NN31545.0746

juni 1973

Instituut voor Cultuurtechniek en Waterhuishouding
Wageningen

BIBLIOTHEEK
STARINGGEBOUW

BIBLIOTHEEK DE HAAFF
Droevendaalsesteeg 3a
Postbus 241
6700 AE Wageningen

EEN INLEIDENDE NOTA OVER

STOCHASTISCHE SIMULATIE

ir Ph.Th. Stol

Nota's van het Instituut zijn in principe interne communicatiemiddelen, dus geen officiële publikaties.
Hun inhoud varieert sterk en kan zowel betrekking hebben op een eenvoudige weergave van cijferreeksen, als op een concluderende discussie van onderzoeksresultaten. In de meeste gevallen zullen de conclusies echter van voorlopige aard zijn omdat het onderzoek nog niet is afgesloten.
Bepaalde nota's komen niet voor verspreiding buiten het Instituut in aanmerking

1792721

CENTRALE LANDBOUWCATALOGUS
0000 0941 0396

I N H O U D

1. INLEIDING

Toevalsgetallen of realisaties van stochastische variabelen kunnen worden gebruikt om de aanwezigheid van toevalsfluctuaties in getallenrijen na te bootsen. Gedacht kan worden aan getallen die, volgens toeval, rond een gemiddelde waarde fluctueren als nabootsing van waarnemingen van dat gemiddelde, maar ook aan reeksen die bedoeld zijn het verloop van een variabele in de tijd als stochastische reeks weer te geven.

Voor elk doel zal men specifieke eisen stellen waaraan de reeks moet voldoen, zoals

- De termen van de reeksen dienen te voldoen aan een gewenste kansverdeling met gegeven verwachtingswaarde en spreiding.
- De termen dienen te voldoen aan deze eerste eis en bovendien aan de eis dat ze onderling, of met een andere reeks, gecorreleerd zijn met een gegeven waarde van de correlatiecoëfficiënt.

In deze nota zullen enkele aspecten van het genereren en gebruiken van dit soort reeksen worden besproken.

2. NOMENCLATUUR

Op de gebruikelijke wijze zullen stochastische variabelen door onderstreping worden aangegeven. Is p de kans (P) dat de stochastische grootheid $\underline{x}$ niet groter zal zijn dan een waarde x dan wordt dit aangegeven met

$$P(\underline{x} \leq x) = p \qquad (1)$$

Is f(x) de kansdichtheidsfunctie van x, dan luidt (1)

$$P(\underline{x} \leq x) = \int_{-\infty}^{x} f(x)\, dx \qquad (2)$$

De verwachtingswaarde van x zal worden aangeduid met μ of $\mu_{\underline{x}}$. Deze verwachtingswaarde kan worden uitgedrukt in de parameters van de kansverdeling (2) door te berekenen

$$E(\underline{x}) = \int_{-\infty}^{+\infty} x\, f(x)\, dx$$

Voor de variantie wordt geschreven σ^2 of $\sigma^2_{\underline{x}}$ en uitdrukking in de parameters van de kansverdeling vindt plaats door oplossing van de integraal

$$E(\underline{x} - \mu)^2 = \int_{-\infty}^{+\infty} (x - \mu)^2\, f(x)\, dx$$

De correlatiecoëfficiënt tussen twee stochastische grootheden $\underline{x}$ en $\underline{y}$ wordt aangegeven met $\rho(\underline{x}, \underline{y})$ of $\rho_{\underline{xy}}$ eventueel met ρ. Berekening in de parameters van de gezamenlijke kansverdeling vindt plaats met de covariantie waarvoor geldt

$$Cov(\underline{x}, \underline{y}) = E(\underline{x} - \mu)\ (\underline{y} - \eta)$$

en

$$\rho(\underline{x}, \underline{y}) = \frac{Cov(\underline{x}, \underline{y})}{\sigma_{\underline{x}} \sigma_{\underline{y}}} \qquad (2a)$$

waarin $\eta = E(\underline{y})$

Voorts zal het gemakkelijk blijken om gebeurtenissen met een apart symbool aan te duiden bijvoorbeeld $\underline{G}$. Is de kans (P) dat de gebeurtenis $\underline{G}$ zich voordoet gelijk aan p dan wordt dat weergegeven met

$$P(\underline{G}) = p$$

Het bepalen van een getal dat uit een populatie met een gewenste verdeling stamt wordt genoemd het t r e k k e n uit die verdeling. De waarde van het getrokken getal is dan een r e a l i s a t i e van de stochastische variabele.

3. GRONDSLAGEN VAN STOCHASTISCHE SIMULATIE

Stochastische grootheden kunnen worden gegenereerd met een toevalsgenerator, bijvoorbeeld door te dobbelen, lootjes te trekken en dergelijke. Ook wordt gebruik gemaakt van tabellen van bijvoorbeeld de normale verdeling.

Voor gebruik in het groot van deze laatste mogelijkheid zou het noodzakelijk zijn een tabel in de computer in te voeren wat een groot nadeel is door het bezetten van veel geheugenruimte met enkele honderden of meer toevasgetallen.

Nadat met een rekensysteem bepaald is op welke 'toevallige' plaats in de tabel zal worden begonnen, ligt de volgorde van de getallen verder vast. Is de tabel ten einde dan ontstaat weer opnieuw dezelfde reeks door herhaling van gebruik van dezelfde waarden uit de tabel.

Het nadeel, dat bij keuze van eenzelfde beginpunt exact dezelfde reeks wordt verkregen kan overigens een praktisch voordeel opleveren. In het teststadium van een programma kan het nuttig zijn in achtereenvolgende bewerkingen dezelfde 'random' reeks te creëren om steeds dezelfde numerieke uitkomsten te krijgen. Nadat het programma voldoende getest is kan men overgaan tot reeksen die niet meer aan elkaar gelijk zijn door andere beginpunten in de tabel als startwaarde te kiezen.

Bij toepassing in het groot worden stochastische variabelen gegenereerd door middel van de computer zelf. Toevalselementen kunnen dan direct in een rekenproces worden opgenomen en verwerkt.

Hoewel strikt genomen de gegenereerde reeksen een deterministi-

sche oorsprong hebben kunnen door een verantwoorde toepassing van de theorie van de kansrekening procedures ontworpen worden die reeksen genereren waarvan de opeenvolgende getallen in hoge mate 'onvoorspelbaar' zijn. Dergelijke reeksen noemt men wel pseudo-random.

De grondslag van dit soort generatieprocessen is in de regel de berekening van een transcedente functie waarvan de laatste decimalen als toevallige grootheden (afrondingsfouten) worden beschouwd. Een dergelijke uitkomst wordt weer in de functie ingevoerd en geeft dan het volgende toevalsgetal.

Hier volgt uit dat bij een vast begingetal exact dezelfde reeks wordt verkregen zodat zeker ook hier niet van geheel-toevallige-reeksen gesproken kan worden. Wel blijkt dat de getallen in zulk een reeks eigenschappen bezitten die sterk overeenkomen met 'echte' toevalsgetallen. Hiermede is het gebruik van deze methoden in de praktijk gerechtvaardigd.

Stel er wordt de laatste decimaal van de uitkomst van een transcedente functie gebruikt. De mogelijke uitkomsten zijn dan

$$\underline{k} = i,\ i = 0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9$$

Voor grote aantallen zal blijken dat de frequentie f_d van voorkomen van $k = i$ bij gebruik van de laatste cijfers voldoet aan

$$f_1(\underline{k} = i) \to \frac{1}{10},\ i = 0,\ 1,\ \ldots,\ 9$$

Evenzo geldt voor de laatste 2 decimalen

$$f_2(\underline{k} = i) \to \frac{1}{100},\ i = 0,\ 1,\ \ldots,\ 99$$

Algemeen kan geschreven worden, met d laatste decimalen

$$f_d(\underline{k} = i) \to \frac{1}{10^d},\ i = 0,\ 1,\ \ldots,\ 10^d - 2,\ 10^d - 1$$

Deze frequenties beschrijven voor de praktijk voldoende correct de zogenaamde uniforme verdeling waarvan de dichtheidsfunctie een

constante c is en dus niet van k afhangt.

4. ENKELE OPMERKINGEN OVER HET GEBRUIK VAN STOCHASTISCHE REEKSEN

Reeds is aangegeven dat van een stochastische reeks geëist moet worden dat deze een gewenste kansverdeling weergeeft. Beter: dat de verkregen trekkingen van toevalsgetallen afkomstig gedacht kunnen worden uit de gewenste kansverdeling.

Het zal duidelijk zijn dat bij elk fenomeen een groot aantal stochastische grootheden ieder met hun eigen kansverdeling kan worden onderscheiden. Voor de neerslag bijvoorbeeld

- k-daagse neerslagsommen per maand
- runs van droge dagen
- het aantal dagen na een bepaalde datum waarin voor het eerst een k-daagse som van gegeven grootte wordt overschreden
- enz.

De benodigde reeksen voor simulatie kunnen pas worden verkregen nadat het te analyseren verschijnsel duidelijk gedefinieerd is, gegevens over voldoend lange tijd verzameld zijn en hiervan de kansverdeling is geschat.

Stochastische simulatie voegt dus geen nieuwe informatie toe. Alle informatie in de waarnememingsreeksen aanwezig wordt gebruikt voor het schatten van de kansverdeling. Deze is dus zelf een steekproefuitkomst die gebruikt gaat worden voor het genereren van nieuwe reeksen uitkomsten die dezelfde kansverdeling als de steekproefverdeling heeft. De kansverdeling van het geanalyseerde fenomeen wordt met de gegenereerde reeksen dus niet betrouwbaarder vastgesteld. Ook de zogenaamde staarten van de verdeling kan men op deze wijze niet verbeteren. Bij het genereren krijgen waarden in de staarten een frequentie die overeenkomt met de kans geschat met de frequentie van de oorspronkelijke reeks.

Toevalsgetallen kunnen gebruikt worden om combinaties van gebeurtenissen op hun frequentie van voorkomen te onderzoeken (zie fig. 1).

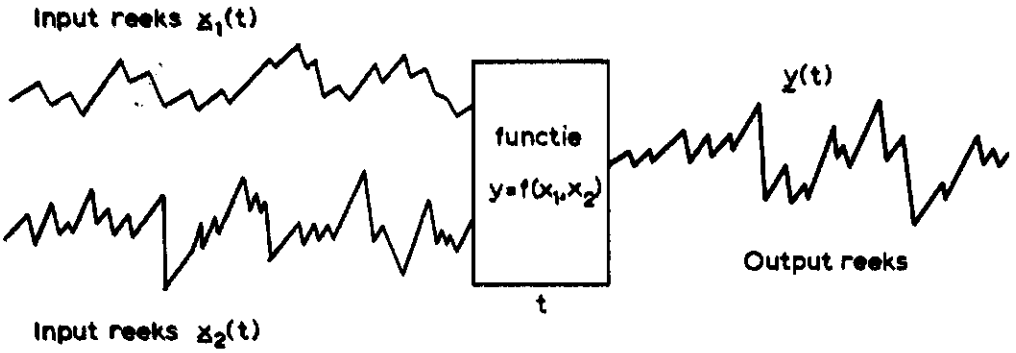


Fig. 1. Twee stochastische reeksen $\underline{x}_1$ en $\underline{x}_2$ worden gegenereerd. Realisaties op tijdstip t worden getransformeerd met de functie $y(t) = f(x_1(t), x_2(t))$ tot een waarde van y. Van $\underline{y}(t)$ kan een empirische frequentieverdeling worden opgesteld

Overigens kan men bijvoorbeeld bij gebruik van neerslagreeksen ook historische (vroeger reeds gemeten) reeksen toepassen op nieuwe situaties. Eventueel zouden bijvoorbeeld steeds de maanden uit het beschikbare materiaal geloot kunnen worden teneinde nieuwe volgorden te creëren. Dit soort overwegingen is gebaseerd op het niet-gecorreleerd zijn van neerslagwaarnemingen voor tijdsintervallen groter dan 2 à 3 dagen.

Voor grootheden die regelmatig aan kunstmatige verandering onderhevig zijn heeft stochastische simulatie geen betekenis. Zo is bijvoorbeeld het genereren van 50 jaar beekafvoeren zinloos indien bekend is dat elke 10 à 15 jaar in het stroomgebied waterhuishoudkundige wijzigingen worden aangebracht die een ander afvoerregime tot gevolg hebben.

5. TOEPASSING BIJ MODELBEREKENING

Stochastische simulatie kan ook worden toegepast in aansluiting op hydrologische modelberekening. Het hydrologisch model verantwoordt de deterministische relatie met de inputgegevens. De parameters van

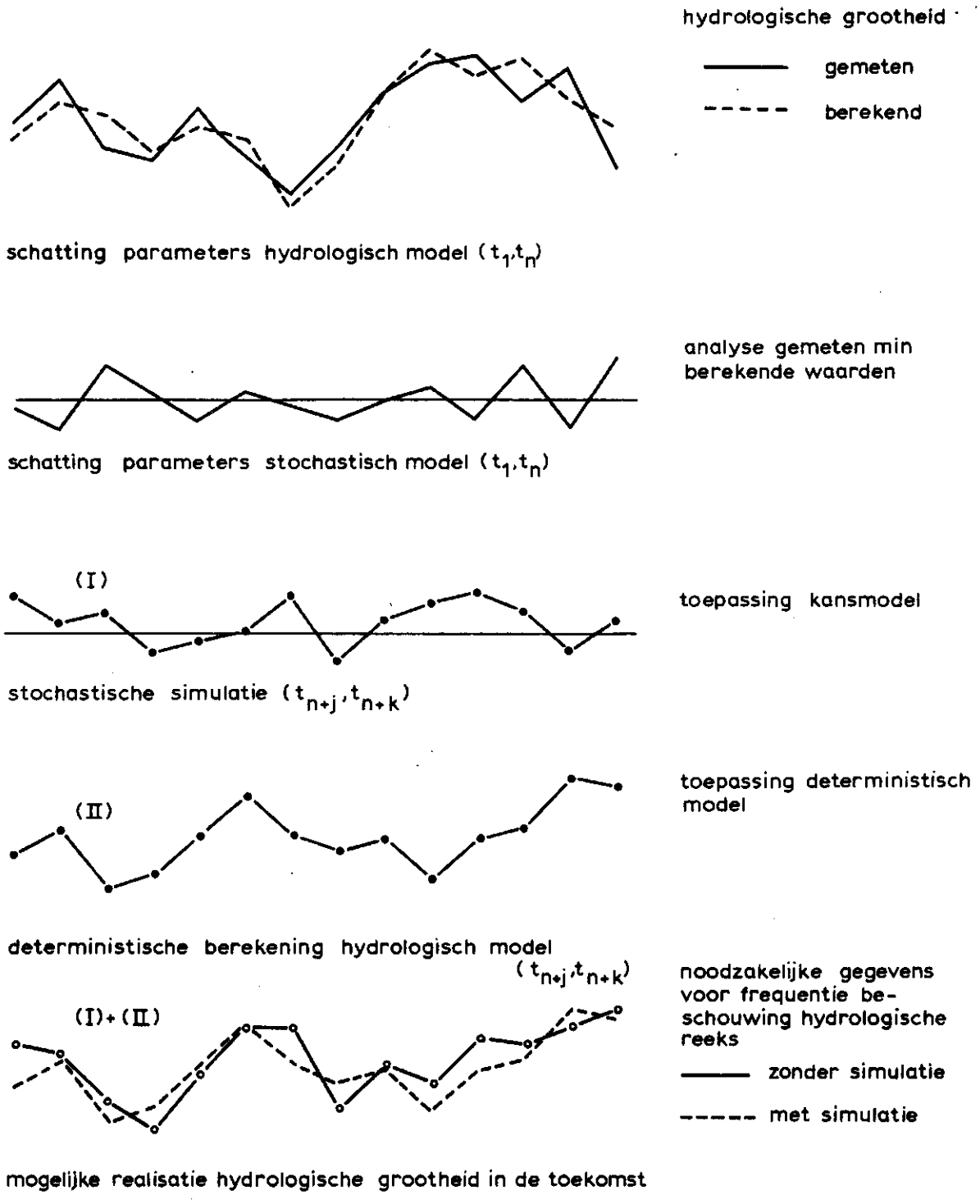


Fig. 2. Voorbeeld toepassing stochastische simulatie op modelberekening

De kansdichtheid van de uniforme verdeling is constant (c) onafhankelijk van de variabele x. Op het interval a, b geldt dus de volgende kansverdeling

$$P(a \leq \underline{x} \leq x) = \int_a^x c \, dx, \quad x \leq b$$

De constante c kan worden bepaald uit het feit dat de totale kans (integratie over het gehele waardenbereik van x) gelijk is aan 1. Dus

$$P(a \leq \underline{x} \leq b) = \int_a^b c \, dx = 1$$

waaruit volgt dat $c = \frac{1}{b - a}$ de dichtheidsfunctie is, uitgedrukt in de eindpunten van het interval $[a, b]$.
In fig. 3 staat de verdeling weergegeven en wordt daar vergeleken met de normale verdeling.

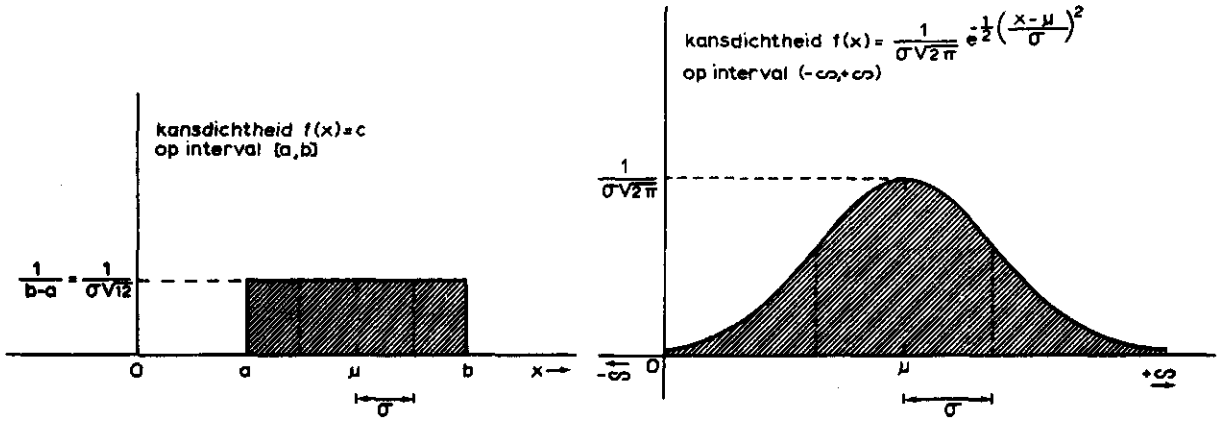


Fig. 3. De uniforme verdeling en de normale verdeling

De verwachting μ volgt uit

$$E(\underline{x}) = \frac{1}{b - a} \int_a^b x \, dx = \frac{1}{2} \frac{b^2 - a^2}{b - a} = \frac{1}{2}(a + b) \qquad (3)$$

het hydrologisch model kunnen worden geschat met behulp van waarnemingsuitkomsten van een hydrologische grootheid zoals bijvoorbeeld afvoer of grondwaterstand.

Op deze wijze ontstaat een gemeten en een berekende waarde (fig. 2) waarvan het verschil geen deterministische componenten meer bevat doch alleen nog toevalsfluctuaties. Deze toevalsfluctuaties kunnen ontstaan gedacht worden doordat de eigenschappen van het actuele meetpunt niet exact door het model beschreven worden maar op onderdelen (toevallig) zal afwijken.

De gemeten waarden in het meetpunt zijn echter realiteiten die een deterministische en een stochastische component hebben.

De stochastische component wordt zichtbaar door gemeten en berekende waarden van elkaar af te trekken. De stochastische component kan nu geanalyseerd worden en na vaststelling van de kansverdeling en schatting van de parameters kan worden overgegaan tot stochastische simulatie en verlenging van de toevalsreeks na het interval van waarneming (t_1, t_n) bijvoorbeeld over een nieuw tijdsinterval (t_{n+j}, t_{n+k}).

Deze stochastische component kan nu bij de berekende waarden bijgeteld worden. Er ontstaat nu een reeks waarin zowel de deterministische als de stochastische component verantwoord zijn. Deze reeks kan weer gebruikt worden voor het vaststellen van overschrijdingsfrequenties zoals in fig. 1 in beeld is gebracht. Behalve door de betrekking $y = f(x_1, x_2)$ wordt de output nu verkregen door $\underline{y} = f(x_1, x_2) + \underline{z}$ waarin $\underline{z}$ de stochastische rest component vertegenwoordigt.

De gehele hier besproken procedure is in fig. 2 schematisch weergegeven.

6. DE UNIFORME VERDELING

Uit het voorgaande (par. 3) valt op te maken dat de uniforme verdeling een belangrijke rol speelt bij het genereren van toevalsreeksen. De eigenschappen van deze verdeling zullen in het kort worden besproken.

wat het zwaartepunt van het interval oplevert.

De variantie σ^2 van deze verdeling wordt gevonden uit

$$E(\underline{x} - \mu)^2 = E(\underline{x}^2) - \mu^2$$

$$= \frac{1}{b - a} \int_a^b x^2 \, dx - \frac{1}{4}(a + b)^2$$

$$= \frac{1}{3} \frac{b^3 - a^3}{b - a} - \frac{1}{4}(a + b)^2$$

$$= \frac{4b^2 + 4ab + 4a^2 - 3a^2 - 6ab - 3b^2}{15} = \frac{(b - a)^2}{12} \qquad (4)$$

Hieruit is op eenvoudige wijze een stochastische grootheid te transformeren met gewenste verwachting en spreiding. Stel

$$\underline{Z} = \frac{\underline{x} - \mu}{\sigma}$$

dan is

$$E(\underline{Z}) = 0 \quad \text{en} \quad E(\underline{Z} - \mu_{\underline{Z}})^2 = 1$$

Neem dus

$$\underline{Z} = \frac{\underline{x} - \frac{1}{2}(a + b)}{\frac{(b - a)^2}{12}} = \frac{12\,\underline{x} - 6(a + b)}{(b - a)^2}$$

Stel dat het gewenst is over een stochastische variabele $\underline{y}$ te beschikken met de volgende eigenschappen.

Verdeling van $\underline{y}$ uniform met

$$E(\underline{y}) = \eta$$

$$E(\underline{y} - \eta)^2 = \tau^2$$

Dan kan gekozen worden als transformatie

$$\underline{y} = \eta + \tau \underline{Z}$$

namelijk

$$E(\underline{y}) = \eta + \tau E(\underline{Z}) = \eta$$

en

$$E(\underline{y} - \eta)^2 =$$

$$= E(\eta + \tau \underline{Z} - \eta)^2 = \tau^2 E(\underline{Z}^2) = \tau^2$$

Zodat uiteindelijk, bij trekking van $\underline{x}$ op [a, b] getransformeerd wordt tot

$$\underline{y} = \eta + \tau \frac{12\underline{x} - 6(a + b)}{(b - a)^2}$$

Het is gebruikelijk de subroutines voor het genereren van toevalsgetallen te conditioneren op het interval $a = 0$, $b = 1$, waardoor verkregen wordt

$$\mu = \frac{1}{2} \quad \sigma = \sqrt{\frac{1}{12}} = 0.2886 \tag{5}$$

en

$$\underline{y} = \eta + \tau(12\underline{x} - 6)$$

$$= \eta - 6\tau + 12\tau\underline{x} \tag{6}$$

waaruit ook rechtstreeks volgt:

$$E(\underline{y}) = \eta - 6\tau + 12\tau E(\underline{x}) = \eta$$

en, met (6)

$$E(\underline{y} - \eta)^2 = E(-6\tau + 12\tau\underline{x})^2$$

$$= 144\tau^2 \, E(\underline{x} - \tfrac{1}{2})^2$$

$$= 144\tau^2 \cdot (\tfrac{1}{12})^2 = \tau^2$$

wat de gewenste eigenschappen zijn.

Opmerking

De betrekking tussen de intervalgrenzen $[a, b]$ en de waarden η en τ is eenduidig. Dat wil zeggen dat met de keuze

. $\underline{y}$ uniform verdeeld op het interval $[a, b]$

dezelfde kansverdeling kan worden verkregen met de juiste keuze van η en τ in

. $\underline{y}$ uniform verdeeld met verwachting η en variantie τ.

Er geldt namelijk met (3) indien a en b gekozen worden

$$\eta = \tfrac{1}{2}(a + b)$$

en

$$\tau = \sqrt{\frac{(b - a)^2}{12}} = \frac{b - a}{\sqrt{12}}$$

waaruit volgt

$$a = \eta - \tfrac{1}{2}\tau\sqrt{12}$$

$$b = \eta + \tfrac{1}{2}\tau\sqrt{12}$$

Indien η en τ gegeven zijn volgt hieruit weer het begin- en eindpunt van het interval, symmetrisch van η, waarop de verdeling geldt. Ook blijkt dat de lengte van het interval $b - a = \tau \sqrt{12}$.

7. HET KIEZEN VAN EEN INITIELE WAARDE

Het proces van genereren van toevalsgetallen moet met een eerste getal begonnen worden. Dit behoeft veelal niet een realisatie uit de gekozen verdeling te zijn, maar in de regel is een getal nodig om de subroutine die in de computer als subprogramma beschikbaar is in werking te stellen.

Voor het C.D.-6600 systeem kan dit met behulp van twee FORTRAN--opdrachten. Deze zijn:

```
CALL RANSET (X)
```

Met behulp van een initiële waarde X wordt de subroutine voor het genereren van random getallen uit de library opgeroepen.

```
Y = RANF(DUM)
```

De random functie RANF voert waarden van de uniforme verdeling op het interval $[0, 1]$ terug via een plaats in het geheugen in dit geval aan te roepen met Y.

Een vaste initiële waarde bijvoorbeeld door op te nemen in het programma

```
X = 123.45
```

heeft tot gevolg dat steeds dezelfde random reeks wordt verkregen. Reeds is uiteengezet (par. 3) dat dit in de testfase van een programma geen bezwaar behoeft te zijn.

Een volgende mogelijkheid is de initiële waarde te laten afhangen van een tussenresultaat verkregen tijdens de uitvoering van het programma.

Wordt bijvoorbeeld berekend een waterstand W = 1.52 dan kan men kiezen

X = W

teneinde de gegenereerde toevalsfluctuatie bij de - deterministisch bepaald - waterstand op te tellen.

Ook deze procedure heeft principiële bezwaren. Nu zou immers elke waterstand van 1.52 eenzelfde toevalsfluctuatie toebedeeld krijgen, terwijl bij herhaling van het gehele rekenprogramma identieke resultaten ontstaan waar het juist de bedoeling is door herhaling van de berekening een inzicht in de frequentie van optreden van bepaalde situaties op de lange duur te verkrijgen.

Een methode om een toevallige beginwaarde te creëeren is het gebruik van datum en verwerkingstijdstip van het programma als initiële waarde. Het C.D.C.-systeem kent hiervoor de volgende subroutines

TIME (A)
DATE (D)

Opgenomen in het rechterlid van een statement geeft A het tijdstip volgens een 10 Hollerith code bijvoorbeeld A = b 11.34.25. waarin b een blanc voorstelt en de tijd is 11 h 34 min 25 sec.

Op eenzelfde wijze geeft het systeem terug D = b 11/24/72 in de volgorde maand, dag, jaar. Van de aldus verkregen informatie kan een getal worden gevormd (zie bijlagen) dat unique is bijvoorbeeld

$$24 \times 113425 = .27222\ E + 07$$

Zou men op de 24e dag van een andere maand dezelfde berekening uitvoeren dan is de kans verwaarloosbaar klein dat op exact hetzelfde tijdstip de tijd aan het systeem wordt opgevraagd. Gelijkheid van toevalsreeksen komt dus voor met een kans P = 0.

Opmerking

Nemen we aan dat de JOB op een willekeurig tijdstip tussen 8.00 h en 18.00 h aan het systeem wordt aangeboden, dan is het tijdstip van verwerking uniform verdeeld op het interval $[8.00\ ,\ 18.00]$. De kans dat de tijd aan het systeem wordt opgevraagd, precies op een gegeven tijdstip in seconden is

$$P(\underline{t} = 11.34.25) = \frac{1}{(18 - 8) \times 60 \times 60}$$

$$= \frac{1}{36\ 000} = 0.003\ \% \simeq 0$$

8. GENEREREN VAN ANDERE VERDELINGEN

Verschillende procedures kunnen worden gevolgd om uit de uniform verdeelde reeksen getallenrijen te construeren die aan een andere kansverdeling voldoen. Hiervoor is het noodzakelijk gebruik te maken van stellingen uit de kansrekening. In het volgende zullen enkele mogelijkheden worden aangegeven.

8a. Transformatie

We beschouwen een stochastische grootheid $\underline{x}$ met cumulatieve kansverdeling $F(x)$ zodat

$$P(\underline{x} < x) = F(x) \qquad (7)$$

Stel er wordt een waarde voor $\underline{x}$ gegenereerd. In dat geval is de bijbehorende onderschrijdingskans volgens (7) een stochastische grootheid. Bij elke trekking voor $\underline{x}$ behoort een realisatie van $\underline{F}$. Het blijkt nu dat $\underline{F}$ uniform verdeeld is op het interval $[0,\ 1]$. Dit is plausibel door uit te gaan van uniform verdeelde trekkingen van $\underline{F}$ (zie fig. 4). Beschouw de gebeurtenis

$$\underline{G}_1 = \underline{F} \in [F_1,\ F_2]$$

dan is

$$P(\underline{G}_1) = P(\underline{F} \in [F_1, F_2]) = F_2 - F_1$$

volgens de eigenschap van de uniforme verdeling (par. 6).

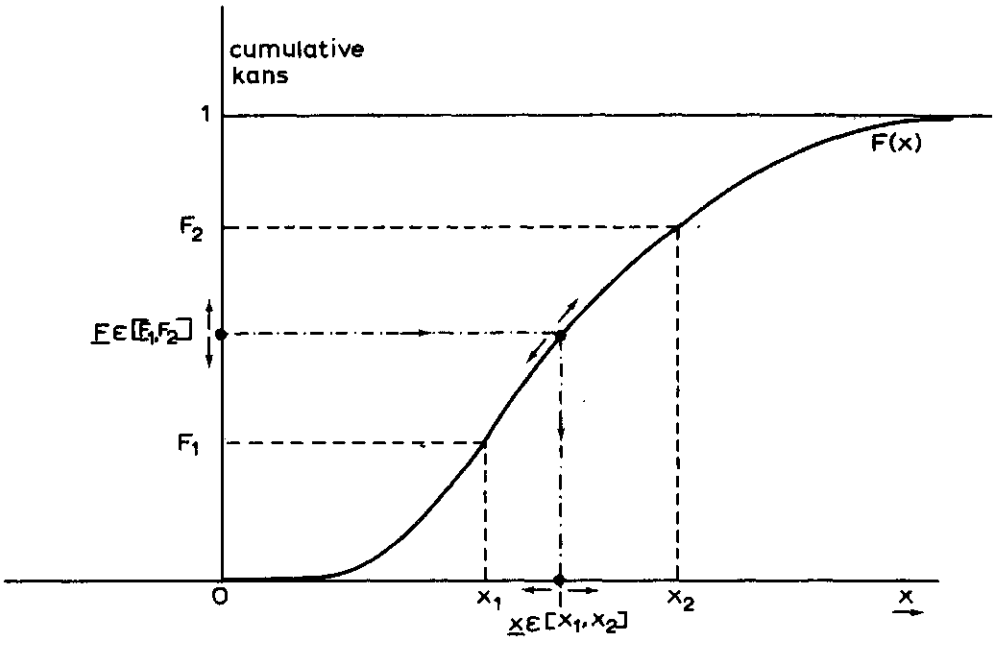


Fig. 4. De cumulatieve kansverdeling van een stochastische grootheid $\underline{x}$ opgevat als transformatie van de uniforme verdeling van $\underline{F}$

Echter uit fig. 4 leiden we ook af dat als

$$\underline{G}_2 \equiv \underline{x} \in [x_1, x_2]$$

dan

$$P(\underline{G}_2) = P(\underline{x} \in [x_1, x_2]) = F(x_2) - F(x_1)$$

volgens de eigenschap van cumulatieve verdelingen. Volgens de definities in fig. 4 geldt

$$F(x_2) \equiv F_2 \quad \text{en} \quad F(x_1) \equiv F_1$$

zodat

$$P(\underline{G}_1) = P(\underline{G}_2)$$

Wordt nu dus $\underline{F}$ uniform getrokken uit het interval $[0, 1]$ dan volgt de bijbehorende drempelwaarde van x uit de inverse functie $F^{-1}(x)$.

Voorbeeld (BUSLENKO und SCHREIDER, 1964 p. 37)

Voor de exponentiële verdeling geldt

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

en dus

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

de inverse functie luidt

$$x = -\frac{1}{\lambda} \ln(1 - F) \quad , \quad 0 \leq F \leq 1 \qquad (8)$$

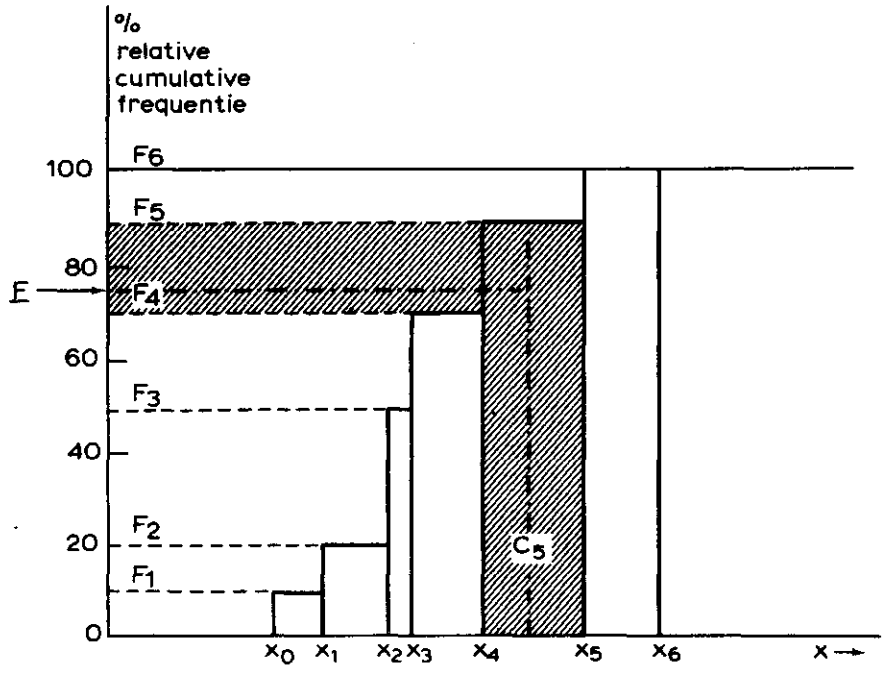
Worden waarden van $\underline{F}$ gegenereerd uit een uniforme verdeling $[0, 1]$ dan geeft de transformatie (8) realisaties van $\underline{x}$ die exponentieel verdeeld zijn op $[0, +\infty)$.

8b. Histogrammen

Is van $\underline{x}$ de kansverdeling niet als functie gegeven maar als empirische relatie uit waarnemingsuitkomsten afgeleid, dan kan eenzelfde procedure gevolgd worden als in par. 7a omschreven. Hiervoor is het nodig een benaderingsfunctie te vinden die de gewenste transformatie

tot stand brengt.

In plaats van met een continue functie kan ook van de cumulatieve stapfunctie, afgeleid uit het betreffende histogram, worden uitgegaan. Zie fig. 5.



continu $P(x_4 \leq \underline{x} < x_5) = 90 - 70 = 20\ \%$

discreet $P(\underline{x} = C_5) = P(F_4 \leq \underline{F} < F_5) = 20\ \%$

voor x geldt $P(x_0 \leq \underline{x} < x_6) = 100\ \%$

Fig. 5. Cumulatief histogram als transformatie van de uniforme verdeling van $\underline{F}$

Op overeenkomstige wijze als in par. 8a kunnen de volgende gebeurtenissen worden gedefinieerd, bijvoorbeeld

$$\underline{G}_1 \equiv \underline{F} \in [F_4, F_5]$$

en

$$\underline{G}_2 \equiv \underline{x} \in [x_4, x_5]$$

Aangezien in een histogram in een interval geen specificatie meer wordt gegeven moet worden overgaan op een vertegenwoordigende waarde bijvoorbeeld het klassemidden, zodat er komt

$$\underline{G}_2 \equiv (\underline{x} = C_5)$$

Voorts geldt (zie 8a)

$$P(\underline{G}_1) = F_5 - F_4$$

Wordt nu $\underline{F}$ volgens de uniforme verdeling getrokken en blijkt dat $F \in [F_4, F_5]$ dan wordt gedefinieerd $x = C_5$. Verder blijkt weer dat

$$P(\underline{G}_1) = P(\underline{G}_2)$$

zodat

$$P(\underline{x} = C_5) = P(\underline{x} \in [x_4, x_5]) = F_5 - F_4$$

waarmede het gewenste resultaat is verkregen.

De procedure is dan: trek $\underline{F}$ uniform uit het interval $[0, 1]$ waarna geldt (fig. 5)

$$\text{als} \quad \underline{F} \in [F_i, F_{i+1}] \quad \text{dan} \quad \underline{x} = c_{i+1} \ , \ i = 0, 1, 2, \ldots , 5$$

In het geheugen van de computer moeten dus de klassegrenzen F_i en de klassemiddens c_i worden opgenomen.

8c. Toepassing kansrekening

Door gebruik te maken van stellingen uit de kansrekening kunnen nieuwe verdelingen worden verkregen. Enkele opmerkingen volgen hieronder voor een continue en voor een discrete verdeling.

Voor discrete verdelingen die berusten op stochastische aantallen zal veelal gebruik moeten worden gemaakt van tests of een realisatie van een continue stochastische grootheid in een kritiek interval is terechtgekomen. In het programma dient dan een teller bijgehouden te worden die aangeeft hoe groot het aantal trekkingen is dat

aan de test voldoet.

. Normale verdeling

De centrale limietstelling van de kansrekening luidt, kort samengevat, dat de som $\underline{S}$ van n stochastische grootheden $\underline{x}$ beter normaal verdeeld zal zijn naarmate n groter wordt. Dit geldt onder vrij ruime voorwaarden ten aanzien van de verdeling van $\underline{x}$. Een bekend voorbeeld is de verdeling van de neerslag die voor 1-daagse sommen scheef is, doch voor jaarsommen de normale verdeling goed benadert M.G. VAN STEENBERGEN,(1972).

Op deze centrale limietstelling berust de eigenschap van gemiddelden om beter normaal verdeeld te zijn dan de oorspronkelijke stochastische variabele. Namelijk

$$\bar{\underline{x}} = \frac{\underline{x}_1 + \underline{x}_2 + \ldots + \underline{x}_n}{n}$$

betekent

$$\underline{S} = n\,\bar{\underline{x}} = \underline{x}_1 + \underline{x}_2 + \ldots + \underline{x}_n$$

aangezien deling door een constante geen invloed heeft op het type verdeling.

De stochastische grootheid $\bar{\underline{x}}$ kan dus opgevat worden als een som van stochastische grootheden $\frac{\underline{x}_i}{n}$ waarop de limietstelling van toepassing is.

Door nu de limietstelling op sommen van termen uit een uniforme verdeling toe te passen kan de normale verdeling benaderd worden. We merken nog op

$\underline{x}$ uniform verdeeld	$\bar{\underline{x}}$ nadert tot normale verdeling
$E(\underline{x}) = \mu$	$E(\bar{\underline{x}}) = \mu$
$E(\underline{x} - \mu)^2 = \sigma^2$	$E(\bar{\underline{x}} - \mu)^2 = \frac{\sigma^2}{n}$

Op het interval $[0, 1]$ wordt dit volgens (5)

$E(\underline{x}) = \frac{1}{2}$	$E(\bar{\underline{x}}) = \frac{1}{2}$
$E(\underline{x} - \frac{1}{2})^2 = \frac{1}{12}$	$E(\bar{\underline{x}} - \frac{1}{2})^2 = \frac{1}{12n}$

Stel men wil een reeks genereren van een stochastische variabel $\underline{y}$ die normaal verdeeld is met gegeven verwachting η en variantie τ^2. Men transformeert dan als volgt

$$\underline{y} = \tau \sqrt{12n}(\bar{\underline{x}} - \tfrac{1}{2}) + \eta$$

Namelijk

$$E(\underline{y}) = \tau \sqrt{12n}\, E(\bar{\underline{x}} - \tfrac{1}{2}) + \eta = \eta$$

en

$$E(\underline{y} - \eta)^2 = \tau^2\, 12n\, E(\bar{\underline{x}} - \tfrac{1}{2})^2 = \tau^2$$

wat de gevraagde eigenschappen zijn.

Met n = 12 wordt al een praktisch bruikbare benadering van de normale verdeling verkregen. In de bijlagen volgen enkele voorbeelden.

. Binomiale verdeling

De binomiale verdeling ontstaat uit de volgende situatie. Indien een gebeurtenis $\underline{A}$ met kans p in het eerstvolgende experiment (in de eerstvolgende waarneming) optreedt, dan wordt de kans P dat in de volgende n experimenten de gebeurtenis $\underline{A}$ in totaal $\underline{k}$ maal optreedt gegeven door

$$P(\underline{A}) = P(\underline{k} = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Voor toepassing van deze verdeling bij cultuurtechnische vraagstukken wordt verwezen naar STOL (1972).

Een dergelijke verdeling kan nu als volgt worden gegenereerd. We definiëren de gebeurtenis

$$\underline{A} \equiv (\underline{x} > x) \ , \ 0 \leq \underline{x} \leq 1, \ \underline{x} \text{ uniform verdeeld}$$

waarbij de kritieke waarde x zo gekozen kan worden dat aan een gewenste elementaire kans p_o wordt voldaan. Dit kan met de keuze van $x_o = 1 - p_o$ op het interval $[0, 1]$ zodat

$$\underline{A} \equiv (\underline{x} > x_o) \ , \ 0 \leq \underline{x} \leq 1, \ \underline{x} \text{ uniform verdeeld}$$

en

$$P(\underline{A}) = P(\underline{x} > x_o) = p_o$$

Door nu met behulp van de subroutine getallen uit een uniforme verdeling te trekken en te testen of deze getallen groter zijn dan x_o is een systeem verkregen waarbij de situatie waarvoor de binomiale verdeling geldt is nagebootst. Elke trekking is een gebeurtenis met kans p_o op succes. De verkregen reeks geeft een toevallige reeks met successen en mislukkingen die binomiaal verdeeld is met kans p_o op succes (gebeurtenis A treedt op).

Dit type reeksen kan worden gebruikt om te simuleren of met de gebeurtenis A op het onderzochte tijdstip t rekening moet worden gehouden, dan wel dat op tijdstip t de gebeurtnis A niet is opgetreden.

In de bijlagen volgt hiervan een voorbeeld.

8d. Gemengde methode

Een methode waarmee normaal verdeelde grootheden door toepassing van stellingen uit de kansrekening - uitlopend in een eenvoudige transformatie - kan worden verkregen is ontworpen door BOX en MULLER (1958).

Hierbij wordt van de volgende stellingen uit de kansverdeling gebruik gemaakt.

- Indien de variabelen $\underline{x}_1$ en $\underline{x}_2$ twee dimensionaal normaal verdeeld zijn, zijn de marginale verdeling van $\underline{x}_1$ en $\underline{x}_2$ eveneens normaal verdeeld.
- Indien de variabelen $\underline{x}_1$ en $\underline{x}_2$ twee dimensionaal normaal verdeeld zijn met cirkelsymmetrische kansdichtheid, dan zijn $\underline{x}_1$ en $\underline{x}_2$ stochastisch onafhankelijk.

Overgegaan wordt op poolcoördinaten (r, θ) zodat

$$x_1 = r \cos \theta \quad \text{en} \quad x_2 = r \sin \theta$$

hieruit volgt dan weer

$$r^2 = x_1^2 + x_2^2$$

$$\theta = \arctan \frac{x_2}{x_1}$$

De twee dimensionale dichtheid op cirkels met r constant is eveneens constant zodat $\underline{\theta}$ uniform verdeeld is op het interval $[0, 2\pi)$.

Verder volgt dat r^2 de som is van de kwadraten van twee normaal verdeelde grootheden zodat $\underline{r}^2$ verdeeld is volgens χ^2 met twee vrijheidsgraden. BOX en MULLER (1958) maken ervan gebruik dat als $\underline{u}$ uniform verdeeld is op $[0, 1]$ dat $-2 \ln \underline{u}$ een χ^2 verdeling heeft met twee vrijheidsgraden. Dit geeft aanleiding tot de volgende transformatie

$$\underline{x}_1 = \sqrt{-2 \ln \underline{u}_1} \cos 2\pi \underline{u}_2$$

$$\underline{x}_2 = \sqrt{-2 \ln \underline{u}_1} \sin 2\pi \underline{u}_2 \qquad (8)$$

Hierin zijn $\underline{u}_1$ en $\underline{u}_2$ onafhankelijk uniform verdeel en bij gevolg $\underline{x}_1$ en $\underline{x}_2$ onafhankelijk en normaal verdeel elk met verwachting 0 en variantie 1.

De formules (8) kunnen dus worden gebruikt om met twee trekkingen uit de uniforme verdeling door transformatie twee trekkingen uit een normale verdeling te bepalen. Deze mogen dan opgevat worden als

- één trekking uit een twee dimensionale cirkelsymmetrische normale verdeling met verwachtingsvector(0, 0) en variantievector (1, 1)
- twee onafhankelijke trekkingen uit een normale verdeling met verwachting = 0 en variantie = 1

Het bewijs van (8) verloopt als volgt.

We veronderstellen dat $\underline{u}$ uniform verdeeld is op $[0, 1]$. De kansverdeling van $\underline{u}$ is dan

$$P(\underline{u}) = \int_0^1 du \tag{9}$$

We vragen nu naar de kansverdeling van $-2 \ln \underline{u}$ en stellen $-2 \ln \underline{u} = \underline{x}$. Dit leidt tot

$$\underline{u} = e^{-\frac{1}{2}\underline{x}}, \quad (0 \leq \underline{u} \leq 1 \ , \ \infty > \underline{x} \geq 0)$$

zodat in (9) ingevuld de gelijkheid ontstaat

$$P(e^{-\frac{1}{2}\underline{x}}) = \int_0^1 d\, e^{-\frac{1}{2}x}$$

Geschreven als kansverdeling van $\underline{x}$ wordt dit

$$P(\underline{x}) = \int_\infty^0 e^{-\frac{1}{2}x}\, d(-\tfrac{1}{2}x)$$

$$= \frac{1}{2}\int_0^\infty e^{-\frac{1}{2}x}\, dx \tag{10}$$

De chi-kwadraatverdeling luidt

$$P(\chi_n^2) = \int_0^\infty \frac{1}{\Gamma(\frac{n}{2})} (\frac{\chi^2}{2})^{\frac{1}{2}n-1} e^{-\frac{1}{2}\chi^2} \frac{1}{2}\, d(\chi^2)$$

wat met $n = 2$ vrijheidsgraden wordt

$$P(\chi_2^2) = \int_0^\infty e^{-\frac{1}{2}\chi^2} \frac{1}{2}\, d(\chi^2)$$

Stellen we nu $\underline{\chi}^2 = \underline{x}$ dan komt er, met dezelfde integratie grenzen,

$$P(\underline{x}) = \frac{1}{2} \int_0^{\infty} e^{-\frac{1}{2}x} \, dx$$

zodat $\underline{x}$ verdeeld is volgens chi-kwadraat met 2 vrijheidsgraden. Door vanaf (10) terug te werken vinden we dat als $\underline{u}$ uniform verdeeld is, dat dan $-2 \ln \underline{u}$ verdeeld is als $\underline{\chi}^2_2$, wat dus de verdeling van $\underline{r}^2 = \underline{x}^2_1 + \underline{x}^2_2$ oplevert.

Tenslotte, als $\underline{\theta}$ uniform verdeeld is op $[0, 2\pi)$ dan is $2\pi\theta$ uniform verdeeld op $[0, 1)$.

Hiermede zijn de gebruikte transformaties verantwoord.

De methode is volgens BOX en MULLER ook in de staarten van de verdeling nauwkeurig. Bovendien is de efficiëncie groot ten opzichte van andere methoden (zie ook TOCHER, 1969).

9. REEKSEN MET GEVRAAGDE CORRELATIE

a. Autocorrelatie

Waarnemingsreeksen die een chronologische volgorde hebben blijken vaak autocorrelatie te vertonen. Bekende voorbeelden zijn afvoerreeksen waarvan de gemeten waarden een aantal achtereenvolgende dagen hoog blijft wanneer een afvoergolf het meetpunt passeert. Voorbeelden worden gegeven in het Tweede Interimrapport van de werkgroep Afvloeiingsfactoren (1970).

Wil men nu als input-gegeven in een rekenmodel een afvoerreeks simuleren, dan zal met de eigenschap van autocorrelatie rekening moeten worden gehouden. Wanneer we uitgaan van een zogenaamd Markov I model dan luidt de samenhang tussen twee opvolgende metingen

$$\underline{x}_t = \beta x_{t-1} + \underline{\varepsilon}_t$$

waarin x_{t-1} een realisatie is van

$$\underline{x}_{t-1} = \beta x_{t-2} + \underline{\varepsilon}_{t-1}$$

enzovoorts.

Een dergelijke rij heet stationair als voor elke $\underline{x}_t$ dezelfde verdeling geldt. In ieder geval is dan

$$E(\underline{x}_t) = \mu \qquad \text{en dus constant voor alle } t$$

en

$$E(\underline{x}_t - \mu)^2 = \sigma^2 \text{ en dus constant voor alle } t$$

Verder wordt aangenomen dat $\underline{\varepsilon}$ verwachting 0 heeft en dat $\underline{\varepsilon}_i$ en $\underline{\varepsilon}_j$ ongecorreleerd zijn als $i \neq j$. Een kleinste kwadratenschatter voor β volgt uit

$$\hat{\beta} = b = \frac{\sum_{t=2}^{n} x_{t-1}\, x_t}{\sum_{t=2}^{n} x_{t-1}\, x_{t-1}}$$

wat tevens de formule is voor de 1e coëfficiënt in het correllogram (MALINVAUD, 1966, p -454-).

Dit geeft aanleiding tot de formule

$$\underline{x}_t = \rho\, \underline{x}_{t-1} + \underline{\varepsilon}_t \qquad (11)$$

als model voor een stochastische reeks met autocorrelatie gelijk aan ρ. Voor ρ, uitgedrukt in de parameters van het kansmodel geldt

$$\rho = \frac{E(\underline{x}_{t-1}\,,\, \underline{x}_t)}{E(\underline{x}_t^2)}$$

Voor de verwachting van $\underline{x}_t$ geldt uit (11)

$$E(\underline{x}_t) = \rho\, E(\underline{x}_{t-1}) + E(\underline{\varepsilon}_t)$$

waarin de laatste term volgens de gegeven aannamen gelijk is aan 0. De reeks (11) is dus stationair met $\rho \neq 0$ indien

$$E(\underline{x}_t) = E(\underline{x}_{t-1}) = 0 \text{ , voor alle } t$$

(KENDALL and STUART, 1966 - p. 405).

Verder kan worden berekend dat de variantie is

$$E(\underline{x}_t^2) = E(\rho^2\, \underline{x}_{t-1}^2 + 2\rho\, \underline{x}_{t-1}\, \underline{\varepsilon}_t + \underline{\varepsilon}_t^2)$$

$$= \rho^2\, E(\underline{x}_{t-1}^2) + 2\rho\, E(\underline{x}_{t-1}\, \underline{\varepsilon}_t) + E(\underline{\varepsilon}_t^2) \qquad (12)$$

Aangezien $\underline{x}_{t-1}$ en $\underline{\varepsilon}_t$ onafhankelijk zijn kan geschreven worden $E(\underline{x}_{t-1}\, \underline{\varepsilon}_t) = E(\underline{x}_{t-1}).E(\underline{\varepsilon}_t) = 0$. Noemen we $E(\underline{\varepsilon}_t^2) = \sigma_{\underline{\varepsilon}}^2$ dan wordt (12), rekening houdend met het stationair karakter van (11) waardoor $E(\underline{x}_t^2) = E(\underline{x}_{t-1}^2)$,

$$E(\underline{x}_t^2) = \frac{\sigma_{\underline{\varepsilon}}^2}{1 - \rho^2}$$

Stel nu dat gevraagd wordt een toevalsreeks te genereren waarvan de autocorrelatie is ρ en de variantie van $\underline{x}_t$ gelijk is aan $\sigma_{\underline{x}}^2$ op basis van een toevalsgenerator met verwachting 0 en variantie $\sigma_{\underline{\varepsilon}}^2$, dan kan gebruik worden gemaakt van het volgende model

$$\underline{x}_t = \rho\, \underline{x}_{t-1} + \frac{\sigma_{\underline{x}}}{\sigma_{\underline{\varepsilon}}} \sqrt{1 - \rho^2}\, \underline{\varepsilon}_t$$

Wordt een toevalsgenerator gebruikt met variantie gelijk aan 1, dan wordt dit

$$\underline{x}_1 = \sqrt{1 - \rho^2}\, \sigma_{\underline{x}}\, \underline{\varepsilon}_1$$

$$\underline{x}_2 = \rho\, \underline{x}_1 + \sqrt{1 - \rho^2}\, \sigma_{\underline{x}}\, \underline{\varepsilon}_1 \qquad (13)$$

en algemeen

$$\underline{x}_t = \rho\, \underline{x}_{t-1} + \sqrt{1 - \rho^2}\, \sigma_{\underline{x}}\, \underline{\varepsilon}_t$$

welke formules kunnen worden gebruikt voor het genereren van reeksen met de gewenste eigenschappen.

b. Gewone correlatie

Voor het geval dat twee reeksen toevalsgetallen $\underline{x}$ en $\underline{y}$ gevraagd worden die aan de volgende eigenschappen voldoen

Verwachting van $\underline{x} = \mu$, van $\underline{y} = \eta$
Variantie van $\underline{x} = \sigma^2$, van $\underline{y} = \tau^2$
Correlatie tussen $\underline{x}$ en $\underline{y} = \rho$

kan een speciaal geval van een algemene methode van YAGIL (1963) worden toegepast. De methode is gebaseerd op de volgende formules

$$\underline{x}_t = \mu + \sigma\, \underline{\varepsilon}_t$$

$$\underline{y}_t = \eta + \frac{\tau}{\sigma}\, \rho(\underline{x}_t - \mu) + \sqrt{1 - \rho^2}\, \tau\, \underline{\varepsilon}'_t \qquad (14)$$

In dit systeem worden $\underline{\varepsilon}_t$ en $\underline{\varepsilon}'_t$ als twee normaal verdeelde grootheden beschouwd met verwachting = 0 en variantie = 1, die ongecorreleerd zijn.

Voor dit systeem geldt nu

$$E(\underline{x}_t) = \mu + \sigma\, E(\underline{\varepsilon}_t) = \mu$$

$$E(\underline{y}_t) = \eta + \frac{\tau}{\sigma}\rho(E(\underline{x}_t) - \mu) + \sqrt{1-\rho^2}\,\tau\, E(\underline{\varepsilon}'_t)$$

$$= \eta$$

$$E(\underline{y}_t - \mu)^2 = \sigma^2\, E(\underline{\varepsilon}_t^2) = \sigma^2$$

$$E(\underline{y}_t - \eta)^2 = \tau^2\{\frac{\rho^2}{\sigma^2} E(\underline{x}_t - \mu)^2 + (1-\rho^2)\, E(\underline{\varepsilon}'_t)^2\}$$

in de laatste vorm is het dubbele produkt door de factor $E(\underline{x}_t - \mu)\,\underline{\varepsilon}'_t$ weggevallen aangezien beide onafhankelijk verdeeld zijn. Er komt dus

$$= \tau^2\left\{\frac{\rho^2}{\sigma^2}\sigma^2 + 1 - \rho^2\right\} = \tau^2$$

Tenslotte berekenen we de teller van (2a) als volgt uit (14)

$$\mathrm{Cov}(\underline{x}_t\ ,\ \underline{y}_t) = E(\underline{x}_t - \mu)(\underline{y}_t - \mu)$$

$$E(\underline{x}_t - \mu)\{\frac{\tau}{\sigma}\rho(\underline{x}_t - \mu) + \sqrt{1-\rho^2}\,\tau\,\underline{\varepsilon}'_t\}$$

We merken op dat $\underline{x}_t$ en $\underline{\varepsilon}'_t$ ongecorreleerd zijn zodat er overblijft

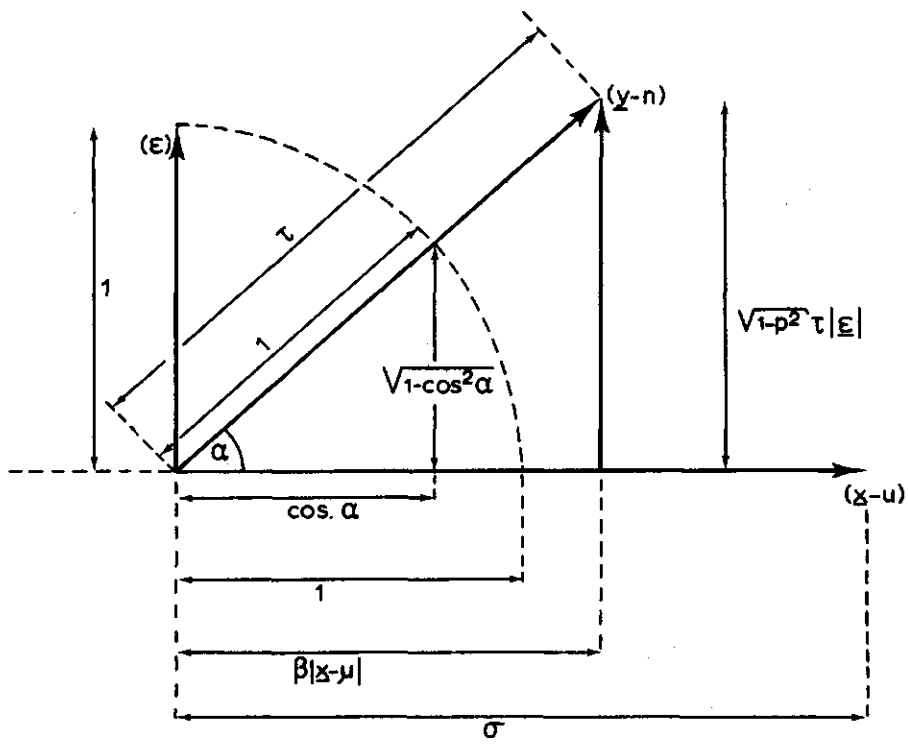
$$\mathrm{Cov}(\underline{x}_t\ ,\ \underline{y}_t) = \frac{\tau}{\sigma}\rho\, E(\underline{x}_t - \mu)^2 = \tau\,\sigma\,\rho$$

en tenslotte voor de correlatiecoëfficiënt, gebruikmakend van deze uitkomst

$$\rho(\underline{x}_t\ ,\ \underline{y}_t) = \frac{\mathrm{Cov}(\underline{x}_t\ ,\ \underline{y}_t)}{\sigma\,\tau} = \rho$$

Hiermede is bewezen dat het systeem (14) aan alle gewenste eigenschappen voldoet.

De gegeven afleidingen zijn eenvoudig te verifiëren met een vectorvoorstelling van de besproken grootheden (fig. 6). Er volgt nog



$\rho = \cos \alpha$

optellen van vectoren geeft

$$(\underline{y} - \eta) = \beta(\underline{x} - \mu) + \sqrt{1 - \rho^2}\,\tau(\underline{\varepsilon})$$

waarin

$$\beta = \frac{|\underline{y} - \eta| \cos \alpha}{|\underline{x} - \mu|} = \frac{\tau}{\sigma}\rho$$

waarmede (14) verkregen is

Fig. 6. Het lineaire regressiemodel voor 2 variabelen in vectorvoorstelling.
Vectoren zijn aangeduid met ( )
Lengten van vectoren met | | en met een scalaire grootheid bij |⟷|

uit dat de regressiecoëfficiënt β gelijk is aan

$$\beta = \frac{\tau}{\sigma} \rho \qquad (15)$$

zodat in (14) geschreven kan worden

$$\underline{y}_t = \eta + \beta(\underline{x}_t - \mu) + \sqrt{1 - \rho^2}\, \tau\, \underline{\varepsilon}'_t$$

Echter met (15) worden de verschillende grootheden met elkaar in verband gebracht zodat ze niet alle vrij gekozen kunnen worden. Met een keuze van σ, τ en ρ ligt β dus vast.

Opmerking

Voor stationaire reeksen met autocorrelatie ρ blijkt volgens par. 9a dat $\mu = \eta = 0$ en $\sigma = \tau = \sigma_{\underline{x}}$. Hiermede is het model (14) herleid tot (13).

Voorts wordt opgemerkt dat bij het genereren van twee gecorrelleerde reeksen volgens (14) toepassing van de methode Box-Muller, weergegeven in (8), voor het genereren van $\underline{\varepsilon}_t$ en $\underline{\varepsilon}'_t$ tot een efficiënte werkwijze leidt.

c. Multipele correlatie

Het algemene systeem waarvan (14) een onderdeel is werd door YAGIL (1963) voor het volgende geval ontworpen. Voor een meer (meer van Tiberias) wordt gevraagd de maandelijkse instroming te genereren en hierbij te voldoen aan de volgende vastgestelde statistische eigenschappen:

. Totale jaarlijkse instromingen zijn ongecorreleerd
. De variantie van de jaarlijkse instroming moet worden benaderd
. De maandelijkse gemiddelden en varianties moeten worden benaderd
. De correlaties tussen alle paren maanden moet worden benaderd
. De maandelijkse instromingen kunnen normaal verdeeld worden verondersteld

Het systeem waarmee dit bereikt wordt luidt nu voor de eerste 3 variabelen (maanden) als volgt (YAGIL, 1963)

$$\underline{x} = \mu + \sigma\, \underline{\varepsilon}_1$$

$$\underline{y} = \eta + \alpha(\underline{x} - \mu) + \sqrt{1 - R_2^2}\, \tau\, \underline{\varepsilon}_2 \qquad (16)$$

$$\underline{Z} = \xi + \beta_1(\underline{x} - \mu) + \beta_2(\underline{y} - \eta) + \sqrt{1 - R_3^2}\, \Psi\, \underline{\varepsilon}_3$$

waarin

$E(\underline{z}) = \xi$, $E(\underline{z} - \xi)^2 = \Psi$

α, β_1 en β_2 zijn regressiecoëfficiënten

R_2 is de multipele correlatie tussen $\underline{y}$ en $\underline{x}$

R_3 is de multipele correlatie tussen $\underline{z}$ en $(\underline{y}, \underline{x})$

In de regressiecoëfficiënten zijn de correlaties tussen tweetallen begrepen maar volgens een meer gecompliceerde structuur dan in (15) werd gegeven. Dit wordt toegelicht met fig. 7.

De berekening van β_1 en β_2 in dit regressiemodel geschiedt met de normaalvergelijkingen van inprodukten

$$\beta_1(\underline{x} - \mu, \underline{x} - \mu) + \beta_2(\underline{x} - \mu, \underline{y} - \eta) = (\underline{x} - \mu, \underline{z} - \xi)$$

$$\beta_2(\underline{y} - \eta, \underline{x} - \mu) + \beta_2(\underline{y} - \eta, \underline{y} - \eta) = (\underline{y} - \eta, \underline{z} - \xi)$$

Door rijen en kolommen door de bijbehorende standaardafwijking te delen ontstaat er

$$\sigma\, \beta_1 + \tau\, \beta_2\, \rho(\underline{x}, \underline{y}) = \Psi\, \rho(\underline{x}, \underline{z})$$

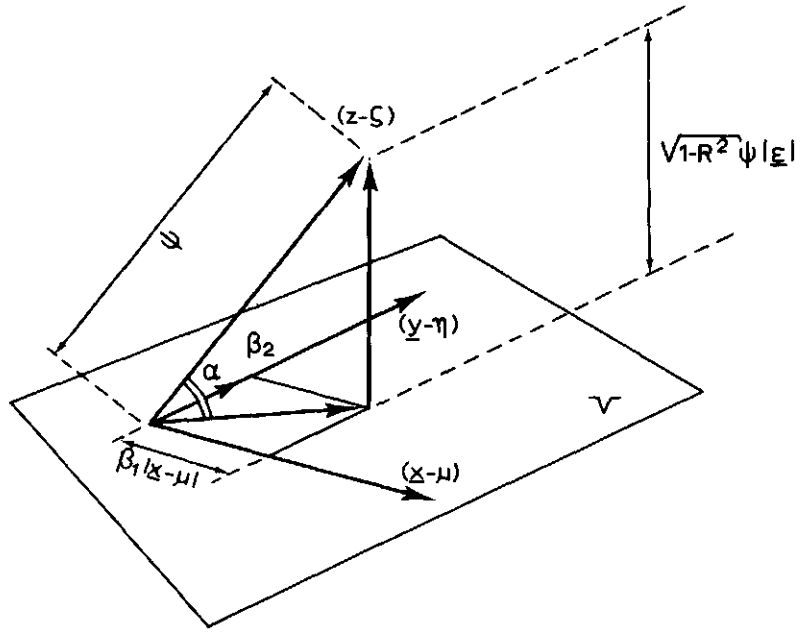
$$\sigma\, \beta_1\, \rho(\underline{x}, \underline{y}) + \tau\, \beta_2 = \Psi\, \rho(\underline{y}, \underline{z})$$

waaruit als oplossing volgt (KENNY and KEEPING, 1959 - p 342 -)

$$\beta_1 = \frac{\psi}{\sigma} \frac{\rho(\underline{x}\ ,\ \underline{z}) - \rho(\underline{x}\ ,\ \underline{y})\ \rho(\underline{z}\ ,\ \underline{y})}{1 - \rho^2(\underline{x}\ ,\ \underline{y})}$$

$$\beta_2 = \frac{\psi}{\tau} \frac{\rho(\underline{y}\ ,\ \underline{z}) - \rho(\underline{y}\ ,\ \underline{x})\ \rho(\underline{z}\ ,\ \underline{x})}{1 - \rho^2(\underline{x}\ ,\ \underline{y})}$$

Alleen met $\rho(\underline{x}\ ,\ \underline{y}) = 0$ wordt een oplossing van de gedaante (15) verkregen.



α is de standhoek van de drager van de vector $(\underline{z} - \xi)$ met multipele correlatie $R = \cos \alpha$

$(\underline{y} - \eta)$ en $(\underline{x} - \mu)$ niet loodrecht dus $\rho(\underline{x}\ ,\ \underline{y}) \neq 0$

De ontbinding in vlak V is scheefhoekig.

Optellen van vectoren geeft

$$(\underline{z} - \xi) = \beta_1(\underline{x} - \mu) + \beta_2(\underline{y} - \eta) + \sqrt{1 - R^2}\ \psi\ \underline{\varepsilon}$$

Fig. 7. Het lineaire regressiemodel voor 3 variabelen in vectorvoorstelling. Zie onderschrift fig. 6

Uit fig. 7 volgt nog dat met de tot nu besproken grootheden de waarden van de multipele correlatiecoëfficiënten R_2 en R_3 vastliggen analoog aan (15). Er geldt namelijk (zie fig. 7)

$$R_3 = \cos \alpha$$

zodat R de verhouding is tussen lengten van de volgende vectoren

$$R_3 = \frac{|\beta_1(\underline{x} - \mu) + \beta_2(\underline{y} - \eta)|}{|\underline{z} - \xi|}$$

$$= \frac{|\beta_1(\underline{x} - \mu) + \beta_2(\underline{y} - \eta)|^2}{|\underline{z} - \xi|^2} \qquad (17)$$

wat resulteert in

$$R_3 = \sqrt{\frac{\psi^2 - s_e^2}{\psi^2}}$$

volgens de stelling van Pythagoras, en waarin s_e^2 de zogenaamde niet verklaarde rest variantie uit het regressiemodel is. Tenslotte

$$R_3 = \sqrt{1 - \frac{s_e^2}{\psi^2}}$$

Uit vergelijking van fig. 6 met fig. 7 valt op te maken dat $R_2 = \rho(\underline{x}, \underline{y})$ wat ook algebraïsch eenvoudig te bewijzen valt.

De gang van zaken met 3 variabelen is nu als volgt

- geef waarden voor verwachtingen μ, η, ξ
- geef waarden voor de varianties σ^2, τ^2, ψ^2
- geef de correlaties $\rho(\underline{x}, \underline{y})$, $\rho(\underline{x}, \underline{z})$, $\rho(\underline{y}, \underline{z})$
- bereken uit $\rho(\underline{x}, \underline{y})$ de waarde voor α
- bereken uit $\rho(\underline{x}, \underline{y})$, $\rho(\underline{x}, \underline{z})$, $\rho(\underline{y}, \underline{z})$ de waarden voor β_1 en β_2
- bereken uit α, β_1 en β_2 waarden voor R_2 en R_3

Hiermede zijn dan alle waarden bekend om het proces (16) tot uitvoer te brengen.

Opmerking

Aan de twee eerste eisen voor jaarsommen wordt met het door YAGIL (1963) gegeven model 'automatisch' voldaan.

Voor het hier beschouwde geval (3 variabelen) geldt namelijk dat, aangezien $\underline{x}$, $\underline{y}$ en $\underline{z}$ normaal verdeeld zijn, ook hun som normaal verdeeld is. De v e r w a c h t i n g luidt dan

$$E(\underline{x} + \underline{y} + \underline{z}) = \mu + \eta + \xi$$

Voor de v a r i a n t i e geldt

$$E(\underline{x} + \underline{y} + \underline{z} - \mu - \eta - \xi)^2 =$$

$$E(\underline{x} - \mu)^2 + E(\underline{y} - \eta)^2 + E(\underline{z} - \xi)^2$$

$$+ 2E(\underline{x} - \mu)\ (\underline{y} - \eta) + 2E(\underline{x} - \mu)\ (\underline{z} - \xi) + 2E(\underline{y} - \eta)\ (\underline{z} - \xi)$$

$$= \sigma^2 + \tau^2 + \Psi^2 + 2\,\rho_{xy}\,\sigma\,\tau + 2\,\rho_{xz}\,\sigma\,\Psi + 2\,\rho_{yz}\,\tau\,\Psi$$

Deze betrekking geldt zowel voor de historische gegevens als voor het model. Aangezien in deze laatste uitdrukking alleen grootheden voorkomen die aan de gestelde eisen voldoen, geldt dit ook voor hun combinatie.

10. CONTROLE OP DE RESULTATEN

Het wezenlijke kenmerk van stochastische reeksen is dat de uitkomsten niet voorspelbaar zijn. Dit houdt tevens in dat het zinloos is naar een methode te vragen waarmede 'controle' op resultaten kan worden verkregen. Een deterministische methode kan principieel niet gebruikt worden evenmin trouwens als een nieuwe stochastische simulatie die ook weer uitkomsten levert die niet voorspelbaar zijn.

Controle van resultaten moet gezocht worden in het statistisch toetsen op statistische eigenschappen van de verkregen uitkomsten. Van de gesimuleerde reeksen verwacht men dat zij steekproeven uit kansverdelingen met gegeven eigenschappen zijn. Deze eigenschappen kan men als nul-hypothese stellen en de reeks hierop toetsen.

Naast het toetsen op de gewenste eigenschappen kan men ook toetsen op ongewenste eigenschappen. Kan in het laatste geval een nul-hypothese niet worden verworpen dan is de gesimuleerde reeks in ieder geval verdacht met betrekking tot de ongewenste eigenschap.

Het zal duidelijk zijn dat theoretisch afgeleide transformaties juist die eigenschappen bezitten die men wenst. De aanname die steeds gedaan wordt is dat de randomgenerator (de computer-subroutine) uniform trekt op het interval $[0\ ,\ 1]$. Dit is in feite het(zwakke)onderdeel dat toetsing noodzakelijk maakt.

Toetsen of de (pseudo) randomreeksen aan de gewenste eigenschappen voldoen hebben betrekking op

. Toets op type verdeling
. Toets op gewenste waarden van de parameters van de verdeling

Toetsen of de gegenereerde reeksen niet-gewenste eigenschappen bezitten zijn onder andere

. Toets tegen trends
. Toets tegen oscillaties
. Toets tegen periodiciteiten

Vele van dit type toetsen zijn parametervrij en hangen dus niet af van de verdeling waarvan men uitgaat. Beschrijvingen ervan vindt men in de statistische literatuur en handboeken.

Bijlage 1

SIMULATIE NORMALE VERDELING

a. Fortran-programma (belangrijkste opdrachten)

Fortran opdracht		Enkele uitkomsten
NDRAW=DRAW=24. $ NUMBER=200	0	
E=10. $ S=2.	0	
KLOCK=TIME(A)	1	b11.34.25.
WHEN=DATE(D)	2	b11/24/72
DECODE(9,2,A)IN1,IN2,IN3	3	11 , 34 , 25
DECODE(6,3,D)IN4	4	24
2 FORMAT(3(1X,I2))	5	
3 FORMAT(4X,I2)	6	
X=IN4*(10000*IN1+100*IN2+IN3)	7	.27222E+07
CALL RANSET(X)	8	
YO=RANF(DUM)	9	.067808
Y1=RANF(DUM)	10	.099113
Y2=RANF(DUM)	11	.428104
	12	
DO 30 J = 1 , NUMBER	13	1 , 200
T=.0	14	
DO 31 IMAAL=1 , NDRAW	15	1 , 24
Y=RANF(DUM)	16	
31 T=T+Y	17	
T=T/DRAW	18	
T=(T-0.5)*SQRT(DRAW*12.)	19	
30 ZRAND(J)=S*T+E	20	15.1932

Toelichting programma

. In (0) worden systeemparameters gedefinieerd en worden gesteld E=10.(=η) en S=4.(=τ).
. In (1) en (2) worden aan het CDC-systeem respectievelijk de tijd en de datum opgevraagd waarop executie van het programma plaats vindt.
. In (3) en (4) worden respectievelijk tijd en datum gedecodeerd tot numeriek verwerkbare getallen zodat IN1=11, IN2=34 en IN3=25, tenslotte IN4=24.
. In (5) en (6) wordt gespecificeerd hoe decodering moet plaatsvinden.
. In (7) wordt de initiële waarde X berekend.
. In (8) wordt de subroutine: 'definieer random initiëring' aangeroepen.
. In (9), (10) en (11) worden drie randomfuncties berekend. Er worden drie trekkingen uit een uniforme verdeling op 0 , 1 verkregen.
. In (13) wordt aangegeven dat de opdrachten tot en met label 30 een aantal keren gelijk aan NUMBER herhaald moeten worden. Gedefinieerd is NUMBER=200 (opdracht (0)).
. In (14) wordt een geheugenplaats ,aangeroepen met T, gelijk aan 0 gemaakt.
. In (15) wordt aangegeven dat de opdrachten tot en met label 31 een aantal keren gelijk aan NDRAW herhaald moeten worden. Gedefinieerd is NDRAW=24 (opdracht (0)).
. In (16) wordt een getal uit de uniforme verdeling op $[0 , 1]$ getrokken. Dit getal is in het geheugen aan te roepen met Y.
. In (17) wordt de waarde van Y bijgeteld bij de waarde van T. Het resultaat wordt weer weggezet op T.
. Nadat (16) en (17) in totaal 24 x zijn herhaald wordt de som T gedeeld door het aantal trekkingen DRAW=NDRAW zodat nu T de betekenis krijgt van $\frac{\sum y}{24} = \bar{y}$.
. In (19) vindt omrekening van $\bar{y}$ plaats naar een normaal verdeelde grootheid met verwachting 0 en spreiding 1. Zie par. 7c.
. In (20) wordt ZRAND(1) gelijk aan een normaal verdeelde grootheid met verwachting E en spreiding S. Zie par. 7c. De samenhang

Tabel 4. Voorbeelden van empirische binomiale verdelingen op basis van het 200 maal verrichten van 20 trekkingen (p = 0.25 met 2 empirische reeksen, p = 0.05 met 3 empirische reeksen)

p = 0.25				p = 0.05				
k successen	Aantal malen van voorkomen van k successen per 20 trekkingen			k successen	Aantal malen van voorkomen van k successen per 20 trekkingen			
	empirisch		theoretisch		empirisch			theoretisch
0	0	2	0.6	0	65	63	67	71.7
1	5	1	4.2	1	75	91	76	75.5
2	16	14	13.4	2	44	40	40	37.7
3	26	25	26.8	3	12	2	15	11.9
4	30	42	37.9	4	2	4	2	2.7
5	43	39	40.5	5	2	0	0	0.4
6	41	37	33.7	6	0	0	0	0.1
7	24	23	22.5	7	0	0	0	0.0
8	8	9	12.2	8	0	0	0	0.0
9	6	5	5.4	9	0	0	0	0.0
10	1	3	2.0	10	0	0	0	0.0
11	0	0	0.6					
12	0	0	0.2					
13	0	0	0.0					
14	0	0	0.0					
15	0	0	0.0					
	200	200	200.0		200	200	200	200.0

LITERATUUR

BOX, G.E.P. and M.E. MULLER (1958). A note on the generation of random normal deviates. Ann. Math. Stat. 28 (pp 610, 611).

BUSLENKO, N.P. und J.A. SCHREIDER (1964). Die Monte-Carlo-Methode und ihre Verwirklichung met Elektronischen Digitalrechnern. Teubner Leipzig (ICW 11/307).

KENDALL, M.G., and A. STUART (1966). The advanced theory of statistics. Vol. 3. Design and analysis, and time-series. Griffin, London (ICW 11/114).

KENNY, J.F. and E.S. KEEPING (1959). Mathematics of Statistics. Vol. II. Nostrand New York (ICW 11/35).

MALINVOUD, E. (1966). Statistical methods of econometrics. North. Holland Publ. Comp. Amsterdam (ICW No. 11/285).

STEENBERGEN, M.G. VAN (1972). Een toepassing van het gebruik van reeksontwikkeling voor empirische frequentieverdelingen op neerslaggegevens. ICW Nota 656.

STOL, Ph.Th. (1972). Een beschouwing over de frequentie van weerkeren van hydrologische gebeurtenissen. Cult. Tijdschr. Jrg. 11 Nr 4 (168-186). ICW Verspreide Overdr. 125.

WERKGROEP AFVLOEIINGSFACTOREN (1970). Tweede Interim Rapport

YAGIL, S. (1963). Generation of input data for simulation. IBM systems Journal Sept.-Dec. (pp 288-296).

TOCHER, K.D. (1969). The art of simulation. English Un. Press. London (ICW 11/401).