KONINKLIJK NEDERLANDS
METEOROLOGISCH INSTITUUT

De Bilt

Verslagen
V - 269

S. Kruizinga

Enkele praktische gegevens voor de Gumbelverdeling

De Bilt, 1975

Publikatienummer: K.N.M.I. V-269 (S.B.)

U.D.C.: 519.2

1. Inleiding

De Gumbelverdeling is een van de drie limietverdelingen die bruikbaar zijn voor de beschrijving van de kansverdeling van extreme waarden. Onder een extreme waarde wordt dan verstaan, het maximum van een steekproef van N onafhankelijke trekkingen uit een willekeurige verdeling. De drie limiet-verdelingen worden uitgebreid behandeld door Gumbel [1]. In de praktijk is gebleken dat in vele gevallen de Gumbelverdeling ook uitstekend voldoet voor de beschrijving van de kansverdeling van het maximum van een continue stochastische functie binnen een gegeven tijdvak. Speciaal voor neerslag wordt de Gumbelverdeling veel toegepast, de grootheid die men hierbij dan bestudeert is bijvoorbeeld de maximale dagsom per jaar. Het is de bedoeling van dit rapport om een aantal theoretische begrippen en formules op een overzichtelijke manier bij elkaar te zetten. Verder worden de op het KNMI aanwezige hulpmiddelen kort besproken.

2. Theorie

2.1. De verdelingsfunctie

De verdelingsfunctie van de Gumbelverdeling wordt in het algemeen als volgt genoteerd,

$$F(x) = \exp(-\exp(-\alpha(x-u))) \qquad (2.1)$$

hierin zijn α en u respectievelijk de locatie- en de schaalparameter van de verdeling. De variabele x is het maximum van de steekproef en $F(x)$ geeft de onderschrijdingskans.

Voor het gemiddelde en de variantie van x gelden de volgende relaties;

$$\bar{x} = u + \gamma/\alpha \qquad (\gamma = .5772166) \qquad (2.2)$$

en

$$\sigma_x^2 = \pi^2/(6\alpha^2) \qquad (2.3)$$

Het is gebruikelijk om een standaardvariabele z in te voeren,

$$z = \alpha \left(x - u \right); \qquad (2.4)$$

deze variabele z wordt in veel relaties gebruikt en is via (2.1) op één-éénduidige manier gekoppeld aan de onderschrijdingskans.

2.2. Herhalingsgetal J.

De zeldzaamheid van een gegeven maximum wordt vaak uitgedrukt in de gemiddelde lengte van het tijdsinterval tussen twee overschrijdingen van het gegeven maximum. Aangezien de maxima die worden bestudeerd in vele gevallen maxima zijn over tijdvakken die slechts eenmaal per jaar voorkomen wordt dit getal vaak uitgedrukt in jaren. Men spreekt dan dus van een maximum dat slechts eens in de J jaren zal worden overschreden. Dit herhalingsgetal is één-éénduidig gekoppeld aan de onderschrijdingskans $F(x)$ en de standaardvariabele z via de volgende relaties:

$$J(x) = 1 / \left(1 - F(x)\right) \qquad (2.5)$$

$$J(z) = 1 / \left(1 - \exp(-\exp(-z))\right) \qquad (2.6)$$

2.3. Risico R

Bij het ontwerp van een installatie is het vaak zinniger om te spreken over het risico dat men loopt dat de installatie tijdens zijn levensduur door een catastrofe zal worden beschadigd dan te stellen dat de installatie gemiddeld eens in de J jaar (met J veel groter dan de levensduur) zal worden beschadigd.

Stel nu dat een installatie door het overschrijden van een grens x_g onherstelbaar wordt beschadigd dan is het risico $R_N(x_g)$ dat dit binnen een tijdvak van N jaren zal gebeuren als volgt gerelateerd aan de onderschrijdingskans

$$R_N(x_g) = 1 - \left(F(x_g)\right)^N \tag{2.7}$$

of aan de standaardvariabele z

$$R_N(x_g) = 1 - \left(e^{-e^{-z_g}}\right)^N \tag{2.8}$$

met dus $z_g = \alpha\,(x_g - u)$.

Het herhalingsgetal J en het risico R als volgt met elkaar verbonden

$$R_N(x_g) = 1 - \left(1 - 1/J(x_g)\right)^N \tag{2.9}$$

2.4. Enige voorbeelden van voorgaande formules

Stel dat de jaarlijks maximale 24 uursom neerslag een Gumbel-verdeling volgt met de parameters

$$u = 30 \text{ mm} \quad , \quad \alpha = 0{,}1 \text{ mm}^{-1}$$

Bij een maximum van 60 mm hoort dan dus een z waarde

$$z = 0{,}1\,(60 - 30) = 3$$

zie (2.4)

De onderschrijdingskans die hierbij hoort is

$$P(\underline{x} < 60\text{mm}) = F(z) = e^{-e^{-3}} = .951 \qquad \text{zie (2.1)}$$

Het herhalingsgetal J is

$$J = 1/(1 - 0{,}951) = 20{,}6 \qquad \text{zie (2.5-2.6)}$$

dit betekent dat we verwachten dat gemiddeld één van de 20.6 jaren een maximum 24 uursom zal hebben groter dan 60 mm. Het risico op overschrijding van 60 mm binnen 10 jaar is dan

$$R_{10}(60) = 1 - \left(e^{-e^{-3}}\right)^{10} = 0{,}39 \qquad \text{zie (2.8)}$$

oftewel er is een kans van 39% dat binnen tien jaar een maximum groter dan 60 mm optreedt.

2.5. Schatten van de parameters

In de praktijk zijn veelal de parameters van de verdeling niet bekend. Deze parameters zal men dan op grond van een steekproef van N maxima moeten schatten. In de literatuur zijn vele schattingsmethoden voor de parameters van de Gumbelverdeling te vinden. Voor grote N is de maximum-likelyhood methode de meest aangewezen schattingsmethode. Deze methode gaat echter bij kleine N systematisch fout schatten. Een gunstig alternatief vormt de schattingsmethode volgens Kimball (Gumbel [1] p.234). Door Sneyers [2] is aangetoond dat deze methode voor grote N ($N \geqslant 30$) praktisch net zo effectief is als de maximum-likelyhood methode.

De parameters worden als volgt geschat;

Zij $x_1, \dots, x_N$ de steekproef van de maxima en $x_1^*, \dots, x_N^*$ de geordende steekproef dan schatten we α volgens

$$1/\hat{\alpha} = c_N \left(\sum_{n=1}^{N} g_n \cdot x_n^* \right) \tag{2.10}$$

waarin g_n de gewichtsfactoren waarvoor geldt,

$$g_n = \left[1 - \sum_{i=n}^{N} 1/i \right] / N \tag{2.11}$$

en c_N een correctiefactor die alleen afhankelijk is van N. Sneyers [2] geeft voor c_n een benaderingsformule welke bruikbaar is voor $N \geqslant 7$. De parameter u wordt dan vervolgens geschat met,

$$\hat{u} = \frac{1}{N} \sum_{n=1}^{N} x_n + \gamma/\hat{\alpha} \tag{2.12}$$

In het verleden zijn, ook binnen het KNMI, meerdere schattingsmethoden in gebruik geweest. Vanaf heden zal door het Statistisch Bureau alleen de Kimball methode, zoals hiervoor beschreven, worden gebruikt. Dit vereenvoudigt het onderling vergelijken van resultaten. In voorkomende gevallen zullen oude resultaten worden aangepast.

2.6. Betrouwbaarheidsbanden

Het gegeven dat de parameters α en u uit de steekproef worden geschat houdt tevens in dat de geschatte waarden niet exact gelijk zullen zijn aan de werkelijke waarden. Op grond van maximum likelyhood theorie kan men aangeven binnen welk interval de werkelijke waarden van α en u zullen liggen. Deze betrouwbaarheidsintervallen zijn slechts geldig voor maximum likelyhood schatters. We kunnen ze in de praktijk ook gebruiken voor Kimball schatters voor $N \geq 20$, voor kleinere N moeten we deze intervallen slechts beschouwen als een eerste orde schatting. Voeren we de volgende grootheden in

$$\Delta\alpha = 1{,}53\,\hat{\alpha}$$
$$\Delta u = 2{,}06/\hat{\alpha} \tag{2.13}$$

dan kunnen we stellen, bij een steekproefgrootte van N, dat de werkelijke α en u met 95% zekerheid binnen de intervallen

$$(\hat{\alpha} - \Delta\alpha/\sqrt{N}\ ,\ \hat{\alpha} + \Delta\alpha/\sqrt{N})$$
$$(\hat{u} - \Delta u/\sqrt{N}\ ,\ \hat{u} + \Delta u/\sqrt{N}) \tag{2.14}$$

zullen liggen.

De onzekerheid in $\hat{\alpha}$ en $\hat{u}$ houdt tevens in dat we slechts met een gegeven onzekerheid de x behorend bij een gegeven z kunnen schatten en omgekeerd eveneens de z behorend bij een gegeven x. Voeren we nu de volgende grootheden in:

$$\hat{z}(x) = \hat{\alpha}\,(x - \hat{u}), \tag{2.15}$$

de schatting van z behorend bij gegeven x en

$$\hat{x}(z) = z/\hat{\alpha} + \hat{u}, \tag{2.16}$$

de schatting van x behorend bij gegeven z.

En verder de grootheden

$$\Delta_z(z) = 1{,}96\sqrt{(,608 \cdot z^2 + ,514 \cdot z + 1{,}109)} \tag{2.17}$$

$$\Delta_x(z) = \Delta_z(z)/\hat{\alpha} \tag{2.18}$$

Dan geldt, bij steekproefgrootte N, dat de werkelijke z en x met 95% zekerheid in de intervallen

$$\begin{gathered}\left(\hat{z}(x) - \Delta_z(\hat{z}(x))/\sqrt{N},\ \hat{z}(x) + \Delta_z(\hat{z}(x))/\sqrt{N}\right)\\ \left(\hat{x}(z) - \Delta_x(z)/\sqrt{N},\ \hat{x}(z) + \Delta_x(z)/\sqrt{N}\right)\end{gathered} \tag{2.19}$$

zullen liggen.

Voorbeelden:

Stel dat de parameters van de voorbeelden uit 2.4. geschat zijn uit een steekproef van honderd waarnemingen. Dan vinden we voor het betrouwbaarheidsinterval van $\hat{z}$

$$\left(0{,}1 - 1{,}53 * 0{,}1/\sqrt{100},\ 0{,}1 + 1{,}53 * 0{,}1/\sqrt{100}\right) = (0{,}085\ ,\ 0{,}115)$$

en voor het betrouwbaarheidsinterval van $\hat{x}$ vinden we (27.4, 32.6). Verder is bijv. x behorend bij een herhalingsgetal van 100 jaar ofte wel $z = 4.6$

$$\hat{x}\,(z = 4.6\,,\ J = 100) = 4.6/0.1 + 30 = 76 \text{ mm};$$

de Δx is dan

$$\Delta x = 1.96\sqrt{\left(0.608 * (4.6)^2 + 0.514 * 4.6 + 1.109\right)}/0.1 = 79 \text{ mm}.$$

Dus het betrouwbaarheidsinterval van $\hat{x}$ is dan (68, 84). Bij een neerslaghoogte van 60 mm hoort volgens (2.4) een z-waarde van 3 en een herhalingsgetal J = 20.6. Toepassing van (2.17) levert voor z een betrouwbaarheidsinterval van (2.44, 3.56). Dit komt overeen met een betrouwbaarheidsinterval van (12, 36) voor het herhalingsgetal J.

2.7. Aanpassingscriteria

Indien men de Gumbelverdeling aanpast aan een gegeven steekproef dan zal men ook moeten nagaan of de gegeven steekproef afkomstig kan zijn uit een Gumbelverdeling. In het algemeen gebruikt men hiervoor een aanpassingstoets. De meest bekende aanpassingstoets is de χ^2-toets die echter bij kleine steekproeven onbruikbaar wordt. In het algemeen zullen we bij $N < 50$ de Kuipertoets hanteren en voor $N \geqslant 50$ de χ^2-toets. (zie ook Kruizinga [3]).

Een visuele indruk van de aanpassing van de gegevens aan de Gumbel-verdeling kan verkregen worden door de geordende steekproef $x_1^*, \ldots, x_N^*$ op lineair grafiekenpapier uit te zetten tegen de medianen van de bijbehorende z-waarden,

$$z_{n,med} = -\ln\left(-\ln\left(\frac{n-.3}{N+.4}\right)\right) \tag{2.20}$$

en in deze figuur ook de rechte lijn

$$\hat{x} = z/\hat{\alpha} + \hat{u}$$

op te nemen.

De punten horen dan redelijk bij de lijn aan te sluiten.

3. Hulpmiddelen

Op het KNMI zijn een aantal hulpmiddelen aanwezig die we hier kort zullen bespreken.

3.1. Nomogram (z, $F(z)$, $\Delta_z(z)$, J, R_5, R_{10}, R_{20}, R_{50}, R_{100}).

Zoals uit de voorgaande formules blijkt bestaat er tussen de grootheden z en respectievelijk $F(z)$, $\Delta_z(z)$, $J(z)$, $R_n(z)$ een één-éénduidige relatie. We hebben deze relaties in een nomogram vastgelegd. Dit nomogram is in figuur 1 afgebeeld.

3.2. Wangprogramma's

a. Programma 11.02, Gumbelaanpassing

Het programma berekent na invoer van de steekproefwaarden, de schattingen van $\hat{\alpha}$ en $\hat{u}$ en de Kuiperafstand $*\sqrt{N}$ voor de Kuipertoets.

Op aanvraag wordt daarna een lijst van z-waarden uitgeprint met bijbehorende x-waarden en hun betrouwbaarheidsinterval.

Verder eveneens op aanvraag een lijst met de geordende steekproef met bijbehorende waarde van $z_{med,n}$ (zie 2.20)

b. Programma 12.01, Gumbelformules.

Door het oproepen van bepaalde functies kan men met dit programma een aantal grootheden berekenen. We geven hier een kort overzicht:

Functie	Berekende grootheid	Invoer
00	$F(z)$	z
01	$z(F)$	F
02	$J(z)$	z
03	$z(J)$	J
04	$J^*(z)$	z
05	$z(J^*)$	J^*
06	$R_N(z)$	N, z
07	$z(R_N)$	R_N, N
08	$\Delta_z(z)$	z
09	$\hat{z}$ met betrouwbaarheidsinterval	$\hat{\alpha}, \hat{u}, N, z$
10	$\hat{z}$ en $\hat{J}$ met betrouwbaarheidsintervallen	$\hat{\alpha}, \hat{u}, N, x$

Voor inlichtingen wordt verwezen naar de handleiding.

Opmerking. Voor kleine z-waarden komen de J-waarden zeer dichtbij één te liggen echter ze worden nooit kleiner dan één. Het is logischer om voor deze kleine z-waarden met een J^* te werken. Deze geeft aan dat slechts eens in de J^* jaar een maximum kleiner dan de gegeven z zal optreden.

3.3. Burroughsprogramma Gumbel

Dit programma doet in feite hetzelfde als het Wangprogramma Gumbel-aanpassing. Het levert ook dezelfde gegevens af.

Literatuur.

1. GUMBEL, E.J. (1958) Statistics of Extremes
Columbia University Press, New York

2. SNEYERS, R. et ISACKER, J. VAN
Revue Belge de Statisque, d'Informatique
et de Recherche operationnelle
Vol. 12, No. 1 Mai 1972

3. KRUIZINGA, S. Aanpassingstoetsen
KNMI, W.R.74-7, De Bilt 1974

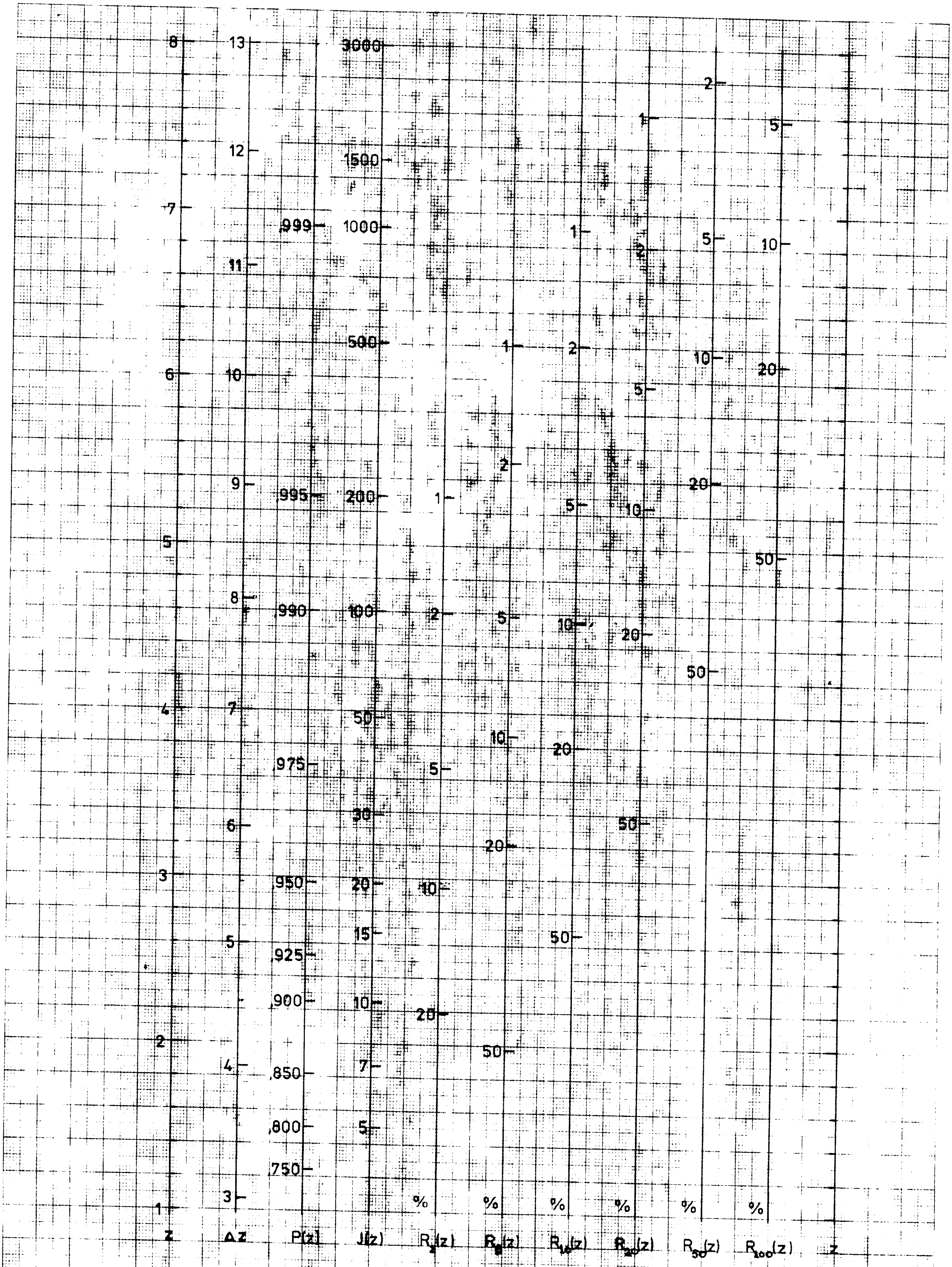

FIG. 1 NOMOGRAM GUMBELVERDELING