 NOTA 724 maart 1973

Instituut voor Cultuurtechniek en Waterhuishouding
Wageningen

HET GEBRUIK VAN DE NEGATIEF-EXPONENTIELE KANS-VERDELING VOOR AFVOERGEGEVENS EN VOOR HET BEPALEN VAN DE MAATGEVENDE AFVOER

DEEL I

THEORETISCHE ACHTERGRONDEN

ir Ph.Th. Stol

INHOUD

1. INLEIDING

Overschrijdingsfrequenties van afvoergegevens kunnen vaak goed met de negatief-exponentiële kansverdeling worden weergegeven. De theorie omtrent deze verdeling komt op verschillende plaatsen verspreid in de literatuur voor. Een aantal eigenschappen van deze verdeling zullen daarom in hun onderlinge verband in deze nota worden toegelicht en besproken. De theorie wordt daarna op een praktijkvraagstuk toegepast.

Achtereenvolgens komen de volgende vragen aan de orde:

- wat zijn de theoretische eigenschappen van de exponentiële kansverdeling,
- wanneer mag de exponentiële kansverdeling worden toegepast,
- indien de exponentiële kansverdeling van toepassing is, wat is dan de betrouwbaarheid van de uitkomsten,
- welke conclusies kunnen aan een aan waarnemingen aangepaste exponentiële kansverdeling worden ontleend, speciaal met betrekking tot een maatgevende afvoer.

2. NOTATIES EN BEGRIPPEN

Kortheidshalve zal over de exponentiële verdeling zonder meer worden gesproken.

Zoals gebruikelijk in statistische beschouwingen zullen stochastische (kans-)variabelen door onderstreping van het betrokken symbool worden weergegeven. Bijvoorbeeld:

$$P(\underline{x} > x) = p_1 \qquad (2.1)$$

wat betekent dat de kans P dat de kansvariabele $\underline{x}$ een waarde aanneemt groter dan een vast getal x, gelijk is aan p_1. De waarden die $\underline{x}$ in een steekproef kan aannemen worden realisaties genoemd. De uitdrukking

$$\underline{x} = x_i$$

betekent dan dat in de i-de waarneming de kansvariabele $\underline{x}$ de waarde x_i heeft aangenomen, of ook x_i is de i-de realisatie van $\underline{x}$.

Wordt een vaste overschrijdingskans p_1 beschouwd, dan wordt de waarde van x die hiermede correspondeert de kritieke of drempelwaarde genoemd. Voor afvoeren is het gebruikelijk dan over de maatgevende afvoer te spreken.

Verder geldt

$$P(\underline{x} \geq x) = 1 - P(\underline{x} < x)$$

in het kort

$$P_{\geq} = 1 - P_{<} \qquad (2.2)$$

Formule (2.1) is slechts een principe formule. Een functie die aan de te overschrijden waarden x_i de kans p_i toevoegt wordt een kansverdeling genoemd.

3. DE EXPONENTIELE KANSVERDELING

De exponentiële kansverdeling berust op een exponentiële afname van de kansdichtheidsfunctie. De algemene vorm hiervan luidt:

$$dP(\underline{x} < x) = ae^{-\lambda x} \text{ , } x \geq 0 \text{ , } \lambda > 0 \qquad (3.1)$$

Voorts moet hierbij worden aangenomen dat

$$P(\underline{x} < 0) = 0$$

Integratie van (3.1) geeft de kansverdeling zelf. Over het waardenbereik van x moet de integraal de waarde 1 opleveren want

$$P(0 \leq \underline{x} < +\infty) = 1$$

zodat

$$\int_0^\infty ae^{-\lambda x}\, dx = 1$$

of

$$\left[-\frac{a}{\lambda} e^{-\lambda x}\right]_0^\infty = \frac{a}{\lambda} = 1$$

De parameter a heeft dus de waarde λ, en de dichtheidsfunctie is

$$dP(\underline{x} < x) = \lambda e^{-\lambda x} \tag{3.2}$$

De kansverdeling luidt dus

$$P(\underline{x} < x) = \int_0^\infty \lambda e^{-\lambda u}\, du$$

$$= 1 - e^{-\lambda x} \tag{3.3}$$

Wordt met overschrijdingskansen gewerkt, dan volgt uit (3.3), aangezien $P_{\geq} = 1 - P_{<}$ volgens (2.2),

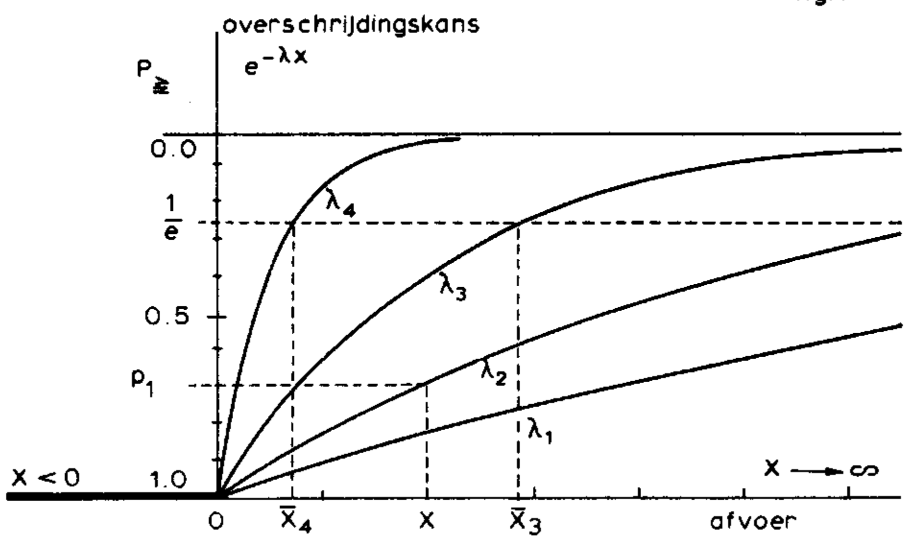
$$P(\underline{x} \geq x) = e^{-\lambda x} \tag{3.4}$$

zie fig. 1.

Uit deze laatste vorm volgt het voodeel van gebruik van half-logarithmisch grafiekpapier, namelijk

$$\log P = \log e^{-\lambda x} = -(0.4343)\, \lambda x \tag{3.5}$$

Door dus de logarithme van de kans uit te zetten tegen de afvoer x lineair, ontstaat een rechte met negatieve helling 0.4343 λ. Door fig.1



cumulatieve kans op lineaire schalen

$P(-\infty < \underline{x} < 0) = 0$; $P(\underline{x} \geqq 0) = 1$; $P(\underline{x} \geqq x) = p_1$

voor λ geldt $\lambda_1 < \lambda_2 < \lambda_3 < \lambda_4$ dus $\bar{x}_1 > \bar{x}_2 > \bar{x}_3 > \bar{x}_4$

bij een kansniveau p_1 de waarde van x in de grafiek af te lezen kan λ berekend worden uit

$$\lambda = \frac{-\log p_1}{(0.4343)\ x} \tag{3.6}$$

Omgekeerd kan bij bekende waarde van λ de bijbehorende kritieke waarde van x bij een kansniveau p_1 gevonden worden met

$$x = \frac{-\log p_1}{(0.4343)\ \lambda}$$

Door voor p_1 bijvoorbeeld 1 % te kiezen wordt dit

$$x = \frac{4.605}{\lambda} \tag{3.7}$$

Hiermede kan een tweede punt gevonden worden om bij gegeven λ de curve te kunnen tekenen. Een eerste punt is steeds $(x, p) = (0,1)$.

4. DE VOORWAARDELIJKE EXPONENTIELE VERDELING

De exponentiële verdeling heeft een eigenschap die de verdeling bij uitstek geschikt maakt voor het toepassen op hoge afvoeren.

Beschouw hoge afvoeren groter dan een gefixeerd minimum x_o (fig. 2). Gezocht wordt de kans dat onder deze voorwaarde een realisatie van $\underline{x}$ de waarde $x \geqq x_o$ overschrijdt.

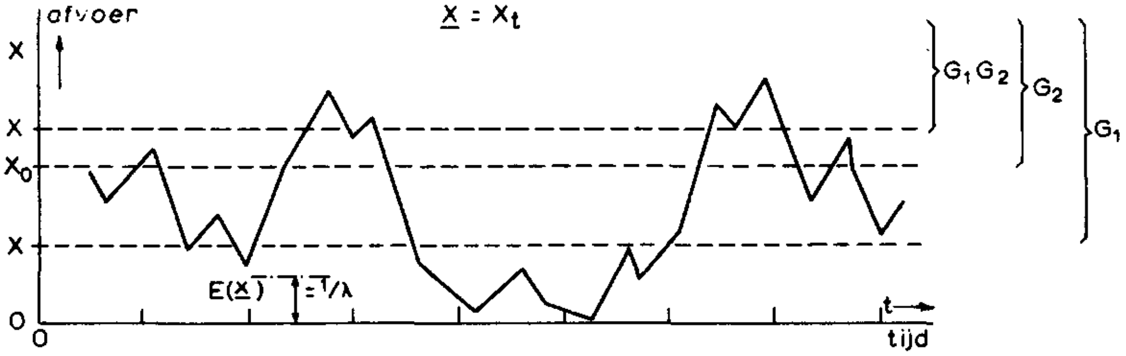
De volgende stochastische gebeurtenissen worden gedefinieerd

$$\left.\begin{aligned} G_1 &= (\underline{x} \geqq x) \\ G_2 &= (\underline{x} \geqq x_o) \end{aligned}\right\} \quad x \geqq x_o \tag{4.1}$$

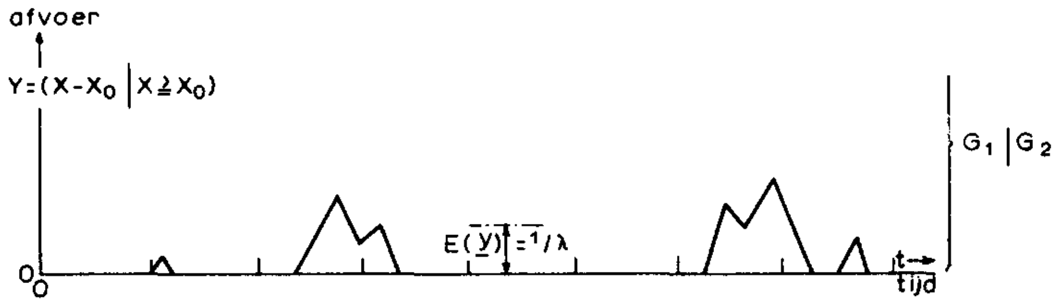
Het gelijktijdig optreden van beide gebeurtenissen wordt voorgesteld door G_1G_2.

De gebeurtenis $G_1|G_2$ is de gebeurtenis dat G_1 optreedt o n d e r d e v o o r w a a r d e G_2, hetgeen betekent dat alleen de gebeurtenis G_2 van belang geacht wordt en dat nagegaan wordt welke rol G_1

fig. 2



voorbeeld oorspronkelijke afvoerreeks $\underline{X} \geqq 0$



voorbeeld voorwaardelijke afvoerreeks $\underline{Y} \geqq 0$

gedefinieerd is $\left.\begin{aligned} G_1 &= (\underline{X} \geqq X) \\ G_2 &= (\underline{X} \geqq X_0) \end{aligned}\right\} X \geqq X_0$

hierin speelt.

De kans $P(G_1|G_2)$ wordt de voorwaardelijke kans genoemd, namelijk de kans dat G_1 optreedt onder de voorwaarde dat G_2 zich heeft voorgedaan.

In het specifieke geval van (4.1) is de combinatie van gebeurtenissen G_1G_2 echter gelijkwaardig aan G_1 zelf. Namelijk als een afvoer groter is dan x, is deze volgens de definitie zeker groter dan x_o (zie fig. 2). De kansberekening geeft nu voor de voorwaardelijke kans

$$P(G_1|G_2) = \frac{P(G_1G_2)}{P(G_2)} = \frac{P(G_1)}{P(G_2)}$$

Passen we dit toe op (4.1) en (3.4) dan komt er

$$P(\underline{x} \geq x|\underline{x} \geq x_o) = \frac{e^{-\lambda x}}{e^{-\lambda x_o}} = e^{-\lambda(x - x_o)} \tag{4.2}$$

Worden nu alle afvoeren uitgedrukt ten opzichte van het gekozen niveau x_o dan blijven de voorgaande formules geldig. Namelijk, definieer

$$\begin{aligned} &\underline{y} < 0|\underline{x} < x_o \\ &\underline{y} = \underline{x} - x_o|\underline{x} \geq x_o \end{aligned} \tag{4.3}$$

waarin voor het waardenbereik geldt

$$(x_o \leq \underline{x} < \infty) \Longleftrightarrow (0 \leq \underline{y} < \infty)$$

Opgemerkt wordt dat (4.3) niet een gewone verschuivingstransformatie is, aangezien afvoeren tussen $0 \leq x < x_o$ in het laatste geval niet meer meegenomen worden in de steekproef.

Voor (4.2) kan nu geschreven worden

$$P(\underline{y} \geq y) = e^{-\lambda(x - x_o)} = e^{-\lambda y}, \quad y \geq 0 \tag{4.4}$$

wat weer dezelfde vorm is als in (3.4). De parameter van de verdeling behoudt dus zijn oorspronkelijke waarde λ en deze kan dus ook berekend worden met behulp van de kansvariabele $\underline{y}$ gedefinieerd volgens (4.3). De exponentiële verdeling is de enige verdeling die deze eigenschap bezit (HEMELRIJK, 1963).

(Teneinde in twijfelgevallen misverstanden uit te sluiten kunnen gewone kansen als a b s o l u t e kansen worden aangeduid in tegenstelling tot v o o r w a a r d e l ij k e kansen).

5. DE KARAKTERISTIEKEN (MOMENTEN) VAN DE EXPONENTIELE VERDELING

De gemiddeld te verwachten afvoer bij de kansverdeling gegeven onder (3.3) kan in de parameter van de verdeling zelf worden uitgedrukt. Deze mathematische verwachting (Mathematical Expectation E) volgt uit

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

$$= - \int_0^\infty x \, d \, e^{-\lambda x}$$

Na partiële integratie wordt verkregen

$$E(\underline{x}) = \left[- x \, e^{-\lambda x} - \frac{1}{\lambda} e^{-\lambda x} \right]_0^\infty = \frac{1}{\lambda} \qquad (5.1)$$

Wordt uit een grafische bewerking met half-logarithmisch grafiekpapier een waarde voor λ gevonden, dan volgt de gemiddeld te verwachten afvoer dus uit reciproke van λ.

De overschrijdingskans van deze waarde wordt eenvoudig gevonden uit (3.4) en bedraagt

$$P(\underline{x} \geq \frac{1}{\lambda}) = e^{-1} = 0.3679 \qquad (5.2)$$

Voor de variantie van de kansverdeling geldt

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

$$= \int_0^\infty x^2 \lambda e^{-\lambda x} dx - (\frac{1}{\lambda})^2 = (\frac{1}{\lambda})^2 \qquad (5.3)$$

De variantie blijkt dus gelijk te zijn aan het kwadraat van de verwachtingswaarde, hetgeen door tweemaal partieel integreren wordt gevonden.

6. VOORWAARDELIJKE VERWACHTING

Indien geldt dat $\underline{x} \geq 0$ dan blijkt uit (5.1) dat de gemiddeld te verwachten afvoer $E(\underline{x}) = \frac{1}{\lambda}$.

Gevraagd kan nu worden welke waarde de afvoer gemiddeld zal aannemen als eenmaal een bepaald niveau wordt overschreden.

Dit probleem is in feite reeds opgelost met het vaststellen van de voorwaardelijke kansverdeling. Zijn we geïnteresseerd in

$$E(\underline{x}|\underline{x} \geq x_o)$$

dan kan volgens (4.3) vastgesteld worden dat

$$E(\underline{x} - x_o|\underline{x} \geq x_o) = E(\underline{y}) = \frac{1}{\lambda} \qquad (6.1)$$

hetgeen uit (4.4) en (5.1) volgt.

Weer uitgedrukt ten opzichte van de oorspronkelijke schaal blijkt nu uit (6.1) dat

$$E(\underline{x}|\underline{x} \geq x_o) = x_o + \frac{1}{\lambda} \qquad (6.2)$$

De voorwaardelijke verwachting wordt dus steeds verkregen door de reeds bekende $\frac{1}{\lambda}$ bij het niveau x_o bij te tellen.

De overschrijdingskans van deze hoeveelheid wordt weer gevonden uit (3.4) en bedraagt

$$P(\underline{x} \geq x_o + \frac{1}{\lambda}) = e^{-\lambda(x_o + \frac{1}{\lambda})}$$

$$= e^{-\lambda x_o - 1}$$

$$= \frac{0.3679}{e^{\lambda x_o}} \qquad (6.3)$$

Voor $x_o \to \infty$ gaat deze kans naar o hetgeen plausibel is.

7. HET SCHATTEN VAN DE PARAMETERWAARDE

In par. 3 is aangegeven dat λ grafisch geschat kan worden door uitzetten op half-log grafiekpapier. Een procedure waarbij λ berekend wordt volgt uit het gelijkstellen van de steekproefmomenten met de populatiemomenten dus volgens (5.1)

$$\frac{1}{\lambda} = \bar{x}$$

waaruit dan volgt dat een schatter voor λ is

$$S_1(\lambda) = \frac{n}{\sum_i x_i} \qquad (7.1)$$

Door nu met de verzamelde waarnemingen x_i $(i = 1, \ldots, n)$ de waarde van λ te berekenen kan de curve voor de kansverdeling worden getekend. Behalve de helling λ, in te vullen in (3.5), is namelijk nog bekend dat

$$P(\underline{x} \geq 0) = 1$$

hetgeen eenvoudig uit (3.4) volgt. Nu zijn een punt van de rechte en de helling bekend, en hiermede kan de rechte dus getekend worden.

Eenvoudiger nog is het een tweede punt te berekenen door bijvoorbeeld $p_1 = 0{,}01$ te nemen en (3.7) toe te passen.

De exponentiële verdeling is door slechts één parameter geheel bepaald. Toepassing van (7.1) is dus voldoende om de verdeling in een specifiek geval te leren kennen. Uit (5.3) blijkt dat gemiddelde en standaardafwijking voor deze verdeling aan elkaar gelijk zijn.

Indien dus de verdeling toegepast mag worden, zal tevens moeten gelden dat

$$S_2(\lambda) = \sqrt{\frac{n-1}{\sum_i (x_i - \bar{x})^2}} = \frac{1}{s} \tag{7.2}$$

Door steekproeffluctuaties zal niet tegelijk aan (7.1) en (7.2) voldaan zijn. Wel zal moeten gelden dat de waarden niet te sterk mogen verschillen. Hierop zal in par. 12 worden teruggekomen.

Behalve met de formules (7.1) en (7.2) waarmede schatters volgens de zogenaamde momenten methode werden verkregen kunnen schatters worden afgeleid met behulp van de methode van de grootste aannemelijkheid, of maximum likelihood.

Uitgegaan wordt hierbij van het feit dat bij n waarnemingen in feite n stochastische variabelen $\underline{x}_1, \ldots, \underline{x}_n$ gerealiseerd zijn. Deze hebben alle dezelfde kansverdeling en er wordt aangenomen dat deze stochastische variabelen onderling onafhankelijk zijn in statistische zin.

Een zogenaamde maximum likelihood (M.L.) schatter wordt nu verkregen met de volgende procedure.

Gegeven een steekproef van n onafhankelijke waarnemingen van $\underline{x}_1, \ldots, \underline{x}_n$. De simultane verdelingsdichtheid van deze steekproef wordt verkregen door de produktregel voor het gelijktijdig optreden van n onafhankelijke gebeurtenissen op (3.2) toe te passen zodat er komt

$$L = \prod_{i=1}^{n} \lambda\, e^{-\lambda x_i}$$

wat de aannemelijkheidsfunctie wordt genoemd. Geschreven kan worden

$$L = \lambda^n e^{-\lambda \sum x_i}$$

Deze uitdrukking is maximaal ('de kans op het optreden van juist deze steekproef is het grootst') als de eerste afgeleide naar λ gelijk is aan 0. Dan is dië waarde van λ gevonden, en daarmede dië kansverdeling, die bij de gevonden steekproef het meest aannemelijk is.

Eenvoudiger is nog te stellen dat $d \ln L/d\lambda = 0$ wat dezelfde oplossing geeft aangezien de logarithme een monotoon stijgende functie van λ is.

Er komt nu

$$\frac{d}{d\lambda} \ln L = \frac{d}{d\lambda}(n \ln \lambda - \lambda \sum x_i) = \frac{n}{\lambda} - \sum x_i = 0$$

zodat

$$S_3(\lambda) = \frac{n}{\sum x_i} = \frac{1}{x} \qquad (7.3)$$

wat eenzelfde resultaat blijkt te geven als de met de momenten methode gevonden schatter S_1.

Opgemerkt wordt nog dat de tweede afgeleide voor alle λ negatief is (zie (9.3)) zodat (7.3) inderdaad de optimale waarde voor het maximum is.

8. ZUIVERE SCHATTERS

Wordt met een nieuwe steekproef de uitkomst van $S_1(\lambda)$ opnieuw berekend, dan zal deze laatste waarde van de vorige verschillen. Dit is een gevolg van steekproeffluctuaties, zodat ook $S_1\ (\lambda)$ een stochastische grootheid is.

Gedefinieerd wordt $S_1\ (\underline{\lambda}) = \underline{\hat{\lambda}}$ met

$$\hat{\underline{\lambda}} = \frac{n}{\underline{x}_1 + \ldots + \underline{x}_n}$$

Hieruit volgt dan tevens dat $\hat{\underline{\lambda}}$ een kansverdeling bezit en dat een verwachtingswaarde en variantie van deze verdeling bepaald kunnen worden. Een belangrijke eigenschap zal eerst behandeld worden.

Een schatter wordt z u i v e r genoemd als de verwachtingswaarde van de schatter de geschatte parameter tot uitkomst heeft. Een v o o r b e e l d is het rekenkundig gemiddelde $\bar{x}$ als schatter van het populatie gemiddelde μ. Dan geldt

$$E(\bar{\underline{x}}) = E\left(\frac{\sum \underline{x}_i}{n}\right) = \frac{1}{n} E\left(\sum \underline{x}_i\right) = \frac{1}{n} \sum E(\underline{x}_i)$$

aangezien verondersteld wordt dat alle x_i trekkingen uit dezelfde populatie zijn, geldt hiervoor

$$E(\bar{\underline{x}}) = \frac{1}{n} \sum (E\underline{x})_i = \frac{1}{n} \sum \mu = \mu$$

Hetgeen inderdaad de gevraagde parameter oplevert. Deze schatter is dus zuiver.

Teneinde na te gaan of $\hat{\lambda} = \frac{1}{\bar{x}}$ een zuivere schatter is moet berekend worden

$$E(\hat{\underline{\lambda}}) = n\, E \frac{1}{\sum \underline{x}_i} \tag{8.1}$$

Teneinde deze verwachtingswaarde te kunnen bepalen moet eerst de kansverdeling van $\frac{1}{\sum \underline{x}_i}$ worden afgeleid. Het is in (8.1) namelijk niet mogelijk door verwisseling van de operatoren E en $\sum$ tot een eenvoudige oplossing te komen zoals voor $\bar{x}$ kon worden toegepast, aangezien $\frac{1}{\sum x_i} \neq \sum \left(\frac{1}{x_i}\right)$

8a. De verdeling van $\underline{x}_1 + \underline{x}_2$

Voor de kansdichtheid van de som van twee onafhankelijke kansvariabelen $\underline{z} = \underline{x} + \underline{y}$ geldt algemeen (STAM, 1964, -251-)

$$h(z) = \int_{-\infty}^{-\infty} p(x)\ q(z - x)\ dx$$

waarin p de kansdichtheid van $\underline{x}$ voorstelt en q die van $\underline{y}$.

Zijn de kansvariabelen $\underline{x}$ en $\underline{y}$ beide beperkt tot het interval $[0, \infty)$ dan wordt dit

$$h(z) = \int_{0}^{z} p(x)\ q(z - x)\ dx \tag{8.2}$$

Stel nu dat de functies p en q voldoen aan de exponentiële verdeling waarvan de dichtheid gegeven is door (3.2), dan wordt (8.2)

$$h(z) = \int_{0}^{z} \lambda\ e^{-\lambda x}\ \lambda\ e^{-\lambda(z - x)}\ dx$$

$$= \lambda^2 \int_{0}^{z} e^{-\lambda(x + z - x)}\ dx$$

$$= \lambda^2\ e^{-\lambda z} \int_{0}^{z} dx = \lambda^2\ z\ e^{-\lambda z}$$

8b. De verdeling van $\sum \underline{x}_i$

De bovenuitgevoerde integraalberekening staat bekend als het nemen van de convolutie van de dichtheidsfunctie met zichzelf. Voor een som van n gelijk verdeelde kansvariabelen wordt de kansverdeling verkregen uit de nde convolutie van de dichtheidsfunctie met zichzelf. Door volledige inductie valt te bewijzen (STAM, 1964, -259-) dat indien

de stochastische variabelen $\underline{x}_1, \ldots, \underline{x}_n$ onafhankelijk zijn en elke $\underline{x}_i$ negatief exponentieel verdeeld is met parameter λ, de verdeling van $\underline{z} = \underline{x}_1 + \ldots + \underline{x}_n$ volgt uit

$$h(z) = 0 \qquad z < 0$$
$$h(z) = \frac{\lambda^n z^{n-1}}{(n-1)!} e^{-\lambda z} \qquad z \geq 0 \tag{8.3}$$

Deze verdeling wordt de Erlang verdeling genoemd. De kansverdeling van $\underline{z} = \sum \underline{x}_i$ is hiermede bepaald. Nu wordt berekend wat de verwachtingswaarde van $\frac{1}{\underline{z}}$ is. Deze volgt uit (8.3) als volgt

$$E(\frac{1}{\underline{z}}) = \int_0^\infty \frac{1}{z} \frac{\lambda^n z^{n-1}}{(n-1)!} e^{-\lambda z} dz$$

$$= \frac{\lambda^n}{(n-1)!} \int_0^\infty z^{n-2} e^{-\lambda z} dz$$

HÜTTE (1959, -116-) geeft als algemene uitkomst voor dit type integralen, te berekenen door herhaalde partiële integratie,

$$\int_0^\infty x^n e^{-ax} dx = \frac{n!}{a^{n+1}}$$

zodat de verwachtingswaarde luidt:

$$E(\frac{1}{\underline{z}}) = \frac{\lambda^n}{(n-1)!} \frac{(n-2)!}{\lambda^{n-1}} = \frac{\lambda}{n-1}$$

en dus, nu weer voor de uitgangsformule (8.1)

$$E(\underline{\bar{\lambda}}) = n\, E \frac{1}{\sum \underline{x}_i} = \frac{n}{n-1} \lambda > \lambda$$

8c. Bespreking resultaten

Uit het bovenstaande valt op te maken dat een functie van een zuivere schatter geen zuivere schatter behoeft op te leveren. Aangetoond werd dat $S(\mu) = \bar{x}$ een zuivere schatter van μ is. Daarentegen blijkt $1/\bar{x}$ geen zuivere schatter voor $1/\mu$. In het geval van de exponentiële verdeling blijkt dus dat met $\frac{1}{\bar{x}}$ de waarde van λ systematisch te groot geschat wordt. De schatter is wel zogenaamd asymptotisch raak aangezien geldt dat voor grote steekproeven

$$\lim_{n \to \infty} \frac{n}{n-1} = \lim_{n \to \infty} \frac{1}{1 - \frac{1}{n}} = 1$$

Vaak wordt aangenomen dat voor $n > 50$ (soms $n > 30$) deze situatie al ontstaat.

Voor kleine steekproeven kan gebruik gemaakt worden van de zuivere schatter

$$\underline{\bar{\lambda}} = \frac{n-1}{\sum \underline{x}_i} \text{ met } E(\underline{\bar{\lambda}}) = \lambda \qquad (8.4)$$

tenzij men de M.L. schatter prefereert waarop de theorie van het construeren van betrouwbaarheidsintervallen is gebaseerd (par. 10).

9. VARIANTIES VAN SCHATTERS

In de vorige paragraaf werd reeds aangegeven dat schatters een kansverdeling hebben, en werd de verwachtingswaarde ervan berekend. Van belang is het tevens de variantie te bepalen aangezien deze informatie verschaft over de variaties die bij de gekozen wijze van schatten te verwachten zijn.

Gevraagd kan nu worden de variantie van deze schatters eveneens in de parameter λ uit te drukken. Hieruit kan een maat voor de betrouwbaarheid van de schatting verkregen worden. Uitgegaan wordt van de reeds verkregen resultaten (met $\underline{z} = \sum \underline{x}_i$)

M.L. schatter : $\hat{\underline{\lambda}} = \frac{n}{\underline{z}}$ $E(\hat{\underline{\lambda}}) = \frac{n}{n-1}\lambda$

Zuivere schatter: $\bar{\underline{\lambda}} = \frac{n-1}{\underline{z}}$ $E(\bar{\underline{\lambda}}) = \lambda$

waaruit dan volgt

Variantie M.L. schatter $E\left(\frac{n}{\underline{z}} - \frac{n}{n-1}\lambda\right)^2$

Variantie zuivere schatter $E\left(\frac{n-1}{\underline{z}} - \lambda\right)^2$

Voor de beide laatste vormen wordt geschreven, volgens een bekende eigenschap van de verwachting van het tweede gereduceerde moment,

$$\sigma^2(\hat{\underline{\lambda}}) = n^2\, E\left(\frac{1}{\underline{z}}\right)^2 - \left(\frac{n\lambda}{n-1}\right)^2$$
$$\sigma^2(\bar{\underline{\lambda}}) = (n-1)^2\, E\left(\frac{1}{\underline{z}}\right)^2 - \lambda^2 \qquad (9.1)$$

Allereerst dient dus de verwachting van $\left(\frac{1}{\underline{z}}\right)^2$ berekend te worden. Dit gaat als volgt, teruggrijpend op (8.3)

$$E\left(\frac{1}{\underline{z}}\right)^2 = \int_0^\infty \left(\frac{1}{z}\right)^2 \frac{\lambda^n z^{n-1}}{(n-1)!} e^{-\lambda z}\, dz$$

$$= \frac{\lambda^n}{(n-1)!} \int_0^\infty z^{n-3} e^{-\lambda z}\, dz$$

Weer volgens HÜTTE (1959, -116-) wordt gevonden

$$E\left(\frac{1}{\underline{z}}\right)^2 = \frac{\lambda^n}{(n-1)!}\, \frac{(n-3)!}{\lambda^{n-2}} = \frac{\lambda^2}{(n-1)\,(n-2)}$$

Van deze vorm moet nog volgens (9.1), na vermenigvuldigen met een vorm in n, het kwadraat van de verwachtingswaarde worden afgetrokken. Er wordt dan verkregen:

$\sigma^2(\hat{\underline{\lambda}})$ M.L. (niet zuiver)	$\sigma^2(\bar{\underline{\lambda}})$ (zuiver)
$(\frac{n}{n-1})^2 \frac{\lambda^2}{n-2}$	$\frac{\lambda^2}{n-2}$

Aangezien $\frac{n}{n-1} > 1$ blijkt hieruit dat de zuivere schatter bovendien nog het voordeel heeft van een grotere efficientie, of wel een kleinere spreiding in uitkomsten wat resulteert in nauwere betrouwbaarheidsintervallen. Voor $n \to \infty$ gaan beide gevallen in elkaar over.

Voor het bovenstaande kan onder andere ook verwezen worden naar KENNY and KEEPING (1959, -375-). Van belang is namelijk de algemene stelling dat de v e r d e l i n g van:

maximum likelihood (M.L.) schatters voor grote waarden van n tot de normale verdeling naderen en in de limiet ($n \to \infty$) een minimale variantie bezitten ten opzichte van andere schatters

Van deze eigenschap kan gebruik worden gemaakt door toepassing van de tabel van de normale verdeling voor het vaststellen van betrouwbaarheidsintervallen. Dit is veelal eenvoudiger dan deze in elk specifiek geval opnieuw te moeten berekenen.

Een methode om de variantie van een M.L. schatter voor $n \to \infty$ te vinden is het toepassen van de informatie-matrix die hier uit één element bestaat en wel

$$\frac{1}{\sigma^2(\hat{\underline{\lambda}})} = - E(\frac{d^2 \ln \underline{L}}{d\lambda^2}) \qquad (9.2)$$

De berekening levert het volgende resultaat

$$\ln \underline{L} = n \ln \lambda - \lambda \sum \underline{x}_i$$

(zie par. 7) zodat

$$\frac{d\underline{L}}{d\lambda} = \frac{n}{\lambda} - \sum \underline{x}_i$$

en

$$\frac{d^2\underline{L}}{d\lambda^2} = -\frac{n}{\lambda^2} \qquad (9.3)$$

zodat (9.2) geeft, aangezien de kansvariabelen $\underline{x}_i$ reeds door differentiatie zijn weggevallen,

$$\frac{1}{\sigma^2(\hat{\underline{\lambda}})} = -E(-\frac{n}{\lambda^2}) = n\,E(\frac{1}{\lambda^2}) = \frac{n}{\lambda^2}$$

en tenslotte

$$\sigma^2(\hat{\underline{\lambda}}) = \frac{\lambda^2}{n}$$

Voor grote waarden van n kan dus worden aangenomen dat $\hat{\underline{\lambda}}$ normaal verdeeld is met respectievelijk

verwachting $\mu = \lambda$

en (9.4)

standaard afwijking $\sigma = \frac{\lambda}{\sqrt{n}}$

10. BETROUWBAARHEIDSINTERVALLEN

In plaats van alleen de parameterwaarde λ te schatten kan ook een intervalschatting worden gemaakt. Het voordeel hiervan is dat een uitspraak omtrent de betrouwbaarheid van het resultaat kan worden gedaan.

10a. Voor de parameter λ

Met behulp van de bovenstaande uitkomsten kunnen betrouwbaarheidsintervallen worden afgeleid voor de verschillende grootheden. De procedure is dan - voor grote waarden van n -

$$\hat{\lambda} = \frac{n}{\sum x_i} \qquad \text{(berekend uit gegevens)}$$

Voor de normale verdeling met parameters (μ, σ) uit (9.4) geldt dan voor schattingen van μ en σ

$$S(\mu) = \hat{\lambda} \quad \text{en} \quad S(\sigma) = \frac{\hat{\lambda}}{\sqrt{n}}$$

zodat een ($1 - \alpha$) betrouwbaarheidsinterval verkregen kan worden uit (α = kans dat interval λ niet bevat)

$$P(\underline{\hat{\lambda}} + T_{\frac{1}{2}\alpha}\frac{\underline{\hat{\lambda}}}{\sqrt{n}} < \lambda < \underline{\hat{\lambda}} + T_{(1-\frac{1}{2}\alpha)}\frac{\underline{\hat{\lambda}}}{\sqrt{n}}) = 1 - \alpha \tag{10.1}$$

Hierin kunnen $T_{\frac{1}{2}\alpha}$ en $T_{(1-\frac{1}{2}\alpha)}$ bij de in de index aangegeven kans in de tabel voor de normale verdeling worden opgezocht.

Indien een onbetrouwbaarheid van α = 5 % aanvaard wordt, dan luidt (10.1)

$$P(\underline{\hat{\lambda}}(1 - \frac{1,96}{\sqrt{n}}) < \lambda < \underline{\hat{\lambda}}(1 + \frac{1,96}{\sqrt{n}})) = 95\ \% \tag{10.2}$$

Een voorstelling van dit betrouwbaarheidsinterval, in relatie tot de normale verdeling, staat weergegeven in fig. 3.

10b. Voor de overschrijdingskansen

De vorm van de exponentiële verdeling namelijk

$$P(\underline{x} \geq x) = e^{-\lambda x} = f(\lambda) \tag{10.3}$$

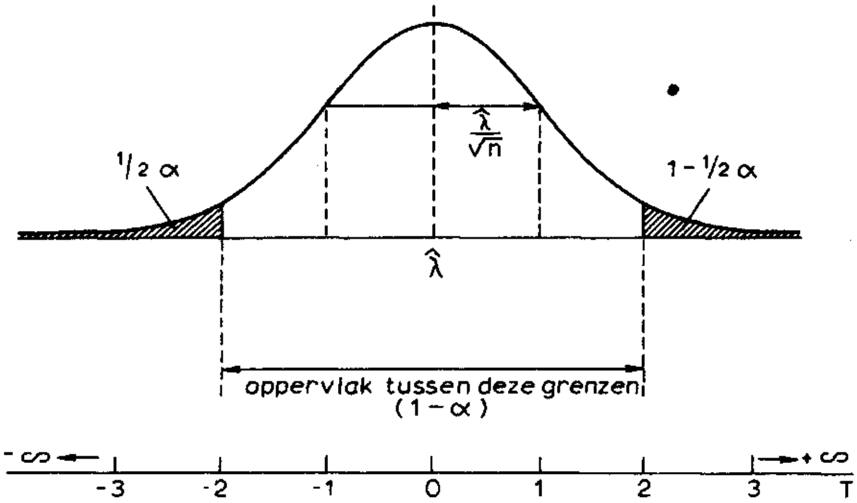
laat zien dat deze functie voor toenemende waarden van λ monotoon dalend is en slechts van één parameter afhangt.

Er kan dus geschreven worden voor (10.2)

$$P(e^{-\underline{\hat{\lambda}}(1+1,96/\sqrt{n})\,x} < e^{-\lambda x} < e^{-\underline{\hat{\lambda}}(1-1,96/\sqrt{n})\,x}) = 95\ \% \tag{10.4}$$

(zie ook Rapport Delta Commissie, 1960)

fig. 3



$\alpha = 10$	$T_{1/2\alpha} = -1.64$	$T_{1-1/2\alpha} = 1.64$
$\alpha = 5$	$T_{1/2\alpha} = -1.96$	$T_{1-1/2\alpha} = 1.96$
$\alpha = 1$	$T_{1/2\alpha} = -2.58$	$T_{1-1/2\alpha} = 2.58$

voorbeeld van het vaststellen van betrouwbaarheids intervallen waarmede een betrouwbaarheidsinterval is verkregen voor de overschrijdingskans

$$p = e^{-\lambda x}$$

en wel

$$e^{-\hat{\lambda}(1+1,96/\sqrt{n})\,x} < p < e^{-\hat{\lambda}(1-1,96/\sqrt{n})\,x} \qquad (10.5)$$

Hierbij is gebruik gemaakt van de stelling dat een één-éénduidige functie van een M.L.-schatter zelf ook weer een M.L. schatter is (MOOD and GRAYBILL, 1963, -185-).

10c. Voor maatgevende afvoeren

Wordt (10.3) opgelost naar de drempelwaarde x, de maatgevende afvoer bij een kansniveau p_1, dan wordt verkregen

$$x = \frac{-\ln p_1}{\lambda} = f(\lambda)$$

Ook deze functie is voor toenemende waarden van λ monotoon dalend, zodat nu geldt:

$$P\left(\frac{-\ln p_1}{\hat{\lambda}(1+1,96/\sqrt{n})} < \frac{-\ln p_1}{\lambda} < \frac{-\ln p_1}{\hat{\lambda}(1-1,96/\sqrt{n})}\right) = 95\ \% \qquad (10.6)$$

waarmede een betrouwbaarheidsinterval is verkregen voor de maatgevende afvoer

$$x = \frac{-\ln p_1}{\hat{\lambda}}$$

en wel

$$\frac{-\ln p_1}{\hat{\lambda}(1+1,96/\sqrt{n})} < x < \frac{-\ln p_1}{\hat{\lambda}(1-1,96/\sqrt{n})} \qquad (10.7)$$

10d. Samenvatting en gebruik

Toepassing van het gebruik van de betrouwbaarheidsintervallen kan het meest praktisch plaatsvinden door uitzetten op half-logarithmisch grafiekpapier zoals besproken in par. 3.

Met behulp van (10.2) kunnen betrouwbaarheidsgrenzen voor λ op dezelfde wijze worden uitgezet als de kansverdeling zelf met behulp van (3.5). De betrouwbaarheidsintervallen worden dan door rechte lijnen weergegeven.

De interpretatie van (10.5) is nu dat met dezelfde lijnen een betrouwbaarheidsinterval voor de kans p wordt gegeven bij constante waarde van de afvoer x. Voor dit doel moet dus bij een van belang zijnde waarde van x de v e r t i c a l e afstand tussen de betrouwbaarheidsgrenzen genomen worden (zie deel II).

De interpretatie van (10.7) is dat met weer dezelfde lijnen een betrouwbaarheidsinterval voor de maatgevende afvoer x wordt verkregen bij een constante waarde van de overschrijdingskans p. Voor dit doel moet nu bij een van belang zijnde waarde van p de h o r i z o n t a l e afstand tussen de betrouwbaarheidsgrenzen genomen worden. (Zie deel II).

11. TOETSEN VAN AANPASSING AAN DE EXPONENTIELE VERDELING

Het is mogelijk na te gaan (te toetsen) of een steekproef door een exponentiële verdeling kan worden beschreven, met andere woorden te toetsen of het aannemelijk is te veronderstellen dat de steekproef inderdaad uit een exponentiële verdeelde populatie komt. Hiertoe kan gebruik worden gemaakt van de χ^2-toets van aanpassing voor in klassen ingedeelde gegevens.

In feite is de χ^2-verdeling voor dit geval een benadering van de exacte multinomiale verdeling die hier toegepast zou moeten worden, doch moeilijk hanteerbaar is. Indien in één of meer klassen het aantal gegevens gering is, is de benadering slecht en wordt de limietverdeling slechts langzaam bereikt. Veelal wordt dan geadviseerd

klassen samen te nemen zodat er per klasse minstens 5 gegevens voorkomen (KENNY and KEEPING, 1959, -118-).
Ook andere minima worden genoemd, bijvoorbeeld 10 of 20. Voor vele praktische toepassingen wordt 5 als laagste waarde aangehouden.

De toets dient uitgevoerd te worden met parameters geschat uit de klasse-indeling (VAN MONTFORT, 1973, -37- e.v.), aangezien geen differentie binnen de klassen wordt aangebracht en als het ware overgegaan wordt op een discrete kansvariabele. Stel het aantal klassen gelijk aan k, de kans op een gegeven in klasse i gelijk aan p_i, waarbij $\sum_i p_i = 1$ het aantal gegevens in klasse i gelijk aan n_i, waarbij $\sum_i n_i = n$ en de klassegrenzen van klasse i gelijk aan g_i en g_{i+1}, met $g_i < g_{i+1}$, dan is de M.L.-functie

$$L = \prod_{i=1}^{k} p_i^{n_i}$$

waarvan de logarithme is

$$\ln L \sum_{i=1}^{k} n_i \ln(p_i) \qquad (11.1)$$

De parameters die p_i bepalen dienen dusdanig gekozen te worden dat ln L maximaal wordt. Voor de exponentiële verdeling wordt dit

$$\ln L = \sum_{i=1}^{k} n_i \ln(e^{-\lambda g_{i+1}} - e^{-\lambda g_i})$$

met eis

$$\frac{dL}{d\lambda} = 0 \qquad (11.2)$$

Met behulp van de hieruit gevonden waarde van λ - bijvoorbeeld met een trial-and-error methode bepaald - wordt voor elke klasse de kans p_i geschat met

$$\hat{p}_i = e^{-\lambda g_i} - e^{-\lambda g_{i+1}} \qquad (11.3)$$

zodat de toetsingsgrootheid wordt

$$\chi_N^2 = \sum_{i=1}^{k} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} \tag{11.4}$$

Het aantal vrijheidsgraden dient gekozen te worden als $N = n - 1 - 1$. Aftrek vindt plaats voor het feit dat $\sum n_i = n$ met n een vast getal, waardoor de aantallen n_i reeds door k - 1 klassen bepaald zijn, en voor het feit dat de ene parameter (λ) uit dezelfde streekproef werd bepaald.

In tabellen voor de χ^2-verdeling kan nagegaan worden of de gevonden waarde voor χ^2 zodanig hoog uitvalt dat eraan getwijfeld moet worden of de steekproef wel uit een exponentieel verdeelde grootheid getrokken gedacht kan worden.

Opmerking

De bewerkelijkheid van het oplossen van de niet-lineaire vergelijking (11.2) zal ertoe leiden dat in de praktijk veelal (11.3) en (11.4) berekend zullen worden met λ gevonden uit (7.1) of (8.4).

In tegenstelling tot de χ^2-toets maakt de Kolmogorov-Smirnov (K.G.-)toets gebruik van alle individuele waarnemingen waardoor geen informatie door het indelen in klassen verloren gaat (SIEGEL, 1956, -51-). Voor het toetsen of een steekproef uit een populatie met een bekende verdeling afkomstig kan zijn (one-sample toets) is de K.G.-toets meer onderscheidend dan de χ^2-toets. Ook de toepassing van de toets is eenvoudiger. Een verder bijkomend voordeel is dat de toets een eenvoudige grafische weergave en gebruik mogelijk maakt. Voorbeelden zijn gegeven door STOL (1970).

Er wordt hier herinnerd aan de formule voor de one-sample toets bij 5 % onbetrouwbaarheid

$$|D| = \frac{1.358}{\sqrt{n}} \tag{11.5}$$

waarin $|D|$ het absolute verschil tussen steekproef en theoretische verdeling is, gemeten langs de kans-as, en dus de meetkundige plaats van de kritieke waarden geeft bij alle kansniveaus.

De kritieke waarde volgens de two-sample toets, bij het toetsen van twee steekproeven, bedraagt bij 5 % onbetrouwbaarheid

$$|d| = \frac{1.358}{\sqrt{\dfrac{n_1 n_2}{n_1 + n_2}}} \tag{11.6}$$

Het alternatief van de toets is dat de getoetste verdeling exponentieel is met andere parameters, respectievelijk een geheel andere verdeling is.

12. SAMENHANG TUSSEN SCHATTEN EN TOETSEN

De in de vorige paragraaf beschreven b e t r o u w b a a r h e i d s i n t e r v a l l e n (I_b) geven aan welke parameterwaarden bij de verkregen w a a r g e n o m e n uitkomsten nog aannemelijk zijn.
De algemene formulering hiervoor is

$$P(\lambda \in I_b(\underline{\hat{\lambda}})) = (1 - \alpha) \tag{12.1}$$

wat wil zeggen dat gezien deze realisatie werkelijke waarden van de onbekende parameter λ die door $\underline{I}_b$ met een kans $(1 - \alpha)$ worden overdekt, nog aannemelijk zijn.

Bij t o e t s i n g is het probleem of een onbekende λ wel gelijk zou kunnen zijn aan een a a n g e n o m e n waarde λ_o. De nul-hypothese is dan H_o: $\lambda = \lambda_o$ tegen bijvoorbeeld het alternatief H_1: $\lambda \neq \lambda_o$. De hypothese H_o wordt verworpen als λ in het kritieke gebied $I_k(H_o, \alpha)$ blijkt te vallen. De hypothese H_o wordt nu onderzocht met een schatting voor λ bijvoorbeeld $\underline{\hat{\lambda}}$. De kansverdeling F van de toetsingsgrootheid $\underline{\hat{\lambda}}$ moet nu bekend zijn voor het geval H_o geldt. Algemeen geformuleerd, bepaal F(a) in

$$P(\underline{\hat{\lambda}} < a \mid \lambda = \lambda_o) = F(a)$$

Vaak kan (wordt) deze verdeling door een normale verdeling benaderd (worden). Bij aan te nemen waarden voor de onbetrouwbaarheid α gaat het in dit type problemen om $F(I_k) = \alpha$, uitgeschreven om de kans

$$P(\hat{\underline{\lambda}} \in I_k \mid \lambda = \lambda_o) = \alpha \qquad (12.2)$$

De nul-hypothese H_o wordt verworpen indien de gevonden schatting $\hat{\underline{\lambda}}$ in het kritieke gebied I_k terecht is gekomen. De kans hierop is gelijk aan de gekozen (kleine) kans α.

De formuleringen gegeven in (12.1) en (12.2) hangen nauw met elkaar samen, zie fig. 4. Zo kunnen betrouwbaarheidsgrenzen worden afgeleid uit toetsen terwijl omgekeerd bij bekende betrouwbaarheidsintervallen een toets kan worden ontworpen.

Wordt in (12.2) het kritieke gebied I_k vervangen door het gebied waarin H_o geaccepteerd wordt, stel I_a, dan komt er

$$P(\hat{\underline{\lambda}} \in I_a \mid \lambda = \lambda_o) = 1 - \alpha \qquad (12.3)$$

Onder H_o, dat wil zeggen indien inderdaad $\lambda = \lambda_o$, wat niet bekend is, stelt (12.3) de steekproefverdeling van $\hat{\underline{\lambda}}$ voor. Alle realisaties $\hat{\lambda}$ die n i e t verworpen worden op basis van (12.3) vallen in I_a zodat er een interval I'_b moet bestaan zodanig dat de volgende stochastische gebeurtenissen identiek zijn, namelijk

$$G_1 = (\hat{\underline{\lambda}} \in I_a)$$

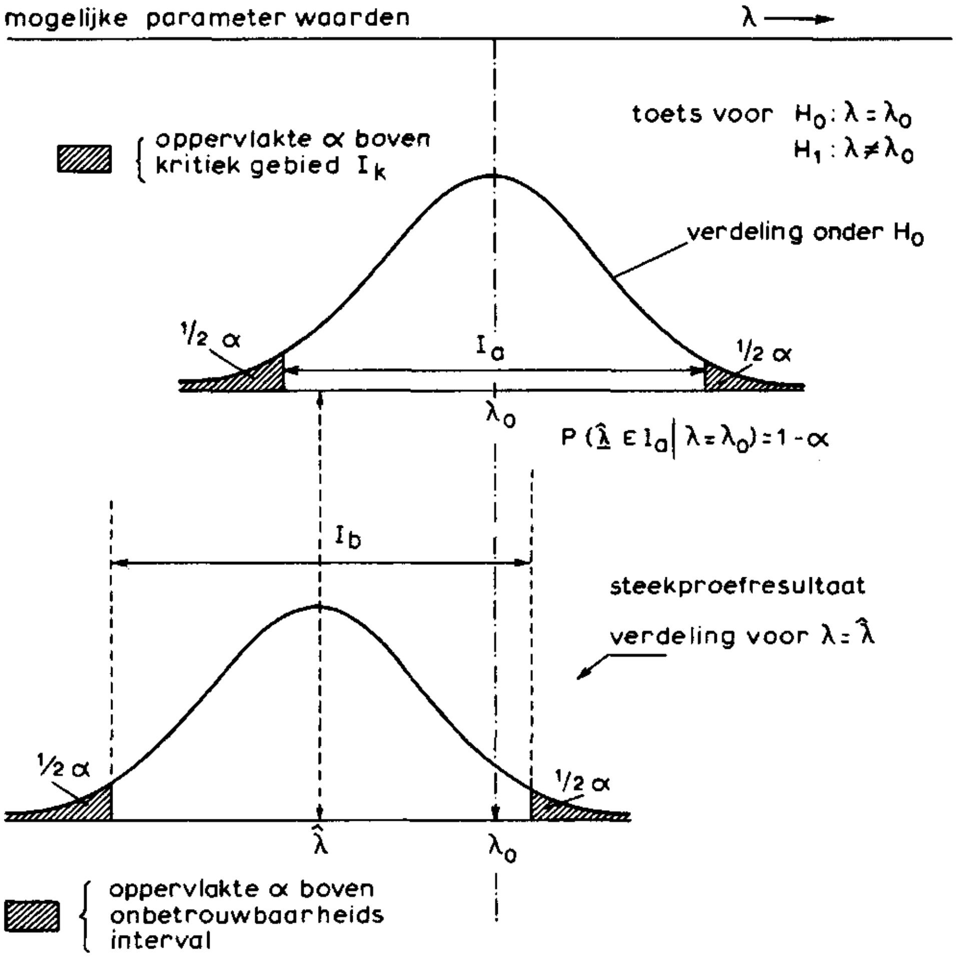
$$G_2 = (\lambda_o \in I'_b(\hat{\underline{\lambda}}))$$

Zodat (12.3) dan overgaat in

$$P(\lambda_o \in I'_b(\hat{\underline{\lambda}})) = 1 - \alpha$$

een vorm overeenkomstig (12.1) waarin $\lambda_o = \lambda$ en de, onbekende, werkelijke waarde van de parameter λ voorstelt.

fig. 4



$P(\lambda \in I_b(\hat{\underline{\lambda}})) = 1 - \alpha$

samenhang tussen schatten en toetsen

Wordt uitgegaan van (12.1) dan wordt eerst opgemerkt dat deze de steekproefverdeling in het punt $\lambda = \hat{\lambda}$ weergeeft. Nu kan $H_o: \lambda = \hat{\lambda}$ als hypothese gelden door waarden van λ_o te verwerpen waarvoor geldt

$$\lambda_o \notin I_b(\hat{\underline{\lambda}}) = I'_a$$

wordt λ_o op waarnemingsuitkomsten gebaseerd dan wordt dit

$$P(\underline{\lambda}_o \in I'_a \mid \lambda = \hat{\lambda}) = 1 - \alpha$$

een vorm overeenkomstig (12.3).

Het voorgaande had betrekking op het toetsen van parameterwaarden. Een dergelijke 'inversie' van toetskriteria naar betrouwbaarheidsintervallen is voor aanpassingstoetsen in het algemeen niet mogelijk. De uitzondering hierop is de Kolmogorov-Smirnov toets door het feit dat het toetsingscriterium een zo directe maat is voor het optreden van afwijkingen van de curve zelf. Stelt D het maximale verschil voor - langs de kans-as gemeten - tussen de cumulatieve steekproefverdeling S(x) en een theoretische verdeling F(x) dan kan voor de K.S.-toets geschreven worden overeenkomstig (12.2) met H_o: de populatieverdeling is F(x) tegen het alternatief H_1: de populatieverdeling is niet F(x)

$$P(|\underline{D}| > d_\alpha \mid F(x)) = \alpha$$

waarin d_α de kritieke waarde bij onbetrouwbaarheid α voorstelt (LINDGREN and McELRATH, 1967). Hieruit volgt dan weer met $|\underline{D}| = |F(x) - \underline{S}(x)|$

$$P(-d_\alpha \leq F(x) - \underline{S}(x) \leq + d_\alpha) = 1 - \alpha$$

zodat

$$P(\underline{S}(x) - d_\alpha \leq F(x) \leq \underline{S}(x) + d_\alpha) = 1 - \alpha$$

hetgeen weer een intervalschatting volgens (12.1) oplevert (KENDALL and STUART, 1961, -457-).

13. HET BENODIGD AANTAL GEGEVENS

Betrouwbaarheidsintervallen lenen zich vaak goed voor het berekenen van het minimum aantal gegevens dat vereist is om een resultaat met een gevraagde betrouwbaarheid te verkrijgen. Dit zal achtereenvolgens voor de drie formules (10.2), (10.5) en (10.7) worden toegelicht.

Aangenomen wordt dat een onbetrouwbaarheid α geaccepteerd wordt met kritieke waarde T. Uit (10.2) volgt dan dat in het algemeen het betrouwbaarheidsinterval I_b groot is

$$I_b = \lambda(1 + \frac{T}{\sqrt{n}}) - \lambda(1 - \frac{T}{\sqrt{n}})$$

$$= \lambda \frac{2T}{\sqrt{n}}$$

Nu wordt gevraagd het interval hoogstens gelijk te maken aan een grootte I_λ. Zodat gesteld wordt

$$I_b \leqq I_\lambda$$

of

$$I_\lambda \geqq \lambda \frac{2T}{\sqrt{n}} \qquad (13.1)$$

waarin de factor in het linker lid gemaximaliseerd kan worden tot rond 4 waardoor zeker aan de ongelijkheid voldaan is*.De oplossing is dan voor λ:

$$n \geqq \left(\frac{2T\lambda}{I_\lambda}\right)^2 \qquad (13.2)$$

Op eenzelfde wijze kan deze berekening worden uitgevoerd voor de kans p bij gegeven maatgevende afvoer x. Er komt dan uit (10.5)

*n.l. $T_{0,05} = 1,96$ zodat $4 > 2 \times 1,96$

$$I_b = e^{-\lambda(1-T/\sqrt{n})x} - e^{-\lambda(1+T/\sqrt{n})x}$$

$$= e^{-\lambda x}(e^{T\lambda x/\sqrt{n}} - e^{-T\lambda x/\sqrt{n}})$$

met

$$I_p \geqq e^{-\lambda x}\, 2\sinh(T\lambda x/\sqrt{n}) \tag{13.3}$$

Zodat met behulp van een tabel van waarden van de sinus hyperbolicus functie dië waarde van n bepaald moet worden waarvoor geldt dat

$$\sinh(T\lambda x/\sqrt{n}) \leqq \frac{1}{2} I_p\, e^{\lambda x}$$

of expliciet

$$n \geqq \frac{(T\lambda x)^2}{(\operatorname{arsinh} \frac{1}{2} I_p\, e^{\lambda x})^2} \tag{13.4}$$

Bij gegeven waarde van λ en x verschaft de formule dus het aantal onafhankelijke waarnemingen n dat gedaan moet worden voor het verkrijgen van een tweezijdig interval voor p niet groter dan I_o, behoudens een kans α = 5 %.

Tenslotte geldt voor een interval rond x bij gegeven waarde van p uit (10.7)

$$I_b = \frac{-\ln p}{\lambda(1 - T/\sqrt{n})} - \frac{-\ln p}{\lambda(1 + T/\sqrt{n})}$$

$$= \frac{-\ln p}{\lambda} \cdot \frac{2T/\sqrt{n}}{1 - (T/\sqrt{n})^2}$$

met

$$I_x \geqq \frac{-\ln p}{\lambda} \frac{2T\sqrt{n}}{n - T^2} \tag{13.5}$$

Deze uitdrukking levert een vierkantsvergelijking in $\sqrt{n}$ welke na oplossing in de volgende vorm kan worden gebracht

$$n \geq T^2 \, e^{2 \, \mathrm{arsinh}\left(\frac{-\ln p}{I_x \lambda}\right)} \tag{13.6}$$

Zoals steeds bij dit type formules moet iets bekend zijn omtrent de verdeling waaruit de steekproef wordt getrokken. Dit is vanzelfsprekend aangezien zonder enige formatie over het te analyseren systeem niet bekend kan zijn hoeveel waarnemingen minstens noodzakelijk zijn om een bepaald doel te bereiken. In een nieuw gebied van onderzoek is λ onbekend en zou n dus niet opgelost kunnen worden. Veelal zal men voor de parameter λ dan niet veel anders kunnen doen dan een betrouwbare waarde $\hat{\lambda}$ van een overeenkomstig gebied in de gegeven formules in te vullen als populatieparameter. Of, weer minder goed, de gegevens van een korte reeks alvast gebruiken voor het berekenen van (13.2), (13.4) respectievelijk (13.6).

Opgemerkt wordt nog dat de grootte van de intervallen I_λ, I_p en I_x niet onafhankelijk van elkaar kan worden gekozen.

Wordt bijvoorbeeld bij een gewenste α de waarde van T bepaald (bijv. 1.96), en wordt bij een gewenste overschrijdingskans p een intervalbreedte I_x als eis gesteld, dan geeft (13.6) voor de populatie waarvoor λ geldt de grootte n van de steekproef. Alle vrij te kiezen waarden liggen hiermede vast en de andere intervalbreedten worden dan gevonden uit (13.3) voor p en uit (13.1) voor λ.

14. HET EFFECTIEVE AANTAL GEGEVENS

Het voorgaande geldt voor het geval dat de gegevens in de steekproef onderling onafhankelijk zijn. Voor veel hydrologische reeksen is dit uitgangspunt niet juist en bestaat er een grote afhankelijkheid tussen opeenvolgende gebeurtenissen (dagafvoeren bijv.). Onder geschikte aannamen een zogenaamd Markov I model, waarvoor geldt dat

$$P(\underline{G}_{t+1} \mid G_t, G_{t-1}, \ldots, G_o) = P(\underline{G}_{t+1} \mid G_t)$$

wat tot uitdrukking brengt dat de voorwaardelijke kans op de gebeurtenis $\underline{G}_{t+1}$ alleen afhangt van wat op tijdstip t werd bereikt kan afgeleid worden dat, met k = (t + k) - t

$$\rho_k = \rho^k \tag{14.1}$$

waarin ρ_k de autocorrelatie van de over een afstand k ten opzichte van zichzelf verschoven waarnemingsreeks voorstelt (KENDALL and STUART, 1966, -405-).

Zijn we nu geinteresseerd in het effectieve aantal waarnemingen n' waarop de variantie van een schatter berust, dan gaat

$$\sigma^2_{\bar{\underline{x}}} = \frac{\sigma^2_{\underline{x}}}{n}$$

over in

$$\sigma^2_{\bar{\underline{x}}} = \frac{\sigma^2_{\underline{x}}}{n'}$$

dus

$$\text{variantie}\left(\frac{\sum \underline{x}}{n}\right) = \frac{\sigma^2_{\underline{x}}}{n'}$$

Voor het linkerlid kan met (15.1) worden afgeleid dat

$$\text{variantie}\left(\frac{\sum \underline{x}}{n}\right) \simeq \frac{\sigma^2_{\underline{x}}}{n} \frac{1 + \rho}{1 - \rho} \qquad \text{(n groot)}$$

zodat

$$n' \simeq n \, . \, \frac{1 - \rho}{1 + \rho}$$

(zie bijv. VAN MONTFORT, 1973).

LLOYD (1970) vond voor hetzelfde model dat de variantie van de herhalingsreeks T in geval van autocorrelatie bij overschrijdingskans p wordt

$$\sigma^2_{\underline{T}} = \frac{1 - p}{p^2} \frac{1 + \rho}{1 - \rho}$$

waarin dezelfde correctieterm voorkomt. Het toepassen van het effectieve aantal gegevens veroorzaakt een vergroting van de variantie en dus een breder worden van betrouwbaarheidsintervallen.

15. GECOMBINEERDE EXPONENTIELE VERDELINGEN

VAN DER MADE (1963) geeft een afleiding van de voorwaardelijke verwachting in het geval dat de populatie met twee exponentiële verdelingen moet worden beschreven. De kansverdeling luidt dan

$$_1F_>(x) = e^{-\lambda_1 x} \qquad 0 \leq \underline{x} < x_s$$

$$_2F_>(x) = e^{-\lambda_2 (x-C)} \qquad x_s \leq \underline{x} < \infty$$

waarin C moet voldoen aan

$$C = \frac{\lambda_2 - \lambda_1}{\lambda_2} x_s$$

waarin x_s het snijpunt van beide verdelingen is.

Voor de voorwaardelijke verwachting geldt dan

$$E(\underline{x} \mid \underline{x} \geq x_o) = \frac{1}{\lambda_1} + e^{-\lambda_1 (x_s - x_o)} \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right), \quad x_o < x_2$$

$$= \frac{1}{\lambda_2}, \qquad x_s \leq x_o$$

In een voorbeeld voor de Maas maakt VAN DER MADE (1963) het bestaan van zulk een combinatie aannemelijk.

Hier kunnen nog enkele aantekeningen worden gemaakt.

In de eerste plaats wordt opgemerkt dat de voorwaardelijke verdeling dezelfde parameterwaarde heeft als de absolute verdeling (zie par. 4). Vindt men verschillen dan zal getoetst moeten worden of deze significant zijn teneinde tot een combinatie van verdelingen te besluiten.

In de tweede plaats moet worden opgemerkt dat de waarde van x_s niet rechtstreeks uit de steekproef mag worden bepaald aangezien men dan een bijzonder punt opspoort en vervolgens toetst of het bijzonder is. De beste procedure lijkt hier aan beide zijden van x_s een aantal punten buiten beschouwing te laten en de overige voor aanpassing respectievelijk toetsing te gebruiken.

APPENDIX 1. DE SAMENHANG MET DE χ^2-VERDELING

De in deze nota behandelde kansverdeling kans als bijzonder geval van de χ^2-verdeling worden beschouwd. Ter completering van de theorie zal dit verband hier nog in het kort worden weergegeven.

De dichtheid van de χ^2-verdeling luidt

$$dP(\underline{\chi}_n^2 < a) = \frac{1}{\Gamma(\frac{n}{2})} \left(\frac{\chi^2}{2}\right)^{\frac{1}{2}n-1} e^{-\frac{1}{2}\chi^2} \frac{1}{2} d(\chi^2)$$

hierin wordt n het aantal vrijheidsgraden genoemd.

We transformeren als volgt, stel n = 2N

$$dP(\underline{\chi}_{2N}^2 < a) = \frac{1}{\Gamma(N)} \left(\frac{\chi^2}{2}\right)^{N-1} e^{-\frac{1}{2}\chi^2} \frac{1}{2} d(\chi^2) \qquad (1)$$

Hierin is N het aantal gesommeerde onafhankelijke kansvariabelen dat beschouwd wordt, dus is N een geheel getal en is $\Gamma(N) = (N-1)!$

Beschouw nu, voor N = 1, de verdeling van $\frac{1}{2}\chi_2^2$, de dichtheid hiervan is

$$dP(\tfrac{1}{2}\underline{\chi}_2^2 < a) = e^{-\frac{1}{2}\chi^2} d(\tfrac{1}{2}\chi^2)$$

wat de exponentiële verdeling oplevert zodat met de transformatie

$$\frac{1}{2}\underline{\chi}_2^2 = \lambda\underline{x}$$

na invulling in de χ^2-verdeling

$$dP(\lambda\underline{x} < a) = dP(\underline{x} < x) = e^{-\lambda x}\lambda dx$$

het resultaat uit (3.2) wordt verkregen. Er volgt dus uit dat

$$\lambda\underline{x} \doteq \frac{1}{2}\underline{\chi}_2^2$$

wat betekent dat $\lambda\underline{x}$ isomoor is met (dezelfde kansverdeling heeft als) $\frac{1}{2}\chi^2$ met 2 vrijheidsgraden.

Ook geldt dan

$$2\ \lambda\underline{x}\ \underline{\underline{i}}\ \underline{\chi}_2^2$$

De momenten van de χ^2-verdeling kunnen in de parameter n worden uitgedrukt. Deze luiden (CRAMER, 1961)

$$\text{verwachting}\quad E(\underline{\chi}^2) = n$$

$$\text{variantie}\quad E(\underline{\chi}^2 - n)^2 = 2n$$

Toepassing hiervan levert op, voor N = 1 dus n = 2

$$E(2\ \lambda\underline{x}) = 2 \qquad \text{dus} \qquad E(\underline{x}) = \frac{1}{\lambda}$$

en

$$E(2\ \lambda\underline{x} - 2)^2 = 4 \qquad \text{dus} \qquad E(\underline{x} - \frac{1}{\lambda})^2 = \frac{1}{\lambda^2}$$

Deze uitkomsten werden reeds afgeleid in (5.1) en (5.3). Opgemerkt wordt dat dus is gebleken dat $\bar{x}$ een zuivere schatter is voor $1/\lambda$, maar dat $\frac{1}{\bar{x}}$ geen zuivere schatter is voor λ. Nu wordt λ systematisch te groot geschat (zie par. 8c). Wel geldt dat $\bar{x}$ een M.L. schatter is voor $1/\lambda$ en dat $\frac{1}{\bar{x}}$ een M.L. schatter is voor λ. Deze eigenschap geldt algemeen voor M.L. schatters (MOOD and GRAYBILL, 1963), maar ze behoeven dus niet beide tegelijk zuiver respectievelijk onzuiver te zijn.

Beschouwen we nu N onafhankelijke kansvariabelen, dan geldt voor hun som

$$2\ \lambda \sum_{i=1}^{N} \underline{x}_i\ \underline{\underline{i}}\ \underline{\chi}_{2N}^2$$

De dichtheid voor deze verdeling luidt dus door de transformatie

$$\underline{\chi}^2_{2N} = 2\,\lambda \sum \underline{x}_i = 2\,\lambda \underline{z}$$

toe te passen op (1)

$$dP(\underline{z} < z) = \frac{1}{(N-1)!}\,\frac{(2\,\lambda z)^{N-1}}{2^{N-1}}\,e^{-\lambda z}\,\frac{1}{2}\,d(2\,\lambda z)$$

$$= \frac{\lambda^N z^{N-1}}{(N-1)!}\,e^{-\lambda z}$$

hetgeen de in (8.3) gevonden Erlang verdeling is.

Opgemerkt wordt nog dat weer eenvoudig blijkt dat

$$E(2\,\lambda \underline{z}) = E(\underline{\chi}^2_{2N}) = n = 2N$$

dus

$$E(\underline{z}) = \frac{N}{\lambda}$$

In paragraaf 8b werd afgeleid dat

$$E(\frac{1}{\underline{z}}) = \frac{\lambda}{N-1}$$

Voor grote steekproeven kan worden afgeleid (KENNY and KEEPING, 1959) dat

$$\underline{\chi}^2_{2N} \doteq 2N + \sqrt{4N}\,\underline{\chi}$$

waarin $\underline{\chi}$ de normaal verdeelde kansvariabele representeert met $(\mu, \sigma) = (0,1)$ (Centrale Limiet stelling).

Betrouwbaarheidsgrenzen kunnen nu als volgt uit de schattingen worden afgeleid (HEMELRIJK, 1963). Kies een onbetrouwbaarheidsdrempel α (bijv. 5 %), dan kan een interval worden bepaald waardoor

$$P(\ a \ < \ \underline{\chi}^2_{2N} \ < \ b\ ) = (1 - \alpha) \tag{2}$$

of $$P(\ a \ < 2\ \lambda \sum \underline{x}_i \ < \ b\ ) = (1 - \alpha)$$

en $$P\left(\frac{a}{2\sum \underline{x}_i} < \ \lambda \ < \frac{b}{2\sum \underline{x}_i}\right) = (1 - \alpha)$$

Er gold

$$\underline{\hat{\lambda}} = \frac{N}{\sum \underline{x}_i}$$

zodat $$P\left(\frac{1}{2}\ a\ \frac{\underline{\hat{\lambda}}}{N} < \ \lambda \ < \frac{1}{2}\ b\ \frac{\underline{\hat{\lambda}}}{N}\right) = (1 - \alpha) \tag{3}$$

Hierin zijn a en b gerelateerd aan de $\underline{\chi}^2_{2N}$-verdeling volgens (2) en is een exacte oplossing verkregen zoals die in het RAPPORT DELTA COMMISSIE (1960) is toegepast. Opgemerkt wordt dat het interval verkregen met (3) een s t o c h a s t i s c h interval is. Gezegd kan worden dat het interval met 95 % kans de werkelijke parameterwaarde λ zal bevatten (N i e t: dat λ met 95 % kans in het interval ligt, want λ is geen stochastische variabele). Algemeen voor een interval I geldt namelijk $P(\lambda \in \underline{I}) \geq (1 - \alpha)$.

De normale benadering met de Centrale Limietstelling levert de volgende praktische oplossing:

$$P(\ a \ < \ \underline{\chi}^2_{2N} \ < \ b\ ) = (1 - \alpha)$$

of $$P(\ a \ < 2N + \sqrt{4N}\ \underline{\chi} < \ b\ ) = (1 - \alpha)$$

zodat $$P\left(\frac{a - 2N}{\sqrt{4N}} < \ \underline{\chi} \ < \frac{b - 2N}{\sqrt{4n}}\right) = (1 - \alpha)$$

Een symmetrisch interval op basis van $\alpha = 0{,}05$ wordt

$$P(-\ 1{,}96 < \underline{\chi} < +\ 1{,}96) = 0{,}95$$

waaruit volgt

$$a = 2N - 1{,}96\ \sqrt{4N}$$

$$b = 2N + 1{,}96\ \sqrt{4N}$$

Ingevuld in (3) wordt verkregen

$$\frac{1}{2}\,\frac{a}{N}\,\hat{\underline{\lambda}} = \left(1 - \frac{1{,}96}{\sqrt{N}}\right)\hat{\underline{\lambda}}$$

$$\frac{1}{2}\,\frac{b}{N} = \left(1 - \frac{1{,}96}{\sqrt{N}}\right)\hat{\underline{\lambda}}$$

Hetgeen voor het betrouwbaarheidsgebied dezelfde waarden voor de grenzen van het interval oplevert als werd gevonden in (10.2)

APPENDIX 2. DE SAMENHANG MET DE VERDELING VAN EXTREMEN

Behalve naar de afvoer die op een enkele dag kan worden overschreden, kan men vragen naar de afvoer die gedurende 2, 3, ..., k dagen wordt overschreden.

Stel de overschrijdingskans op p_o, dan is de kans op herhaling van de gebeurtenis p_o^2, voor optreden van het verschijnsel op 3 dagen p_o^3 enz. Aangenomen wordt dat een zodanige keuze gedaan wordt dat de opeenvolgende gebeurtenissen onafhankelijk zijn. Nu geldt dus voor het éénmaal optreden van de gebeurtenis $\underline{G} = (\underline{x} > x)$

$$P(\underline{x}_1 \geq x) = e^{-\lambda x}$$

Voor 2 malen het optreden van $\underline{G}$

$$P(\underline{x}_1 \geq x \cap \underline{x}_2 \geq x) = (e^{-\lambda x})^2$$

Deze laatste gebeurtenis is identiek met $\underline{G}_2 = \underline{x}_{(1)} \geq x$, waarin $\underline{x}_{(1)}$ de kleinste (eerste) waarde is van de naar volgorde van grootte geordende steekproef. Evenzo is

$$P(\underline{x}_1 \geq x \cap \ldots \cap \underline{x}_k \geq x)$$

$$= P(\underline{x}_{(1)} \geq x) = (e^{-\lambda x})^k = e^{-k\lambda x} \qquad (1)$$

Wordt in het laatste geval geschreven

$$P_k(\underline{x}_{(1)} \geq x) = e^{-\lambda_k x} \qquad (2)$$

dan zal blijken dat bij keuze van de kleinste afvoer uit k dagen en bij uitzetten volgens (3.5) de waarde van λ_k evenredig is met het aantal beschouwde dagen k volgens

$$\lambda_k \doteq k\lambda \qquad (3)$$

waarin λ de parameterwaarde van de oorspronkelijke ($k = 1$) verdeling is.

Vergelijking (3) geldt echter voor niet te grote waarden van k. Uit (1) zou volgen dat $\lim_{k \to \infty} P_k = 0$. Dit zou dan gelden voor alle $x > 0$ en men zou uiteindelijk moeten besluiten dat met kans 1 $\underline{x}_{(1)} \to$ constante voor $k \to \infty$, zodat uiteindelijk met grote waarden van k op deze wijze berekend, de kansverdeling degenereert. Van belang is nu te onderzoeken of voor de uitdrukking een limiet-waarde bestaat. Deze wordt gevonden door een stabiliteitsvoorwaarde te stellen welke kan luiden dat het minimum van de kleinste waarde dezelfde verdeling moet hebben als de kleinste zelf, op mogelijk een lineaire transformatie na. Hieruit ontstaat de functie vergelijking

$$F^k(x) = F(a_k x + b_k)$$

Bewezen kan worden dat deze relatie slechts drie oplossingen kent, de zogenaamde asymptotische verdelingen (GUMBEL, 1954; 1966). Voor variabelen die een grenswaarde bereiken, zoals hier voor minimum afvoeren het geval is (bijv. $\underline{x}_{(1)} \to 0$ respectievelijk $\underline{x}_{(1)} \to x_o$ als $k \to \infty$) geldt dan

$$\lim_{k \to \infty} P_k(\underline{x}_{(1)} \geq x) \to e^{-\left(\frac{x-\gamma}{\beta-\gamma}\right)^{\alpha}} = P, \quad x \geq \gamma, \; \alpha > 0, \; \beta > \gamma$$

$$\gamma \geq 0$$

(GUMBEL, 1960, -165-, -278-).

Geschreven kan worden voor de kans P

$$\ln(-\ln P) = \alpha \ln(x - \gamma) - \alpha \ln(\beta - \gamma)$$

Deze uitkomst geeft aan dat $(x - \gamma)$ tegen P uitgezet moet worden op logarithmisch - Gumbel - grafiekpapier, dus de kans dubbellogarithmisch en $(x - \gamma)$ logarithmisch. Aangezien $(x - \gamma)$ onder de logarithme voorkomt moet door trial-and-error die waarde van γ gezocht worden die tot een rechte aanleiding geeft waarmee dan de parameters kunnen worden bepaald (YEVJEVICH, 1972).

LITERATUUR

CRAMER, H., 1961. Mathematical methods of statistics. Princeton University Press, Princeton.

GUMBEL, E.J., 1954. Statistical Theory of Extreme values and some Practical Applications. Nat. Bureau of Standards. Applied Mathem. Series 33. U.S. Dept. of Commerce, Washington.

——— 1960. Statistics of Extremes. Columbia University Press, New York.

——— 1966. Extreme Value Analysis. In: Statistical Methods in Hydrology. Proceedings of Hydrology Symposium No. 5. McGill University, Montreal.

HARNETT, D.L., 1970. Introduction to statistical methods. Addison-Wesley, London.

HEMELRIJK, J., 1963. Waarschijnlijkheidsrekening en statistiek. In: Handboek der Wiskunde, Kuipers, L. et al Ed. Hfdst. XIV pp 749-833. Scheltema, Amsterdam.

HÜTTE, 1959. Mathematische Formeln und Tafeln. Wilhelm Ernst und Sohn. Berlin

KENDALL, M.G. and A. STUART, 1961. The advanced theory of statistics. Vol 2. Inference and relationship. Griffin, London.

——— 1966. Idem Vol 3. Design and analysis and time series.

KENNY, J.F. and KEEPING, E.S., 1959. 'Mathematics of Statistics, Prt II'. D. van Nostrand Comp. New York

LINDGREN, B.W. and McELRATH, G.W., 1967. Introduction to probability and statistics (2nd ed) Collier MacMillan, London.

LLOYD, E.H., 1970. Return periods in the presence of persistence. J. of Hydrology 10, pp 291-298.

MADE, J.W. VAN DER, 1963. Grondslagen en toepassingen van de rechtlijnig-exponentiële verdelingsfunctie. Rijkswaterstaat, Directie Waterhuishouding en waterbeweging, 's-Gravenhage.

MOOD, A.M. and F.A. GRAYBILL, 1963. Introduction to the theory of statistics McGraw-Hill, New York.

MONTFORT, M.A.J. VAN, 1973. Bijzondere kansverdelingen. Syllabus L.H. Wageningen.

RAPPORT DELTA COMMISSIE, 1960. Beschouwingen over stormvloeden en getijbeweging. Deel 3. Bijdrage II.1 Extrapolatie van de overschrijdingslijn der hoogwaterstanden te Hoek van Holland met behulp van geselecteerde stormen (pp. 8-56) Mathematisch Centrum Amsterdam. Staatsdrukkerij 's-Gravenhage.

SIEGEL, S., 1956. Non-parametric statistics for the behavioral sciences. McGraw-Hill. New York.

STAM, A.J., 1964. Inleiding tot de waarschijnlijkheidsrekening. De Technische Uitgeverij H. Stam., Haarlem.

STOL, Ph.Th., 1970. Het vergelijken van empirische frequentie-verdelingen met een toepassing op reeksen neerslaggegevens uit de Gelderse Achterhoek. Cult. Techn. Tijdschrift 9 Nr 4. ICW Med. 129 (1-16).

——— 1972. Een beschouwing over de frequentie van weerkeren van hydrologische gebeurtenissen. Cult. Techn. Tijdschrift 11 Nr 4 (168-185) ICW Verspr. Overdr. 125.

WERKGROEP AFVLOEIINGSFACTOREN, 1970. Tweede Interimrapport. Sectie voor Cultuurtechniek, Kon. Inst. Ing. en Studiekring voor Cultuurtechniek. Kon. Gen. Landb. Wet.

YEVJEVICH, V., 1972. Probability and statistics in hydrology. Water Resources Publications. Colorado.