



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

Verwachte tijd tot verwachting
(Engelse titel: Expected time till expectancy)

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

Sofie van den Hoogen

Delft, Nederland
Juli 2014



BSc verslag TECHNISCHE WISKUNDE

“Verwachte tijd tot verwachting”

(Engelse titel: “Expected time till expectancy”)

Sofie van den Hoogen

Technische Universiteit Delft

Begeleider

Prof. dr. ir. G. Jongbloed

Overige commissieleden

Prof. dr. ir. A.W. Heemink

Drs. E.M. van Elderen

Juli, 2014

Delft

Inhoudsopgave

Inleiding	7
1 Current duration model	9
1.1 Beschrijving van het model	9
1.2 Verdeling tijdsduur gegeven vrouw in steekproef	12
1.3 Verdeling current duration	15
1.4 Verdeling tijdsduur	19
2 Niet-parametrische maximum likelihood schatter	21
2.1 Parametrisch versus niet-parametrisch schatten	21
2.2 Constructie	23
2.3 Consistentie op $(0, \infty)$	25
2.4 Inconsistentie rond 0	28
2.5 Simulatie inconsistentie	32
3 Penalized maximum likelihood schatter	35
3.1 Constructie	35
3.2 Parameter α	38
3.3 Simulatie optimale parameter α	39
4 Alternatieve schatter	43
4.1 Constructie	43
4.2 Consistentie	45
4.3 Simulatie optimale parameter h_n	50
5 Analyse data	53
5.1 Methode 1	54
5.2 Methode 2	56
5.3 Methode 3	59
6 Conclusie	61
7 Appendix	63
7.1 Parameter α	63
7.1.1 Standaard halve normale verdeling	63
7.1.2 Standaard exponentiële verdeling	65
7.2 R-codes	67
7.2.1 Simulatie inconsistentie (2.5)	67
7.2.2 Simulatie optimale parameter α (3.3)	68

7.2.3	Simulatie optimale parameter h_n (4.3)	69
7.2.4	Methode 1 (5.1)	70
7.2.5	Methode 2 manier 1 (5.2)	72
7.2.6	Methode 2 manier 2 (5.2)	74
7.2.7	Methode 3 (5.3)	76
7.2.8	Functies	78

Inleiding

Medici zijn geïnteresseerd in de tijdsduur die een vrouw nodig heeft om zwanger te worden. De tijdsduur die een vrouw nodig heeft om zwanger te worden, wordt gezien als een belangrijk instrument voor het vaststellen van de vruchtbaarheid van de desbetreffende vrouw [1]. In dit verslag wordt gezocht naar de verdeling van deze tijdsduur.

De meest voorkomende methoden om tot een verdeling van de tijdsduur te komen zijn de prospectieve en retrospectieve cohort studie. De prospectieve cohort studie volgt de vrouw vooruit in de tijd. De vrouw wordt gevolgd vanaf het moment waarop zij stopt met het gebruik van anticonceptie. Eindpunt van de studie is het tijdstip waarop de vrouw daadwerkelijk in verwachting is. Het verkrijgen van de data wordt in het geval van een prospectieve cohort studie bemoeilijkt, doordat het tijdstip waarop besloten wordt om zwanger te willen worden per vrouw varieert. Voor de één is dit kort van tevoren, voor de ander lang van tevoren. De retrospectieve cohort studie volgt de vrouw terug in de tijd. Beginpunt van de studie is nu het moment waarop de vrouw in verwachting raakt. Gekeken wordt wanneer de vrouw is gestopt met het gebruik van anticonceptie. Het interpreteren van een retrospectieve cohort studie wordt bemoeilijkt, doordat onvruchtbare vrouwen niet in de studie worden opgenomen.

Het mag duidelijk zijn dat beide methoden moeilijkheden met zich meebrengen. In het artikel van Keiding [1] wordt een alternatieve methode aangedragen: het current duration model. Het current duration model zit als volgt in elkaar. Aan een vrouw wordt gevraagd of zij zwanger probeert te worden, en zo ja sinds wanneer. Deze tijdsduur wordt de current duration genoemd. Het blijkt mogelijk te zijn om op basis van de verdeling van deze current duration, de verdeling van de tijdsduur die een vrouw nodig heeft om zwanger te worden vast te stellen.

In hoofdstuk 1 wordt het current duration model uitgebreid beschouwd en geanalyseerd. Vervolgens worden in hoofdstuk 2, 3 en 4 verschillende schatters beschouwd. Tenslotte worden in hoofdstuk 5 deze schatters toegepast op data om zo de verdeling van de tijdsduur die een vrouw nodig heeft om zwanger te worden te verkrijgen.

1 — Current duration model

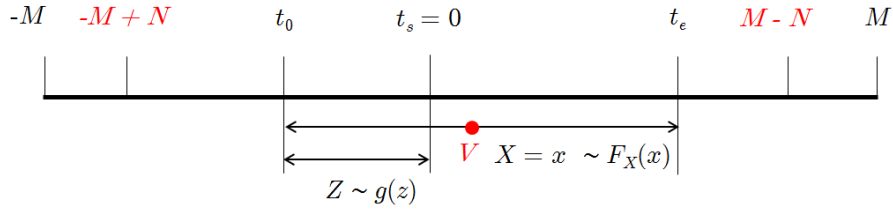
In dit hoofdstuk wordt het current duration model, zoals beschreven in de inleiding, geanalyseerd. Allereerst wordt het current duration model uitgebreid beschouwd. Vervolgens wordt de verdeling van de tijdsduur die een vrouw nodig heeft om zwanger te worden, gegeven dat de vrouw in de steekproef voorkomt, afgeleid. Daarna wordt de verdeling van de current duration afgeleid. Tot slot wordt de verdeling van de tijdsduur die een vrouw nodig heeft om zwanger te worden afgeleid.

1.1 Beschrijving van het model

In dit verslag gezocht naar de verdeling van de tijdsduur die een vrouw nodig heeft om zwanger te worden. Hiertoe wordt het specifieke current duration model, dat staat beschreven in het artikel van Keiding [1], beschouwd. In figuur 1.1 is dit model schematisch weergegeven door middel van een tijdlijn. De tijdlijn loopt van $-M$ tot M , waarbij M staat voor een groot positief getal. Merk op dat het tijdstip t_s gelijk is gesteld aan 0. Laat het punt V , het tijdstip $-M + N$ en het tijdstip $M - N$ buiten beschouwing.

Het tijdstip t_0 toont het tijdstip waarop een vrouw besluit zwanger te willen worden. Denk hierbij aan het stoppen met het gebruik van anticonceptie zoals de pil of het spiraaltje. Het tijdstip t_e toont het tijdstip waarop deze vrouw daadwerkelijkheid zwanger wordt. Het verschil tussen de tijdstippen t_0 en t_s wordt weergegeven met de stochastische variabele X . De stochastische variabele X geeft de tijdsduur weer die een vrouw nodig heeft om zwanger te worden. De bijbehorende cumulatieve verdelingsfunctie van X wordt aangeduid met F_X . Het uiteindelijke doel is om deze cumulatieve verdelingsfunctie F_X te bepalen.

Het tijdstip t_s toont het tijdstip waarop de steekproef wordt afgenomen. Op dit tijdstip wordt de volgende vraag aan de vrouw gesteld: Probeert u zwanger te worden, en zo ja sinds wanneer? Enkel indien deze vraag met het antwoord ja wordt beantwoord, wordt de desbetreffende vrouw in de steekproef opgenomen. Het verschil tussen de tijdstippen t_0 en t_s wordt weergegeven met de stochastische variabele Z . De stochastische variabele Z is de current duration, en geeft slechts een fractie van de tijdsduur weer die een vrouw nodig heeft om zwanger te worden. De bijbehorende kansdichtheid van Z wordt aangeduid met g . Echter kan worden aangetoond dat uit deze kansdichtheid g de cumulatieve verdelingsfunctie F_X kan worden verkregen [2]. Dit zal in de volgende paragrafen worden aangetoond.



t_0 = tijdstip zwanger willen worden

t_e = tijdstip zwanger worden

t_s = tijdstip steekproef

X = tijdsduur nodig om zwanger te worden

Z = fractie van tijdsduur nodig om zwanger te worden (current duration)

Figuur 1.1: Schematische weergave van het current duration model.

Verder kan worden aangetoond dat de kans dat een vrouw in de steekproef voorkomt groter wordt, naarmate de tijdsduur die een vrouw nodig heeft om zwanger te worden groter wordt. Hiertoe wordt figuur 1.1 nogmaals beschouwd. Laat nu echter het punt V , het tijdstip $-M + N$ en het tijdstip $M - N$ niet meer buiten beschouwing. Merk op dat punt V het middelpunt vormt van de lijn $X = x$. Hierbij wordt aangenomen dat het punt V uniform is verdeeld op het interval $[-M + N, M - N]$. Hieronder is deze aanname nogmaals kort genoteerd.

$$V \sim f_V, \text{ waarbij geldt } f_V(v) = \begin{cases} \frac{1}{2(M - N)} & \text{als } -M + N \leq v \leq M - N \\ 0 & \text{als anders} \end{cases}$$

Op basis van deze aanname kan de kans dat een vrouw in de steekproef voorkomt, gegeven dat $X = x$, worden berekend. Onderstaand is deze berekening genoteerd.

$$\begin{aligned} P(\text{vrouw in steekproef} \mid X = x) &= P\left(0 \in \left[V - \frac{x}{2}, V + \frac{x}{2}\right]\right) \\ &= P\left(V \in \left[-\frac{x}{2}, \frac{x}{2}\right]\right) \\ &= P\left(-\frac{x}{2} \leq V \leq \frac{x}{2}\right) \\ &= \int_{v=-\frac{x}{2}}^{\frac{x}{2}} f_V(v) dv \end{aligned}$$

$$\begin{aligned}
&= \int_{v=-\frac{x}{2}}^{\frac{x}{2}} \frac{1}{2(M-N)} dv \\
&= \left[\frac{v}{2(M-N)} \right]_{v=-\frac{x}{2}}^{\frac{x}{2}} \\
&= \frac{\frac{x}{2}}{2(M-N)} + \frac{\frac{x}{2}}{2(M-N)} \\
&= \frac{x}{2(M-N)}
\end{aligned}$$

De berekening toont inderdaad dat de kans dat een vrouw in de steekproef voorkomt groter wordt, naarmate de tijdsduur die een vrouw nodig heeft om zwanger te worden groter wordt.

1.2 Verdeling tijdsduur gegeven vrouw in steekproef

De tijdsduur die een vrouw nodig heeft om zwanger te worden, gegeven dat de vrouw in de steekproef voorkomt, wordt beschouwd. Deze tijdsduur wordt vanaf nu aangeduid met de stochastische variabele $\bar{X}$. In deze paragraaf wordt de cumulatieve verdelingsfunctie van deze stochastische variabele $\bar{X}$ afgeleid. Deze cumulatieve verdelingsfunctie wordt $\overline{F_X}$ genoemd.

Om de cumulatieve verdelingsfunctie $\overline{F_X}$ te kunnen afleiden, wordt gekeken naar de verdelingen van de stochasten V en X . De stochast V is uniform verdeeld op het interval $[-M + N, M - N]$. Dit houdt in dat voor ieder deelinterval van $[-M + N, M - N]$ waarvan de lengte gelijk is, de kans dat de stochast V in het interval ligt ook gelijk is. Simpel gezegd hangt de kans dat de stochast V in het interval ligt, af van de lengte van het desbetreffende interval. De stochast X kan gezien worden als een trekking uit de onbekende kansdichtheid f_X . Hieronder is dit nogmaals kort genoteerd.

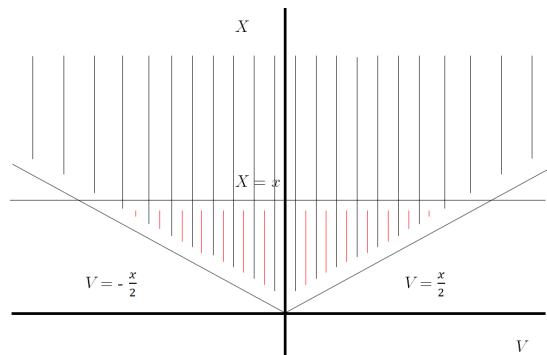
$$V \sim f_V \text{ waarbij geldt } f_V(v) = \begin{cases} \frac{1}{2(M-N)} & \text{als } -M + N \leq v \leq M - N \\ 0 & \text{als anders} \end{cases}$$

$$X \sim f_X$$

Gewenst is om meer inzicht te krijgen in het gedrag van de stochasten V en X . Om dit te bewerkstelligen is in figuur 1.2 V uitgezet tegen X . Merk op dat V en X onafhankelijk zijn. Dit wordt genoteerd als $V \perp X$. Merk tevens op dat de onderstaande relaties gelden.

$$\left[-\frac{x}{2} \leq V \leq \frac{x}{2} \text{ en } 0 \leq X \leq x\right] \iff [(V, X) \in \text{||||}]$$

$$\left[-\frac{x}{2} \leq V \leq \frac{x}{2} \text{ en } 0 \leq X < \infty\right] \iff [(V, X) \in \text{||||}]$$



Figuur 1.2: Globale schets waarbij V is uitgezet tegen X .

Door gebruik te maken van achtereenvolgens de definitie van voorwaardelijke kans, de zojuist gegeven relaties en de onafhankelijkheid van V en X kan de gezochte cumulatieve verdelingsfunctie $\overline{F}_X$ worden afgeleid. Onderstaand is deze afleiding uitgewerkt.

$$\begin{aligned}
\overline{F}_X(x) &= P(0 \leq X \leq x \mid \text{vrouw in steekproef}) \\
&= \frac{P(0 \leq X \leq x \text{ en vrouw in steekproef})}{P(\text{vrouw in steekproef})} \\
&= \frac{P(0 \leq X \leq x \text{ en vrouw in steekproef})}{P(0 \leq X < \infty \text{ en vrouw in steekproef})} \\
&= \frac{P\left(0 \leq X \leq x \text{ en } -\frac{x}{2} \leq V \leq \frac{x}{2}\right)}{P\left(0 \leq X < \infty \text{ en } \frac{x}{2} \leq V \leq \frac{x}{2}\right)} \\
&= \frac{P\left(-\frac{x}{2} \leq V \leq \frac{x}{2} \text{ en } 0 \leq X \leq x\right)}{P\left(-\frac{x}{2} \leq V \leq \frac{x}{2} \text{ en } 0 \leq X < \infty\right)} \\
&= \frac{P([(V, X) \in \text{||||})]}{P([(V, X) \in \text{||||})]} \\
&= \frac{\int_{x=0}^x \int_{v=-\frac{x}{2}}^{\frac{x}{2}} f_X(x) f_V(v) dv dx}{\int_{x=0}^{\infty} \int_{v=-\frac{x}{2}}^{\frac{x}{2}} f_X(x) f_V(v) dv dx} \\
&= \frac{\int_{x=0}^x \int_{v=-\frac{x}{2}}^{\frac{x}{2}} f_X(x) \frac{1}{2(M-N)} dv dx}{\int_{x=0}^{\infty} \int_{v=-\frac{x}{2}}^{\frac{x}{2}} f_X(x) \frac{1}{2(M-N)} dv dx} \\
&= \frac{\int_{x=0}^x \left[f_X(x) \frac{v}{2(M-N)} \right]_{v=-\frac{x}{2}}^{\frac{x}{2}} dx}{\int_{x=0}^{\infty} \left[f_X(x) \frac{v}{2(M-N)} \right]_{v=-\frac{x}{2}}^{\frac{x}{2}} dx} \\
&= \frac{\int_{x=0}^x \left(f_X(x) \frac{\frac{x}{2}}{2(M-N)} + f_X(x) \frac{\frac{x}{2}}{2(M-N)} \right) dx}{\int_{x=0}^{\infty} \left(f_X(x) \frac{\frac{x}{2}}{2(M-N)} + f_X(x) \frac{\frac{x}{2}}{2(M-N)} \right) dx}
\end{aligned}$$

$$\begin{aligned}
&= \frac{\int_{x=0}^x f_X(x) \frac{x}{2(M-N)} dx}{\int_{x=0}^{\infty} f_X(x) \frac{x}{2(M-N)} dx} \\
&= \frac{\frac{1}{2(M-N)} \int_{x=0}^x f_X(x) x dx}{\frac{1}{2(M-N)} \int_{x=0}^{\infty} f_X(x) x dx} \\
&= \frac{\int_{x=0}^x f_X(x) x dx}{\int_{x=0}^{\infty} f_X(x) x dx}
\end{aligned}$$

Deze verdelingsfunctie $\overline{F}_X$ wordt ook wel de length biased verdeling corresponderend met de kansdichtheid f_X genoemd. Merk op dat de integraal $\int_{x=0}^{\infty} f_X(x) x dx$ gelijk is aan een constante. De verdelingsfunctie F_X wordt vervormd door deze constante.

1.3 Verdeling current duration

De current duration wordt nu beschouwd. In paragraaf 1.1 is terug te vinden dat de current duration overeenkomt met de stochastische variabele Z . In deze paragraaf wordt de kansdichtheid van deze stochastische variabele Z afgeleid. Deze kansdichtheid wordt g genoemd. Aangenomen wordt dat $Z = U\bar{X}$.

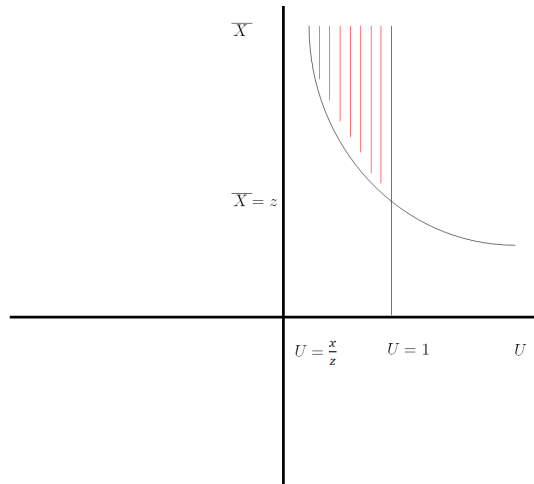
Om de kansdichtheid g te kunnen afleiden, wordt gekeken naar de verdelingen van de stochasten U en $\bar{X}$. Hierbij wordt de aanname gedaan dat de stochast U uniform is verdeeld op het interval $[0, 1]$. Dit houdt in dat voor ieder deelinterval van $[0, 1]$ waarvan de lengte gelijk is, de kans dat de stochast U in het interval ligt ook gelijk is. Simpel gezegd hangt de kans dat de stochast U in het interval ligt, af van de lengte van het desbetreffende interval. De stochast $\bar{X}$ kan gezien worden als een trekking uit de kansdichtheid $\bar{f}_X$. Deze kansdichtheid wordt gevonden door de cumulatieve verdelingsfunctie $\bar{F}_X$ af te leiden naar x . Hieronder is dit nogmaals kort genoteerd.

$$U \sim f_U \text{ waarbij geldt } f_U(u) = \begin{cases} 1 & \text{als } 0 \leq u \leq 1 \\ 0 & \text{als anders} \end{cases}$$

$$\bar{X} \sim \bar{f}_X \text{ waarbij geldt } \bar{f}_X(x) = \frac{d}{dx} \bar{F}_X(x)$$

Gewenst is om meer inzicht te krijgen in het gedrag van de stochast Z . Om dit te bewerkstelligen is in figuur 1.3 U uitgezet tegen $\bar{X}$. Merk op dat U en $\bar{X}$ onafhankelijk zijn. Dit wordt genoteerd als $U \perp \bar{X}$. Merk tevens op dat de onderstaande relatie geldt.

$$[Z > z] \iff [(U, \bar{X}) \in \text{III}]$$



Figuur 1.3: Globale schets waarbij U is uitgezet tegen $\bar{X}$.

Door gebruik te maken van achtereenvolgens de zojuist gegeven relatie en de onafhankelijkheid van U en $\overline{X}$ kan de overlevingsfunctie van Z worden afgeleid. Onderstaand is deze afleiding uitgewerkt.

$$\begin{aligned}
P(Z > z) &= P([(U, \overline{X}) \in \text{III}]) \\
&= \int_{x=z}^{\infty} \int_{u=\frac{z}{x}}^1 \overline{f_X}(x) f_U(u) du dx \\
&= \int_{x=z}^{\infty} \int_{u=\frac{z}{x}}^1 \overline{f_X}(x) 1 du dx \\
&= \int_{x=z}^{\infty} \int_{u=\frac{z}{x}}^1 \overline{f_X}(x) du dx \\
&= \int_{x=z}^{\infty} [\overline{f_X}(x)u]_{u=\frac{z}{x}}^1 dx \\
&= \int_{x=z}^{\infty} \left(\overline{f_X}(x) - \overline{f_X}(x) \frac{z}{x} \right) dx \\
&= \int_{x=z}^{\infty} \left(1 - \frac{z}{x} \right) \overline{f_X}(x) dx \\
&= \int_{x=z}^{\infty} \left(1 - \frac{z}{x} \right) \frac{d}{dx} \overline{F_X}(x) dx \\
&= \int_{x=z}^{\infty} \left(1 - \frac{z}{x} \right) \frac{d}{dx} \left(\frac{\int_{x=0}^x x f_X(x) dx}{\int_{x=0}^{\infty} x f_X(x) dx} \right) dx \\
&= \int_{x=z}^{\infty} \left(1 - \frac{z}{x} \right) \frac{x f_X(x)}{\int_{x=0}^{\infty} x f_X(x) dx} dx \\
&= \int_{x=z}^{\infty} (x - z) \frac{f_X(x)}{\int_{x=0}^{\infty} x f_X(x) dx} dx \\
&= \left[(x - z) \frac{F_X(x)}{\int_{x=0}^{\infty} x f_X(x) dx} \right]_{x=z}^{\infty} - \int_{x=z}^{\infty} \frac{F_X(x)}{\int_{x=0}^{\infty} x f_X(x) dx} dx
\end{aligned}$$

$$\begin{aligned}
&= \lim_{x \rightarrow \infty} \left((x - z) \frac{F_X(x)}{\int_{x=0}^{\infty} x f_X(x) dx} \right) - \int_{x=z}^{\infty} \frac{F_X(x)}{\int_{x=0}^{\infty} x f_X(x) dx} dx \\
&= \lim_{x \rightarrow \infty} \left((x - z) \frac{1}{\int_{x=0}^{\infty} x f_X(x) dx} \right) - \int_{x=z}^{\infty} \frac{F_X(x)}{\int_{x=0}^{\infty} x f_X(x) dx} dx \\
&= \lim_{x \rightarrow \infty} \left(\frac{x - z}{\int_{x=0}^{\infty} x f_X(x) dx} \right) - \int_{x=z}^{\infty} \frac{F_X(x)}{\int_{x=0}^{\infty} x f_X(x) dx} dx \\
&= \int_{x=z}^{\infty} \frac{1}{\int_{x=0}^{\infty} x f_X(x) dx} dx - \int_{x=z}^{\infty} \frac{F_X(x)}{\int_{x=0}^{\infty} x f_X(x) dx} dx \\
&= \int_{x=z}^{\infty} \left(\frac{1}{\int_{x=0}^{\infty} x f_X(x) dx} - \frac{F_X(x)}{\int_{x=0}^{\infty} x f_X(x) dx} \right) dx \\
&= \int_{x=z}^{\infty} \frac{1 - F_X(x)}{\int_{x=0}^{\infty} x f_X(x) dx} dx
\end{aligned}$$

Merk op dat de overlevingsfunctie van Z kan worden uitgedrukt in de gezochte kansdichtheid g . Door van dit gegeven gebruik te maken kan deze kansdichtheid g worden afgeleid. Hieronder is deze afleiding uitgewerkt.

$$\begin{aligned}
P(Z > z) &= \int_{x=z}^{\infty} \frac{1 - F_X(x)}{\int_{x=0}^{\infty} x f_X(x) dx} dx \\
\iff \int_{x=z}^{\infty} g(z) dz &= \int_{x=z}^{\infty} \frac{1 - F_X(x)}{\int_{x=0}^{\infty} x f_X(x) dx} dx \\
\iff g(z) &= \frac{1 - F_X(z)}{\int_{x=0}^{\infty} x f_X(x) dx}
\end{aligned}$$

Merk op dat de kansdichtheid g begrensd is. Dit wordt duidelijk door de grenzen van het interval $[0, \infty)$ te beschouwen.

$$\begin{aligned} 1. \quad g(0) &= \frac{1 - F_X(0)}{\int_{x=0}^{\infty} x f_X(x) dx} \\ &= \frac{1}{\int_{x=0}^{\infty} x f_X(x) dx} \end{aligned}$$

$$\begin{aligned} 2. \quad \lim_{z \rightarrow \infty} g(z) &= \lim_{z \rightarrow \infty} \frac{1 - F_X(z)}{\int_{x=0}^{\infty} x f_X(x) dx} \\ &= 0 \end{aligned}$$

Merk tevens op dat de kansdichtheid g dalend is. Dit wordt duidelijk door de onderstaande eigenschap van de cumulatieve verdelingsfunctie F_X te beschouwen.

$$\begin{aligned} z_1 < z_2 &\implies P(Z < z_1) \leq P(Z < z_2) \\ &\implies F_X(z_1) \leq F_X(z_2) \\ &\implies \frac{1 - F_X(z_1)}{\int_{x=0}^{\infty} x f_X(x) dx} \geq \frac{1 - F_X(z_2)}{\int_{x=0}^{\infty} x f_X(x) dx} \\ &\implies g(z_1) \geq g(z_2) \end{aligned}$$

1.4 Verdeling tijdsduur

De tijdsduur die een vrouw nodig heeft om zwanger te worden wordt nu beschouwd. In paragraaf 1.1 is terug te vinden dat deze tijdsduur overeenkomt met de stochastische variabele X . In deze paragraaf wordt de cumulatieve verdelingsfunctie van deze stochastische variabele X afgeleid. Deze cumulatieve verdelingsfunctie wordt F_X genoemd.

In het paragraaf 1.3 is de kansdichtheid g afgeleid. Door deze kansdichtheid g om te schrijven wordt de cumulatieve verdelingsfunctie F_X verkregen. Hierbij wordt gebruik gemaakt van het gegeven dat $g(0) = \frac{1}{\int_{x=0}^{\infty} x f_X(x) dx}$. Onderstaand is dit omschrijvingsproces uitgewerkt.

$$g(z) = \frac{1 - F_X(z)}{\int_{x=0}^{\infty} x f_X(x) dx}$$

$$\iff g(z) \int_{x=0}^{\infty} x f_X(x) dx = 1 - F_X(z)$$

$$\iff g(z) \int_{x=0}^{\infty} x f_X(x) dx - 1 = -F_X(z)$$

$$\iff F_X(z) = 1 - g(z) \int_{x=0}^{\infty} x f_X(x) dx$$

$$\iff F_X(z) = 1 - \frac{g(z)}{g(0)}$$

Merk op dat de cumulatieve verdelingsfunctie F_X al kan worden vastgesteld als enkel de kansdichtheid g bekend is. Interessant is nu om te bekijken hoe deze kansdichtheid g kan worden geschat. Hiertoe worden in de volgende hoofdstukken verschillende schatters bestudeerd. In het bijzonder wordt aandacht besteed aan het gedrag van de schatters in het punt 0, gezien de waarde in dit punt vereist is.

2 — Niet-parametrische maximum likelihood schatter

Het voorafgaande hoofdstuk toont dat de verdeling van de tijdsduur die een vrouw nodig heeft om zwanger te worden al kan worden vastgesteld als enkel de verdeling van de current duration bekend is. Met specifiekere woorden: de cumulatieve verdelingsfunctie F_X behorend bij de stochast X kan al worden vastgesteld als de kansdichtheid g behorend bij de stochast Z bekend is. Gekeken wordt nu hoe deze kansdichtheid g kan worden geschat op basis van een steekproef. In dit hoofdstuk wordt allereerst onderscheid gemaakt tussen parametrisch en niet-parametrisch schatten. Vervolgens wordt de niet-parametrische maximum likelihood schatter beschouwd, zoals voorgesteld door Grenander (1956) [3]. Aangetoond wordt dat deze schatter consistent is op het interval $(0, \infty)$, en inconsistent is rond het punt $z = 0$. Tot slot wordt het gedrag van de schatter bekeken door middel van een simulatie.

2.1 Parametrisch versus niet-parametrisch schatten

Laat $Z_1, \dots, Z_n$ onafhankelijke, identiek verdeelde stochastische variabelen zijn met een gemeenschappelijke kansdichtheid g . Getracht wordt om deze kansdichtheid te schatten op basis van deze stochastische variabelen. Bij het schatten van deze dichtheid wordt onderscheid gemaakt tussen parametrisch en niet-parametrisch schatten.

Bij parametrisch schatten wordt aangenomen dat de stochastische variabelen trekkingen zijn uit een bekende verdeling, welke enkel nog afhangt van een onbekende parameter. Hierbij wordt gezocht naar de onbekende parameter. Kortom, bij parametrisch schatten wordt aangenomen dat de te schatten kansdichtheid g behoort tot een verdeling die wordt gekarakteriseerd door een parameter. Een kansdichtheid g die wordt gekarakteriseerd door een parameter θ wordt vanaf nu aangeduid met $g(\cdot|\theta)$. Op basis van de steekproef wordt door middel van de schatter de onbekende parameter benaderd. Merk overigens op dat een kansdichtheid ook kan worden gekarakteriseerd door meerdere parameters.

Bij niet-parametrisch schatten wordt echter niet aangenomen dat de stochastische variabelen trekkingen zijn uit een bekende verdeling, die wordt gekarakteriseerd door een parameter. Kortom, bij niet-parametrisch schatten wordt zo min mogelijk aangenomen over de te schatten kansdichtheid g . Hierbij wordt niet aangenomen dat de kansdichtheid g behoort tot een specifieke verdeling die wordt gekarakteriseerd door

een parameter. Er wordt nu enkel een aanname gedaan over de structuur van de kansdichtheid g . Op basis van de steekproef wordt door middel van de schatter de onbekende kansdichtheid benaderd.

Ter verduidelijking worden twee voorbeelden bekeken: de een dient als voorbeeld voor het parametrisch schatten, de ander dient als voorbeeld voor het niet-parametrisch schatten.

1. De maximum likelihood schatter dient als voorbeeld voor het parametrisch schatten. Hierbij wordt aangenomen dat de kansdichtheid g wordt gekarakteriseerd door een parameter θ . De likelihood van de data $Z_1, \dots, Z_n$ wordt gevonden door het product te nemen van de kansdichtheden $g(Z_1|\theta), \dots, g(Z_n|\theta)$. Simpel gezegd toont de likelihood hoe waarschijnlijk de gekozen waarde van de parameter is ten opzichte van de geobserveerde data. Voor de maximum likelihood schatter $\hat{\theta}$ wordt die waarde gekozen, waarvoor de likelihood maximaal is. Dat wil zeggen dat de maximum likelihood schatter $\hat{\theta}$ wordt gevonden door de likelihood te maximaliseren over de onbekende parameter θ .
2. De niet-parametrische maximum likelihood schatter dient als voorbeeld voor het niet-parametrisch schatten. Hierbij wordt enkel aangenomen dat de kansdichtheid g dalend is. De likelihood van de data $Z_1, \dots, Z_n$ wordt gevonden door het product te nemen van de kansdichtheden $g(Z_1), \dots, g(Z_n)$. De likelihood toont hoe waarschijnlijk de gekozen kansdichtheid is ten opzichte van de geobserveerde data. Voor de maximum likelihood schatter $\hat{g}$ wordt die kansdichtheid gekozen, waarvoor de likelihood maximaal is. Dat wil zeggen dat de maximum likelihood schatter $\hat{g}$ wordt gevonden door deze likelihood te maximaliseren over de onbekende kansdichtheid g . In de volgende paragrafen wordt dieper ingegaan op deze niet-parametrische maximum likelihood schatter.

2.2 Constructie

Laat $Z_1, \dots, Z_n$ onafhankelijke, identiek verdeelde stochastische variabelen zijn met een gemeenschappelijke kansdichtheid g . Bekend is dat deze kansdichtheid dalend is. Getracht wordt om deze kansdichtheid g te schatten op basis van de data. Het niet-parametrische maximum likelihood probleem om deze kansdichtheid $g(z)$ te schatten is als volgt gedefinieerd.

$$\text{maximaliseer} \quad \ell(g) = \frac{1}{n} \sum_{i=1}^n \log(g(Z_i))$$

$$\text{onderhevig aan} \quad g \in \mathcal{D} = \{f : [0, \infty) \rightarrow [0, \infty) : f \text{ is dalend, } \int_0^\infty f(x)dx = 1\}$$

Het oplossen van het bovenstaande probleem geeft de niet-parametrische maximum likelihood schatter $\hat{g}_n$, welke door Grenander werd gevonden in 1956. Grenander pleit dat deze schatter gelijk is aan een stapfunctie, welke volgens het onderstaande stappenplan kan worden verkregen.

1. Construeer de empirische cumulatieve verdelingsfunctie G_n behorend bij de data $Z_1, \dots, Z_n$. Deze functie wordt als volgt verkregen. Sorteert de data zodanig dat $Z_{1;n} < \dots < Z_{n;n}$. Begin in het punt $Z_{1;n}$, en maak in dit punt een sprong van $\frac{1}{n}$. Herhaal het voorafgaande voor de punten $Z_{i;n}$ met $2 < i < n$. De definitie luidt als volgt [4].

$$G_n(z) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{Z_i \leq z\}$$

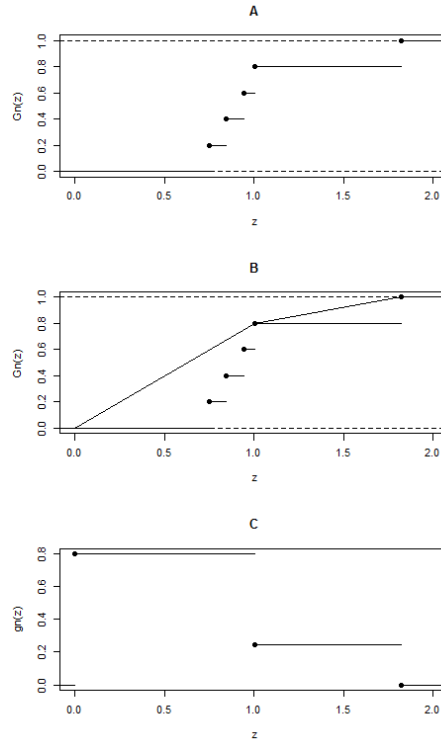
2. Construeer de kleinste concave majorant $\hat{G}_n$ behorend bij de empirische cumulatieve verdelingsfunctie G_n . Deze functie wordt als volgt verkregen. Ook hier geldt dat de data gesorteerd dient te worden zodanig dat $Z_{1;n} < \dots < Z_{n;n}$. Begin in het punt $(0, 0)$, waarbij lijnen worden getrokken naar de punten $(Z_{i;n}, G_n(Z_{i;n}))$ voor $1 < i < n$. Kies het punt $(Z_{i;n}, G_n(Z_{i;n}))$, waarbij de lijn de richtingscoëfficiënt maximaal is. Teken de desbetreffende lijn. Herhaal het voorafgaande voor dit punt $(Z_{i;n}, G_n(Z_{i;n}))$. De definitie luidt als volgt [5].

$$\hat{G}_n(z) = \inf\{\phi(z) : \phi(z) \text{ is een concaaf, } \phi(z) \geq G_n(z) \text{ voor alle } 0 < z < \infty\}$$

3. Neem de linker afgeleide van de kleinste concave majorant $\hat{G}_n$. Merk op dat de kleinste concave majorant $\hat{G}_n$ knikpunten bevat, waardoor het niet mogelijk is om de afgeleide te nemen. De niet-parametrische maximum likelihood schatter $\hat{g}_n$ is nu gevonden. De definitie luidt als volgt [6].

$$\hat{g}_n(z) = \lim_{h \uparrow 0} \frac{\hat{G}_n(z+h) - \hat{G}_n(z)}{h}$$

Ter illustratie wordt een kleine steekproef van $n = 5$ trekkingen uit de halve standaard normale verdeling beschouwd. Dit wordt bewerkstelligd door de absolute waarde te nemen van de trekkingen uit de standaard normale verdeling. De kansdichtheid behorend bij de halve standaard normale verdeling wordt gevonden door de kansdichtheid behorend bij de standaard normale verdeling te vermenigvuldigen met twee. Merk op dat de kansdichtheid behorend bij de halve standaard normale verdeling een dalende dichtheid is. Op basis van deze trekkingen zijn achtereenvolgens de empirische cumulatieve verdelingsfunctie G_n , de kleinste concave majorant $\hat{G}_n$ en de niet-parametrische maximum likelihood schatter $\hat{g}_n$ bepaald. In figuur 2.1 is dit weergegeven. Figuur 2.1A toont de empirische cumulatieve verdelingsfunctie G_n . Figuur 2.1B toont de kleinste concave majorant $\hat{G}_n$. Figuur 2.1C toont de niet-parametrische maximum likelihood schatter $\hat{g}_n$. Merk op dat de schatter inderdaad een stapfunctie is.



Figuur 2.1: A, B en C tonen respectievelijk de empirische cumulatieve verdelingsfunctie G_n , de kleinste concave majorant $\hat{G}_n$ en de niet-parametrische maximum likelihood schatter $\hat{g}_n$ op basis van een steekproef van $n = 5$ trekkingen uit de halve standaard normale verdeling.

2.3 Consistentie op $(0, \infty)$

De niet-parametrische maximum likelihood schatter $\hat{g}_n$ blijkt consistent te zijn op het interval $(0, \infty)$. Dit houdt in dat de niet-parametrische maximum likelihood schatter $\hat{g}_n$ naar de kansdichtheid g convergeert op dit interval met kans 1, indien de grootte van de steekproef naar oneindig wordt gestuurd. Om dit aan te kunnen tonen worden allereerst de stelling van Glivenko Cantelli en het lemma van Marshall beschouwd [7].

1. De stelling van Glivenko Cantelli toont een eigenschap van de empirische cumulatieve verdelingsfunctie G_n . De stelling zegt dat de supremumfout van de empirische cumulatieve verdelingsfunctie G_n ten opzichte van de cumulatieve verdelingsfunctie G naar 0 convergeert met kans 1, indien de grootte van de steekproef naar oneindig wordt gestuurd. Dat wil zeggen dat de empirische cumulatieve verdelingsfunctie G_n een consistente schatter is voor de cumulatieve verdelingsfunctie G op het interval $[0, \infty)$. Onderstaand is de stelling genoteerd.

$$\lim_{n \rightarrow \infty} \|G_n(z) - G(z)\|_{\infty} = 0 \quad \text{met kans 1}$$

$$\iff \lim_{n \rightarrow \infty} \sup_{z \in [0, \infty)} |G_n(z) - G(z)| = 0 \quad \text{met kans 1}$$

2. Het lemma van Marshall toont een eigenschap van de kleinste concave majorant $\hat{G}_n$. Het lemma zegt dat de supremumfout van de kleinste concave majorant $\hat{G}_n$ ten opzichte van de cumulatieve verdelingsfunctie G kleiner dan of gelijk is aan de supremumfout van de empirische cumulatieve verdelingsfunctie G_n ten opzichte van de cumulatieve verdelingsfunctie G . Onderstaand is het lemma genoteerd.

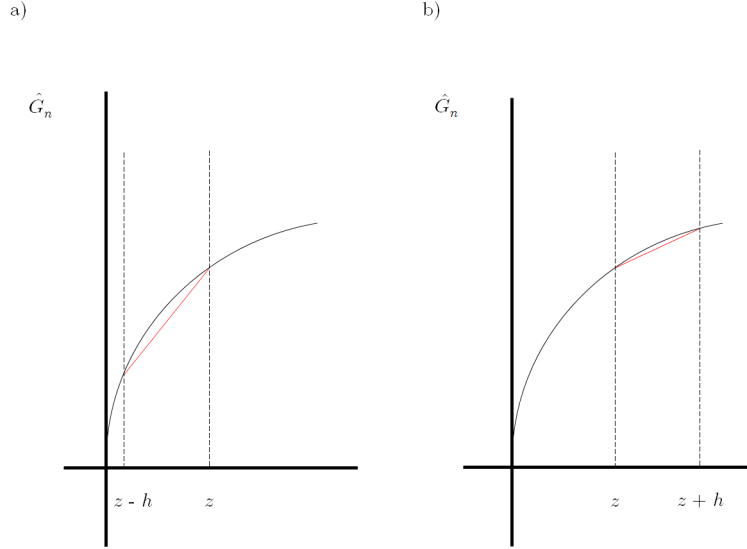
$$\|\hat{G}_n(z) - G(z)\|_{\infty} \leq \|G_n(z) - G(z)\|_{\infty}$$

$$\iff \sup_{z \in [0, \infty)} |\hat{G}_n(z) - G(z)| \leq \sup_{z \in [0, \infty)} |G_n(z) - G(z)|$$

Combineren van de stelling van Glivenko Cantelli en het lemma van Marshall geeft nu een nieuwe eigenschap voor de kleinste concave majorant $\hat{G}_n$. De supremumfout van de kleinste concave majorant $\hat{G}_n$ ten opzichte van de cumulatieve verdelingsfunctie G convergeert naar 0 met kans 1, indien de grootte van de steekproef naar oneindig wordt gestuurd. Dat wil zeggen dat de kleinste concave majorant $\hat{G}_n$ een consistente schatter is voor de cumulatieve verdelingsfunctie G op het interval $[0, \infty)$. Dit gevolg is hieronder genoteerd.

$$\lim_{n \rightarrow \infty} \sup_{z \in [0, \infty)} |\hat{G}_n(z) - G(z)| = 0$$

Merk op dat het gewenst is om de zojuist gevonden eigenschap van de kleinste concave majorant $\hat{G}_n$ uit te breiden naar haar afgeleide. Om dit te bewerkstelligen wordt gebruik gemaakt van de concaviteit van de kleinste concave majorant. In figuur 2.2 is een globale schets van de kleinste concave majorant weergegeven. Figuur 2.2A toont de lijn tussen $G(z - h)$ en $G(z)$, figuur 2.2B toont de lijn tussen $G(z)$ en $G(z + h)$. In de onderstaande afleiding wordt gebruik gemaakt van de richtingscoëfficiënten van deze lijnen. Merk op dat hierbij geldt dat $h > 0$.



Figuur 2.2: Globale schets van de kleinste concave majorant $\hat{G}_n$.

$$\begin{aligned}
& \frac{\hat{G}_n(z) - \hat{G}_n(z-h)}{h} \geq \hat{G}_n^l(z) \geq \frac{\hat{G}_n(z+h) - \hat{G}_n(z)}{h} \\
& \implies \lim_{n \rightarrow \infty} \frac{\hat{G}_n(z) - \hat{G}_n(z-h)}{h} \geq \limsup_{n \rightarrow \infty} \hat{G}_n^l(z) \geq \liminf_{n \rightarrow \infty} \hat{G}_n^l(z) \geq \lim_{n \rightarrow \infty} \frac{\hat{G}_n(z+h) - \hat{G}_n(z)}{h} \\
& \iff \frac{G(z) - G(z-h)}{h} \geq \limsup_{n \rightarrow \infty} \hat{G}_n^l(z) \geq \liminf_{n \rightarrow \infty} \hat{G}_n^l(z) \geq \frac{G(z+h) - G(z)}{h} \\
& \implies \lim_{h \downarrow 0} \frac{G(z) - G(z-h)}{h} \geq \limsup_{n \rightarrow \infty} \hat{G}_n^l(z) \geq \liminf_{n \rightarrow \infty} \hat{G}_n^l(z) \geq \lim_{h \downarrow 0} \frac{G(z+h) - G(z)}{h} \\
& \iff G^l(z) \geq \limsup_{n \rightarrow \infty} \hat{G}_n^l(z) \geq \liminf_{n \rightarrow \infty} \hat{G}_n^l(z) \geq G^r(z) \\
& \iff G^l(z) \geq \limsup_{n \rightarrow \infty} \hat{g}_n(z) \geq \liminf_{n \rightarrow \infty} \hat{g}_n(z) \geq G^r(z)
\end{aligned}$$

Aangenomen wordt nu de dat kansdichtheid g continu is op het interval $(0, \infty)$. Hieruit volgen de onderstaande eigenschappen. Merk op dat de tweede eigenschap wordt gevonden door middel van de insluitstelling.

$$1. G^l(z) = G^r(z) = \frac{d}{dz}G(z) = g(z)$$

$$2. G^l(z) = G^r(z) = \limsup_{n \rightarrow \infty} \hat{g}_n(z) = \liminf_{n \rightarrow \infty} \hat{g}_n(z) = \lim_{n \rightarrow \infty} \hat{g}_n(z)$$

Combineren van deze twee eigenschappen geeft nu een nieuwe eigenschap voor de niet-parametrische maximum likelihood schatter $\hat{g}_n$. De niet-parametrische maximum likelihood schatter $\hat{g}_n$ convergeert puntsgewijs naar de kansdichtheid g . Echter kan worden aangetoond dat de niet-parametrische maximum likelihood schatter $\hat{g}_n$ ook uniform convergeert naar de kansdichtheid g op het interval $[\epsilon, \infty)$ voor $\epsilon > 0$. Dat wil zeggen dat de niet-parametrische maximum likelihood schatter $\hat{g}_n$ een consistente schatter is voor de kansdichtheid g op het interval $[\epsilon, \infty)$ voor $\epsilon > 0$. Dit gevolg is hieronder genoteerd.

$$\lim_{n \rightarrow \infty} \sup_{z \in [\epsilon, \infty)} \hat{g}_n(z) = g(z) \quad , \quad \epsilon > 0$$

$$\iff \lim_{n \rightarrow \infty} \sup_{z \in [\epsilon, \infty)} |\hat{g}_n(z) - g(z)| = 0 \quad , \quad \epsilon > 0$$

2.4 Inconsistentie rond 0

De niet-parametrische maximum likelihood schatter $\hat{g}_n$ blijkt echter inconsistent te zijn in de rechteromgeving van het punt $z = 0$. Dit houdt in dat de niet-parametrische maximum likelihood schatter $\hat{g}_n$ dus niet naar de kansdichtheid g convergeert in deze omgeving, indien de grootte van de steekproef naar oneindig wordt gestuurd. De volgende notatie wordt in deze paragraaf gehanteerd.

$$1. Z_1, \dots, Z_n \stackrel{i.i.d.}{\sim} g$$

$$2. Z_{1;n} < \dots < Z_{n;n}$$

$$\implies Z_{1;n} = \min_{1 \leq i \leq n} Z_i$$

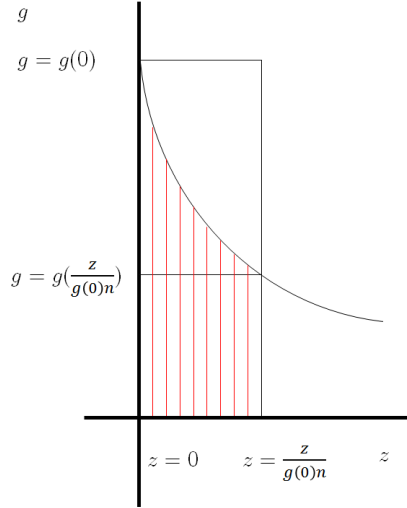
$$\implies Z_{n;n} = \max_{1 \leq i \leq n} Z_i$$

Om aan te kunnen tonen dat de niet-parametrische maximum likelihood schatter $\hat{g}_n$ inconsistent is, wordt allereerst de cumulatieve verdelingsfunctie van de $g(0)nZ_{1;n}$ afgeleid. Hierbij wordt de grootte van de steekproef naar oneindig gestuurd. Onderstaand is deze afleiding uitgewerkt. Merk op dat dit overeenkomt met de cumulatieve verdelingsfunctie behorend bij de standaard exponentiële verdeling.

$$\begin{aligned} P(g(0)nZ_{1;n} \leq z) &= 1 - P(g(0)nZ_{1;n} > z) \\ &= 1 - P\left(Z_{1;n} > \frac{z}{g(0)n}\right) \\ &= 1 - P\left(Z_{1;n} > \frac{z}{g(0)n}, \dots, Z_{n;n} > \frac{z}{g(0)n}\right) \\ &= 1 - P\left(Z_1 > \frac{z}{g(0)n}, \dots, Z_n > \frac{z}{g(0)n}\right) \\ &= 1 - P\left(Z_1 > \frac{z}{g(0)n}\right) \dots P\left(Z_n > \frac{z}{g(0)n}\right) \\ &= 1 - P\left(Z_1 > \frac{z}{g(0)n}\right)^n \\ &= 1 - \left(1 - P\left(Z_1 \leq \frac{z}{g(0)n}\right)\right)^n \\ &= 1 - \left(1 - \int_{z=0}^{\frac{z}{g(0)n}} g(z) dz\right)^n \end{aligned}$$

$$\begin{aligned}
\implies \lim_{n \rightarrow \infty} P(g(0)nZ_{1;n} \leq z) &\stackrel{(*)}{=} \lim_{n \rightarrow \infty} \left(1 - \left(1 - \frac{z}{n}\right)^n\right) \\
&= 1 - e^{-z}
\end{aligned}$$

De stap weergegeven met $(*)$ vereist verdere toelichting. Deze stap wordt daarom toegelicht door middel van figuur 2.3. In deze figuur is een globale schets van de kansdichtheid g weergegeven. Hierbij is de oppervlakte behorend bij de integraal $\int_{z=0}^{\frac{z}{g(0)n}} g(z)dz$ gemarkeerd. Te zien is dat deze oppervlakte groter is dan het vierkant $\frac{z}{g(0)n}g\left(\frac{z}{g(0)n}\right)$ en kleiner is dan het vierkant $\frac{z}{g(0)n}g(0)$. Onderstaand wordt dit gegeven in combinatie met de insluitstelling toegepast. Hierbij wordt aangenomen dat de kansdichtheid g rechtscontinu is in het punt $z = 0$.



Figuur 2.3: Globale schets van de kansdichtheid g .

$$\begin{aligned}
\frac{z}{g(0)n}g\left(\frac{z}{g(0)n}\right) &\leq \int_{z=0}^{\frac{z}{g(0)n}} g(z)dz \leq \frac{z}{g(0)n}g(0) \\
\iff \frac{z}{g(0)n}g\left(\frac{z}{g(0)n}\right) &\leq \int_{z=0}^{\frac{z}{g(0)n}} g(z)dz \leq \frac{z}{n} \\
\iff \frac{z}{g(0)}g\left(\frac{z}{g(0)n}\right) &\leq n \int_{z=0}^{\frac{z}{g(0)n}} g(z)dz \leq z \\
\iff \lim_{n \rightarrow \infty} \frac{z}{g(0)}g\left(\frac{z}{g(0)n}\right) &\leq \lim_{n \rightarrow \infty} n \int_{z=0}^{\frac{z}{g(0)n}} g(z)dz \leq \lim_{n \rightarrow \infty} z
\end{aligned}$$

$$\iff \lim_{n \rightarrow \infty} \frac{z}{g(0)} g(0) \leq \lim_{n \rightarrow \infty} n \int_{z=0}^{\frac{z}{g(0)^n}} g(z) dz \leq \lim_{n \rightarrow \infty} z$$

$$\iff \lim_{n \rightarrow \infty} z \leq \lim_{n \rightarrow \infty} n \int_{z=0}^{\frac{z}{g(0)^n}} g(z) dz \leq \lim_{n \rightarrow \infty} z$$

$$\iff \lim_{n \rightarrow \infty} n \int_{z=0}^{\frac{z}{g(0)^n}} g(z) dz = \lim_{n \rightarrow \infty} z$$

Gewenst is nu om de waarde van $\hat{g}_n(0+)$ af te schatten. Om dit te bewerkstelligen wordt gebruik gemaakt van de definitie van de kleinste concave majorant. Onderstaand is deze afchatting weergegeven.

$$\begin{aligned} \hat{g}_n(0+) &= \max_{1 \leq i \leq n} \frac{G_n(Z_{i;n})}{Z_{i;n}} \\ &= \max_{1 \leq i \leq n} \frac{\frac{i}{n}}{Z_{i;n}} \\ &= \max_{1 \leq i \leq n} \frac{i}{n Z_{i;n}} \\ &\geq \frac{1}{n Z_{1;n}} \end{aligned}$$

Door gebruik te maken van de zojuist gevonden afchatting van $\hat{g}_n(0+)$ en de cumulatieve verdelingsfunctie van $g(0)nZ_{1;n}$ wordt de kans verkregen dat de niet-parametrische maximum likelihood schatter $\hat{g}_n$ in de rechteromgeving van het punt $z = 0$ groter dan of gelijk is aan twee keer de kansdichtheid g in het punt $z = 0$. Hierbij wordt de grootte van de steekproef naar oneindig gestuurd. Deze kans is groter dan of gelijk aan 0.393. Dat wil zeggen dat de niet-parametrische maximum likelihood schatter $\hat{g}_n$ systematisch een te grote waarde aanneemt in de de rechteromgeving van het punt $z = 0$. Dit laat zien dat de niet-parametrische maximum likelihood schatter inconsistent is in deze omgeving. Hieronder is dit genoteerd.

$$\begin{aligned} P(\hat{g}_n(0+) \geq 2g(0)) &\geq P\left(\frac{1}{nZ_{1;n}} \geq 2g(0)\right) \\ &= P\left(\frac{1}{g(0)nZ_{1;n}} \geq 2\right) \\ &= P\left(g(0)nZ_{1;n} \leq \frac{1}{2}\right) \end{aligned}$$

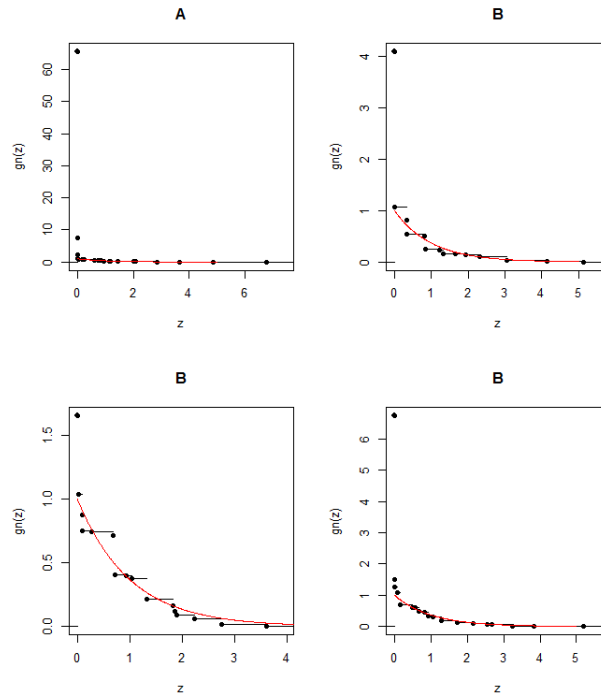
$$\implies \lim_{n \rightarrow \infty} P(\hat{g}_n(0+) \geq 2g(0)) \geq 1 - e^{-\frac{1}{2}}$$

$$\approx 0.393$$

2.5 Simulatie inconsistentie

In paragraaf 2.4 is aangetoond dat de niet-parametrische maximum likelihood schatter inconsistent is in de rechteromgeving van het punt $z = 0$. De schatter neemt systematisch een te grote waarde aan in deze omgeving. Interessant is nu om de fout van de schatter ten opzichte van de te schatten kansdichtheid in kaart te brengen. In deze paragraaf is een simulatie opgezet om deze fout te bekijken, waarbij in het bijzonder de fout in het punt $z = 0$ wordt beschouwd.

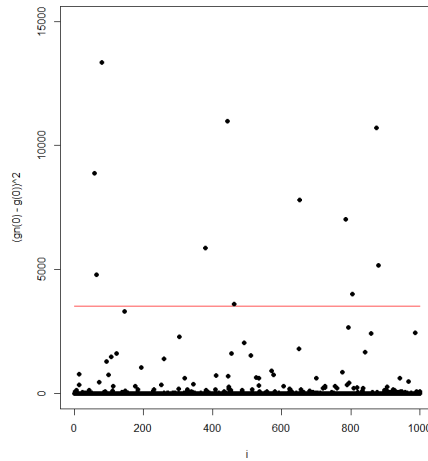
Allereerst wordt een steekproef van $n = 200$ trekkingen uit de standaard exponentiële verdeling beschouwd. Op basis van deze steekproef wordt de niet-parametrische maximum likelihood schatter $\hat{g}_n$ geconstrueerd. In totaal wordt deze handeling vier maal uitgevoerd. In figuur 2.4 is de niet-parametrische maximum likelihood schatter $\hat{g}_n$ dan ook vier maal weergegeven. De stapfunctie toont de niet-parametrische maximum likelihood schatter $\hat{g}_n$. De rode lijn toont de te schatten kansdichtheid g . De figuur toont dat de niet-parametrische maximum likelihood schatter $\hat{g}_n$ en de te schatten kansdichtheid g op het blote oog niet veel van elkaar verschillen op het interval $[\epsilon, \infty)$ voor $\epsilon > 0$. Zoals verwacht, vormt de rechteromgeving van het punt $z = 0$ een uitzondering. Gekeken wordt nu hoe groot de fout in het punt $z = 0$ van de niet-parametrische maximum likelihood schatter $\hat{g}_n$ ten opzichte van de te schatten kansdichtheid g is, wanneer het bovenstaande een groot aantal keer wordt uitgevoerd.



Figuur 2.4: Niet-parametrische maximum likelihood schatter $\hat{g}_n$ op basis van een steekproef van 200 trekkingen uit de standaard exponentiële verdeling.

Hiertoe is de volgende simulatie opgezet. Uit de standaard exponentiële verdeling worden 1000 steekproeven gedaan. Elke steekproef bestaat uit $n = 200$ trekkingen. Op basis van elke steekproef wordt de niet-parametrische maximum likelihood schatter $\hat{g}_n$ geconstrueerd. Merk op dat de schatter dus 1000 maal wordt geconstrueerd. Vervolgens wordt de fout in het punt $z = 0$ van de niet-parametrische maximum likelihood schatter $\hat{g}_n$ ten opzichte van de te schatten kansdichtheid g bepaald.

De mean squared error is een manier om deze fout te bepalen. De mean squared error meet het gemiddelde over de 1000 steekproeven van het gekwadrateerde verschil tussen de niet-parametrische maximum likelihood schatter $\hat{g}_n$ en de te schatten kansdichtheid g . In figuur 2.5 wordt dit weergegeven. De punten tonen het gekwadrateerde verschil tussen de niet-parametrische maximum likelihood schatter $\hat{g}_n$ en de te schatten kansdichtheid g . De rode lijn toont de mean squared error. De figuur toont dat het gedrag van de niet-parametrische maximum likelihood schatter $\hat{g}_n$ zeer onvoorspelbaar is. In de figuur zijn zowel behoorlijke uitschieters naar boven als naar beneden te vinden. De mean squared error is gelijk aan $3.53 \cdot 10^4$.



Figuur 2.5: Mean squared error van de niet-parametrische maximum likelihood schatter $\hat{g}_n$ op basis van 1000 steekproeven van 200 trekkingen uit de standaard exponentiële verdeling.

Merk op dat de bijbehorende R-code van de simulatie is terug te vinden in de appendix.

3 — Penalized maximum likelihood schatter

In dit hoofdstuk wordt de penalized maximum likelihood schatter beschouwd, zoals voorgesteld door Woodroffe en Sun (1993) [8]. Aangetoond wordt dat deze schatter, in tegenstelling tot de niet-parametrische maximum likelihood schatter, wel consistent is op het gehele interval $[0, \infty)$. Tot slot wordt het gedrag van de schatter bekeken door middel van een simulatie.

3.1 Constructie

Laat $Z_1, \dots, Z_n$ onafhankelijke, identiek verdeelde stochastische variabelen zijn met een gemeenschappelijke kansdichtheid g . Bekend is dat deze kansdichtheid dalend is. Getracht wordt om deze kansdichtheid g te schatten op basis van de data. Het penalized maximum likelihood probleem om deze kansdichtheid g te schatten is als volgt gedefinieerd. Hierbij staat de parameter α voor een constante, welke afhangt van de verkregen data. In de volgende paragraaf wordt hier dieper op ingegaan.

$$\text{maximaliseer} \quad \ell_p(g) = \frac{1}{n} \sum_{i=1}^n \log(g(Z_i)) - n\alpha g(0+)$$

$$\text{onderhevig aan} \quad g \in \mathcal{D} = \{f : [0, \infty) \longrightarrow [0, \infty) : f \text{ is dalend, } \int_0^\infty f(x)dx = 1\}$$

Het oplossen van het bovenstaande probleem geeft de penalized maximum likelihood schatter $\hat{g}_n$, welke door Woodroffe en Sun werd gevonden in 1993. Woodroffe en Sun pleiten dat deze schatter wederom gelijk is aan een stapfunctie, welke volgens het onderstaande stappenplan kan worden verkregen.

1. Definieer de parameter $\gamma = \min_{1 \leq i \leq n} \left(\frac{1}{2} \left(1 - \frac{\alpha}{Z_i} \right) + \sqrt{\left(\frac{\alpha}{2Z_i} \right)^2 + \frac{\alpha}{2Z_i} \left(1 - \frac{2i}{n} \right) + \frac{1}{4}} \right)$.
2. Definieer de getransformeerde data $\tilde{Z}_i = \alpha + \gamma Z_i$.

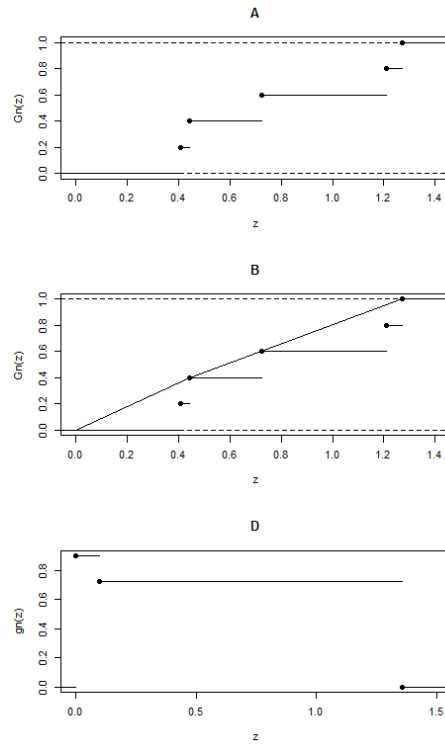
3. Construeer de empirische cumulatieve verdelingsfunctie G_n behorend bij de getransformeerde data $\tilde{Z}_1, \dots, \tilde{Z}_n$. De constructie plus definitie van deze functie zijn terug te vinden in paragraaf 2.2.
4. Construeer de kleinste concave majorant $\hat{G}_n$ behorend bij de empirische cumulatieve verdelingsfunctie G_n . Ook hier geldt weer dat de constructie plus definitie van deze functie zijn terug te vinden in paragraaf 2.2.
5. Neem de linker afgeleide van de kleinste concave majorant $\hat{G}_n$. Wederom geldt dat de constructie plus definitie van deze schatter zijn terug te vinden in paragraaf 2.2. Noem deze linker afgeleide $\tilde{g}_n$. Merk op dat het volgende geldt.

$$\tilde{g}_n(z) = \tilde{g}_i \quad \text{als } \tilde{Z}_{i-1} < z \leq \tilde{Z}_i$$

6. Definieer nu de penalized maximum likelihood schatter $\hat{g}_n$ als volgt.

$$\hat{g}_n(z) = \tilde{g}_i \quad \text{als } Z_{i-1} < z \leq Z_i$$

Ter illustratie wordt ook voor de penalized maximum likelihood schatter een kleine steekproef van $n = 5$ trekkingen uit de halve standaard normale verdeling beschouwd. Op basis van deze trekkingen zijn achtereenvolgens de empirische cumulatieve verdelingsfunctie G_n , de kleinste concave majorant $\hat{G}_n$ en de penalized maximum likelihood schatter $\hat{g}_n$ bepaald. Hierbij is de keuze gemaakt voor $\alpha = n^{-\frac{3}{5}}$. De onderbouwing van deze keuze is terug te vinden in de appendix. In figuur 3.1 is dit weergegeven. Figuur 3.1A toont de empirische cumulatieve verdelingsfunctie G_n . Figuur 3.1B toont de kleinste concave majorant $\hat{G}_n$. Figuur 3.1C toont de penalized maximum likelihood schatter $\hat{g}_n$. Merk op dat de schatter inderdaad een stapfunctie is. Tevens lijkt deze schatter weinig te verschillen van de in het vorige hoofdstuk besproken niet-parametrische maximum likelihood schatter. Echter in de volgende paragrafen zal worden aangetoond dat er wel degelijk een groot verschil tussen deze twee schatters bestaat.



Figuur 3.1: A, B en C tonen respectievelijk de empirische cumulatieve verdelingsfunctie G_n , de kleinste concave majorant $\hat{G}_n$ en de penalized maximum likelihood schatter $\hat{g}_n$ op basis van een steekproef van $n = 5$ trekkingen uit de halve standaard normale verdeling.

3.2 Parameter α

In paragraaf 3.1 staat beschreven dat de penalized maximum likelihood schatter wordt verkregen door het stappenplan van de niet-parametrische maximum likelihood schatter (grotendeels) toe te passen op de getransformeerde data. Het transformeren van de data vindt plaats door middel van de parameters α en γ . Welke waarde voor de parameter γ moet worden gekozen is terug te vinden in paragraaf 3.1. Welke waarde voor de parameter α moet worden gekozen ligt echter wat gecompliceerder.

Woodroffe en Sun stellen dat de parameter α van de vorm cn^{-pq} dient te zijn. De parameter α is dan optimaal. Laat $Z_1, \dots, Z_n$ onafhankelijke, identiek verdeelde stochastische variabelen zijn met een gemeenschappelijke dalende kansdichtheid g . Merk op dat de bijbehorende cumulatieve verdelingsfunctie wordt aangeduid met G . De waarden van de parameters p, q en c dienen nu als volgt te worden bepaald. Merk op dat de waarden van de parameters enkel kunnen worden bepaald, indien de gemeenschappelijke kansdichtheid g bekend is.

1. De waarde van de parameter p wordt gevonden door de cumulatieve verdelingsfunctie G om te schrijven tot de onderstaande vorm. Hierbij geldt dat $1 < p < \infty$.

$$G(z) = g_0 z - g_1 z^p + o(z^p) \text{ voor } z \downarrow 0$$

2. De waarde van de parameter q wordt gevonden door de zojuist gevonden waarde van de parameter p in te vullen in de onderstaande formule. Hierbij geldt dat $0 < q < \infty$.

$$q = \frac{1}{2p - 1}$$

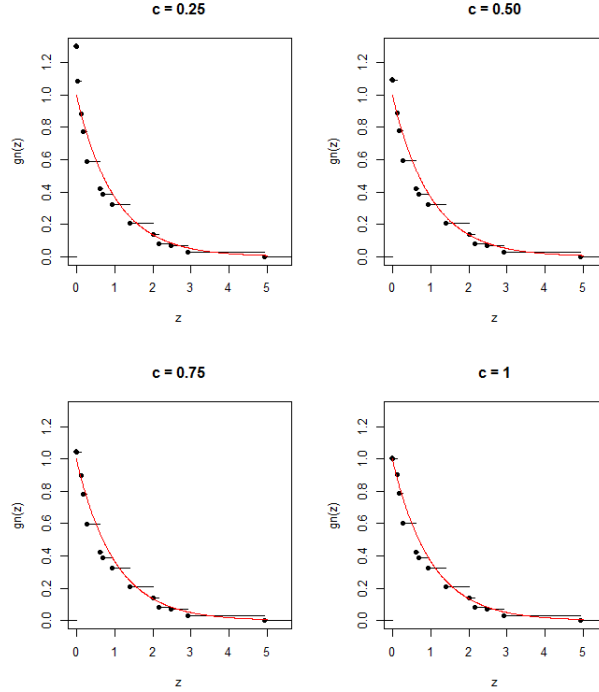
3. Op welke wijze de waarde van de parameter c moet worden bepaald is echter onbekend. In paragraaf 3.3 wordt hier dieper op ingegaan. In deze paragraaf wordt een methode aangedragen om tot een geschikte waarde voor deze parameter te komen.

Woodroffe en Sun vermelden tevens dat de penalized maximum likelihood schatter niet gevoelig blijkt te zijn voor de waarde van de parameter α op het interval $[\epsilon, \infty)$ voor $\epsilon > 0$. Echter blijkt de schatter wel gevoelig te zijn voor de waarde van de parameter α in de rechteromgeving van het punt $z = 0$.

3.3 Simulatie optimale parameter α

In paragraaf 3.2 is opgemerkt dat de parameter α van de vorm cn^{-pq} dient te zijn. De wijze waarop de waarden van de parameters p en q moeten worden bepaald is bekend. Echter de wijze waarop de waarde van de parameter c moet worden bepaald is onbekend. Interessant is nu om de fout van de penalized maximum likelihood schatter ten opzichte van de te schatten kansdichtheid voor verschillende waarden van de parameter c in kaart te brengen. In deze paragraaf is een simulatie opgezet om deze fout te bekijken, waarbij in het bijzonder de fout in het punt $z = 0$ wordt beschouwd. Dit wegens het feit dat de schatter enkel in de rechteromgeving van het punt $z = 0$ gevoelig blijkt te zijn voor de waarde van de parameter α , en dus ook voor de waarde van de parameter c .

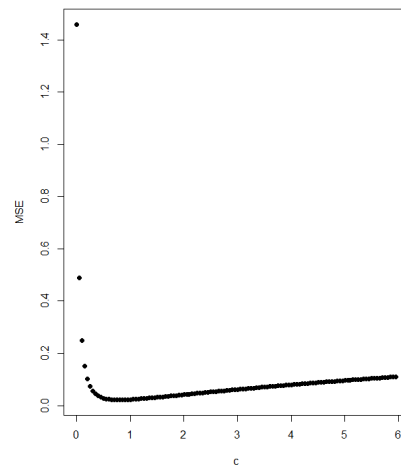
Allereerst wordt een steekproef van $n = 200$ trekkingen uit de standaard exponentiële verdeling beschouwd. Op basis van deze steekproef wordt de penalized maximum likelihood schatter $\hat{g}_n$ geconstrueerd voor de waarden $c = 0.25, 0.50, 0.75$ en 1 . Hierbij is de keuze gemaakt voor $\alpha = cn^{-\frac{2}{3}}$. De onderbouwing van deze keuze is terug te vinden in de appendix. In figuur 3.2 is de penalized maximum likelihood schatter $\hat{g}_n$ voor deze waarden weergegeven. De stapfunctie toont de penalized maximum likelihood schatter $\hat{g}_n$. De rode lijn toont de te schatten kansdichtheid g . De figuur toont dat de schatter inderdaad gevoelig is voor de waarde van de parameter c in de rechteromgeving van het punt $z = 0$. De waarde $c = 0.25$ lijkt het slechtst uit de test te komen, gezien het verschil tussen de penalized maximum likelihood schatter $\hat{g}_n$ en de te schatten kansdichtheid g bij deze waarde maximaal is. De waarde $c = 1$ lijkt het best uit de test te komen, gezien het verschil tussen de penalized maximum likelihood schatter $\hat{g}_n$ en de te schatten kansdichtheid g bij deze waarde minimaal is. Gekeken wordt nu voor welke waarde van de parameter c de fout in het punt $z = 0$ van de penalized maximum likelihood schatter $\hat{g}_n$ ten opzichte van de te schatten kansdichtheid g minimaal is, wanneer het bovenstaande een groot aantal keren wordt uitgevoerd.



Figuur 3.2: Penalized maximum likelihood schatter $\hat{g}_n$ op basis van een steekproef van 200 trekkingen uit de standaard exponentiële verdeling.

Hiertoe is de volgende simulatie opgezet. Uit de standaard exponentiële verdeling worden 1000 steekproeven gedaan. Elke steekproef bestaat uit $n = 200$ trekkingen. Op basis van elke steekproef wordt de penalized maximum likelihood schatter $\hat{g}_n$ voor de waarden $c = 0.05, 0.10, \dots, 5.95, 6$ geconstrueerd. Hierbij is wederom de keuze gemaakt voor $\alpha = cn^{-\frac{2}{3}}$. Merk op dat de schatter dus 1000 maal wordt geconstrueerd voor elke waarde van de parameter c . Vervolgens wordt de fout in het punt $z = 0$ van de penalized maximum likelihood schatter $\hat{g}_n$ ten opzichte van de te schatten kansdichtheid g bepaald voor elke waarde van de parameter c .

Zoals in het vorige hoofdstuk staat vermeld is de mean squared error een manier om deze fout te bepalen. De mean squared error wordt nu bepaald voor elke waarde van de parameter c . In figuur 3.3 is de mean squared error uitgezet tegen de waarde van de parameter c . De figuur toont dat de mean squared error een minimum bereikt. De mean squared error bereikt haar minimum voor de waarde $c = 0.85$. Dat wil zeggen dat $\alpha = 0.85n^{-\frac{2}{3}}$ optimaal is, indien een steekproef van n trekkingen uit de standaard exponentiële verdeling wordt gedaan. De mean squared error is dan gelijk aan $2.16 \cdot 10^{-2}$.



Figuur 3.3: Mean squared error van de penalized maximum likelihood schatter $\hat{g}_n$ op basis van 1000 steekproeven van 200 trekkingen uit de standaard exponentiële verdeling.

Merk op dat de bijbehorende R-code van de simulatie is terug te vinden in de appendix.

4 — Alternatieve schatter

In dit hoofdstuk wordt een alternatieve schatter beschouwd om de kansdichtheid in het punt $z = 0$ te schatten. Deze alternatieve schatter wordt voorgesteld in [10]. Aangetoond wordt dat deze schatter consistent is. Tot slot wordt het gedrag van de schatter bekeken door middel van een simulatie.

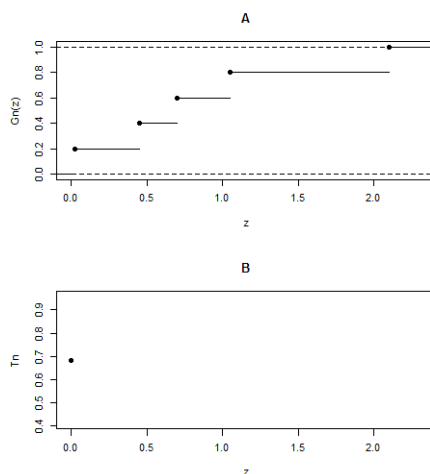
4.1 Constructie

Laat $Z_1, \dots, Z_n$ onafhankelijke, identiek verdeelde stochastische variabelen zijn met een gemeenschappelijke kansdichtheid g . Bekend is dat deze kansdichtheid dalend is. Getracht wordt om deze kansdichtheid g te schatten op basis van de data. Het schatten van de kansdichtheid g op het interval $[\epsilon, \infty)$ voor $\epsilon > 0$ heeft tot nu toe weinig problemen opgeleverd. Echter het schatten van deze kansdichtheid g in de rechteromgeving van het punt $z = 0$ levert wel problemen op.

De kansdichtheid g in het punt $z = 0$ kan worden geschat door middel van de alternatieve schatter T_n . Deze alternatieve schatter wordt volgens het onderstaande stappenplan verkregen. Merk op dat de schatter, in tegenstelling tot de niet-parametrische maximum likelihood schatter en de penalized maximum likelihood schatter, niet gelijk is aan een stapfunctie. De schatter geeft slechts één waarde terug.

1. Definieer de parameter h_n zodanig dat deze parameter naar 0 convergeert, indien de grootte van de steekproef naar oneindig wordt gestuurd. In [10] wordt gepleit dat de parameter h_n van de vorm $cn^{-\frac{1}{3}}$ dient te zijn. De mean squared error van de schatter T_n ten opzichte van de te schatten kansdichtheid g in het punt $z = 0$ blijkt dan minimaal te zijn. In paragraaf 4.3 wordt een methode aangedragen om de waarde van de parameter c te bepalen.
2. Construeer de empirische cumulatieve verdelingsfunctie G_n behorend bij de data $Z_1, \dots, Z_n$. De constructie plus definitie is terug te vinden in paragraaf 2.2.
3. Definieer nu de alternatieve schatter $T_n = \frac{1}{h_n} G_n(h_n)$.

Ter illustratie wordt ook voor deze alternatieve schatter een kleine steekproef van $n = 5$ trekkingen uit de halve standaard normale verdeling beschouwd. Op basis van deze trekkingen zijn achtereenvolgens de empirische cumulatieve verdelingsfunctie G_n en de alternatieve schatter T_n bepaald. Hierbij is de keuze gemaakt voor $h_n = n^{-\frac{1}{3}}$. In figuur 4.1 is dit weergegeven. Figuur 4.1A toont de empirische cumulatieve verdelingsfunctie G_n . Figuur 4.1B toont de alternatieve schatter T_n .



Figuur 4.1: A en B tonen respectievelijk de empirische cumulatieve verdelingsfunctie G_n en de alternatieve schatter T_n op basis van een steekproef van $n = 5$ trekkingen uit de halve standaard normale verdeling.

4.2 Consistentie

De alternatieve schatter T_n blijkt een consistente schatter voor de kansdichtheid $g(0)$ te zijn. Dit houdt in dat de alternatieve schatter naar de kansdichtheid $g(0)$ convergeert met kans 1, indien de grootte van de steekproef naar oneindig wordt gestuurd. De volgende notatie wordt in deze paragraaf gebruikt.

1. $Z_1, \dots, Z_n \stackrel{i.i.d.}{\sim} g$

2. $T_n = \frac{1}{h_n} G_n(h_n)$

Om aan te kunnen tonen dat de schatter T_n consistent is, wordt allereerst de ongelijkheid van Chebyshev afgeleid. In de onderstaande afleiding is gebruik gemaakt van het gegeven dat de indicatorfunctie $\mathbb{1}\{|T_n - E(T_n)| > \epsilon\}$ kan worden gezien als een stochastische variabele uit de Bernoulli verdeling. Hierbij is de kans op slagen p gelijk aan $P(|T_n - E(T_n)| > \epsilon)$. Merk op dat de verwachtingsverwaarde behorend bij een stochastische variabele uit de Bernoulli verdeling gelijk is aan deze kans p .

$$\begin{aligned}
 V(T_n) &= E((T_n - E(T_n))^2) \\
 &= E((T_n - E(T_n))^2 \mathbb{1}\{|T_n - E(T_n)| > \epsilon\} + (T_n - E(T_n))^2 \mathbb{1}\{|T_n - E(T_n)| \leq \epsilon\}) \\
 &\geq E((T_n - E(T_n))^2 \mathbb{1}\{|T_n - E(T_n)| > \epsilon\}) \\
 &\geq E(\epsilon^2 \mathbb{1}\{|T_n - E(T_n)| > \epsilon\}) \\
 &= \epsilon^2 E(\mathbb{1}\{|T_n - E(T_n)| > \epsilon\}) \\
 &= \epsilon^2 P(|T_n - E(T_n)| > \epsilon) \\
 \\
 &\implies P(|T_n - E(T_n)| > \epsilon) \leq \frac{V(T_n)}{\epsilon^2}
 \end{aligned}$$

De ongelijkheid van Chebyshev toont dat de kans groot is dat de schatter T_n niet veel zal verschillen van haar verwachtingswaarde, indien haar variantie zeer klein is [9]. De ongelijkheid van Chebyshev wordt vaak gebruikt bij het bewijzen van limietstellingen. Een voorbeeld hiervan is de zwakke wet van de grote aantallen.

Gewenst is nu om de verwachtingswaarde en de variantie van de schatter T_n af te leiden. In de onderstaande afleiding is gebruik gemaakt van de onafhankelijkheid van de stochastische variabelen $Z_1, \dots, Z_n$. Tevens is gebruik gemaakt van het gegeven dat de indicatorfunctie $\mathbb{1}\{Z_i \leq h_n\}$ kan worden gezien als een stochastische variabele uit de Bernoulli verdeling. Hierbij is de kans op slagen p gelijk aan $P(Z_i \leq h_n)$. Merk op dat de verwachtingswaarde en de variantie behorend bij een stochastische variabele uit de Bernoulli verdeling respectievelijk gelijk zijn aan p en $p(1 - p)$.

$$\begin{aligned}
 1. \quad E(T_n) &= E\left(\frac{1}{h_n} G_n(h_n)\right) \\
 &= E\left(\frac{1}{nh_n} \sum_{i=1}^n \mathbb{1}\{Z_i \leq h_n\}\right) \\
 &= \frac{1}{nh_n} \sum_{i=1}^n E(\mathbb{1}\{Z_i \leq h_n\}) \\
 &= \frac{1}{nh_n} \sum_{i=1}^n P(Z_i \leq h_n) \\
 &= \frac{1}{nh_n} \sum_{i=1}^n G(h_n) \\
 &= \frac{1}{nh_n} n G(h_n) \\
 &= \frac{1}{h_n} G(h_n)
 \end{aligned}$$

$$\begin{aligned}
 2. \quad V(T_n) &= V\left(\frac{1}{h_n} G_n(h_n)\right) \\
 &= V\left(\frac{1}{nh_n} \sum_{i=1}^n \mathbb{1}\{Z_i \leq h_n\}\right) \\
 &= \frac{1}{(nh_n)^2} \sum_{i=1}^n V(\mathbb{1}\{Z_i \leq h_n\}) \\
 &= \frac{1}{(nh_n)^2} \sum_{i=1}^n P(Z_i \leq h_n) (1 - P(Z_i \leq h_n)) \\
 &= \frac{1}{(nh_n)^2} \sum_{i=1}^n G(h_n) (1 - G(h_n))
 \end{aligned}$$

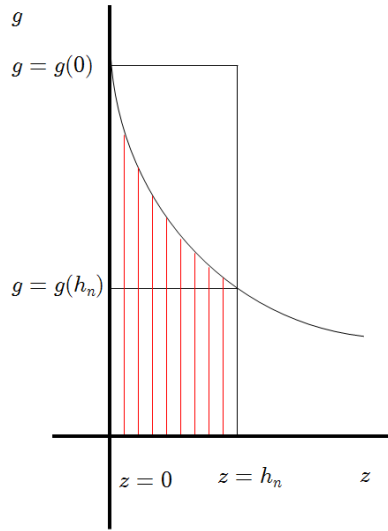
$$\begin{aligned}
&= \frac{1}{(nh_n)^2} nG(h_n)(1 - G(h_n)) \\
&= \frac{1}{nh_n^2} G(h_n)(1 - G(h_n))
\end{aligned}$$

Gekeken wordt nu naar de kans dat het verschil tussen de schatter T_n en de kansdichtheid $g(0)$ groter is dan een willekeurige $\epsilon > 0$. Hierbij wordt de grootte van de steekproef naar oneindig gestuurd. Toepassen van de ongelijkheid van Chebyshev en invullen van de variantie van de schatter T_n toont dat deze kans gelijk is aan 0. Dat wil zeggen dat de schatter T_n een consistente schatter is voor de kansdichtheid $g(0)$.

$$\begin{aligned}
P(|T_n - g(0)| > \epsilon) &= P(|T_n - E(T_n) + E(T_n) - g(0)| > \epsilon) \\
&\leq P(|T_n - E(T_n)| + |E(T_n) - g(0)| > \epsilon) \\
&= P\left(|T_n - E(T_n)| > \frac{\epsilon}{2} \text{ of } |E(T_n) - g(0)| > \frac{\epsilon}{2}\right) \\
&\leq \left(P\left(|T_n - E(T_n)| > \frac{\epsilon}{2}\right) + P\left(|E(T_n) - g(0)| > \frac{\epsilon}{2}\right)\right) \\
&\leq \left(\frac{V(T_n)}{\left(\frac{\epsilon}{2}\right)^2} + P\left(|E(T_n) - g(0)| > \frac{\epsilon}{2}\right)\right) \\
&= \left(\frac{\frac{1}{nh_n^2} G(h_n)(1 - G(h_n))}{\left(\frac{\epsilon}{2}\right)^2} + P\left(|E(T_n) - g(0)| > \frac{\epsilon}{2}\right)\right) \\
&\leq \left(\frac{\frac{1}{nh_n^2} G(h_n)}{\left(\frac{\epsilon}{2}\right)^2} + P\left(|E(T_n) - g(0)| > \frac{\epsilon}{2}\right)\right) \\
&\leq \left(\frac{\frac{1}{nh_n^2} g(0)h_n}{\left(\frac{\epsilon}{2}\right)^2} + P\left(|E(T_n) - g(0)| > \frac{\epsilon}{2}\right)\right) \\
&= \left(\frac{\frac{1}{nh_n} g(0)}{\left(\frac{\epsilon}{2}\right)^2} + P\left(|E(T_n) - g(0)| > \frac{\epsilon}{2}\right)\right)
\end{aligned}$$

$$\implies \lim_{nh_n \rightarrow \infty} P(|T_n - g(0)| > \epsilon) \stackrel{(*)}{=} 0$$

De stap weergegeven met $(*)$ vereist verdere toelichting. Deze stap wordt daarom toegelicht door middel van figuur 4.2. In deze figuur is een globale schets van de kansdichtheid g weergegeven. Hierbij is de oppervlakte behorend bij de integraal $\int_{z=0}^{h_n} g(z)dz$ gemarkeerd. Te zien is dat deze oppervlakte groter is dan het vierkant $h_n g(h_n)$ en kleiner is dan het vierkant $h_n g(0)$. Onderstaand wordt dit gegeven in combinatie met de insluitstelling toegepast. Hierbij wordt tevens aangenomen dat de kansdichtheid g rechtscontinu is in het punt $z = 0$.



Figuur 4.2: Globale schets van de kansdichtheid g .

$$h_n g(h_n) \leq \int_{z=0}^{h_n} g(z)dz \leq h_n g(0)$$

$$\iff h_n g(h_n) \leq G(h_n) \leq h_n g(0)$$

$$\iff g(h_n) \leq \frac{1}{h_n} G(h_n) \leq g(0)$$

$$\iff g(h_n) \leq E(T_n) \leq g(0)$$

$$\iff \lim_{n \rightarrow \infty} g(h_n) \leq \lim_{n \rightarrow \infty} E(T_n) \leq \lim_{n \rightarrow \infty} g(0)$$

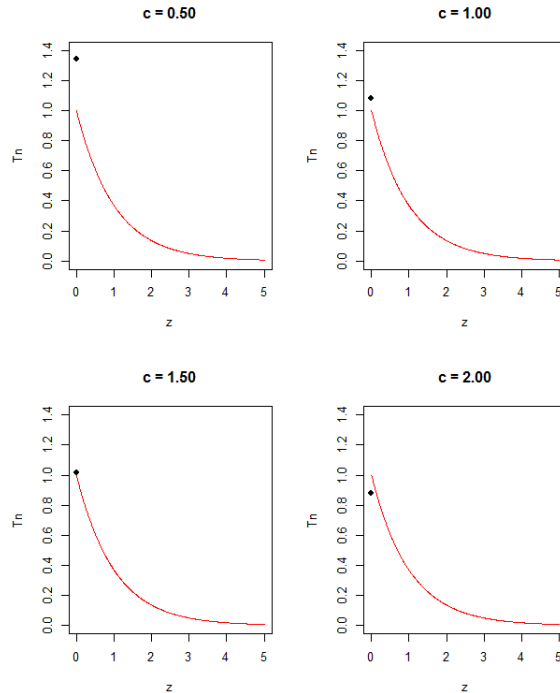
$$\Longleftrightarrow g(0) \leq \lim_{n \rightarrow \infty} E(T_n) \leq g(0)$$

$$\Longleftrightarrow \lim_{n \rightarrow \infty} E(T_n) = g(0)$$

4.3 Simulatie optimale parameter h_n

In paragraaf 4.1 is opgemerkt dat de parameter h_n van de vorm $cn^{-\frac{1}{3}}$ dient te zijn. Interessant is nu om de fout van schatter ten opzichte van de te schatten kansdichtheid voor verschillende waarden van de parameter c in kaart te brengen. In deze paragraaf is een simulatie opgezet om deze fout te bekijken. Merk op dat de opzet van de simulatie equivalent is aan simulatie in paragraaf 3.3.

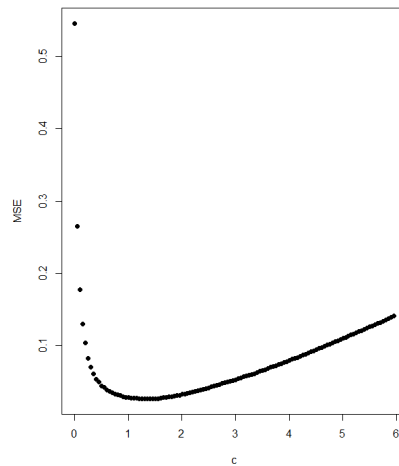
Allereerst wordt een steekproef van $n = 200$ trekkingen uit de standaard exponentiële verdeling beschouwd. Op basis van deze steekproef wordt de schatter T_n geconstrueerd voor de waarden $c = 0.50, 1.00, 1.50$ en 2.00 . In figuur 4.3 is de schatter T_n voor deze waarden weergegeven. De punt toont de schatter T_n . De rode lijn toont de te schatten kansdichtheid g . De waarde $c = 0.50$ lijkt het slechtst uit de bus te komen, gezien het verschil tussen de schatter T_n en de te schatten kansdichtheid $g(0)$ bij deze waarde maximaal is. De waarde $c = 1.50$ lijkt het best uit de bus te komen, gezien het verschil tussen de schatter T_n en de te schatten kansdichtheid $g(0)$ bij deze waarde minimaal is. Gekeken wordt nu voor welke waarde van de parameter c de fout van de schatter T_n ten opzichte van de te schatten kansdichtheid $g(0)$ minimaal is, wanneer het bovenstaande een groot aantal keren wordt uitgevoerd.



Figuur 4.3: Schatter T_n op basis van een steekproef van 200 trekkingen uit de standaard exponentiële verdeling.

Hiertoe is de volgende simulatie opgezet. Uit de standaard exponentiële verdeling worden 1000 steekproeven gedaan. Elke steekproef bestaat uit $n = 200$ trekkingen. Op basis van elke steekproef wordt de schatter T_n voor de waarden $c = 0.05, 0.10, \dots, 5.95, 6$ geconstrueerd. Merk op dat de schatter dus 1000 maal wordt geconstrueerd voor elke waarde van de parameter c . Vervolgens wordt de fout van de schatter T_n ten opzichte van de te schatten kansdichtheid $g(0)$ bepaald voor elke waarde van de parameter c .

Net zoals in het vorige hoofdstuk wordt de mean squared error gebruikt om deze fout te bepalen. De mean squared error wordt nu bepaald voor elke waarde van de parameter c . In figuur 4.4 is de mean squared error uitgezet tegen de waarde van de parameter c . De figuur toont dat de mean squared error een minimum bereikt. De mean squared error bereikt haar minimum voor de waarde $c = 1.35$. Dat wil zeggen dat $h_n = 1.35n^{-\frac{1}{3}}$ optimaal is, indien een steekproef van n trekkingen uit de standaard exponentiële verdeling wordt gedaan. De mean squared error is dan gelijk aan $2.66 \cdot 10^{-2}$.



Figuur 4.4: Mean squared error van de schatter T_n op basis van 1000 steekproeven van 200 trekkingen uit de standaard exponentiële verdeling.

Merk op dat de bijbehorende R-code van de simulatie is terug te vinden in de appendix.

5 — Analyse data

In het artikel van Slama [11] is de data behorend bij het current duration model, zoals beschreven in hoofdstuk 1, verzameld. De data bestaat uit $n = 867$ current durations afkomstig van vrouwen in Frankrijk met een leeftijd tussen de 18 en 44 jaar. De data is uitgedrukt in maanden. Voor verdere informatie over de precieze selectieprocedure van de vrouwen wordt verwezen naar het artikel van Slama [11].

Hoofdstuk 1 toont hoe uit de verdeling van de current duration de verdeling van de tijdsduur die een vrouw nodig heeft om zwanger te worden kan worden bepaald. Hoofdstuk 2, 3 en 4 tonen hoe deze verdeling van de current duration kan worden geschat. In dit hoofdstuk zal deze kennis worden toegepast. Zodoende zullen drie verschillende methoden worden toegepast op de data. Simpel gezegd is elke methode als volgt opgebouwd.

1. De verdeling van de current duration wordt geschat. Dit houdt in dat de schatter $\hat{g}_n$ voor de kansdichtheid g wordt bepaald.
2. De verdeling van de tijdsduur die een vrouw nodig heeft om zwanger te worden wordt geschat. Dit houdt in dat de schatter $\hat{F}_X$ voor de cumulatieve verdelingsfunctie F_X wordt bepaald. Deze schatter wordt verkregen door de schatter $\hat{g}_n$ in te vullen in de cumulatieve verdelingsfunctie F_X . Hieronder is dit genoteerd.

$$F_X(x) = 1 - \frac{g(x)}{g(0)}$$

$$\Longleftrightarrow \hat{F}_X(x) = 1 - \frac{\hat{g}_n(x)}{\hat{g}_n(0)}$$

5.1 Methode 1

In deze paragraaf wordt de verdeling van de tijdsduur die een vrouw nodig heeft om zwanger te worden, verkregen door middel van de penalized maximum likelihood schatter. Deze schatter staat beschreven in hoofdstuk 3. De penalized maximum likelihood schatter wordt toegepast op de data van Slama. Om de penalized maximum likelihood schatter te vinden, dient de parameter α zo optimaal mogelijk te worden vastgesteld. Ter herinnering: $\alpha = cn^{-pq}$. Dit houdt in dat allereerst de waarden van de parameters c , p en q dienen te worden bepaald.

Gewenst is om de optimale vorm van de parameter α op dezelfde wijze te verkrijgen als in paragraaf 3.3. In deze paragraaf is de onderliggende verdeling van de data bekend. Dit maakt het betrekkelijk eenvoudig om de waarden van de parameters c , p en q vast te stellen. Echter nu is de onderliggende verdeling van de data onbekend.

Hiertoe wordt de volgende methode bekeken, zoals voorgesteld door Silverman [12] Silverman pleit dat het geoorloofd is om de zogenaamde smoothing parameter van een niet-parametrische schattingsmethode te bepalen door een parametrisch model aan te nemen. Kortom, enkel voor het vaststellen van de smoothing parameter wordt een parametrisch model aangenomen. Deze methode wordt nu toegepast om de waarden van de parameters c , p en q vast te stellen. Aangenomen wordt nu dat de data uit de exponentiële verdeling is getrokken. Hierbij wordt de bijbehorende parameter λ geschat door middel van de parametrische maximum likelihood schatter $\hat{\lambda}$. Er blijkt te gelden $\hat{\lambda} = 3.21 \cdot 10^{-2}$. Onderstaand is dit nogmaals kort genoteerd.

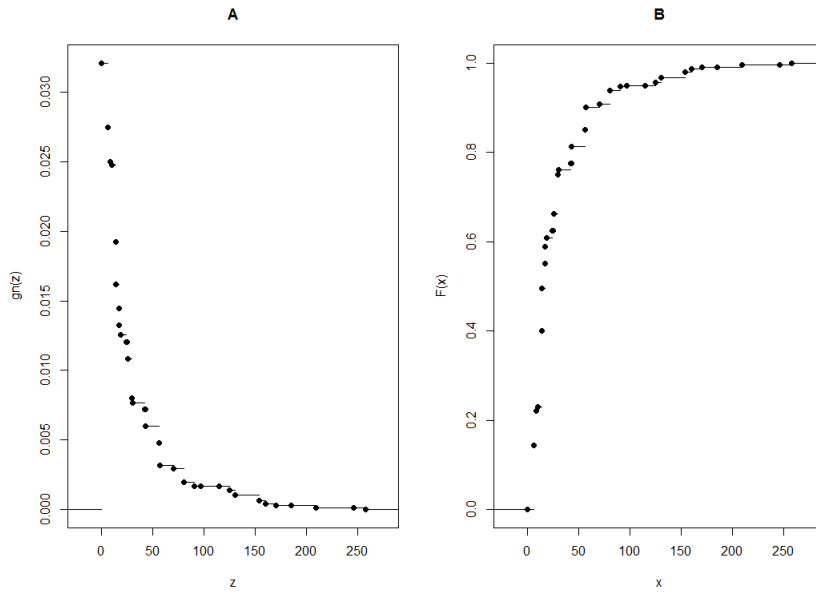
$$\begin{aligned} Z_1, \dots, Z_n &\stackrel{i.i.d.}{\sim} g \\ &\stackrel{i.i.d.}{\sim} \text{Exp}(3.21 \cdot 10^{-2}) \\ &\stackrel{i.i.d.}{\sim} \begin{cases} (3.21 \cdot 10^{-2})e^{-(3.21 \cdot 10^{-2})z} & \text{als } z \geq 0 \\ 0 & \text{als } z < 0 \end{cases} \end{aligned}$$

Merk op dat de waarden van de parameters p en q direct vastliggen door de aanname dat de data uit de exponentiële verdeling is getrokken. De waarde van de parameter c wordt nu als volgt gevonden. Voor de waarden $c = 0.05, 0.10, \dots, 999.95, 1000$ wordt de penalized maximum likelihood schatter $\hat{g}_n$ geconstrueerd. Voor elke waarde van de parameter c wordt vervolgens de mean squared error in het punt $z = 0$ van de penalized maximum likelihood schatter $\hat{g}_n$ ten opzichte van de kansdichtheid g bepaald. Gekeken wordt voor welke waarde van de parameter c deze mean squared error minimaal is. Deze waarde wordt gekozen. De mean squared error blijkt minimaal te zijn voor de waarde $c = 5.69 \cdot 10^2$. De bijbehorende mean squared error is dan gelijk aan $3.68 \cdot 10^{-14}$. Hieruit volgt nu dat $\alpha = (5.69 \cdot 10^2)n^{-\frac{2}{3}}$ optimaal is. Op de volgende pagina zijn de resultaten kort genoteerd.

1. $c_{opt} = 5.69 \cdot 10^2$, $p = 2$, $q = 3$

2. $\alpha_{opt} = (5.69 \cdot 10^2)n^{-\frac{2}{3}}$

De zojuist gevonden waarde van de parameter α wordt nu gebruikt om de penalized maximum likelihood schatter $\hat{g}_n$ en de bijbehorende schatter $\hat{F}_X$ te construeren. In figuur 5.1 zijn deze schatters weergegeven. Figuur 5.1A toont de penalized maximum likelihood schatter $\hat{g}_n$. Figuur 5.1B toont de schatter $\hat{F}_X$. Figuur 5.1B toont dat de kans dat een vrouw minder dan 15 maanden nodig heeft om in verwachting te raken gelijk is aan 0.5.



Figuur 5.1: A en B tonen respectievelijk de penalized maximum likelihood schatter $\hat{g}_n$ en de schatter $\hat{F}_X$ op basis van de data van Slama (gebruik gemaakt van methode 1).

Merk op dat de bijbehorende R-code van de methode is terug te vinden in de appendix.

5.2 Methode 2

In deze paragraaf wordt de verdeling van de tijdsduur die een vrouw nodig heeft om zwanger te worden, verkregen door middel van de niet-parametrische maximum likelihood schatter in combinatie met de alternatieve schatter. Deze schatters staan beschreven in respectievelijk hoofdstuk 2 en 4. De niet-parametrische maximum likelihood schatter in combinatie met de alternatieve schatter wordt toegepast op de data van Slama. Om deze alternatieve schatter te vinden, dient de parameter h_n zo optimaal mogelijk te worden vastgesteld. Ter herinnering: $h_n = cn^{-\frac{1}{3}}$. Dit houdt in dat allereerst de waarde van de parameter c dient te worden bepaald.

Gewenst is om de optimale vorm van de parameter h_n op dezelfde wijze te verkrijgen als in paragraaf 4.3. In deze paragraaf is de onderliggende verdeling van de data bekend. Dit maakt het betrekkelijk eenvoudig om de waarde van de parameter c vast te stellen. Echter nu is de onderliggende verdeling van de data onbekend.

Ook hier wordt de methode, zoals voorgesteld door Silverman, toegepast om de waarde van de parameter c vast te stellen. Aangenomen wordt dat de data uit de exponentiële verdeling is getrokken. Hierbij wordt de bijbehorende parameter λ geschat door middel van de parametrische maximum likelihood schatter $\hat{\lambda}$. Er blijkt te gelden $\hat{\lambda} = 3.21 \cdot 10^{-2}$. Onderstaand is dit nogmaals kort genoteerd.

$$Z_1, \dots, Z_n \stackrel{i.i.d.}{\sim} g$$

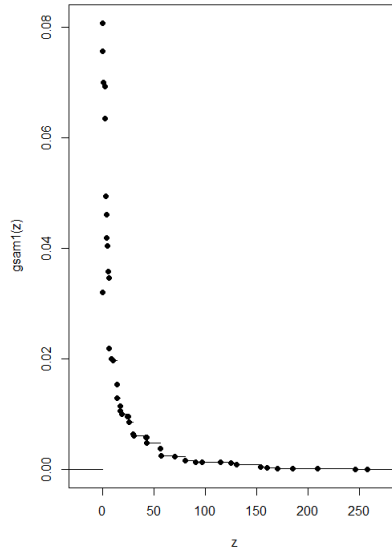
$$\stackrel{i.i.d.}{\sim} \text{Exp}(3.21 \cdot 10^{-2})$$

$$\stackrel{i.i.d.}{\sim} \begin{cases} (3.21 \cdot 10^{-2})e^{-(3.21 \cdot 10^{-2})z} & \text{als } z \geq 0 \\ 0 & \text{als } z < 0 \end{cases}$$

De waarde van de parameter c wordt nu als volgt gevonden. Voor de waarden $c = 0.05, 0.10, \dots, 999.95, 1000$ wordt de schatter T_n geconstrueerd. Voor elke waarde van de parameter c wordt vervolgens de mean squared error van de schatter T_n ten opzichte van de kansdichtheid $g(0)$ bepaald. Gekeken wordt voor welke waarde van de parameter c deze mean squared error minimaal is. Deze waarde wordt gekozen. De mean squared error blijkt minimaal te zijn voor de waarde $c = 1.69 \cdot 10^2$. De bijbehorende mean squared error is dan gelijk aan $2.77 \cdot 10^{-12}$. Hieruit volgt nu dat $h_n = (1.69 \cdot 10^2)n^{-\frac{1}{3}}$ optimaal is. Hieronder zijn de resultaten kort genoteerd.

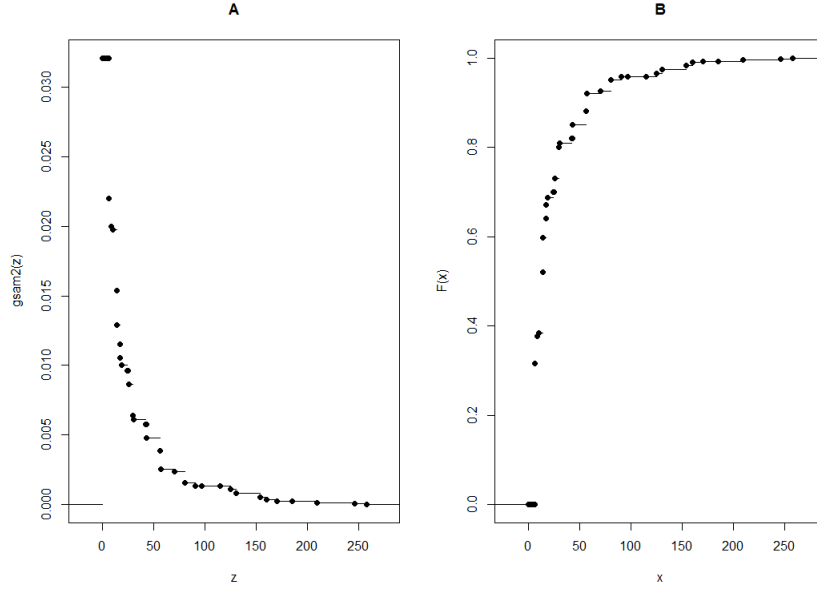
1. $c_{opt} = 1.69 \cdot 10^2$
2. $h_{n \text{ opt}} = (1.69 \cdot 10^2)n^{-\frac{1}{3}}$

De zojuist gevonden waarde van de parameter h_n wordt nu gebruikt om de schatter T_n te construeren. De schatter T_n is gelijk aan $3.21 \cdot 10^{-2}$. Tevens wordt de niet-parametrische maximum likelihood schatter $\hat{g}_n$ geconstrueerd. De waarde van de niet-parametrische maximum likelihood schatter $\hat{g}_n$ in het punt $z = 0$ wordt vervangen door de schatter T_n . Deze samengestelde schatter wordt aangeduid met $\hat{g}_{sam1}$. In figuur 5.2 is deze samengestelde schatter $\hat{g}_{sam1}$ weergegeven. Merk echter op dat de samengestelde schatter $\hat{g}_{sam1}$ geen dalende kansdichtheid is. Dit is te wijten aan het feit dat de niet-parametrische maximum likelihood schatter $\hat{g}_n$ systematisch een te grote waarde aanneemt in de rechteromgeving van het punt $z = 0$.



Figuur 5.2: De samengestelde schatter $\hat{g}_{sam1}$ op basis van de data van Slama (gebruik gemaakt van methode 2).

De volgende oplossing is gevonden voor dit problem. Alle waarden van de niet-parametrische maximum likelihood schatter $\hat{g}_n$ die groter zijn dan de schatter T_n , worden nu vervangen door de schatter T_n . Deze samengestelde schatter wordt aangeduid met $\hat{g}_{sam2}$. Tevens wordt de bijbehorende schatter $\hat{F}_X$ geconstrueerd. In figuur 5.3 zijn deze schatters weergegeven. Figuur 5.3A toont de samengestelde schatter $\hat{g}_{sam2}$. Figuur 5.3B toont de schatter $\hat{F}_X$. Figuur 5.3B toont dat de kans dat een vrouw minder dan 15 maanden nodig heeft om zwanger te worden gelijk is aan 0.5. Dit komt overeen met methode 1 uit paragraaf 5.1.



Figuur 5.3: A en B tonen respectievelijk de samengestelde schatter $\hat{g}_{sam2}$ en de schatter $\hat{F}_X$ op basis van de data van Slama (gebruik gemaakt van methode 2).

Merk op dat het vervangen van waarden van de niet-parametrische maximum likelihood door de schatter T_n tot gevolg heeft dat de eigenschap $\int_{z=0}^{\infty} \hat{g}_{sam2}(z)dz = 1$ niet meer geldt. Dit houdt in dat de samengestelde schatter $\hat{g}_{sam2}$ geen kansdichtheid is. Echter de schatter $\hat{F}_X$ voldoet aan alle eigenschappen waaraan een cumulatieve verdelingsfunctie behoort te voldoen. Onderstaand zijn deze eigenschappen genoteerd.

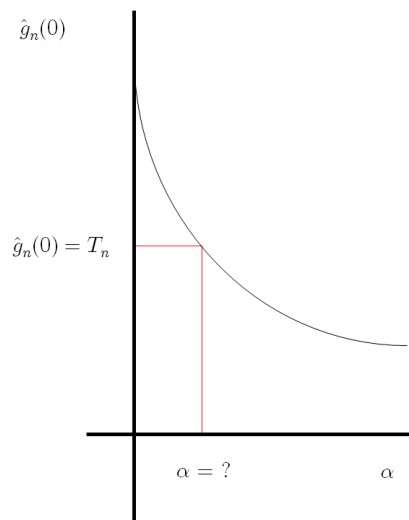
1. De schatter $\hat{F}_X$ heeft domein $[0, \infty]$ en bereik $[0, 1]$.
2. De schatter $\hat{F}_X$ is rechtscontinu.
3. De schatter $\hat{F}_X$ is monotoon stijgend.

Merk op dat de bijbehorende R-code van de methode is terug te vinden in de appendix.

5.3 Methode 3

In deze paragraaf wordt de verdeling van de tijdsduur die een vrouw nodig heeft om zwanger te worden, verkregen door middel van de penalized maximum likelihood schatter in combinatie met de alternatieve schatter. Deze schatters staan beschreven in respectievelijk hoofdstuk 3 en 4. De penalized maximum likelihood schatter in combinatie met de alternatieve schatter wordt toegepast op de data van Slama. Om de penalized maximum likelihood schatter te vinden, dient de parameter α zo optimaal mogelijk te worden vastgesteld.

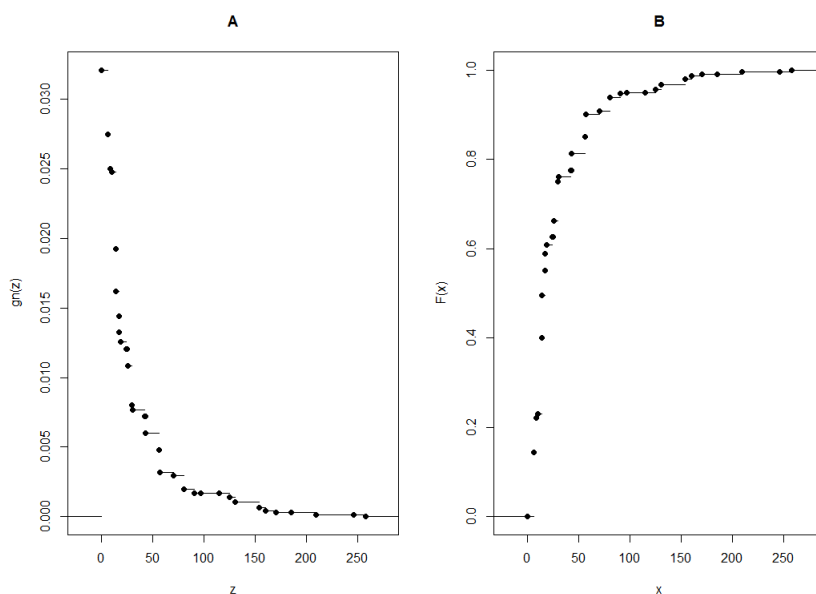
Een alternatieve methode om de optimale vorm van de parameter α te verkrijgen, is door handig gebruik te maken van de schatter T_n . In de vorige paragraaf is bepaald dat de schatter T_n gelijk is aan $3.21 \cdot 10^{-2}$. Gezocht wordt naar de waarde van de parameter α zodanig dat de waarde van de penalized maximum likelihood schatter $\hat{g}_n(0)$ gelijk is aan deze schatter T_n . In figuur 5.4 is deze methode schematisch weergegeven.



Figuur 5.4: Schematische weergave van methode 3.

De waarde van de parameter α wordt nu als volgt gevonden. Voor de waarden $\alpha = 0.05, 0.10, \dots, 199.95, 200$ wordt de penalized maximum likelihood schatter $\hat{g}_n$ geconstrueerd. Voor elke waarde van de parameter α wordt vervolgens de mean squared error van de penalized maximum likelihood schatter $\hat{g}_n(0)$ ten opzichte van schatter T_n bepaald. Gekeken wordt voor welke waarde van de parameter α deze mean squared error minimaal is. Deze waarde wordt gekozen. De mean squared error blijkt minimaal te zijn voor de waarde $\alpha = 6.25 \cdot 10^2$. De bijbehorende mean squared error is dan gelijk aan $3.50 \cdot 10^{-11}$.

De zojuist gevonden waarde van de parameter α wordt nu gebruikt om de penalized maximum likelihood schatter $\hat{g}_n$ en de bijbehorende schatter $\hat{F}_X$ te construeren. In figuur 5.5 zijn deze schatters weergegeven. Figuur 5.5A toont de penalized maximum likelihood schatter $\hat{g}_n$. Figuur 5.5B toont de schatter $\hat{F}_X$. Figuur 5.5B toont dat de kans dat een vrouw minder dan 15 maanden nodig heeft om in verwachting te raken gelijk is aan 0.5. Dit komt overeen met methode 1 uit paragraaf 5.1 en methode 2 uit paragraaf 5.2.



Figuur 5.5: A en B tonen respectievelijk de penalized maximum likelihood schatter $\hat{g}_n$ en de schatter $\hat{F}_X$ op basis van de data van Slama (gebruik gemaakt van methode 3).

Merk op dat de bijbehorende R-code van de methode is terug te vinden in de appendix.

6 — Conclusie

Het current duration model, dat staat beschreven in het artikel van Keiding [1], is een alternatieve methode om tot de verdeling van de tijdsduur die een vrouw nodig heeft om zwanger te worden te komen. Het current duration model is uitgebreid beschouwd en geanalyseerd. Aangetoond is dat de verdeling van de tijdsduur die een vrouw nodig heeft om zwanger te worden, al kan worden vastgesteld als enkel de verdeling van de current duration bekend is. De waarde in het punt 0 speelt hierbij een cruciale rol.

Vervolgens is gekeken hoe de verdeling van de current duration kan worden geschat. Hiertoe zijn de niet-parametrische maximum likelihood schatter, de penalized maximum likelihood schatter en een alternatieve schatter bestudeerd. Gekeken is naar onder andere de constructie, de consistentie en de inconsistentie van de schatters. Tevens zijn verschillende simulaties uitgevoerd om het gedrag van de schatters in kaart te brengen. In het bijzonder is gekeken naar het gedrag in het punt 0.

In het artikel van Slama is de data behorend bij het current duration model verzameld. Drie verschillende methodes, welke bestaan uit verschillende combinaties van de bestudeerde schatters, zijn toegepast op deze data. Zodoende zijn de penalized maximum likelihood schatter, de niet-parametrische maximum likelihood schatter in combinatie met de alternatieve schatter en de penalized maximum likelihood schatter in combinatie met de alternatieve schatter bekeken. Alle methoden lijken een adequate verdeling van de tijdsduur die een vrouw nodig heeft om zwanger te worden voort te brengen.

Geconcludeerd kan worden dat het current duration model een goed alternatief is om de verdeling van de tijdsduur die een vrouw nodig heeft om zwanger te worden te schatten, mits de juiste schatters op de juiste manier worden toegepast.

7 — Appendix

7.1 Parameter α

7.1.1 Standaard halve normale verdeling

Een steekproef uit de halve standaard normale verdeling wordt beschouwd. De bijbehorende kansdichtheid wordt weergegeven met g , de bijbehorende cumulatieve verdelingsfunctie wordt weergegeven met G . Door gebruik te maken van paragraaf 3.2 kan de optimale vorm van de parameter α worden afgeleid. Onderstaand is deze afleiding uitgewerkt.

1. Door de cumulatieve verdelingsfunctie G om te schrijven wordt de waarde van de parameter p verkregen. Onderstaand is dit omschrijvingsproces uitgewerkt. Merk op dat de waarde van de parameter p gelijk is aan 3.

$$\begin{aligned} G(z) &= P(0 \leq Z \leq z) \\ &= \int_{z=0}^z g(z) dz \\ &= \int_{z=0}^z \frac{2}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= \int_{z=0}^z \left(\frac{2}{\sqrt{2\pi}} - \frac{2z^2}{2\sqrt{2\pi}} + \frac{2z^4}{8\sqrt{2\pi}} - \frac{2z^6}{48\sqrt{2\pi}} + \frac{2z^8}{384\sqrt{2\pi}} + O(z^9) \right) dz \\ &= \int_{z=0}^z \left(\frac{2}{\sqrt{2\pi}} - \frac{z^2}{\sqrt{2\pi}} + \frac{z^4}{4\sqrt{2\pi}} - \frac{z^6}{24\sqrt{2\pi}} + \frac{z^8}{192\sqrt{2\pi}} + O(z^9) \right) dz \\ &= \left[\frac{2z}{\sqrt{2\pi}} - \frac{z^3}{3\sqrt{2\pi}} + \frac{z^5}{20\sqrt{2\pi}} - \frac{z^7}{168\sqrt{2\pi}} + \frac{z^9}{1728\sqrt{2\pi}} + O(z^{10}) \right]_{z=0}^z \\ &= \frac{2z}{\sqrt{2\pi}} - \frac{z^3}{3\sqrt{2\pi}} + \frac{z^5}{20\sqrt{2\pi}} - \frac{z^7}{168\sqrt{2\pi}} + \frac{z^9}{1728\sqrt{2\pi}} + O(z^{10}) \\ &= \frac{2}{\sqrt{2\pi}} z - \frac{1}{3\sqrt{2\pi}} z^3 + \frac{1}{20\sqrt{2\pi}} z^5 - \frac{1}{168\sqrt{2\pi}} z^7 + \frac{1}{1728\sqrt{2\pi}} z^9 + O(z^{10}) \end{aligned}$$

2. Door de zojuist gevonden waarde van de parameter p in te vullen in de onderstaande formule wordt de waarde van de parameter q verkregen. Hieronder is dit weergegeven.

$$q = \frac{1}{2p-1}$$
$$= \frac{1}{5}$$

3. Door de zojuist gevonden waarden van de parameters p en q in te vullen in de onderstaande formule wordt de optimale vorm van de parameter α verkregen. Op de volgende pagina is dit weergegeven.

$$\alpha = cn^{-pq}$$
$$= cn^{-\frac{3}{5}}$$

7.1.2 Standaard exponentiële verdeling

Een steekproef uit de standaard exponentiële verdeling wordt beschouwd. De bijbehorende kansdichtheid wordt weergegeven met g , de bijbehorende cumulatieve verdelingsfunctie wordt weergegeven met G . Door gebruik te maken van paragraaf 3.2 kan de optimale vorm van de parameter α worden afgeleid. Onderstaand is deze afleiding uitgewerkt.

1. Door de cumulatieve verdelingsfunctie G om te schrijven wordt de waarde van de parameter p verkregen. Onderstaand is dit omschrijvingsproces uitgewerkt. Merk op dat de waarde van de parameter p gelijk is aan 2.

$$\begin{aligned} G(z) &= P(0 \leq Z \leq z) \\ &= \int_{z=0}^z g(z) dz \\ &= \int_{z=0}^z e^{-z} dz \\ &= \int_{z=0}^z \left(1 - z + \frac{z^2}{2} - \frac{z^3}{6} + \frac{z^4}{24} - \frac{z^5}{120} + O(z^6) \right) dz \\ &= \left[z - \frac{z^2}{2} + \frac{z^3}{6} - \frac{z^4}{24} + \frac{z^5}{120} - \frac{z^6}{720} + O(z^7) \right]_{z=0}^z \\ &= z - \frac{z^2}{2} + \frac{z^3}{6} - \frac{z^4}{24} + \frac{z^5}{120} - \frac{z^6}{720} + O(z^7) \\ &= z - \frac{1}{2}z^2 + \frac{1}{6}z^3 - \frac{1}{24}z^4 + \frac{1}{120}z^5 - \frac{1}{720}z^6 + O(z^7) \end{aligned}$$

2. Door de zojuist gevonden waarde van de parameter p in te vullen in de onderstaande formule wordt de waarde van de parameter q verkregen. Hieronder is dit weergegeven.

$$\begin{aligned} q &= \frac{1}{2p-1} \\ &= \frac{1}{3} \end{aligned}$$

3. Door de zojuist gevonden waarden van de parameters p en q in te vullen in de onderstaande formule wordt de optimale vorm van de parameter α verkregen. Onderstaand is dit weergegeven.

$$\begin{aligned}\alpha &= cn^{-pq} \\ &= cn^{-\frac{2}{3}}\end{aligned}$$

7.2 R-codes

7.2.1 Simulatie inconsistentie (2.5)

```
1 | # Het laden van package fdrtool
2 | library("fdrtool")
3 |
4 | # Het laden van functies gn, MSError
5 | source('~ / Bachelorproject / R-Codes (SIMULATIE) / gn.R')
6 | source('~ / Bachelorproject / R-Codes (SIMULATIE) / MSError.R')
7 |
8 | s <- 1000
9 | onthoud <- c(1:s)
10 |
11 | # Het simuleren van 1000 steekproeven
12 | for(i in 1:s){
13 |
14 |   rij = 0
15 |
16 |   # Het genereren van een steekproef met 200 trekkingen
17 |   n <- 200
18 |   z <- sort(rexp(n, rate = 1))
19 |
20 |   # Het construeren van de niet-parametrische maximum likelihood schatter gren
21 |   gren <- gn(z)
22 |
23 |   onthoud[i] <- gren(0)
24 | }
25 |
26 | # Het berekenen van de mean squared error MSE
27 | MSE <- MSError(onthoud)
28 |
29 | # Het afdrukken van de mean squared error MSE
30 | MSE
31 |
32 | # Het plotten van de mean squared error MSE
33 | plot(c(1, 1000), c(MSE, MSE), col = 2, type = 'l', xlim = c(0,1000), ylim = c
34 |       (0, 15000), xlab = "i", ylab = "(gn(0) - g(0))^2")
35 |
36 | for(j in 1: s){
37 |   points(j, (onthoud[j]- dexp(0,rate=1))^2, pch=16)
```

7.2.2 Simulatie optimale parameter α (3.3)

```
1 # Het laden van package fdrtool
2 library("fdrtool")
3
4 # Het laden van functies gam, gn_pen, MSError, vind
5 source('~ / Bachelorproject / R-Codes (SIMULATIE) / gam.R')
6 source('~ / Bachelorproject / R-Codes (SIMULATIE) / gn_pen.R')
7 source('~ / Bachelorproject / R-Codes (SIMULATIE) / MSError.R')
8 source('~ / Bachelorproject / R-Codes (SIMULATIE) / vind.R')
9
10 # Het specificeren van k
11 begin <- 0.05
12 eind <- 6
13 stapgrootte <- 0.05
14 stappen <- eind / stapgrootte
15
16 s <- 1000
17 onthoud <- matrix(1, nrow = s, ncol = stappen)
18
19 # Het simuleren van 1000 steekproeven
20 for(i in 1:s){
21
22   rij = 0
23
24   # Het genereren van een steekproef met 200 trekkingen
25   n <- 200
26   z <- sort(rexp(n, rate = 1))
27   z_nieuw <- matrix(1, nrow = n, ncol = stappen)
28
29   for(k in seq(from=begin, to=eind, by=stapgrootte)){
30
31     rij = rij + 1
32
33     # Het declareren van de variabelen
34     alpha <- k*n^(-2/3)
35     gamma <- gam(n, z, alpha)
36
37     # Het transformeren van de data
38     for(j in 1:n){
39       z_nieuw[j, rij] <- alpha + gamma*z[j]
40     }
41
42     # Het construeren van de penalized maximum likelihood schatter pen
43     pen <- gn_pen(z_nieuw[,rij], alpha, gamma)
44
45     # Het opslaan van de waarde van de penalized maximum likelihood schatter pen
46     # in het punt z = 0
47     onthoud[i, rij] <- pen(0)
48   }
49
50 MSE <- matrix(1, nrow = 1, ncol = stappen)
51
52 # Het berekenen van de mean squared error MSE
53 for(l in 1:stappen){
54   MSE[1,l] <- MSError(onthoud[,l])
55 }
56
57 # Het plotten van de mean squared error MSE
58 plot(1*stapgrootte:stappen*stapgrootte, MSE[1,], pch=16, xlab = "c", ylab = '
59 MSE', main = 'MSE')
60
61 # Het afdrukken van de kleinste mean squared error MSE (met bijbehorende waarde
62 # van k)
63 vind(MSE)
```

7.2.3 Simulatie optimale parameter h_n (4.3)

```
1 | # Het laden van package fdrtool
2 | library("fdrtool")
3 |
4 | # Het laden van functies schatTn, MSError, vind
5 | source('~ / Bachelorproject / R-Codes (SIMULATIE) / schatTn.R')
6 | source('~ / Bachelorproject / R-Codes (SIMULATIE) / MSError.R')
7 | source('~ / Bachelorproject / R-Codes (SIMULATIE) / vind.R')
8 |
9 | # Het specificeren van k
10 | begin <- 0.05
11 | eind <- 6
12 | stapgrootte <- 0.05
13 | stappen <- eind / stapgrootte
14 |
15 | s <- 1000
16 | onthoud <- matrix(1, nrow = s, ncol = stappen)
17 |
18 | # Het simuleren van 1000 steekproeven
19 | for(i in 1:s){
20 |
21 |   rij = 0
22 |
23 |   # Het genereren van een steekproef met 200 trekkingen
24 |   n <- 200
25 |   z <- sort(rexp(n, rate = 1))
26 |   z_nieuw <- matrix(1, nrow = n, ncol = stappen)
27 |
28 |   for(k in seq(from=begin, to=eind, by=stapgrootte)){
29 |
30 |     rij = rij + 1
31 |
32 |     # Het declareren van de variabelen
33 |     h <- k * n^(-1/3)
34 |
35 |     # Het construeren van de alternatieve schatter Tn
36 |     Tn <- schatTn(z, h)
37 |
38 |     # Het opslaan van de waarde van de alternatieve schatter Tn
39 |     onthoud[i, rij] <- Tn
40 |   }
41 | }
42 |
43 | MSE <- matrix(1, nrow = 1, ncol = stappen)
44 |
45 | # Het berekenen van de mean squared error MSE
46 | for(l in 1:stappen){
47 |   MSE[1,l] <- MSError(onthoud[,l])
48 | }
49 |
50 | # Het plotten van de mean squared error MSE
51 | plot(1*stapgrootte:stappen*stapgrootte, MSE[1,], pch=16, xlab = "c", ylab = '
52 |   MSE', main = 'MSE')
53 | # Het afdrukken van de kleinste mean squared error MSE (met bijbehorende waarde
54 |   van k)
54 | vind(MSE)
```

7.2.4 Methode 1 (5.1)

```
1 | # Het laden van packages fdrtool, fitdistrplus
2 | library("fdrtool")
3 | library("fitdistrplus")
4 |
5 | # Het laden van functies gam, gn_pen, vind, omkeer
6 | source("~/Bachelorproject/R-Codes (SLAMA1)/gam.R")
7 | source("~/Bachelorproject/R-Codes (SLAMA1)/gn_pen.R")
8 | source("~/Bachelorproject/R-Codes (SLAMA1)/vind.R")
9 | source("~/Bachelorproject/R-Codes (SLAMA1)/omkeer.R")
10 |
11 |
12 | ##### DEEL 1 (het vinden van optimale k)
13 |
14 | # Het specificeren van k
15 | begin <- 0.05
16 | eind <- 1000
17 | stapgrootte <- 0.05
18 | stappen <- eind/stapgrootte
19 |
20 | onthoud <- c(1:stappen)
21 |
22 | # Het genereren van de data (Slama)
23 | z <- sort(scan("~/Bachelorproject/R-Codes (SLAMA1)/Slama.dat"))
24 | n <- length(z)
25 |
26 | # Het zoeken van de parameter lambda (exponentile verdeling) dmv parametrische
    maximum likelihood methode
27 | fit <- fitdist(z, "exp")
28 | lambda = fit$estimate[1]
29 |
30 | rij = 0
31 |
32 | for(k in seq(from=begin, to=eind, by=stapgrootte)){
33 |
34 |   rij = rij + 1
35 |
36 |   # Het declareren van de variabelen
37 |   alpha <- k*n^(-2/3)
38 |   gamma <- gam(n, z, alpha)
39 |   z_nieuw <- c(1:n)
40 |
41 |   # Het transformeren van de data
42 |   for(j in 1:n){
43 |     z_nieuw[j] <- alpha + gamma*z[j]
44 |   }
45 |
46 |   # Het construeren van de penalized maximum likelihood schatter pen
47 |   pen <- gn_pen(z_nieuw, alpha, gamma)
48 |
49 |   # Het opslaan van de waarde van de penalized maximum likelihood schatter pen in
    het punt z = 0
50 |   onthoud[rij] <- pen(0)
51 | }
52 |
53 | MSE <- c(1:stappen)
54 |
55 | # Het berekenen van de mean squared error MSE
56 | for(l in 1:stappen){
57 |
58 |   MSE[l] = (onthoud[l] - dexp(0, rate=lambda))^2
59 | }
60 |
61 | # Het plotten van de mean squared error MSE
```

```

62 plot(1*stapgrootte:stappen*stapgrootte, MSE, pch=16, xlab = "c", ylab = 'MSE',
63      main = 'MSE')
64 # Het afdrukken van de kleinste mean squared error MSE (met bijbehorende waarde
65   van k)
66 vind(MSE)
67
68 ##### DEEL 2 (het plotten met optimale k)
69
70 # Het genereren van de data (Slama)
71 z <- sort(scan("~/Bachelorproject/R-Codes (SLAMA1)/Slama.dat"))
72 n <- length(z)
73
74 # Het declareren van de variabelen
75 alpha <- vind(MSE)[2]*n^(-2/3)
76 gamma <- gam(n, z, alpha)
77 z_nieuw <- c(1:n)
78
79 # Het transformeren van de data
80 for(w in 1:n){
81   z_nieuw[w] <- alpha + gamma*z[w]
82 }
83
84 # Het construeren van de penalized maximum likelihood schatter pen
85 pen <- gn_pen(z_nieuw, alpha, gamma)
86
87 # Het construeren van de cumulatieve verdelingsfunctie F
88 finito <- omkeer(z_nieuw, alpha, gamma)
89
90 # Het plotten
91 par(mfrow=c(1,2))
92 plot(pen, pch=16, verticals=FALSE, xlab = "z", ylab = "gn(z)", main = "A")
93 plot(finito, pch=16, verticals=FALSE, xlab = "x", ylab = "F(x)", main = "B")

```

7.2.5 Methode 2 manier 1 (5.2)

```
1 # Het laden van packages fdrtool, fitdistrplus
2 library("fdrtool")
3 library("fitdistrplus")
4
5 # Het laden van functies schatTn, samen, vind, omkeer
6 source('~ / Bachelorproject / R-Codes (SLAMA2) / schatTn.R')
7 source('~ / Bachelorproject / R-Codes (SLAMA2) / samen.R')
8 source('~ / Bachelorproject / R-Codes (SLAMA2) / vind.R')
9 source('~ / Bachelorproject / R-Codes (SLAMA2) / omkeer.R')
10
11
12 ##### DEEL 1 (het vinden van optimale k)
13
14 # Het specificeren van k
15 begin <- 0.05
16 eind <- 1000
17 stapgrootte <- 0.05
18 stappen <- eind / stapgrootte
19
20 onthoud <- c(1:stappen)
21
22 # Het genereren van de data (Slama)
23 z <- sort(scan("~/Bachelorproject/R-Codes (SLAMA2)/Slama.dat"))
24 n <- length(z)
25
26 # Het zoeken van de parameter lambda (exponentiele verdeling) dmv parametrische
    maximum likelihood methode
27 fit <- fitdist(z, "exp")
28 lambda = fit$estimate[1]
29
30 rij = 0
31
32 for(k in seq(from=begin, to=eind, by=stapgrootte)){
33
34     rij = rij + 1
35
36     # Het declareren van de variabelen
37     h <- k*n^(-1/3)
38
39     # Het construeren van de alternatieve schatter Tn
40     Tn <- schatTn(z, h)
41
42     # Het opslaan van de waarde van de alternatieve schatter Tn
43     onthoud[rij] <- Tn
44 }
45
46 MSE <- c(1:stappen)
47
48 # Het berekenen van de mean squared error MSE
49 for(l in 1:stappen){
50
51     MSE[l] = (onthoud[l] - dexp(0, rate=lambda))^2
52 }
53
54 # Het plotten van de mean squared error MSE
55 plot(1*stapgrootte:stappen*stapgrootte, MSE, pch=16, xlab = "c", ylab = 'MSE',
    main = 'MSE')
56
57 # Het afdrukken van de kleinste mean squared error MSE (met bijbehorende waarde
    van k)
58 vind(MSE)
59
60
61 ##### DEEL 2 (het plotten met optimale k)
```

```

62 |
63 # Het genereren van de data (Slama)
64 z <- sort(scan("~/Bachelorproject/R-Codes (SLAMA2)/Slama.dat"))
65 n <- length(z)
66
67 # Het declareren van de variabelen
68 h <- vind(MSE)[2]*n^(-1/3)
69
70 # Het construeren van de alternatieve schatter Tn
71 Tn <- schatTn(z, h)
72
73 # Het construeren van de samengestelde schatter
74 samenvoegen <- samen(z, Tn)
75
76 # Het construeren van de cumulatieve verdelingsfunctie F
77 finito <- omkeer(z, Tn)
78
79 # Het plotten
80 par(mfrow=c(1,2))
81 plot(samenvoegen, pch=16, verticals=FALSE, xlab = "z", ylab = "gsam1(z)", main
      = "A")
82 plot(finito, pch=16, verticals=FALSE, xlab = "x", ylab = "F(x)", main = "B")

```

7.2.6 Methode 2 manier 2 (5.2)

```
1 # Het laden van packages fdrtool, fitdistrplus
2 library("fdrtool")
3 library("fitdistrplus")
4
5 # Het laden van functies schatTn, samenalt, vind, omkeeral
6 source("~/Bachelorproject/R-Codes (SLAMA2)/schatTn.R")
7 source("~/Bachelorproject/R-Codes (SLAMA2)/samenalt.R")
8 source("~/Bachelorproject/R-Codes (SLAMA2)/vind.R")
9 source("~/Bachelorproject/R-Codes (SLAMA2)/omkeeral.R")
10
11
12 ##### DEEL 1 (het vinden van optimale k)
13
14 # Het specificeren van k
15 begin <- 0.05
16 eind <- 1000
17 stapgrootte <- 0.05
18 stappen <- eind/stapgrootte
19
20 onthoud <- c(1:stappen)
21
22 # Het genereren van de data (Slama)
23 z <- sort(scan("~/Bachelorproject/R-Codes (SLAMA2)/Slama.dat"))
24 n <- length(z)
25
26 # Het zoeken van de parameter lambda (exponentile verdeling) dmv parametrische
    maximum likelihood methode
27 fit <- fitdist(z, "exp")
28 lambda = fit$estimate[1]
29
30 rij = 0
31
32 for(k in seq(from=begin, to=eind, by=stapgrootte)){
33
34     rij = rij + 1
35
36     # Het declareren van de variabelen
37     h <- k*n^(-1/3)
38
39     # Het construeren van de alternatieve schatter Tn
40     Tn <- schatTn(z, h)
41
42     # Het opslaan van de waarde van de alternatieve schatter Tn
43     onthoud[rij] <- Tn
44 }
45
46 MSE <- c(1:stappen)
47
48 # Het berekenen van de mean squared error MSE
49 for(l in 1:stappen){
50
51     MSE[l] = (onthoud[l] - dexp(0, rate=lambda))^2
52 }
53
54 # Het plotten van de mean squared error MSE
55 plot(1*stapgrootte:stappen*stapgrootte, MSE, pch=16, xlab = "c", ylab = 'MSE',
    main = 'MSE')
56
57 # Het afdrukken van de kleinste mean squared error MSE (met bijbehorende waarde
    van k)
58 vind(MSE)
59
60
61 ##### DEEL 2 (het plotten met optimale k)
```

```

62 |
63 # Het genereren van de data (Slama)
64 z <- sort(scan("~/Bachelorproject/R-Codes (SLAMA2)/Slama.dat"))
65 n <- length(z)
66
67 # Het declareren van de variabelen
68 h <- vind(MSE)[2]*n^(-1/3)
69
70 # Het construeren van de consistente schatter Tn
71 Tn <- schatTn(z, h)
72
73 # Het construeren van de samengestelde schatter
74 samenvoegen <- samenalt(z, Tn)
75
76 # Het construeren van de cumulatieve verdelingsfunctie F
77 finito <- omkeeralt(z, Tn)
78
79 # Het plotten
80 par(mfrow=c(1,2))
81 plot(samenvoegen, pch=16, verticals=FALSE, xlab = "z", ylab = "gsam2(z)", main
      = "A")
82 plot(finito, pch=16, verticals=FALSE, xlab = "x", ylab = "F(x)", main = "B")

```

7.2.7 Methode 3 (5.3)

```
1 # Het laden van packages fdrtool, fitdistrplus
2 library("fdrtool")
3 library("fitdistrplus")
4
5 # Het laden van functies gam, gn_pen, vind, omkeer
6 source('~ / Bachelorproject / R-Codes (SLAMA3) / gam.R')
7 source('~ / Bachelorproject / R-Codes (SLAMA3) / gn_pen.R')
8 source('~ / Bachelorproject / R-Codes (SLAMA3) / vind.R')
9 source('~ / Bachelorproject / R-Codes (SLAMA3) / omkeer.R')
10
11
12 ##### DEEL 1 (het vinden van optimale alpha)
13
14 # Het specificeren van k
15 begin <- 0.05
16 eind <- 200
17 stapgrootte <- 0.05
18 stappen <- eind / stapgrootte
19
20 onthoud <- c(1:stappen)
21
22 # Het genereren van de data (Slama)
23 z <- sort(scan("~/Bachelorproject/R-Codes (SLAMA3)/Slama.dat"))
24 n <- length(z)
25
26 # Het declareren van de alternatieve schatter Tn
27 Tn <- 0.03209432
28
29 rij = 0
30
31 for(k in seq(from=begin, to=eind, by=stapgrootte)){
32
33   rij = rij + 1
34
35   # Het declareren van de variabelen
36   alpha <- k
37   gamma <- gam(n, z, k)
38   z_nieuw <- c(1:n)
39
40   # Het transformeren van de data
41   for(j in 1:n){
42     z_nieuw[j] <- alpha + gamma*z[j]
43   }
44
45   # Het construeren van de penalized maximum likelihood schatter pen
46   pen <- gn_pen(z_nieuw, alpha, gamma)
47
48   # Het opslaan van de waarde van de penalized maximum likelihood schatter pen in
49   # het punt z = 0
50   onthoud[rij] <- pen(0)
51 }
52 MSE <- c(1:stappen)
53
54 # Het berekenen van de mean squared error MSE
55 for(l in 1:stappen){
56
57   MSE[l] = (onthoud[l] - Tn)^2
58 }
59
60 # Het plotten van de mean squared error MSE
61 plot(1*stapgrootte:stappen*stapgrootte, MSE, pch=16, xlab = "alpha", ylab = '
62   MSE', main = 'MSE')
```

```

63 # Het afdrukken van de kleinste mean squared error MSE (met bijbehorende waarde
    van k)
64 vind(MSE)
65
66
67 ##### DEEL 2 (het plotten met optimale alpha)
68
69 # Het genereren van de data (Slama)
70 z <- sort(scan("~/Bachelorproject/R-Codes (SLAMA3)/Slama.dat"))
71 n <- length(z)
72
73 # Het declareren van de variabelen
74 alpha <- vind(MSE)[2]
75 gamma <- gam(n, z, alpha)
76 z_nieuw <- c(1:n)
77
78 # Het transformeren van de data
79 for(w in 1:n){
80   z_nieuw[w] <- alpha + gamma*z[w]
81 }
82
83 # Het construeren van de penalized maximum likelihood schatter pen
84 pen <- gn_pen(z_nieuw, alpha, gamma)
85
86 # Het construeren van de cumulatieve verdelingsfunctie F
87 finito <- omkeer(z_nieuw, alpha, gamma)
88
89 # Het plotten
90 par(mfrow=c(1,2))
91 plot(pen, pch=16, verticals=FALSE, xlab = "z", ylab = "gn(z)", main = "A")
92 plot(finito, pch=16, verticals=FALSE, xlab = "x", ylab = "F(x)", main = "B")

```

7.2.8 Functies

1.

```
1 gam <- function(n, z, alpha){
2
3   gamma <- 1000
4
5   for(i in 1:n){
6     if(((1/2)*(1-(alpha/z[i])) + sqrt((alpha/(2*z[i]))^2 + (alpha/(2*z[i]
7       ))*(1-((2*i)/n))+ 1/4)) < gamma){
8       gamma <- (1/2)*(1-(alpha/z[i])) + sqrt((alpha/(2*z[i]))^2 + (alpha/
9         (2*z[i]))*(1-((2*i)/n))+ 1/4)
10    }
11  }
12 }
```

2.

```
1 gn <- function(z){
2
3   # Het construeren van de empirische verdelingsfunctie Gn
4   Gn <- ecdf(z)
5
6   # Het construeren van de kleinste concave majorant LCM
7   LCM <- gcmlcm(unique(c(0,z)), Gn(unique(c(0,z))), type = c("lcm"))
8
9   # Het construeren van de niet-parametrische maximum likelihood schatter
10  gn <- stepfun(LCM$x.knots, c(0,LCM$slope.knots,0))
11
12  gn
13 }
```

3.

```
1 gn_pen <- function(z, alpha, gamma){
2
3   # Het construeren van de empirische verdelingsfunctie Gn
4   Gn <- ecdf(z)
5
6   # Het construeren van de kleinste concave majorant LCM
7   LCM <- gcmlcm(unique(c(0,z)), Gn(unique(c(0,z))), type = c("lcm"))
8
9   # Het construeren van de niet-parametrische maximum likelihood schatter
10  gn <- stepfun(LCM$x.knots, c(0,LCM$slope.knots,0))
11
12  # Het construeren van de penalized maximum likelihood schatter gn_pen
13  gn_pen <- stepfun(c(0, (LCM$x.knots[2:length(LCM$x.knots)] - alpha)/
14    gamma), c(0,LCM$slope.knots,0))
15
16  gn_pen
17 }
```

4.

```
1 MSEError <- function(onthoud){
2
3   MSE = 0
4
5   for(i in 1: s){
6     MSE = MSE + (1/s)*(onthoud[i] - dexp(0, rate=1))^2
7   }
8
9   MSE
10 }
```

5. (a) Gebruikt bij methode 1 (5.1), methode 3 (5.3).

```
1 omkeer <- function(z, alpha, gamma){
2
3   # Het construeren van de empirische verdelingsfunctie Gn
4   Gn <- ecdf(z)
5
6   # Het construeren van de kleinste concave majorant LCM
7   LCM <- gcmlcm(unique(c(0,z)), Gn(unique(c(0,z))), type = c("lcm"))
8
9   # Het construeren van de niet-parametrische maximumlikelihood
   schatter gn
10  gn <- stepfun(LCM$x.knots, c(0,LCM$slope.knots,0))
11
12  # Het construeren van de penalized maximum likelihood schatter gn_
   pen
13  gn_pen <- stepfun(c(0, (LCM$x.knots[2:length(LCM$x.knots)] - alpha)
   /gamma), c(0,LCM$slope.knots,0))
14
15  # Het construeren van de cumulatieve verdelingsfunctie F
16  omkeer <- stepfun(c(0, (LCM$x.knots[2:length(LCM$x.knots)] - alpha)
   /gamma), c(0, (1 - LCM$slope.knots/LCM$slope.knots[1]),1))
17
18  omkeer
19 }
```

(b) Gebruikt bij methode 2 (5.2).

```
1 omkeer <- function(z, Tn){
2
3   # Het construeren van de empirische verdelingsfunctie Gn
4   Gn <- ecdf(z)
5
6   # Het construeren van de kleinste concave majorant LCM
7   LCM <- gcmlcm(unique(c(0,z)), Gn(unique(c(0,z))), type = c("lcm"))
8
9   LCM$slope.knots[1] <- Tn
10
11  # Het construeren van de niet-parametrische maximum likelihood
   schatter gn
12  gn <- stepfun(LCM$x.knots, c(0,LCM$slope.knots,0))
13
14  # Het construeren van de cumulatieve verdelingsfunctie F
```

```

15 | omkeer <- stepfun(LCM$x.knots, c(0, (1 - LCM$slope.knots/LCM$slope.
    | knots[1]),1))
16 |
17 | omkeer
18 | }

```

6.

```

1 | omkeeralt <- function(z, Tn){
2 |
3 | # Het construeren van de empirische verdelingsfunctie Gn
4 | Gn <- ecdf(z)
5 |
6 | # Het construeren van de kleinste concave majorant LCM
7 | LCM <- gcmlcm(unique(c(0,z)), Gn(unique(c(0,z))), type = c("lcm"))
8 |
9 | for(i in 1:length(LCM$slope.knots)){
10 |   if(LCM$slope.knots[i] > Tn){
11 |     LCM$slope.knots[i] <- Tn
12 |   }
13 | }
14 |
15 | # Het construeren van de niet-parametrische maximum likelihood schatter
    | gn
16 | gn <- stepfun(LCM$x.knots, c(0,LCM$slope.knots,0))
17 |
18 | # Het construeren van de cumulatieve verdelingsfunctie F
19 | omkeer <- stepfun(LCM$x.knots, c(0, (1 - LCM$slope.knots/LCM$slope.
    | knots[1]),1))
20 |
21 | omkeer
22 | }

```

7.

```

1 | samen <- function(z, Tn){
2 |
3 | # Het construeren van de empirische verdelingsfunctie Gn
4 | Gn <- ecdf(z)
5 |
6 | # Het construeren van de kleinste concave majorant LCM
7 | LCM <- gcmlcm(unique(c(0,z)), Gn(unique(c(0,z))), type = c("lcm"))
8 |
9 | LCM$slope.knots[1] <- Tn
10 |
11 | # Het construeren van de niet-parametrische maximum likelihood schatter
    | gn
12 | gn <- stepfun(LCM$x.knots, c(0,LCM$slope.knots,0))
13 |
14 | gn
15 | }

```

8.

```
1 | samenalt <- function(z, Tn){
2 |
3 |   # Het construeren van de empirische verdelingsfunctie Gn
4 |   Gn <- ecdf(z)
5 |
6 |   # Het construeren van de kleinste concave majorant LCM
7 |   LCM <- gcmlcm(unique(c(0,z)), Gn(unique(c(0,z))), type = c("lcm"))
8 |
9 |   for(i in 1:length(LCM$slope.knots)){
10 |     if(LCM$slope.knots[i] > Tn){
11 |       LCM$slope.knots[i] <- Tn
12 |     }
13 |   }
14 |
15 |   # Het construeren van de niet-parametrische maximum likelihood schatter
16 |   gn <- stepfun(LCM$x.knots, c(0,LCM$slope.knots,0))
17 |
18 |   gn
19 | }
```

9.

```
1 | schatTn <- function(z, h){
2 |
3 |   # Het construeren van de empirische verdelingsfunctie Gn
4 |   Gn <- ecdf(z)
5 |
6 |   # Het construeren van de schatter Tn
7 |   schatTn <- (1/h)*Gn(h)
8 |
9 |   schatTn
10 | }
```

10.

- (a) Gebruikt bij simulatie inconsistentie (2.5), simulatie optimale parameter α (3.3), simulatie optimale parameter h_n (4.3).

```
1 | vind <- function(MSE){
2 |
3 |   vind <- c(1:2)
4 |   vind[1] <- 1000
5 |   vind[2] <- 0
6 |
7 |   for(i in 1:stappen){
8 |     if(MSE[1,i] < vind[1]){
9 |       vind[1] <- MSE[1, i]
10 |       vind[2] <- i*stapgrootte
11 |     }
12 |   }
13 |
14 |   vind
15 | }
```

(b) Gebruikt bij methode 1 (5.1), methode 2 (5.2), methode 3 (5.3).

```
1 vind <- function(MSE){  
2  
3   vind <- c(1:2)  
4   vind[1] <- 1000  
5   vind[2] <- 0  
6  
7   for(i in 1:stappen){  
8     if(MSE[i] < vind[1]){  
9       vind[1] <- MSE[i]  
10      vind[2] <- i*stapgrootte  
11    }  
12  }  
13  
14  vind  
15 }
```

Bibliografie

- [1] N. Keiding (2012): *The Current Duration Approach to Estimating Time to Pregnancy*, Scandinavian Journal of Statistics, Vol. 39, No. 2, pp. 185 - 204.
- [2] G. Jongbloed (2013): *Bio- and Environmental Statistics: Modeling and analysis of time-to-event data*, TU Delft, pp. 44 - 46.
- [3] U. Grenander (1956): *On the theory of mortality measurement II*, Skandinavisk Aktuarietidskrift, Vol. 39, pp. 125 - 153.
- [4] J. A. Rice (2007): *Mathematical Statistics and Data Analysis*, Brooks/Cole, pp. 378 - 380.
- [5] P. P. B. Eggermont en V. N. LaRiccia (2001): *Maximum Penalized Likelihood Estimation Volume I: Density Estimation*, Springer Series in Statistics, pp. 225.
- [6] G. Jongbloed (2013): *Bio- and Environmental Statistics: Modeling and analysis of time-to-event data*, TU Delft, pp. 75 - 78.
- [7] G. Jongbloed (2013): *Bio- and Environmental Statistics: Modeling and analysis of time-to-event data*, TU Delft, pp. 79 - 83.
- [8] M. Woodroffe en J. Sun (1993): *A penalized maximum likelihood estimate of $f(0+)$ when f is non-increasing*, Statistica Sinica, Vol. 3, No. 2, pp. 501-515.
- [9] J. A. Rice (2007): *Mathematical Statistics and Data Analysis*, Brooks/Cole, pp. 133- 134.
- [10] G. Jongbloed (2013): *Bio- and Environmental Statistics: Modeling and analysis of time-to-event data*, TU Delft, pp. 84 - 85.

- [11] R. Slama (2012): *Estimation of the frequency of involuntary infertility on a nation-wide basis*, Human Reproduction, Vol. 27, No. 5, pp. 1489 - 1498.
- [12] B.W. Silverman (1986): *Density estimation for statistics and data analysis*, CRC Press, Vol. 26.