STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

ZW 1950-010

# Over "upcrosses" en "downcrosses" in tijdreeksen

"Actualiteiten"

H. Theil

1950

Voordracht door H. Theil in de serie
Actualiteiten op 20 Mei 1950.

## Over "upcrosses" en "downcrosses" in tijdreeksen.

1. Onder tijdreeks verstaan we hier een getallenrij $x_t$, gedefinieerd op een discrete verzameling van waarden van de parameter t. Deze parameter wordt als de tijd geïnterpreteerd. We beperken ons tot het geval, dat de waarden van t aequidistant zijn; zonder verlies van algemeenheid kunnen we dan het verschil = 1 stellen. Verder beperken we ons tot waarden $t \geq t_o$, waarbij we (zolang $t_o$ constant gehouden wordt) zonder beperking kunnen aannemen, dat $t_o = 0$ is.

   In de literatuur worden ook algemene tijdreeksen beschouwd, bijv. die welke gedefinieerd zijn voor alle reële waarden van t.

2. Beschouw in het bijzonder de tijdreeksen, gegeven door een lineaire differentievergelijking met constante coëfficiënten. In deze van de orde h, dan heeft deze de vorm

$$(1) \qquad x_t + a_1 x_{t-1} + \ldots + a_h x_{t-h} = \varepsilon_t \ . \qquad (a_h \neq 0)$$

Hierbij is $\varepsilon_t$ een gegeven tijdreeks en zijn $a_1,\ldots,a_h$ gegeven getallen. Meer in het bijzonder beschouwen we het geval h = 1

$$(2) \qquad x_t = \alpha x_{t-1} + \varepsilon_t$$

Indien geldt: $\varepsilon_t \equiv 0$, dan hebben we een <u>homogene</u> differentievergelijking, waarvan de oplossing luidt

$$(3) \qquad x_t = \alpha^t x_o \ .$$

Indien geldt: $|\alpha| < 1$, hebben we : $\lim_{t \to \infty} x_t = 0$.

De <u>inhomogene</u> differentievergelijking heeft tot oplossing:

$$(4) \qquad x_t = \sum_{1}^{t} \alpha^{t-j} \varepsilon_j + \alpha^t x_o$$

indien we $x_o = \varepsilon_o$ stellen, kunnen we voor (4) schrijven:

$$(5) \qquad x_t = \sum_{0}^{t} \alpha^{t-j} \varepsilon_j \ .$$

3. Op allerlei gebieden van toegepaste wiskunde komen tijdreeksen voor bijv. op meteorologisch gebied (de dagelijkse of jaarlijkse regenval op een bepaalde plaats, enz.), op agrarisch gebied (de gemiddelde jaarlijkse opbrengsten per HA van een bepaald gewas in een be

paald gebied, enz.), op demografisch gebied (de loop van een bevolking, geboortecijfers, enz.), op economisch gebied (maandelijkse productie-indices, enz.). De theorie van de tijdreeksen stelt zich ten doel, deze empirische reeksen wiskundig te analyseren. Dit doel zou bereikt zijn, indien bijv. de empirische reeks een oplossing bleek te zijn van een differentievergelijking. Aangezien het aantal waarnemingen van dergelijke reeksen steeds eindig is, is het altijd mogelijk tot een dergelijke differentievergelijking te komen. Echter, wanneer men op grond hiervan voorspellingen doet, komen zij (behalve in de astronomie of de natuurkunde) doorgaans niet uit.

Hieraan kan tegemoet gekomen worden, doordat we de empirische reeks beschouwen als een "toevallige" keuze uit een verzameling van eventuele tijdreeksen. We nemen dus aan, dat voor iedere t $\underline{x}_t$ een stochastische grootheid ') is, d.w.z. dat $\underline{x}_t$ een waarschijnlijkheidsverdeling heeft. $\underline{x}$ is dus een stochastische functie van t. Op de algemene theorie van de stochastische functies en de stochastische processen gaan we hier niet in.

We zullen ons ook hier beperken tot het geval, dat een tijdreeks kan worden voorgesteld door een differentievergelijking van de eerste orde met constante $\alpha$. $\underline{\varepsilon}_t$ is dan een stochastische functie van t. We nemen aan, dat $\underline{\varepsilon}_0, \ldots, \underline{\varepsilon}_t$ onafhankelijk verdeeld zijn volgens dezelfde verdeling.

De oplossing (5) wordt nu

$$(6) \qquad \underline{x}_t = \sum_0^t \alpha^{t-j}\, \underline{\varepsilon}_j \,.$$

4. Bij gegeven $\alpha$ en gegeven verdeling van $\underline{\varepsilon}_t$ kan men de verdeling vinden van $\underline{x}_t$ of van bepaalde functies daarvan. Nu heeft men in empirische reeksen vaak gezocht naar periodiciteiten, bijv. met behulp van de zgn. periodigram-analyse. Voor een onderzoek naar deze periodiciteiten zullen wij ons hier echter richten naar Kendall's begrippen "upcrosses" en "downcrosses". We nemen daartoe aan, dat de verwachting van $\underline{\varepsilon}_t$ en dus ook van alle $\underline{x}_t$ gelijk 0 is. Dan is een "upcross" resp. een "downcross" een punt t+1 langs de tijdas, waarvoor geldt

$$(7) \qquad \begin{array}{ccc} \underline{x}_t < 0 & & \underline{x}_t > 0 \\ & \text{resp.} & \\ \underline{x}_{t+1} > 0 & & \underline{x}_{t+1} < 0 \end{array}$$

---

') Stochastische grootheden worden door onderstreping onderscheiden van grootheden, die geen verdeling hebben.

We veronderstellen, dat de verdelingsfunctie van $\underline{\varepsilon}_t$ en dus van alle $\underline{x}_t$ continu is. Dan is $P[\underline{x}_t = 0] = 0$. Hierdoor sluiten we uit de mogelijkheden: $\underline{x}_{t-1} > 0 = \underline{x}_t < x_{t+1}$ en $x_{t-1} < 0 = x_t > x_{t+1}$.

Nu volgt uit (6), dat het zinvol is, te spreken over de waarschijnlijkheid, dat voor een tweetal punten t,t+1 de ongelijkheden (7) gelden. Langs de tijdas zal men dan een verdeling van punten t, die "upcrosses" resp. "downcrosses" zijn, vinden. De afstand tussen twee opeenvolgende van dergelijke punten is dus een stochastische grootheid. We zullen ons nu bezig houden met de verdeling van deze afstand.

5. M.G. Kendall[λ]) heeft zich bezig gehouden met de verwachting van deze afstand, onder de voorwaarde, dat alle $\underline{\varepsilon}_t$ onafhankelijk normaal verdeeld zijn met verwachting 0 en variantie $\sigma^2$. Daartoe berekende hij de waarschijnlijkheid, dat zich op t+1 een "upcross" zou voordoen, d.w.z. dat geldt

$$(8) \qquad \underline{x}_t = \sum_{0}^{t} \alpha^{t-j} \underline{\varepsilon}_j = \sum_{-1}^{t} \alpha^{t-j} \underline{\varepsilon}_j < 0$$

$$(9) \qquad \underline{x}_{t+1} = \sum_{0}^{t+1} \alpha^{t+1-j} \underline{\varepsilon}_j = \sum_{-1}^{t} \alpha^{t-j} \underline{\varepsilon}_{j+1} > 0,$$

waarbij we ter vereenvoudiging van de formules $\underline{\varepsilon}_{-1} \equiv 0$ gesteld hebben. Deze waarschijnlijkheid is gelijk aan de massa van

$$(10) \qquad d\,F = K \exp\left[-\frac{1}{2\sigma^2} \sum_{0}^{t+1} \varepsilon_j^2\right] d\,\varepsilon_0 \ldots d\,\varepsilon_{t+1}$$

(K is een constante), die ligt tussen de hypervlakken $\sum_{-1}^{t} \alpha^{t-j} \varepsilon_j = 0$ en $\sum_{-1}^{t} \alpha^{t-j} \varepsilon_{j+1} = 0$; immers, de projectie van d F op het $\varepsilon_1, \ldots, \varepsilon_{t+1}$-hypervlak is alzijdig symmetrisch. Deze massa is evenredig met de hoek $\theta_t$ tussen eerstgenoemde hypervlakken; hierbij geldt

$$\cos\,\theta_t = \frac{\sum_{-1}^{t-1} \alpha^{t-j}\,\alpha^{t-j-1}}{\sum_{-1}^{t} (\alpha^{t-j})^2} =$$

$$(11) \qquad = \frac{1-\alpha^{2(t+1)}}{1-\alpha^{2(t+2)}}\,\alpha$$

Hieruit volgt:

$$(12) \qquad \lim_{t\to\infty} \cos\,\theta_t = \alpha$$

---

λ) The advanced theory of statistics, II, p. 381 e.v.

Op dezelfde wijze vindt men, als men de veronderstelling $t_o = 0$ laat varen, voor iedere constante t:

$$(13) \qquad \lim_{t_o \to -\infty} \cos \theta_t = \alpha .$$

De asymptotische waarschijnlijkheid van een upcross is dus

$$(14) \qquad p = \frac{\text{arc cos}\,\alpha}{2\pi}$$

Hieruit concludeert Kendall, dat de verwachting van de afstand tussen twee opeenvolgende "upcrosses" gelijk is aan $2\pi/\text{arc cos}\,\alpha$. Wij zullen trachten dit te interpreteren.

Voeg aan de reeks $\underline{x}_t$ een tweede reeks $\underline{y}_t$ toe met de eigenschap, dat als $\underline{x}_t$ voor $t = t_1$ een "upcross" geeft, $\underline{y}_{t_1}$ de waarde 1 heeft, terwijl $\underline{y}_t = 0$ is voor alle waarden van t, die deze eigenschap niet hebben. Dan hebben alle $\underline{y}_t$ een simultane verdeling met de eigenschap $\mathcal{E}\,\underline{y}_t = p$. Wanneer we nu een tijdreeks $\underline{x}_t$ van de lengte n beschouwen, dan krijgen we

$$(15) \qquad \lim_{n \to \infty} \mathcal{E}\,\frac{1}{n} \sum_{1}^{n} \underline{y}_t = p .$$

Wanneer we het aantal "upcrosses" in de reeks $\underline{x}_t$ noemen: $\underline{d}_n$, dan geldt dus

$$(16) \qquad \lim_{n \to \infty} \mathcal{E}\, n^{-1}\, \underline{d}_n = p .$$

Noem verder de gemiddelde afstand tussen opvolgende "upcrosses" $\underline{\bar{l}}_n$, dan geldt

$$(17) \qquad \lim_{n \to \infty} \underline{\bar{l}}_n = \lim_{n \to \infty} n \cdot \underline{d}_n^{-1} .$$

Nu is $d_n^{-1}$ een positief-convexe functie van $d_n$ voor $d_n > 0$, d.w.z. dat iedere koorde van deze kromme boven de kromme ligt.[3]) In dat geval zegt de ongelijkheid van Jensen:

$$(19) \qquad \mathcal{E}\,\underline{d}_n^{-1} \geqslant (\mathcal{E}\,\underline{d}_n)^{-1} = \frac{1}{np} ,$$

zodat $\lim_{n \to \infty} \mathcal{E}\,\underline{\bar{l}}_n \geqslant p^{-1}$ is.

Kendall gaat zelfs verder, en concludeert dat

$$(20) \qquad \lim_{n \to \infty} \mathcal{E}\,\underline{\bar{l}}_n = \lim_{n \to \infty} \mathcal{E}\, n\, \underline{d}_n^{-1} = p^{-1}$$

is. Dat deze limiet $= p^{-1}$ is, is echter niet bewezen. In het geval, dat de opeenvolgende $\underline{y}_t$ onderling onafhankelijk zijn, is de overeen-

---

[3]) Cf. D. van Dantzig, Capita Selecta der Waarschijnlijkheidsrekening. Caput II: Momenten en Ongelijkheden; pp. 93-96.

komstige stelling wel te bewijzen.[4]) Dit is hier echter zeker niet het geval, daar de mogelijkheden $y_t = 1$ en $y_{t+1} = 1$ elkaar zelfs uitsluiten. Hoewel de stelling van Kendall dus niet bewezen is, is echter wel te verwachten, dat zij juist zal blijken te zijn.

6. Ter vergelijking zullen we de verdeling voor de tijdreeks met genoemde stochastische differentievergelijking bestuderen onder de voorwaarde, dat $\varepsilon_t$ onafhankelijk homogeen verdeeld is in het interval $(-\frac{1}{2}, \frac{1}{2})$[5]), d.w.z. dat geldt:

$$
\begin{aligned}
P\left[\varepsilon_t \leqslant s\right] &= 0 \quad \text{voor } s \leqslant -\tfrac{1}{2} \\
&= \tfrac{1}{2} + s \text{ voor } -\tfrac{1}{2} \leqslant s \leqslant \tfrac{1}{2} \\
&= 1 \quad \text{voor } s \geqslant \tfrac{1}{2} .
\end{aligned} \tag{21}
$$

We nemen verder aan, dat $|\alpha|$ zo klein is, dat $\alpha^2$ en hogere machten van verwaarloosd kunnen worden. In dat geval geldt bij benadering[6]):

$$x_t = \varepsilon_t + \alpha \varepsilon_{t-1} \tag{22}$$

Dan beschouwen we een tweetal termen in de reeks, bijv. $x_0$ en $x_1$, waarvan we veronderstellen, dat de eerste positief, de tweede negatief is. In dat geval begint met $x_1$, een iteratie ("run") van termen met negatief teken. Wij zullen nu waarschijnlijkheid berekenen, dat deze iteratie minstens de lengte k heeft, d.w.z. dat geldt

$$x_0 > 0,\ x_1 < 0,\ x_2 < 0, \ldots,\ x_k < 0. \tag{23}$$

---

4) Prof. van Dantzig deelt mede, dat zijn uitspraak (Sur la méthode des fonctions génératrices, Coll. de Lyon, 1949, 26-46) "La distribution ..... approximativement" (l.c. p. 42) niet juist is. Met de l.c. gevolgde methode kan bewezen worden, dat voor $k = 2$, $p_1 = p$, $p_2 = q = 1 - p$ deze verwachting gelijk is aan

$$p^{-1}\left[\, 1 + \frac{p^{n+2} - q^{n+2}}{p - q} \,\right],$$

dus voor n tot $p^{-1}$ nadert mits $pq \neq 0$ is.

5) Voor de verdeling van de afstand tussen "upcrosses" maakt het uiteraard geen verschil, of $\varepsilon_t$ homogeen verdeeld is tussen $-a$ en $+a$, dan wel tussen $-\frac{1}{2}$ en $+\frac{1}{2}$.

6) Alle verdere vergelijkingen in deze paragraaf gelden behoudens machten van de tweede en hogere graad in $\alpha$.

In symbolen luidt deze waarschijnlijkheid

$$(24) \qquad P\left[\underline{x}_2 < 0, \ldots, \underline{x}_k < 0 \mid \underline{x}_0 > 0, \underline{x}_1 < 0\right]$$

Volgens het vermenigvuldigingsaxioma is zij gelijk aan

$$(25) \qquad \frac{P\left[\underline{x}_0 > 0, \underline{x}_1 < 0, \ldots, \underline{x}_k < 0\right]}{P\left[\underline{x}_0 > 0, \underline{x}_1 < 0\right]}$$

De noemer hiervan is gelijk aan

$$(26) \qquad \begin{aligned} & P\left[\underline{\varepsilon}_1 + \alpha \underline{\varepsilon}_0 < 0, \underline{\varepsilon}_0 + \alpha \underline{\varepsilon}_{-1} > 0\right] = \\ & = \int_{-\frac{1}{\lambda}}^{\frac{1}{\lambda}} d\varepsilon_{-1} \int_{-\alpha\varepsilon_{-1}}^{\frac{1}{\lambda}} d\varepsilon_0 \left(\tfrac{1}{2} - \alpha \varepsilon_0\right) = \\ & \tfrac{1}{4} - \frac{\alpha}{8} \ . \end{aligned}$$

Vervolgens de teller van (20). Stel k = 2. Dan geldt:

$$(27) \qquad \begin{aligned} & P\left[\underline{\varepsilon}_2 + \alpha \underline{\varepsilon}_1 < 0, \underline{\varepsilon}_1 + \alpha \underline{\varepsilon}_0 < 0, \underline{\varepsilon}_0 + \alpha \underline{\varepsilon}_{-1} > 0\right] = \\ & = \int_{-\frac{1}{\lambda}}^{\frac{1}{\lambda}} d\varepsilon_{-1} \int_{-\alpha\varepsilon_{-1}}^{\frac{1}{\lambda}} d\varepsilon_0 \int_{-\frac{1}{\lambda}}^{-\alpha\varepsilon_0} d\varepsilon_1 \left(\tfrac{1}{2} - \alpha \varepsilon_0\right) = \\ & = \int_{-\frac{1}{\lambda}}^{\frac{1}{\lambda}} d\varepsilon_{-1} \int_{-\alpha\varepsilon_{-1}}^{\frac{1}{\lambda}} d\varepsilon_0 \left(\tfrac{1}{4} + \frac{\alpha}{8} - \frac{\alpha}{2} \varepsilon_0\right) \end{aligned}$$

Voor willekeurige positieve gehele k geldt algemeen:

$$(28) \qquad \begin{aligned} & P\left[\underline{\varepsilon}_k + \alpha \underline{\varepsilon}_{k-1} < 0, \ldots, \underline{\varepsilon}_0 + \alpha \underline{\varepsilon}_{-1} > 0\right] = \\ & = \int_{-\frac{1}{\lambda}}^{\frac{1}{\lambda}} d\varepsilon_{-1} \int_{-\alpha\varepsilon_{-1}}^{\frac{1}{\lambda}} d\varepsilon_0 \left(a_k + b_k - c_k \alpha \varepsilon_0\right) = \\ & = \tfrac{1}{2} a_k + \left(\tfrac{1}{2} b_k - \frac{1}{8} c_k\right)\alpha . \end{aligned}$$

Hierbij is

$$(29) \qquad a_k = \tfrac{1}{2} a_{k-1}$$

$$(30) \qquad b_k = \tfrac{1}{2} b_{k-1} + \frac{1}{8} c_{k-1}$$

$$(31) \qquad c_k = a_{k-1}$$

Hieruit en uit de waarden $a_1 = \frac{1}{2}$, $b_1 = 0$, $c_1 = 1$ (zie (21)) volgt:

$$(32) \qquad a_k = \left(\tfrac{1}{2}\right)^k$$

$$(33) \qquad b_k = \left(\tfrac{1}{2}\right)^{k+1} (k-1)$$

$$(34) \qquad c_k = \left(\tfrac{1}{2}\right)^{k-1}$$

Dus is de waarschijnlijkheid, dat de iteratie minstens de lengte k heeft

$$P\left[\underline{1}' \geqslant k\right] = \frac{(\frac{1}{2})^{k+1} + (\frac{1}{2})^{k+2}\,(k-2)\,\alpha}{\frac{1}{4} - \frac{\alpha}{8}} = \tag{35}$$

$$= \frac{2 + (k-1)\alpha}{2^k}$$

De complementsfunctie (gedefinieerd als 1 minus de verdelingsfunctie) van $\underline{1}'$ wordt gegeven door: (36) $P[\underline{l}' > k] = \frac{1+\frac{1}{2}k\alpha}{2^k}$ ; de frequentie door:

$$P\left[\underline{1}' = k\right] = \frac{1 + (\frac{1}{2}k - 1)\alpha}{2^k} \tag{37}$$

De verwachting van $\underline{1}'$ is

$$\mathcal{E}\underline{1}' = \sum_{1}^{\infty} P\left[\underline{1}' \geqslant k\right] = 2 + \alpha . \tag{38}$$

Voor de variantie vindt men

$$\text{var}\,(\underline{1}') = 2 + 3\alpha . \tag{39}$$

7. In de vorige paragraaf beschouwden we de afstanden van "downcross" tot "upcross"; noem de eerste dezer afstanden $\underline{1}'_1$, de tweede $\underline{1}'_2$ enz. Uit symmetrie-overwegingen volgt, dat de afstanden van "upcross" tot "downcross" $\underline{1}''_1$ , $\underline{1}''_2$ ,... dezelfde verdeling hebben. Dan geldt voor een "gehele periode" van "upcross" tot "upcross" of van "downcross" tot "downcross" $\underline{1}_i = \underline{1}'_i + \underline{1}''_i$ :

$$\mathcal{E}\underline{1}_i = 4 + 2\alpha + \mathcal{O}(\alpha^2) \tag{40}$$

Wanneer de reeks nu bestaat uit m afstanden van "upcross" tot "upcross", dan geldt voor de gemiddelde afstand, dat zijn voorwaardelijke verwachting bij gegeven m gegeven wordt door:

$$\mathcal{E}\,(\bar{\underline{1}} \mid m) = \mathcal{E}\underline{1}_i . \tag{41}$$

We geven de voorwaardelijke verwachting bij gegeven n van de gemiddelde afstand van "upcross" tot "upcross" in geval van normaal verdeelde $\underline{\varepsilon}_t$ aan door $\lambda_N^{(n)}$ en haar limiet voor $n\to\infty$ door $\lambda_N$ . Verder geven we de voorwaardelijke verwachting bij gegeven m van genoemde afstand in geval van homogeen verdeelde $\underline{\varepsilon}_t$ aan door $\lambda_H$. Dan geldt, indien Kendall's formule juist is:

$$\lambda_N = 4 + \frac{8}{\pi}\alpha + \mathcal{O}(\alpha^2) \tag{42}$$

$$\lambda_H = 4 + 2\alpha + \mathcal{O}(\alpha^2) . \tag{43}$$

Hieruit volgt (althans voor voldoend· kleine waarden van $|\alpha|$ ) bij eenzelfde waarde van $\alpha$, dat $\lambda_N$ meer afwijkt van de verwachting 4 in geval van onafhankelijk verdeelde $x_t$ dan $\lambda_H$. [7]) Men moet in het oog houden, dat de verwachtingen $\lambda_N$ en $\lambda_H$ onder verschillende voorwaarden berekend zijn. Het lijkt ons echter plausibel, dat de betekenis van deze verschillende voorwaarden asymptotisch voor $n\to\infty$ verdwijnt.

---

[7]) Is Kendall's formule onjuist, dan moet het gelijkheidsteken in de vergelijking (42) vervangen worden door een groter-teken, zodat deze conclusie gehandhaafd blijft.