MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

1953 - 32(1)

Colloquium over hoge waterstanden.

Theorie der uiterste waarden I

door

H.Kesten en J.Th.Runnenburg

1953

. Inleiding.

Dit verslag bevat een kort overzicht van de opzet van de theorie der uiterste waarden, ontwikkeld door FRÉCHET en GUMBEL Het is geschreven naar aanleiding van een voordracht over dit onderwerp, gehouden door Dr C. LEVERT op het Mathematisch Centrum op 2 Juni 1953. Alleen de grootste waarneming van een reeks onafhankelijke waarnemingen is beschouwd.

§1.1. Limietverdeling van de grootste waarneming.

Als $w(x)$ de verdelingsdichtheid van een stochastische grootheid $\underline{x}$ is, met verdelingsfunctie $W(x)$, is de verdelingsfunctie $F_N(x)$ van de grootste onder N onderling onafhankelijke waarnemingen van $\underline{x}$:

$$F_N(x) = W^N(x)$$

Het probleem van GUMBEL is, voor grote N voor $F_N(x)$ een benadering te vinden, die onafhankelijk is van $W(x)$.

GUMBEL citeert het volgende van FRÉCHET (1927) afkomstige resultaat (hier en verder veronderstellen we dat $\underline{x}$ een naar $+\infty$ onbegrensde verdeling heeft. Deze onderstelling is niet noodzakelijk):

Als voldaan is aan:

$$\lim_{x\to\infty} x\,\alpha(x) = k > 0 \qquad (1.1.1)$$

waarin $\alpha(x) \overset{\text{def}}{=} \dfrac{w(x)}{1 - W(x)}$ en k een positieve constante is, geldt:

$$\lim_{N\to\infty} F_N(x)\, e^{\left(\frac{u}{x}\right)^k} = 1 \qquad (1.1.2)$$

Hierbij wordt u bepaald door $N(1 - W(u)) = 1$.

GUMBEL zelf beschouwt die gevallen waarin voldaan is aan:

$$\lim_{x\to\infty} \frac{d}{dx}\,\frac{1}{\alpha(x)} = 0 \qquad (1.1.3)$$

Hiervoor geldt:

$$\lim_{N\to\infty} F_N(x)\, e^{e^{-\alpha(u)(x-u)}} = 1 \qquad (1.1.4)$$ (GUMBEL (1941))

Het bewijs voor (1.1.2) en (1.1.4) wordt in het genoemde artikel door GUMBEL niet gegeven, zodat niet blijkt of zijn voorwaarden volledig zijn.

Als we eisen:

$$x\,\alpha(x) = k > 0 \qquad (1.1.1,5)$$

$$\text{of} \qquad x\,W'(x) = k(1 - W(x))$$

dan is

$$W(x) = 1 - \frac{c}{x^k} \qquad (x \geq c \text{ , een constante}).$$

Deze verdelingsfunctie is bekend onder de naam Paretoverde-

ling.

Evenzo vinden we, bij het weglaten van het limietteken in (1.1.3)

$$(1.1.3,5)\qquad \frac{d}{dx}\,\frac{1-W(x)}{W(x)} = 0 \qquad \text{of}$$

$$1 - W(x) = a\,W'(x)$$

dus

$$W(x) = 1 - e^{-a(x-b)} \qquad (x \geq b;\ a \text{ en } b \text{ constanten}, a > 0).$$

Dit geeft ons dus de exponentiële verdeling.

De staart van de verdeling $W(x)$, d.w.z. $1 - W(x)$, gaat bij (1.1.1,5) veel langzamer naar o , dan bij (1.1.3,5). We kunnen bovendien aantonen, dat een $W(x)$, waarvoor geldt:

$$(1.1.5)\qquad 1 - W(x) = x^{-f(x)} \qquad (x \geq d)$$

waarbij $f(x)$ voor voldoende grote x een monotoon stijgende functie van x is, zodanig, dat $f(x) \to \infty$ voor $x \to \infty$ en bovendien $W(d) = 0$, niet aan (1.1.1) kan voldoen. Om $\alpha(x)$ te kunnen bepalen, moeten we $f(x)$ differentieerbaar veronderstellen. Nu is:

$$x\,\alpha(x) = \frac{x\,W(x)}{1 - W(x)} = \frac{x\{x^{-f(x)} f(x) + f'(x)\log x\}\,e^{-f(x)\log x}}{e^{-f(x)\log x}} =$$

$$= f(x) + x f'(x)\log x$$

Voor monotoon stijgende $f(x)$ is $f'(x) \geq 0$, zodat dus $x\,\alpha(x)$ voor $x \to \infty$ naar ∞ gaat en dus zeker niet naar een constante $k > 0$ convergeert.

Dat daarentegen aan (1.1.3) wel te voldoen is, met een verdeling $W(x)$ van het type (1.1.5) die langzamer tot 1 nadert voor $x \to \infty$ dan de exponentiële verdeling, blijkt aldus:

$$\alpha(x) = \frac{f(x) + x f'(x)\log x}{x}$$

$$\frac{d}{dx}\,\frac{1}{\alpha(x)} = \frac{1}{f(x) + x f'(x)\log x} - \frac{x\{2 f'(x) + f'(x)\log x + f''(x)\,x\log x\}}{\{f(x) + x f'(x)\log x\}^2}$$

Dus is $$\lim_{x\to\infty} \frac{d}{dx}\,\frac{1}{\alpha(x)} = -\lim_{x\to\infty} \frac{x^2 f''(x)\log x}{\{f(x) + x f'(x)\log x\}^2}$$

Om aan (1.1.3) te voldoen, kunnen we b.v. eisen:

$$\lim_{x\to\infty}\left|\frac{d}{dx}\,\frac{1}{\alpha(x)}\right| = \lim_{x\to\infty}\left|\frac{x^2 f''(x)\log x}{\{f(x) + x f'(x)\log x\}^2}\right| \leq$$

$$\leq \lim_{x\to\infty}\left|\frac{f''(x)}{\{f'(x)\}^2 \log x}\right| = 0$$

Hieraan voldoet bv. $f(x)=x^k$ voor $k>0$ (met $d=1$). Voor $k<1$ nadert $x^{-x^k} = e^{-x^k \log x}$ langzamer tot o dan $e^{-d(x-b)}$ voor $x \to \infty$.

Voor $f(x) = \frac{x}{\log x}$ vinden we de exponentiële verdeling zelf.

We kunnen als volgt inzien, dat (1.1.1) en (1.1.3) niet tegelijkertijd vervuld kunnen zijn.

Stel $\frac{1}{\alpha(x)} = \frac{x}{k} + x f(x)$, waarin $f(x)$ voorlopig onbekend is.

Om aan (1.1.1) te voldoen, moeten we eisen:

$$\lim_{x\to\infty} \frac{1}{x\,\alpha(x)} = \frac{1}{k} + \lim_{x\to\infty} f(x) = \frac{1}{k}$$

dus

$$\lim_{x\to\infty} f(x) = 0$$

$$\frac{d}{dx}\frac{1}{\alpha(x)} = \frac{1}{k} + f(x) + x f'(x)$$

$$\lim_{x\to\infty} \frac{d}{dx}\frac{1}{\alpha(x)} = \frac{1}{k} + \lim_{x\to\infty} x f'(x) = 0$$

als we bovendien aan (1.1.3) willen voldoen.

Voor voldoende grote x (bv. $x \geq x_0$) moet dus gelden voor $0<\varepsilon<\frac{1}{k}$:

$$-\frac{1}{k} - \varepsilon \leq x f'(x) \leq -\frac{1}{k} + \varepsilon$$

of

$$\int_{x_0}^{x} \frac{(-\frac{1}{k}-\varepsilon)}{t}\,dt \leq \int_{x_0}^{x} f'(t)\,dt \leq \int_{x_0}^{x} \frac{(-\frac{1}{k}+\varepsilon)}{t}\,dt$$

waaruit: $(-\frac{1}{k}-\varepsilon)\log x \leq f(x) + \text{constante} \leq (-\frac{1}{k}+\varepsilon)\log x$

zodat $f(x)$ voor $x \to \infty$ naar $-\infty$ moet gaan, hetgeen onverenigbaar is met $\lim_{x\to\infty} f(x) = 0$

Uit het bovenstaande blijkt dus, dat het type verdelingen, dat aan (1.1.1) voldoet, een veel dikkere staart heeft dan het type dat aan (1.1.3) voldoet. Tevens hebben we gezien, dat niet aan (1.1.1) en (1.1.3) tegelijkertijd voldaan kan zijn.

In de practijk hebben we misschien met grote, maar zeker altijd met eindige N te doen. Om nu van de limietrelaties gebruik te kunnen maken, veronderstelt GUMBEL, dat die relaties ook voor grote N geldig zijn, dus resp. als aan (1.1.1) en aan (1.1.3) voldaan is:

(1.1.2,5) $F_N(x) \approx e^{-\left(\frac{u}{x}\right)^k}$ voor grote N

(1.1.4,5) $F_N(x) \approx e^{-e^{-\alpha(u)(x-u)}}$ voor grote N

We zullen ons verder evenals GUMBEL beperken tot het tweede geval, dus tot (1.1.4,5). In principe is het echter wel mogelijk de

benadering (1.1.2,5) te gebruiken. De parameters u en k kunnen uit het materiaal geschat worden.

In de kadercursus van Prof. Dr D.VAN DANTZIG (1947) (hfdst. 6, § 2) wordt een precisering gegeven van de vervanging van $F_N(x)$ door $e^{-e^{-\alpha(u)(x-u)}}$ volgens (1.1.4,5) en wel de volgende:

Als $1 - W(x) = e^{-S(x)}$ continu differentieerbaar is en $S'(x)$ voor voldoende grote x positief is en

$$\lim_{x\to\infty} \frac{d}{dx} \frac{1-W(x)}{w(x)} = \lim_{x\to\infty} \frac{S''(x)}{\{S'(x)\}^2} = 0 \quad \text{is,}$$

kunnen we voor voldoend grote N ($N \geq n_0(\alpha,\delta,\varepsilon)$) stellen:

$$dF_N(x) = e^{-y-e^{-y}}\,dy$$

Voor deze benadering geldt in een interval $\delta \leq 1-F_N(x) \leq 1-\alpha$ (α en δ willekeurig, $\alpha > 0$, $0 < \delta < 1-\alpha$), dat y hoogstens ε (een van te voren gekozen, positief getal) afwijkt van $(x-\hat{x})S'(\hat{x})$. Hierbij is $\hat{x}$ die waarde van x waarvoor geldt $W(\hat{x}) = 1 - \frac{1}{N} + O(N^{-2})$.

Tevens wordt voor een aantal verdelingen een betere benadering gegeven dan die van GUMBEL, maar hiervoor moet men de beschikking hebben (althans numeriek) over $W(x)$.

Als men met behulp van de benadering van GUMBEL de waarde X bepaalt, die door de grootste waarde uit een steekproef met uitgebreidheid N behoudens een kans α niet overschreden wordt, zal de ware overschrijdingskans van die X niet precies α zijn, maar $\alpha + \Delta\alpha$ ($\Delta\alpha$ kan negatief zijn). Bij een niet te gunstige verdeling, b.v. de normale, en een kleine α (b.v. 0,05) hebben we een vrij grote N nodig als we $\left|\frac{\Delta\alpha}{\alpha}\right|$, b.v. $\leq \frac{1}{5}$ wensen. In genoemd voorbeeld van de normale verdeling moet $N \geq 73790$ zijn.

In het algemeen kunnen we zeggen dat voor een exponentiële verdeling, d.w.z. $W(x) = 1 - e^{-\alpha(x-b)}$ (voor α positief, $x \geq b$), de benadering goed is; voor een verdeling, waarbij $1-W(x)$ voor x naar ∞ sneller naar nul gaat dan $e^{-\alpha x}$ is de benadering minder goed (of slecht, b.v. bij normale verdeling en de verdeling van GOMPERTZ) maar aan de <u>veilige</u> kant. Met veilig wordt hier bedoeld, dat bij deze benadering de kans op overschrijding van een bepaalde waarde te groot genomen wordt ($\Delta\alpha < 0$). Bij een verdeling, waarbij $1-W(x)$ langzamer naar nul gaat dan $e^{-\alpha x}$ is de benadering eveneens minder goed of zelfs niet toepasbaar, indien niet aan (1.1.3) voldaan is. De voorwaarden, die we moeten opleggen, wil de benadering werkelijk voldoen, zijn van uitgesproken mathematisch karakter, zodat ze niet getoetst kunnen worden.

§ 1.2. Schatting van de parameters van de verdeling van de uiterste waarde.

Indien we toch de benadering (1.1.3) willen gebruiken, is het nodig over een schattingsmethode voor u en $\alpha(u)$ te beschikken. We kunnen een schatting vinden, door gebruik te maken van de voor grote N en grote x geldende relatie:

(1.2.1) $$\log \frac{1}{1 - F_N(x)} \approx \alpha(u)(x-u) \qquad \text{dus}$$

door $\log \frac{1}{1 - F_N(x)}$ uit te zetten als functie van x en door de waargenomen punten een rechte te trekken. De parameters van de rechte geven dan direct $\alpha(u)$ en u. We veronderstellen hier dus dat we over een aantal waarnemingen van hoogste waarden uit steekproeven van N waarnemingen beschikken. (1.2.1) is een gevolg van (1.1.4,5). $\alpha(u)$ is positief, dus voor grote x is $-\alpha(u)(x-u)$ negatief en groot in absolute waarde, dus $e^{-\alpha(u)(x-u)}$ weinig van nul verschillend, zodat:

$$F_N(x) \approx e^{-e^{-\alpha(u)(x-u)}} \approx 1 - e^{-\alpha(u)(x-u)}$$

dus $$\log \frac{1}{1 - F_N(x)} \approx \alpha(u)(x-u)$$

GUMBEL heeft voor de schatting van $\alpha(u)$ en u een speciaal waarschijnlijkheidspapier ontworpen.

Hierbij worden op de ordinaat de waargenomen uiterste waarden (x) uitgezet en op de abscis een schatting voor $F_N(x)$. Door de uitgezette punten trekken we een rechte, die ons een schatting geeft voor de rechte, die we (door de constructie van het waarschijnlijkheidspapier) zouden krijgen, wanneer $\alpha(u)$ en u exact bekend waren.

We kunnen op verschillende manieren een schatting voor $F_N(x)$ bepalen.

GUMBEL bespreekt een aantal manieren in een van zijn artikelen (zie GUMBEL(1945)).

Als we de waarnemingen naar opklimmende grootte ordenen: $x_1, \ldots\ldots x_m$, kunnen we b.v. nemen:

a) $$F_N(x_k) = \frac{k - \frac{1}{2}}{m}$$

als compromis tussen $\frac{k}{m}$ en $\frac{k-1}{m}$.

Als verwachting van het aantal nodige waarnemingen tot de eerstvolgende overschrijding van de grootste waarde vinden we dan

$$T(x_m) \overset{\text{def}}{=} \frac{1}{1 - F_N(x_m)} = \frac{1}{1 - \frac{m-\frac{1}{2}}{m}} = 2m$$

zodat een gebeurtenis, die volgens de steekproef eens in de m keer voorkomt, nu verondersteld wordt gemiddeld slechts eens in $2m$ keer voor te komen. Op grond hiervan verwerpt GUMBEL deze methode.

Opgemerkt dient te worden dat het gebruik van de schatting $1 - \frac{1}{m}$ (waarbij m het aantal waarnemingen is) voor $F_N(x_m)$ het volgende bezwaar heeft. De kans, dat de grootste waarneming verder in de staart van de verdeling valt dan het punt waarvoor theoretisch geldt $F_N(x_m) = 1 - \frac{1}{m}$, is voor grote m belangrijk groter dan $\frac{1}{2}$. Hoewel bij deze keuze wel $T(x_m) = m$ wordt, is nu $T(x_1) = 1$; bij $F_N(x_m) = 1$ is $T(x_m) = \infty$, zodat geen der genoemde methoden erg aantrekkelijk is.

b) De verdelingsfunctie $G_{i,m}$ van $\underline{x}_i$, de i-de waarde (van beneden af) van een steekproef van m onafhankelijke waarnemingen van uiterste waarden, is een functie van $F_N(x_i)$.

Als schatting voor $F_N(x_i)$ kunnen we de mediaan van $G_{i,m}(F_N(x_i))$ nemen, dus die waarde waarvoor

$$G_{i,m}(F_N(x_i)) = \frac{1}{2}$$

is. De waarde x_i waarvoor dit geldt, geven we aan met $\breve{x}_i$. Dit geeft voor de grootste uiterste waarde:

$$\left\{F_N(\breve{x}_m)\right\}^m = \frac{1}{2}$$

Dus

$$\frac{1}{1 - F_N(\breve{x}_m)} = \frac{1}{1 - 2^{\frac{1}{m}}} = \frac{1}{1 - e^{-\frac{\log 2}{m}}} = \frac{1}{\frac{\log 2}{m} - \frac{(\log 2)^2}{2m^2}} =$$

$$= \frac{m}{\log 2} + \frac{1}{2} + \cdots = 1{,}443\, m + \frac{1}{2} + \cdots .$$

Om een analoge reden als onder a) genoemd wordt deze methode door GUMBEL eveneens verworpen.

c) In geval we te doen hebben met waarnemingen van de dagelijks door een dwarsdoorsnede van een rivier stromende hoeveelheid water $\underline{z}_i$ en die waarnemingen voor een aantal jaren beschouwen, beveelt GUMBEL de volgende methode aan:

De verdelingsfunctie van $\underline{x}$, de grootste $\underline{z}_i$ van een jaar is

(1.2.2) $$F(x) = e^{-e^{-y}}$$

(1.2.3) $$f(x) = F'(x) = \alpha(u)\, e^{-y - e^{-y}}$$

(1.2.4) $$f'(x) = \alpha(u)\, f(x)\,(-1 + e^{-y})$$

(1.2.5) met $$y = \alpha(u)(x - u)$$

De modus $\tilde{x}_m$ van $\underline{x}_m$ de grootste x van m jaar kunnen we als volgt bepalen:

De verdelingsdichtheid van de stochastische grootheid $\underline{x}_m$ is $m\,F^{m-1}(x)\,f(x)$, dus de modus $\tilde{x}_m$ voldoet aan

$$m(m-1)\,F^{m-2}(\tilde{x}_m)\,f^2(\tilde{x}_m) + m\,F^{m-1}(\tilde{x}_m)\,f'(\tilde{x}_m) = 0$$

zodat

$$m-1 = -\frac{F(\tilde{x}_m)\,f'(\tilde{x}_m)}{\{f(\tilde{x}_m)\}^2};$$

dit geeft met (1.2.2), (1.2.3), (1.2.4) en (1.2.5):

$$m-1 = -\frac{\alpha(u)\,f(\tilde{x}_m)(e^{-\tilde{y}}-1)e^{-e^{-\tilde{y}}}}{f(\tilde{x}_m)\,\alpha(u)\,e^{-\tilde{y}}-e^{-\tilde{y}}} = -1+e^{\tilde{y}}$$

of

$$m = e^{\alpha(u)(\tilde{x}_m-u)} = e^{\tilde{y}}$$

dus

$$F(\tilde{x}_m) = e^{-\frac{1}{m}}$$

$$\frac{1}{1-F(\tilde{x}_m)} = \frac{1}{1-e^{-\frac{1}{m}}} = m+\tfrac{1}{2}+\cdots$$

Nu vinden we dus een aanvaardbare verwachting van het benodigde aantal waarnemingen tot de eerstvolgende overschrijding van $\tilde{x}_m$. Voor $F(\tilde{x}_m)$ hebben we nu een waarde gevonden. Analoog kunnen we voor de modus van de kleinste x van m jaren $(\tilde{x}_1)$ het volgende verband vinden:

$$m = \frac{1}{F(\tilde{x}_1)} + \frac{1}{F(\tilde{x}_1)\log F(\tilde{x}_1)} - \frac{1}{\log F(\tilde{x}_1)}$$

Hieruit kunnen we door $F(\tilde{x}_1)$ te kiezen en daaruit m te bepalen en vervolgens te interpoleren met behulp van de verkregen m-waarden, $F(\tilde{x}_1)$ als functie van m berekenen.

De abscis van het waarschijnlijkheidspapier verdelen we nu aldus bij m waarnemingen:

Als eerste punt kiezen we $F(\tilde{x}_1)$, als laatste $F(\tilde{x}_m)$, de tussenliggende punten ($m-2$ in aantal) bepalen we door lineaire interpolatie naar $F(x)$, dus voor de k^{de} waarneming kiezen we:

$$F(x_k) = F(\tilde{x}_1) + \frac{(k-1)\,(F(\tilde{x}_m)-F(\tilde{x}_1))}{m-1}$$

Als we de grootste waarde uit een steekproef van de uitgebreidheid N voor m verschillende van die steekproeven uitzetten, verwachten we op het papier van GUMBEL een rechte lijn. Bij het K.N.M.I. heeft men opgemerkt, dat ook wanneer de uitgebreidheid N_i (van de i^{de} steekproef) voor verschillende i verschil-

lend is, de uitgezette punten ongeveer op een rechte lijn liggen. Dit is misschien te verklaren door grote uitgebreidheid N_i, waarbij de N_i onderling weinig verschillen.

Indien de waarnemingen dat toestaan (dit moet dus onderzocht worden), kunnen we ook direct $W(x)$ exponentieel kiezen en de parameters van de exponentiële verdeling schatten. Ir P.J.WEMELSFELDER heeft dit gedaan voor de waarnemingen van hoge waterstanden te Hoek van Holland (WEMELSFELDER (1939)).

De betekenis van de schattingen van de parameters kan pas worden overzien als we een betrouwbaarheidsinterval voor die parameters kunnen geven.

Men kan met behulp van de binomiale verdeling een betrouwbaarheidsinterval voor de bij enkele x waarden (de waargenomen waterpeilen) behorende $F_N(x)$ construeren. Deze intervallen zijn afhankelijk. In hoeverre hieruit dan ook conclusies kunnen worden getrokken over betrouwbaarheidsintervallen voor $F_N(x)$ met een x buiten het waarnemingsgebied zal nader onderzocht moeten worden.

Literatuur.

D.van Dantzig (1947), Kadercursus Mathematische Statistiek.

M.Fréchet (1927), Sur la loi de probabilité de l'écart maximum. Annales Soc. Polon. Math. 6 (1927).

E.J.Gumbel (1941), The return period of flood flows. Ann. Math. Stat. 12 (1941), p. 163-190.

E.J.Gumbel (1945), Simplified plotting of statistical observations. Trans. American Geophysical Union 26 (1945), p. 69-82.

P.J.Wemelsfelder (1939), Wetmatigheden in het optreden van stormvloeden. "De Ingenieur" 9 (1939), Bouw- en Waterbouwkunde 3.