekere student beheerst de stof niet al te best, maar kan wel bij ieder vraag een van de antwoorden elimineren. Van de overige 3 antwoorden kiest de student er één *op goed geluk*. De eis is dat tenminste 12 van de vragen correct zijn beantwoord.

a) Wat is de kans dat de student slaagt?

b) Bepaal de kans op slagen nog eens, maar nu onder de aaname dat de student twee van de mogelijke antwoorden kan elimineren.

26. (3.7(3)) Door drie extra bits aan een vier-bit woord toe te voegen op een bepaalde manier (een Hamming code), kan men tot één fout in een bit detecteren en corrigeren. Als elk 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 het woord correct wordt ontvangen (d.w.z. nul of één bit is fout)? Wat is de kans dat het woord correct wordt ontvangen als er geen check bits zijn?

27. (3.7(6)) Het aantal e-mailtjes dat per uur binnenkomt is Poisson verdeeld met parameter $\lambda = 2$.

a) Bereken de kans dat er een e-mailtje is tijdens een koffiepauze van 10 minuten.

b) Hoe lang kan men pauze nemen als men eist dat de kans dat er geen e-mailtje tijdens de pauze binnenkomt tenminste 0.5 is?

28. (3.7(6)) Een zeldzame ziekte heeft een incidentie van één op de 1000. Stel dat de individuen in een populatie onafhankelijk van elkaar al of niet geinfecteerd raken. Bepaal de kans op x gevallen in een populatie van 10 000 individuen, voor $x = 0, 1, 2, \ldots$.

29. (1.9, 3.7(3), 3.12) Laat X binomiaal verdeeld zijn met parameters n en p, en Y binomiaal verdeeld met parameters m en p. Veronderstel verder dat X en Y o.o. zijn. Bepaal de verdeling van $X + Y$.

30. (1.9, 3.7(6), 3.12) Laat X en Y o.o. Poisson verdeeld zijn met parameters resp. μ en ν. Bepaal de verdeling van $X + Y$. Bereken $P(X = x|X + Y = N)$.

31. (2.7, 3.8(3)) Stel X is exponentieel verdeeld met parameter λ. Laat zien dat

$$P(X \leq a + x|X > a) = P(X \leq x),\ x > a > 0.$$

32. (3.8(2)) Veronderstel dat bij telefoongesprekken de gespreksduur normaal verdeeld is met $\mu = 2$ min. en $\sigma = 30$ seconden. Bereken de kans dat een gesprek

(a) langer duurt dan 3 minuten,

(b) korter duurt dan 30 seconden,

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

Opmerking: De aaname van normaliteit is in dit voorbeeld nogal vreemd, omdat gepreksduren niet negatief kunnen zijn. Een meer realistische aanname is de z.g. *lognormale* verdeling. (X is lognormaal verdeeld als $\log X$ normaal verdeeld is.)

33. (3.8(2), 3.10) Stel $X_1, \ldots, X_4$ zijn o.o. en $N(0,1)$-verdeeld. Noem $\bar{X} = (X_1 + \ldots + X_4)/4$. Bereken

$$P(\bar{X} > \frac{1}{2}).$$

34. (3.8(2), 3.10) Stel X en Y zijn o.o. en $X \sim N(0,25)$, $Y \sim N(-1,9)$. Bereken $P(X - Y > 2)$ en $P(2X + 3Y > 5)$.

35. (3.8(3)) Stel X en Y zijn o.o. exponentieel verdeeld met parameter λ. Bepaal de verdeling van $\min(X, Y)$.

36. (3.5) Stel X heeft verdelingsfunctie F. Bereken de dichtheid f in de volgende gevallen:
(a) $F(x) = x^2$, $0 \leq x \leq 1$,
(b) $F(x) = 1 - (1+x)^{-4}$, $x \geq 0$,
(c) $F(x) = \sin(x)$, $0 \leq x \leq \frac{\pi}{2}$.

37. (3.5) Stel X heeft dichtheid f. Bereken de verdelingsfunctie F in de volgende gevallen:
(a) $f(x) = 12x^2(1-x)$, $0 \leq x \leq 1$,
(b) $f(x) = \frac{1}{2}x + \frac{1}{2}$, $-1 \leq x \leq 1$,
(c) $f(x) = \sin(x)$, $0 \leq x \leq \frac{\pi}{2}$.

38. (3.11) Genereer een steekproef ter grootte n uit F, met F gegeven in 36(a),(b) of (c). Teken de empirische verdelingsfunctie F_n en de theoretische verdelingsfunctie F in één plaatje.

Opgaven bij de hoofdstukken 4 en 5.

Opgave 1. Stel X heeft verdeling

$$P(X=1)=\frac{1}{6},\ P(X=2)=\frac{2}{6},\ P(X=3)=\frac{1}{6},\ P(X=6)=\frac{2}{6}.$$

Bereken EX en $\mathrm{var}(X)$.

Opgave 2. Gegeven de stochastische grootheden X en Y, met verdeling

$$P(X=1,Y=-1)=0.25,\ P(X=1,Y=1)=0.35,\ P(X=2,Y=-1)=0.20,\ P(X=2,Y=1)=0.20.$$

a) Bepaal de verdeling van X en van Y.
b) Zijn X en Y o.o.?
c) Bereken EX, EY, en $\mathrm{var}(X)$, $\mathrm{var}(Y)$.
d) Bereken EXY en $\mathrm{cov}(X,Y)$.

Opgave 3. Een vaas bevat 2 groene, 4 blauwe en 4 rode knikkers. Men trekt aselect knikkers uit de vaas. Hoe lang duurt het gemiddeld totdat men een blauwe knikker heeft getrokken, bij
a) trekkingen met terugleggen,
b) trekkingen zonder terugleggen?

Opgave 4. Men gooit met twee dobbelstenen. Bereken de verwachting en de variantie van het aantal ogen.

Opgave 5. Beschouw een stochastische grootheid met dichtheid

$$f(x)=\frac{1}{2}\mathrm{e}^{-|x|},\ -\infty<x<\infty.$$

Bepaal EX en $\mathrm{var}(X)$.

Opgave 6. Stel X bezit de Poisson verdeling met parameter μ. Laat zien dat voor $k\in\{0,1,2,\ldots\}$,

$$E(X(X-1)(X-2)\ldots(X-k))=\mu^{k+1}.$$

Opgave 7. Laat X een discrete s.g. zijn met waarden in $\{0,1,\ldots\}$. Laat zien dat

$$EX=\sum_{k=0}^{\infty}P(X>k)$$

(als de oneindige som aan de rechterhand convergeert).

Opgave 8. Laat X het aantal keren gooien met een muntje zijn, totdat men n keer achter elkaar kruis heeft gegooid. Bepaal EX.

Opgave 9. Stel X bezit de standaard normale verdeling. Bepaal $E\mathrm{e}^{2X}$.

Opgave 10. Laat $(X_1,Y_1),\ldots,(X_n,Y_n)$ een steekproef zijn uit (X,Y), en zij S_X^2 resp. S_Y^2 de steekproefvariantie van X resp. Y, en S_{XY} de steekproefcovariantie. Noem S_{X+Y}^2 de steekproefvariantie van $X+Y$. Laat zien dat

$$S_{X+Y}^2=S_X^2+S_Y^2+2S_{XY}.$$

Opgave 11. Laat $a>0$ en $b>0$ positieve getallen zijn, en X en Y twee stochastische grootheden. Ga na dat de correlatie tussen aX en bY gelijk is aan de correlatie tussen X en Y.

Opgave 12. Toon aan dat $|\rho_{XY}|\leq 1$. (Hint: neem (zonder verlies van algemeenheid: zie opgave 11) aan dat $\mathrm{var}(X)=\mathrm{var}(Y)=1$, en bekijk $\mathrm{var}(X+Y)$ en $\mathrm{var}(X-Y)$.)

Opgave 13. Stel X en Y zijn twee stochastische grootheden met gelijke variantie σ^2. Bepaal $\mathrm{cov}(X+Y, X-Y)$.

Opgave 14. Laat $X_1, \ldots, X_n$ een steekproef zijn uit één of andere verdeling, zeg F. Noem het steekproefgemiddelde $\bar{X}_n = \sum_{i=1}^n X_i/n$. Voor $m \leq n$, zij $\bar{X}_m = \sum_{i=1}^m X_i/m$ het steekproefgemiddelde gebaseerd op de eerste m waarnemingen. Laat zien dat

$$E\frac{\bar{X}_m}{\bar{X}_n} = 1.$$

Bepaal ook $\mathrm{cov}(X_n, X_m)$.

Opgave 15. Beschouw een steekproef $(X_1, Y_1), \ldots, (X_n, Y_n)$ uit (X, Y), met steekproefgemiddelden $(\bar{X}, \bar{Y})$. Laat zien dat

$$\mathrm{cov}(\bar{X}, \bar{Y}) = \frac{1}{n}\mathrm{cov}(X, Y),$$

en

$$\rho_{\bar{X}\bar{Y}} = \rho_{XY}.$$

Opgave 16. Bereken m.b.v. de normale benadering de kans dat er bij 200 experimenten 50 successen zijn, bij o.o. experimenten met kans $p = 1/5$ op succes.

Opgave 17. Stel $X_1, \ldots, X_n$ zijn o.o. alternatief verdeeld met kans $p = P(X_i = 1) = 1 - P(X_i = 0)$ op succes, $i = 1, \ldots, n$. (Er geldt dus $EX_i = p$, $\mathrm{var}(X_i) = p(1-p)$, $i = 1, \ldots, n$.) Noem $\hat{p}$ de fractie successen. Ga na dat $\bar{X} = \hat{p}$ en $\hat{\sigma}^2 = \hat{p}(1-\hat{p})$, waarbij $\bar{X}$ het steekproefgemiddelde, en $\hat{\sigma}^2 = \frac{n-1}{n}S^2$, met S^2 de steekproefvariantie (zie 4.11).

Opgave 18. Laat X binomiaal verdeeld zijn met parameters $n = 100$ en $p = \frac{1}{3}$. Bepaal de waarde c zodat het verschil tussen X en EX hoogstens c is,

(a) met kans (ongeveer) 95 %,

(b) met kans (ongeveer) 99 %.

Opgave 19. Beschouw $n = 4$ o.o. waarnemingen $X_1, \ldots, X_4$ uit één of andere X. De gevonden waarden zijn

$$x_1 = 0.26,\ x_2 = 5.12,\ x_3 = -0.16,\ x_4 = 3.91.$$

Bepaal een (asymptotisch!) 95 % betrouwbaarheidsinterval voor EX. (N.B. We nemen hier weinig waarnemingen zodat het rekenwerk beperkt is en men het zonder computer kan uitrekenen. Met $n = 4$ waarnemingen is het asymptotische betrouwbaarheidsinterval natuurlijk i.h.a. geen goede benadering.)

Opgave 20. Neem een steekproef $X_1, \ldots, X_n$ uit de uniforme verdeling op $[0, 1]$. Bepaal een 95 % betrouwbaarheidsinterval voor de verwachting $\mu = \frac{1}{2}$. (Neem $n = 100$.)

Opgaven bij hoofdstuk 6.

Opgave 1. Laat $X_1, \ldots, X_n$ een steekproef zijn uit een verdeling met verwachting μ en variantie σ^2.

(a) Toon aan dat elke lineaire combinatie $T = \sum_{i=1}^n a_i X_i$ met $\sum_{i=1}^n a_i = 1$ een zuivere schatter is van μ.

(b) Toon aan dat van deze zuivere schatters, de schatter $\bar{X}$ de kleinste variantie heeft.

Opgave 2. Laat $X_1, \ldots, X_n$ een steekproef zijn uit de Poisson verdeling met parameter μ. Bepaal de meest-aannemelijke schatter van μ.

Opgave 3. Laat $X_1, \ldots, X_n$ een steekproef zijn uit de exponentiële verdeling met parameter λ. Bepaal de meest-aannemelijke schatter van λ.

Opgave 4. Laat $X_1, \ldots, X_n$ een steekproef zijn uit een verdeling met dichtheid

$$f_\theta(x) = \theta(\theta+1)x^{\theta-1}(1-x),\ 0 < x < 1,$$

waarbij $\theta > 0$ een onbekende parameter is.

(a) Bepaal de meest-aannemelijke schatter voor θ.

(b) Bereken de meest-aannemelijke schatting $\hat{\theta}$, als de waarnemingen zijn

0.53 0.71 0.62

0.41 0.58 0.57

0.30 0.28 0.39

0.89 0.79 0.98

0.43 0.23 0.75.

Maak een histogram en teken $f_{\hat{\theta}}$ in dezelfde figuur.

Opgave 5. Veronderstel het lineaire regressiemodel $Y_i = \alpha + \beta x_i$ voor de volgende waarnemingen:

x	y
1	3.2
2	6.9
2	6.0
3	8.0
4	10.4
4	11.4
5	14.2
6	14.9

Bereken de kleinste-kwadratenschattingen van α en β, en zet de gegevens en de geschatte lijn in één plaatje.