

VERGELIJKING VAN VERDELINGEN

*concludering bij dubbel-exponentiële verdelingen met dezelfde
schaalparameter bij multinomiale data*

Martin A.J. van Montfort^{*)}

SAMENVATTING

Een gerieflijke eigenschap van de verdeling van kansvariabelen kan zijn, dat hun verdelingsfuncties in elkaar uit te drukken zijn d.m.v. een exponent. Dit komt voor de dubbel-exponentiële verdeling, ook wel Gumbelverdeling genoemd, neer op gelijkheid in schaalparameter bij eventueel verschil in plaatsparameter. De Gumbelverdeling, die sprekend op de log-Normale verdeling lijkt, kan ook gebruikt worden voor het beschrijven van inkomensverdelingen, die doorgaans gegeven worden in de vorm van frequenties in intervallen (aan te duiden met: multinomiale data).

Bij multinomiale data worden de ML-schatters afgeleid voor de plaats- en schaal-parameter bij één steekproef, en voor de s plaatsparameters en één gemeenschappelijke schaalparameter bij s steekproeven; ook wordt de likelihood ratio toets afgeleid op onderlinge gelijkheid van de schaalparameters.

Een voorbeeld van toepassing wordt gegeven. Tevens wordt aandacht besteed aan de kwaliteit van de schatters van de parameters in afhankelijkheid van het aantal klassen bij tellingen in intervallen in vergelijking tot continue meting.

^{*)} Vakgroep Wiskunde, Landbouwhogeschool, Wageningen.

1. INLEIDING

Bij vergelijkingen van twee kansvariabelen x_1 en x_2 is het gerieflijk dat voor iedere x geldt

$$p_1 = p_2^k \quad (1a)$$

met $p_i = P\{x_i \leq x\}$ en waarbij k (met $k > 0$) niet afhangt van x , maar eventueel wel van de parameters van x_1 en x_2 . Met (1a) is namelijk gelijkwaardig

$$(-\log p_1) = k \cdot (-\log p_2) \quad (1b)$$

evenals

$$\left(-\log (-\log p_1)\right) = -\log k + \left(-\log (-\log p_2)\right) \quad (1c)$$

Formule (1c) geeft (geldt op de kansschaal van waarschijnlijkheidspapier voor de dubbel-exponentiële of Gumbel-verdeling) t.o.v. elkaar verschoven curven in een Gumbel-plot met verschuiving in kansschaal-richting.

Naast model (1a) kan men

$$(1 - p_1) = (1 - p_2)^k \quad (2)$$

beschouwen met vergelijkbare eigenschappen. De exponenten in (1a) en (2) worden indien nodig aangeduid met resp. k_1 en k_2 .

Met geschikt gekozen restricties aan de parameters kan elk paar Gumbel-verdelingen aan (1) voldoen en kan elk paar afkomstig uit de klasse van Weibull- en Pareto-verdelingen aan (2) voldoen, zoals uit het volgende blijkt.

Voor de *Gumbel-verdeling* geldt het volgende: $\phi(x)$, met ϕ toenemend, heeft een Gumbelverdeling met parameters μ_ϕ en σ_ϕ (kortweg: μ en σ) als geldt

$$p = P\{x \leq x\} = P\{\phi(x) \leq \phi(x)\} = \exp\{-e^{-(\phi(x)-\mu)/\sigma}\} \quad (3a)$$

dus

$$-\log p = \exp \left\{ -\frac{\phi(x) - \mu}{\sigma} \right\}. \quad (3b)$$

Uit (1b) en (3b) volgt dan

$$k = \exp \left\{ -\phi(x) \left(\frac{1}{\sigma_1} - \frac{1}{\sigma_2} \right) + \frac{\mu_1}{\sigma_1} - \frac{\mu_2}{\sigma_2} \right\} \quad (4)$$

Nu hangt k niet van x af als $\sigma_1 = \sigma_2 (= \sigma)$; voor $\sigma_1 = \sigma_2$ geldt

$$k = \exp \left\{ \frac{\mu_1 - \mu_2}{\sigma} \right\} \quad (5)$$

Hieruit volgt dat (1) geldt als er een zodanige monotoon toenemende ϕ bestaat, dat de kansvariabelen $\phi(x_1)$ en $\phi(x_2)$ een Gumbelverdeling hebben met dezelfde schaalparameter (evenwijdige rechte lijnen in een Gumbelplot); voor de niet-getransformeerde kansvariabelen vallen in de Gumbelplot na verschuiving in de kansschaal-richting de (eventueel niet rechte) lijnen samen.

De *Weibull-verdeling* is als volgt te definiëren: de kansvariabele x heeft een Weibull-verdeling als er een μ bestaat zō, dat $-\log(x - \mu)$ een Gumbel-verdeling heeft. Daar $-\log(x - \mu)$ een afnemende functie in x is, geldt (2) voor elk paar Weibull-verdelingen met dezelfde vormparameter.

De *Pareto-verdeling* van de tweede soort, zie Johnson en Kotz (1970, p. 234), wordt gedefinieerd door

$$p = P\{x \leq x\} = 1 - \left(\frac{x - \mu}{\sigma} \right)^{-\alpha} \quad \text{met} \quad \alpha > 0, \sigma > 0, x \geq \mu + \sigma \quad (6)$$

(en $p = 0$ voor $x < \mu + \sigma$); dan geldt

$$\frac{-\log(1 - p_1)}{-\log(1 - p_2)} = \frac{\alpha_1 \log \left(\frac{x - \mu_1}{\sigma_1} \right)}{\alpha_2 \log \left(\frac{x - \mu_2}{\sigma_2} \right)}, \quad (7)$$

en voor $\sigma_1 = \sigma_2$ en $\mu_1 = \mu_2$ is dit quotiënt (k_2) gelijk aan α_1/α_2 .

Voor de Pareto-verdeling van de eerste soort, d.w.z. $\mu = 0$, geldt hetzelfde.

Elke 1-parameter familie van verdelingen die aan eis (2) voldoet is, op reparametrisering na en uitgedrukt in verdelingsfuncties, van de vorm

$$F_k(x) = 1 - (1 - G(x))^k \quad \text{voor} \quad -\infty < x < +\infty \quad (8a)$$

met k een positieve parameter en G een verdelingsfunctie, die de familie bepaalt. Door differentiëren van (8a) en met $h(x) = -\log(1 - G(x))$ vindt men als uitdrukking voor de kansdichtheid in x op de drager, d.i. $\{x: G'(x) > 0\}$,

$$f_k(x) = k h'(x) e^{-k \cdot h(x)} \quad (8b)$$

waarin k een positieve parameter is en h op de drager monotoon stijgt van 0 naar $+\infty$, verg. McCullagh (1979). Voor de Pareto-verdeling geldt dan $h(x) = \log(x)$. Deze 1-parameter families kunnen tot meer-parameter families uitgebreid worden door G (of h) te doen afhangen van een of meer parameters. Voor G in (8a) kan men bijv. formule (6) kiezen.

Analoog is elke 1-parameter familie van verdelingen, die aan eis (1) voldoet, op reparametrisering na van de vorm

$$F_k(x) = G^k(x), \quad (9)$$

en met $h(x) = -\log(G(x))$ vindt men voor de kansdichtheid (op een - teken na) weer (8b) maar nu met $h(x)$ monotoon dalend van $+\infty$ naar 0.

Door in (9) voor G in te vullen $\exp(-\exp(-x))$ vindt men

$$F_k(x) = \exp \left[-e^{-(x - \log k)} \right] \quad (10)$$

hetgeen neerkomt op een verandering van de plaatsparameter, ongeacht de schaal waarin x uitgedrukt is.

2. PROBLEEMSTELLING

Bovenvermeld model (1a) kwam ter sprake bij het vergelijken van inkomensverdelingen; in dat geval zijn de gegevens meestal in de vorm van een frequentie-verdeling over inkomensklassen (multinomiale data). Als model voor de verdeling van inkomens wordt vaak de log-Normale verdeling

gekozen, verg. Kloek en Van Dijk (1976, 1977); de Gumbel-verdeling lijkt daar sprekend op, verg. Johnson en Kotz (1970, deel 1, p. 279). De Gumbelverdeling, gebruikt bij levensduur-verdelingen, wordt Gompertz-verdeling genoemd, verg. Johnson and Kotz (1970, deel 1, p. 274; 1970, deel 2, p. 271).

Hier wordt de ML-schatting voor de parameters afgeleid voor multinomiale data en voor data zonder indeling tot klassen (bijlagen 1 en 2). Ook de situatie van verdelingen met gelijke schaalparameter wordt behandeld; de likelihood ratio toets op gelijkheid in schaalparameter volgt direkt bij gebruik van de ML-methode. De asymptotische covariantie-matrices van de schatters bij multinomiale en individuele data worden afgeleid. Ook wordt ingegaan op de vergroting van de variantie van de schatters voor de parameters t.g.v. het indelen in klassen.

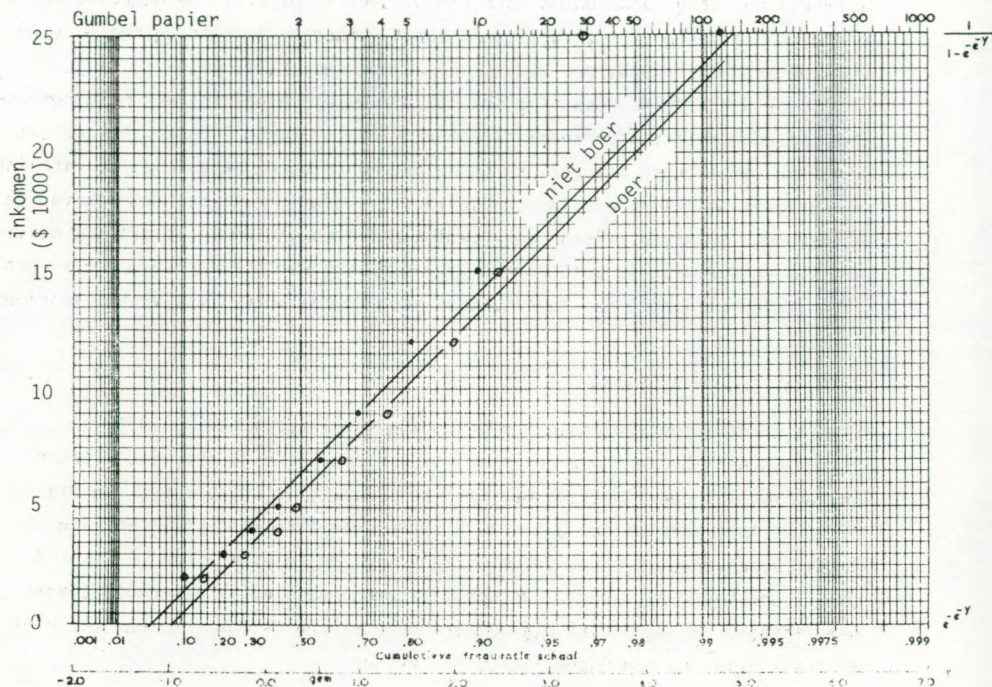
3. SCHATTING EN TOETSING

Als inleidend voorbeeld worden de inkomensverdelingen in 1974 van boeren en niet-boeren onderzocht en wel van negers in de Verenigde Staten in rurale gebieden ('outside metropolitan areas'), zie Table 640 van Statistical Abstracts of the U.S. (1975), welke deels vermeld staat in Tabel 1 dezes. Deze verdelingen worden enerzijds snel en (niet te) ruw (quick and dirty) bekeken (QD); anderzijds wordt als methode die van de grootste aannemelijkheid (ML) gebruikt bij de veronderstelling dat de inkomens (al/niet na transformatie) een Gumbel-verdeling hebben.

Inkomen van VS-negers in 1974 van boeren en niet-boeren in ruraal gebied.

Legenda: $\hat{p}$ is fractie met inkomen kleiner dan grenswaarde;
 $\hat{k}$ is $(-\log(\hat{p}_{\text{boer}})) / (-\log(\hat{p}_{\text{niet-boer}}))$

bovengrens (\$1000)	2	3	4	5	7	9	12	15	25
$\hat{p}$ niet-boer	.099	.217	.313	.415	.573	.691	.808	.895	.992
$\hat{p}$ boer	.144	.289	.409	.483	.641	.763	.868	.919	.965
$\hat{k}$	1.19	1.23	1.30	1.21	1.25	1.37	1.51	1.31	0.23



Figuur 1 Gumbelplot van de data uit Tabel 1 met de bijbehorende ML-geschatte verdelingsfuncties.

Tabel 2 ML-schatting van de parameters van de twee (Box/Cox-getransformeerde) inkomensverdelingen van niet-boeren (1) en boeren (2) uit Tabel 1. Links staan de ML-schattingen van $\mu_1, \sigma_1, \mu_2, \sigma_2$. Naast de ML-schatting van μ_1, μ_2 en $\sigma_1 = \sigma_2 (= \sigma)$ staat rechts ook $k = \exp(\hat{\mu}_1 - \hat{\mu}_2)/\hat{\sigma}$ vermeld (zie formule (5))

λ	transf.	niet-boer		boer		$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}$	k
		$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$				
1	$x - 1$	3.96	4.08	3.17	4.08	3.96	3.17	4.08	1.21
$\frac{1}{2}$	$2\sqrt{x} - 2$	2.23	1.63	1.89	1.65	2.23	1.89	1.64	1.23
0	$\log(x)$	1.39	0.71	1.23	0.73	1.39	1.23	0.72	1.25

Bij QD schat men bij de grenzen van de inkomensintervallen de parameter k in samenhang met (1b); iedere grens geeft een k -schatting, en deze k -schattingen middelt men op een of andere manier. In Tabel 1 staan de gegevens en de k -schattingen; de mediaan van de (gecorrleerde) k -schattingen is 1.25. Naar schatting geldt dan voor ieder inkomen

$$p_{\text{niet-boer}} = p_{\text{boer}}^{1.25}$$

De gegevens staan tevens uitgebeeld in een Gumbelplot, zie figuur 1.

Bij de methode van de grootste aannemelijkheid (ML) is enig differentiatieën nodig. Bijlage 1 bevat, naast de ML-schattingen voor de plaats- en schaalparameter bij één steekproef, ook de ML-schatting voor de plaatsparameter en de gemeenschappelijke schaalparameter bij meer dan één steekproef.

Als men als mogelijke transformatie de Box/Cox-transformatie kiest, wordt een resultaat gevonden, dat in Tabel 2 vermeld staat. De Box/Cox-transformatie voor positieve data met als transformatie-parameter λ luidt: $(x^\lambda - 1)/\lambda$ voor $\lambda \neq 0$ en $\log(x)$ voor $\lambda = 0$. Voor de parameter λ werden enige waarden gekozen, nl. 0, $\frac{1}{2}$ en 1. Opvallend is de overeenkomst van de k -schattingen bij de verschillende methodes; de k -schattingen behoren bij ML-schattingen bij $\sigma_1 = \sigma_2$.

Uit de log-likelihood van de gecombineerde steekproeven en die van de afzonderlijke steekproeven volgt dan de likelihood ratio test op gelijkheid in schaalparameter. Tevens geeft bijlage 1 de asymptotische covariantie-matrix van de schatters van de parameters, waarmee lineaire combinaties van de plaatsparameters beoordeeld kunnen worden.

4. EEN VOORBEELD VAN TOEPASSING

Tabel 3 bevat de verdeling van gezinsinkomens in de V.S. in 1972, 1973 en 1974 in de drie regio's NE (Northeast), NC (North Central) en W (West), en is ontleend aan Table 638 van het Statistical Abstract of the U.S. (1975). Tabel 3 bevat tevens parameterschattingen. De uitkomsten suggereren een verschil in plaatsparameter van NE en NC versus W in ieder van de jaren 1972, 1973 en 1974 bij per jaar nagenoeg constante schaalparameter.

Tabel 3. Verdeling (in ‰) van het gezinsinkomen in de V.S. in de jaren 1972, 1973 en 1974 in de regio's NE (Northeast), NC (North Central) en W (West) en enige aanvullende statistische grootheden. De steekproefomvang wordt genoteerd als N

	inkomensinterval (\$ 1000)							max $ F_N - \hat{F} $	regio		tezamen	Pearson's χ^2 (N = 1000)
	< 3	3-5	5-7	7-10	10-12	12-15	>15		$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	
1972 NE	32	71	75	126	96	150	450	.015	10.24	7.12	10.28	14.36
NC	44	64	74	127	99	157	435	.010	10.01	7.19	10.01	5.72
W	49	71	83	135	89	145	429	.013	9.69	7.27	9.66	7.21
											$\hat{\sigma}=7.19$	27.41
1973 NE	37	72	75	123	92	147	454	.016	10.23	7.23	10.22	13.91
NC	39	63	69	123	100	153	452	.010	10.32	7.28	10.33	6.72
W	44	72	84	129	93	139	439	.017	9.86	7.30	9.86	10.05
											$\hat{\sigma}=7.30$	30.70
1974 NE	32	73	76	131	100	145	442	.015	10.10	7.02	10.11	14.34
NC	39	63	79	127	102	155	435	.011	10.04	7.05	10.05	6.00
W	48	72	94	141	96	129	421	.017	9.47	7.11	9.45	9.33
											$\hat{\sigma}=7.06$	29.71

Tabel 4 Invloed van indelen in klassen op de asymptotische covariantie-matrix van de ML-schatters van de parameters bij steekproefomvang N; zie tekst.

aantal klassen	N . var($\hat{\mu}$)		N . var($\hat{\sigma}$)		$\rho_{\hat{\mu}, \hat{\sigma}}$	N ² . determinant	
	abs.	rel.	abs.	rel.		abs.	rel.
4	1.3603	1.23	1.5349	2.52	-0.001	2.088	3.43
6	1.2330	1.11	1.0782	1.77	0.117	1.311	2.16
12	1.1526	1.04	0.8068	1.33	0.217	0.886	1.46
20	1.1306	1.02	0.7244	1.19	0.255	0.766	1.26
∞	1.1087	1	0.6079	1	0.313	0.608	1

Als de steekproefomvang voor ieder der verdelingen op ($N=$) 1000 gesteld wordt (deze omvang is niet aan Table 638 te ontleen) vindt men voor ieder van de verdelingen ongeveer $\text{var}(\hat{\mu}) = 0.26^2$, $\text{var}(\hat{\sigma}) = 0.26^2$ en 0.40 als correlatiecoëfficiënt voor $\hat{\mu}$ en $\hat{\sigma}$.

Het grootste absolute verschil tussen aangepaste en waargenomen verdelingsfunctie is circa 0.015; inspectie van deze verschillen toont opwaartse kromming in de Gumbelplot (dus empirisch meer hoge inkomens dan volgens het model verwacht wordt).

Simultaan schatten van de plaatsparameters en de gemeenschappelijke parameter levert ongeveer $\text{var}(\hat{\mu}) = 0.25^2$, $\text{var}(\hat{\sigma}) = 0.15^2$ ($\approx 0.26^2/3$) met correlatiecoëfficiënt +0.26 (de correlatiecoëfficiënt van de μ -schatters is circa +0.06). Voor 1973 bijvoorbeeld geldt naar schatting

$$1000 \cdot \text{cov} \begin{pmatrix} \hat{\mu}_{NE}, \hat{\mu}_{NC}, \hat{\mu}_W, \hat{\sigma} \end{pmatrix}' = \left(\begin{array}{ccc|c} 61 & 4 & 4 & 9 \\ 4 & 61 & 4 & 9 \\ 4 & 4 & 61 & 9 \\ \hline 9 & 9 & 9 & 22 \end{array} \right).$$

Hieruit volgt in samenhang met Tabel 3

$$\frac{\hat{\mu}_{NE} + \hat{\mu}_{NC} - 2\hat{\mu}_W}{\sqrt{\hat{\text{var}}(\hat{\mu}_{NE} + \hat{\mu}_{NC} - 2\hat{\mu}_W)}} = \frac{+0.83}{\sqrt{0.342}} = +1.42;$$

dit bedrag is nagenoeg constant over de jaren 1972 t/m 1974, hetgeen een systematisch verschil tussen $N(\text{NE}, \text{NC})$ en W suggereert. Dit systematische verschil is te herleiden tot de exponent k met de waarde $\exp(((10.22 + 10.33)/2 - 9.86)/7.30) = 1.06$; derhalve $p_N \approx p_W^{1.06}$.

5. OPMERKINGEN

Een waarschuwende opmerking bij het kiezen van een type inkomensverdeling is de volgende. Men zou de neiging kunnen hebben te denken dat bij toenemend inkomen de dichtheid aanvankelijk toeneemt en vervolgens afneemt. Beschouw nu de inkomensverdeling van het weinig mobiele personeel van een instelling voor hoger onderwijs; de verdeling is veeltoppig, want bij de maxima van de verschillende schalen zit ieder die een bepaalde opleiding heeft en een zekere ambtelijke leeftijd overschreden heeft; tussen

de schaalmaxima zitten slechts degenen die nog even wachten moeten. Het aanpassen van een ééntoppige verdeling is dan een twijfelachtige zaak.

Een steeds terugkerend probleem is de keuze van het type verdeling in samenhang met het aantal parameters: weinig parameters (bijv. voor plaats en schaal), interpreteerbare parameters en slechte aanpassing versus veel, niet-interpreteerbare parameters bij (te) goede aanpassing. Kloek and Van Dijk (1976, 1977) bijv. voegen een derde parameter toe door van log-Gauss over te gaan op log-Student (met een positief reëel getal ν voor Students t_ν) ter bijsturing van de staartdikte. Zulks is ook mogelijk bij de Gumbelverdeling, nl. overgaan op de GEV (general extreme value distribution) met plaats-, schaal- en vormparameter; voor de ML-schatting bij individuele data wordt verwezen naar Otten and Van Montfort (1980).

Voor het bestuderen van de invloed van het aantal klassen op de verdeling van de ML-schattingen van de parameters van de Gumbelverdeling zijn de asymptotische covariantie-matrices van belang. Als model worden klassen gekozen ieder met kans $1/(\text{aantal klassen})$. Bijlage 1 geeft de asymptotische covariantie-matrix bij wel indelen tot klassen, en bijlage 2 bij niet indelen tot klassen. Bijlage 2 behandelt nl. de ML-schatting bij niet indelen tot klassen. Tabel 4 bevat elementen van de matrix bij wel-indelen, zowel relatief (t.o.v. niet-indelen) als absoluut. Men ziet dat indelen in klassen veel meer verlies aan informatie over de schaalparameter oplevert dan over de plaatsparameter.

LITERATUUR

- Johnson, N.L. and S. Kotz (1970). Distributions in statistics: Continuous univariate distributions-1,2. Houghton Mifflin Company, Boston.
- Kloek, T. and H.K. van Dijk (1976). Efficient estimation of income distribution parameters. Report 7616/E, Erasmus Universiteit Rotterdam.
- Kloek, T. and H.K. van Dijk (1977). Further results on efficient estimation of income distribution parameters. *Economie Appliquée*, Tome XXX, 1-21.
- McCullagh, P. (1979). The use of the logistic function in the analysis of ordinal data. *Invited Paper 17.2 at the 42nd meeting of the I.S.I.*
- Otten, A. and M.A.J. van Montfort (1980). Maximum-likelihood estimation of the general extreme value distribution parameters. *Journal of the Royal Statistical Society B*, 42, 1-19.

Bijlage 1

ML-SCHATTING VAN DE PARAMETERS VAN DE GUMBELVERDELING BIJ MULTINOMIALE DATA

Notatie en likeliheid

Laat n_k de frequentie zijn in klasse k ($k = 1, \dots, m+1$) met ondergrens g_{k-1} en bovengrens g_k ; definieer $g_0 = -\infty$, $g_{m+1} = +\infty$, $N = \sum_{k=1}^{m+1} n_k$

In figuur:

klasse k met	$k = 1$	2	$\dots$	m	$m+1$	
frequentie	n_1	n_2	$\dots$	n_m	n_{m+1}	met som: N
grens	$g_0 = -\infty$	g_1	g_2	g_{m-1}	g_m	$g_{m+1} = +\infty$

Voor de Gumbelverdeling geldt

$$p = P\{\underline{x} \leq x\} = \exp(-e^{-(x-\mu)/\sigma}), \quad -\infty < x, \mu < +\infty; \sigma > 0.$$

Met

$$F_k = P\{\underline{x} \leq g_k\}$$

(waaruit volgt $F_0 = 0$ en $F_{m+1} = 1$) wordt de log likeliheid op een niet relevante constante na (bij stochastisch onafhankelijke waarnemingen):

$$L = \log \prod_{k=1}^{m+1} (F_k - F_{k-1})^{n_k} = \sum_{k=1}^{m+1} n_k \log(F_k - F_{k-1}).$$

L moet met betrekking tot de plaatsparameter μ en de schaalparameter σ (met $\sigma > 0$) gemaximaliseerd worden bij gegeven $m, g_1, \dots, g_m, n_1, \dots, n_{m+1}$.

Beginschattingen

Beginschattingen kunnen worden verkregen door middel van een zodanige parameterkeuze in de verdelingsfunctie, dat deze in twee punten overeenstemt met de empirische verdelingsfunctie. Voor de inverse van de Gumbelverdelingsfunctie geldt

$$x = \mu + \sigma (-\log(-\log p))$$

Voor twee x -waarden wordt p hierin vervangen door de fractie ($\hat{p}$) van kleinere waarden in de steekproef. Deze twee fracties kan men rond 0.25 en 0.75 kiezen.

Afgeleiden van L met betrekking tot de parameters μ en σ

De index ij duidt op de i^e afgeleide naar μ en de j^e afgeleide naar σ .

$$y = (x - \mu)/\sigma$$

$$y_{10} = -1/\sigma; y_{01} = \frac{-(x - \mu)}{\sigma^2} = \frac{-y}{\sigma}; y_{20} = 0; y_{02} = \frac{2(x - \mu)}{\sigma^3} = \frac{2y}{\sigma^2}; y_{11} = \frac{1}{\sigma^2}$$

$$F = \exp(-e^{-y})$$

$$F_{10} = y_{10} e^{-y} \cdot F$$

$$F_{20} = y_{20} e^{-y} \cdot F + y_{10} \left(-y_{10} e^{-y} \cdot F + F_{10} e^{-y} \right) = e^{-y} \cdot F \left\{ y_{20} + y_{10}^2 (e^{-y} - 1) \right\}$$

$$F_{01} = y_{01} e^{-y} \cdot F$$

$$F_{02} = y_{02} e^{-y} \cdot F + y_{01} \left(-y_{01} e^{-y} \cdot F + e^{-y} F_{01} \right) = e^{-y} \cdot F \left\{ y_{02} + y_{01}^2 (e^{-y} - 1) \right\}$$

$$F_{11} = y_{11} e^{-y} \cdot F + y_{10} \left(-y_{01} e^{-y} \cdot F + e^{-y} y_{01} e^{-y} F \right) = e^{-y} \cdot F \left\{ y_{11} + y_{10} y_{01} (e^{-y} - 1) \right\}$$

Voor F_0 en F_{m+1} zijn alle afgeleiden nul wegens $F_0 = 0$ en $F_{m+1} = 1$.

Met $\Delta_{ijk} = F_{ij}(g_k) - F_{ij}(g_{k-1})$ en $\Delta_{00k} = F_k - F_{k-1}$ vindt men

$$L_{10} = \sum_k \Delta_{10k} / \Delta_{00k}$$

$$L_{01} = \sum_k \Delta_{01k} / \Delta_{00k}$$

$$L_{20} = \sum_k (\Delta_{20k} \Delta_{00k} - \Delta_{10k}^2) / \Delta_{00k}^2$$

$$L_{02} = \sum_k (\Delta_{02k} \Delta_{00k} - \Delta_{01k}^2) / \Delta_{00k}^2$$

$$L_{11} = \sum_k (\Delta_{11k} \Delta_{00k} - \Delta_{10k} \Delta_{01k}) / \Delta_{00k}^2$$

waarbij $\sum$ is $\sum_{k=1}^{m+1}$.

Het iteratieschema

Met $\Theta = (\mu, \sigma)'$ en ℓ het volgnummer van de iteratieronde ($\ell = 0, 1, 2, \dots$) ontstaat de rij schattingen Θ_ℓ met Θ_∞ de ML-schatting voor Θ .

Met

$$C = \left(- \begin{pmatrix} L_{20} & L_{11} \\ L_{11} & L_{02} \end{pmatrix} \right)^{-1} = \frac{1}{L_{20} L_{02} - L_{11}^2} \begin{pmatrix} -L_{02} & L_{11} \\ L_{11} & -L_{20} \end{pmatrix}$$

wordt het iteratie schema

$$\Theta_{\ell+1} = \Theta_\ell + C_\ell \begin{pmatrix} L_{10} \\ L_{01} \end{pmatrix}_\ell$$

C_∞ is dan de geschatte asymptotische covariantie-matrix.

Schatting bij twee of meer verdelingen met zelfde schaalparameter

Als 2 verdelingen dezelfde schaalparameter hebben ($\sigma_1 = \sigma_2 = \sigma$) en de parameters μ_1, μ_2, σ geschat moeten worden, volgen de ML-schattingen uit

$L = L^{(1)} + L^{(2)}$ met L de log likelijkheid van het tweetal onafhankelijke steekproeven. Merk op dat $L^{(1)}$ geen functie van μ_2 en $L^{(2)}$ geen functie van μ_1 is. Voor het iteratie-schema volgt daaruit

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \sigma \end{pmatrix}_{\ell+1} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \sigma \end{pmatrix}_\ell + \left(- \left[\begin{array}{cc|cc} L_{20}^{(1)} & 0 & L_{11}^{(1)} & 0 \\ 0 & L_{20}^{(2)} & L_{11}^{(2)} & 0 \\ \hline L_{11}^{(1)} & L_{11}^{(2)} & L_{02}^{(1)} + L_{02}^{(2)} & 0 \end{array} \right]_\ell \right)^{-1} \cdot \begin{pmatrix} L_{10}^{(1)} \\ L_{10}^{(2)} \\ L_{01}^{(1)} + L_{01}^{(2)} \end{pmatrix}_\ell$$

Met behulp van de maximale waarden van L , $L^{(1)}$ en $L^{(2)}$ is een likelihood ratio toets op $\sigma_1 = \sigma_2$ uit te voeren. De statistische concludering bij meer dan 2 steekproeven verloopt analoog. Partitioneren van de matrix $-L''$ bij s steekproeven in links boven-eendiagonale $s \times s$ -matrix etc., is bij het interpreteren nuttig. Met de inverse kunnen dan weer lineaire combinaties van de plaatsparameters bestudeerd worden.

ML-SCHATTING VAN DE PARAMETERS VAN DE GUMBELVERDELING ZONDER INDELING TOT KLASSEN

Afleiding van de asymptotische covariantie-matrix

$$F(x) = P\{\underline{x} \leq x\} = \exp \left\{ -e^{-(x-\mu)/\sigma} \right\} \quad (-\infty < x, \mu < +\infty, 0 < \sigma < +\infty)$$

$$f(x) = \frac{1}{\sigma} \exp(-y - e^{-y}) \quad \text{met} \quad y = (x - \mu)/\sigma$$

$$L = \log \prod_{i=1}^n f(x_i) = -n \log \sigma - \Sigma(y + e^{-y})$$

$$y_{10} = \frac{-1}{\sigma}, \quad y_{01} = \frac{-(x - \mu)}{\sigma^2} = \frac{-y}{\sigma}$$

$$L_{10} = -\Sigma(y_{10} - y_{10} e^{-y}) = \frac{1}{\sigma} \Sigma(1 - e^{-y})$$

$$L_{01} = \frac{-n}{\sigma} - \Sigma(y_{01} - y_{01} e^{-y}) = \frac{1}{\sigma} \Sigma(-1 + y - ye^{-y})$$

Uit $L_{10} = L_{01} = 0$ volgen de ML-schattingen, zie beneden. Voor de asymptotische covariantie-matrix is het volgende nodig.

$$L_{20} = \frac{1}{\sigma} \Sigma y_{10} e^{-y} = \frac{-1}{\sigma^2} \Sigma e^{-y}$$

$$\begin{aligned} L_{02} &= \frac{-1}{\sigma^2} \Sigma(-1 + y - ye^{-y}) + \frac{1}{\sigma} \Sigma(y_{01} - y_{01} e^{-y} + y y_{01} e^{-y}) = \\ &= \frac{-1}{\sigma^2} \Sigma(-1 + 2y - 2ye^{-y} + y^2 e^{-y}) \end{aligned}$$

$$L_{11} = \frac{-1}{\sigma^2} \Sigma(1 - e^{-y}) + \frac{1}{\sigma} \Sigma y_{01} e^{-y} = \frac{-1}{\sigma^2} \Sigma(1 - e^{-y} + ye^{-y})$$

Met behulp van

$$\mathbb{E}y = \gamma$$

$$\mathbb{E}(e^{-y}) = 1 \quad \text{wegens} \quad e^{-y} \quad \text{standaard exponentieel verdeeld}$$

$$\mathbb{E}(ye^{-y}) = -\Gamma'(2) = \gamma - 1$$

$$\mathbb{E}(y^2 e^{-y}) = -\Gamma''(2) = (1 - \gamma)^2 + \frac{\pi^2}{6} - 1$$

vindt men

$$-\mathbb{E} L_{20} = \frac{n}{\sigma^2}$$

$$-\mathbb{E} L_{11} = \frac{n}{\sigma^2} (\gamma - 1)$$

$$-\mathbb{E} L_{02} = \frac{n}{\sigma^2} \left\{ \frac{\pi^2}{6} + (1 - \gamma)^2 \right\}$$

Hieruit volgt de asymptotische covariantie-matrix als inverse van de 2x2-matrix $-\mathbb{E} L''$:

$$\frac{6\sigma^2}{n\pi^2} \begin{pmatrix} \frac{\pi^2}{6} + (1 - \gamma)^2 & 1 - \gamma \\ 1 - \gamma & 1 \end{pmatrix}$$

met als correlatie-coëfficiënt van de schatters $\left\{ 1 + \frac{\pi^2}{6(1 - \gamma)^2} \right\}^{-\frac{1}{2}} = +0.31313 \dots$

Oplossen van $L_{01} = L_{01} = 0$

$$\text{Uit } L_{10} = 0 \text{ volgt } \mu = \sigma \log(n/S_0) \text{ met } S_k = \sum x_i^k \exp(-x_i/\sigma) .$$

Daarmee vindt men

$$L_{01} = \frac{-n}{\sigma} (\sigma - \bar{x} + S_1/S_0)$$

en de oplossing voor σ van $\phi(\sigma) = \sigma - \bar{x} + S_1/S_2 = 0$ moet gevonden worden.

Voor de Newton-Raphson methode is nodig

$$\phi'(\sigma) = 1 + \frac{S_0 S_2 - S_1^2}{\sigma^2 \cdot S_0^2}$$

Iteratieve correctie volgens

$$\sigma_{\ell+1} = \sigma_{\ell} - \frac{\phi(\sigma_{\ell})}{\phi'(\sigma_{\ell})} \quad (\ell = 0, 1, 2, \dots)$$

levert σ_{∞} als ML-schatting voor σ . Als beginschatting kan men op grond van $\text{var}(\underline{x}) = \pi^2 \sigma^2 / 6$ kiezen

$$\sigma_0 = \left\{ 6 \left(\sum x_i^2 / n - (\sum x_i)^2 / n^2 \right) \right\}^{\frac{1}{2}} / \pi .$$

De ML-schatting voor μ is dan $\sigma_{\infty} \log(n/S_0)$ wegens $L_{10} = 0$.