K O N I N K L I J K N E D E R L A N D S
M E T E O R O L O G I S C H I N S T I T U U T

Wetenschappelijk Rapport W.R. 57-001 (III-197)

Dr. C. Levert

Het normaliseren van een niet-normale verdeling.
Normalization of a non-normal distribution.

De Bilt, 1957

KONINKLIJK NEDERLANDS
METEOROLOGISCH INSTITUUT

Wetenschappelijk Rapport
W.R. 57-001 (III-197)
U.D.C.: 519.2

Dr. C. Levert

Het normaliseren van een niet-normale verdeling.
Normalization of a non-normal distribution.

INHOUD

0. Inleiding; waarom "normaliseren"?

Het komt herhaaldelijk voor, dat men "statistisch gehandicapt" is door het feit, dat de onderhavige variable $\underline{x}$ *) niet een normale verdeling volgt, ook niet in voldoende benadering. In vele gevallen moet men weten, dat aan voorwaarde van normaliteit voldaan is om bepaalde statistische "methodes" (toetsingen enz.) te mogen toepassen. Enige voorbeelden:

0.1 Men wil een grootheid $\underline{x}$ met een andere grootheid $\underline{y}$ correleren. Niets belet natuurlijk een berekening van de klassieke correlatiecoëfficiënt r (c.c.) , berustende op de sommen $\sum x, \sum y, \sum x^2, \sum y^2, \sum x.y$, doch wil men weten welke de "betekenis" deze r heeft, dan moet men de synificantietabel raadplegen. Deze onderstelt echter, dat men reeds weet, dat zowel $\underline{x}$ - als het $\underline{y}$ - universum een normale verdeling heeft.

0.2 Men wil onderzoeken of een tijdreeks $\underline{x}$ vrij van persistentie, d.i. auto- of interne correlatie, is en berekent daartoe de c.c- tussen de reeksen $x_1, x_2, x_3, \ldots$ en $x_2, x_3, x_4 \ldots$ en die tussen de reeksen $x_1, x_2, x_3, \ldots$ en $x_3, x_4, x_5 \ldots$ enz.(d.w.z. autocorrelatiecoëfficiënten van orden 1, 2, enz.) Weer eist het raadplegen van de significantie-tabel, dat $\underline{x}$ normaal verdeeld is.

0.3 Men wil berekenen tussen welke grenzen de overschrijdingskans van zekere x ligt in een $\underline{x}$-populatie, waarover wij slechts op basis van N metingen informatie bezitten. Nodig is dan dat de $\underline{x}$-verdeling normaal is.

0.4 Men wil de t-toets op het verschil van de gemiddelden van twee reeksen metingen toepassen. Elk der populaties, waaruit de twee reeksen komen, moet dan een normale verdeling hebben.

0.5 Men wil de variantie-analyse toepassen. Weer is de normaliteit der verdeling een eis.

0.6 Men wil de toets van Bartlett toepassen (: is $\sigma_1 = \sigma_2 = \ldots \sigma_n$?), d.w.z. men wil onderzoeken of de n populaties, waaruit de n reeksen van metingen, t.w. $x_{11}, x_{12} \ldots x_{1,a_1}$ (a_1 stuks); $x_{21}, x_{22} \ldots x_{2,a_2}$ (a_2 stuks) ... $x_{n1}, x_{n2} \ldots x_{n,a_n}$ (a_n stuks) stammen, ongelijke varianties bezitten. De toets onderstelt, dat men weet dat deze populaties normale verdelingen hebben.

---

*) De grootheid $\underline{x}$ is onderstreept als ze een waarschijnlijkheidsverdeling volgt (een stochastische of statistische variable, een "stochastiek") Een speciale waarde ervan is niet onderstreept.

0.7 Men wil de z - of F-toets toepassen. (Zie "Statistische Toetsingsmethoden" R III 120 1953)

0.8 De toepassing van vele "uitwerpingskriteria" eist, dat het universum normaal verdeeld is (de parametervrije kriteria eisen dit niet). Voor een uitschieterskriterium en een daarop gebaseerd nomogram, zie (1), waarin tevens een lijst met artikelen over uitschieterskriteria opgenomen is.

0.9 Men wil aan de vorm van de transformatiefunctie $z=\psi(x)$, zie verderop, zekere physische beschouwingen vastknopen.

In sommige gevallen, althans bij "toetsingen", is het mogelijk om parameter-vrije (d.i. verdelingsvrije) toetsen te gebruiken, waarbij geen (of zeer weinige en weinig strenge) onderstellingen gemaakt worden, maar men dient steeds te bedenken, dat deze toetsingen dientengevolge minder onderscheidend zijn dan de niet-parametervrije toetsen, die men toepassen mag wanneer men op grond van neveninformatie of van afzonderlijke onderzoekingen wèl weet dat aan deze voorwaarden gehoorzaamd wordt.

1. Het normaliseren heeft twee aspecten.

Het normaliseren heeft twee verschillende aspecten:

1. de kansverdeling in het universum is exact bekend en
2. de kansverdeling is niet exact bekend.

Wij verkeren in de praktijk gewoonlijk in het tweede geval. Omdat de behandeling van het tweede geval het duidelijkst is na die van het eerste, beginnen wij met aan het eerste aandacht te schenken, ook al maakt dit wellicht een zeer theoretische, academische indruk.

1.1 Het eerste geval; $f(\underline{x})d\underline{x}$ is exact bekend.

Denk gegeven de exacte distributieve kansverdeling $f(\underline{x})d\underline{x}$ (de z.g. kansdichtheid) of de (cumulatieve) kansverdeling (onderschrijdingskans) $\mathbb{P}(\underline{x} \leq x) = \int_a^x f(\underline{x})dx = F(x)$ in het universum van de $\underline{x}$-variable, beide gedefinieerd voor een gegeven traject: $a \leq \underline{x} \leq b$. Met $-\infty < a < \infty$ en $-\infty < b < \infty$ Dit betekent: $\int_a^b f(\underline{x})d\underline{x} = 1$, d.i. $F(b)=1$ en $F(a)=0$.; $F(x)$ zal een monotoon stijgende functie zijn en natuurlijk is $f(x) \geq 0$ voor elke $a \leq x \leq b$. Beide verdelingen kunnen gegeven zijn op twee manieren:

1. d.m.v. een analystische uitdrukking òf voor de kansdichtheid òf voor de kansverdeling. Enkele voorbeelden:

a) $f(\underline{x})d\underline{x} = \frac{1}{3\sqrt{2\pi}} \exp.\left\{-(\underline{x}-8)^2/18\right\} dx$, d.i. een normale verdeling $\mathcal{N}(8;3)$, waarvoor $-\infty < x < \infty$ en $\int_{-\infty}^{\infty} f(\underline{x})d\underline{x} = 1$; gemiddelde $= \mu = \mathcal{E}_{\underline{x}} = \int_{-\infty}^{\infty} \underline{x} f(\underline{x})dx = 8$ en variantie $= \sigma^2 = \mathcal{E}(\underline{x} - \mu)^2 = \int_{-\infty}^{\infty} (\underline{x}-8)^2 f(\underline{x})d\underline{x} = 3^2$

b) $f(\underline{x})d\underline{x} = \frac{\underline{x}^{\frac{1}{2}n-1} e^{-\frac{1}{2}\underline{x}}}{(\sqrt{2})^n \Gamma(\frac{1}{2}n)} d\underline{x}$,d.i. de χ^2-verdeling met n graden van vrijheid, waarvoor $o \leqslant x \leqslant \infty$ en $\int_{\infty}^{\infty} f(\underline{x})d\underline{x} = 1$; $\mu = \mathcal{E}_{\underline{x}} = \int_{-\infty}^{\infty} \underline{x}\, f(\underline{x})d\underline{x} = n$ en $\sigma^2 = \int_{\infty}^{\infty} (\underline{x}-\mu)^2 f(\underline{x})d\underline{x} = 2n$

c) $f(\underline{x})dx = \frac{\exp\{-(\underline{x}^2-3\underline{x}+2)\}}{\int_1^{10} \exp\{-(\underline{x}^2-3\underline{x}+2)\}\,dx}$ voor $1 \leqslant \underline{x} \leqslant 10$

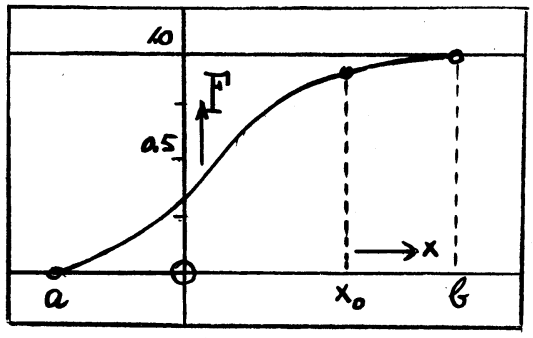
d) $F(\underline{x})dx = 1-e^{-(\underline{x}-3)^2/4}$, d.i. een der typen goodrich-verdelingen, gedefinieerd in het traject $3 \leqslant x < \infty$, zodat $\int_3^x f(\underline{x})d\underline{x} = 1-\exp.\{-(\underline{x}-3)^2/4\}$

2. als een kromme, getekend in een dubbellineair net (langs de ene as de x in een lineaire schaal, langs de andere as de F eveneens in lineaire schaal) of in een lineair-waarschijnlijkheidsnet (langs de ene as de x in een lineaire schaal, langs de andere as de F in een niet-lineaire, z.g. gausz-schaal, waarover verderop meer).

In geval 1ste kunnen wij bij iedere gegeven x_o, waarvoor $a \leqslant x_o \leqslant b$, de <u>onderschrijdingskans berekenen</u> òf als $\int_a^{x_o} f(\underline{x})d\underline{x}$, indien $f(\underline{x})$ gegeven is, òf als $F(\underline{x}_o)$, indien F (x) gegeven is, zo nauwkeurig als men maar wil. In geval 2de lezen wij de $F(x_o)$ af van het prentje (zie hieronder, A). Daarbij denken wij ons de kromme F(x) tegen x over het gehele traject $a \rightarrow b$ getekend (zeker als a én b eindig is). Het kan echter zijn, bijv. als a eindig en b positief oneindig is, dat de kromme F(x) tegen x niet over het gehele traject (dat dan immers oneindig uitgestrekt is) getekend gegeven is (zie hieronder, B). Beschouwt men nu een x_o-waarde, groter dan de grootste met de kromme corresponderen x-waarde (x_m), dan wordt er niet een scherp bepaald punt op de kromme gefixeerd. In dit geval moet de getekende curve uit de hand geëxtropoleerd worden, hetgeen over kleine afstanden wellicht in voldoende mate betrouwbaar is, over grote echter niet.

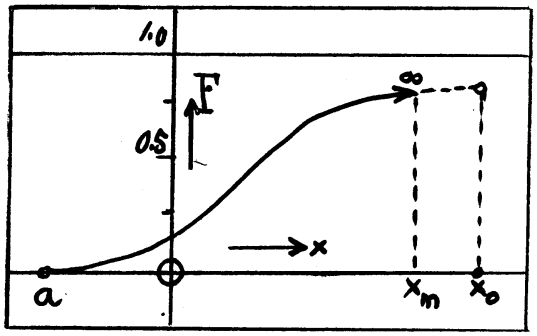
De volgende schetsjes mogen een en ander verduidelijken.

A



a én b eindig

B



a eindig ; $b = +\infty$

Definitie van normale verdeling.

Onderstel x is zelf niet normaal verdeeld. Wij spreken af een variabele y normaal verdeeld te noemen , als er tenminste één stel waarden α en β bestaat, zodanig dat voor iedere specifieke waarde y_i uit het traject $-\infty \leqslant y \leqslant \infty$ de kans op y gelegen tussen y_i en $y_i + dy_i$ geschreven kan worden als

$$g(y_i)dy_i = \frac{1}{\beta\sqrt{2\pi}} e^{\frac{-(y_i-\alpha)^2}{2\beta^2}} dy_i$$

Dat daarbij deze constanten α en β^2 de speciale, eenvoudige betekenis hebben, t.w. gemiddelde $= \mathcal{E}_{\underline{y}} = \int_{-\infty}^{\infty} \underline{y} g(\underline{y})\, d\underline{y}$ resp. variantie $= \mathcal{E}(\underline{y} - \mathcal{E}_{\underline{y}})^2 = \sigma_y^2 = \int_{-\infty}^{\infty} (\underline{y}-\alpha)^2 g(\underline{y})\, d\underline{y}$, is eigenlijk min of meer bijkomstig.

Wat is normaliseren?[1])

Gesteld, $\underline{x}$ is zelf niet normaal verdeeld, is er dan tenminste één functie van x, $z = \psi(x)$, zodanig, dat $\underline{z}$ wel normaal verdeeld is? Het transformeren d.m.v. $z = \psi(x)$ heet dan normaliseren; deze $\psi(x)$ heet de normalisatie-transformatie. Kapteyn en Van Uven, zie [4] ,spreken van "normale functie", welke term wij liever vermijden.

N.B. Als $z = \psi(x)$ en $x = h(z)$ = inversie functie, dan is $f(\underline{x})d\underline{x} = \left[f\{h(z)\}\frac{dh(z)}{dz}\right] d\underline{z}$. De onderschrijdingskans $\int_a^{x_i} f(\underline{x})d\underline{x}$ van één gegeven x_i in de gegeven x-verdeling moet gelijk zijn aan die van $z_i = \psi(x_i)$ in de normale $\underline{z}$-verdeling $g(\underline{z})d\underline{z}$, d.i. $\int_{-\infty}^{z_i} g(\underline{z})d\underline{z}$, en wel voor iedere x_i uit het traject $a \leqslant x \leqslant b$. Door dit voorschrift is aan iedere x-waarde één z-waarde toegevoegd en omgekeerd.

1.2 Het tweede geval; $f(\underline{x})d\underline{x}$ te schatten uit steekproef

Er zijn N "metingen" $x_1, x_2 \ldots x_n$ verricht; de kansverdeling $f(\underline{x})d\underline{x}$ van $\underline{x}$ in het universum is onbekend. Wij hebben misschien een vermoeden, dat ze van een bekend type is (met onbekende parameters), of zelfs dit vermoeden is er niet. Laten wij onderstellen, dat deze $f(\underline{x})d\underline{x}$ geen normale verdeling voorstelt.

---

1) Men zou, zo men wil (maar dit is zeer ongebruikelijk), ook kunnen "Pearson-niseren", d.w.z. zoeken naar zulk een transformatie $z = \phi(x)$, dat de z verdeeld is volgens de Pearson-verdeling van type III, gedefinieerd door $p(\underline{z})d\underline{z} = \frac{2}{\beta\sqrt{2\pi}} \underline{z}^{\alpha-1} e^{-\underline{z}^{2\alpha}/2\beta^2} d\underline{z}$ met $0 \leqslant z \leqslant \infty$, en α en β reële constanten. ($\beta > 0$) De $z = \phi(x)$ zou dan vastgelegd zijn door het voorschrift:
$\int_a^x f(\underline{x})d\underline{x} = \int_{\phi(a)}^{\phi(x)} \frac{2\alpha}{\beta\sqrt{2\pi}} \underline{z}^{\alpha-1} e^{-\underline{z}^{2\alpha}/2\beta^2} d\underline{z}$, voor iedere $a \leqslant x \leqslant b$.

(Voortzetting volgende blz.)

De vraag of $f(\underline{x})$ zulk een vorm hebben kan, dat normaliseren überhaupt mogelijk is behoeft niet gesteld te worden, aangezien iedere kansverdeling $f(x)dx$ genormaliseerd kan worden, wat reeds onder 1.1. bleek en in 2 duidelijker worden zal. Wel heeft het zin te vragen hoe $f(\underline{x})$ eruit moet zien opdat de normalisatie-transformatie $z=\psi(x)$ van eenvoudige gedaante (dit nader te definiëren) zal zijn. De beantwoording eist de ontwikkeling van een kriterium om uit te maken of $f(x)$ zodanig kan zijn, dat van een gegeven $z=\psi(x)$ sprake kan zijn. Zulk een kriterium zal gebaseerd moeten zijn op bepaalde functies der N metingen en relaties daartussen.

2. Eerste geval; $f(\underline{x})\,d\underline{x}$ is exact bekend.

$f(\underline{x})d\underline{x}$ is gegeven over $a \leqslant x \leqslant b$, zodat $\int_a^b f(\underline{x})d\underline{x}=1$.
Bij ieder paar waarden μ_z, σ_z (met $-\infty < \mu_z < +\infty$; $0 < \sigma < \infty$) is de normalisatie-transformatie $z=\psi(x)$ vastgelegd (en hiermee impliciet bekend) door het voorschrift:

1. $$\int_a^x f(\underline{x})d\underline{x} = \int_{\psi(a)}^{\psi(x)} \frac{1}{\sigma_z\sqrt{2\pi}} \exp.\left[-\frac{(z-\mu_z)^2}{2\sigma_z^2}\right] dz$$

Gezien het feit, dat deze betrekking moet gelden voor iedere $a \leqslant x \leqslant b$, moet $\psi(a)=-\infty$ en $\psi(b)=+\infty$ zijn, waardoor reeds twee voorwaarden aan $z=\psi(x)$ opgelegd zijn. Tegelijk zijn μ_z en σ_z het gemiddelde en de standaarddeviatie van de normale $\underline{z}$-verdeling. Door ieder ander stel σ, μ [1]) is een andere $z=\psi(x)$ vastgelegd, tussen welke transformaties een eenvoudige relatie bestaat, waarover straks meer.

---

zodat ook gelden moet $\phi(a)=0$ en $\phi(b)=\infty$. Nu is $\phi(x)$ bekend, zodra α en β gegeven zijn. Ieder ander stel α en β bepaalt een andere $\phi(x)$ en het is de vraag of ook nu tussen al deze $\phi(x)$'en een eenvoudige relatie bestaat, zo simpel als bij het normaliseren.

1) Misschien lijkt het logisch te eisen, dat de gemiddelde waarden van x en z (μ_x en μ_z) gelijk zullen zijn, evenals hun varianties (σ_x^2 en σ_z^2). Men kan immers μ_z en σ_z kiezen zoals men wil; d.w.z. $\mu_z=\mu\equiv\int_a^b \underline{x} f(\underline{x})d\underline{x}$ en $\sigma_z^2 = \sigma_x^2 \equiv \int_a^b (\underline{x}-\mu)^2 f(\underline{x})d\underline{x}$. Dit is een kwestie van smaak. Men kan ook prefereren $\mu_z=0$ en $\sigma_x=1$ of μ_z =modus of mediaan in $f(\underline{x})d\underline{x}$; enz. enz.

Met dit te zeggen is natuurlijk $z=\psi(x)$ als functie van x <u>expliciet</u> nog niet bekend. Het kan zijn, dat wij juist deze functie willen kennen.

2. Hoe vinden wij z als functie van x, exact of in voldoende goede benadering?

Men bedenke:

3. $$\int_a^x f(\underline{x})d\underline{x}=\int_{-\infty}^{\psi(x)} \frac{1}{\sigma\sqrt{2\pi}} e^{-(\underline{z}-\mu)^2/2\sigma^2} d\underline{z}=\int_{-\infty}^{\tau(x)} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\underline{\tau}^2} d\underline{\tau},$$ waarin

$$\phi(\underline{\tau})d\underline{\tau}=\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\underline{\tau}^2} d\underline{\tau}=\mathcal{N}(0;1)=$$ de gestandaardiseerde normale verdeling op de variabele $\underline{\tau}$ (gemiddelde = 0; standaarddeviatie: 1) voorstelt. Hierbij is $\underline{\tau}=(\underline{z}-\mu):\sigma$ (de indices z aan μ_z en σ_z laten we maar weg.)

Men berekent voor elk van een groot aantal (=N) waarden van x: $a\leqslant x_1 < x_2 \ldots < x_n \leqslant b$, de onderschrijdingskans $F(\underline{x}\leqslant x_i)=\int_{\ldots}^{x_i} f(\underline{x})d\underline{x}$ en zet deze op lineair-gauszisch papier uit. Dit papier heeft twee assen: een lineaire x-as en een niet lineaire kansen-as G. Langs de laatste kan men zich ook een lineaire τ-as voorstellen, met $\tau=0$ in het punt G = 50%, zodat bijv. $\tau = \pm 1$ overeenkomst met de kansen 0.841 resp. 0.159 en $\tau = \pm 2$ met de kansen 0.977 resp. 0,023 etc. (zie de tabel van de gestandaarddiseerde normale verdeling, volgens welke $\int_{-\infty}^{1}\phi(\underline{\tau})d\underline{\tau}=0.841$ en $\int_{-\infty}^{-1}\phi(\underline{\tau})d\underline{\tau}=1-0.841=0.159$ enz.

Vindt men bijv. bij $x=x_3$ een $F_3=0.0107$, dan komt daarmee overeen het punt $G_3=0.0107$; $x=x_3$, doch ook het punt $\tau_3=-2.3$; $x=x_3$, aangezien $\int_{-\infty}^{-2.3}\phi(\underline{\tau})d\underline{\tau}=0.0107$. Geeft een andere waarde $x=x_8$ een $F_8=0.9332$, dan wordt hiermede het punt $G_8=0.9332$; $x=x_8$ vastgelegd, d.i. tevens $\tau_8=1.5$; $x=x_8$, omdat $\int_{-\infty}^{1.5}\phi(\underline{\tau})d\underline{\tau}=0.9332$. Aldus ontstaan er N punten $\tau_i;x_i$. Het verband tussen $\underline{\tau}$ en x zal kunnen worden beschreven door $\tau=\varphi(x)$, zo nauwkeurig als men wil, aangezien men N onbeperkt kan laten toenemen. (het "hoe" bespreken wij niet; dit is een afzonderlijk probleem). Men kan natuurlijk met zekere graad van <u>benadering</u> tevreden zijn teneinde $\varphi(x)$ niet te ingewikkeld te maken.

Voor zekere $\tilde{x}$ (de z.g. mediaan) zal $F(x\leqslant\tilde{x})=\frac{1}{2}$ zijn. Gesteld dat men deze $\tilde{x}$ voldoende nauwkeurig berekenen kan (de berekening is soms zeer moeilijk), dan moet in elk geval $\varphi(x)$ een dusdanige vorm hebben, dat $\varphi(\tilde{x})=0$, of ook, de normalisatie-transformatie $z=z(x)$ geeft voor $x=\tilde{x}$ het gemiddelde μ in de normale $\underline{z}$-verdeling, d.w.z. $\mu=z(\tilde{x})$. D.w.z. de kromme $\tau=\varphi(x)$ in het dubbellineaire τ-x-vlak moet door het punt $\tau=0$, $x=\tilde{x}$ gaan.

Men moet bedacht zijn, in het algemeen, op de ongelijkheid der schaaleenheden langs de τ- en x-as op het gauszisch papier. (e_τ resp. e_x). Op het gebruikelijke waarschijnlijkheidspaier is de τ-eenheid ca. 31.5 mm papier, terwijl men voor de x-eenheid nemen kan, naar believen, 42 ,21, $10\frac{1}{2}$, 2,1 enz. mm. Als men gelijke eenheden wenst, moet men schrijven $\tau^* = q \cdot \varphi(x)$ als $q = e_\tau : e_x$. Deze $\varphi(x)$ bepaalt de gezochte normalisatie-transformatie, echter niet éénduidig. Dit blijkt a.v.: zodra men $\varphi(x)$ schrijven kan in de vorm $\varphi(x) = \frac{\psi(x) - \mu}{\sigma}$ is deze $z = \psi(x)$ de gezochte normalisatie-transformatie, waarbij de normaal verdeelde $\underline{z}$ het gemiddelde μ en de standaarddeviatie σ heeft. Maar dan is er ook een andere $z^* = \psi_1(x)$, die normaal verdeeld is rondom een ander gemiddelde μ_1 met een andere σ_1, die zich a.v. laten bepalen. Het is altijd mogelijk te schrijven $\mu = A\,\mu_1 + B$ en $\sigma = A\,\sigma_1$.

Nodig is $F(x) \int_a^x f(\underline{x})d\underline{x} = \int_{-\infty}^{z(x)} (\sigma\sqrt{2\pi})^{-1} \exp.\left\{-(\underline{z}-\mu)^2/2\sigma^2\right\} d\underline{z} = \int_{-\infty}^{z^*(x)} (\sigma_1\sqrt{2\pi})^{-1} \exp.\left\{-(\underline{z}-\mu_1)^2/2\sigma_1^2\right\} d\underline{z}$. Maar

$$\int_{-\infty}^{z(x)} (\sigma\sqrt{2\pi})^{-1} \exp.\left\{-(\underline{z}-\mu)^2/2\sigma^2\right\} d\underline{z} = \int_{-\infty}^{z(x)} (A\sigma_1\sqrt{2\pi})^{-1} \exp.\left[-\left\{\underline{z}-(A\mu_1+B)\right\}^2 / 2A^2\sigma_1\right] d\underline{z} =$$

$$\int_{-\infty}^{?} (\sigma_1\sqrt{2\pi})^{-1} \exp.\left\{-(\underline{t}-\mu_1)^2/2\sigma_1^2\right\} d\underline{t}\ ,\ \text{als}\ \underline{t} = \frac{\underline{z}-B}{A}.$$

De nieuwe bovengrens ? wordt dan $\left\{z(x)-B\right\}$: A. Opdat dit waar zal zijn voor elke x uit $a \leqslant x \leqslant b$ moet gelden

4. $$\boxed{z^*(x) \equiv \left\{z(x)-B\right\}/A = \frac{\sigma_1}{\sigma}\cdot z(x) + \mu_1 - \frac{\sigma_1}{\sigma}\mu.}$$

Zodra dus één normaliserende transformatie $z = \varphi(x)$ bekend is, bij een bepaald paar μ en σ, is een ∞^2 verzameling van andere normalisatie transformaties bekend, want geeft men μ_1 en σ_1, dan is z^* volgens (4) in z uit te drukken.

Getallenvoorbeelden

a) $$f(\underline{x})d\underline{x} = \frac{1}{(\underline{x}+25)\sqrt{2\pi}} \exp.\left[-\left\{\lg(\underline{x}+25)\right\}/2\right] d\underline{x},\ \text{voor}\ -25 \leqslant x \leqslant \infty.$$

Nu is zeker $z = \lg(x+25)$ een normalisatie-transformatie, waarbij z normaal verdeeld is met gemiddelde 0 en standaarddeviatie 1. ($-\infty \leqslant z \leqslant \infty$, als $-25 \leqslant x < \infty$). Wil men een normale verdeling met gemiddeld 4 en standaarddeviatie 5 dan is (volgens (4)) de nieuwe norm. transformatie $z^*(x) = 5\,z(x)+4$, d.i. $z^*(x) = 5\,\lg(x+25)+4$ ($-\infty \leqslant z^* \leqslant \infty$ als $-25 \leqslant x \leqslant \infty$) ; kennelijk is de mediaan $\tilde{x} = -24$, waarvoor $z = \lg(x+25) = 0$.

b) $f(\underline{x})d\underline{x} = \frac{e^{\underline{x}} + \frac{1}{x^2}}{\sqrt{2\pi}}\, e^{-(e^{\underline{x}} - \frac{1}{\underline{x}} - 4)^2/2}\, d\underline{x}$, voor $0 \leqslant \underline{x} \leqslant \infty$.

Nu is zeker $z = e^{x} - \frac{1}{x}$ één der ∞^2 normalisatie-transformaties, waarvoor $-\infty \leqslant z \leqslant \infty$, als $0 \leqslant x \leqslant \infty$; de normale z-verdeling heeft een gemiddelde 4 en een standaarddeviatie 1. Wil men liever een gemiddelde 0 en standaarddeviatie 10, dan is volgens (4) de nieuwe norm. transformatie $z^{*}(x) = 10(e^{x} - \frac{1}{x}) - 40$. De mediaan $\tilde{x} \cong 1.54$, want hiervoor is $e^{x} - \frac{1}{x} = 4$

Deze twee voorbeelden lijken sterk van academische aard, omdat zij voldoende doorzichtig zijn. Het kan echter voorkomen, dat $f(\underline{x})$ zelf zo ingewikkeld is, dat men de behoefte gevoelt een betrekkelijk eenvoudige normalisatie-transformatie $z = \psi(x)$ te vinden, zodanig, dat $\underline{z}$ voldoende goed normaal verdeeld is, terwijl $\underline{x}$ zelf het beslist niet is. Is eenmaal $z = \psi(x)$ gevonden, dan kan het zijn, dat het minder rekenwerk vergt om voortaan voor iedere andere x - waarde van de gebruikte N stuks niet direct $\int_a^x f(\underline{x})d\underline{x}$ te berekenen, maar d.m.v. de z, d.w.z. als $\int_{-\infty}^{z(x)} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{(\underline{z} - \mu)^2 : 2\sigma^2\right\} d\underline{z} = \int_{-\infty}^{\tau(x)} \phi(\underline{\tau})d\underline{\tau}$ met $z(x) = \psi(x)$, en $\tau(x) = \{z(x) - \mu\} : \sigma$, waarbij $z = \psi(x)$ en daarmede $\tau(x)$ berekend moet worden en verder de tabel van de gestandaardiseerde normale verdeling geraadpleegd moet worden.

Bovendien: zoals reeds gezegd werd, is het tweede aspect der normalisatie veel belangrijker dan het eerste, aangezien wij nooit met universa, doch met steekproeven te maken hebben en vrijwel nooit de $f(x)dx$ exact bekend is. Het kwam ons systematischer voor toch met de behandeling van dit eerste aspect te beginnen.

3. <u>Tweede geval; $f(\underline{x})d\underline{x}$ moet worden geschat uit de steekproef.</u>

Met dit geval hebben we gewoonlijk te maken. Immers van $f(\underline{x})d\underline{x}$ kennen wij veelal niet eens het type, en zo ja, dan toch niet de numerieke waarden der bepalende parameters. Er zijn N (voorlopig verschillend gedachte) metingen $x_1, x_2 \ldots x_n$ verricht, welke waarden tot een populatie met een <u>onbekende</u> waarschijnlijkheidsverdeling $f(\underline{x})d\underline{x}$ behoren. Gewoonlijk kent men wel a priori de kleinst en de grootst mogelijke x - waarde (onder geval 1 resp. a en b genoemd) Dagsommen neerslag : a = o, b = ∞; zonneschijnpercentages : a = o, b = 100; maximum-temperatuur : a = $-273^{\circ}C$ 3 b= ∞ $^{\circ}C$; aantal ijsdagen in december: a = 0, b = 31, enz. Wij kennen derhalve ook <u>niet</u> de <u>ware</u> waarde van $F = \int_a^{x_i} f(\underline{x})d\underline{x}$, voor $x_i = x_2 \ldots x_n$. Wij kennen <u>slechts schattingen</u> $\hat{F}$. Over de <u>beste</u> schattingen verschillen de meningen.

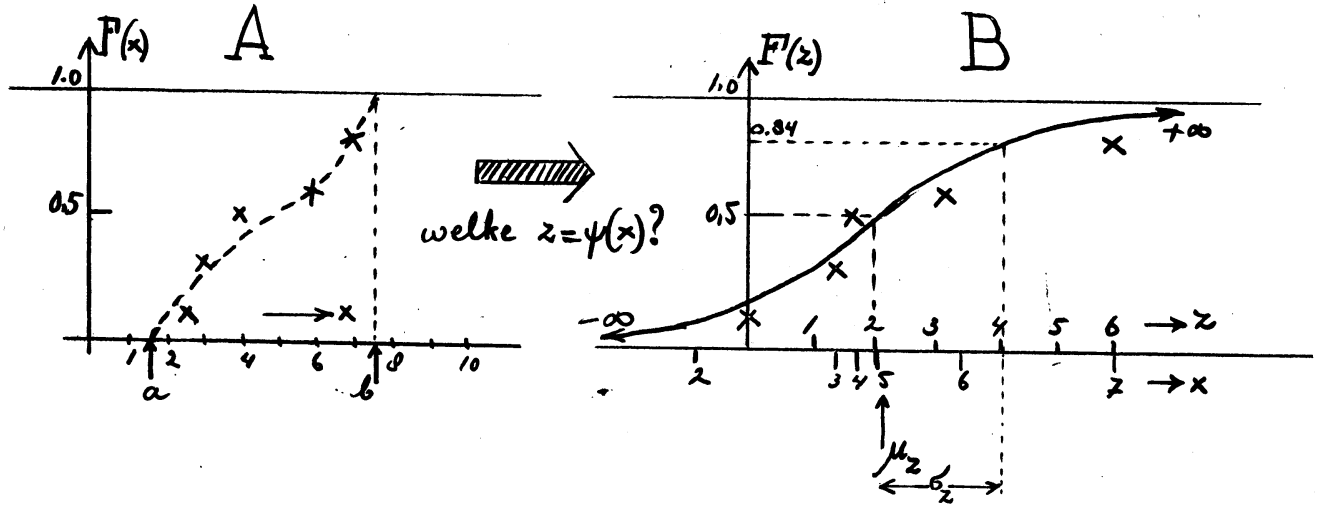
Er is veel over geschreven; dikwijls ligt het knellende punt in de definitie van "beste". Laat de x-waarden reeds gerangschikt zijn: $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$. (waaronder eventueel gelijke waarden). Men neemt wel: $\hat{F}(x_i) = \frac{i}{N}$, waarbij met het eventuele meervoudig optreden van zekere x-waarden rekening gehouden is. Komt bijv. x_1 n_1 keren voor; x_2 n_2 keren;.... x_n n_n keren, zodat $\sum_1^n n_j = N$, dan is $\hat{F}(x_i) = \frac{\sum_1^i n_j}{N}$. Anderen prefereren $\frac{i - \frac{1}{2}}{N}$ (d.i. $\frac{2i - 1}{2N}$); $\frac{i - 1}{N}$; $\frac{i + 1}{N}$ enz.

In [2] wordt uiteengezet, dat, als men te voren weet, dat $f(x)dx$ van het type der normale verdeling is, het best $\hat{F}(x_i) = \frac{i - 0.3}{N + 0.4}$ genomen kan worden. Verondersteld, men heeft op zekere gronden redenen om de normaliteit van de x - verdeling te poneren. Nadat dan de n punten in het lineair-gauszisch net aangebracht zijn, kan men, bijv. op het oog, de beste rechte tussen deze punten door trekken, waardoor weer de beste μ en σ vastgelegd worden. Doch dit geval interesseert ons hier niet.

Wij onderstellen nl., dat we niets over de gedaante van $f(x)$ weten (althans niet meer dan wat de steekproef daarover suggereert) en dat wij willen onderzoeken of $f(x)$ kan behoren tot die verzameling van functies, die zich met simpele vormen van transformaties $z = \psi(x)$ laten normaliseren.

Ten behoeve van dit onderzoek kan het nuttig zijn kriteria (toetsen) te ontwikkelen. Zulk een kriterium en wel t.a.v. $z = \psi(x) = x^a$ werd door Neumann en Kotz in [6] ontwikkeld.

De volgende schetsjes illustreren een en ander. Thans is in het x-F-vlak niet een ononderbroken kromme gegeven, maar een (klein) aantal punten (bij N metingen: N of N-1 stuks, al naar de opvatting omtrent de beste $\hat{F}(x)$). Tussen die punten door loopt de bij het universum behorende, onbekende, onderschrijdingscurve, tussen de eindpunten a en b van het variatiegebied $a \leqslant x \leqslant b$. Kan deze curve zulk een gedaante hebben, dat er een simpele transformatie $z = \psi(x)$ bestaat, zodanig, dat z wèl normaal verdeeld is, rondom μ en met standaarddeviatie σ ?



De verdere procedure is als onder geval 1, doch nu liggen de punten τ, x in het τ-x- veld "meer verspreid", d.w.z. ze liggen niet precies op, doch door toevalseffecten (steekproefeffecten) rondom de onbekende kromme $\tau = \varphi(x)$, hetgeen het vinden van deze $\varphi(x)$ bemoeilijkt. Nadat dan een bevredigende $\varphi(x)$ gevonden is, hebben wij tegelijk de beste schatting van $f(\underline{x})d\underline{x}$ gevonden, nl. $\hat{f}(\underline{x})d\underline{x} = \frac{dz/dx}{\sigma\sqrt{2\pi}} \exp.\left\{-(\underline{z}-\mu)^2/2\sigma^2\right\} d\underline{x}$,
waarin $\underline{z} = \varphi(x)$ normaal verdeeld is, rondom $\mathcal{E}\underline{z}=\mu$ en met variantie σ^2, waarna met de χ^2-toets kan worden onderzocht hoe goed empirische en theoretische distributie frequentieverdeling "overeenkomen". Hierbij is $\varphi(x)$ geschreven gedacht als $(\varphi(x)-\mu):\sigma$

Verschillende auteurs gaan van de andere kant uit, d.w.z. zij denken $z = \varphi(x)$ gegeven en vragen zich dan af welke vorm $f(\underline{x})d\underline{x}$ bezit en welke klimatologische grootheden een distributie hebben die goed met deze typen $f(\underline{x})d\underline{x}$ te benaderen zijn. Men neemt allicht eenvoudige functies $\varphi(x)$, zoals $a+bx$; $(a+bx)^c$; $a_0+a_1x+a_2x^2+$ ... ; $d+\lg(x+c)$ enz.

Helaas ontwikkelen ze geen kriterium, door middel waarvan men met behulp der N metingen kan uitmaken (met voorgeschreven betrouwbaarheid) of de ene normalisatie-transformatie wel , de andere niet in aanmerking komt.

4 Numerieke voorbeelden.

4.1 Uit de praktijk.

Wij citeren een voorbeeld uit het rapport van Rijkoort [7]. Hij beschouwde in elk van 35 winters de grootste diepte x (beneden aardniveau), waartoe de vorst doordrong. Er waren winters, waarin de vorst niet in de grond drong. Voor deze werd x nul genoemd. De 35 x-waarden waren 8 keer 0 ($x_1 = \ldots x_8 = 0$); $x_9=3$; $x_{10}=x_{11}=4$; $x_{12}=5$; $x_{13}=x_{14}=6$; $x_{15}=7$; $x_{16}=10$; $x_{17}=1$; $x_{18}=12$; $x_{19}=x_{20}=13$; $x_{21}=15$; $x_{22}=16$; $x_{23}=17$ $x_{24}=19$ $x_{25}=21$ $x_{26}=22$ $x_{27}=x_{28}=31$ $x_{29}=34$ $x_{30}=35$ $x_{31}=36$ $x_{32}=38$ $x_{33}=53$ $x_{34}=54$ $x_{35}=60$ (24 ongelijke x - waarden)

Als men log-gausz-papier gebruikt en met de x' en zelf werkt, komen de punten allesbehalve lineair te liggen, zie in [7]. Het ligt dus voor de hand het met x + b (b te zoeken) te proberen, wat impliceert, dat men wil onderstellen, dat x + b normaal verdeeld is (van welke normale verdeling men verder de μ en σ weten wil). Aldus ging Rijkoort te werk. Naar wat waarden geprobeerd te hebben kwam hij tot $20 < b < 30$, terwijl μ en σ van het papier afgelezen kunnen worden. Deze uitkomst houdt in, dat men voor de onbekende kansverdeling van x meent gevonden te hebben in voldoend goede benadering?

$$f(\underline{x})d\underline{x} = \frac{1}{\sigma(\underline{x}+b)\sqrt{2\pi}} \exp.\left[-\lg(\underline{x}+b) - \mu^2/2\sigma^2\right] d\underline{x}$$

Toch kan dit niet exact juist zijn, want alleen over het traject $-b \leqslant x < \infty$ is $\int_{-b}^{\infty} f(\underline{x})d\underline{x} = 1$, terwijl $\underline{x}$ alleen waarden $o \leqslant x < oo$ aannemen kan. Men kan aan de moeilijkheid ontkomen door uit te gaan van $f^{*}(\underline{x})dx = \{f(\underline{x})d\underline{x}\}/K$, als $K = \int_{o}^{\infty} f(\underline{x})dx < 1$, want nu is $\int_{o}^{\infty} f^{*}(\underline{x})d\underline{x} = 1$. Vermoedelijk ligt K zeer dicht bij één. Doet men dit, dan zal de aanpassing allicht iets minder goed zijn en zullen wij op iets andere waarden van b, μ en σ moeten overgaan om haar te herstellen [1].

Deze complicaties kunnen ook vermeden worden door allereerst alleen die winters (derhalve minder dan 35 stuks) te beschouwen, waarin x òf precies nul òf positief is en de andere winters, waarin zelfs niet het aardniveau op 0°C-temperatuur kwam, uit te zonderen. Alleen voor de dan overblijvende x-waarden (een aantal $\leqslant 35$) zoeken we de kansverdeling in het universum. Bijgevolg moeten wij dan eigenlijk de logarithmisch-normale verdeling met $b > o$ uitsluiten en direct naar andere uitzien [2]. Het is niet de bedoeling hier verder op door te gaan; wel willen wij, louter bij wijze van numeriek voorbeeld, het probleempje van Rijkoort hier uitwerken:
Er geldt $\int_{-b}^{\infty} f(\underline{x})d\underline{x} = \int_{-\infty}^{\infty} \psi(\underline{z})dz = 1$; zet $t = (z-\mu)/\sigma$ en $\phi(\underline{t}) = \frac{e^{-\frac{1}{2}\underline{t}^2}}{\sqrt{2\pi}} d\underline{t}$
Er zijn 24 verschillende x-waarden x_i. Voor elk geldt:

$$\int_{-b}^{x_i} f(\underline{x})d\underline{x} = \int_{-\infty}^{z_i} \psi(\underline{z})d\underline{z} = \int_{-\infty}^{t_i} \phi(\underline{t})d\underline{t} \quad \text{met } z_i = \lg(x_i+a) \text{ en } t_i = (z_i-\mu)/\sigma$$

Door dit voorschrift is aan elke x_i één t_i toegevoegd. We moeten de beste â, $\hat{\mu}$, $\hat{\sigma}$ berekenen; de beste schatting van $\int_{-b}^{x_i} f(\underline{x})d\underline{x}$ (onbekend, aangezien $f(\underline{x})d\underline{x}$ onbekend is) is

$\hat{F}(\underline{x} \leqslant o) = \frac{8}{35} = 0.228$, zodat $t = -0.75$;
die van $\int_{-b}^{3} f(\underline{x})d\underline{x}$ is $\hat{F}(\underline{x} \leqslant 3) = \frac{8+1}{35} = 0.257$, zodat $t = -0.65$;
die van $\int_{-b}^{4} f(\underline{x})d\underline{x}$ is $\hat{F}(\underline{x} \leqslant 4) = \frac{8+1+2}{35} = 0.314$, zodat $t = -0.485$ enz.

Aldus ontstaan er 23 vergelijkingen van de vorm $t = \frac{\lg(x+b)-\mu}{\sigma}$ of $\lg(x+b) = t\sigma + \mu$, met onbekenden b, σ en u. Hoe deze oplossen? Neem enkel-log-papier. Kies een b, bijv. 10 en zet de punten bij $\lg(10+x)$ tegen x uit (fig. 1). Zie of ze lineair liggen.

---

1) Maar deze finesse is van geen betekenis als het ons alleen gaat om een schatting van de overschrijdingskans $\int_{m}^{\infty} f(\underline{x})d\underline{x}$ voor grote m en hier ging het Rijkoort inderdaad in hoofdzaak om.

2) Zie mijn ander rapport [5] , waarin ik normaliseer d.m.v. $z = x^a$.

Zo niet, neem een wat groter b , bijv. 20. Misschien liggen de punten nu op een wat minder sterk gebogen curve. Vergroot b weer wat en ga zo voort tot de punten zo goed mogelijk lineair liggen. Dit op het oog te beoordelen (want hoe objectief?) Wij vonden $\hat{b}=25$; $\hat{\mu}=3.6$ en $\hat{\sigma}=0.5$ cm. De normalisatie transformatie is dus $z=\lg(x+25)$ en de kansverdeling

$$f(\underline{x})d\underline{x} = \frac{e^{-[\lg(\underline{x}+25)-3.6]^2/2(\frac{1}{2})^2}}{\frac{1}{2}(\underline{x}+25)\sqrt{2\pi}}\,d\underline{x} \quad \text{, waarvoor} \int_{-25}^{\infty} f(\underline{x})d\underline{x}=1$$

Hoe goed is de benadering?
Zie het volgende tabelletje.

interval voor x	aantallen	
	geteld	met formule
$\leqslant$ 10 cm	16	16.2
11–20	8	7.2
21–30	2	4.6
31–40	6	2.9
41–50	0	1.6
51–60	3	1.4
$\geqslant$ 61	0	1.1
Som	35	35.0

De χ^2-toets is helaas niet toe te passen omdat (afgezien van de reden, dat hier de beste schattingen grafisch gebeurden en niet volgens de Methode der Maximum Likelihood) het aantal graden van vrijheid tenminste 1 moet zijn; d.w.z. het aantal intervallen moet tenminste 1+4=5 zijn (3 parameters en sommen gelijk), terwijl in elk interval het gemiddeld te verwachten aantal liefst tenminste 5 moet zijn. Voor dit alles is het totale aantal 35 te klein. Maar wij zien ook wel op het oog hoe(goed) de aanpassing is.

N.B. In wezen is de hier behandelde "methode" gelijk aan die van Rijkoort. In beide gevallen ontbreekt het "lineariteitskriterium"

4.2 Gefantaseerd voorbeeld.

Er werden 10 metingen verricht; met behulp van $\hat{P}=i/10$ (i=1.2 ... 9) werden op lineair-gauszisch papier de 9 kruisjes verkregen. (fig. 2) Na wat proberen lijkt de kromme $\tau = e^{x} - \frac{1}{x} - 4$ (tussen -oo en + oo) een goede benadering van de ware cumulatieve kanscurve.

Is men met een aanpassing op het oog niet tevreden en wenst men het begrip "voldoende goede benadering" objectief mathematisch te definiëren, dan is dit een allesbehalve eenvoudig vraagstuk, ondermeer vanwege het feit, dat de punten min of meer afhankelijk zijn, d.w.z. met toenemende x een groter gewicht hebben. Uit genoemde $\tau(x)$ volgt weer, dat de beste schatting van de onbekende kansverdeling in het x-universum is

$$\hat{f}(\underline{x})d\underline{x} = \frac{e^{\underline{x}} + \frac{1}{\underline{x}^2}}{\sqrt{2\pi}} \text{ exp.}\left[-\frac{(e^{\underline{x}} - \frac{1}{\underline{x}} - 4)^2}{2}\right]dx \text{ , voor } o \leqslant \underline{x} \leqslant \infty.$$

De normalisatie-transformatie is dus $z = \psi(x) = e^{\underline{x}} - \frac{1}{\underline{x}}$.

De normale $\underline{z}$- verdeling is $\frac{1}{\sqrt{2\pi}}$ exp. $\{-(z-4)^2/2\}dz$, met $\mu_z = 4$ en $\sigma_z = 1$.

5. Literatuur.

[1] Levert, C. Beschouwing over betrouwbaarheidsintervallen en een uitschieterscriterium K.N.M.I. W.R. 56-005, 1956.

[2] Benard, A. en Bos-Levenbach, E.C. Het uitzetten van waarnemingen op waarschijnlijkheidspapier. Statistica 7 163 1953.

[3] Iterson, G. van Het nut van waarschijnlijkheidspapier voor de variantiestatistiek. Statistica 4 129 1950.

[4] Exalto, L.J.H. Gebruik en toepassing van waarschijnlijkheidspapier. Statistica 4 121 1950.

[5] Levert, C. Normaliseren door middel van de transformatie $z=x^a$. K.N.M.I. W.R. 57-002, 1957.

[6] Neumann, J. and Kotz, S. Some Pearsonion-like frequency functions capable of a modified normalization. In "The scientific basis of weather modification studies", Arizona 1956.

[7] Bijkoort, P. Bijdrage tot het bepalen van de meest gunstige diepte voor het leggen van waterleidingbuizen in verband met het bevriezingsrisico. K.N.M.I. W.R. 56-003, 1957.

6. Summary

0. In this chapter 8 examples for the necessity of normalization are mentioned:

0.1 Both $\underline{x}$ and $\underline{y}$ must be distributed normally if one wants to verify the statistical reality of the common correlation coefficient by means of the common correlation table.

0.2 The members of a time series $x_1, x_2, x_3 \ldots$ must all be distributed normally if one wants to investigate the significance of the serial correlation coefficients (autocorrelation) of lag 1, 2 and so on, as in 0.1.

0.3 Having at our disposal only N measurements $x_1 \leqslant x_2, \ldots \leqslant x_n$ it is possible to compute the confidence limits for the exceedance probability of a given value x_g (even if $x_g < x_1$ or $> x_n$) almost only if the $\underline{x}$ - distribution is normal.

0.4 Application of the t-test needs a normal distribution in both populations.

0.5 Application of the Variance Analysis needs a normal distribution.

0.6 Application of the Bartlett-test (is $\sigma_1 = \sigma_2 = \ldots \sigma_n$?) supposes that the n universa all have normal distributions.

0.7 Applications of the z- or F- test needs the same fact.

0.8 Several criteria for rejection of "outlyers" require that the distribution in the population is normal [1] .

0.9 In special theories the form of the transformation $z = \varphi(x)$ is seen in relation to special theoretical considerations.

1.1. $f(\underline{x})d\underline{x}$ is known exactly.

Either $f(\underline{x})d\underline{x}$ or $F(\underline{x} \leqslant x) = \int_a^x f(\underline{x})d\underline{x}$ is given analytically, or $F(\underline{x} \leqslant x)$ is given as a curve on double linear paper ($\underline{x}$ and F in linear scales) or e.g. on linear-probability paper ($\underline{x}$ in linear scale and F in gausz-scale). The variation range of x should be $a \leqslant \underline{x} \leqslant b$.

Definition: The variable $\underline{y}$ is distributed normally if at least one couple of values η and σ exists so that for each value y_i out of the range $-\infty < y < \infty$ the probability of a value situated between y_i and $y_i + dy_i$ can be written as

$$(1) \qquad g(y_i)dy_i = (\sigma\sqrt{2\pi})^{-1} \exp\left[-(y_i - \eta)^2/2\sigma^2\right] dy_i$$

Definition of normalizing: suppose $\underline{x}$ itself is not distributed normally. We ask: Is there at least one normalizing transformation $z = \varphi(x)$ so that $\underline{z}$ is distributed normally? This involves:

$$(2) \qquad \int_a^{x_i} f(\underline{x})d\underline{x} = \int_{-\infty}^{z_i} g(\underline{z})d\underline{z} \quad \text{with } z_i = \varphi(x_i)$$

As soon as the mean and standard deviation μ_z and σ_z are chosen the function $z = \varphi(x)$ is fixed uniquely by the relation (2) implicitely.

1.2 $f(\underline{x})d\underline{x}$ is not known, but N values $x_1, x_2 \dots x_n$ have been "measured". Suppose the $\underline{x}$ population is not distributed normally. Stress is laid on the following question: is it possible that the unknown $f(\underline{x})$ has such a (simple) form that it can be normalized by a simple normalization transformation $z=\psi(x)$ (from a given class)? Of course it is desirable and useful to derive a criterion with the aid of which it is possible to investigate wholly statistically whether the first question can be answered in the affirmative. Such a criterion will be based on special functions of the N measurements and relations between these functions.

2. Suppose $f(\underline{x})d\underline{x}$ is known, defined for $a \leqslant x \leqslant b$, with $\int_a^b f(\underline{x})d\underline{x} = 1$. Then $z=\psi(x)$ must be found (and is defined) so that

(3) $$\int_a^{x_i} f(\underline{x})d\underline{x} = \int_{\psi(a)}^{\psi(x_i)} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-(z-\mu)^2/2\sigma^2\right\} dz$$ for each $a \leqslant x \leqslant b$ and given values of μ and σ.

(4) Hence $\psi(a)=-\infty$; $\psi(b)=\infty$ and $f(x) = \frac{dz/dx}{\sigma\sqrt{2\pi}}\, e^{-(z-\mu)^2/2\sigma^2}$ Further:

(5) $$\int_a^{x_i} f(\underline{x})d\underline{x} = \int_{-\infty}^{\psi(x_i)} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(z-\mu)^2/2\sigma^2} dz = \int_{-\infty}^{\tau(x_i)} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\tau^2} \cdot d\tau$$

(6) where $$\phi(\tau)d\tau = \frac{e^{-\frac{1}{2}\tau^2}}{\sqrt{2\pi}}\, d\tau$$

is the normal probability distribution in standard form ($\mu=0$; $\sigma=1$). The solution is as follows: Choose a large number N of values $x_1\ x_2 \dots x_N$, compute the values $F(\underline{x} \leqslant x_i)$ and plot these pairs x,F on linear-probability paper. The non linear F-scale of this paper is at the same time the linear τ scale defined by the probability integral $F = \int_{-\infty}^{\tau} \phi(\tau)d\tau$. Suppose the relation between x and τ can be described sufficiently well by a certain function $\tau = \varphi(x)$, with $\varphi(\tilde{x})=0$ if $\tilde{x}$=median, defined by $F(x \leqslant \tilde{x}) = \frac{1}{2}$. Now write

(7) $\varphi(x) = \{\psi(x)-\mu\} : \sigma$, in which expression μ and σ are not functions of x. Then $z=\psi(x)$ is distributed normally in good approximation with mean μ and standarddeviation σ. Each other function

(8) $$z^{*} = \frac{\sigma^{*}}{\sigma} z + \mu^{*} - \frac{\sigma^{*}}{\sigma}\mu$$ is also distributed normally with mean μ^{*} and standarddeviation σ^{*}. Hence the functions $z^{*}(x)$ represent all normalising transformations functions, approximative solutions of (4).

(9) Then $$f(\underline{x})d\underline{x} = (\sigma\sqrt{2\pi})^{-1} \exp\left\{-(\underline{z}-\mu)^2/2\sigma^2\right\} d\underline{z}$$

3. Suppose $f(\underline{x})d\underline{x}$ is not known, but we only have the disposal of N "measurements" $x_1 \leqslant x_2 \leqslant x_3 \dots \leqslant x_n$. Mostly we do know lower and upper limits of the variable $\underline{x}$, e.g. daily rainfall amounts $o \leqslant x < \infty$; daily sunshine percentages $o \leqslant x \leqslant 100$; number of freezing days per december $o \leqslant x \leqslant 31$.

Now we must make estimations $\hat{F}$ for the n unknown values of $F = \int_a^{x_i} f(\underline{x})d\underline{x}$. Let the x values be ordered as follows $x_1 < x_2 \ldots < x_n$; n_1 times x_1; n_2 times x_2 ;.. n_n times x_n, with $N = \sum_1^n n_j$. Some authors then take $\hat{F}(\underline{x} < x_i) = \sum_1^i n_j/N$. Other authers take $(\sum_1^i n_j - \frac{1}{2}):N$ or $(\sum_1^i n_j - 1):N$ or $(\sum_1^i n_j + 1):N$ or especially $(\sum_1^i n_j - 0.3):(N+0.4)$ if the distribution may be supposed to be normal [2]. Plot then n points x_i, $\hat{F}$ on linear-probability paper (the non linear probability axis contains the linear τ scale) Try to find the best fitting function $\tau = \varphi(x)$. With this function corresponds a curve. The curve should be situated in such a way that the points scatter around this curve as narrowly as possible. Write $\varphi(x)$ as $\{\varphi(x) - \mu\} : \sigma$. Hence the best estimation $\hat{f}(x)dx$ of $f(x)dx$ becomes:

$$(9) \quad \hat{f}(\underline{x})d\underline{x} = \frac{dz/dx}{\sigma\sqrt{2\pi}} \exp\left\{-(z-\mu)^2/2\sigma^2\right\} dx$$

with $z = \varphi(x) = \mu + \sigma.\varphi(x)$

4. In this chapter two numerial examples are treated numerically. In the first one in each winter the maximum frost depth is considered. (de Bilt), see [7]. On the basis of the results obtained in 35 winters the following distribution is found:

$$f(\underline{x})d\underline{x} = \frac{e^{-\{\lg(\underline{x}+25)-3.6\}^2/2(\frac{1}{2})^2}}{\frac{1}{2}(\underline{x}+25)\sqrt{2\pi}} d\underline{x}$$

with $\int_{-25}^{\infty} f(\underline{x})d\underline{x} = 1$.

The agreement between the calculated frequency distribution (in the sample of 35 x values) and the theoretical one, based on the distribution mentioned above, is very good.

Vraag :Voor welke b liggen de 24 punten zo„lineaïr mogelijk"?

Antw: b≅25

Dus $\lg(X+25)=t\sigma+\mu$

hier $\lg(X+25)=\frac{1}{2}t+3.6$ of $t=2\lg(X+25)-3.6$

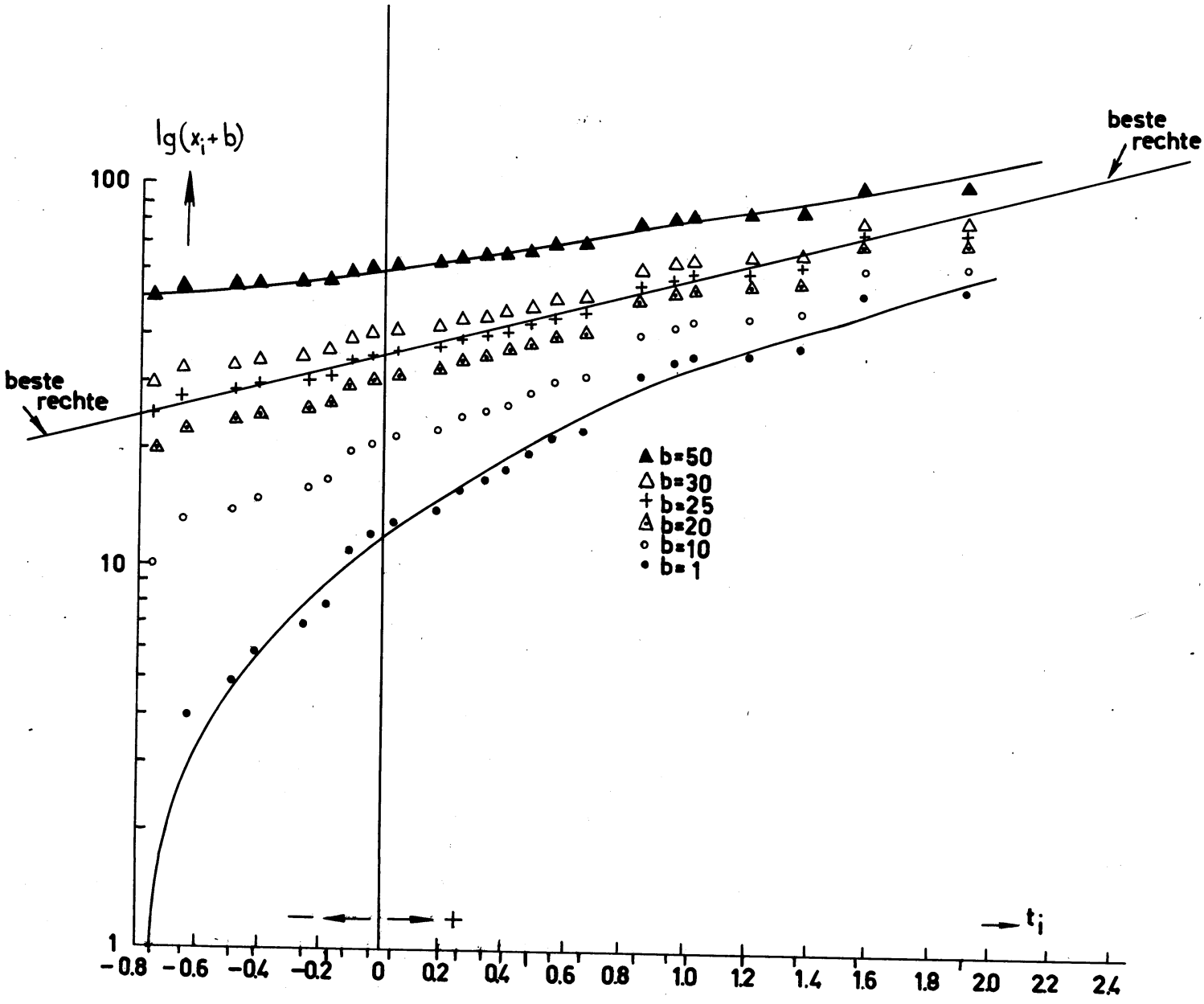


Fig 1

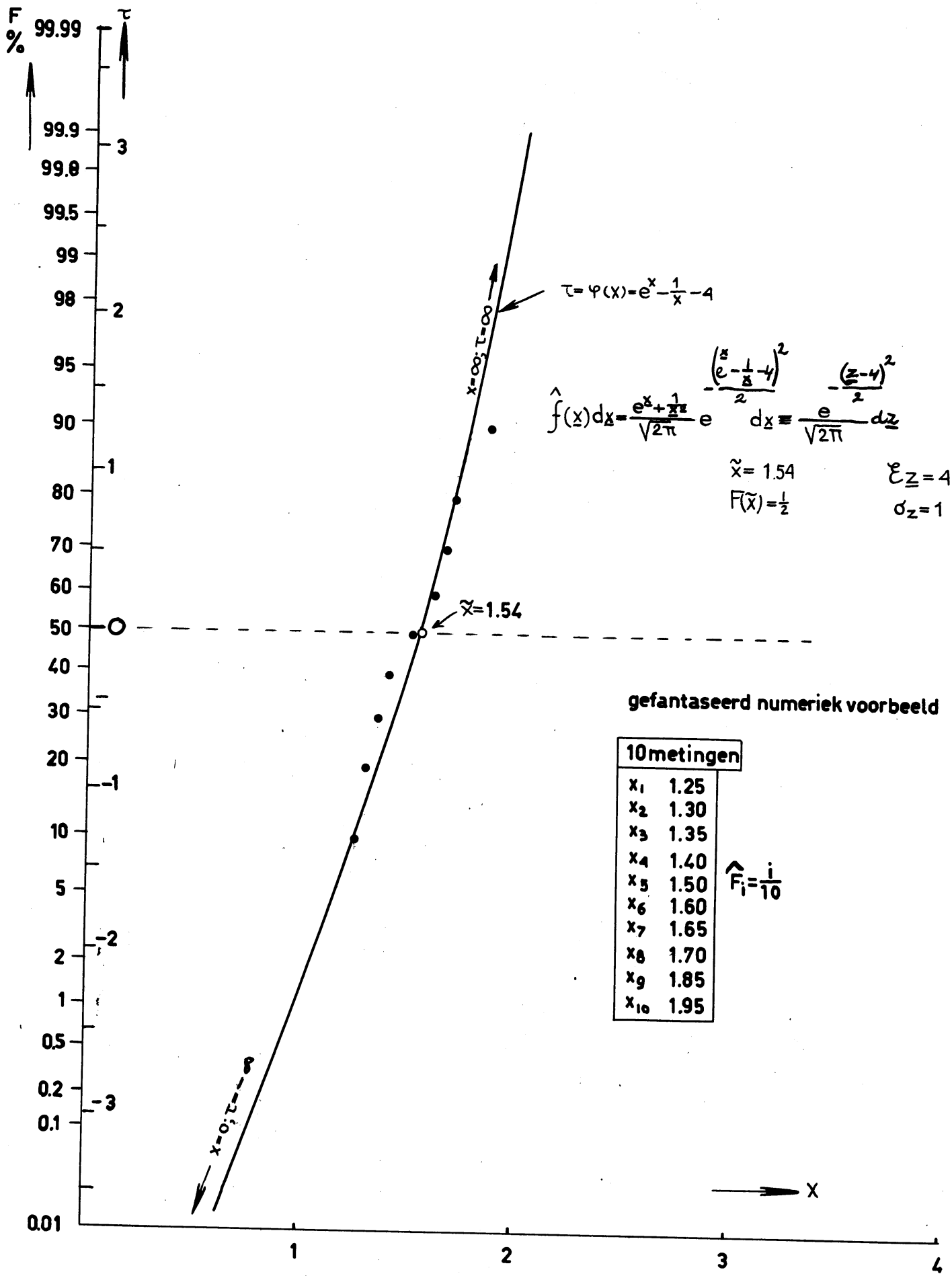

F %
τ
99.99
99.9
99.8
99.5
99
98
95
90
80
70
60
50
40
30
20
10
5
2
1
0.5
0.2
0.1
0.01
3
2
1
0
-1
-2
-3
x=8; τ=8
x=0; τ=-8
τ= φ(x)=e^x − 1/x − 4
f̂(x)dx = e^{x+1/x²}/√2π · e^{−(e^x − 1/x − 4)²/2} dx ≡ e/√2π · e^{−(z−4)²/2} dz
x̃= 1.54
F(x̃)=½
ε_z = 4
σ_z = 1
x̃=1.54
gefantaseerd numeriek voorbeeld
10 metingen
x1 1.25
x2 1.30
x3 1.35
x4 1.40
x5 1.50
x6 1.60
x7 1.65
x8 1.70
x9 1.85
x10 1.95
F̂i = i/10
X
1
2
3
4

Fig 2