



Technische Universiteit Delft
Faculteit Elektrotechniek, Wiskunde en Informatica
Delft Institute of Applied Mathematics

De supremumfout van de Grenander schatter
(Engelse titel: The supremum error of the
Grenander estimator)

Verslag ten behoeve van het
Delft Institute of Applied Mathematics
als onderdeel ter verkrijging

van de graad van

BACHELOR OF SCIENCE
in
TECHNISCHE WISKUNDE

door

Said el Bouhaddani

Delft, Nederland
Augustus 2012



BSc verslag TECHNISCHE WISKUNDE

“De supremumfout van de Grenander schatter” (Engelse titel: “The supremum error of the Grenander estimator”)

SAID EL BOUHADDANI

Technische Universiteit Delft

Begeleider

Dr. H. P. Lopuhaä

Overige commissieleden

Dr.ing. D. Jeltsema

Dr. J. G. Spandaw

Augustus, 2012

Delft

Inhoudsopgave

1	Probleemstelling	2
2	De methode van Maximum Likelihood	4
2.1	Inleiding in Maximum Likelihood	4
2.2	Parametrisch schatten	4
2.3	Niet-parametrisch schatten	5
3	De Grenander schatter	7
3.1	Inleiding	7
3.2	Uitrekenen	7
3.3	Consistentie van de Grenander schatter	10
4	De supremumafstand	13
4.1	De limietverdeling van het supremum	13
4.2	Betrouwbaarheidsband	14
5	Het experiment	16
5.1	Inleiding	16
5.2	De afgeknotte exponentiele verdeling	16
5.3	De Grenander schatter	18
5.4	De supremumafstand	19
5.5	Resultaten	24
6	Conclusie	30
A	Functie code	32
B	Uitvoer code	38

1 Probleemstelling

In de statistiek is het probleem van de verdeling van een dataset schatten een fundamenteel probleem. Er zijn hiervoor verschillende methoden bedacht die hiermee proberen om te gaan. We kunnen ons beperken tot een klasse van onderliggende verdelingen gekarakteriseerd door een of meerdere parameters, en dan deze parameters schatten. Aan de andere kant kunnen we aannames over de verdeling links laten liggen, en andere intuïtieve schattingsmethoden toepassen. Een tussenweg kan zijn om alleen aannames te doen over de structuur van de verdeling. Denk hierbij aan beperkingen op de vorm van de kansdichtheid. Dit is wat beperkter dan helemaal geen aannames doen, maar toch wel veel vrijer dan onze toevlucht te zoeken alleen in één klasse van verdelingen. Voor ons is deze tussenweg van belang.

Stel we beschikken over een dataset van n punten. We nemen aan dat deze realisaties zijn uit een onbekende verdeling met een dalende dichtheid f . Er bestaat dan een maximum likelihoodschatter voor dit niet-triviale probleem. Grenander liet in 1956 zien hoe we een schatter kunnen maken voor deze f , en dat deze de likelihood maximaliseert, zie [5]. Deze schatter wordt, ter ere van de uitvinder, de Grenander schatter genoemd. De eerste resultaten over convergentie van de Grenander schatter zijn gepubliceerd in 1969 door P. Rao in [12]. Deze behandelde vooral puntsgewijze convergentie van de Grenander schatter. De jaren daarna is er onderzoek gedaan naar de normconvergentie van de Grenander schatter. In 1999 werd door Groeneboom et al [7] voor de limietverdeling van de L_1 -afstand een resultaat afgeleid. In 2005 en 2008 werd dit gedaan voor L_p , zie [11] resp. [3]. Voor informatie over L_p -afstanden, zie [14]. In 2011 hadden Durot, Kulikov en Lopuhaä een artikel geschreven, zie [4], met daarin een aantal zeer interessante eigenschappen van de Grenander schatter. De supremum fout van de schatter convergeert namelijk, mits herschaald, naar een bekende verdeling. Het blijkt dat deze verdeling de Gumbel verdeling is. Met deze resultaten kunnen we een betrouwbaarheidsband construeren, die aangeeft hoe betrouwbaar onze schatter is. De betrouwbaarheid zal niet exact zijn, in de praktijk hebben we namelijk maar eindig veel datapunten. We weten nog niets over de toepasbaarheid van de stelling in het geval van een eindige dataset. Daarom zullen we de resultaten verifiëren met een gekozen dalende dichtheid.

We behandelen ten eerste hoe we een dalende dichtheid moeten schatten, met behulp van maximum likelihood. We gebruiken de resultaten van Grenander, zie [5], waarbij we eerst de empirische verdelingsfunctie bepalen. Dan construeren we de kleinste concave majorant. De linker afgeleide van deze functie is de Grenander schatter. En zijn stelling is dat deze de maximum likelihood schatter is van de dalende dichtheid f . Een kort bewijs gedaan door Groeneboom en Lopuhaä in [8] zal geleverd worden.

Ten tweede willen we weten hoe betrouwbaar onze gemaakte schatting is. We kunnen een betrouwbaarheidsband maken, die aangeeft hoe nauwkeurig onze schatter is. Dit is een asymptotische betrouwbaarheidsband, omdat de supremumafstand een Gumbel verdeling heeft in de limiet.

Het doel van dit verslag is het nagaan in hoeverre de stelling klopt in het eindige geval. Voor een vaste n simuleren we dus uit onze dalende dichtheid, en

bekijken in hoeverre de stelling inderdaad klopt. We gebruiken hierbij hulpmiddelen die ons een indicatie geven van convergentiesnelheid in n naar de Gumbelverdeling en de kwaliteit van de benadering van de Gumbelverdeling voor vaste n . Dit zal interessant zijn voor diegenen die deze stelling willen toepassen op een dataset uit de praktijk.

2 De methode van Maximum Likelihood

2.1 Inleiding in Maximum Likelihood

De methode van Maximum Likelihood werd veel gebruikt en bekend gemaakt door R. A. Fisher (1890-1962) tussen 1912 en 1922, een Engelsman die de grondlegger voor moderne statistiek wordt genoemd. Hij heeft deze methode niet uitgevonden. Onder andere C. F. Gauss in 1809, en F. Y. Edgeworth in 1908, hadden al een basis gemaakt voor Fisher's ideeën. Zie het artikel van J. Aldrich, [1], of A. Hald in [9] voor meer informatie hierover.

De Maximum Likelihood methode werkt vrij intuïtief. We willen, gegeven een dataset met een aantal getallen, de schatting voor de oorspronkelijke verdeling hebben die de *kans op de data*¹ zo groot mogelijk maakt. Omdat we vaak alleen een aantal getalletjes zien, moeten we wat aannames maken over deze getallen. We kunnen twee kanten op met deze aannames. Namelijk dat we aannemen dat deze getallen trekkingen zijn uit een expliciet gemaakte verdeling, die weliswaar nog afhangt van een of meer reëelwaardige onbekende parameters. Dit wordt ook wel parametrisch schatten genoemd, vanwege de aanname dat we alleen de parameter hoeven te vinden.

Anderzijds kunnen we de veel zwakkere aanname maken dat we alleen de structuur van de dichtheid of verdeling weten. Denk aan de aanname dat de getallen uit een onbekende *dalende* kansdichtheid komen. Dit is dan niet-parametrisch schatten. Deze aanname is zwakker omdat we ons niet beperken tot een expliciet geparametriseerde verdelingsfunctie.

In het verdere verloop van dit project gaan we er vanuit dat de getallen realisaties zijn van onafhankelijk en gelijk verdeelde stochasten (iid stochasten²). Dit vergemakkelijkt het daadwerkelijk uitrekenen van de maximum likelihood schatter. Er zijn ook manieren om met afhankelijkheid om te gaan, maar daar gaan we niet op in.

2.2 Parametrisch schatten

We hebben een aantal getallen, zeg $x_1, \dots, x_n$. Stel we willen een parametrische schatting maken van de dichtheid, en we nemen aan dat de getallen uit een dichtheid f komen, die nog afhangt van een parameter θ . Dus $f(x|\theta)$ is expliciet gegeven. We kunnen de likelihood (waarschijnlijkheid) opschrijven, deze is namelijk simpelweg de gezamenlijke dichtheid berekend in de data. Omdat we iid data hebben kunnen we deze uitfactoren in een product van dezelfde dichtheid:

$$l(\theta) = \prod_{i=1}^n f(x_i|\theta).$$

We zeggen dat $l(\theta)$ de likelihood van de data is. Deze hangt nog af van de parameter θ . Als de schatting voor deze parameter niet redelijk overeenkomen met de data, is de kans klein dat we deze data trekken. Dit betekent een kleine likelihood. Een concreet geval is een normale verdeling met parameter μ , en data waarvan het gemiddelde 3 is. Als $\mu = 1000$ dan zal het onwaarschijnlijk zijn dat we data trekken met gemiddelde 3. De dichtheid gegeven dat $\mu = 1000$ is dan

¹Ook wel de waarschijnlijkheid of likelihood van de data genoemd.

²iid=independent and identically distributed.

klein en de likelihood ook. We moeten dus de θ hebben die l maximaliseert. In formulevorm:

$$\hat{\theta}_n := \arg \max_{\theta \in D} l(\theta).$$

We zoeken meestal een getal θ in de reële getallen ($D = \mathbb{R}$), maar een vector van getallen is evengoed mogelijk, ($D = \mathbb{R}^k$). Denk aan de normale verdeling met $\vec{\theta} = (\mu, \sigma)$, $D = \mathbb{R} \times [0, \infty)$. Een eindige set van mogelijke θ komt tenslotte ook voor.

2.3 Niet-parametrisch schatten

Dit is eenzelfde verhaal als parametrisch, alleen nu hebben we minder aannames. We willen een eis op de structuur leggen van de dichtheid, dat wil zeggen dat we zoeken in een klasse van functies die aan deze structuur voldoen. We geven de likelihood, deze is weer

$$l(f) = \prod_{i=1}^n f(x_i).$$

Deze hangt dus af van de hele functie, in plaats van alleen een eindig dimensionale parameter. Dit is natuurlijk logisch, omdat we niet over een parameter van een functie maximaliseren, maar over een klasse van algemene functies.

We willen dus over een klasse van functies, zeg $\mathcal{F}$, maximaliseren, en kiezen die functie die de likelihood maximaliseert. In formulevorm

$$\hat{f}_n := \arg \max_{f \in \mathcal{F}} l(f).$$

De klasse waarin je zoekt kan van alles zijn, en hangt geheel af van wat je aanneemt. Een voorbeeld is de klasse van alle dalende functies op $[0, \infty)$. Deze komt vaak voor in vraagstukken waar levensduur een grote rol speelt. Denk aan verzekeringsmaatschappijen of fabrieken van apparaten met een beperkte levensduur. Het is in het algemeen veel moeilijker om de schatter te vinden die de likelihood maximaliseert in het niet-parametrische geval. Dit komt omdat we de dichtheid in een klein interval drastisch kunnen veranderen, zonder dat de cumulatieve verdelingsfunctie zelf erg verandert. De dichtheid geeft namelijk de helling aan in een punt, en integreren over een klein interval rond een punt met veel grotere helling geeft nagenoeg dezelfde integraal. We hebben dan ook een langzamere convergentie van de niet-parametrische schatter naar de echte dichtheid. De convergentiesnelheid die we hebben bij een niet-parametrisch schatten blijkt met $\sqrt[3]{n}$ te gaan. Dit in tegenstelling tot een convergentiesnelheid van $\sqrt{n}$ bij parametrisch schatten. Dit is de prijs die we betalen om zwakkere restricties te leggen op ons model.

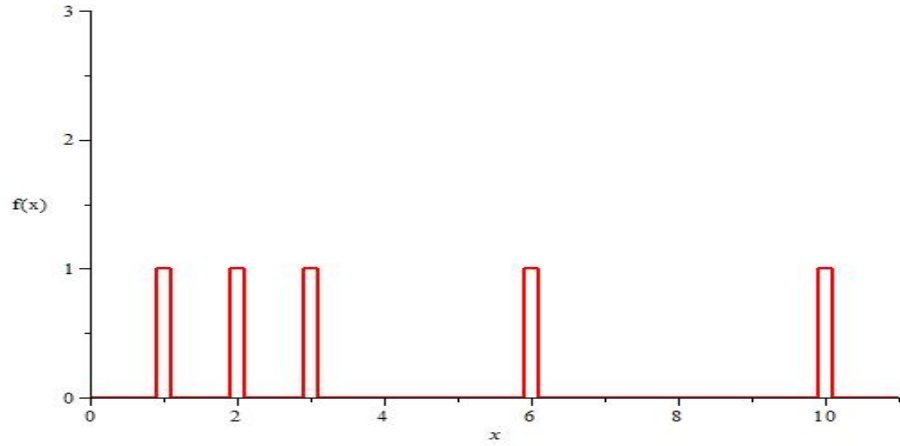
In het meest algemene probleem, maximaliseer over alle dichtheden, bestaat er geen maximum likelihood schatter. Intuïtief is het duidelijk, kies een dichtheid die ‘oneindig’ is in de datapunten. Precies gezegd:

Laat $X_i, i = 1 \dots n$, n trekkingen uit een verdeling.

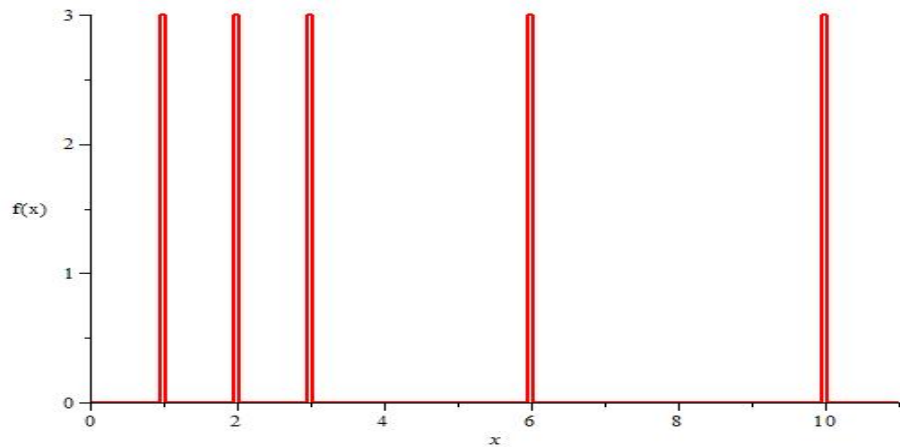
Er bestaat dan er een rij dichtheden $(f_k)_{k=1}^\infty$ die divergeert:

$$f_k(x) = \frac{k}{2n} \quad \text{als } \exists i \in \{1, 2, \dots, n\} : x \in \left[X_i - \frac{1}{k}, X_i + \frac{1}{k} \right],$$

en 0 elders. De functie is stuksgewijs constant met n intervallen waar deze niet 0 is. Hier is de functiewaarde gelijk aan $k/(2n)$, en de intervalbreedte is $2k$. In figuur 1 zien we een concreet voorbeeld, met datapunten genomen op 1, 2, 3, 6, 10. We moeten even nagaan dat f_k , bij vaste k , een dichtheid is. Dit is gemakkelijk in te zien. De oppervlakte onder f_k is $n \cdot (2k \cdot 2k/n) = 1$ (breedte maal hoogte). Bovendien is $f_k(x) \geq 0$ voor alle x .



(a)



(b)

Figuur 1: eerst $k = 10$, dan $k = 30$.

De likelihood van deze dichtheid divergeert naar ∞ als k naar ∞ gaat. Dus een maximum bestaat inderdaad niet.

We zien dus dat een niet-parametrisch probleem niet altijd een oplossing heeft. Als we maximaliseren over *dalende* dichtheden, kunnen we wel een maximum opschrijven! De manier waarop we dit zullen doen, zien we in Hoofdstuk 3.

3 De Grenander schatter

3.1 Inleiding

De Grenander schatter heeft zijn naam te danken aan de bedenker, Ulf Grenander. Hij heeft veel gepubliceerd in zijn vakgebied en had onder andere in 1956 een tweetal artikelen, [5] en [6], uitgebracht. In deze artikelen staat ons hoe we een dalende dichtheid kunnen schatten, gegeven een aantal getallen uit een dalende dichtheid. Dit resultaat gebruiken we in dit verslag.

Stel we beschikken over een aantal datapunten. We weten dat ze uit een dalende dichtheid komen, maar we weten niet uit welke klasse van verdelingen ze komen. Een natuurlijke schatter voor de echte cumulatieve verdelingsfunctie, is de empirische cumulatieve verdelingsfunctie (ecdf). Deze gebruikt men als er geen extra informatie is over de onderliggende verdeling. We weten echter dat de dichtheid van onze datapunten, dalend is. Deze informatie gebruiken we om een schatter te vinden voor de dalende dichtheid. Zo maken we slim gebruik van de ecdf, die op het eerste gezicht toch niet zo bruikbaar is.

We kunnen uit Hoofdstuk 2 halen dat de Grenander schatter een niet-parametrische schatter is. Het uitrekenen van deze schatter is betrekkelijk eenvoudig, zoals zal blijken.

3.2 Uitrekenen

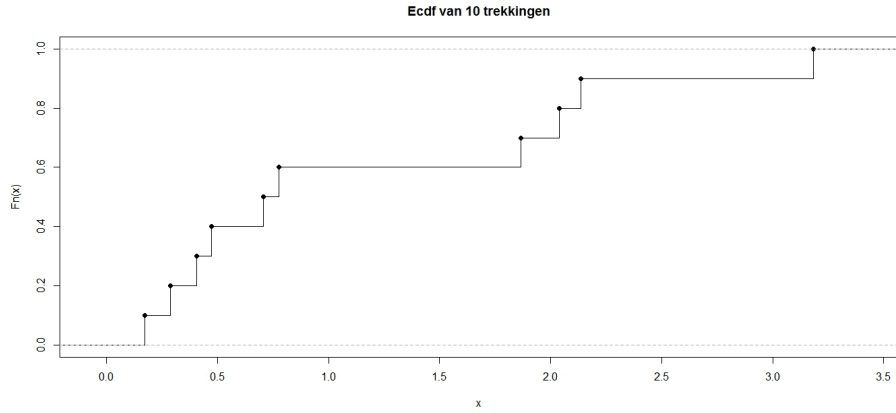
We zitten in de setting dat we over een aantal, zeg n , positieve datapunten beschikken. We nemen aan dat deze uit een dalende kansdichtheid f komen, en doen geen aannames over de familie van dichtheden waarin deze zit, zoals de normale of de exponentiële familie.

We beginnen met het construeren van de empirische cumulatieve verdelingsfunctie. De ecdf is de stapfunctie

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{X_i \leq x},$$

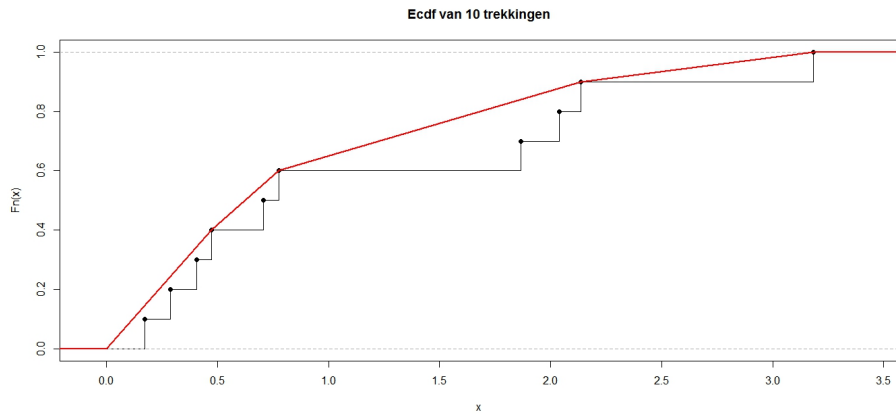
die springt bij de volgende gesorteerde waarneming. We beginnen dus bij 0 en springen telkens een stap $1/n$ omhoog, tot we bij de grootste datapunt zijn, daar springen we nogmaals naar de waarde 1. En de ecdf blijft ook 1 voor alle $x \geq X_{(n)}$. Hierbij is $X_{(i)}$ het i -de gesorteerd datapunt. De ecdf is rechtscontinu, en maakt eindig veel sprongen. Zie figuur 2 voor de ecdf van 10 datapunten. We zien duidelijk dat deze een sprong maakt van $1/10$, precies op de datapunten. Merk op dat er geldt $F_n(X_{(i)}) = i/n$ voor elke gesorteerd datapunt.

In het artikel van Grenander wordt nu gesuggereerd dat we naar de kleinste concave majorant moeten kijken. Grafisch is dit voor te stellen als een touw bij 0 vast te pinnen in het netgenoemde somdiagram, en vervolgens vanuit een verticale lijn omlaag te buigen tot we een horizontale lijnstuk op het eind krijgen. We krijgen dan een stuksgewijs lineaire functie, de kleinste concave majorant. Deze wordt genoteerd met $\hat{F}_n$. Deze concave majorant kan je berekenen door vanuit een vaste punt $(X_{(i)}, \frac{i}{n})$ lijnen te trekken naar andere punten $(X_{(j)}, \frac{j}{n})$, met $j > i$. Dan kiezen we de lijn met de grootste helling. Zodra we een lijn tussen $(X_{(i)}, \frac{i}{n})$ en $(X_{(j)}, \frac{j}{n})$ voor een zekere j hebben geconstrueerd, nemen we



Figuur 2: De empirische verdelingsfunctie van 10 trekkingen

$i = j$ en herhalen we de procedure. In Figuur 3 zien we de kleinste concave majorant van de ecdf in Figuur 2. We zien inderdaad de concaviteit van $\hat{F}_n$ en dat deze boven F_n ligt.



Figuur 3: De ecdf, in het Rood de kleinste concave majorant

We zien ook dat deze concave majorant knikpunten heeft. We kunnen dus niet de afgeleide nemen, maar wel de linker-afgeleide nemen. Dit is dan de Grenander schatter:

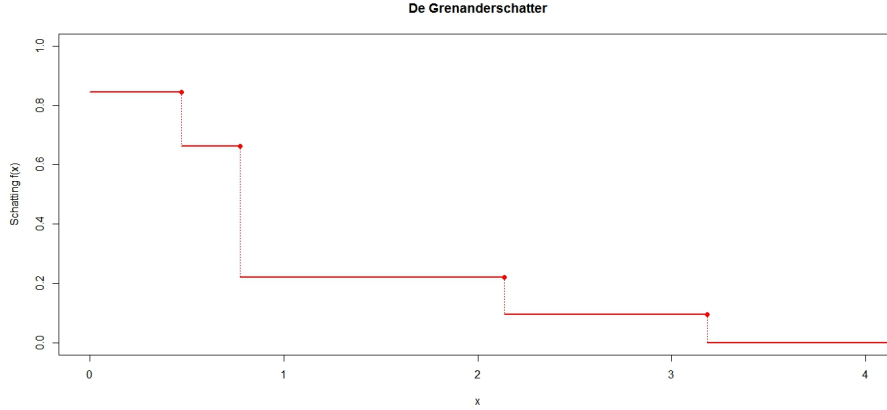
Definitie 1. We noemen de linker afgeleide van de kleinste concave majorant,

$$\hat{f}_n(x) := \lim_{h \uparrow 0} \frac{\hat{F}_n(x+h) - \hat{F}_n(x)}{h} \quad \forall x,$$

de Grenander schatter voor f .

In [2] en [13] wordt besproken waarom de Grenander schatter linkscontinu is, en constant op deelintervallen. Dit heeft te maken met het maximaliseren

van de likelihood. In Figuur 4 is de Grenander schatter te zien van de kleinste concave majorant zoals in Figuur 3. Een goede oefening voor het oog is om na te gaan dat de waarde van de Grenander schatter in een bepaald punt x gelijk is aan de helling van de kleinste concave majorant in x .



Figuur 4: De Grenander schatter

Grenander's stelling, zie [5], is dan

Stelling 3.1 (Stelling van Ulf Grenander (1956)). *Laat $X_i, i = 1, \dots, n$, random variabelen met een dalende dichtheid. Laat $\hat{f}_n$ zoals in Definitie 1. Deze maximaliseert de likelihood van de data, oftewel:*

$$\hat{f}_n = \arg \max_{f \text{ dalend}} \prod_{i=1}^n f(X_i).$$

Bewijs. Het bewijs zoals gegeven door Groeneboom en Lopuhaä, in [8], geeft een goed inzicht in waarom deze functie de likelihood maximaliseert. Het probleem is als volgt gedefinieerd: Vind een dalende dichtheid die de volgende afbeelding maximaliseert

$$f \mapsto \prod_{i=1}^n f(X_i).$$

Om het rekenwerk te vergemakkelijken bekijken we de loglikelihood, met een factor $1/n$ ervoor:

$$f \mapsto \frac{1}{n} \sum_{i=1}^n \ln f(X_i) = \int \ln f(x) dF_n(x).$$

We hebben de Grenander schatter, en we willen bewijzen dat deze de loglikelihood, en dus de likelihood zelf, maximaliseert. A. Zaanen had al bewezen in [14] (blz. 80) dat voor elke niet-negatieve functie g een rij stapfuncties (g_n) bestaat die monotoon convergeren naar g : $0 \leq f_n \uparrow g$. Daarna volgt het bewijs dat de integralen meeconvergeren, $\int f_n dF_n(x) \uparrow \int g dF_n(x)$. We bekijken daarom eerst de klasse van stapfuncties,

$$f(x) = \sum_{i=1}^n a_i 1_{[0, X_{(i)}]}.$$

Omdat f positief is, hebben we dat $a_i \geq 0, i = 1, \dots, n-1$ en $a_n > 0$. Herinner dat $X_{(i)}$ het i -de geordende datapunt is. Om de stelling te bewijzen bekijken we het volgende.

$$\begin{aligned} \int \left\{ \ln(f(x)) - \ln(\hat{f}_n(x)) \right\} d\hat{F}_n(x) &= \int \ln \left(1 + \frac{f(x) - \hat{f}_n(x)}{\hat{f}_n(x)} \right) \hat{f}_n(x) dx \\ &\leq \int \left\{ f(x) - \hat{f}_n(x) \right\} dx \\ &= 0 \end{aligned}$$

Dit geldt omdat $\ln(1+x) \leq x$, als $x > -1$. Hieruit volgt de ongelijkheid

$$\int \ln(f(x)) d\hat{F}_n(x) \leq \int \ln(\hat{f}_n(x)) d\hat{F}_n(x). \quad (1)$$

Verder merken we op dat $\hat{F}_n \geq F_n$, omdat $\hat{F}_n$ de kleinste concave *majorant* is van F_n . Er geldt dus, voor elke $i = 1, \dots, n$

$$\int 1_{[0, X_{(i)}]}(x) d\hat{F}_n(x) = \hat{F}_n(X_{(i)}) \geq F_n(X_{(i)}) = \int 1_{[0, X_{(i)}]}(x) dF_n(x).$$

En dit geeft

$$\int \ln(f(x)) d\hat{F}_n(x) \geq \int \ln(f(x)) dF_n(x). \quad (2)$$

Omdat $\hat{f}_n$ constant is tussen de sprongpunten geldt ten slotte dat

$$\int \ln(\hat{f}_n) d\hat{F}_n(x) = \int \ln(\hat{f}_n) dF_n(x). \quad (3)$$

Hieruit halen we nu

$$\frac{1}{n} \sum_{i=1}^n \ln \hat{f}_n(X_i) \stackrel{(3)}{=} \int \ln(\hat{f}_n) d\hat{F}_n(x) \stackrel{(1),(2)}{\geq} \int \ln f(x) dF_n(x) = \frac{1}{n} \sum_{i=1}^n \ln f(X_i). \quad (4)$$

En dit voor alle stapfuncties f .

We hebben dus inderdaad dat de Grenander schatter, $\hat{f}_n$, de likelihood van $(X_{(i)})_{i=1}^n$ maximaliseert $\square$

3.3 Consistentie van de Grenander schatter

De kleinste concave majorant $\hat{F}_n$ heeft een interessante eigenschap. Er geldt namelijk dat de supremumfout van $\hat{F}_n$ kleiner of gelijk is aan de supremumfout van de ecdf. Dit is bewezen in Robertson et al, zie [13] en wordt Marshall's lemma genoemd.

Lemma 3.2 (Marshall's lemma). *Laat $X_i, i = 1, \dots, n$ iid stochasten met een dalende dichtheid op $[0, \infty)$. Laat F de cumulatieve verdelingsfunctie zijn van deze stochasten. Laat nu $\hat{F}_n$ de kleinste concave majorant van F_n , en F_n de ecdf van $X_1, \dots, X_n$. Dan geldt er dat*

$$\sup_{x \in [0, \infty)} |\hat{F}_n(x) - F(x)| \leq \sup_{x \in [0, \infty)} |F_n(x) - F(x)|.$$

Bewijs. Definieer $e_n = \sup_x |F_n(x) - F(x)|$, $n \in \mathbb{N}$. Merk op dat F concaaf is, zijn afgeleide is namelijk dalend. De functie $F + e_n$ is ook concaaf voor elke n . Bovendien is $F(x) + e_n \geq F_n(x)$ voor alle n en $x \in [0, \infty)$. We weten dat $\hat{F}_n$ de kleinste concave majorant van F_n is, dus

$$F_n \leq \hat{F}_n \leq F + e_n, \quad \forall n.$$

Deze ongelijkheden geven voor alle $x \in [0, \infty)$

$$-e_n \leq F_n(x) - F(x) \leq \hat{F}_n(x) - F(x) \leq e_n.$$

We zien dus dat de fout $\hat{F}_n(x) - F(x)$ van boven wordt begrensd door e_n en van beneden door $-e_n$ voor alle $x \in [0, \infty)$. Dus de absolute waarde is begrensd door e_n voor alle x . Dan moet ook het supremum over alle x begrensd zijn door deze e_n :

$$\sup_{x \in [0, \infty)} |\hat{F}_n(x) - F(x)| \leq e_n = \sup_{x \in [0, \infty)} |F_n(x) - F(x)|.$$

□

Een interessante eigenschap van de Grenander schatter is de volgende. We hebben graag dat schatters consistent zijn, dat wil zeggen de schatter convergeert naar de echte functie als n naar oneindig gaat. In ons geval geldt dat ook, maar niet op het gehele domein van f .

Stelling 3.3. *Laat $X_1, \dots, X_n$ i.i.d. random variabelen met een dalende kansdichtheid op $[0, \infty)$. De Grenander schatter, zoals gegeven in definitie 1, is consistent op het inwendige van het domein van f . Dat wil zeggen, voor elke $\epsilon > 0$ geldt*

$$P\left(\lim_{n \rightarrow \infty} \hat{f}_n(t) = f(t), \forall t \in [\epsilon, \infty)\right) = 1.$$

Bewijs. Het bewijs staat in Barlow et al in [2]. De consistentie van de Grenander schatter zal volgen uit de consistentie van de kleinste concave majorant. We noteren voor een functie h , $h(x^-) = \lim_{h \uparrow 0} h(x + h)$ en $h(x^+)$ analoog, $\lim_{h \downarrow 0} h(x + h)$. De stelling van Glivenko-Cantelli zegt dat

$$\sup_{t \in [0, \infty)} |F_n(t) - F(t)| \rightarrow 0, \quad \text{als } n \rightarrow \infty,$$

met kans 1. Er geldt volgens Lemma 3.2 dat

$$\sup_{t \in [0, \infty)} |\hat{F}_n(t) - F(t)| \leq \sup_{t \in [0, \infty)} |F_n(t) - F(t)|.$$

Dus inderdaad

$$\sup_{t \in [0, \infty)} |\hat{F}_n(t) - F(t)| \rightarrow 0, \quad n \rightarrow \infty,$$

met kans 1. Nu rest ons de volgende opmerking. $\hat{F}_n$ is concaaf op het open interval $(0, \infty)$, en convergeert uniform op het interval naar F als $n \rightarrow \infty$. Vanwege de concaviteit is het zo dat, voor elke $k > 0$

$$\frac{\hat{F}_n(t) - \hat{F}_n(t - k)}{k} \geq \hat{f}_n(t^-) \geq \hat{f}_n(t^+) \geq \frac{\hat{F}_n(t + k) - \hat{F}_n(t)}{k}.$$

Dit geldt voor elke n . Laat nu n naar ∞ en k naar 0. We krijgen vanwege de uniforme convergentie van $\hat{F}_n$

$$f(t^-) \geq \limsup_n \hat{f}_n(t^-) \geq \liminf_n \hat{f}_n(t^+) \geq f(t^+).$$

In het bijzonder als f continu is in t , dan geldt $f(t^-) = f(t^+) = \lim_n \hat{f}_n(t^-)$.

We merken op dat $\hat{f}_n(t^-)$ de linkerafgeleide is van $\hat{F}_n(t)$ in het punt t . Dit is precies de Grenander schatter van f . We concluderen dat de Grenander schatter consistent is op het inwendige van het domein van f . $\square$

We sluiten dit hoofdstuk met de opmerking dat bovenstaande stelling ook geldt voor begrensde gesloten domeinen. Stel dat f gedefinieerd is op $[0, 1]$. Dan geldt dat $\hat{f}_n$ consistent is op het inwendige van $[0, 1]$.

4 De supremumafstand

4.1 De limietverdeling van het supremum

Pas zeer recentelijk is in 2012 een artikel uitgebracht door Durot, Kulikov en Lopuhaä, zie [4]. Deze bevat een resultaat over de supremumafstand tussen de Grenander schatter en de echte dichtheid, f .

Het resultaat zegt dat de supremumafstand, op schaling na, convergeert in verdeling naar een standaard Gumbelverdeling. Het bewijs zullen we niet behandelen. Het belang van deze stelling is vooral voor het maken van zogeheten simultane betrouwbaarheidsbanden een belangrijke bouwsteen.

Om tot de stelling te komen, hebben we een aantal aannames nodig. Deze staan ook in [4]. De stelling kan worden uitgebreid naar $[0, M]$, maar we bekijken het interval $[0, 1]$ en nemen het volgende aan:

- De functie f is dalend en differentiëerbaar op $[0, 1]$ en

$$\inf_{t \in [0, 1]} |f'(t)| > 0, \quad \sup_{t \in [0, 1]} |f'(t)| < \infty.$$

- Er is een $C_0 > 0$ en $\sigma \in (0, 1]$ zo dat voor alle $t, u \in [0, 1]$,

$$|f'(u) - f'(t)| \leq C_0 |u - t|^\sigma.$$

De volgende constanten zijn nodig voor de stelling, ze zijn gehaald uit [10].

$$\begin{aligned} \mu_n &= 1 - \frac{\kappa}{2^{1/3}(\log(n))^{2/3}} + \frac{1}{\log(n)} \left[\frac{1}{3} \log(\log(n)) + \log(\theta C_f) \right] \\ \kappa &= 2^{1/3} \cdot 2.3381 = 2.945821 \\ \theta &= \frac{4^{1/3}}{0.7022} = 2.260611 \\ C_f &= 2 \int_0^1 \left(\frac{|f'(t)|^2}{f(t)} \right)^{1/3} dt. \end{aligned}$$

Let op dat in [10] onze θ juist λ wordt genoemd. De stelling luidt dan als volgt:

Stelling 4.1 (De Supremumafstand, 2012). *Laat $(\alpha_n), (\beta_n)$ twee reëelwaardige rijen met*

$$\frac{\alpha_n}{n^{-\frac{1}{3}} (\log n)^{-\frac{2}{3}}} \rightarrow \infty, \text{ als } n \rightarrow \infty,$$

en

$$\frac{\beta_n}{n^{-\frac{1}{3}} (\log n)^{-\frac{2}{3}}} \rightarrow \infty, \text{ als } n \rightarrow \infty.$$

Definieer

$$T_n = \log(n) \left\{ \left(\frac{n}{\log(n)} \right)^{1/3} \sup_{t \in (\alpha_n, 1 - \beta_n]} \frac{|\hat{f}_n(t) - f(t)|}{|2f'(t)f(t)|^{1/3}} - \mu_n \right\}.$$

Dan geldt voor alle $x \in \mathbb{R}$ het volgende:

$$P(T_n \leq x) \rightarrow e^{-e^{-x}}, \quad \text{als } n \rightarrow \infty.$$

Hier zien we dat na herschaling de stochast, de herschaalde supremumafstand, naar een standaard Gumbel verdeling convergeert.

4.2 Betrouwbaarheidsband

Zoals eerder gezegd is het belang van deze stelling het maken van zogenaamde betrouwbaarheidsbanden.

Definitie 2. *We definiëren een paar van twee functies, (L_n, U_n) gedefinieerd op dezelfde domein D als f , waarvoor het volgende geldt:*

$$P(L_n(t) \leq f(t) \leq U_n(t), \forall t \in D) = p$$

We zeggen dan dat (L_n, U_n) een 100p%-betrouwbaarheidsband is.

Merk op dat de band afhangt van n , de steekproefsgrootte. Dit is natuurlijk logisch, omdat bij een grotere steekproef we een betere benadering hebben en een nauwere band in het algemeen. Het doel is dus om een paar (L_n, U_n) te vinden waarvoor bovenstaande geldt voor onze dalende dichtheid f .

Ten eerste moeten we opmerken dat we in het geval van de Grenander schatter niet het gehele domein kunnen voorzien van een betrouwbaarheidsband. Het is namelijk zo dat de Grenander schatter inconsistent is in de randpunten, en daarom doen deze niet mee in de stelling. We kunnen een betrouwbaarheidsband maken op het interval $(\alpha_n, 1 - \beta_n)$, met α_n, β_n zoals in Stelling 4.1 gekozen.

We definiëren ten eerste de zogenaamde herschaalfunctie

$$H_n(x) = \log(n) \left\{ \left(\frac{n}{\log(n)} \right)^{1/3} x - \mu_n \right\}, \quad (5)$$

die de herschaling uit Stelling 4.1 toepast op een getal x . De inverse wordt gegeven door

$$H^{-1}(x) = \left\{ \frac{x}{\log n} + \mu_n \right\} \left(\frac{\log n}{n} \right)^{1/3}$$

Dan geldt er

$$P \left(H_n \left(\sup_{t \in (\alpha_n, 1 - \beta_n]} \frac{|\hat{f}_n(t) - f(t)|}{|2f'(t)f(t)|^{1/3}} \leq x \right) \right) \rightarrow e^{-e^{-x}}.$$

We willen een 100p%-betrouwbaarheidsband maken voor het supremum. Kies x_p zodanig dat

$$P \left(\sup_{t \in (\alpha_n, 1 - \beta_n]} \frac{|\hat{f}_n(t) - f(t)|}{|2f'(t)f(t)|^{1/3}} \leq H_n^{-1}(x_p) \right) \rightarrow p.$$

Kies dus $x_p = -\log(-\log(p))$, dan geldt

$$\sup_{t \in (\alpha_n, 1 - \beta_n]} \frac{|\hat{f}_n(t) - f(t)|}{|2f'(t)f(t)|^{1/3}} \leq H_n^{-1}(x_p),$$

met een kans die gaat naar p , als n naar ∞ . Bijvoorbeeld voor een 95%-betrouwbaarheidsband kiezen we $x_{0.95} = -\log(-\log(0.95)) = 2.97$. Dan is $H_n^{-1}(x_{0.95}) = 0.0235$.

We maken nu een $100p\%$ -betrouwbaarheidsband voor f . Neem de volgende functies:

$$\begin{aligned} L_n(t) &= \hat{f}_n(t) - |2f'(t)f(t)|^{1/3} \cdot H^{-1}(x_p), \\ U_n(t) &= \hat{f}_n(t) + |2f'(t)f(t)|^{1/3} \cdot H^{-1}(x_p), \end{aligned} \quad (6)$$

voor $t \in (\alpha_n, 1 - \beta_n]$. Merk op dat we in het project de functie f zelf kiezen. In tegenstelling tot in de praktijk, waar deze waardevolle informatie niet beschikbaar is. Hier gebruiken we dan schattingsmethoden om f en f' te schatten. Dan hebben we dat de kans dat dit interval f bevat voor alle t in $(\alpha_n, 1 - \beta_n]$ ongeveer gelijk is aan p .

Om daadwerkelijk via een computer te bepalen of f tussen twee functies ligt, moeten we in principe alle t in $[\alpha_n, 1 - \beta_n]$ nagaan. Dit is niet exact door een computer te doen, dus moeten we het anders doen. We presenteren een makkelijkere manier om deze check uit te voeren.

Lemma 4.2. *Laat L_n, U_n zoals in (6). De volgende twee uitspraken zijn equivalent:*

$$L_n(t) \leq f(t) \leq U_n(t), \quad \forall t \in (\alpha_n, 1 - \beta_n], \quad (7)$$

$$H_n \left(\sup_{t \in (\alpha_n, 1 - \beta_n]} \frac{|\hat{f}_n(t) - f(t)|}{|2f'(t)f(t)|^{1/3}} \right) \leq x_p. \quad (8)$$

Bewijs. Het volgende geldt

$$\begin{aligned} L_n(t) \leq f(t) \leq U_n(t), \quad \forall t \in (\alpha_n, 1 - \beta_n] &\stackrel{(*)}{\iff} \\ |\hat{f}_n(t) - f| \leq |2f'(t)f(t)|^{1/3} \cdot H^{-1}(x_p), \quad \forall t \in (\alpha_n, 1 - \beta_n] &\iff \\ \frac{|\hat{f}_n(t) - f|}{|2f'(t)f(t)|^{1/3}} \leq H^{-1}(x_p), \quad \forall t \in (\alpha_n, 1 - \beta_n] &\iff \\ \sup_{t \in (\alpha_n, 1 - \beta_n]} \frac{|\hat{f}_n(t) - f(t)|}{|2f'(t)f(t)|^{1/3}} \leq H^{-1}(x_p) &\iff \\ H_n \left(\sup_{t \in (\alpha_n, 1 - \beta_n]} \frac{|\hat{f}_n(t) - f(t)|}{|2f'(t)f(t)|^{1/3}} \right) \leq x_p & \end{aligned}$$

Omdat $f - \hat{f}_n$ begrensd is door de uitdrukking $|2f'f|^{1/3} \cdot H^{-1}(x_p)$, is ook $|f - \hat{f}_n|$ begrensd door deze uitdrukking. Dus (*) geldt inderdaad. $\square$

5 Het experiment

5.1 Inleiding

We hebben in het algemeen gekeken naar de Grenander schatter [5], en hoe deze tot stand kwam, zie Hoofdstuk 3. Daarnaast hadden we ook wat algemene eigenschappen besproken die gelden voor de L_p -fout van de Grenander schatter.

Ten slotte is in 2012 een resultaat bewezen die voor ons van belang is voor simultane betrouwbaarheidsbanden. Naarmate de steekproefsgrootte steeds groter wordt, zal de herschaalde L_∞ -fout van de Grenander schatter naar een Gumbel verdeling convergeren [4]. Deze stelling is gegeven in Hoofdstuk 4, samen met een stuk over de constructie van een dergelijke betrouwbaarheidsband in het algemeen. We weten dat dit resultaat zeker waar is als de steekproefsgrootte naar oneindig gaat, maar waar wij ons mee bezig zullen houden is het geval dat we een eindige steekproefsgrootte hebben. Dit is vanuit een praktisch oogpunt natuurlijk interessant.

We zullen met een zelfgekozen dalende kansdichtheid een aantal trekkingen doen. Daarmee kunnen we de Grenander schatter maken. Omdat we daadwerkelijk de echte dichtheid weten, kunnen we de L_∞ -fout exact berekenen. Dit zal ons dan een realisatie geven waarvan verondersteld wordt dat deze bij benadering uit een Gumbel verdeling komt. Dit kunnen we vaker doen, en met behulp van statistische hulpmiddelen onderzoeken we of de L_∞ -fout inderdaad bij benadering Gumbel verdeeld is. Hiermee bespreken we in hoeverre Stelling 4.1 bruikbaar is in het eindige geval.

5.2 De afgeknotte exponentiele verdeling

Er is al eerder gesproken over een afgeknotte exponentiële verdeling. We laten nu zien welke afgeknotte exponentiële verdeling, aev, we gebruiken en we leggen uit hoe we deze maken. Laat eerst $f_{exp}(x)$ de exponentiële kansdichtheid met parameter λ en $x \in [0, \infty)$. Dus

$$f_{exp}(x) = \begin{cases} 0 & x < 0, \\ \lambda e^{-\lambda x} & x \geq 0. \end{cases}$$

De cumulatieve verdelingsfunctie is gegeven door

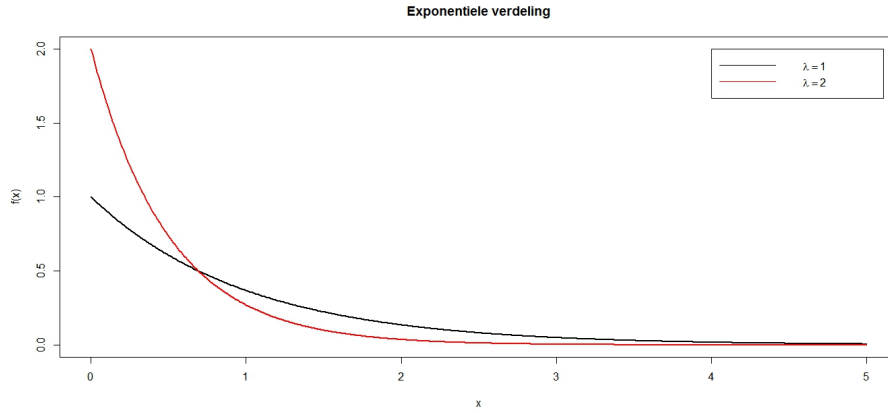
$$F_{exp}(x) = \begin{cases} 0 & x < 0, \\ 1 - e^{-\lambda x} & x > 0. \end{cases}$$

In Figuur 5 zien we twee exponentiële dichtheden. De exponentiële dichtheid heeft géén begrensde domein, dus we kunnen deze dichtheid niet gebruiken om Stelling 4.1 te testen. We beschouwen een gesloten interval $[0, M]$, waarbij $M > 0$, en beschouwen het volgende:

Laat $X \sim Exp(\lambda)$ en definieer $E = X|X \leq M$, dan geldt voor alle $x \in [0, M]$ dat

$$P(E \leq x) = P(X \leq x|X \leq M) = \frac{P(X \leq x, X \leq M)}{P(X \leq M)} \stackrel{(*)}{=} \frac{P(X \leq x)}{P(X \leq M)}.$$

De gelijkheid (*) geldt inderdaad, omdat de gebeurtenis $X \leq x$ bevat is in $X \leq M$. We vinden dus een uitdrukking voor de verdelingsfunctie van E ,

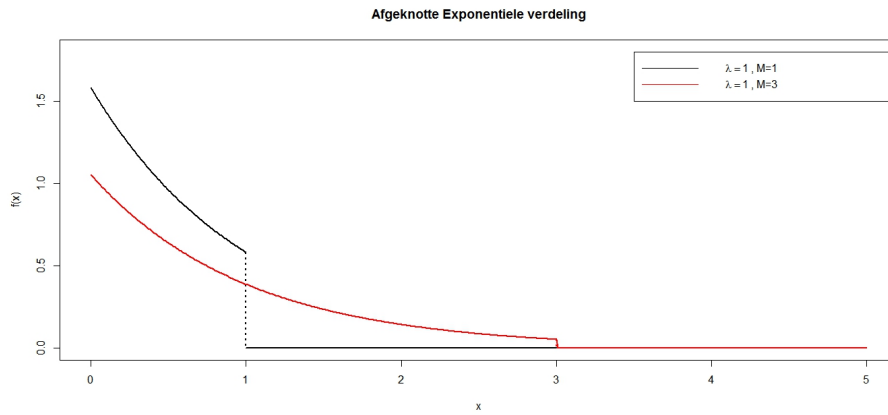


Figuur 5: De exponentiële dichtheid, met parameter λ .

namelijk

$$E \sim F(x) = \begin{cases} 0 & x < 0, \\ \frac{1-e^{-\lambda x}}{1-e^{-\lambda M}}, & 0 \leq x \leq M, \\ 1 & x > M. \end{cases}$$

Een illustratie van de grafiek van de dichtheid f van E is in Figuur 6 te zien. We



Figuur 6: De afgeknotte exponentiële dichtheid, met parameter M, λ .

hebben nu een verdelingsfunctie geconstrueerd die op een compacte³ verzameling is gedefinieerd. We willen nu trekkingen doen uit de verdeling van E . Daarvoor moeten we nagaan hoe we kunnen simuleren uit de verdeling van E .

Definieer $U = F(E)$, dan heeft $U \sim Unif(0, 1)$, een standaard uniforme verdeling. Het is namelijk duidelijk dat $U \in [0, 1]$, en bovendien

$$P(U \leq x) = P(F(E) \leq x) = P(E \leq F^{inv}(x)) = F(F^{inv}(x)) = x. \quad (9)$$

³Een verzameling in $\mathbb{R}$ is compact $\leftrightarrow$ de verzameling is begrensd en gesloten

Er geldt

$$U = F(E) \iff E = F^{inv}(U).$$

We kunnen E dus simuleren door trekkingen uit een standaard uniforme verdeling te doen, en de inverse verdelingsfunctie van E toe te passen. Dit heet ook wel de INVERSE VERDELINGSFUNCTIE METHODE. We moeten natuurlijk wel nagaan of onze verdelingsfunctie wel inverteerbaar is, en wat de inverse is.

De kansdichtheid van de aev is strikt dalend op $[0, M]$ en natuurlijk positief, en de cumulatieve verdelingsfunctie gedefinieerd door

$$F(x) = \int_0^x f(s) ds, \quad 0 \leq x \leq M,$$

is strikt stijgend op dit interval. We weten dat functies die strikt stijgend zijn op een interval, een inverse functie hebben op dat interval. Deze wordt gegeven door het volgende op te lossen naar $y \in [0, 1]$:

$$\begin{aligned} y &= \frac{1 - e^{-\lambda x}}{1 - e^{-\lambda M}}, \\ y(1 - e^{-\lambda M}) &= 1 - e^{-\lambda x}, \\ \lambda^{-1} \ln(1 - y(1 - e^{-\lambda M})) &= x, \\ F^{inv}(y) &= \lambda^{-1} \ln(1 - (y(1 - e^{-\lambda M}))). \end{aligned}$$

We hebben dus de inverse, gedefinieerd op $[0, 1]$ met bereik $[0, M]$. We passen de inverse verdelingsfunctie methode toe en vinden

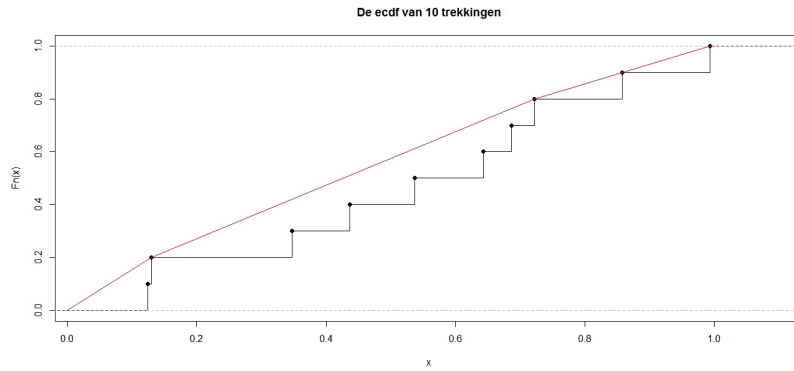
$$E = F^{inv}(U) = \lambda^{-1} \ln(1 - (U(1 - e^{-\lambda M}))).$$

Waarbij U dus standaard uniform is. Nu zijn we in staat om E te simuleren.

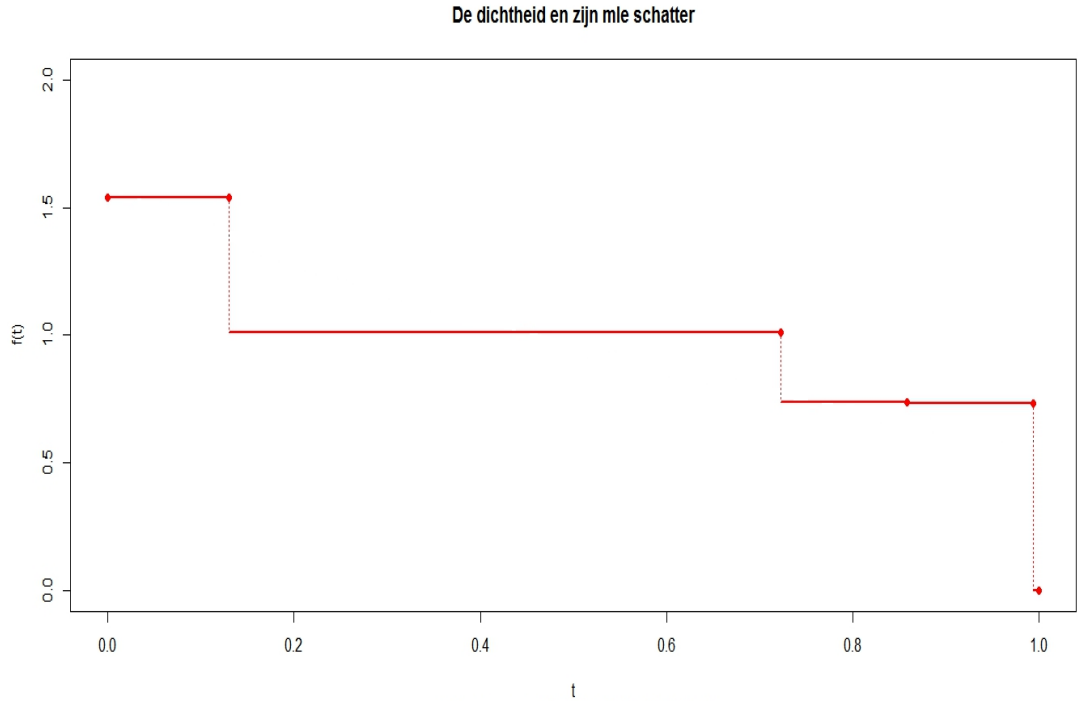
5.3 De Grenander schatter

We beschikken over een dataset, zeg $x_1, \dots, x_n > 0$, van n trekkingen uit de afgeknotte exponentiële verdeling (aev). De empirische verdelingsfunctie en de kleinste concave majorant, $\hat{F}_n$, zijn in Figuur 7 weergegeven. Zie Hoofdstuk 3 voor de precieze constructie van deze functies.

We weten ook dat de linkerafgeleide van de kleinste concave majorant de Grenander schatter is, $\hat{f}_n$. Deze is afgebeeld in Figuur 8.



Figuur 7: De ecdf en de kleinste concave majorant van 10 trekkingen uit een afgeknotte exponentiele, $\lambda = 1, M = 1$.



Figuur 8: De Grenander schatter van $f, n = 10, \lambda = 1, M = 1$.

5.4 De supremumafstand

De volgende stap is het uitrekenen van de L_∞ -fout van $\hat{f}_n$. We weten bij voorbaat al dat

$$f(t) = \frac{\lambda}{1 - e^{-\lambda M}} e^{-\lambda t}, \quad t \in [0, M],$$

en 0 elders. De afgeleide wordt gegeven door

$$f'(t) = \frac{-\lambda^2}{1 - e^{-\lambda M}} e^{-\lambda t}, \quad t \in [0, M],$$

en 0 elders.

Er wordt automatisch aan de voorwaarden voldaan met deze keuze van f . Zie Sectie 4 voor de aannames. Ten eerste moet f differentiëerbaar zijn en $\inf |f'(t)| > 0$ en $\sup |f'(t)| < \infty$. De functie f' is strikt stijgend op een gesloten interval, dus $\inf_t |f'(t)| = \min_t |f'(t)| = \frac{\lambda^2}{1 - e^{-\lambda M}} e^{-\lambda M}$. Verder is f' begrensd en het supremum van f' dus ook. Aan de tweede aanname is automatisch voldaan, omdat f tweemaal differentiëerbaar is.

Nu f gegeven is, berekenen we de constanten vermeld in Stelling 4.1. We berekenen ten eerste C_f .

$$\begin{aligned} C_f &= 2 \int_0^M \left(\frac{|\lambda^2 e^{-\lambda t}|^2}{(\lambda e^{-\lambda t})(1 - e^{-\lambda M})} \right)^{1/3} dt \\ &= 2 \int_0^M \left(\frac{\lambda^4 e^{-2\lambda t}}{(\lambda e^{-\lambda t})(1 - e^{-\lambda M})} \right)^{1/3} dt \\ &= \frac{2}{(1 - e^{-\lambda M})^{1/3}} \int_0^M (\lambda^3 e^{-\lambda t})^{1/3} dt \\ &= \frac{2}{(1 - e^{-\lambda M})^{1/3}} \int_0^M \lambda e^{-\frac{\lambda}{3} t} dt \\ &= \frac{2}{(1 - e^{-\lambda M})^{1/3}} \left(-3e^{-\frac{\lambda}{3} t} \Big|_0^M \right) \\ &= (-6e^{-\frac{1}{3}\lambda \cdot M} + 6)/(1 - e^{-\lambda M})^{1/3} \end{aligned}$$

Hieruit volgt dan ook μ_n , deze is namelijk

$$\begin{aligned} \mu_n &= 1 - \frac{2.95}{2^{1/3}(\log(n))^{2/3}} \\ &\quad + \frac{1}{\log(n)} \left[\frac{1}{3} \log(\log(n)) + \log(-2.26 \cdot 6e^{-\frac{1}{3}\lambda \cdot M} + 13.6) \right] \end{aligned}$$

Een voorbeeld met $M = \lambda = 1$: dan is $C_f \approx 1.98$ en $\mu_{1000} \approx 0.67$

De herschaalde L_∞ -fout van $\hat{f}_n$ wordt gegeven door

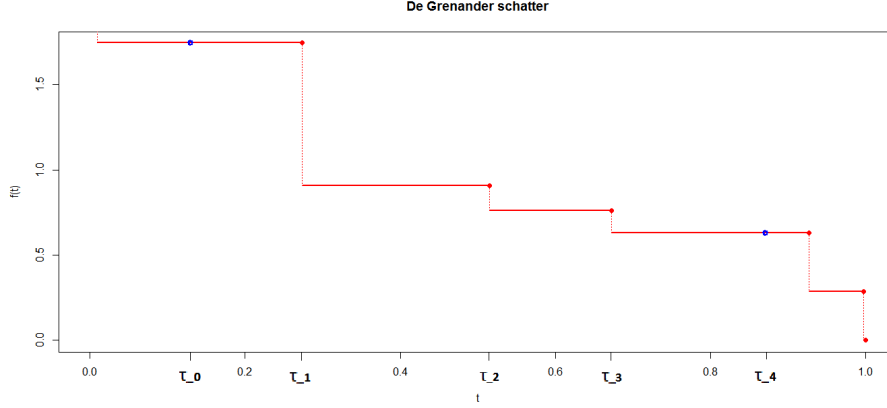
$$\sup_{t \in (\alpha_n, M - \beta_n]} \frac{|\hat{f}_n(t) - f(t)|}{|2f(t)f'(t)|^{1/3}}. \quad (10)$$

De noemer in (10) schrijven we dan als

$$h(t) := |2f(t)f'(t)|^{1/3} = 2^{1/3} \lambda (1 - e^{-\lambda M})^{-2/3} e^{-\frac{2}{3}\lambda t}, t \in [0, M]. \quad (11)$$

We kiezen $\alpha_n = \beta_n$ hetzelfde, zodat we een symmetrisch interval krijgen, rond $\frac{1}{2}M$, waarover we het supremum nemen. Dit getal α_n noemen we dan ϵ_n .

Voor elke n is $\hat{f}_n$ een stapfunctie met K sprongen in $(\epsilon_n, M - \epsilon_n]$. Laat n vast zijn en definieer $\tau_0 = \epsilon_n$ en $\tau_{K+1} = M - \epsilon_n$. Laat $\tau_1 \leq \tau_2 \leq \dots \leq \tau_K$ nu



Figuur 9: De sprongpunten van de Grenander schatter

de t -coördinaten van de sprongpunten zijn met $\tau_0 \leq \tau_i \leq \tau_{K+1}$. In Figuur 9 is een voorbeeld van een dergelijke ordening, waarbij $M = 1$ en $\epsilon_n = 0.13$. Hier zijn er 3 sprongpunten tussen ϵ_n en $1 - \epsilon_n$, dus $K = 3$.

We kunnen (10) nu ook schrijven als

$$\max_{i=1, \dots, K+1} \sup_{t \in (\tau_{i-1}, \tau_i]} \frac{|a_i - f(t)|}{h(t)}$$

waarbij $a_i = \hat{f}_n(\tau_i)$. Ons probleem reduceert nu tot het vinden van het supremum per interval waarop $\hat{f}_n$ constant is. Merk op dat de $(\tau_{i-1}, \tau_i]$ onderling disjunct zijn.

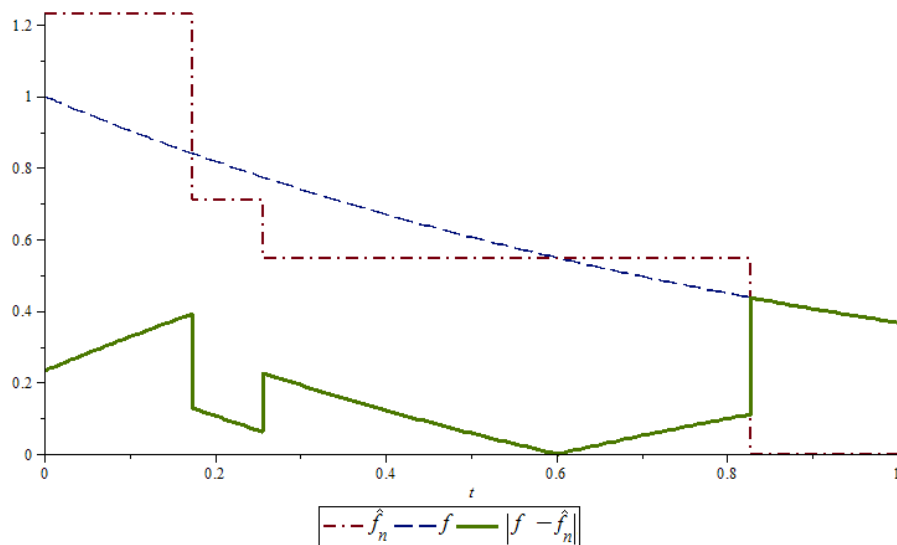
We gebruiken nu dat f de afgeknotte exponentiële dichtheid is. Laat i nu vast tussen 1 en $K + 1$. Dan krijgen we op het interval $(\tau_{i-1}, \tau_i]$

$$\sup_{t \in (\tau_{i-1}, \tau_i]} \frac{|a_i - f(t)|}{h(t)} = \sup_{t \in (\tau_{i-1}, \tau_i]} \frac{|a_i - \lambda e^{-\lambda t} / (1 - e^{-\lambda M})|}{2^{1/3} \lambda (1 - e^{-\lambda M})^{-2/3} e^{-\frac{2}{3} \lambda t}}.$$

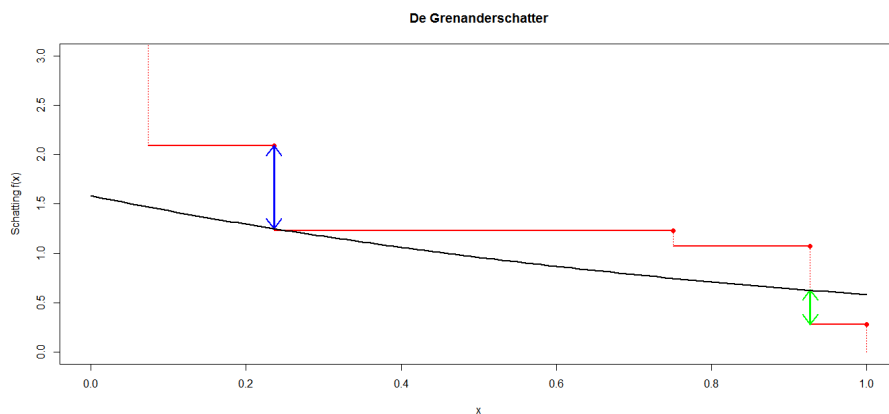
Omdat $(a_i - f(t))/h(t)$ een strikt stijgende functie is, wordt het supremum alleen op het linkerrandpunt van het interval $(\tau_{i-1}, \tau_i]$ aangenomen. De absolute waarde van deze uitdrukking heeft deze eigenschap niet, maar gelukkig neemt deze wel altijd zijn supremum aan op een van de randpunten van het interval. Voor de duidelijkheid is in Figuur 10 een voorbeeld opgenomen waarin de afstand wordt geplot. We zien dat $|a_i - f(t)|/h(t)$ niet per se strikt stijgend is, maar wel altijd een maximum heeft op τ_{i-1} of τ_i voor zekere i . Er geldt dus op een interval

$$\sup_{t \in (\tau_{i-1}, \tau_i]} \frac{|a_i - f(t)|}{h(t)} = \max \left\{ \frac{|a_i - f(\tau_{i-1})|}{h(\tau_{i-1})}, \frac{|a_i - f(\tau_i)|}{h(\tau_i)} \right\}. \quad (12)$$

Om te laten zien waarom we spreken over een supremum en niet over een maximum in (10) laten we een informatief plaatje zien: In Figuur 11 is een voorbeeld te zien van de afstand $|a_3 - f(\tau_3)|$, aangegeven met de blauwe pijl. Hier wordt de afstand daadwerkelijk aangenomen. De afstand $|a_6 - f(\tau_5)|$ komt in het



Figuur 10: De afstand tussen $\hat{f}_n$ en f



Figuur 11: In het rood de Grenander schatter.

plaatje overeen met de groene pijl, hier betreft het een afstand die willekeurig dicht benaderd wordt, maar niet aangenomen wordt.

Hier zien we dus dat het, op het eerste oog ingewikkelde, supremum reduceert tot een maximum dat alleen op $K + 2$ punten aangenomen kan worden:

$$\sup_{t \in (\alpha_n, M - \beta_n]} \frac{|\hat{f}_n(t) - f(t)|}{h(t)} = \max_{i=0, \dots, K+1} \left[\max \left\{ \frac{|a_i - f(\tau_{i-1})|}{h(\tau_{i-1})}, \frac{|a_i - f(\tau_i)|}{h(\tau_i)} \right\} \right] \quad (13)$$

Het maximum in vergelijking (13) kunnen we gemakkelijk berekenen met een computer, en zodoende vinden we het supremum over het gehele interval $(\epsilon_n, M - \epsilon_n]$.

Na toepassing van de herschaling in (5) vinden we een realisatie van een stochast die Gumbel verdeeld zou moeten zijn, bij benadering. Om te controleren of dit inderdaad het geval is, herhalen we dit bovenstaande experiment vaker en krijgen we een indruk hoe deze stochast verdeeld is.

5.4.1 Samengevat

1. Kies de afkapping M en de parameter λ van de afgeknotte exponentiële verdeling en trek hier n getallen uit, zie paragraaf 5.2.
2. Construeer de Grenander schatter $\hat{f}_n$, door de linker afgeleide van $\hat{F}_n$ te nemen, zie paragraaf 5.3.
3. Kies ϵ_n en gebruik de formule in gelijkheid (13) om het supremum te berekenen over het interval $(\epsilon_n, M - \epsilon_n]$.
4. Herschaal het supremum met de functie in (5).

Nu hebben we een realisatie van de stochast zoals in Stelling 4.1 die asymptotisch Gumbel verdeeld is. Om wat te kunnen zeggen over de verdeling van deze stochast willen we natuurlijk meer dan één realisatie. Het stappenplan kunnen we nog een keer doorlopen, en we krijgen een andere realisatie van dezelfde stochast. Dit kunnen we vaker doen. Laat m het aantal keer zijn dat we het stappenplan doorlopen. We kunnen dan meer zeggen over de stochast, en nagaan in hoeverre deze Gumbel verdeeld is en Stelling 4.1 voor eindige n te gebruiken is.

5.5 Resultaten

Voor we verder gaan volgt voor het gemak van de lezer een opsomming van alle te gebruiken parameters; f is de afgeknotte exponentiële dichtheid:

- λ, M , de parameters van f . In het bijzonder is M de afkapping.
- n , het aantal trekkingen uit f .
- ϵ_n , afstand tot de randpunten van het domein van f .
- m , het aantal herhalingen van het experiment, zie 5.4.1.
- p , de te kiezen asymptotische betrouwbaarheid van de betrouwbaarheidsband.

We weten dat de supremumafstand tussen f en zijn maximum likelihood schatter, mits herschaald, naar een Gumbel verdeling gaat als n toeneemt. In de praktijk hebben we maar een beperkte steekproef. Hoe goed is dan de benadering van de Gumbel verdeling? Kunnen we dan ook wat zeggen over de nauwkeurigheid van de Grenander schatter? Ook de vraag hoe goed een betrouwbaarheidsband zou zijn op basis van de Gumbel verdeling is zeer interessant. In deze paragraaf gaan we in op dit soort vragen en geven we antwoorden en resultaten van het experiment.

5.5.1 Consistentie

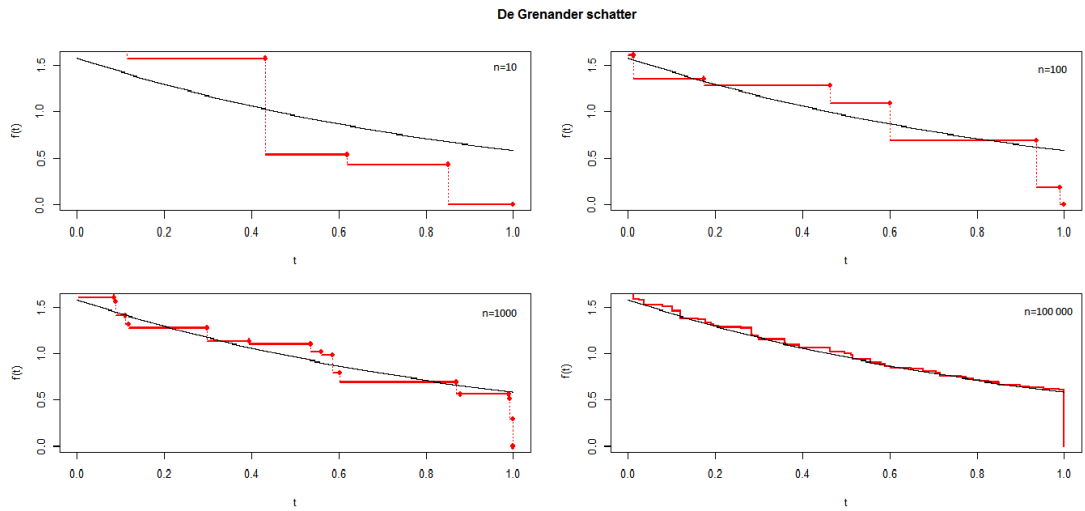
We kiezen alvast de volgende parameters: $\lambda = 1, M = 1$. Om een indruk te krijgen hoe goed de Grenander schatter de echte f schat, simuleren we het experiment één keer. In figuur 12 zien we in het rood de Grenander Schatter voor toenemende n . Ook visueel is de consistentie van de schatter bevestigd, indien we weg van de randpunten blijven. Op de randpunten kan het nog steeds misgaan. Op $t = 0$ zien we dan ook een overschatting, en op $t = 1$ zien we een onderschatting. Herinner dat $f(t) = 0$ buiten het interval $[0, 1]$.

Het is duidelijk dat de supremumfout van $\hat{f}_n$, gekeken op $[0, 1]$, bijna altijd wordt aangenomen in 0 of in 1. Dit zal ongewenste invloed hebben op de verdeling van het herschaalde supremum. Om dit te vermijden kiezen we ϵ_n zo dat we voldoende wegblijven van de randpunten. Hoe ver we moeten wegblijven staat in Stelling 4.1.

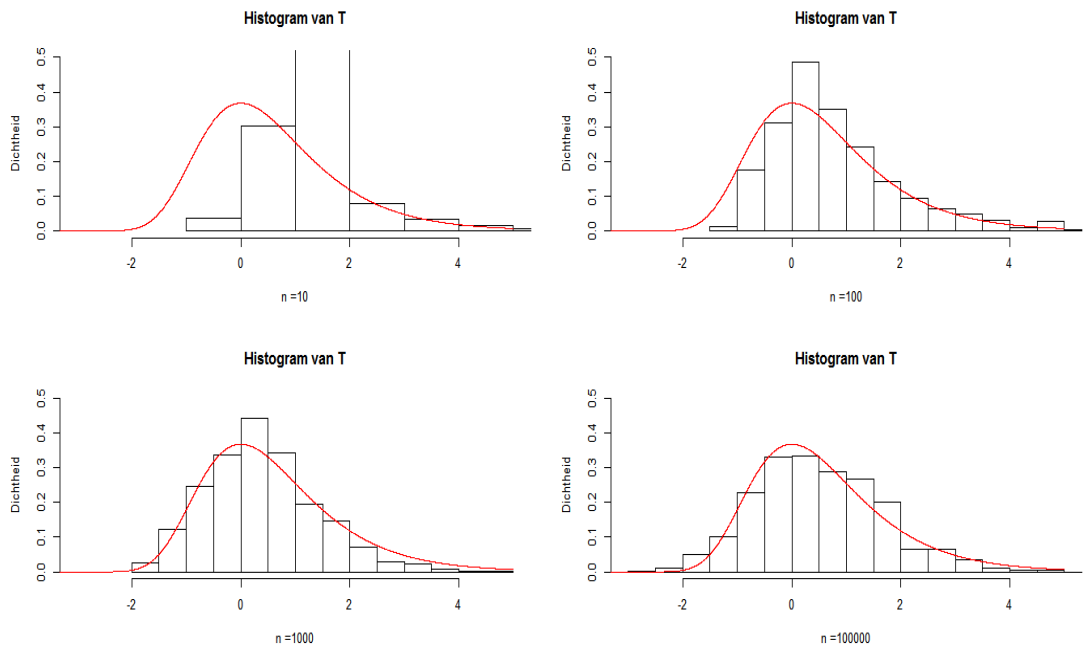
5.5.2 Herhaling van het experiment, vaste ϵ_n

Laat $\epsilon_n = 0.1$ voor alle n . We nemen het herschaalde supremum, zoals in Stelling 4.1, over $[0.1; 0.9]$ en simuleren $m = 1000$ keer. We noemen de stochast waaruit we simuleren T_n , zoals in Stelling 4.1. Om te kijken hoe goed we de Gumbel verdeling benaderen, gebruiken we een aantal grafische hulpmiddelen. Hierbij kijken we dus of er in redelijke mate geldt dat $T \approx \text{Gumbel}(0,1)$.

We maken voor variërende n een histogram op basis van de 1000 herschaalde supremumafstanden, in Figuur 13. We zetten in het rood ook de Gumbel dichtheid erin. Voor toenemende n zien we dat de histogram meer op de Gumbel dichtheid lijkt, alhoewel er niet veel onderscheid is.



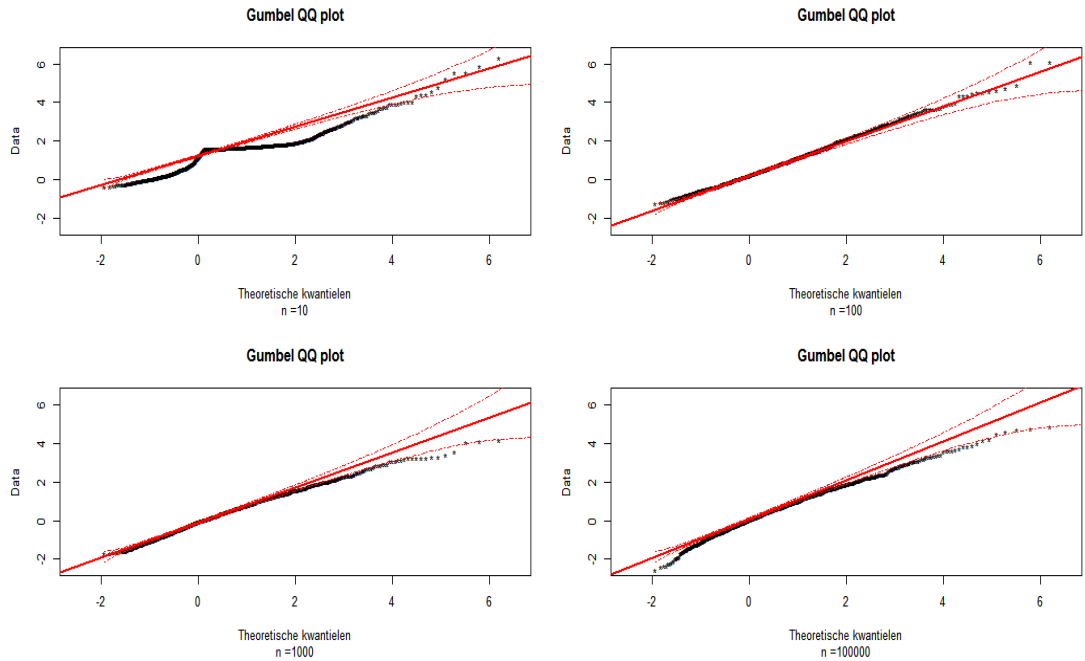
Figuur 12: De Grenander schatter benadert inderdaad de echte f .



Figuur 13: De histogram nadert de Gumbel dichtheid redelijk.

Tot slot vergelijken we de kwantielen van de steekproef met de Gumbel kwantielen. Dit doen we in een qq-plot. Een rechte lijn geeft aan dat de steekproef inderdaad uit een Gumbel verdeling komen, maar de helling van de lijn is niet noodzakelijk gelijk aan 1. Zie Figuur 14 voor het resultaat.

Voor de volledigheid is er een puntsgewijze 95%-betrouwbaarheidsband ge-



Figuur 14: De staarten liggen niet altijd in het 95%-betrouwbaarheidsgebied.

tekening. We zien links dat de datapunten boven de lijn liggen. Dit duidt aan dat de verdeling van de suprema links een dunnere staart heeft dan de Gumbel verdeling. Aan de rechterkant vallen de datapunten juist onder de lijn. Dit duidt ook op een dunnere staart rechts. Hier zien we voor $n = 10$ een bijna horizontale lijn ter hoogte van 1.5. Blijkbaar zijn er veel datapunten die een herschaalde supremumafstand hebben rond de 1.5. Om het oog wat te ondersteunen hebben we bijgehouden hoeveel punten buiten de band vallen. Dit is met 95% betrouwbaarheid gedaan, dus we verwachten dat 50 punten niet tussen de twee lijnen vallen. Omdat we niet precies de Gumbel verdeling hebben, zal dit aantal waarschijnlijk groter zijn. Het aantal punten buiten de band zijn

n	10	100	1000	100.000
aantal	794	447	314	202

5.5.3 Herhaling van het experiment, vaste ϵ_n

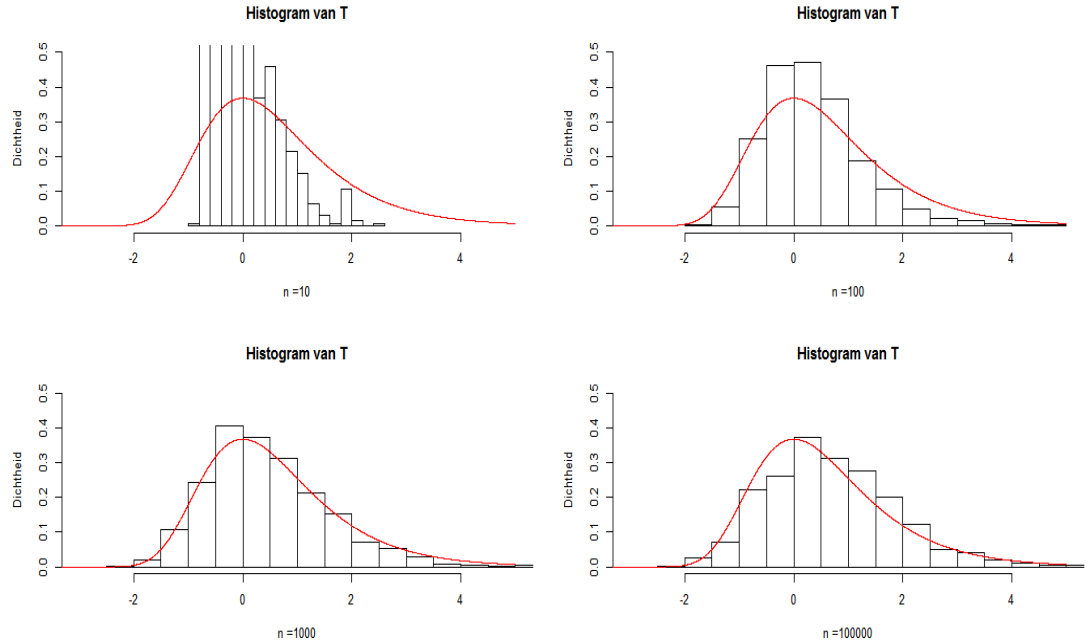
We hebben een vaste ϵ_n bekeken, maar we hebben niet echt een goede benadering. Een keuze voor ϵ_n die het beter doet is een ϵ_n die daadwerkelijk van n afhangt. Een optimale subinterval vinden is een interessante vraag, waar nog niet over nagedacht is. Omdat we weten uit Stelling 4.1 dat ϵ_n niet sneller naar nul mag gaan dan $n^{-1/3}(\log n)^{-2/3}$, kiezen we ϵ_n wat groter:

$$\epsilon_n = n^{-1/4}(\log n)^{-1/3}.$$

We laten de resultaten zien, weer voor $n = 10, 100, 1000, 10^5$. Hierbij horen bij benadering $\epsilon_n = 0.43; 0.19; 0.09; 0.02$. Hier zien we terug waarom het fout

ging bij kleine n . Toen, met $\epsilon_n = 0.10$, bekeken we een interval dat te ruim was, waardoor ook de inconsistentie een rol speelde.

De histogrammen in Figuur 15 zien er voor $n = 10, 100$ beter uit dan eerst, en we zitten meteen dicht op de Gumbel dichtheid. Voor grotere n neemt zoals verwacht de nauwkeurigheid toe.



Figuur 15: Het histogram lijkt voor elke geplote n op de Gumbel dichtheid.

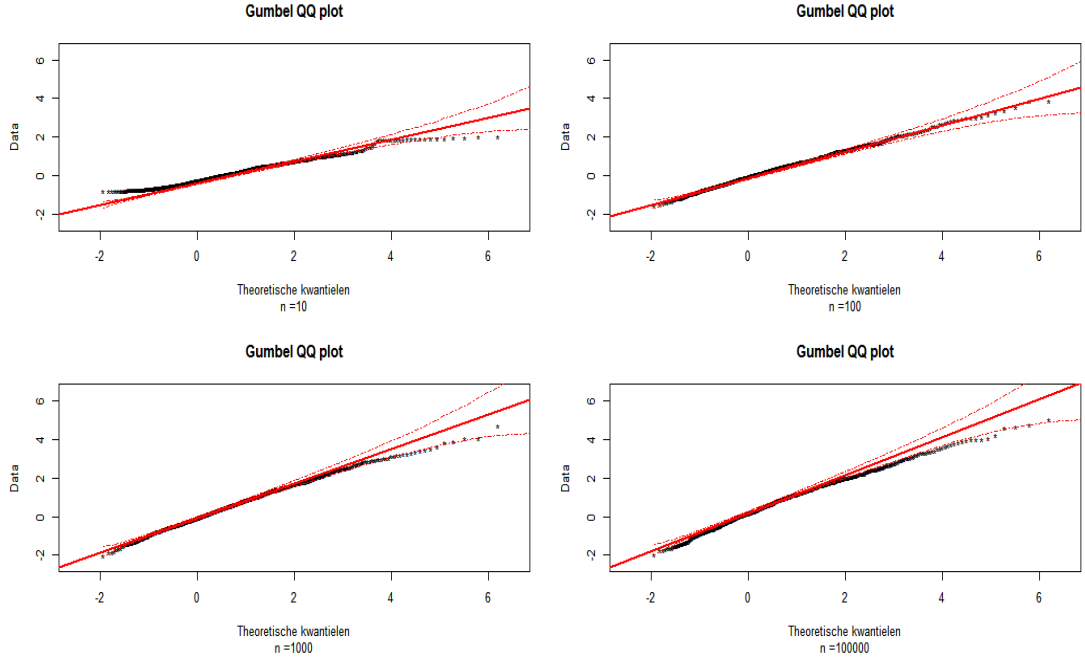
Voor de volledigheid voegen we de qq-plots erbij, in Figuur 16. We zien in de plots meteen dat de helling van de rechte lijnen niet gelijk is aan 1. Dit wijst erop dat T_n niet een standaard Gumbel verdeling heeft. Uit de qq-plot voor $n = 100.000$ zien we wel een helling van ongeveer 1. Dit betekent dat T_{10^5} redelijkerwijs een standaard Gumbel verdeling heeft. Merk op dat voor $n = 1000$, we heel goed tussen de banden blijven. Dit zou betekenen dat er sprake is van een verschoven Gumbel verdeling in dit geval. Hier zijn het aantal punten buiten de band

n	10	100	1000	100.000
aantal	861	663	10	165

5.5.4 Betrouwbaarheidsbanden

We hebben gezien dat pas bij grotere n (denk aan $n \geq 1000$) we een Gumbel verdeling redelijk benaderen. De algemene constructie van een betrouwbaarheidsband op basis van Gumbel vinden we terug in Sectie 4. Hier gebruiken we het gegeven dat f een afgeknotte exponentiële dichtheid is. We willen twee functies L_n, U_n waarvoor geldt

$$P(L_n(t) \leq f(t) \leq U_n(t), \forall t \in [\epsilon_n, 1 - \epsilon_n]) = p.$$



Figuur 16: De staarten liggen steeds meer in het 95%-betrouwbaarheidsgebied.

In genoemde Sectie staan ook de uitdrukkingen voor de functies die hieraan voldoen, namelijk

$$L_n(t) = \hat{f}_n(t) - |2f'(t)f(t)|^{1/3} \cdot H^{-1}(x_p), \quad U_n(t) = \hat{f}_n(t) + |2f'(t)f(t)|^{1/3} \cdot H^{-1}(x_p).$$

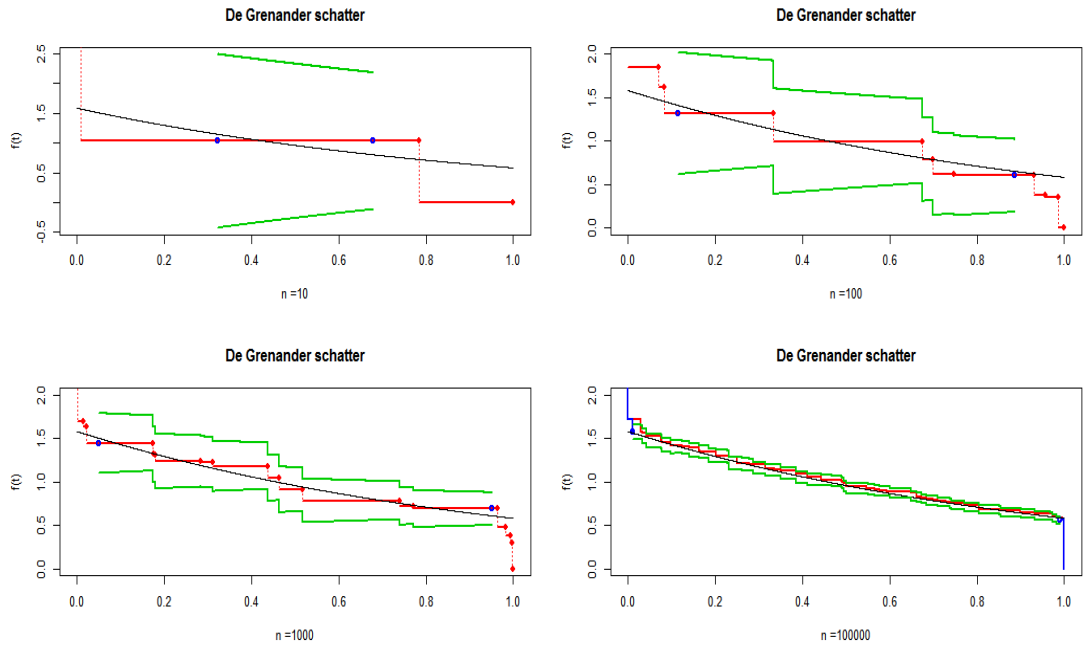
Laat nu $p = 0.95$, dan maken we een 95%-betrouwbaarheidsband op basis van een Gumbel verdeling, en vinden we $x_{0.95} = 2.97$ en $H^{-1}(x_{0.95}) = 0.10$. We krijgen

$$L_n(t) = \hat{f}_n(t) - 0.10 h(t), \quad U_n(t) = \hat{f}_n(t) + 0.10 h(t),$$

met $h(t)$ zoals in (11)

De functies L_n, U_n worden functies die symmetrisch rond de Grenander schatter liggen. Hiervoor geldt dus dat als n naar oneindig gaat, dat deze banden de echte dichtheid omvatten met kans 95%. Om te kijken of voor alle t in $[\epsilon_n, 1 - \epsilon_n]$ de functie f tussen de twee functies ligt, moeten we in principe alle t bekijken. Volgens Lemma 4.2 hoeven we veel minder te doen. We houden bij hoeveel realisaties van het herschaalde supremum, zoals in voorgaande paragraaf, groter is dan het herschaalde 95% kwantiel. De betrouwbaarheidsband hangt natuurlijk ook nog af van ϵ_n . We vervolgen met de keuze van $\epsilon_n = n^{-1/4}(\log n)^{-1/3}$. In Figuur 17 is een dergelijke betrouwbaarheidsband geplot in het groen, merk dus op dat de betrouwbaarheidsband tot en met een afstand ϵ_n van de randen loopt. Let op dat in het eerste plaatje de y -as anders is.

De vraag die we willen beantwoorden is: “Hoe vaak valt f tussen de 95%-betrouwbaarheidsgrenzen?” Als het goed is, moet het percentage dat f tussen die grenzen ligt ongeveer 95% bedragen. We gaan dit met simulatie na. We



Figuur 17: De banden worden steeds nauwer, en het interval ruimer.

houden $m = 1000$ nog aan en bekijken de resultaten.

n	10	100	1000	100.000
ϵ_n	0.43	0.19	0.09	0.02
percentage	100%	99.6%	97.2%	95.3%

Het is duidelijk te zien dat het overdekkingspercentage monotoon wordt benaderd met de keuze van ϵ_n . We zien bij $n = 10$ dat f altijd tussen de banden blijft, dit komt door het krappe interval $[0.43, 0.57]$ welke we bekijken. Het is nog onbekend waarom we consequent een te hoog percentage vinden. Desalniettemin vinden we een convergentie naar 95%.

Het is ook interessant om te zien hoe deze percentages zich gedragen voor vaste $\epsilon_n = 0.1$. Hier zijn de resultaten

n	10	100	1000	100.000
percentage	93.5%	94.2%	98.2%	97.1%

Hier zien we dat we voor kleine n niet voldoen aan 95%, maar dat we er daarna overheen schieten en weer te vaak overdekken. Waarom dit fenomeen voorkomt is onduidelijk, maar het is zeker het gevolg van de keuze voor ϵ_n .

6 Conclusie

We hebben wat eigenschappen van de Grenander schatter aangestipt, denk aan de vorm van de functie en (in)consistentie. Er is al veel onderzoek gedaan, maar resultaten over de supremumfout bleven uit. Met het artikel van Durot, Kulikov en Lopushaä [4] is ook op dit deel licht geworpen. In dit verslag deden we onderzoek naar de praktische toepasbaarheid en gebreken van Stelling 4.1. Omdat we werken met benaderingen, is het niet anders dan de conclusie in vage termen als ‘dichtbij’ en ‘goed’ te formuleren. Uit het oog van de toepasser proberen we toch zo goed mogelijk de conclusie van dit verslag te vermelden.

Ten eerste is gebleken dat de Grenander schatter, voor eindige n , dicht zit op de echte functie. Toch is er altijd een afstand te vinden, voor toenemende n . Deze is in de limiet van n naar oneindig standaard Gumbel verdeeld, mits we het supremum herschalen. Ook voor eindige n is de herschaalde supremumfout een stochast, zeg T_n die een onbekende verdeling heeft. We hebben gezien dat voor toenemende n deze verdeling steeds meer op een Gumbel verdeling lijkt. De histogrammen lijken steeds meer op de Gumbel dichtheid. De qq-plots geven ook aan dat de kwantielen meer in het betrouwbaarheidsgebied liggen voor toenemende n . Maar ook hier is geen sluitend bewijs dat T_n een standaard Gumbel verdeling heeft. Het is ook duidelijk dat de convergentie ook lijkt af te hangen van de waarde van ϵ_n . Voor een toepasser zou $n = 1000$ al redelijk zijn om de Gumbel verdeling te gebruiken.

De betrouwbaarheidsbanden lijken meer succes op te leveren. We zagen dat voor kleine n het overdekkingspercentage te hoog is. Dit is niet erg als de toepasser toch al conservatief is en zekerheid prefereert. Voor toenemende n convergeert het percentage gelukkig wel naar de gewenste betrouwbaarheid. Let op dat het wel uitmaakt welke ϵ_n we nemen. Ondanks alles is $n = 1000$ een voldoende grote steekproef om op basis van Gumbel te werken.

Al met al zien we een convergentie naar de standaard Gumbel verdeling, alhoewel deze niet heel snel gaat. Verder maakt het wel degelijk uit welk interval we nemen, maar of dit alleen zo is voor $n \leq 100.000$ of dat dit altijd zo is voor elke eindige n is niet duidelijk. Verder onderzoek zou ook hier antwoord op kunnen geven.

Referenties

- [1] John Aldrich. R. A. Fisher and the making of maximum likelihood 1912–1922. *Statist. Sci.*, 12(3):162–176, 1997.
- [2] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical inference under order restrictions. The theory and application of isotonic regression*. John Wiley & Sons, London-New York-Sydney, 1972. Wiley Series in Probability and Mathematical Statistics.
- [3] Cécile Durot. On the $\mathbb{L}_p$ -error of monotonicity constrained estimators. *Ann. Statist.*, 35(3):1080–1104, 2007.
- [4] Cécile Durot, Vladimir N. Kulikov, and Hendrik P. Lopuhaä. The limit distribution of the l_∞ -error of grenander-type estimators. *The Annals of Statistics*, 2012.
- [5] Ulf Grenander. On the theory of mortality measurement. I. *Skand. Aktuarietidskr.*, 39:70–96, 1956.
- [6] Ulf Grenander. On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, 39:125–153 (1957), 1956.
- [7] Piet Groeneboom, Gerard Hooghiemstra, and Hendrik P. Lopuhaä. Asymptotic normality of the L_1 error of the Grenander estimator. *Ann. Statist.*, 27(4):1316–1347, 1999.
- [8] Piet Groeneboom and Hendrik P. Lopuhaä. Isotonic estimators of monotone densities and distribution functions: basic facts. *Statist. Neerlandica*, 47(3):175–183, 1993.
- [9] Anders Hald. *A history of mathematical statistics from 1750 to 1930*. Wiley Series in Probability and Statistics: Texts and References Section. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.
- [10] Gerard Hooghiemstra and Hendrik P. Lopuhaä. An extremal limit theorem for the argmax process of Brownian motion minus a parabolic drift. *Extremes*, 1(2):215–240, 1998.
- [11] Vladimir N. Kulikov and Hendrik P. Lopuhaä. Asymptotic normality of the L_k -error of the Grenander estimator. *Ann. Statist.*, 33(5):2228–2255, 2005.
- [12] B. L. S. Prakasa Rao. Estimation of a unimodal density. *Sankhyā Ser. A*, 31:23–36, 1969.
- [13] Tim Robertson, F. T. Wright, and R. L. Dykstra. *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester, 1988.
- [14] Adriaan C. Zaanen. *Continuity, integration and Fourier theory*. Universitext. Springer-Verlag, Berlin, 1989.

A Functie code

```

library(fdrtool)
library(evd)
library(car)

##### functie runif ZONDER duplicates #####
## bijgevoegd omdat runif bij grote getallen duplicates teruggeeft
runiform <- function(numb=1,min=0,max=1){
U=runif(numb,min,max)
test=duplicated(U)
while( any(test) ){ # haalt alle "dubbele" weg
  U[test]=runif(sum(test),min,max)
  test=duplicated(U)
}
return(U)
}

##### functie trekking #####
trekking <- function(n=1,Lambda=lambda,Max=M){
## Geeft n realisaties uit de modified exponential density
## Waarbij geldt  $X_i \in [0, Max]$   $i=1..n$ 
fact=pexp(Max,Lambda)
# trekkingen
U=runiform(n,0,1)
#  $T_n$  bestaat uit n trekkingen uit "afgeknotte exp"
#  $F(t)=(1-\exp(-\Lambda t))/(1-\exp(-\Lambda Max))$ 
T=qexp(U*fact,Lambda); # inv-cdf

return(T) # 1xn
}

#####

##### functie h(t) #####
absh <- function(x){
h=2^(1/3)*lambda*exp(-2/3*lambda*x)/fact^(2/3) #is al absoluut
return(h) #1xlength(x)
}

#####

##### functie aexp #####
aexp <- function(x,L=lambda,Max=M){
h=dexp(x,L)/pexp(Max,L)
h[x>Max]=0
return(h) #1xlength(x)
}

#####

##### piecewise functie hat{f}_n #####
gren <- function(x,xco=xgn,yco=ygn){

y=0*x;
yco=c(yco[1],yco) # check dat gren(xco[i])==yco[i-1]

for(j in 1:length(x))
{
  i=1
  while(i<=length(xco) && x[j]>=0 && x[j]<=M)#blijf binnen domein en bereik
  {
    if(x[j]<=xco[i]){ y[j]=yco[i]; break }#loop tot x>xco
  }
}
}

```

```

    i=i+1 #uitvoeren alleen als vorige regel niet waar is
  }
}
return(y) #1xlength(x)
}
#####

##### functie afstand tussen hat{f} en f #####
afstand <- function(an=eps,xco=xgn,yco=ygn){
  distance=NA*1:length(xco)
  temp=rep(NA,length(xco))
  temp2=temp
  xval=temp
  #structuur: check tussen an en M-an en daarbuiten NA
  #evaluatie: afstand als t=an, dan t=M-an, dan vergelijk tussen tau(i) en tau(i-1)
  #supremum vanaf eps tot en met M-eps
  for(i in 2:(length(xco)))
  {
    ifelse(xco[i]>=an & xco[i-1]<=(M-an),
    {
      x1=max(an,xco[i-1])
      xr=min(M-an,xco[i])
      xval[i-1]=x1
      xval[i]=xr # ==xval[i-1] als xco[i]>=an
      temp[i]=x1 ### debug
      temp2[i]=xr ### debug
      h1=absh(x1)
      hr=absh(xr)
      tau_l = abs(gren(xr,xco,yco)-aexp(x1))/h1
      tau_r = abs(gren(xr,xco,yco)-aexp(xr))/hr
      distance[i]=max(tau_l,tau_r)
    },{
      distance[i]=NA
    })
  }
  distance=na.omit(distance)
  list(supremum=distance,xvalues=xval,debug=rbind(temp,temp2))
}
#####

##### functie herschaalde sup-afstand #####
## gegeven de sup uitdrukking, herschaal deze tot de stochast in artikel Durot et al
rescale <- function(sup, inverse=FALSE){
  herschaal=0
  if(!inverse){
    h1=sup
    h2=(n/log(n))^(1/3)*h1 - mun ## let op! n,mun uit de parameter lijst
    herschaal=log(n)*h2
  }
  #---- inverse ----#
  herschaalinv=0*sup
  if(inverse){
    kwant=sup
    herschaalinv = (kwant/log(n) + mun)*(log(n)/n)^(1/3)
  }
  return(herschaal*(inverse==FALSE)+herschaalinv*(inverse==TRUE)) #1x1
}
#####

```

```

###----- functies PLOTS ----###
devoff<-function(){
#sluit huidige plot
if(dev.cur()!=1){ invisible(dev.off()) }
}

plotkernel<-function(Tn=NULL, aant=NULL, Lambda=lambda, Max=M, close=TRUE)
{
  if(close){devoff()}
  if(is.null(Tn)){
    Tn=trekking(aant, Lambda, Max)
  }
  fact=pexp(Max, Lambda)
  # Plot van de kernschatter en de echte dichtheid
  Tnplot=density(Tn) #kernschatter van Tn
  plot(Tnplot,xlim=c(0,Max),ylim=c(0,1.1*dexp(0,Lambda)/fact),
  xlab='t',ylab='Dichtheid f(t)',main='Kernschatter tegen echte dichtheid')
  # echte dichtheid
  xseq=seq(from=0,to=Max,by=Max*1e-4);
  lines(xseq,dexp(xseq,Lambda)/fact,col='red')
}

plotFn<-function(Fn=NULL, aant=NULL, Lambda=lambda, Max=M, close=TRUE)
{ # plot de ecdf van een steekproef uit aexp
  if(close){devoff()}
  # als Fn==NULL maakt een willekeurige Tn en Fn
  if(is.null(Fn)){
    Tn=trekking(aant, Lambda, Max)
    Fn=ecdf(Tn)
  }
  # LCM van Fn inclusief 0 en 1
  Gn=gcmlcm(c(0,sort(Tn),Max),c(0,Fn(sort(Tn)),1),type="lcm")

  # Plot de ecdf van Tn
  plot(Fn,ylim=c(0,1),xlim=c(0,Max),verticals=TRUE, main='De ecdf van n trekkingen',
  ylab='Femp(x)',sub='Zwart: De ecdf, Rood: De kleinste Concave majorant')
  lines(Gn$x.knots, Gn$y.knots, col=2, lwd=2)
}

plotmle<-function(Gren=NULL, aant=NULL, Lambda=lambda, Max=M,
  ylim_vect=NULL,optie="1",close=TRUE,suparea=FALSE,xlab='t',ylab='f(t)')
# De Grenander schatter met een mogelijkheid om de echte dichtheid niet te plotten
{
  if(close){devoff()}
  if(is.null(Gren)){
    Tn=trekking(aant, Lambda, Max)
    Fn=ecdf(Tn)
    Gren=gcmlcm(c(0,sort(Tn),Max),c(0,Fn(sort(Tn)),1),type="lcm")
  }
  xn=Gren$x.knots;
  sn=Gren$slope.knots;
  xseq=seq(from=0,to=Max,length=10000);
  if(is.null(ylim_vect)){ylim_vect=c(0,1.1*Lambda/fact)} # stel ylim in, VECTOR
# voor de overzichtelijkheid 2 opties voor groot en klein aantal sprongpunten
if(length(sn)<=20)

```

```

{ # als weinig sprongpunten (i.e. K klein)
plot(xn,gren(xn,xn,sn),col='red',ylim=ylim_vect,type="n",
lwd=2,main='De Grenander schatter',xlab=xlab,ylab=ylab)
for(i in 1:length(xn))
{
segments(xn[i],sn[i],xn[i+1],sn[i],lwd=2,col=2)
points(xn[i+1],sn[i],col=2,pch=16)
segments(xn[i+1],sn[i],xn[i+1],sn[i+1],col=2,lty="dotted")
}
lines(xseq,aexp(xseq),type=optie) #plot als optie!="n"
if(suparea){
points(c(eps,Max-eps),gren(c(eps,Max-eps),xn,sn),col='blue',lwd=2)
}
}
if(length(xn)>20)
{ # als veel sprongpunten (i.e. K groot)
#plot MLE(f) tegen f
plot(xn,gren(xn,xn,sn),col='red',ylim=ylim_vect,type="s",
lwd=2,main='De Grenander schatter',xlab=xlab,ylab=ylab)
if(suparea){
# interval dat niet meegenomen wordt in supremum
lines(c(Max-eps,xn[xn>(Max-eps)]),gren(c(Max-eps,xn[xn>(Max-eps)]),xn,sn)
,col='blue',lwd=2,type="s")
lines(c(xn[xn<eps],eps),gren(c(xn[xn<eps],eps),xn,sn)
,col='blue',lwd=2,type="s")
points(c(eps,Max-eps),gren(c(eps,Max-eps),xn,sn),col='blue',lwd=2)
}
lines(xseq,aexp(xseq),type=optie) #plot als optie!="n"
}
}

plotline<-function(T=NULL,close=FALSE,xlim=c(-3,5),ylim=c(-3,5),xlab="Theoretische
kwantielen",
ylab="Data",main="Gumbel QQ plot",sub=sub,pch='*'){
# plot de kwantielen van T tegen Gumbel kwantielen
if(close){devoff()}
yvar=seq(1/length(T),1-1/length(T),length=length(T)) # de kwantielen

qqplot(qgumbel(yvar),quantile(T,yvar),xlim=xlim,ylim=ylim,pch=pch,ylab=ylab,xlab=xlab,
main=main,sub=sub,lwd=2) # de bouwsteen van de uiteindelijke qqplot

muhat=median(T)-( -log(log(2)) );
sigmahat=(quantile(T,prob=3/4)-quantile(T,prob=1/4))/(qgumbel(3/4)-qgumbel(1/4));
abline(a=muhat,b=sigmahat,col=2,lwd=2) #referentielijn, y=x

# de 95% betr.band voor de kwantielen
g=qgumbel(yvar)
x.hat<-muhat+sigmahat*g
se=sigmahat/dgumbel(g)*sqrt(yvar*(1-yvar)/length(T))
lines(g,x.hat+2*se,lty=4,col=2)
lines(g,x.hat-2*se,lty=4,col=2)

#teller voor aantal punten buiten band
fallen=quantile(T,yvar)<x.hat-2*se|quantile(T,yvar)>x.hat+2*se
print(paste("aantal buiten band = ",sum(fallen),sep="")) #teller voor aantal punten buiten
band
print(muhat);print(sigmahat)
}

```

```
plotdens<-function(T=NULL,M=m,close=TRUE){ #dichtheid Gumbel vs kernschatter T
  if(close){devoff()}
  plot(seq(min(T),max(T),0.01),dgumbel(seq(min(T),max(T),0.01)),type="l",
xlab="X ~ Gumbel(0,1)",ylab="p(x)",col=2,main="Zwart:kernschatter")
  lines(density(T),lwd=2)
}
```

```
#####----- functie betrouwbaarheids band -----#####
confband<-function(x=xgn,y=ygn,supr=sup,betr=0.95,loc=FALSE,debug=FALSE){
t=seq(eps,M-eps,length=50000)
kw1=qgumbel(betr) #p-kwantiel van Gumbel
Hinv_kw1=rescale(kw1,inverse=TRUE) #inverse herschaal
h=absh(t) # = |2 f' f|^{\frac{1}{3}} (t)
## de band, Un bovengrens en Ln ondergrens
Un=gren(t,x,y)+h*Hinv_kw1
Ln=gren(t,x,y)-h*Hinv_kw1
```

```
list(tvalues=t,upperbound=Un,lowerbound=Ln,testconf=Test,location=argt)
}
```

```
#####
```

```
##### END #####
```

B Uitvoer code

```

## LIBRARY ##
library(fdrtool)
library(evd)
library(car)

##### PARAMETERS #####
lambda=1;          # de lambda van de exponentiele verdeling
M=1;               # de afkapping waarvoor geldt t in [0,M]
n2=c(10,100,1e3,1e5); # aantal trekkingen
m=1e3;             # aantal realisaties Tn
fact=pexp(M,lambda) # geconditioneerd op Tn<M
theta = 4^(1/3)/0.7022 # constante
kappa = 2^(1/3)*2.3381 # constante
Cf = (-6*exp(-1/3*lambda*M)+6)/(fact)^(1/3) # van 0 naar M geïntegreerd
mun2 = 1 - kappa/(2^(1/3)*(log(n2))^(2/3)) +
      1/log(n2)*(1/3*log(log(n2))+log(theta*Cf))
#eps2 = rep(0.1,4) # sup over interval [eps,M-eps]
eps2 = n2^(-1/4)*log(n2)^(-1/3)

betr=0.95

##### Grenander Schatter
par(mfrow=c(2,2))
for(i in 1:4){
  Tn=trekking(n2[i])
  # Een schatting voor Gn(y)=P(Xn<=y) #
  Fn=ecdf(Tn) # empirische cdf
  # LCM van Fn inclusief 0 en 1 #
  Gn=gcm1cm(c(0,sort(Tn),M),c(0,Fn(sort(Tn)),1),type="lcm")
  # De afgeleide van Gn #
  xgn=Gn$x.knots
  ygn=Gn$slope.knots
  mun=mun2[i]
  eps=eps2[i]
  sup=afstand(eps,xgn)
  supremum=max(sup$supr)
  n=n2[i]

  ## De betrouwbaarheidsband
  band=confband(supr=sup,betr=0.95,loc=FALSE)
  stap=band$tval
  Upp=band$upper
  Low=band$lower
  #onderdrukt wetenschappelijke notatie
  options(scipen=99);waarde=paste("n =",n2[i],sep="");options(scipen=0)
  #andere assen voor i==1
  if(i==1)plotmle(Gren=Gn,ylim_vect=c(min(Low),max(Upp)),suparea=TRUE,close=FALSE,
    optie='n',xlab=waarde)
  if(i!=1)plotmle(Gren=Gn,ylim_vect=c(0,2),suparea=TRUE,close=FALSE,
    optie='n',xlab=waarde)
  lines(stap,Upp,col=3,lwd=2,type='s')
  lines(stap,Low,col=3,lwd=2,type='s')
}

##### Vergelijken van dichtheden
T=matrix(0,4,m)

for(i in 4){ # loopt door verschillende n
  for (k in 1:m){ # berekent m suprema

```

```

n=n2[i]
Tn=trekking(n)
mun=mun2[i]
# Een schatting voor  $G_n(y)=P(X_n \leq y)$  #
Fn=ecdf(Tn) # empirische cdf

# LCM van Fn inclusief 0 en 1 #
Gn=gcm1cm(c(0,sort(Tn),M),c(0,Fn(sort(Tn)),1),type="lcm")

# De afgeleide van Gn #
xgn=Gn$x.knots
ygn=Gn$slope.knots

sup=afstand(eps2[i],xgn)
sup_xval=(sup$xval)

T[i,k]=rescale(max(sup$supr)) # waarde van het supremum
}
print(ks.test(T[i,],"pgumbel",0,1)) #kolmogorov-smirnov toets
}

## de dichtheid van T onderzoeken
par(mfrow=c(2,2))
for(i in 1:4){
#onderdrukt wetenschappelijke notatie
options(scipen=99);waarde=paste("n =",n2[i],sep="");options(scipen=0)
hist(T[i,],20,prob=TRUE,ylim=c(0,0.5),xlim=c(-3,5),
  main="Histogram van T",xlab=waarde,ylab="Dichtheid" )
lines(seq(-5,5,0.001),dgumbel(seq(-5,5,0.001)),lwd=1,col=2)
}

##### QQ Plots
par(mfrow=c(2,2))
for(i in 1:4){
#onderdrukt wetenschappelijke notatie
options(scipen=99);waarde=paste("n =",n2[i],sep="");options(scipen=0)
plotline(T[i,],xlim=c(-2.5,6.5),ylim=c(-2.5,6.5),sub=waarde)
}

##### Index Plots
par(mfrow=c(2,2))
for(i in 1:4){
plot(T[i,],pch='*',ylim=c(min(T)-0.05,max(T)+0.05))
abline(h=max(T[i,]),col=2);abline(h=min(T[i,]),col=2)
abline(h=qgumbel(0.999),col=3,lty=4);abline(h=qgumbel(0.001),col=3,lty=4)
}

##### Betrouwbaarheidsband
test=rep(0,4)
eps2
for(i in 1:4){
test[i] = 1-mean( T[i,] > qgumbel(betr) ) # fractie binnen de band
print(test[i])
}

```