Ministerie van Verkeer en Waterstaat

Directoraat-Generaal Rijkswaterstaat

RIZA Rijksinstituut voor Integraal Zoetwaterbeheer en Afvalwaterbehandeling

Bayesiaanse statistiek voor de analyse van extreme waarden

RIZA rapport 2002.006
ISBN 9036954231
Auteurs: E.H. Chbab (RIZA)
J.M. van Noortwijk (HKV)

RIZA, HKVLIJN IN WATER
Lelystad, januari 2002

HKV LIJN IN WATER

Inhoudsopgave

Samenvatting

Omdat tweederde van Nederland beneden de zeespiegel ligt, kunnen grote gebieden als gevolg van extreme hoogwaterstanden op de rivier of stormvloeden op zee overstromen. Een groot deel van Nederland is tegen overstromingen beschermd door middel van dijken, duinen of andere waterkerende constructies. Dat overstromingen in Nederland voor kunnen komen en schade kunnen veroorzaken, is gebleken tijdens de stormvloed van 1953 en tijdens de hoogwaterperiodes van 1993 en 1995.

In toenemende mate wordt voor de bescherming tegen overstromingen, zoals het ontwerpen en toetsen van dijken, gebruik gemaakt van probabilistische methoden. Soms worden deze methoden direct toegepast, soms wordt eerst in een voorschrift een vertaling gegeven naar relatief eenvoudige ontwerpregels met behulp van veiligheidscoëfficiënten. In beide gevallen liggen aan de berekening een probabilistische analyse en statistische verdelingen ten grondslag. Met name waar het gaat om overschrijdingskansen van zeldzame natuurverschijnselen, zoals extreem hoge afvoeren, windsnelheden, zeewaterstanden e.d. zijn statistische verdelingen, die de relatie tussen deze grootheden en de daarbij behorende overschrijdingskansen beschrijven, onmisbaar. Voor de toepassing van probabilistische berekeningen is het dus nodig dat men beschikt over bruikbare rekenmethoden en adequate statistische verdelingen. Een groot probleem voor de toepassing van probabilistische berekeningsmethoden is het ontbreken van goed onderbouwde statistische verdelingen voor de statistische analyses. Dit wordt met name veroorzaakt door een gebrek aan gegevens.

Het voorliggende rapport ‘Bayesiaanse statistiek voor de analyse van extreme waarden’ levert een bijdrage aan de kennisontwikkeling ten behoeve van het uitvoeren van een statistische analyse. Daarbij gaat het met name om de methodiek voor het selecteren van de juiste kansverdeling die een fysisch proces beschrijft. Van belang is daarbij dat er meestal weinig waarnemingsmateriaal is en dat men bijzondere interesse heeft in de staart van verdelingen. Een directe consequentie van een beperkt waarnemingsmateriaal is dat de met behulp van klassieke methoden gemaakte schattingen behept zijn met, soms grote, onzekerheden. Het omgaan met statistische onzekerheden is dan ook het essentiële onderdeel van dit rapport. De Bayesiaanse statistiek is een zeer goed middel hiervoor. Met deze methodiek kunnen op een nieuwe manier overschrijdingskansen worden bepaald van onder meer extreme rivierafvoeren, zeewaterstanden, windsnelheden, etc. Zo wordt een beter inzicht verkregen in de onzekerheid in die overschrijdingskansen. Dit moet uiteindelijk leiden tot een meer nauwkeurige onderbouwing van de vereiste veiligheidsniveaus van de waterkeringen.

Bij het bepalen van bijvoorbeeld de maatgevende rivierafvoer of zeewaterstand spelen onzekerheden een grote rol. Zo is voor het schatten van de maatgevende Rijnafvoer - met een overschrijdingskans van 1/1250 per jaar - een meetreeks beschikbaar van amper 100 jaar. Zo’n maatgevende afvoer kan bepaald worden door middel van statistische extrapolatie. De twee belangrijkste onzekerheden, waarmee rekening moet worden gehouden bij het bepalen van zo’n afvoer, zijn de inherente en statistische onzekerheid.

Inherente onzekerheid betreft de onzekerheid die van nature in het fysisch proces - bijvoorbeeld een rivierafvoer - aanwezig is. Vanwege de natuurlijke fluctuaties in de tijd is het zo niet mogelijk om de afvoer van volgende week of volgend jaar exact te voorspellen.
Statistische onzekerheid ontstaat indien er voor het schatten van de parameters van een kansverdeling te weinig gegevens beschikbaar zijn. Hoe meer waarnemingen, des te kleiner is de statistische onzekerheid. Ook de onzekerheid met betrekking tot de keuze van het type kansverdeling behoort tot de statistische onzekerheid. In de waterbouwkunde werd tot nu toe niet of nauwelijks rekening gehouden met statistische onzekerheden.

De Bayesiaanse statistiek is gebaseerd op een - in 1763 postuum verschenen - essay van de Engelse dominee Thomas Bayes. In dit 'Essay towards solving a problem in the Doctrine of Chances' presenteert Bayes een theorema, waarmee zowel waarnemingen als persoonlijke meningen bijdragen tot de onzekerheid in te bepalen parameter. Succesvolle toepassingen van de Bayesiaanse statistiek zijn onder meer te vinden in de geneeskunde, economie, astronomie, archeologie, ruimtevaart en kernfysica. Maar binnenkort ook in de overstromingsrisico's en waterbouwkunde. De Bayesiaanse methodiek is gebaseerd op de stelling van Bayes die kwantificeert wat de nieuwe a posteriori onzekerheden zijn, gegeven een verzameling van a priori onzekerheden en nieuwe gegevens. Met behulp van de stelling van Bayes kunnen verschillende bronnen van informatie worden gecombineerd. Zo wordt het nu bijvoorbeeld mogelijk om, bij het bepalen van de kansverdeling van afvoeren, subjectieve informatie over de extreme rivierafvoeren van 1643 en 1809 te combineren met de vanaf 1901 waargenomen afvoeren.

1 Inleiding

Omdat tweederde van Nederland beneden de zeespiegel ligt, kunnen grote gebieden als gevolg van extreem hoge waterstanden op de rivier of stormvloeden op zee overstromen. Een groot deel van Nederland is tegen overstromingen beschermd door middel van dijken, duinen of andere waterkerende constructies. Dat overstromingen in Nederland voor kunnen komen en schade kunnen veroorzaken, is gebleken tijdens de stormvloed van 1953 en tijdens de hoogwaterperiodes van 1993 en 1995.

In toenemende mate wordt voor de bescherming tegen overstromingen, zoals het ontwerpen van dijken, gebruik gemaakt van probabilistische methoden. Soms worden deze methoden direct toegepast, soms wordt eerst in een voorschrift een vertaling gegeven naar relatief eenvoudige ontwerpregels met behulp van veiligheidscoëfficiënten. In beide gevallen liggen aan de berekening een probabilistische analyse en statistische verdelingen ten grondslag. Met name waar het gaat om de overschrijdingskansen van zeldzame natuurverschijnselen, zoals extreem hoge afvoeren, windsnelheden, zeestanden e.d. zijn statistische verdelingen, die de relatie tussen deze grootheden en de daarbij behorende overschrijdingskansen beschrijven, onmisbaar.

Voor de toepassing van probabilistische berekeningen is het dus nodig dat men beschikt over bruikbare rekenmethoden en adequate statistische verdelingen. Een groter probleem voor de toepassing van probabilistische berekeningsmethoden is het ontbreken van goed onderbouwde statistische verdelingen voor de statistische analyses.

Het is de bedoeling dat dit rapport een bijdrage levert aan de kennis ten behoeve van het uitvoeren van een statistische analyse. Daarbij gaat het met name om de methodiek van het selecteren van kansverdelingen. Van belang is daarbij het kenmerk dat er meestal weinig waarnemingsmateriaal is en men bijzondere interesse heeft in de staart van verdelingen. Een directe consequentie van een beperkt waarnemingsmateriaal is dat de met behulp van klassieke methoden gemaakte schattingen behept zijn met, soms grote, onzekerheden. Het omgaan met statistische onzekerheden is dan ook het essentiële onderdeel van dit rapport. De Bayesiaanse aanpak is het (beste) middel hiervoor. De onzekerheid in de verdelingsparameters, ook *parameteronzekerheid* genoemd, en de onzekerheid in het type verdeling, *verdelingstypeonzekerheid* genoemd, zullen hierbij centraal staan.

Het rapport is als volgt ingedeeld. In hoofdstuk 2 wordt een blik geworpen op de klassieke methoden die tot nu toe toegepast worden voor het uitvoeren van statistische analyses. Bijzondere aandacht wordt daarbij gegeven aan de tekortkomingen van die methoden.
Hoofdstuk 3 geeft beknopte definities van verschillende onzekerheden.
In hoofdstuk 4 wordt ingegaan op conditionele kansen en de stelling van Bayes. Belangrijke begrippen als de waarschijnlijkheid (likelihood) van een steekproef en de waarschijnlijkheidsfunctie (likelihoodfunctie) worden gedefinieerd. Ook wordt aandacht besteed aan de begrippen a priori en a posteriori onzekerheden.

Het schatten van de verdelingsparameters, alsmede het schatten van kwantielen van een bepaalde kansverdeling met behulp van de Bayesiaanse methode zijn onderwerpen van hoofdstuk 5.
A priori kansverdelingen, zowel informatief als niet-informatief, worden behandeld in hoofdstuk 6.
Een aantal numerieke methoden voor het oplossen van meerdimensionale integralen komt in hoofdstuk 7 aan de orde.
Hoofdstuk 8 is bestemd voor Bayesfactors en het bepalen van Bayesiaanse gewichten voor verschillende verdelingen: het Bayesiaans schatten van het kansverdelingstype.

2 Klassieke statistiek

2.1 Inleiding

Om overschrijdingskansen van extreme waarden van natuurverschijnselen te schatten, kunnen de hoogste waarden in een tijdreeks van de metingen worden gebruikt. Hiertoe kan men twee alternatieve dataselectiemethoden gebruiken: de methode van de periodieke maxima en de 'Peaks Over Threshold' of POT-methode. Beide selectiemethoden en de bijbehorende alternatieve klassieke schattingsmethoden voor kansverdelingen worden hierna bondig besproken. Vervolgens komen achtereenvolgens aan de orde de betrouwbaarheid van de schattingen en statistische toetsen.

2.2 Methode van de periodieke maxima

Bij de methode van de periodieke maxima reduceert men de tijdreeksmetingen tot maximale waarden die voorkomen in vooraf gekozen tijdsintervallen van gelijke lengte, meestal een jaar. In dit geval spreken we van een serie jaarmaxima.

Vervolgens dient een keuze te worden gemaakt van een model in de vorm van een kansdichtheidsfunctie dat de verdeling van deze gegevens op een correcte wijze weergeeft. Gezien de selectiemethode is het redelijk te veronderstellen dat de verschillende gegevens onafhankelijk van elkaar zijn. Bovendien kan men aantonen dat het maximum van een groot aantal onafhankelijke identieke verdeelde toevalsvariabelen convergeert naar een bepaalde verdelingsvorm: de Gegeneraliseerde Extreme-Waardenverdeling (GEV). Deze heeft de volgende vorm:

$$F(x) = \exp\left[-\left\{1 + \gamma\frac{x - a}{b}\right\}^{-1/\gamma}\right], \quad 1 + \gamma\frac{x - a}{b} \geq 0, \quad b > 0 \tag{2.1}$$

De GEV-verdeling wordt bepaald door 3 parameters: de parameter γ bepaalt de vorm van de verdeling (de vormparameter), b is een schaalfactor (schaalparameter) en a is een plaatsingsfactor (locatieparameter). De parameter γ is de belangrijkste parameter en bepaalt het type extreme-waardenverdeling. Wanneer de vormparameter gelijk aan nul is, reduceert vergelijking (2.1) tot (d.i. de limiet voor $\gamma \to 0$):

$$F(x) = \exp\left[-\exp\left(-\frac{x - a}{b}\right)\right], \; b > 0 \tag{2.2}$$

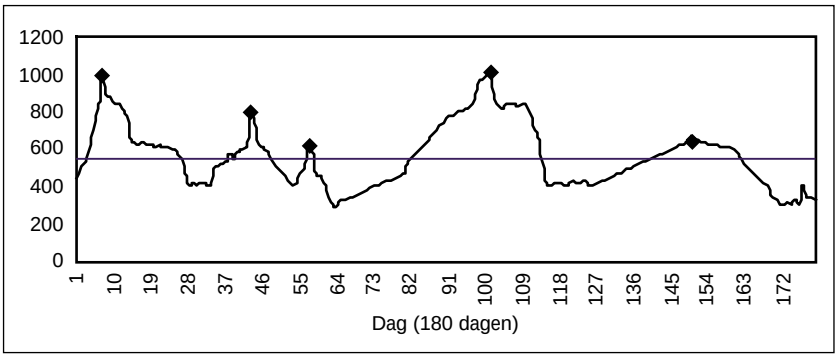
en men spreekt in dit geval van een extreme-waardenverdeling van het type I of een Gumbelverdeling. Wanneer de vormparameter positief is, spreekt men van een extreme-waardenverdeling van het type II of een Fréchetverdeling. Deze verdeling wordt gekenmerkt door een zwaardere rechterstaart en heeft een ondergrens; d.w.z. $a + b / \gamma < x < \infty$. Ten slotte, voor een negatieve vormparameter spreekt men van een extreme-waardenverdeling van het type III of een Weibullverdeling. Dit type verdeling heeft een bovengrens; d.w.z. $-\infty < x < a + b / \gamma$. Voor nadere details wordt verwezen naar Benjamin *et al.* (1970) en Castillo (1988).

Overigens dient vermeld te worden dat verscheidende andere verdelingsvormen worden toegepast in de praktijk: bijvoorbeeld de log-Pearson III verdeling (aanbevolen door de U.S. Water Resources Council) en een combinatie van de lognormale, gamma en Gumbelverdeling (aanbevolen door de commissie Boertien).

2.3 'Peaks Over Threshold'-methode

Bij de 'Peaks Over Threshold'-methode worden de maximale waarden in de tijdreeks van de metingen geselecteerd die een bepaalde grens- of drempelwaarde overschrijden. Op deze manier is het mogelijk om meer gegevens in de analyse te betrekken dan bij de serie jaarmaxima, terwijl bij geschikt gekozen drempelwaarde de maxima nog steeds redelijk onafhankelijk mogen worden verondersteld. Hoe de gegevens worden geselecteerd, wordt geïllustreerd in onderstaande figuur.

Figuur 2.1
Voorbeeld data selectiemethode.



Bovenstaande figuur geeft een hypothetisch voorbeeld van dagelijkse afvoeren in een winterhalfjaar (180 dagen). De hoogste afvoer in dat (half) jaar is 1.025 m^3/s en zal tevens het jaarmaximum zijn in de serie jaarmaxima. Vervolgens merken we op dat in hetzelfde (half)jaar de afvoer 5 keer de drempelwaarde van 550 m^3/s heeft overschreden. Bij de POT-selectiemethode met een drempelwaarde van 550 m^3/s zal de hoogste piekafvoer van elke overschrijding meegenomen worden zodat in dat jaar 5 afvoerwaarden geselecteerd zullen worden, met uiteraard de veronderstelling dat deze onafhankelijk zijn.

Net zoals bij jaarlijkse maxima, dient men ook in dit geval een kansverdelingsfunctie te formuleren die de POT-serie goed beschrijft. Het aantal waarden waarover men het maximum neemt is in de POT-methode typisch lager dan bij de methode van de jaarlijkse maxima. De convergentie naar een GEV-verdeling is daardoor in mindere mate van toepassing. Bij de schatting van de kansverdeling wordt daarom gebruik gemaakt van een alternatieve benadering. Voor de berekening van de kans van overschrijden van extreme waarden, zijn enkel de hoogste waargenomen waarden van belang. Men kan aantonen [Pickands, 1975] dat ongeacht de volledige verdelingsvorm, de staart van een willekeurige verdeling (d.w.z. voor die waarden x waarvoor geldt $x > x_0$) convergeert naar een Gegeneraliseerde Pareto of GPV-verdeling, indien x_0 voldoende groot wordt gekozen. De GPV-verdeling heeft de volgende vorm:

$$F(x) = 1 - \left(1 + \gamma \frac{x - x_0}{b}\right)^{-1/\gamma} \qquad (2.3)$$

Voor alle x waarvoor geldt $1 - \gamma (x - x_0) / b > 0$. De parameter γ is een vormparameter, b is een schaalparameter en x_0 is de drempelwaarde (locatieparameter) waarbij de convergentie van de staart naar de GPV voldoende nauwkeurig is.
Wanneer de vormparameter negatief is heeft de GPV-verdeling een bovengrens, te weten $x_0 \leq x < x_0 + b / \gamma$. Wanneer de vormparameter positief is heeft de GPV-verdeling geen bovengrens.
De situatie $\gamma = 0$ is een limietgeval voor $\gamma \rightarrow 0$; $(1 - \gamma(x - x_0) / b)^{-1/\gamma}$ gaat in de limiet naar $\exp(-(x - x_0) / b)$; dit is de exponentiële verdeling. De exponentiële verdeling is dus een speciaal geval van de GPV-verdeling.

2.4 Parameterschattingsmethoden

In deze paragraaf zullen drie veel gebruikte schattingsmethoden kort aan de orde komen. Achtereenvolgens zullen behandeld worden: de momentenmethode, de maximum-likelihoodmethode en de kleinste-kwadratenmethode. De tekst is deels overgenomen uit het CUR-rapport, bijlage C van 'Kansen in Civiele Techniek, deel 1', CUR (1997). Opgemerkt dient te worden dat er andere schattingsmethoden zijn, maar die zullen hier niet aan de orde komen. Voor andere schattingsmethoden wordt verwezen naar o.a van Gelder (1999).

2.4.1 Momentenmethode

Als de parameters van de verdeling samenvallen met de momenten (gemiddelde, variantie, etc.) kunnen deze parameters rechtstreeks worden bepaald door de momenten van de waarnemingen uit te rekenen. In het geval dat deze niet samenvallen worden zowel de momenten van de waarnemingen bepaald als de momenten van de te fitten kansdichtheidsfunctie. Door de momenten van de waarnemingen en die van de kansdichtheidsfunctie aan elkaar gelijk te stellen kunnen de bijbehorende parameters worden uitgerekend.

In het algemeen is het k^{de} moment van een kansdichtheidsfunctie f (ten opzichte van de as $x = 0$) gedefinieerd als:

$$E(x^k) = \int_{-\infty}^{+\infty} x^k f(x) dx \qquad (2.4)$$

Evenzo wordt het k^{de} centrale moment van f (ten opzichte van de as $x = \mu_x$) gedefinieerd als:

$$m_k = \int_{-\infty}^{+\infty} (x - \mu_x)^k f(x) dx = E\left((x - \mu_x)^k\right) \qquad (2.5)$$

Het eerste moment en het tweede centrale moment komen overeen met respectievelijk de verwachting en de variantie van de kansdichtheidsfunctie f. De momentenschatters zijn vaak onzuiver, niet efficiënt en soms zelfs niet betrouwbaar. Ter illustratie het volgende voorbeeld: de gegeneraliseerde extreme-waardenverdeling heeft geen verwachting voor een vormparameter kleiner dan -1 ($\gamma < -1$); en geen variantie voor een vormparameter kleiner dan -1/2 ($\gamma < -1/2$). De waarnemingen hebben uiteraard wel altijd een eindig gemiddelde en een eindige standaardafwijking. Men komt zodoende bij voorbaat op verkeerde waarden terecht.

2.4.2 Maximum-likelihoodmethode

Het uitgangspunt van de maximum-likelihoodschattingsmethode is de simultane kansdichtheidsfunctie van de waarnemingen: de waarschijnlijkheid van de waarnemingen, ook de likelihoodfunctie genoemd. De likelihood is een functie van de verdelingsparameters. De maximum-likelihoodmethode houdt dan in dat als schatters voor de parameters de waarden worden gekozen met de grootste waarden van de likelihoodfunctie.

Ter illustratie nemen we een algemeen geval en beschouwen de n stochastische grootheden $X_1,\ldots, X_n$ met kansdichtheidsfunctie f. De functie f heeft $\theta_1,\ldots, \theta_k$ als parameters. De likelihoodfunctie van de n waarnemingen $\mathbf{x} = (x_1,\ldots, x_n)$ is hiermee gedefinieerd als een functie van $\theta_1,\ldots, \theta_k$ en geeft de waarschijnlijkheid weer van de verkregen waarnemingen. De likelihoodfunctie wordt nu gegeven door:

$$l(x_1,\ldots,x_n \mid \theta_1,\ldots,\theta_k) = f(x_1,\ldots,x_n \mid \theta_1,\ldots,\theta_k) \tag{2.6}$$

Voorbeeld 1 (likelihood onafhankelijke waarnemingen): Laat $X_1,\ldots, X_n$ onafhankelijke stochastische grootheden zijn met kansdichtheidsfunctie f. De functie f heeft parameter θ. De likelihoodfunctie van de n waarnemingen $\mathbf{x} = (x_1,\ldots, x_n)$ wordt gegeven door:

$$l(x_1,\ldots,x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{2.7}$$

Voorbeeld 2 (exponentiële verdeling): Stel de onafhankelijke waarnemingen $\mathbf{x} = (x_1,\ldots, x_n)$ vormen een steekproef uit de exponentiële verdeling met onbekende verwachting θ, $\theta > 0$. De exponentiële kansdichtheidsfunctie van X is $f(x \mid \theta) = \exp\{-x/\theta\}/\theta$. De likelihoodfunctie van $\mathbf{x} = (x_1,\ldots, x_n)$ wordt dan gegeven door:

$$l(x_1,\ldots,x_n \mid \theta) = f(x_1,\ldots,x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = \left[\frac{1}{\theta}\right]^n \exp\left\{-\frac{\sum_{i=1}^{n} x_i}{\theta}\right\} \tag{2.8}$$

Het vinden van de maximum-likelihoodschatter voor de parameter θ in voorbeeld 2 is equivalent aan het vinden van het maximum van de likelihoodfunctie gegeven in (2.8).

In het geval er verschillende parameters van de verdeling moeten worden geschat, kan dit ook met de maximum-likelihoodmethode. De likelihoodfunctie is in zo'n geval multidimensionaal, zie ook de algemene formule (2.6) voor de likelihoodfunctie. De maximum-likelihoodschatters zijn efficiënt en uitputtend en zijn verder niet altijd zuiver.

2.4.3 Kleinste-kwadratenmethode

De methode van de kleinste kwadraten is een algemeen recept voor het fitten van functies op waarnemingen. Bij deze methode wordt verondersteld dat bij de beste passing de som van de kwadraten van de verschillen tussen de berekende en de waargenomen waarden minimaal is.
Als van de waarnemingen het kansverdelingstype F bekend is, kan de som van de kwadraten worden beschreven als functie van de parameters van de verdeling:

$$f(\theta_1,\ldots,\theta_k) = \sum_{i=1}^{n} \{P(x_i) - F(x_i|\theta_1,\ldots,\theta_k)\}^2 \tag{2.9}$$

Waarin:
$P(x_i)$ = fractie van de waarnemingen dat kleiner is dan x_i;
$F(x_i)$ = berekende onderschrijdingskans;
n = het aantal waarnemingen.

De fractie van de waarnemingen wordt doorgaans bepaald door alle waarnemingen op volgorde te zetten, waarbij de laagste waarneming het rangnummer 1 krijgt en de hoogste waarneming het rangnummer n. Vervolgens geldt dan:

$$P(x_i) = \frac{i}{n+1} \tag{2.10}$$

Formule (2.10) wordt de plotpositie genoemd. Er dient opgemerkt te worden dat er verschillende plotposities bestaan, zie Castillo (1988) voor meer details.

De maximale waarden van de functie *f* worden verkregen door partiële differentiatie naar de parameters $\theta_1,..., \theta_k$. Vervolgens moeten de parameters zodanig worden gekozen dat de partiële afgeleiden nul zijn.

2.5 Betrouwbaarheid van de geschatte parameters

Om inzicht te krijgen in de nauwkeurigheid, waarmee een parameter kan worden geschat, kunnen met de klassieke methoden betrouwbaarheidsintervallen worden gedefinieerd. Het een en ander kan het beste geïllustreerd worden aan de hand van het volgende voorbeeld.
Stel dat van een stochastische variabele bekend is dat deze normaal verdeeld is met een bekende standaard afwijking σ. Het gemiddelde μ van de verdeling wordt benaderd met als schatter het steekproefgemiddelde *m*. Dit steekproefgemiddelde is een normaal verdeelde stochastische variabele (centrale limietstelling) met een gemiddelde μ en standaardafwijking $\sigma/\sqrt{n}$.
Met behulp van de standaardnormale verdeling kan de kans worden gedefinieerd dat het steekproefgemiddelde binnen bepaalde grenzen ligt:

$$P\left(\mu - u\frac{\sigma}{\sqrt{n}} < m < \mu + u\frac{\sigma}{\sqrt{n}}\right) = \Phi(u) - (1 - \Phi(u)) \tag{2.11}$$

Voor het geval $u = 1.64$ geldt:

$$P\left(\mu - 1.64\frac{\sigma}{\sqrt{n}} < m < \mu + 1.64\frac{\sigma}{\sqrt{n}}\right) = 0.90 \tag{2.12}$$

Hieruit mag overigens niet worden geconcludeerd dat het gemiddelde μ tussen $m-\mu.\sigma / \sqrt{n}$ en $m+\mu.\sigma / \sqrt{n}$ ligt. Het gemiddelde is immers een deterministische parameter en kan dus niet als stochast worden behandeld. Om toch betrouwbaarheidsgrenzen te kunnen aangeven, volgt de klassieke statistiek de redenering met het verwerpen van hypothesen. Ten aanzien van het onbekende gemiddelde kan een aantal hypothesen worden gemaakt, bijvoorbeeld:

$$\begin{aligned} &H_1: && \mu < m - 1.64\sigma / \sqrt{n} \\ &H_2: && \mu - 1.64\frac{\sigma}{\sqrt{n}} \le m < \mu + 1.64\frac{\sigma}{\sqrt{n}} \\ &H_3: && m + 1.64\sigma / \sqrt{n} \le \mu \end{aligned} \tag{2.13}$$

De kans dat hypothese H_1 waar is en tegelijkertijd het steekproefgemiddelde gelijk is aan *m*, is slechts 5%. Hetzelfde geldt voor H_3. Op grond hiervan worden de hypothesen H_1 en H_3 verworpen. Wat overblijft is het betrouwbaarheidsinterval.

2.6 Keuze verdelingstype: statistische toetsen

In het voorgaande is verondersteld dat het kansverdelingstype van de beschouwde stochastische variabele reeds bekend is. Meestal is dit kansverdelingstype echter niet bekend. De keuze van het type is afhankelijk van de kennis met betrekking tot de stochastische variabele.
Als waarnemingen beschikbaar zijn, kan een schatting van het verdelingstype worden gemaakt met behulp van klassieke methoden. Hieronder wordt behandeld hoe de keuze van het type verdeling gemaakt wordt.

2.6.1 Chi-kwadraattoets

De χ^2-toets vergelijkt het histogram van de waarnemingen met de kansdichtheid van de gekozen kansverdeling. Bij deze methode wordt een toetsingsgrootheid y berekend volgens de formule:

$$y = \sum_{i=1}^{k} \frac{\left[N_i - n \int_{a_i}^{b_i} f(x)dx \right]^2}{n \int_{a_i}^{b_i} f(x)dx} \qquad (2.14)$$

Waarin: i = intervalnummer van het histogram;
a_i = ondergrens van interval *i;*
b_i = bovengrens van interval *i;*
k = aantal intervallen;
N_i =aantal waarnemingen in interval *i;*
n = het totale aantal waarnemingen;
f = gekozen kansdichtheid.

Bewezen is dat *y* voor grote *n* een χ^2 -verdeling heeft met *k-r*-1 vrijheidsgraden. Hierin is *r* het aantal parameters van f dat op basis van de waarnemingen is geschat.
In de literatuur zijn waarden van χ_v^2 te vinden voor verschillende overschrijdingskansen *P* en vrijheidsgraden *v.* Als groter is dan $\chi^2_{k\text{-}r\text{-}1}$ bij een gekozen overschrijdingskans *P*, dan moet de gemaakte keuze van $f(x)$ worden verworpen. Een nadeel van de χ^2-toets is dat het resultaat afhankelijk is van de gekozen histogramindeling.

2.6.2 Kolmogorov-Smirnovtoets

De Kolmogorov-Smirnovtoets vergelijkt de gekozen verdelingsfunctie met de waarnemingen. Bij deze methode wordt wederom een toetsingsgrootheid γ berekend. Ditmaal met de volgende formule:

$$y = \max\left(\left|\hat{F}(x_i) - F(x_i)\right|\right) \qquad (2.15)$$

Waarin: i = rangnummer van de waarneming, nadat alle waarnemingen Zijn gerangschikt;
n = aantal waarnemingen;
x_i = i^{de} waarneming na rangschikking (van klein naar groot);

$\hat{F}$ = empirische verdelingsfunctie gedefinieerd als i/n;
F = gekozen verdelingsfunctie.

De Kolomogorov-Smirnovtoets, maar ook de χ^2-toets, is puur gebaseerd op de waarnemingen en deze houdt geen rekening met andere kennis die relevant kan zijn voor de keuze van het verdelingstype.
We kunnen nu de hypothese $F = F_0$ toetsen. Als toetsingsgrootheid nemen we de stochast y. De verdeling van y is theoretisch te bepalen en hangt niet af van de gekozen verdeling F, maar alleen van het aantal waarnemingen n. Als na het beschikbaar komen van de waarnemingen blijkt dat y groter is dan een of andere (bij onbetrouwbaarheid α bepaalde) kritieke waarde $y_{n,\alpha}$, dan wordt de hypothese $F = F_0$ verworpen. Hieruit is dus $y_{n,\alpha}$ bepaald door $P(y > y_{n,\alpha})=\alpha$.

2.7 Nadelen statistische toetsen

Zoals reeds vermeld, kunnen bij een klassieke analyse bovengenoemde toetsen worden gebruikt voor het al dan niet 'verwerpen' van kansverdelingen. Er wordt dus eerst een hypothese gedaan dat een bepaalde stochastische grootheid een zekere kansverdeling heeft. Uitgaande van die hypothese wordt vervolgens de kritieke waarde voor een gegeven toetsingsgrootheid en een onbetrouwbaarheid α berekend. Opgemerkt dient te worden dat de onbetrouwbaarheid alleen iets zegt over een verwerping ten onrechte en niets zegt over de kans dat die kansverdeling juist is. Bovendien zijn de toetsen voor aanpassing puur gebaseerd op statistische gegevens en deze houden geen rekening met eventuele andere kennis die relevant is voor de keuze van het verdelingstype. Mede omdat in de praktijk het aantal waarnemingen beperkt is, kunnen meerdere kansverdelingen door de toetsing komen. In dit geval zijn de statistische toetsen minder kritisch te noemen.

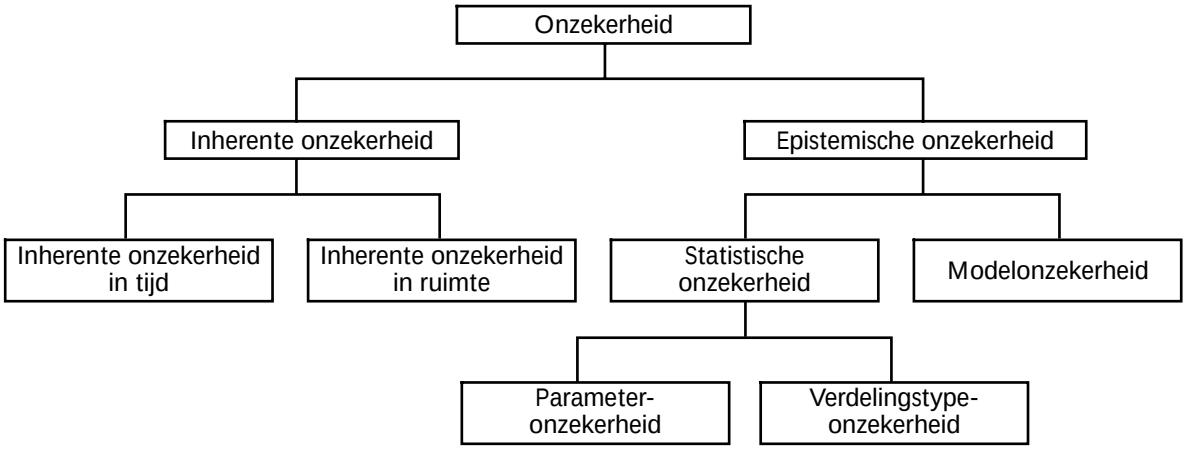
Een krachtige methode voor de keuze van het verdelingstype is de Bayesiaanse methode. In tegenstelling tot de klassieke methoden, selecteert men in de Bayesiaanse aanpak een aantal kansverdelingen vooraf, kent daaraan a priori kansen toe en bepaalt vervolgens op grond van de waarnemingen en de stelling van Bayes de a posteriori kansen. De methode wordt behandeld in de volgende hoofdstukken.

3 Onzekerheden

Bij statistische analyses zoals tot nu toe zijn uitgevoerd ter bepaling van de maatgevende afvoer van de Rijn, kan een aantal onzekerheden worden onderscheiden. Deze onzekerheden zijn onder te verdelen in twee belangrijke groepen: inherente onzekerheid en epistemische onzekerheid.

Inherente onzekerheid (natuurlijke of fysische onzekerheid); dit type onzekerheid is het gevolg van de natuurlijke fluctuatie in de tijd (en ruimte) van bijvoorbeeld de hydraulische ruwheid, de afvoer, het klimaat en het broeikaseffect. Deze van nature aanwezige onzekerheid kan niet worden gereduceerd, maar wel worden gemodelleerd door middel kansverdelingsfuncties: de likelihoodfuncties.

Figuur 3.1
Typen onzekerheden.



Epistemische onzekerheid; dit type onzekerheid is onder te verdelen in 2 typen: modelonzekerheid en statistische onzekerheid.

Modelonzekerheid; dit type onzekerheid is het gevolg van wiskundige recepten die worden gehanteerd om bepaalde fysiche verschijnselen te modelleren, zoals waterbewegings- en golfmodellen. Dit type onzekerheid valt buiten het kader van dit rapport.

Statistische onzekerheid; twee belangrijke bronnen spelen hierbij een rol:
- Parameter-onzekerheid: onzekerheid in de statistische parameters van de kansverdelingsfuncties ten gevolge van een beperkt aantal gegevens of metingen.
- Kansverdelingstype-onzekerheid: onzekerheid in het te kiezen kansverdelingstype ten gevolge van een beperkt aantal gegevens of metingen. In tegenstelling tot parameter-onzekerheid is dit type onzekerheid vaak moeilijk te kwantificeren; toch kan het worden verwerkt in de onzekerheid.

Parameter- en kansverdelingstype-onzekerheid kunnen worden gerepresenteerd door middel van kansverdelingen. Dat statistische parameters een onzekerheidsverdeling hebben is een voorbeeld van het zogenaamde subjectivistische kansbegrip. Globaal kunnen er twee kansbegrippen worden onderscheiden: het objectivistische kansbegrip en het subjectivistische kansbegrip.

Het objectivistische kansbegrip gaat uit van het mathematische concept van een oneindige reeks onafhankelijk van elkaar optredende identieke gebeurtenissen, met ‘kans’ gedefinieerd als de limiet van waargenomen frequenties. Vandaar dat dit ook wel de frequentistische benadering wordt genoemd. Een frequentistische benadering leidt tot kansverdelingen met deterministische parameters.

Het subjectivistische kansbegrip beschouwt kans als een persoonlijke opvatting over de onzekerheid dat een bepaalde gebeurtenis zal optreden. Deze a priori onzekerheid kan eventueel met beschikbare gegevens of metingen worden bijgewerkt tot de a posteriori onzekerheid door middel van Bayesiaanse statistiek. Een subjectivistische benadering leidt tot kansverdelingen met onzekere parameters die op hun beurt weer een gezamenlijke kansverdeling hebben. Het subjectivistische kansbegrip omvat het frequentistische kansbegrip.

Zowel parameter- als verdelingstype-onzekerheid worden in dit rapport bestudeerd.

4 Stelling van Bayes

4.1 Conditionele kansen

Centraal in de Bayesiaanse analyse staat de stelling van Bayes. Opgemerkt wordt dat de stelling als zodanig een gangbare wiskundige stelling is, waarbij de interpretatie in de vorm van het bijwerken met gegevens vooralsnog nog niet aan de orde is.

Stel men wil de kans op een gebeurtenis A weten gegeven het feit dat een gebeurtenis B is opgetreden. Dit betekent dat men alleen geïnteresseerd is in die uitkomsten van A, gegeven dat ze ook tot B behoren. Het gaat dus om de kans op $A \cap B$ gedeeld door de kans op B. De conditionele of voorwaardelijke kans op de gebeurtenis A gegeven het optreden van B is hiermee gedefinieerd als:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{4.1}$$

Voor de kans op B gegeven A geldt analoog:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{4.2}$$

Als met behulp van (4.2) de kans op $A \cap B$ uit (4.1) wordt geëlimineerd volgt er ten slotte:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{4.3}$$

Hiermee is de eenvoudigste vorm van de stelling van Bayes afgeleid. Deze stelling laat zien hoe de kanstoekenning van A (eerst $P(A)$) onder invloed van gegeven B gewijzigd wordt in $P(A \mid B)$. Men noemt $P(A)$ de a priori kans en $P(A \mid B)$ de a posteriori kans. Deze laatste is altijd een conditionele kans.

Voorbeeld (twee-vazenprobleem):
Gegeven zijn twee identieke vazen met oneindig veel rode en witte ballen. In de ene vaas, vaas V_1, zitten 67% rode en 33% witte ballen. De samenstelling van de tweede vaas, vaas V_2, is juist andersom, 33% rood en 67% wit. Stel dat random één van de vazen wordt gekozen en uit deze vaas, eveneens random met teruglegging, drie ballen worden getrokken: tweemaal een rode en eenmaal een witte. Gevraagd: wat is de kans dat deze steekproef afkomstig is uit vaas V_1 en wat is de kans dat deze afkomstig is uit vaas V_2?

Hiertoe voeren we de volgende notatie in:
R : er wordt een rode bal getrokken;
W : er wordt een witte bal getrokken;
V_1 : de vaas is V_1 met $P(R \mid V_1) = 2/3$ en $P(W \mid V_1) = 1/3$;
V_2 : de vaas is V_2 met $P(R \mid V_2) = 1/3$ en $P(W \mid V_2) = 2/3$.

We beginnen met de vaststelling van de a priori kansen van de twee vazen. Omdat niet meer informatie beschikbaar is, nemen we aan dat beide vazen even kansrijk zijn. Dit betekent:

$$P(V_1) = P(V_2) = 0.5 \tag{4.4}$$

De kans op de steekproef $R \cap R \cap W$, gegeven dat we vaas V_1 gekozen hebben, wordt gegeven door:

$$P(R \cap R \cap W \mid V_1) = (2/3)x(2/3)x(1/3) = 4/27 \tag{4.5}$$

Analoog gegeven vaas V_2:

$$P(R \cap R \cap W \mid V_2) = (1/3)x(1/3)x(2/3) = 2/27 \tag{4.6}$$

Met behulp van de stelling van Bayes in vergelijking (4.3) volgt:

$$P(V_1 \mid R \cap R \cap W) = \frac{P(R \cap R \cap W \mid V_1)P(V_1)}{P(R \cap R \cap W)} = \frac{(4 \,/\, 27) \text{ x } 0.5}{P(R \cap R \cap W)} \tag{4.7}$$

en

$$P(V_2 \mid R \cap R \cap W) = \frac{P(R \cap R \cap W \mid V_2)P(V_2)}{P(R \cap R \cap W)} = \frac{(2 \,/\, 27) \text{ x } 0.5}{P(R \cap R \cap W)} \tag{4.8}$$

De som van de kansen op V_1 en V_2, moet natuurlijk ook na de steekproef nog steeds 1 zijn. Deze eis maakt het mogelijk $P(R \cap R \cap W)$ te bepalen (gebruikmakend van de stelling van de totale waarschijnlijkheid):

$$P(R \cap R \cap W) = P(R \cap R \cap W \mid V_1)P(V_1) + P(R \cap R \cap W \mid V_2)P(V_2) = 3/27 \tag{4.9}$$

Daarmee volgt als eindresultaat:

$$P(V_1 \mid R \cap R \cap W) = 2/3 \text{ en } P(V_2 \mid R \cap R \cap W) = 1/3 \tag{4.10}$$

De kans dat de vaas van type V_1 is, stijgt dus als gevolg van de waarnemingen $R \cap R \cap W$ van 1/2 naar 2/3; de kans op vaas V_2 daalt van 1/2 naar 1/3. Indien de reeks waarnemingen oneindig wordt doorgezet, zal in beginsel de kans op één van de twee vazen naar 1 gaan en de ander naar 0.

Een interessante vraag is verder bijvoorbeeld wat de kans is om een rode bal uit de (onbekende) vaas te trekken, zowel voor als na de steekproef. Beschouw eerst de situatie vóór het doen van de steekproef. De kans op een rode bal is 2/3 in geval van vaas V_1, die op zich 50% kans heeft, en 1/3 in geval van vaas V_2, die ook 50% kans heeft, derhalve:

$$P(R) = P(R \mid V_1)P(V_1) + P(R \mid V_2)P(V_2) = 1/2 \tag{4.11}$$

Na de steekproef is de kans op een rode bal in geval van vaas V_1 nog steeds 2/3, maar de kans op vaas V_1 is nu geen 50% maar 67%; analoog is de kans op rood gegeven vaas V_2 nog steeds 1/3, maar de kans op vaas V_2 is gedaald naar 33%. De kans op rood gegeven de steekproef $R \cap R \cap W$ wordt dus:

$$P(R \mid R \cap R \cap W) = P(R \mid V_1)P(V_1 \mid R \cap R \cap W) + P(R \mid V_2)P(V_2 \mid R \cap R \cap W) = 5/9 \tag{4.12}$$

De kans op het trekken van een rode bal is dus onder invloed van de steekproef $R \cap R \cap W$ gestegen van 1/2 naar 5/9.

Het twee-vazen-probleem is een volledige Bayesiaanse analyse in een notendop. Ten behoeve van verdere analyse is het echter zinvol de Bayesiaanse formules verder te ontwikkelen.

4.2 Likelihoodfunctie van de waarnemingen

In het voorbeeld van de twee vazen hebben we gezien dat $P(R \cap R \cap W \mid V)$ meerdere keren is gebruikt. Dat is de kans op rood, rood en wit gegeven vaas V: de waarschijnlijkheid van de steekproef gegeven dat er getrokken wordt uit vaas V. Dit noemen we de likelihood van de steekproef.

Zoals reeds in hoofdstuk 2 geïntroduceerd, bestuderen we nu een algemeen geval en beschouwen de n stochastische grootheden $X_1,..., X_n$ met kansdichtheidsfunctie f. De functie f heeft $\theta_1,..., \theta_k$ als parameters. De likelihoodfunctie van de n waarnemingen $\mathbf{x} = (x_1,..., x_n)$ is hiermee gedefinieerd als een functie van $\theta_1,..., \theta_k$ en geeft de waarschijnlijkheid weer van de verkregen waarnemingen. De likelihoodfunctie wordt nu gegeven door:

$$l(x_1,..., x_n \mid \theta_1,..., \theta_k) = f(x_1,..., x_n \mid \theta_1,..., \theta_k) \tag{4.13}$$

Zoals reeds opgemerkt wordt de likelihood beschouwd als een functie van de parameters $\theta_1,..., \theta_k$ en niet van de waarnemingen $\mathbf{x} = (x_1,..., x_n)$.

4.3 Stelling van Bayes: a priori en a posteriori onzekerheid

Hierboven is reeds een simpele variant van de stelling van Bayes gegeven en is uitgewerkt voor het voorbeeld van twee vazen met rode en witte ballen. In deze paragraaf wordt een meer algemene formulering afgeleid; hoewel vergelijking (4.3) hierbij het uitgangspunt zal zijn.

We beschouwen een praktisch voorbeeld: de afvoeren van de Rijn. Stel we beschikken over dagelijkse afvoermetingen van de laatste n jaren. Uit deze reeks selecteren we verder de hoogste gemeten afvoer in elk hydrologisch jaar van die n jaren. Indien we verder aannemen dat de reeks homogeen is, krijgen we een serie n onafhankelijke en identiek verdeelde jaarmaxima met cumulatieve kansverdelingsfunctie F en kansdichtheidsfunctie f. De functie f heeft een aantal statistische parameters, die wij in het vervolg simpelweg ‘parameters’ noemen.

De parameters van een verdelingsfunctie behorende bij een stochastische grootheid (afvoeren in dit geval) zijn vanwege het beperkte aantal waarnemingen vaak onzeker. We nemen verder aan dat de onzekerheid in de parameters beschreven kan worden door een kansdichtheidsfunctie π. Recht toe recht aan toepassing van formule (4.3) levert:

$$\pi(\textit{parameters} \mid \textit{data}) = \frac{l(\textit{data} \mid \textit{parameters})\pi(\textit{parameters})}{f(\textit{data})} \tag{4.14}$$

met:

$l(.|.)$ = de likelihoodfunctie van de jaarmaxima van de afvoeren gegeven de parameters;

$\pi(.|.)$ = de kansdichtheidsfunctie van de parameters gegeven de waargenomen afvoeren (de a posteriori kansdichtheidsfunctie);
$\pi(\cdot)$ = de kansdichtheidsfunctie van de parameters voordat waarnemingen beschikbaar waren (de a priori kansdichtheidsfunctie) en
f = de likelihoodfunctie van de waarnemingen waarbij de onzekerheid in de parameters is uitgeïntegreerd.

Met behulp van de stelling van de totale waarschijnlijkheid kan de noemer in vergelijking (4.14) geschreven worden als functie van de likelihood *l* en kansdichtheidsfunctie π (na integreren over alle mogelijke waarden van de parameters):

$$f\ (data) = \int l(data \mid parameters)\pi(parameters)d(parameters) \qquad (4.15)$$

Vergelijking (4.14) krijgt daarmee de volgende vorm:

$$\pi(parameters \mid data) = \frac{l(data \mid parameters)\pi(parameters)}{\int l(data \mid parameters)\pi(parameters)d(parameters)} \qquad (4.16)$$

Indien we vervolgens de waargenomen afvoeren aanduiden met $q_1,..., q_n$ en de parameters met $\theta_1,..., \theta_k$, dan kunnen we (4.16) schrijven in de volgende vorm:

$$\pi(\theta_1,..., \theta_k \mid q_1,..., q_n) = \frac{l(q_1,..., q_n \mid \theta_1,..., \theta_k)\pi(\theta_1,..., \theta_k)}{\int ..\int l(q_1,..., q_n \mid \theta_1,..., \theta_k)\pi(\theta_1,..., \theta_k)\, d\theta_1,..., d\theta_k} \qquad (4.17)$$

De vergelijkingen (4.16) en (4.17) zijn een variant van de stelling van Bayes en vormen de basis voor Bayesiaanse analyses in het algemeen en Bayesiaanse parameterschattingen in het bijzonder. Het mag duidelijk zijn dat de 'parameters' in vergelijking (4.16) als stochastische grootheden worden beschouwd en derhalve een kansverdeling hebben. Zoals eerder vermeld zijn er twee soorten kansverdelingen te onderscheiden: de a priori kansverdeling en de a posteriori kansverdeling. In het volgende hoofdstuk wordt verder ingegaan op de Bayesiaanse methodiek.

5 Bayesiaanse Methode: het schatten van verdelings-parameters

5.1 Inleiding

Variabelen die aan toeval onderhevig zijn, kunnen veranderen in de tijd of kunnen ruimtelijk (i.e. afhankelijk van de plaats) fluctueren. In het eerste geval spreekt men van een stochastisch proces, in het laatste geval van een stochastisch veld. In beide gevallen zijn de variabelen niet in deterministische maar in statistische termen te beschrijven. Bedoelde variabelen zijn stochastische grootheden.

Een belangrijk onderdeel van een risico-analyse is het vaststellen van de kansverdelingen van de stochastische grootheden. De keuze van het kansverdelingstype en de kansverdelingsparameters bepalen voor een groot deel de uitkomst van de analyse. Als een groot aantal statistische gegevens bekend is, kan gebruik worden gemaakt van frequentistische methoden uit de klassieke statistiek. In veel gevallen is de hoeveelheid statistische gegevens echter ontoereikend en moet toevlucht worden genomen tot meer subjectieve methoden voor het schatten van kansverdelingstypen en statistische parameters. In dit hoofdstuk wordt aangenomen dat het kansverdelingstype wel bekend is, maar de kansverdelings voor de parameters niet. De vraag is hoe men hiervan een schatting maakt met behulp van de Bayesiaanse methodiek.

5.2 Bayesiaanse schatting van de verdelingsparameters

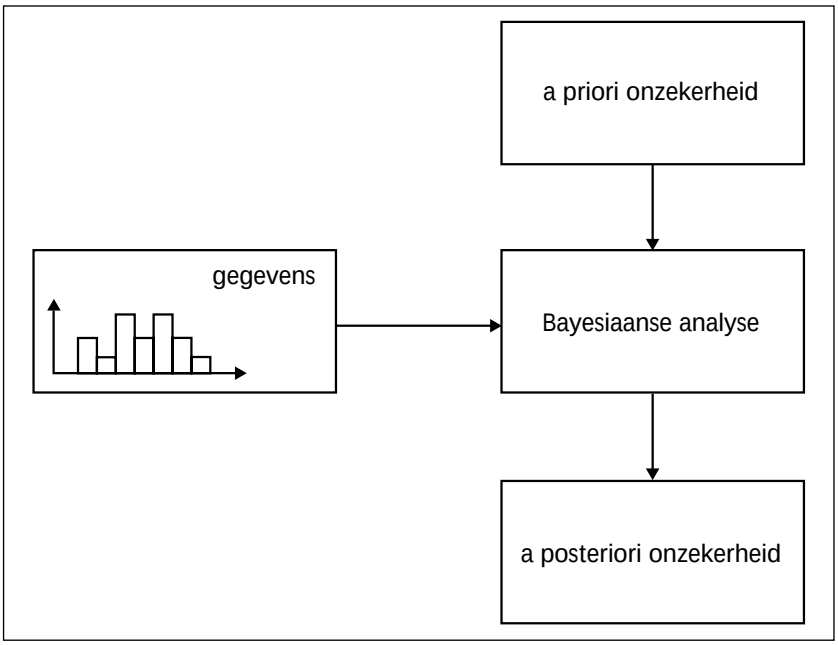
Wanneer de kansverdeling van een stochastische grootheid bekend wordt verondersteld, kunnen met behulp van waarnemingen de verdelingsparameters geschat worden door gebruik te maken van klassieke methoden. Bij deze methoden worden schattingen gemaakt die vervolgens deterministisch worden verdisconteerd in de analyse. De onzekerheid rondom de gemaakte schattingen wordt in beeld gebracht door een betrouwbaarheidsinterval (bijvoorbeeld 90%) te definiëren. De 95%-bovengrens en 5%-ondergrens van de parameterschattingen worden gebruikt voor het bepalen van een boven- en ondergrens van de te schatten waarden van de beschouwde stochastische grootheid (bijvoorbeeld de boven- en ondergrens van de schatting van een maatgevende afvoer). Dit in tegenstelling tot de Bayesiaanse methoden waarbij de parameters stochastisch worden verondersteld. Alle mogelijke parameterwaarden dragen, met een bepaalde kans van optreden, bij aan de schattingen.

Dikwijls is de hoeveelheid statistische gegevens klein; men heeft dan behoefte aan het meenemen van alle vormen van onzekerheid en het inbrengen van alle aanwezige informatie. In een Bayesiaanse analyse is dit mogelijk, zij het op een subjectieve manier. De Bayesiaanse methodiek is een mix van de subjectieve en de frequentistische parameterschatting. De eerste stap in een Bayesiaanse aanpak is het toekennen van kansen aan het juist zijn van een aantal hypothesen met betrekking tot de te schatten statistische parameters. Deze kanstoekenning resulteert in feite in een onzekerheidsverdeling voor de parameters, meestal a priori verdeling genoemd. De onzekerheidsverdeling bevat dus informatie (omtrent statistische parameters), die beschikbaar is nog voordat de statistische gegevens of waarnemingen beschikbaar zijn.

Deze informatie is subjectief en berust veelal op expertmeningen. Op het vaststellen van onzekerheidsverdelingen komen we later gedetailleerd terug.

Zodra (nieuwe) waarnemingen beschikbaar komen, kan de a priori verdeling bijgewerkt worden met behulp van de stelling van Bayes. De subjectieve informatie uit de a priori verdeling wordt gecombineerd met de objectieve waarnemingen om uiteindelijk tot een goede beschrijving van de model-parameters te komen, waarin zowel subjectieve als objectieve informatie is verdisconteerd. Via de likelihoodfunctie van de waarnemingen wordt de a priori kansverdeling van de parameters bijgewerkt tot een a posteriori kans-verdeling. A posteriori omdat de kansverdeling wordt bepaald nádat waar-nemingen beschikbaar zijn. De a posteriori kansverdeling bevat dus zowel informatie uit de a priori kansverdeling als informatie afkomstig uit de waarnemingen.

De a posteriori kansverdeling geeft de waarschijnlijkheid van de gestelde hypothesen met betrekking tot de te schatten parameters en wordt tevens beschouwd als model voor de parameteronzekerheid. Centraal in boven-genoemd proces staat de stelling van Bayes gegeven in vergelijking (4.16) en (4.17). De wijze waarop statistische gegevens met subjectieve kanstoekenning worden gecombineerd is schematisch weergegeven in figuur 4.1.



Figuur 4.1
Principe van een Bayesiaanse analyse: de a priori (subjectieve) onzekerheid wordt met behulp van de stelling van Bayes gecombineerd met (objectieve) gegevens tot de zogenaamde a posteriori onzekerheid.

Samenvatting: Bayesiaanse statistiek wordt gebruikt om schattingen te maken van de statistische parameters van een kansmodel dat een stochastische grootheid karakteriseert. Hierbij wordt gebuik gemaakt van zowel intuïtie en subjectiviteit als het beschikbare statistische materiaal (waarnemingen). De Bayesiaanse methode is dan ook een mix van subjectieve en frequentistische methoden. Het mixen van deze methoden gaat volgens een zekere procedure die in figuur 4.1 staat beschreven. Centraal daarbij staat de stelling van Bayes. Met behulp van de beschikbare waarnemingen, gerepresenteerd door middel van de likelihoodfunctie, wordt de a priori kansverdeling bijgewerkt tot een a posteriori kansverdeling van de statistische parameters. De a posteriori kansverdeling wordt beschouwd als model voor de parameter-onzekerheid.

5.3 Bayesiaanse kwantielschattingen

Stochastische grootheden worden gedefinieerd door een kansmodel of een kansverdelingsfunctie met een aantal statistische parameters. Een kansmodel geeft de relatie weer tussen waarden van die grootheid en de kansen van optreden van die waarden. Als het kansmodel bekend is en een schatting voor de parameters beschikbaar is, kan in principe voor elke gegeven kans tussen 0 en 1 de bijbehorende waarde van de stochastische grootheid worden bepaald. In het geval van een één-dimensionaal probleem en een kansmodel met één parameter wordt bij de klassieke/frequentistische methoden de volgende relatie gebruikt:

$$P(X > x_0 \mid \hat{\theta}) = 1 - F(x_0 \mid \hat{\theta}) \tag{5.1}$$

waarin $P(X > x_0 \mid \hat{\theta})$ de kans is dat de stochastische grootheid X de waarde x_0 overschrijdt gegeven de parameterwaarde $\hat{\theta}$, F het kansmodel of de cumulatieve kansverdelingsfunctie van X en $\hat{\theta}$ een geschatte parameter van F is. De parameter $\hat{\theta}$ is bijvoorbeeld de Maximum-Likelihood-schatter of een momentenschatter. Een analoge relatie geldt voor meer dimensies en/of meer parameters.

Bij vergelijking (5.1) hebben we impliciet aangenomen dat we beschikken over een steekproef $x_1,..., x_n$ van de stochastische grootheid X aan de hand waarvan de deterministische schatting $\hat{\theta}$ is gemaakt; $\hat{\theta}$ is dus een functie van $x_1,..., x_n$ die constant verondersteld wordt.

Dit in tegenstelling tot de Bayesiaanse aanpak die de parameter θ als een stochastische grootheid beschouwt. Het fundamentele verschil tussen bovengenoemde klassieke of frequentistische methode en de Bayesiaanse methode is dat in het eerste geval de parameters van het statistische model in deterministische termen worden beschreven en in het tweede geval in statistische termen. Een a posteriori kansverdeling gegeven een aantal waarnemingen is de representatie van de (onzekere) informatie over de parameters.

Als we de a posteriori kansverdeling van θ gegeven de n waarnemingen $\mathbf{x} = (x_1,..., x_n)$ in het vervolg $\pi(\theta \mid \mathbf{x})$ noemen, betekent dit voor het schatten van de overschrijdingskans van een bepaalde waarde x_0 van X dat er over alle mogelijke waarden van θ geïntegreerd moet worden. In formulevorm:

$$P(X > x_0) = \int P(X > x_0 \mid \theta)\pi(\theta \mid \mathbf{x})\, d\theta \tag{5.2}$$

Deze overschrijdingskans wordt de Bayes-schatter van de kans op overschrijden van x_0 genoemd. Substitutie van vergelijking (5.1) in (5.2) geeft:

$$P(X > x_0) = \int [1 - F(x_0 \mid \theta)]\pi(\theta \mid \mathbf{x})\, d\theta \tag{5.3}$$

Een analoge formule geldt voor meer dimensies en/of meer parameters. In het geval van meer dan één parameter wordt immers de één-dimensionale integraal in vergelijking (5.3) vervangen door een meerdimensionale integraal. De parameter θ in vergelijking (5.3) wordt dan een parametrische vector.

Volgens vergelijking (4.17) is de a posteriori kansdichtheidsfunctie gedefinieerd als:

$$\pi(\theta \mid \mathbf{x}) = \frac{l(\mathbf{x} \mid \theta)\pi(\theta)}{\int l(\mathbf{x} \mid \theta)\pi(\theta)\, d\theta} \tag{5.4}$$

Substitutie hiervan in vergelijking (5.3) geeft:

$$P(X > x_0) = \int [1 - F(x_0 \mid \theta)\pi(\theta \mid \mathbf{x})\, d\theta$$
$$= \frac{\int [1 - F(x_0 \mid \theta)] l(\mathbf{x} \mid \theta)\pi(\theta)\, d\theta}{\int l(\mathbf{x} \mid \theta)\pi(\theta)\, d\theta} \qquad (5.5)$$

In vergelijking (5.5), die gebruikt wordt voor het schatten van overschrijdingskansen van een bepaald quantiel, komt de a posteriori kansverdeling voor de parameter θ expliciet voor. Deze a posteriori fungeert als model voor parameteronzekerheid. In de te verkrijgen quantielschattingen is die onzekerheid dan ook expliciet in rekening gebracht.

Voorbeeld (Gumbelverdeling): Ter illustratie van de Bayesiaanse analyse ten behoeve van parameteronzekerheid nemen we de afvoeren van de Rijn bij Lobith. Dagelijkse afvoermetingen sinds 1901 zijn beschikbaar. Hieruit worden jaarmaxima geselecteerd en wel volgens de definitie van het hydrologisch jaar (1 oktober t/m 31 maart). De nieuwe reeks bevat dus onafhankelijke en identiek verdeelde jaarmaxima, die we $q_1,..., q_n$ (n = 99) noemen. Als kansmodel voor $q_1,..., q_n$ wordt de Gumbelverdeling aangenomen. De Gumbelverdeling heeft twee onbekende parameters: een locatieparameter μ en een schaalparameter σ. De Gumbelverdeling wordt gegeven door:

$$P(Q \le q \mid \mu, \sigma) = F(q \mid \mu, \sigma) = \exp\left\{-\exp\left\{-\frac{q - \mu}{\sigma}\right\}\right\} \qquad (5.6)$$

De kansdichtheidsfunctie is dienovereenkomstig te schrijven als:

$$f(q \mid \mu, \sigma) = \frac{dF(q \mid \mu, \sigma)}{dq} = \frac{1}{\sigma} \exp\left\{-\frac{q - \mu}{\sigma}\right\} \exp\left\{-\exp\left\{-\frac{q - \mu}{\sigma}\right\}\right\} \qquad (5.7)$$

De reeks jaarmaxima bij Lobith bevat n onafhankelijke en identiek verdeelde afvoeren $\mathbf{q} = (q_1,..., q_n)$; de likelihoodfunctie wordt in dit geval gedefinieerd als:

$$l(\mathbf{q} \mid \mu, \sigma) = l(q_1,...,q_n \mid \mu, \sigma) = \frac{1}{\sigma^n} \exp\left\{-\sum_{i=1}^{n} \frac{q_i - \mu}{\sigma}\right\} \exp\left\{-\sum_{i=1}^{n} \exp\left\{-\frac{q_i - \mu}{\sigma}\right\}\right\} \qquad (5.8)$$

Zoals reeds opgemerkt is de likelihoodfunctie zoals hierboven gedefinieerd een functie van de parameters van de verdeling. De stelling van Bayes gegeven in vergelijking (4.17) wordt nu gebruikt om tot een a posteriori kansdichtheidsfunctie te komen voor de parameters μ en σ, door de a priori kansdichtheidsfunctie $\pi(\mu,\sigma)$ te combineren met de waargenomen afvoeren $\mathbf{q} = (q_1,..., q_n)$. Het resultaat is:

$$\pi(\mu, \sigma \mid \mathbf{q}) = \frac{\pi(\mu, \sigma) l(\mathbf{q} \mid \mu,\sigma)}{\iint_{\mu,\sigma} \pi(\mu, \sigma) l(\mathbf{q} \mid \mu,\sigma)\, d\mu d\sigma} \qquad (5.9)$$

Hoe de a priori kansdichtheidsfunctie $\pi(\mu,\sigma)$ gekozen wordt, is een apart onderwerp en wordt in het volgende hoofdstuk uitvoerig behandeld.

Zoals reeds vermeld, is uitgegaan van gemeten afvoeren van 1901 t/m 1999. De meetperiode is relatief kort om te extrapoleren naar het extreme gebied. Dit is het gebied waarvoor geen waarnemingen beschikbaar zijn: met andere woorden extreme afvoeren met gemiddelde overschrijdingskansen van orde 10^{-3} en 10^{-4} per jaar. Het feit dat er weinig afvoerdata beschikbaar zijn, leidt tot onnauwkeurige en onzekere schattingen van de parameters μ en σ. Een model voor deze onzekerheid wordt gegeven door vergelijking (5.9).

Stel we willen de gemiddelde overschrijdingskans van een extreme afvoer q_0 (zeg q_0 = 16.000 m^3/s) bepalen, waarbij rekening wordt gehouden met de onzekerheid rondom μ en σ. De twee-dimensionale versie van vergelijking (5.2) biedt de mogelijkheid om dit te doen. Voor de Gumbelverdeling wordt die kans berekend door te integreren over alle mogelijke waarden van μ en σ die, voor de Gumbelverdeling, tot overschrijding van q_0 leiden. In formulevorm geschreven:

$$P(Q>q_0) = 1 - \iint_{\mu,\sigma} F(q_0|\mu,\sigma)\pi(\mu,\sigma|\mathbf{q})\, d\mu d\sigma = 1 - \iint_{\mu,\sigma} \exp\left\{-\exp\left\{-\frac{q_0-\mu}{\sigma}\right\}\right\}\pi(\mu,\sigma|\mathbf{q})d\mu d\sigma \qquad (5.10)$$

met $\pi(\mu,\sigma \mid \mathbf{q})$ gegeven in vergelijking (5.9).

De analyse van andere verdelingen, en natuurlijk ook van andere parameters, gaat op analoge wijze. Voor een tweede voorbeeld (met betrekking tot de exponentiële verdeling) zie hierna.

Voorbeeld 2 (exponentiële verdeling): Stel de onafhankelijke waarnemingen $\mathbf{x} = (x_1,..., x_n)$ vormen een steekproef uit de exponentiële verdeling met onbekende verwachting θ, $\theta > 0$. De Bayesiaanse schatter van de kans op overschrijden van het niveau x_0 is, gegeven de waarnemingen $\mathbf{x} = (x_1,..., x_n)$, te schrijven als:

$$P(X > x_0) = \int_{\theta=0}^{\infty} \exp\left\{-\frac{x_0}{\theta}\right\}\pi(\theta \mid \mathbf{x})\, d\theta \qquad (5.11)$$

met als a priori en a posteriori kansverdeling van θ respectievelijk $\pi(\theta)$ en

$$\pi(\theta \mid x) = \frac{\pi(x)l(x|\,\theta)}{\int_{\theta=0}^{\infty}\pi(\theta)l(x|\theta)d\theta} = \frac{\pi(\theta)\theta^{-n}\exp\{-\sum_{i=1}^{n} x_i / \theta\}}{\int_{\theta=0}^{\infty}\pi(\theta)\theta^{-n}\exp\{-\sum_{i=1}^{n} x_i / \theta\}d\theta} \qquad (5.12)$$

6 A priori kansverdelingen

6.1 Inleiding

In het voorbeeld van de twee vazen met een aantal rode en witte ballen zijn we ervan uitgegaan dat de twee vazen, voordat er ballen getrokken werden, even kansrijk zijn. De a priori kans op vaas V_1 is dus gelijk aan de kans op vaas V_2; daarbij is geen extra informatie gebruikt. We zeggen dat de a priori kansen *niet-informatief* zijn. Indien we een of andere a priori informatie, anders dan de getrokken ballen rood-rood-wit, hebben gebruikt om de a priori kansen op één van de vazen vast te stellen dan spreken we van een *informatieve* a priori kansverdeling.

Uit figuur 4.1 is duidelijk te zien dat de eerste stap in een Bayesiaanse analyse is het vaststellen van de a priori verdeling voor de onbekende modelparameters. Zoals hierboven reeds geïntroduceerd onderscheiden we daarbij twee typen a priori kansverdelingen: niet-informatieve en informatieve kansverdelingen.

6.2 Niet-informatieve a priori kansverdelingen

Bij een niet-informatieve a priori kansverdeling wordt geen a priori informatie meegenomen; de a priori onzekerheid is in dit geval ‘zeer groot’. Met andere woorden: er wordt verondersteld dat wij a priori, voordat de steekproef of metingen beschikbaar zijn, niets weten over de modelparameters (of de vazen in bovengenoemd voorbeeld). Voor het doel om zeer grote onzekerheid in probabilistische termen (in kansverdelingsfuncties) te definiëren, is een aantal niet-informatieve a priori kansverdelingen geschikt. Een voorbeeld van een niet-informatieve kansverdeling op een begrensd interval is de uniforme verdeling.

Voordat we een overzicht geven van de bekendste niet-informatieve a priori kansverdelingen, moet worden opgemerkt dat niet-informatieve a priori verdelingen eigenlijk niet bestaan. Hoe vaag een a priori kansverdeling ook is, er zit altijd wel enige informatie in verscholen. Een interessante discussie over het al dan niet bestaan van niet-informatieve a priori verdelingen is te vinden in een interview met Bernardo (Irony en Singpurwalla, 1997). In dit interview zegt hij dat ‘vaagheid’ van een a priori verdeling voornamelijk wordt bepaald door hoe dominant de waarnemingen zijn in vergelijking met de a priori informatie. Niet-informatieve a priori verdelingen hangen hierbij *niet* af van de waarnemingen zelf, maar wel van de beschouwde stochastische grootheid en het gekozen kansmodel (likelihood).

Zonder een volledig overzicht te geven, kunnen als niet-informatieve a priori kansverdelingen de volgende vier kansverdelingen toegepast worden:
- a priori uniforme verdeling (‘uniform prior’);
- a priori Jeffreys verdeling (‘Jeffreys’ prior’);
- a priori referentieverdeling (‘reference prior’);
- a priori maximale-data-informatie verdeling (‘maximal data information prior’).

Voor een uitgebreid literatuurovezicht over niet-informatieve a priori kansverdelingen zij verwezen naar Kass en Wasserman (1996).

Bovenstaande vier typen a priori verdelingen sluiten elkaar overigens niet uit. Zo kan de a priori Jeffreys verdeling gelijk zijn aan de referentieverdeling of de a priori maximale-data-informatie verdeling. Deze verdelingen blijken vaak oneigenlijk te zijn ('improper priors'), omdat ze niet tot één integreren en daardoor geen kansdichtheidsfunctie zijn. Eventuele oneigenlijkheid wordt in de a posteriori verdeling echter weer opgeheven, omdat de normalisatieconstante bij toepassing van de stelling van Bayes wegvalt (deze normalisatieconstante zit immers zowel in de noemer als de teller van vergelijking (4.17). Het is dan ook eigenlijk beter om te praten over een "maximaal informatieve a posteriori verdeling" dan over een "minimaal informatieve (niet-informatieve) a priori verdeling". Het grootste nadeel van niet-informatieve kansverdelingen is misschien wel dat er zoveel zijn (Berger, 1985, blz. 89).

6.2.1 A priori uniforme verdeling

Het meest recht toe recht aan is de aanname dat een bepaalde onzekere parameter uniform verdeeld is. Om ervoor te zorgen dat we te maken hebben met een kansdichtheidsfunctie (waarvan de integraal gelijk is aan één) dient de uniforme verdeling dan wel gedefinieerd te zijn op een *begrensd* interval. De aanname van een uniforme verdeling is indertijd gebruikt door zowel Bayes (1763) als Laplace (1812) voor het modelleren van 'vage' a priori informatie. Bij toepassing van een uniforme a priori kansverdeling van statistische parameters moet er verder voor worden gezorgd dat het gedefinieerde interval groot genoeg is. Indien een begrensde uniforme verdeling wordt toegepast voor het representeren van de onzekerheid in een onbegrensde parameter moet uiteraard worden voorkomen, dat een groot deel van de a posteriori kansmassa buiten het interval valt.

6.2.2 A priori Jeffreys verdeling

De bekendste methode voor het afleiden van niet-informatieve a priori verdelingen is de Jeffreys verdeling (Jeffreys, 1961, Hoofdstuk 3-4). Dit type niet-informatieve a priori verdeling is theoretisch gezien het best onderbouwd. Een voordeel van een a priori Jeffreys verdeling is dat de zo verkregen a priori verdeling invariant is onder één-op-één-transformaties. De aanname die ten grondslag ligt aan de Jeffreys verdeling is dat indien het aantal waarnemingen naar oneindig nadert, dat dan de a posteriori kansverdeling bij benadering een normale verdeling is. De één-dimensionale Jeffreys verdeling is gelijk aan de één-dimensionale a priori referentieverdeling. In een recent artikel betoogt Dawid (1999), dat de a priori Jeffreys verdeling als niet-informatieve verdeling de voorkeur verdient. Naast de invariantie onder transformatie heeft de Jeffreys verdeling het voordeel dat deze altijd dimensieloos is. Dit in tegenstelling tot andere mogelijke niet-informatieve a priori verdelingen, zoals de a priori referentieverdeling of de maximale-data-informatie verdeling. In dit verband zij opgemerkt, dat meerdimensionale Jeffreys verdelingen kunnen verschillen van de referentie- en de maximale-data-informatie verdeling. Voor een duidelijke uitleg over de a priori Jeffreys verdeling zij verwezen naar Box en Tiao (1973, Hoofdstuk 1).

6.2.3 A priori referentieverdeling

Een aangepaste vorm van de a priori Jeffreys verdeling is de zogenoemde referentieverdeling ('reference prior'), die uitvoerig beschreven is in Bernardo en Smith (1994). Het idee is dat een a priori kansverdeling wordt afgeleid waarvoor de informatie van de waarnemingen wordt gemaximaliseerd. Het verschil tussen de a priori Jeffreys verdeling en de a priori referentieverdeling is, dat de tweede in vergelijking met de eerste methode eventuele a priori

afhankelijkheid reduceert. Aangezien het te ver voert om de theorie van a priori referentieverdelingen hier volledig te behandelen, zal worden volstaan met de a priori verdeling in het één-dimensionale geval. Zoals hierboven reeds vermeld is in het één-dimensionale geval de a priori Jeffreys verdeling gelijk aan de referentieverdeling. Onder bepaalde voorwaarden (waaronder asymptotische normaliteit van de a posteriori verdeling) geldt dat de a priori kansverdeling van de parameter θ van een kansmodel $l(x \mid \theta)$ de volgende vorm heeft:

$$\pi(\theta) = \propto \{I(\theta)\}^{1/2} \text{ met } I(\theta) = -\int l(x \mid \theta)\left(\frac{\partial^2}{\partial\theta^2}\log l(x \mid \theta)\right)dx \qquad (6.1)$$

met $I(\theta)$ ook wel bekend als Fisher's informatie voor een enkele waarneming. Een voordeel van een a priori referentieverdeling is dat ze meestal invariant is onder één-op-één-transformaties van θ. Een nadeel is dat ze niet altijd dimensieloos is [voor een discussie, zie Dawid (1999)].

Voorbeeld (exponentiële verdeling): De a priori Jeffreys verdeling en de a priori referentieverdeling van de schaalparameter van een exponentiële verdeling met onbekende verwachting θ is:

$$\pi(\theta) \propto \{I(\theta)\}^{1/2} = \frac{1}{\theta} \text{ met}$$

$$I(\theta) = -\int_{x=0}^{\infty} \frac{1}{\theta}\exp\left\{-\frac{x}{\theta}\right\}\left(\frac{\partial^2}{\partial\theta^2}\left[-\log\theta - \frac{x}{\theta}\right]\right)dx = -\frac{1}{\theta^2} + \frac{2\theta}{\theta^3} = \frac{1}{\theta^2} \qquad (6.2)$$

6.2.4 A priori maximale-data-informatie verdeling

Een andere methode voor het afleiden van een niet-informatieve a priori verdeling is de maximale-data-informatie verdeling voorgesteld door Zellner (1977). Ook de methode van Zellner gaat uit van het maximaliseren van informatie afkomstig van waarnemingen, zij het dat hij uitgaat van een andere formulering namelijk:

$$\pi(\theta) = \propto \exp\{\hat{I}(\theta)\} \text{ met } \hat{I}(\theta) = \int l(x \mid \theta)\log l(x \mid \theta)\, dx \qquad (6.3)$$

Een nadeel van een a priori maximale-data-informatie verdeling is dat ze niet altijd invariant is onder één-op-één-transformaties van θ (in het voorbeeld van de exponentiële verdeling is de maximale-data-informatie verdeling overigens wel invariant onder transformaties). Zonder in te gaan op de wiskundige details verschillen de referentieverdeling en de maximale-data-informatie verdeling voornamelijk van elkaar in uitzonderlijke meer-dimensionale kansmodellen. Ook voor de parameter van een Bernoulli-verdeling verschillen beide typen niet-informatieve a priori verdelingen (zie voorbeeld Bernoulli-verdeling).

Voorbeeld (exponentiële verdeling): De a priori maximale-data-informatie verdeling van de schaalparameter van een exponentiële verdeling met onbekende verwachting θ is:

$$\pi(\theta) \propto \exp\{\hat{I}(\theta)\} \propto \frac{1}{\theta}$$

met

$$\hat{I}(\theta) = \int_{x=0}^{\infty} \frac{1}{\theta}\exp\left\{-\frac{x}{\theta}\right\}\left[-\log\theta - \frac{x}{\theta}\right]dx = -\log(\theta) - \frac{\theta}{\theta}$$

Voorbeeld (Bernoulli-verdeling): Een stochastische grootheid X heeft een Bernoulli verdeling met parameter θ ($0 < \theta < 1$) indien X alleen de waarden 0 en 1 kan aannemen. De kans op succes, dat wil zeggen de kans dat X gelijk is aan 1, is hierbij gelijk aan θ. Stel er zijn n onafhankelijke waarnemingen beschikbaar, waarvan n successen en n-r geen successen. De likelihood van deze waarnemingen is dan (gegeven de parameter θ) gedefinieerd als

$$l(\mathbf{x}|\,\theta) = l(x_1,\ldots,x_n \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^r\,(1-\theta)^{n-r}$$

De vier meest bekende niet-informatieve a priori verdelingen van de parameter θ zijn (Berger, 1985):

1. $\pi(\theta) = \theta^{\circ}(1-\theta)^{\circ} = 1$, a priori uniforme verdeling (uniform prior);
2. $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$, a priori oneigenlijke verdeling (improper prior);
3. $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, a priori Jeffreys- en referentieverdeling (Jeffreys- en reference prior);
4. $\pi(\theta) \propto \theta^{\theta}(1-\theta)^{1-\theta}$, a priori maximale-data-informatie verdeling ('maximal data information prior').

Een voordeel van toepassing van de eerste drie a priori kansverdelingen (uniform, oneigenlijk en Jeffreys/referentie) is dat de a posteriori kansverdeling van θ analytisch te bepalen is. Nadelen van de a priori oneigenlijke verdeling zijn, dat deze niet integreert tot één en dat zelfs de a posteriori verdeling oneigenlijk kan zijn (bijvoorbeeld bij nul waargenomen successen). Een voordeel van de a priori oneigenlijke verdeling van θ is echter wel dat de a posteriori schatting van het aantal te verwachten successen gelijk is aan de maximum-likelihood-schatter. Voor alle vier typen a priori kansverdelingen valt dus wat te zeggen. De a priori Jeffreys verdeling is theoretisch gezien het best onderbouwd, omdat zij invariant is onder één-op-één-transformaties van de beschouwde statistische parameter. Bovendien is de a priori Jeffreys verdeling altijd dimensieloos.

6.2.5 Keuze niet-informatieve a priori verdeling

Zonder in te gaan op de wiskundige details zij opgemerkt dat niet-informatieve a priori verdelingen ook kunnen worden afgeleid voor kansmodellen met meer dan één parameter. De meest geschikte methode is de a priori Jeffreys verdeling (Dawid, 1999).

6.4 Informatieve a priori kansverdelingen

Er wordt in dit geval wel a priori informatie meegenomen, namelijk beschikbare informatie voordat de steekproef genomen is of metingen verricht zijn. Een informatieve a priori verdeling voor de statistische parameters van de kansverdeling van het jaarmaximum van de Rijnafvoer zou bijvoorbeeld informatie kunnen bevatten over opgetreden afvoeren vóór 1901 in de vorm van gecensureerde waarnemingen zoals bijvoorbeeld voorgesteld door Stedinger & Cohn (1986). Hierbij is het jaar 1901 de begindatum van de periode waarover gemeten is en waarvoor betrouwbare afvoeren beschikbaar zijn. Er zijn verschillende mogelijkheden om die a priori informatie mee te nemen: via expertmeningen of een analyse van opgetreden data en informatie uit vergelijkbare situaties. In het geval van afvoeren zou immers ook informatie gebruikt kunnen worden van een andere rivier met dezelfde karakteristieken als de Rijn. Er moet wel op gelet worden dat de a priori kansverdeling onafhankelijk is van de waarnemingen waarmee de a posteriori kansverdeling bepaald wordt.

Tijdens de uitvoering van het project zal ruime aandacht besteed worden aan het vaststellen van niet-informatieve en informatieve a priori kansverdelingen.

6.5 Geconjugeerde a priori kansverdelingen

We hebben gezien dat toepassing van een Bayesiaanse analyse voor het vaststellen van een a posteriori kansverdeling resulteert in een complexe integraal die opgelost dient te worden. Helaas bestaat een analytische uitdrukking voor de a posteriori dikwijls niet en moet men zijn toevlucht nemen tot numerieke benaderingen. Echter, indien de a priori kansverdeling in combinatie met een bepaalde likelihoodfunctie van een bepaald type is, dan is het wel mogelijk om de a posteriori kansverdeling analytisch uit te drukken. We spreken in dit geval van *geconjugeerde* a priori verdelingen. Geconjugeerde a priori kansverdelingen kunnen zowel informatief als niet-informatief zijn. Als illustratief voorbeeld zal de kansverdeling van de verwachting van een exponentieel verdeelde variabele worden beschouwd. De familie van geïnverteerde gammaverdelingen vormt een geconjugeerde familie van verdelingen voor trekkingen uit een exponentiële verdeling met onbekende verwachting. Merk op dat als een stochastische grootheid een gammaverdeling heeft, dan heeft de reciproque van deze grootheid een geïnverteerde gammaverdeling.

Voorbeeld (exponentiële verdeling): Stel $x_1,\ldots, x_n$ zijn onafhankelijke waarnemingen uit een exponentiële verdeling met onbekende verwachting θ. We nemen verder aan dat de parameter θ ($\theta > 0$) als a priori verdeling een geïnverteerde gammaverdeling heeft met parameters $a > 0$ en $b > 0$. De likelihoodfunctie en de a priori kansdichtheidsfunctie van θ zijn respectievelijk:

$$l(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \frac{1}{\theta} \exp\left\{-\frac{x_i}{\theta}\right\} \tag{6.4}$$

en

$$\pi(\theta) = \frac{b^a}{\Gamma(a)}\ \theta^{-(a+1)} \exp\left\{-\frac{b}{\theta}\right\} \tag{6.5}$$

De a posteriori kansverdeling van θ kan dientengevolge geschreven worden als:

$$\pi(\theta \mid \mathbf{x}) = \frac{\frac{b^a}{\Gamma(a)}\ \theta^{-(a+n+1)}\ \exp\left\{-\frac{b + n\bar{x}}{\theta}\right\}}{\int_0^{\infty} \frac{b^a}{\Gamma(a)}\ \theta^{-(a+n+1)}\ \exp\left\{-\frac{b + n\bar{x}}{\theta}\right\} d\theta} \quad \text{met } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{6.6}$$

De laatste uitdrukking kan vervolgens herschreven worden als:

$$\pi(\theta \mid \mathbf{x}) = \frac{[b + n\bar{x}]^{a+n}}{\Gamma(a + n)}\ \theta^{-(a+n+1)} \exp\left\{-\frac{b + n\bar{x}}{\theta}\right\} \tag{6.7}$$

Dit is weer een geïnverteerde gammaverdeling, maar nu met parameters $a + n$ en $b + n\bar{x}$. Ook de a posteriori Bayesiaanse schatter van de kans op overschrijden van het niveau x_0 is in analytische vorm uit te drukken:

$$P(X > x_0) = \int_{\theta=0}^{\infty} \exp\left\{-\frac{x_0}{\theta}\right\} \pi(\theta \mid \mathbf{x})\, d\theta = \left[\frac{b}{b + n\bar{x} + x_0}\right]^{a+n} \tag{6.8}$$

Merk op dat de niet-informatieve a priori verdeling van de schaalparameter θ een limietgeval is van de geïnverteerde gammaverdeling en wel voor $a \to 0$ en $b \to 0$ zonder beschouwing van de normalisatieconstante.

Voorbeeld (Bernoulli-verdeling): Stel de stochastische grootheid X heeft een Bernoulli-verdeling met parameter θ, $0 < \theta < 1$. Stel er zijn n onafhankelijke waarnemingen beschikbaar, waarvan r successen en $n - r$ geen successen. We nemen vervolgens aan dat de a priori kansverdeling van θ een betaverdeling is met parameters $a > 0$ en $b > 0$ ofwel

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} \tag{6.9}$$

Het is eenvoudig in te zien dat de betaverdeling een geconjugeerde verdeling is voor Bernoulli-verdeelde waarnemingen. De a posteriori kansverdeling van θ is namelijk weer een betaverdeling:

$$\pi(\theta \mid \mathbf{x}) = \frac{\Gamma(a+b+n)}{\Gamma(a+r)\Gamma(b+n-r)} \theta^{a+r-1}(1-\theta)^{b+n-r-1} \tag{6.10}$$

met parameters $a + r$ en $b + n - r$. Merk verder op dat de drie niet-informatieve a priori kansverdelingen $\pi(\theta) = \theta^0(1 - \theta)^0$ ('uniform prior'), $\pi(\theta) \propto \theta^{-1}(1 - \theta)^{-1}$ ('improper prior') en $\pi(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2}$ ('Jeffreys prior' en 'reference prior') speciale of limietgevallen zijn van de geconjugeerde betaverdeling.

7 Numerieke integratiemethoden

7.1 Inleiding

Gegeven de a priori kansverdeling en de likelihoodfunctie die de beschikbare data beschrijft, kan de a posteriori kansverdeling bepaald worden. Alleen indien de a priori en de a posteriori verdeling geconjugeerd is met betrekking tot de likelihoodfunctie kan a posteriori analytisch worden uitgedrukt. In andere gevallen is het noodzakelijk om numerieke benaderingen toe te passen. Integralen als in vergelijkingen (4.17) en (5.3) dienen dan numeriek benaderd te worden.

De a posteriori kansverdeling heeft immers de volgende algemene gedaante:

$$\pi(\theta \mid \mathbf{x}) = \frac{l(\mathbf{x} \mid \theta)\pi(\theta)}{\int_{\theta} l(\mathbf{x} \mid \theta)\pi(\theta)\, d\theta} \tag{7.1}$$

met $l(\mathbf{x} \mid \theta)$ de likelihoodfunctie van de waarnemingen $\mathbf{x}$ gegeven θ en $\pi(\theta)$ de a priori kansverdeling van θ. De vector $\theta = (\theta_1,..., \theta_k)'$ is een vector van parameters van het kansmodel. De integraal in vergelijking (7.1) is vaak meerdimensionaal en heeft dezelfde dimensie als de parametrische vector θ. Het rechter lid van (7.1) is derhalve niet altijd analytisch op te lossen.

In dit hoofdstuk worden numerieke methoden behandeld die gebruikt kunnen worden voor het schatten van het rechterlid in vergelijking (7.1). De volgende numerieke integratiemethoden worden hierbij behandeld:
1. Laplace-benadering.
2. Gaussische kwadratuurformule.
3. Monte Carlo Importance Sampling.
4. Markov Chain Monte Carlo.

7.2 Laplace-benadering

Beschouw nu de volgende één-dimensionale integraal waarin geldt: $\pi(\theta)$ is de a priori verdeling van θ en $l(\mathbf{x} \mid \theta)$ is de likelihoodfunctie van de waarnemingen $\mathbf{x}$ gegeven θ:

$$I_1 = \int \pi(\theta) l(\mathbf{x} \mid \theta)\, d\theta \tag{7.2}$$

De integrand in vergelijking (7.2) kan herschreven worden als $\exp\{n\nu(\theta)\}$ waarvoor geldt:

$$\nu(\theta) = -\frac{1}{n}\left[\log\pi(\theta) + \log l(\mathbf{x} \mid \theta)\right] \tag{7.3}$$

Substitutie in vergelijking (7.2) levert:

$$I_1 = \int \exp\{-n\nu(\theta)\}\, d\theta \tag{7.4}$$

Introduceer nu de volgende functies:

$$-\nu(\hat{\theta}) = \sup_{\theta}\{-\nu(\theta)\} \text{ en } \hat{\sigma} = \left\{\frac{\partial^2 \nu(\theta)}{\partial \theta^2}\right\}^{-1/2}\Bigg|_{\theta=\hat{\theta}} \tag{7.5}$$

Voor grote waarden van *n* wordt de Laplace-benadering voor de integraal in vergelijking (7.2) dan gegeven door, zie Bernardo en Smith (1994, §5.5.1):

$$I_1 \approx \frac{\sqrt{2\pi}\hat{\sigma}}{\sqrt{n}} \exp\{-n\nu(\hat{\theta})\} \tag{7.6}$$

De Laplace-benadering zoals hierboven beschreven hoort tot de klasse van krachtige benaderingstechnieken binnen de statistiek. De methode kan voor grote *n* goed gebruikt worden voor het schatten van de a posteriori kansverdeling, gegeven de a priori kansverdeling $\pi(\theta)$ en de likelihoodfunctie $l(\mathbf{x} \mid \theta)$.

De Laplace-benadering kan ook worden toegepast om schattingen te maken van integralen zoals in vergelijking (5.3). Het probleem dat dan moet worden opgelost is van de vorm:

$$E(k(\theta) \mid \mathbf{x}) = \frac{I_2}{I_1} = \frac{\int k(\theta)\pi(\theta) l(\mathbf{x} \mid \theta)\, d\theta}{\int \pi(\theta) l(\mathbf{x} \mid \theta)\, d\theta} \tag{7.7}$$

Vergelijking (7.7) is een quotiënt van twee integralen. Gebruik makend van de benadering gegeven in (7.6) kan de integraal in vergelijking (7.7) benaderd worden door:

$$E(k(\theta) \mid \mathbf{x}) = \frac{I_2}{I_1} \approx \frac{\sigma^*}{\hat{\sigma}} \exp\{-n[\nu^*(\theta^*) - \nu(\hat{\theta})]\} = \hat{E}(k(\theta) \mid \mathbf{x}) \tag{7.8}$$

waarin, naast (7.5), geldt:

$$\nu^*(\theta) = -\frac{1}{n}\left[\log k(\theta) + \log \pi(\theta) + \log l(\mathbf{x} \mid \theta)\right]$$

$$-\nu^*(\theta^*) = \sup_{\theta}\{-\nu^*(\theta)\} \text{ en } \sigma^* = \left\{\frac{\partial^2 \nu^*(\theta)}{\partial \theta^2}\right\}^{-1/2}\Bigg|_{\theta=\theta^*} \tag{7.9}$$

Wat betreft de nauwkeurigheid van de benadering, kan er worden aangetoond dat:

$$E(k(\theta) \mid \mathbf{x}) = \hat{E}(k(\theta) \mid \mathbf{x}) \cdot [1 + O(n^{-2})] \tag{7.10}$$

De Laplace-benadering zoals hierboven toegepast op een probleem met één parameter kan worden uitgebreid naar problemen met meer dimensies (meer parameters); zie Bernardo en Smith (1994, §5.5.1). Neem aan dat de parameters van het beschouwde kansmodel zijn gegeven door de parametrische vector $\theta = (\theta_1, \ldots, \theta_k)$. De Laplace-benadering van de noemer van $E(k(\theta) \mid x)$ is gegeven door:

$$\int \exp\{-n\nu(\theta)\}\, d\theta = (2\pi)^{k/2} \left| n\nabla^2 \nu(\hat{\theta}) \right|^{-1/2} \exp\{-n\nu(\hat{\theta})\} \tag{7.11}$$

waarin $\hat{\theta}$ is gedefinieerd door

$$-\nu(\hat{\theta}) = \sup_{\theta}\{-\nu(\theta)\} \text{ en } [\nabla^2\nu(\hat{\theta})]_{ij} = \left.\frac{\partial^2\nu(\theta)}{\partial\theta_i\partial\theta_j}\right|_{\theta=\hat{\theta}} \tag{7.12}$$

de Hessiaanse matrix van h is geëvalueerd in $\hat{\theta}$. Een analoge uitdrukking bestaat er voor de teller van $E(k(\theta) \mid x)$, gedefinieerd in termen van υ^* en θ^*. Indien we schrijven dat:

$$\hat{\sigma} = \left|n\nabla^2\nu(\hat{\theta})\right|^{-1/2} \text{ en } \sigma^* = \left|n\nabla^2\nu^*(\theta^*)\right|^{-1/2} \tag{7.13}$$

dan is, geheel analoog aan het één-dimensionale geval, de Laplace-benadering voor grote n gegeven door:

$$E(k(\theta) \mid \mathbf{x}) \approx \frac{\sigma^*}{\hat{\sigma}} \exp\{-n[\nu^*(\theta^*) - \nu(\hat{\theta})]\} = \hat{E}(k(\theta) \mid \mathbf{x}) \tag{7.14}$$

Voor een heldere uitleg van de Laplace-methode zij overigens verwezen naar de klassieker van de Bruijn (1981, Hoofdstuk 4).

7.3 Gaussische kwadratuurformule

De algemene gedaante van de binnen een Bayesiaanse aanpak op te lossen integralen zijn van de vorm gegeven in vergelijking (7.2). In de vorige paragraaf is de Laplace-methode geïntroduceerd om zulke integralen te benaderen. Een andere benaderingsmethode is de Gaussische kwadratuur.

De methode is gebaseerd op de theorie van orthogonale functies in het algemeen en orthogonale polynomen in het bijzonder. Zonder verder in detail te gaan vermelden we een belangrijk punt uit die theorie, dat wil zeggen:

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} w(x)g(x)dx \approx \sum_{i=1}^{m} w(x_i)g(x_i) \tag{7.15}$$

De integraal in het linker lid van vergelijking (7.15) wordt, na opsplitsing van de functie f in het produkt van de functies w en g, benaderd door het gewogen optellen (met gewichtsfunctie w) van waarden van g in de punten x_i, $i = 1,..., m$. Deze punten zijn precies de nulpunten van Hermite-polynomen gegeven door:

$$H_{i+1}(x) = 2xH_i(x) - 2iH_{i-1}(x) \quad \text{met} \quad H_0(x) = 1 \text{ en } H_1(x) = 2x \tag{7.16}$$

De gewichtsfunctie $w(x)$ is gedefinieerd als $\exp\{-x^2\}$. De benadering gegeven in vergelijking (7.16) is exact in het geval dat f een polynoom is van graad $2n$-1. Zie verder Bernardo en Smith (1994, §5.5.2).

De kunst van het goed toepassen van de Gaussische kwadratuurformule is nu de keuze van de gewichtsfunctie, welke vergelijkbaar is met de keuze van de zogenoemde ‘importance function’ bij ‘importance sampling’ (zie de volgende paragraaf). In wezen moeten de verwachting μ en de standaard deviatie σ van de getransformeerde variabele $t = (\theta - \mu)/\sigma$ zodanig zijn dat de integraal zo efficiënt mogelijk wordt uitgerekend.

7.4 Monte Carlo Importance Sampling

Monte-Carlo-simulatie maakt gebruik van de mogelijkheid om randomgetallen te trekken uit een uniforme kansdichtheidsfunctie tussen 0 en 1. Gegeven het feit dat de onderschrijdingskans van een willekeurige stochastische variabele uniform verdeeld is tussen 0 en 1, kunnen random getallen uit een willekeurige kansverdelingsfunctie *F* worden gegenereerd. Dit gebeurt door de inverse van *F* uit te rekenen voor de waarden getrokken uit de uniforme verdeling.

Neem ter illustratie de integraal gegeven in vergelijking (7.2). Er worden in eerste instantie *m* trekkingen, zeg $u^{(1)},\ldots, u^{(m)}$, gedaan uit een uniforme verdeling tussen 0 en 1. Vervolgens worden *m* waarden uit de a priori kansdichtheidsfunctie $\pi(\theta)$ gegenereerd door de volgende formule toe te passen:

$$\theta^{(i)} = \Pi^{-1}(u^{(i)}) \text{ , } i = 1,\ldots,m \tag{7.17}$$

met $\Pi(\theta)$ de cumulatieve kansverdelingsfunctie van $\pi(\theta)$.

De integraal in vergelijking (7.2) wordt dan benaderd door:

$$I_1 = \int l(x \mid \theta)\pi(\theta)d\theta \approx \frac{1}{m}\sum_{i=1}^{m} l(x \mid \theta^{(i)}) \tag{7.18}$$

De Monte-Carlo-methode is zeer gemakkelijk toe te passen, maar heeft wel een aantal nadelen. Zo bestaat de inverse van de a priori kansverdelingsfunctie niet altijd en kan het aantal benodigde trekkingen zeer hoog zijn om redelijk nauwkeurige schattingen te krijgen.

Een methode om met voldoende nauwkeurigheid het aantal simulaties te reduceren is het toepassen van een variant binnen Monte-Carlo-simulatie: 'Monte Carlo Importance Sampling'. Het idee van de 'importance sampling' is dat de mogelijkheid wordt gecreëerd om trekkingen te doen uit een andere kansdichtheidsfunctie $s(\theta)$, de 'importance sampling function'. Hiertoe wordt de integraal in (7.2) herschreven als:

$$I_1 = \int l(x \mid \theta)\pi(\theta)d\theta = \int \frac{l(x \mid \theta)\pi(\theta)}{s(\theta)}\, s(\theta)d\theta \tag{7.19}$$

De laatste integraal in (7.19) is nu een uitdrukking voor de verwachtingswaarde van de functie $[l(x \mid \theta)\pi(\theta)]/s(\theta)$.

Net zoals bij de Gaussische kwadratuurformule is het ten behoeve van de efficiëntie nu de kunst om de 'importance function' zo te kiezen dat er relatief veel trekkingen vallen in het gebied waar de functiewaarden van $[l(x \mid \theta)\pi(\theta)]/s(\theta)$ relatief groot zijn. In wezen is het aan te bevelen dat de vorm van de functies $[l(x \mid \theta)\pi(\theta)]/s(\theta)$ en $s(\theta)$ redelijk met elkaar overeen komen. Dit slim kiezen vereist echter wel enig speurwerk (zie verder Bernardo en Smith, 1994, §5.5.3).

7.5 Markov Chain Monte Carlo

Een alternatieve Monte-Carlo-simulatiemethode is de Markov-Chain-Monte-Carlo-methode (MCMC-methode). Zonder in te gaan op de wiskundige details is het idee ruwweg als volgt. In plaats van trekkingen te doen uit de

a priori kansverdeling, worden trekkingen gedaan uit de a posteriori kansverdeling. Beschouw een kansmodel met parametrische vector $\theta = (\theta_1,..., \theta_k)$ en stel er zijn n waarnemingen $\mathbf{x} = x_1,..., x_n$ beschikbaar. Als basis voor de MCMC-methode merken we op dat de conditionele kansdichtheidsfuncties:

$$p(\theta_i \mid \mathbf{x}, \theta_j, j \neq i), \; i = 1,...,k \tag{7.20}$$

vaak relatief eenvoudig te bepalen zijn door bestudering van $p(\theta \mid \mathbf{x}) \propto p(\mathbf{x} \mid \theta)p(\theta)$. Stel we beschikken over een willekeurige reeks startwaarden

$$\theta^{(0)} = (\theta_1^{(0)},..., \theta_k^{(0)}) \tag{7.21}$$

van de onzekere statistische parameters. Vervolgens implementeren we de volgende iteratieve procedure:

$$\begin{aligned}
&\text{trek } \theta_1^{(1)} \text{ uit } p(\theta_1 \mid \mathbf{x}, \theta_2^{(0)},..., \theta_k^{(0)}),\\
&\text{trek } \theta_2^{(1)} \text{ uit } p(\theta_2 \mid \mathbf{x}, \theta_1^{(1)}, \theta_3^{(0)},..., \theta_k^{(0)}),\\
&\text{trek } \theta_3^{(1)} \text{ uit } p(\theta_3 \mid \mathbf{x}, \theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)},..., \theta_k^{(0)}),
\end{aligned} \tag{7.22}$$

$$\begin{aligned}
&...\\
&\text{trek } \theta_k^{(1)} \text{ uit } p(\theta_k, \mid \mathbf{x}, \theta_1^{(1)},..., \theta_{k-1}^{(1)}),\\
&\text{trek } \theta_1^{(2)} \text{ uit } p(\theta_1, \mid \mathbf{x}, \theta_2^{(1)},..., \theta_k^{(1)}),\\
&...\\
&\text{enzovoort.}
\end{aligned}$$

Voer bovenstaande procedure uit gedurende t iteraties en voer deze procedure m keer onafhankelijk van elkaar uit. We hebben dan dus m trekkingen van de vector $\theta^t = (\theta_1^{(t)},..., \theta_k^{(t)})$, waarbij θ^t een realisatie is van de Markov-keten met overgangskansen die zijn gegeven door

$$p(\theta^t, \; \theta^{t+1}) = \prod_{l=1}^{k} p(\theta_l^{t+1} \mid \mathbf{x}, \; \theta_j^t, \; j > l, \theta_j^{t+1}, \; j < l) \tag{7.23}$$

Dan kan er wiskundig worden bewezen dat θ^t voor $t \to \infty$ in verdeling nadert naar een stochastische grootheid met kansdichtheidsfunctie $p(\theta \mid \mathbf{x})$.

Op basis van deze MCMC-simulatie kunnen nu onder meer marginale a posteriori kansverdelingen en Bayesschatters worden benaderd. Bovenstaande methode is een variant van MCMC-simulatie en luistert naar de naam ‘Gibbs sampling’. Voor meer details over MCMC zie bijvoorbeeld Bernardo en Smith (1994, §5.5.4), Gelman *et al.* (1995, Hoofdstuk II) en Carlin en Louis (2000, Hoofdstuk 5).

8 Bayesiaanse schatting van het verdelingstype

8.1 Inleiding

De commissie Boertien heeft de maatgevende afvoer van de Rijn bij Lobith bepaald uit een *ongewogen* middeling van een aantal op jaarmaxima gefitte kansverdelingen. Dit betekent dat elk van deze gefitte kansverdelingen eenzelfde (subjectieve) kans heeft dat zij juist is; derhalve zijn de gehanteerde weegfactoren aan elkaar gelijk.

Toepassing van de Bayesiaanse statistiek stelt ons echter in staat om een *gewogen* middeling te hanteren door gebruikmaking van de zogenaamde *Bayesiaanse factoren* of *Bayes factors*. Deze factoren kunnen worden toegepast om, uitgaande van a priori kansen dat bepaalde hypothetische kansverdelingen juist zijn, de a posteriori kansen te bepalen wanneer waarnemingen beschikbaar zijn. In dit proces staan de historische gegevens (data) weer centraal. In het navolgende wordt beschreven hoe de Bayes factoren bepaald kunnen worden.

8.2 Bayes factoren

Zoals reeds vermeld, kent de traditionele analyse toetsen voor het al dan niet 'verwerpen' van kansverdelingen. Er wordt eerst een hypothese gedaan dat een bepaalde stochastische grootheid een zekere kansverdeling heeft. Uitgaande van die hypothese wordt vervolgens de kritieke waarde voor een gegeven toetsingsgrootheid en een onbetrouwbaarheid α berekend. Opgemerkt dient te worden dat de onbetrouwbaarheid alleen iets zegt over een verwerping ten onrechte en niets zegt over de kans dat die kansverdeling juist is. Bovendien zijn de toetsen voor aanpassing puur gebaseerd op statistische gegevens en houden deze geen rekening met eventuele andere kennis die relevant is voor de keuze van het verdelingstype. Mede omdat in de praktijk het waarnemingsmateriaal beperkt is, kunnen meerdere kansverdelingen door de toetsing komen. Door de onbetrouwbaarheid steeds kleiner te kiezen, kan nagegaan worden wanneer een bepaalde kansverdeling afvalt of meerdere kansverdelingen afvallen. Langs die weg zou men gebruik makend van de 'maximum likelihood'-aanpak uiteindelijk tot een oordeel kunnen komen.

Deze methode is anders dan de Bayesiaanse methode voor verdelingskeuze. Bij de Bayesiaanse aanpak selecteert men vooraf een aantal kansverdelingen (die bijvoorbeeld niet verworpen zijn of kunnen worden), kent daaraan a priori kansen toe en bepaalt vervolgens op grond van de waarnemingen en de stelling van Bayes de a posteriori kansen. De selectie van verdelingstypen en de a priori kansen kan deels subjectief zijn.

Procedureel is er geen onderscheid tussen de Bayesiaanse parameterschatting en de procedure voor de keuze van het verdelingstype. Beide methodes zijn in elkaar verweven.

De Bayesiaanse procedure voor de keuze van het verdelingstype kan als volgt samengevat worden:

1. Stel een aantal hypothesen op, bijvoorbeeld:
   H_1 : de betreffende stochastische grootheid is Gumbel verdeeld;
   H_2 : de betreffende stochastische grootheid is lognormaal verdeeld;
   H_3 : de betreffende stochastische grootheid is Gamma verdeeld.
2. Kies de a priori kansen $p(H_i)$, i = 1,2,3.
3. Bepaal de likelihoodfunctie van de waarnemingen **x**: $\pi(\mathbf{x} \mid H_i)$, i = 1,2,3.
4. Bereken de a posteriori kansen $p(H_i|\mathbf{x})$, i = 1,2,3.

Verondersteld worden de waarnemingen $\mathbf{x} = (x_1,..., x_n)$, die onderling onafhankelijk en identiek verdeeld zijn. Wanneer n oneindig is en de 'echte' kansverdeling deel uitmaakt van de beschouwde k kansverdelingen, zal de a posteriori kans op deze verdeling naar 1 naderen. In de praktijk is echter ten eerste het aantal waarnemingen beperkt en is ten tweede de 'echte' kansverdeling van afvoeren bijvoorbeeld onbekend. Bijgevolg kunnen op basis van statistische toetsen, meerdere kandidaat-verdelingen worden geselecteerd. Statistische toetsen zeggen immers niets over de kans dat een specifieke verdeling de meest waarschijnlijke is. Bayesiaanse statistiek doet dat in principe wel.

Zij H_i, i = 1,..., m, een aantal kansverdelingen volgens welke de waarnemingen **x** (waarschijnlijk) verdeeld kunnen zijn. We kennen aan H_i een (subjectieve) a priori kans $p(H_i)$ toe dat deze verdeling inderdaad de juiste is. Het beginsel van de Bayesiaanse analyse kan nu benut worden opdat de beschikbare data de zogenoemde a posteriori kans $p(H_i \mid \mathbf{x})$ produceren: dat wil zeggen de kans dat H_i juist is gegeven de waarnemingen **x**. Gebruikmakend van de stelling van Bayes kunnen deze a posteriori kansen als volgt worden geschreven:

$$p(H_i \mid \mathbf{x}) = \frac{\pi(\mathbf{x} \mid H_i)p(H_i)}{\sum_{j=1}^{m} \pi(\mathbf{x} \mid H_j)\mathrm{p}(H_i)} \quad , \quad i = 1,\ldots,m, \tag{8.1}$$

waarbij de a priori kansen uiteraard dienen te voldoen aan:

$$\sum_{i=1}^{m} p(H_i) = 1 \tag{8.2}$$

Om het berekenen van deze kansen enigszins overzichterlijker te maken, wordt gekeken naar de verhouding tussen twee verschillende a posteriori kansen, zeg $p(H_i \mid \mathbf{x})$ en $p(H_j \mid \mathbf{x})$. Dit resulteert in:

$$\frac{p(H_i \mid \mathbf{x})}{p(H_j \mid \mathbf{x})} = \frac{\pi(\mathbf{x} \mid H_i)}{\pi(\mathbf{x} \mid H_j)} \times \frac{p(H_i)}{p(H_j)} \tag{8.3}$$

De eerste term in het rechterlid van vergelijking (8.3) noemt men de Bayes factor, welke wordt genoteerd als

$$B_{ij} = \frac{p(\mathbf{x} \mid H_i)}{p(\mathbf{x} \mid H_j)} \tag{8.4}$$

Als functie van de Bayesfactoren kunnen de a posteriori kansen, na enig rekenwerk, als volgt worden herschreven:

$$p(H_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid H_i)p(H_i)}{\sum_{j=1}^{m} p(\mathbf{x} \mid H_j)p(H_j)} = \frac{\alpha_i B_{i1}}{\sum_{j=1}^{m} \alpha_j B_{j1}}; \quad \alpha_i = \frac{p(H_i)}{p(H_1)} \tag{8.5}$$

Elke kansverdeling H_i heeft een aantal onzekere statistische parameters $\theta = (\theta_1,..., \theta_k)$, zodat de eerste term in de teller van vergelijking (8.1) wordt gegeven door:

$$\pi(\mathbf{x} \mid H_i) = \int l(\mathbf{x} \mid \theta, H_i)\pi(\theta \mid H_i)d\theta \qquad (8.6)$$

Hierin is $l(\mathbf{x} \mid \theta, H_i)$ de likelihoodfunctie behorende bij kansverdeling H_i en is $\pi(\theta \mid H_i)$ de a priori kansdichtheidsfunctie van de statistische parameters θ.

In het geval van een informatieve, eigenlijke, a priori verdeling zullen er bij het uitrekenen vaak geen echte problemen optreden. Eén van de rekentechnieken behandeld in hoofdstuk 5 (zoals de Laplace-benadering, Gaussische kwadratuur, Monte Carlo Importance Sampling en MCMC) kan hierbij worden gebruikt. Is de a priori verdeling echter een niet-informatieve, oneigenlijke verdeling dan behoeft de integraal in vergelijking (8.6) helaas niet altijd te bestaan. Een mogelijke oplossing is dan om niet de kans op de waarnemingen te benaderen, maar de Bayesfactoren. Een veel toegepaste benaderingsmethode is in dit verband het zogenoemde Schwarz-criterium, zie Schwarz (1978):

$$-2\log B_{12} \approx -2\log \left[\frac{\pi(\mathbf{x} \mid H_1, \hat{\theta}_1)}{\pi(\mathbf{x} \mid H_2, \hat{\theta}_2)}\right] - (k_2 - k_1)\log n \qquad (8.7)$$

met $\hat{\theta}_i$ de maximum-likelihoodschatter voor kansmodel H_i, k_i het aantal parameters van kansmodel H_i en n het aantal waarnemingen. In dit criterium fungeert de term $(k_2 - k_1) \log n$ als een soort correctieterm, die corrigeert voor de verschillende aantallen statistische parameters van de bestudeerde kansmodellen. Voordelen van het Schwarz-criterium zijn dat er gebruikt kan worden gemaakt van standaarduitvoer van klassiek-statistische software (maximum-likelihoodschatters) en dat het niet afhangt van de a priori kansverdeling (wat natuurlijk ook als een nadeel kan worden gezien). Ondanks het feit dat de ralatieve fout in de, met behulp van het Schwarz-criterium benaderde, Bayesfactor in het algemeen een nauwkeurigheid heeft van $O(1)$, blijkt de benadering in de praktijk toch vrij goed te werken. Het Schwarz-criterium wordt dan ook vaak toegepast.

Voor meer details over de Bayes factoren zij verwezen naar Kass en Raftery (1995) en Bernardo en Smith (1994, Hoofdstuk 6).

8.3 Tanggewichten

Een andere methode voor het bepalen van gewichten van kansverdelingen is voorgesteld door Tang (1980). Deze methode is gebaseerd op een Bayesiaanse lineaire regressie van waarnemingen die (na transformatie) zijn geplot op ‘probability paper’. Op basis van discrepanties tussen de waarnemingen enerzijds en de modelvoorspellingen (gefitte kansverdeling) anderzijds worden allereerst de verwachting en de variantie van een bepaalde ontwerpgrootheid (bijvoorbeeld maatgevende afvoer) bepaald. Vervolgens worden deze, per kansmodel berekende, verwachtingen en varianties gecombineerd tot gewichten van kansmodellen.

Stel $E(Y_1)$ en $\mathrm{Var}(Y_1)$ zijn de verwachting en variantie van de ontwerpgrootheid gebruik makend van het eerste kansmodel, terwijl $E(Y_2)$ en $\mathrm{Var}(Y_2)$ de verwachting en variantie zijn van de ontwerpgrootheid volgens het tweede kansmodel. Beide modellen worden onafhankelijk verondersteld. Volgens

Tang kan de gecombineerde verwachting van de ontwerpgrootheid van twee onafhankelijke kansmodellen worden geschreven als:

$$E(X_2) = \frac{\text{Var}(Y_2)E(Y_1) + \text{Var}(Y_1)E(Y_2)}{\text{Var}(Y_2) + \text{Var}(Y_1)} \tag{8.8}$$

en de gecombineerde variantie van de ontwerpgrootheid als:

$$\text{Var}(X_2) = \frac{\text{Var}(Y_2)\text{Var}(Y_1)}{\text{Var}(Y_2) + \text{Var}(Y_1)} \tag{8.9}$$

Het is gemakkelijk in te zien dat de gecombineerde schatting $E(X_2)$ naar $E(Y_1)$ als $\text{Var}(Y_1)$ extreem klein is in vergelijking met $\text{Var}(Y_2)$. Met andere woorden, als het eerste model resulteert in een excellente fit en het tweede model niet, dan wordt de informatie uit het tweede model genegeerd.

Uit het bovenstaande lijkt het alsof de stelling van Bayes niet is toegepast. We zullen in het navolgende uitleggen dat de stelling van Bayes impliciet wel een rol heeft gespeeld bij de definitie van de Tanggewichten. Gebruik makend van de stelling van Bayes, kunnen de vergelijkingen (8.8)-(8.9) als volgt worden bepaald. We beschouwen een steekproef uit een normale verdeling met een onzekere (onbekende) verwachting en zekere (bekende) variantie. Stel dat de a priori kansverdeling van de onzekere verwachting ook een normale verdeling heeft, dan is de a posteriori kansverdeling eveneens een normale verdeling. Met andere woorden: de normale verdeling is een geconjugeerde verdeling voor trekkingen uit een normale verdeling met onbekende verwachting en bekende variantie. De parameters van de a posteriori verdeling van de onzekere verwachting zijn analoog aan vergelijkingen (8.8)-(8.9). In wezen kunnen het eerste en het tweede model worden beschouwd als respectievelijk de a priori en de waargenomen informatie. De a posteriori verwachting is zo een gewogen gemiddelde van de a priori verwachting en de steekproefverwachting, omgekeerd gewogen op basis van de respectieve varianties. Voor details, zie Ang en Tang (1975, Hoofdstuk 8).

Alhoewel Tang in zijn artikel alleen de gewichten heeft gepresenteerd ten behoeve van het vergelijken van ontwerpgrootheden op basis van slechts twee kansmodellen, kunnen zij gemakkelijk worden gegeneraliseerd voor n kansmodellen waarbij $n \geq 2$. Indien (8.8) en (8.9) namelijk gelden voor alle mogelijke paren van n onafhankelijke kansmodellen, dan kan zowel de verwachting $E(X_n)$ als de variantie $\text{Var}(X_n)$ worden afgeleid door volledige inductie te gebruiken. Voor dit resultaat zij verwezen naar Van Gelder *et al.* (1999).

Met betrekking tot de methode van Tang voor het toekennen van gewichten aan kansverdelingen kan worden gesteld, dat de methode in de aard van de zaak geen Bayesiaanse analyse. Zo worden de parameters van kansmodellen gefit met behulp van klassieke statistiek; er wordt geen onzekerheid in de parameters meegenomen en de stelling van Bayes wordt niet gebruikt voor het bijwerken van a priori informatie met waarnemingen. Op basis van de discrepantie tussen de waarnemingen en de voorspelling zijn vervolgens gewichten gedefinieerd, waarbij gebruik is gemaakt van een wiskundig trucje (welk trucje overigens wel volgt uit een Bayesiaanse berekening). Daarnaast is op basis van een Monte-Carlo-simulatie gebleken dat de Bayesfactoren sneller convergeren naar de juiste kansverdeling dan de Tanggewichten (zie Van Gelder *et al.* 1999). In het project Bayesiaanse statistiek zal de aandacht zich zodoende voornamelijk op de Bayesfactoren richten.

8.4 Software ter bepaling van Bayesschatters en -factoren

Wat betreft het gecombineerd bepalen van Bayesschatters en Bayesfactoren (gewichten kansverdelingen) is voor zover wij weten nog geen software beschikbaar. Wel is er op de website van de International Society of Bayesian Analysis (ISBA) een aantal publicaties en software ten behoeve van andere toepassingen te vinden (zoals bijvoorbeeld Gegeneraliseerde Lineaire Modellen, Proportional Hazard Models, Linear Regression en Logistic Regression). Het adres van deze ISBA-website is

"http://www.research.att.com/~volinsky/bma.html".

De rekentechnieken die, indien er geen geconjugeerde kansverdelingen te vinden zijn, het meest toegepast worden, zijn de Laplace-benadering, Markov Chain Monte Carlo (MCMC) en (voor Bayesfactoren) het Schwarz-criterium. Een veel gebruikt softwarepakket voor het uitvoeren van Bayesiaanse analyse is het programma BUGS (Bayesian inference Using Gibbs Sampling). Gibbs sampling is een vorm van de MCMC-simulatiemethode. Voor informatie over BUGS zie:

"http://www.mrc-bsu.cam.ac.uk/bugs/Welcome.shtml".

Voor algemene informatie over Bayesiaanse statistiek is verder de website "Bayesian sites" interessant:

"http://www.isds.duke.edu/sites/bayes.html".

Literatuur

Ang, A.H.S., en W.H. Tang (1995). *Probability Concepts in Engineering Planning and Design*, Volume 1, New York: John Wiley & Sons.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. P*hil. Trans. Roy. Soc. London*, 53, 370-418.

Benjamin, J.R. and Cornell, C.A. (1970), *Probability, Statistics and Decision for Civil Engineers*. McGraw-Hill, Inc.

Bernardo, J.M., en A.F.M. Smith (1994). *Bayesian Theory*. Chichester: John Wiley & Sons.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis; Second Edition*. New York: Springer-Verlag.

Box, G.E.P., and G.C. Tiao (1973). *Bayesian Inference in Statistical Analysis*. New York: John Wiley and Sons.

Bruijn, N.G. de (1981). *Asymptotic Methods in Analysis*. New York: Dover Publications Inc.

Carlin, B.P. and Louis, T.A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall/CRC.

Castillo, E. (1988), *Extreme Value Theory in Engineering*. Academic Press, Inc.

CUR, *Kansen in de civiele techniek, deel 1: probabilistisch ontwerpen in theorie en praktijk*, 1997.

Dawid, A.P. (1999). The trouble with Bayes factors. Research Report No. 202, *Department of Statistical Science, University College London*.

Gelder, P.H.A.J.M. van, J.M. van Noortwijk, and M.T. Duits. (1999). Selection of probability distributions with a case study on extreme Oder river discharges. In G.I. Schuëller and P. Kafka, editors, *Safety and Reliability, Proceedings of ESREL '99 - The Tenth European Conference on Safety and Reliability*, Munich-Garching, Germany, 1999, pages 1475-1480. Rotterdam: Balkema.

Gelder, P.H.A.J.M. van (1999), *Statistical Methods for the Risk-Based Design of Civil Structures*. Ph-D thesis, Delft University of Tecnology, Delft.

Gelman, A., Carlin J.B., Stern, H.S. and Rubin D.B. (1995), *Bayesian Data Analysis*. London Chapman & Hall.

Irony, T.Z., en N.D. Singpurwalla (1997). Noninformative priors do not exist: A dialogue with José M. Bernardo. *Journal of Statistical Planning and Inference*, 65:159-189.

Jeffreys, H.J. (1961). *Theory of Probability; Third Edition*. Oxford: Clarendon Press.

Kass, R.E., en A.E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773-795.

Kass, R.E., and L. Wasserman (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343-1370.

Kok, M., N. Douben, J.M. van Noortwijk en W. Silva (1996). *Integrale Verkenning inrichting Rijntakken; Rapportnummer 12*: Veiligheid. Arnhem: Ministerie van Verkeer en Waterstaat.

Laplace, P.S. (1812). *Théorie Analytique des Probabilités*. Paris: Courcier.

Ministerie van verkeer en Waterstaat (Commissie Boertien, 1993), *Toetsing Uitgangspunten rivierdijkversterkingen, deelrapport 2*, Waterloopkunig Laboratorium, WL, en European-American Center for Policy Analysis, EAC RAND.

Schwarz, G. (1978), Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461-464.

Stedinger, J.R., and T.A. Cohn (1986). Flood frequency analysis with historical and paleoflood information. *Water Resources Research*, 22(5):785-793.

Tang, W.H. (1980). Bayesian frequency analysis. Journal of the Hydraulics Division, Vol. 106, No. HY7, pp. 1203-1218.

Zellner, A. (1977). Maximal data information prior distributions. In A. Aykac en C. Brumat, editors, New *Methods in the Applications of Bayesian Methods*, pp. 211-232. Amsterdam: North-Holland.