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. 57-002 (III-198)

Dr. C. Levert

Het normaliseren van een niet-normale verdeling door middel van de transformatie $z = x^a$.

Normalization of a non-normal distribution by means of the transformation $z = x^a$.

De Bilt, 1957.

KONINKLIJK NEDERLANDS
METEOROLOGISCH INSTITUUT

Wetenschappelijk Rapport
W.R. 57-002 (III-198)
U.D.C.: 519.2

Dr. C. Levert

Het normaliseren van een niet-normale verdeling door middel van de transformatie $z = x^a$.

Rapport, geschreven naar aanleiding van:
J. Neumann and S. Kotz,
"Some Pearsonian-like frequency functions capable of a modified normalization".

INHOUD

1. Inleiding.

Het in de titel genoemde artikel van J. Neumann en S. Kotz (Israël), werd aangeboden aan de deelnemers aan de conferentie: "The scientific basis of Weather Modification Studies", georganiseerd door de Universiteit van Arizona en gehouden in het Institute of Atmospheric Physics, Tucson, Arizona 10-12 april 1956. Neumann vroeg mij om commentaar.

De inhoud van het artikel én mijn commentaar worden beide in dit rapport opgenomen.

Hoewel de schrijvers zich opzettelijk beperken tot een continue reële variabele $\underline{x}$, die slechts waarden tussen 0 en ∞ aannemen kan, d.i. $0 \leqslant x \leqslant \infty$, zou ik willen beginnen met de definitie van "normaliseren" zoveel mogelijk te generaliseren. *)

1. Zij gegeven een waarschijnlijkheidsverdeling $f(\underline{x})\, d\underline{x}$, gedefinieerd voor een reële continue variabele $\underline{x}$, waarvoor $p \leqslant x \leqslant q$.
2. Dan geldt p.d. $\mu_0 = \int_p^q f(\underline{x})\, d\underline{x} = 1$.
3. $\mu_i = \int_p^q \underline{x}^i f(\underline{x})\, d\underline{x} = i^{de}$ moment; $i = 0, 1, 2....$; $\mu_1 = \mathcal{E}x = \bar{x}$
4. $\sigma^2 = \mu_2 - \mu_1^2 =$ variantie.

5. De opgave is te onderzoeken of wij deze verdeling over het gehele traject, $p \leqslant x \leqslant q$ kunnen normaliseren d.m.v. een transformatie $z = \varphi(x)$, (verondersteld, dat binnen dit traject de verdeling van x zelf niet gauszisch is).
Dit betekent p.d. het volgende: uit $z = \varphi(x)$ volgt $x = \alpha(z)$, de inverse functie en $dx = \alpha^1(z)dz$.
Substitueer beide in (1), zodat $f(x)\, dx = \psi(z)\, dz$
Als er 2 waarden μ_z en σ_z bestaan, zódanig, dat voor iedere $p \leqslant x \leqslant q$ zou gelden, dat
6. $\int_p^x f(\underline{x})\, d\underline{x} = \int_{z_1}^z \psi(\underline{z})\, d\underline{z}$, dan heet $f(\underline{x})d\underline{x}$ binnen het traject $p \leqslant x \leqslant q$ normaliseerbaar d.m.v. de transformatie $z = \varphi(x)$.
Hierbij is $z_i = \varphi(x_i)$ en $\varphi(p) < \varphi(x) < \varphi(q)$ of $\varphi(p) > \varphi(x) > \varphi(q)$

---

*) Een grootheid, die een waarschijnlijkheidsverdeling heeft, wordt onderstreept (een stochastische, statistische variabele; een "stochastiek"). Een speciale waarde ervan wordt niet onderstreept.

Verder

7. heet $\varphi(\underline{z})\, d\underline{z} = \left(\sigma_{\underline{z}}\sqrt{2\pi}\right)^{-1} \exp.\left\{-(\underline{z}-\mu_{\underline{z}})^2 : 2\sigma_{\underline{z}}^2\right\} d\underline{z}$ de normale verdeling op $\underline{z}$, met

8. een gemiddelde $\mathcal{E}\,\underline{z} \equiv \int_{-\infty}^{\infty} \underline{z}\,\varphi(\underline{z})\, d\underline{z} = \mu_{\underline{z}}$ en een variantie $\sigma_z^2 = \mathcal{E}\underline{z}^2 - \left(\mathcal{E}_{\underline{z}}\right)^2 = \int_{-\infty}^{\infty} (\underline{z}^2 - \underline{z})\varphi(\underline{z})\, d\underline{z}$.

Wij hebben opzettelijk aan p en q nog geen speciale numerieke waarden toegekend. Er zijn vele mogelijkheden. Hier volgen er vier

a) p = 0 ; q = ∞ ; dagsommen neerslag

b) p = 0 ; q = 100; dagelijkse zonneschijnpercentages

c) p = -273; q = ∞ ; gemiddelde wintertemperaturen

d) p = 0 ; q = 365; aantal regendagen per jaar.

In deze gevallen moet aan $\varphi(x)$ de eis gesteld worden:
a) $\varphi(0) = -\infty$; $\varphi(\infty) = \infty$; b) $\varphi(0) = -\infty$ $\varphi(100) = \infty$; c) $\varphi(-273) = -\infty$ $\varphi(\infty) = \infty$; d) $\varphi(0) = -\infty$; $\varphi(365) = +\infty$ (ofschoon, zoals blijkt in 2.1, ook wel $\varphi(0) = 0$ gesteld mag worden, doch dan normaliseert men tot de half-normale frequentie-verdeling).

Wanneer de gegeven f (x) te schrijven is, over het gehele traject $p \leqslant x \leqslant q$, in de vorm

9. $\boxed{f(x) = \left(b\sqrt{2\pi}\right)^{-1} \cdot \varphi'(x)\, exp.\left[-\{\varphi(x)-a\}^2/2b^2\right]}$

dan is deze $\varphi(x)$ tevens de normalisatie-transformatie functie $z = \varphi(x)$. Want dan is

10. $f(\underline{x})d\underline{x} = \left(b\sqrt{2\pi}\right)^{-1} \exp.\left\{-(\underline{z}-a)^2 : 2b^2\right\} d\underline{z}$ en dus $\mu_{\underline{z}} = a$; $\sigma_z = b$.

In de leerboeken treft men allerlei functies aan, zoals

11. $\varphi(x) = c + dx$; $(m + n\,x)^2$; $a_o + a_1 x + a_2 x^2 + \ldots a_k x^k$ (k = 1, 2, 3...); d + log (c + x), enz., met constanten c, d, m, n, q, a_o, a_1....

2. Het artikel van Neumann en Kotz.

2.1. Welke vorm bezit de normaliseerbare verdelingsfunctie?

Gegeven een waarschijnlijkheidsverdeling f (x) d x, gedefinieerd voor $0 \leqslant x \leqslant \infty$. Gezocht wordt die exponent a, waarvoor $z = \varphi(x) = x^a$ wèl normaal verdeeld is (a reëel en $\neq 0$). Wij spreken af z niet negatief te nemen; bijv. bij $a = \frac{1}{2}$ en x = 9 is de z zowel -3 als +3; toch slechts beschouwen : $0 \leqslant z < \infty$

De f (x) dx, die aldus normaliseerbaar is, móet de gedaante hebben

12. $\boxed{f(\underline{x})d\underline{x} = \frac{2|a|}{\sigma_{\underline{z}}\sqrt{2\pi}}\, \underline{x}^{a-1}\, e^{-\underline{x}^{2a}/2\sigma_{\underline{z}}^2}\, d\underline{x}}$ (2 parameters: a en $\sigma_{\underline{z}}$)

In de exponent zal zeker niet kunnen staan $(\underline{x}^a - \mu)^2 : 2\sigma^2$, met $\mu \neq 0$, want de eis, dat $\int_0^\infty f(\underline{x})d\underline{x} = 1$ brengt mee, dat

13. ook $2\int_{\varphi(0)}^{\varphi(\infty)} \frac{e^{-(\underline{z}-\mu)^2/2\sigma_z^2}}{\sigma_z\sqrt{2\pi}} d\underline{z} = 2 \times \frac{1}{2} = 1$ en dit kan slechts als $\mu = 0$, omdat $z_0 = \varphi(0) = 0$ en $z_\infty = \varphi(\infty) = \infty$ (of ∞ resp. 0, al naar gelang $a \gtrless 0$). De $\underline{z}$-verdeling, waarin $f(\underline{x})d\underline{x}$ overgaat door op z te transformeren, heet <u>half-normaal</u>.

De σ_z draagt de index z, omdat σ_z^2 de variantie van de (gehele)

14. normale $\underline{z}$-verdeling is, d.w.z. $\int_{-\infty}^{\infty} z^2 \varphi(\underline{z})d\underline{z} = \sigma_z^2$, aangezien $\int_{-\infty}^{\infty} \varphi(\underline{z})\, d\underline{z} = 1$, met $\varphi(\underline{z}) = (\sigma_z\sqrt{2\pi})^{-1} \exp.(-\underline{z}^2/2\sigma_z^2)$

Vanzelfsprekend hangt de x-variantie σ_x^2, gedefinieerd door $\int_0^\infty (x-\mu_x)^2 f(x)dx$, samen met genoemde σ_z.

Het verband is

15. $\sigma_z = (\sigma_x \frac{\sqrt{\pi}}{\sqrt{G}})^a$, waarin $G = \sqrt{\left(\frac{1}{2} + \frac{1}{a}\right)}$

Alle door middel van $z = x^a$ normaliseerbare waarschijnlijkheidsverdelingen $f(\underline{x})d\underline{x}$, gedefinieerd voor $0 \leqslant x < \infty$, hebben blijkbaar de gedaante

16. $$f(\underline{x})d\underline{x} = \frac{2|a|\, G^{\frac{1}{2}a}}{\pi^{\frac{1}{4}(a+2)}\sigma_x^a} \underline{x}^{a-1} \exp.\left\{-\frac{G^a}{\pi^{\frac{1}{2}a}\sigma_x^{2a}} \underline{x}^{2a}\right\} d\underline{x}$$

, met twee constanten a en σ_x. (deze σ_x is de standaarddeviatie van x).

Het derde type waarschijnlijkheidsverdeling van Pearson is: $p(\underline{x})d\underline{x} = A\,(1 + \frac{\underline{x}}{b})\exp\{-\gamma\underline{x}\}d\underline{x}$ waarin A een normeringsconstante is; $-b \leqslant x < \infty$; a en γ zijn constanten. Men ziet, dat (16) tot dit type behoort, nl. $a = \frac{1}{2}$; $\gamma b = \frac{1}{2}$)

2.2. <u>Hoe onderzoeken wij of er een normalisatie-transformatie $z = x^a$ bestaat?</u>

Nadat aldus de meest algemene vorm voor $f(\underline{x})d\underline{x}$, gedefinieerd voor $0 \leqslant x < \infty$ en normaliseerbaar door $z = x^a$, gevonden is, rijst het probleem hoe wij moeten onderzoeken of zulk een a bestaat, terwijl de $f(\underline{x})d\underline{x}$ zelf niet exact bekend is, omdat namelijk slechts N x-waarden $\geqslant 0$ gegeven zijn.

Neumann ontwikkelt de volgende 2 kriteria: 2.2.1 en 2.2.2.

2.1.1 <u>De momenten-methode.</u>

De momenten μ_n (t.o.v. $x = 0$) van $f(\underline{x})\,d\underline{x}$ (zie 12) zijn gedefinieerd door

17. $\mu_r \equiv \int_0^\infty \underline{x}^r f(\underline{x})\, d\underline{x} = \frac{(2b^2)^{r/2a}}{\sqrt{\pi}} \Gamma\left(\frac{r}{2a} + \frac{1}{2}\right)$ met r = 0, 1, 2, 3, ...

$(\mu_0 = 1;\ \mu_1 = \mathcal{E}\underline{x} = \bar{x})$

N.B. De index z in σ_z laten wij nu maar weg .

Voor alle positieve a-waarden bestaan alle momenten.
Voor alle negatieve a-waarden zijn alleen de momenten met $r < |a|$ eindig en bepaald, d.w.z. voor alle $r \geq |a|$ oneindig .

Wij gebruiken het symbool μ_r voor het r^{de} moment in het universum en het symbool M_r voor het r^{de} moment in de empirische verdeling (zijnde een steekproef), gedefinieerd door

18. $M_r = \frac{1}{N} \sum_1^N x_i^r$, rekening houdende met eventuele multipliciteiten van x.

Kies 4 gehele, positieve getallen p, q, r, s zodanig dat p + q = r + s, dan is

19. $\frac{\mu_p \mu_q}{\mu_r \mu_s} = \frac{\Gamma\left(\frac{p+a}{2a}\right)\Gamma\left(\frac{q+a}{2a}\right)}{\Gamma\left(\frac{r+a}{2a}\right)\Gamma\left(\frac{s+a}{2a}\right)}$

De onbekende b is op deze wijze geëlimineerd en alleen de parameter a is overgebleven.

Kies r = s en q = 0 en vervolgens s = 1, 2, 3, 4..., dan is p = 2, 4, 6 ... enz.

20. $\frac{\mu_{2s}}{\mu_s^2} = \sqrt{\pi}\, \frac{\Gamma\left(\frac{s}{a} + \frac{1}{2}\right)}{\left\{\Gamma\left(\frac{s}{2a} + \frac{1}{2}\right)\right\}^2}$

(N.B.: $\mu_0 = 1$)
Substitueer s = 1, 2, 3, 4..... dan komt er

21. $\mu_2/\mu_1^2 = \sqrt{\pi}\, \frac{\Gamma\left(\frac{1}{a} + \frac{1}{2}\right)}{\left\{\Gamma\left(\frac{1}{2a} + \frac{1}{2}\right)\right\}^2} = \phi(a) = \Phi_1$

22. $\mu_4/\mu_2^2 = \sqrt{\pi}\, \frac{\Gamma\left(\frac{2}{a} + \frac{1}{2}\right)}{\left\{\Gamma\left(\frac{1}{a} + \frac{1}{2}\right)\right\}^2} = \phi\left(\frac{1}{2}a\right) = \Phi_2$

23. $\mu_6/\mu_3^2 = \sqrt{\pi}\, \frac{\Gamma\left(\frac{3}{a} + \frac{1}{2}\right)}{\left\{\Gamma\left(\frac{1}{2a} + \frac{1}{2}\right)\right\}^2} = \phi\left(\frac{1}{3}a\right) = \Phi_3$

enz. Op deze wijze is derhalve ook een (ingewikkelde) relatie tussen Φ_1 en Φ_2 vastgelegd en eveneens een tussen Φ_1 en Φ_3 of tussen Φ_1 en Φ_4 etc. en het bestaan van deze relaties is geheel het gevolg van het feit, dat de f (x)dx een d.m.v. $z = x^a$ normaliseerbare vorm heeft, d.w.z. van type (16) is.

Neumann verving a door $\frac{1}{m}$ en berekende $\phi(m) = \frac{\Gamma(m+\frac{1}{2})}{\left\{\Gamma\left(\frac{m}{2}+\frac{1}{2}\right)\right\}^2}$ en bracht deze functie in een grafiek van ϕ tegen m.

Hier volgt de tabel:

tabel 1

$$\phi(m) = \sqrt{\pi}\,\frac{\Gamma\left(m+\frac{1}{2}\right)}{\left\{\Gamma\left(\frac{m}{a}+\frac{1}{2}\right)\right\}^2}$$

m	$\phi(m)$	m	$\phi(m)$	m	$\phi(m)$
- 0.50	∞	0.0	1.000	2.4	3.927
- 0.48	7.2410	0.2	1.038	2.6	4.491
- 0.45	3.2050	0.4	1.150	2.8	5.141
- 0.40	1.8835	0.6	1.244	3.0	5.890
- 0.35	1.4540	0.8	1.398	3.2	6.751
- 0.30	1.2551	1.0	1.571	3.5	8.284
- 0.25	1.1543	1.2	1.779	3.8	10.171
- 0.20	1.0776	1.4	2.022	4.0	11.641
- 0.15	1.0370	1.6	2.303	4.2	13.383
- 0.10	1.0149	1.8	2.626	4.5	16.444
- 0.05	1.0035	2.0	3.000	4.8	20.208
0.00	1.0000	2.2	3.430	5.0	23.194

Zie ook de grafiek fig. 1.

Waarden van m, sterker negatief dan $-\frac{1}{2}$, behoeven niet beschouwd te worden. Immers dan ligt a tussen 0 en -2 en $|a|$ tussen 0 en 2, hetgeen betekent, dat slechts het eerste moment μ_1 bestaat. (zie de regels even boven formule (18)). Voorwaarden van m tussen 0 en $-\frac{1}{2}$ (waarbij $\phi(m)$ varieert tussen resp. 1 en $+\infty$) ligt a tussen $+\infty$ en -2 of $|a|$ tussen ∞ en + 2, zodat alleen momenten μ_r bestaan, mits $r < |a|$.

Om de formules (21) en (22) te kunnen toepassen, moet bij negatieve a-waarden $|a| > 4$ zijn, d.w.z. bij normaliserende functies $z = x^{-a^1}$, met $a^1 > 4$.

De procedure is als volgt. Bereken de momenten M_1, M_2 en M_4 uit het empirische materiaal. Het ene quotiënt $q_1 = \frac{M_2}{M_1^2}$ levert, via de kromme in fig. 1, zekere a_1 en het andere $q_2 = \frac{M_4}{M_2^2}$ zekere a_2. Indien $a_2 = \frac{1}{2}a_1$, (d.i. $m_2 = 2\,m_1$) in voldoende

benadering, is daarmee de mogelijkheid tot normalisatie d.m.v. $z = x^a$ bewezen en tevens de exponent a gevonden.

24. Uit $M_2 = \frac{1}{N} \sum_1^N x_i^2 = \frac{(2\sigma^2)^{1/a}}{\sqrt{\pi}} \Gamma\left(\frac{1}{a} + \frac{1}{2}\right)$ volgt dan σ.
Men kan eveneens σ berekenen uit

25. $M_1 = \frac{1}{N} \sum_1^N x_i = \frac{(2\sigma^2)^{1/2a}}{\sqrt{\pi}} \Gamma\left(\frac{1}{2a} + \frac{1}{2}\right)$ of uit

26. $M_3 = \frac{1}{N} \sum_1^N x_i^3 = \frac{(2\sigma^2)^{1/3a}}{\sqrt{\pi}} \Gamma\left(\frac{3}{2a} + \frac{1}{2}\right)$ enz.

De methode der Maximum Likelihood leert, dat de beste schatting van σ^2 volgt uit

27. $\boxed{\sigma^2 = \frac{1}{N} \sum x_i^{2a}}$ (met de multipliciteiten der x-waarden rekening houdende)

2.2.2 <u>De methode der Maximum Likelihood</u>.

Er zijn N x-waarden gemeten en wel n_1 keren x_1, n_2 keren x_2, ... n_j keren x_j, zodat $N = \sum_{i=1}^{j} n_i$. Zij de waarschijnlijkheidsverdeling in het universum $f(\underline{x})d\underline{x}$, met

28. $f(\underline{x}) = \frac{|a|}{\sigma\sqrt{2\pi}} \underline{x}^{a-1} \exp.\left\{-\underline{x}^{2a}/2\sigma^2\right\}$, normaliseerbaar d.m.v. $z = x^a$

De onbekende parameters zijn a en σ.
De kans op een steekproef, bestaande uit n_1 waarden x tussen x_1 en $x_1 + dx_1$, n_2 waarden x tussen x_2 en $x_2 + dx_2$ .... n_j-waarden x tussen x_j en $x_j + dx_j$ is

29. $P(x_1, x_2, \ldots x_j) = \left[f(x_1)dx_1\right]^{n_1} \cdot \left[f(x_2)dx_2\right]^{n_2} \ldots \left[f(x_j)dx_j\right]^{n_j}$

De Likelihood is (volgens Fisher) gedefinieerd door:

30. $L = \prod_{i=1}^{j} \left\{f(x_i)\right\}^{n_i}$ = functie van $x_1, x_2 \ldots x_j$, a en σ.

De beste numerieke waarden â en $\hat{\sigma}$ van a en σ zijn die, waarvoor L, bij het gegeven stelsel $x_1, x_2 \ldots x_j$, maximaal is. Voor deze

31. waarden is derhalve $\left(\partial L/\partial a\right)_{\hat{a},\hat{\sigma}} = 0$ en $\left(\partial L/\partial \sigma\right)_{\hat{a},\hat{\sigma}} = 0$ (men zou ook nog het teken van de tweede afgeleide moeten bekijken om te onderscheiden tussen maximaliseren en minimaliseren van L).

{ en $n_1, n_2 \ldots n_j$

32. Als L maximaal is dan ook lgL en dus: $\left(\frac{\partial \lg L}{\partial a}\right)_{\hat{a},\hat{\sigma}} = \left(\frac{\partial \lg L}{\partial \sigma}\right)_{\hat{a},\hat{\sigma}} = 0$

Door deze twee vergelijkingen zijn $\hat{\alpha}$ en $\hat{\sigma}$ vastgelegd en wel

33. $$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{j} n_i x_i^{2a}$$

34. $$\frac{\overline{x^{2a} \lg x}}{\overline{x^{2a}}} - \overline{\ln x} - \frac{1}{a} = 0$$ ← zie II blz 10

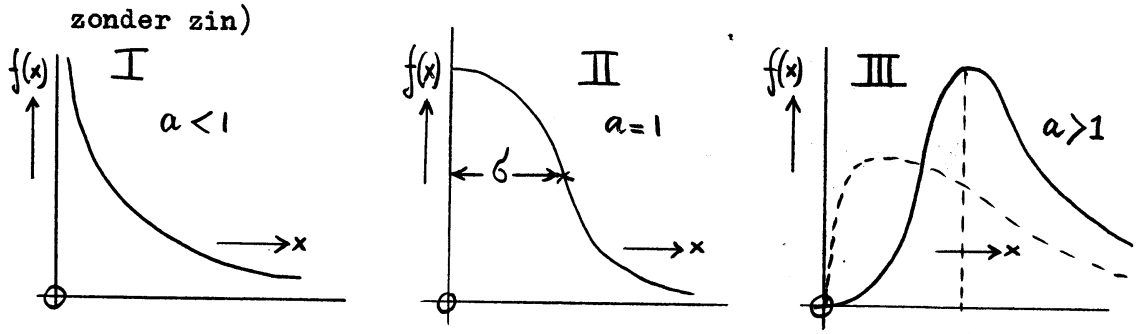
Hierbij duidt de streep ——— op een middelwaarde, rekening houdende met de multipliciteiten, bijv.

$$\overline{x^{2a} \lg x} = \frac{1}{N} \sum_{i=1}^{j} (n_i x_i^{2a} \lg x_i).$$

3. Commentaar.

Gedrag van $f(x)$ in $x = 0$

3.1. Hoe begint $f(x)$ in het punt $x = 0$? Zie (12).
Wij zien, dat $x^{a-1} e^{-x^{2a}/2\sigma^2}$ gelijk is aan òf $\frac{1}{0 \times 1} = \infty$, als $a < 1$, òf $\frac{0}{1} = 0$, als $a > 1$ en 1 als $a = 1$ (, dan staat $f(x)dx$ gelijk met de halve normale verdeling).
Voorts gaat voor iedere $a \neq 0$ $x^{a-1} e^{-x^{2a}/2\sigma^2}$ naar 0 als $x \to \infty$
Bijgevolg kunnen wij 3 typen van (12) onderscheiden ($a = 0$ is zonder zin)



geen max. bij	geen max. bij	een max. bij
een $x > 0$;	een $x > 0$;	een $x > 0$;
$f(0) = \infty$	$f(0)$ = eindig	$f(0) = 0$
$f'(0) = \infty$	$f'(0) = 0$	$f'(0) = 0$ of ∞

Verdelingen, waarin bepaalde momenten ontbreken.

3.2. Voor $a < 0$ zijn alleen de momenten μ_r met $r < |a|$ bepaald en eindig.
Hier volgen enkele voorbeelden.

1. $a = -\frac{1}{2} \to f(\underline{x})d\underline{x} = \frac{1}{\sigma\sqrt{2\pi}} \underline{x}^{-\frac{3}{2}} \cdot e^{-\underline{x}/2\sigma^2} \cdot d\underline{x}$ met $0 \leq x < \infty$.
Voor deze is $\mu_0 = 1$, terwijl $\mu_1, \mu_2 \ldots$ niet eindig en bepaald zijn.

2. $a = -1 \rightarrow f(\underline{x})d\underline{x} = \frac{2\underline{x}^{-2}}{\sigma\sqrt{2\pi}} e^{-\underline{x}^{-2}/2\sigma^2} d\underline{x}$ met $0 \leqslant x < \infty$. Voor deze is wel $\mu_0 = \int_0^\infty f(x)dx = 1$, en weer bestaan μ_1, μ_2, enz. niet. Deze verdeling heeft dus geen gemiddelde, geen standaarddeviatie enz.

3. $a = -2\frac{1}{2} \rightarrow f(\underline{x})d\underline{x} = \frac{5}{\sigma\sqrt{2\pi}} \underline{x}^{-3\frac{1}{2}} e^{-\underline{x}^{-5}/2\sigma^2} d\underline{x}$ met $0 \leqslant x < \infty$. Voor deze is $\mu_0 = 1$, bestaan μ_1 én μ_2 (zodat $\bar{x}$ en σ_x bestaan), doch bestaan μ_3, μ_4 enz. niet. Zo is $\mu_1 = \Gamma(3/10) : 2^{1/5} . \sigma^{1/5} \sqrt{\pi}$

3.3. Bestaat er een "transformatie-exponent" a?

I. De eerste procedure is: bereken de twee quotiënten $q_1 = \frac{\sum x_i^2}{(\sum x_i)^2} N$ en $q_2 = \frac{\sum x_i^4}{(\sum x_i^2)^2} \cdot N$ Gebruik de kromme $\phi(m)$ tegen $m(=\frac{1}{a})$ en zie of m_2, volgende uit $\phi_2 = \phi(m_2)$ het dubbele is van m_1, volgende uit $\phi_1 = \phi(m_1) = q_1$.

Wij vragen: Wat betekent: "zie of $m_2 = 2 m_1$ in voldoende mate"? Hierover zegt Neumann niets. In een concreet numeriek geval vonden wij $m_1 = 1.4$ en $m_2 = 2.2 \neq 2 \times 1.4$. Moet hieruit besloten worden, dat er geen $a = \frac{1}{m}$ bestaat? Of moeten we nemen $\bar{m} = \frac{1}{2}(m_1 + \frac{1}{2}m_2) = 1.25$ en $a = 0.8$? En eigenlijk is het o.i. niet voldoende, dat $m_2 = 2 m_1$, want nodig is ook, dat $q_3 = N(\Sigma x^6)/(\Sigma x^3)^2$ een $m_3 = 3m_1$ levert en dat $q_4 = N(\Sigma x^8)/(\Sigma x^4)^2$ een waarde $m_4 = 4m_1$ levert enz. Blijkbaar acht Neumann het voldoende, dat $m_2 = 2 m_1$, wellicht omdat door steekproefeffecten zeer gemakkelijk de quotiënten q_3, q_4 enz. zo sterk kunnen variëren, dat de kans, dat door het toeval aan de eisen $m_3 = 3 m_1$, $m_4 = 4 m_1$ enz. volstrekt niet voldaan wordt zeer groot is (een reden, waarom men niet graag met momenten van orden 4 en hoger rekent). Wij kunnen de vraag stellen hoe de zinsnede "zie of $m_2 = 2 m_1$ in voldoende mate geldt" statistisch gepreciseerd zou kunnen worden. Daarvoor is nodig, lijkt ons, dat wij de "sampling distributions" van a en σ kennen. Wij bedoelen: gegeven een $f(\underline{x})d\underline{x}$ van de gedaante (12), doch vervang a en σ daarin door twee gegeven numerieke waarden α en β. Denk talloos vele malen een aselecte steekproef uit dit universum genomen en iedere keer via de waarnemingen in de steekproef de beste schattingen $\hat{a}$ en $\hat{b}$ berekend. Dan gehoorzaamt zowel $\hat{a}$ als $\hat{\ell}$ aan een bepaalde waarschijnlijkheidsverdeling, die de α, β en N als parameters zal bevatten, d.w.z.: $w_1(\hat{a}/\alpha, \beta, N)d\hat{a}$ en $w_2(\hat{b}/\alpha, \beta, N)d\hat{b}$. N.B. w_1 bevat óók de β en w_2 bevat ook de α als parameter, hetgeen betekent, dat $\hat{a}$ en $\hat{b}$ niet afhankelijk van elkaar verdeeld zijn.

Zodra deze w_1 en w_2 bekend zijn (en wij hopen tevens, dat elk dezer twee verdelingen voor grotere N steeds minder van een normale verschilt) kan men ook zeggen binnen welke grenzen de theoretische α gelegen zal moeten zijn op basis van het empirische materiaal, met een voorgeschreven zekerheid (bijv. 0.95). Aldus is wellicht zowel voor de m_1 volgende uit q_1, als voor de $\frac{1}{2}m_2$ volgende uit q_2, een betrouwbaarheidsmarge af te leiden en als deze een gemeenschappelijk deel hebben, mogen wij zeggen dat aan $m_2 = 2m_1$ in voldoende mate voldaan is en doen wij inderdaad wellicht het beste om $\bar{m} = \frac{1}{2}(m_1 + \frac{1}{2}m_2)$ te nemen, waaruit weer $\bar{a} = 1/\bar{m}$ volgt.

Het lijkt me niet eenvoudig deze betrouwbaarheidsmarges te berekenen, doch eerst dan is dit kriterium van Neumann werkelijk een kriterium (d.i. toets).

N.B. Een geval, waarin de a zeker niet bestaat, is dat, waarin tenminste één der 2 quotiënten q_1 en q_2 kleiner dan 1 is; bij deze is dan nl. de m, zie de kromme in fig. 1 imaginair.

II. De andere procedure is om de vergelijking (34) op te lossen. Heeft deze vergelijking, zo ze een reële oplossing heeft, slechts één oplossing? Hoe op te lossen? Vermoedelijk slechts grafisch? Kan het zijn, dat er geen reële wortel a is louter vanwege het steekproef-effect (d.w.z. terwijl tóch het universum een verdeling heeft, weergegeven door (12)? Deze vragen behandelt Neumann niet. Ook hier mist het kriterium de "finishing-touch", die wel zeer moeilijk te geven zal zijn.

Wat is het gevolg van $x_i = 0$ waarden? Hoe dan te handelen met vergelijking (34)? Dat Neumann de mogelijkheid $x_i = 0$ niet uitsluit, blijkt uit zijn onderstelling $0 \leqslant x < \infty$.

De moeilijkheden ontstaan vanwege het feit, dat in de L, zie (30), geen functie-waarden $f(x_i)$ gelijk nul of gelijk ∞ mogen optreden (ontaarding), en dat nu juist wèl het geval voor $x_i = 0$ bij $a > 1$ resp. < 1. De kwestie is, dat wij voor de kans, dat x gelegen is tussen x en x +dx, schrijven f(x)dx, hetgeen juist erg fout kan zijn voor $f(o) = 0$ resp. ∞. Als er tenminste één functiewaarde $f(x_i)$ oneindig groot is voor zekere x_i (hier 0), dan is L ongeacht de waarden van a (mits < 1) en σ altijd maximaal, ∞ namelijk, en is er geen maximaal-probleem. Hoe dan te handelen? Neumann spreekt er niet over. Wellicht doen we goed alle gemeten x-waarden te vermeerderen met een klein bedrag ε, dat klein is t.o.v. de eenheid, waarin x wordt uitgedrukt, - bijv. x in cm,

dan $\varepsilon = 0.2$ cm. Daardoor komen wij tot een verg. in a, zie weer (34), waarin nu iedere x_i door $x_i + \varepsilon$ vervangen moet worden. Los deze op; wortel a. Kies nu een iets andere ε^1. Deze leidt tot een wat andere a^1. Misschien is het mogelijk de waarde â te vinden, waarin de reeks a, a^1, a^{11}, a^{111} .... asymptotisch overgaat bij $\varepsilon \to 0$. Dan is dit de beste â.

Dezelfde moeilijkheid ondervinden wij bij de berekening van $\hat{\sigma}$, zie (33), zodra a bekend is. Hier zou $\sigma = \infty$ (en 0) zijn, zodra er één $x_i = 0$ zou zijn en indien $a < 0$ (> 0).

4. Een derde normalisatie-kriterium bij de normalisatie-transformatie $z = x^a$.

Numeriek voorbeeld.

Dit derde kriterium willen wij verduidelijken aan een concreet getallen-voorbeeld, dat wij vernemen uit Rijkoort*). Men beschouwt in ieder van 35 winters de diepte (beneden aardniveau), waartoe de vorst maximaal doordrong. Er waren 8 winters, waarin de vorst niet binnendrong in de grond; in dit geval werd de x nul genoemd. De 35x-waarden zijn $x_1 = x_2 = x_3 = x_4 = x_5 = x_6 = x_7 = x_8 = 0$; $x_9 = 3$; $x_{10} = x_{11} = 4$; $x_{12} = 5$; $x_{13} = x_{14} = 6$; $x_{15} = 7$; $x_{16} = 10$; $x_{17} = 11$; $x_{18} = 12$; $x_{19} = x_{20} = 13$; $x_{21} = 15$; $x_{22} = 16$; $x_{23} = 17$; $x_{24} = 19$; $x_{25} = 21$; $x_{26} = 22$; $x_{27} = x_{28} = 31$; $x_{29} = 34$; $x_{30} = 35$ $x_{31} = 36$ $x_{32} = 38$ $x_{33} = 53$ $x_{34} = 54$ $x_{35} = 60$ (24 ongelijke waarden)

Er geldt $0 \leqq x < \infty$, d.w.z. voor de gezochte $f(\underline{x})d\underline{x}$ is $\int_0^\infty f(\underline{x})d\underline{x} = 1$. Opdat $z = x^a$ een normalisatie-transformatie zal zijn, zal

$$f(\underline{x})d\underline{x} = \frac{2|a|x^{a-1}}{\sigma\sqrt{2\pi}} \exp.\left[-\underline{x}^{2a}/2\sigma^2 \right] d\underline{x};$$

$$\int_0^\infty f(\underline{x})d\underline{x} = 2\int_0^\infty \psi(\underline{z})d\underline{z} = 2\int_0^\infty s(\underline{t})\,d\underline{t} = 1, \text{ als } \psi(z) = (\sigma\sqrt{2\pi})^{-1}$$
$$\exp.\left\{ -\underline{z}^2/2\sigma^2 \right\} \text{ en } s(\underline{t}) = (\sqrt{2\pi})^{-1} \exp.\left\{ -\tfrac{1}{2}\underline{t}^2 \right\} \text{ als } t = z/\sigma.$$

Beschouw $x_9 = 3$. Dan is $\int_0^3 f(\underline{x})d\underline{x} = 2 \cdot \int_0^{3^a} \psi(\underline{z})\,d\underline{z} = 2 \cdot \int_0^{3^a/\sigma} s(\underline{t})d\underline{t}$.
De beste schatting van $\int_0^3 f(\underline{x})d\underline{x}$ wordt gegeven door $\hat{F}[\underline{x} \leqq 3] = \frac{9}{35}$,
zodat $\frac{9}{35} = 2\int_0^{\cdot} s(\underline{t})d\underline{t}$ en $\frac{3^a}{\sigma} = 0.328$.
Vervolgens: $x_{10} = x_{11} = 4$. Nu $\frac{11}{35} = 2\int_0^{\cdot} s(\underline{t})d\underline{t}$ en $\frac{4^a}{\sigma} = 0.405$ enz.

*) P. Rijkoort: Bijdrage tot het bepalen van de meest gunstige diepte voor het leggen van waterleidingsbuizen in verband met het bevriezingsrisico.
W.R. 56-003 1956.

Aldus ontstaan 22 vergelijkingen $t_1 = 0.328 = 3^a/\sigma$; $t_2 = 0.405 = 4^a/\sigma$; ..... $t_{22} = 54^a/\sigma$ of ook $a \log 3 = \log t_1 + \sigma$; $a \log 4 = \log t_2 + \sigma$; ..... $a \log 54 = \log t_{22} + \sigma$ met onbekenden a en σ.
Men neemt dubbellogarithmisch papier en zet daarop de 22 punten $3, t_1$; $4, t_2$; .... $54, t_{22}$ uit.
Daarna wordt gekeken of zij lineair liggen. Zo ja, of in voldoende mate, zo trekke men de beste rechte, die de a en σ levert. Hier rijzen weer enige vragen: Wanneer heet de ligging "voldoende lineair"? Is er een objectief kriterium? Hoe de "beste" rechte te trekken? Wat is "beste"? (de Meth. der Kl. Kw. kan niet toegepast worden). Zolang wij deze vragen niet kunnen beantwoorden, blijft er niet anders over dan de lineariteit op het oog te beoordelen.
Er komt (fig. 2) $a = 0.64$; $\sigma = 0.66$, zodat $z = x^{0.64}$ en

35. $$f(\underline{x})d\underline{x} = 2\,\frac{0.64}{6.66\sqrt{2\pi}}\,\underline{x}^{0.36}\cdot e^{-(\underline{x}^{0.8})^2/2\cdot 6.66^2}\cdot d\underline{x}, \text{ voor } 0 \leqslant x < \infty.$$

Hoe goed de aanpassing is blijkt uit het volgende tabelletje.

interval x cm	aantallen berekend T	aantallen geteld W
0–10	17.0	16
11–10	7.2	8
21–30	4.3	2
31–40	2.9	6
41–50	1.4	0
51–60	0.8	3
rest.	1.4	0
som	35.0	35.0

x cm	T	W	X^2
0–10	17.0	16	0.0588
11–20	7.2	8	0.0890
21–40	7.2	8	0.0890
rest	3.6	3	0.1000
som	35.0	35.0	0.3368

$$\chi^2 = \sum_1^4 \frac{(T-W)^2}{T} = 0.34 \quad \text{bij } \nu = 4 - 3 = 1 \text{ g.v.v., dus } P \sim 0.60$$

Deze $P \gg 0.05$, zodat de aanpassing goed is.

De Momentenmethode van Neumann levert een $q_1 = M_2/M_1^2 = 2.05$ en $q_2 = M_4/M_2^2 = 3.40$, waarmede wij van de grafiek, fig. 1, aflezen: $m_1 = 1.4$ en $m_2 = 2.2$. Helaas is niet $m_2 = 2m_1$, zodat wij maar nemen $\bar{m} = \frac{1}{2}\{1.4 + \frac{1}{2}\cdot 2{,}2\} = 1.25$, zodat $\bar{a} = 1/\bar{m} = 0.80$. Onze eigen procedure geeft 0.64. Misschien mogen wij het verschil tussen 0.80 en 0.64 niet noemenswaard noemen. De andere methode van Neumann leidde tot teveel rekenwerk (mede in verband met de vele nulwaarden van x).

N.B. Men kan natuurlijk ook van lineair waarschijnlijkheidspapier gebruik maken en langs de lineaire schaal een aantal schalen $x^{0.1}$, $x^{0.2}$, $x^{0.3}$ ....x^{1} aanleggen. Zouden wij dan tewerk gaan volgens ons eerste geval, behandeld onder 4, d.w.z. $F(\underline{x} \leqslant 3) = \frac{9}{35}$, en $F(\underline{x} \leqslant 4) = 11/35$; $F(\underline{x} \leqslant 5) = \frac{13}{35}$, enz., zouden wij op dit papier een reeks punten verkregen hebben met coördinaten x_i^a en $\frac{1}{2}F(x \leqslant x_i)$ + 0.5. De beste keuze van a zou dan die geweest zijn, waarvoor deze punten zo goed mogelijk lineair gelegen zouden hebben. Zou dit voor geen enkele a bereikt kunnen worden (louter door "trial and error" dus), dan zou de conclusie geweest zijn, dat een normalisatie-transformatie $z = x^a$ niet bestaat. Hier wordt weer de lineaire ligging op het oog beoordeeld. Is er een beter kriterium?

Zo blijkt duidelijk hoe de moeilijkheid in de beslissing slechts verschoven worden door van de ene procedure op de andere over te gaan. Moet men kiezen, dan zouden wij voorkeur hechten aan de momenten-procedure van Neumann en de onze, ontwikkeld in 4.

5. Summary

1. In this report an article by J. Neumann and S. Kotz, entitled "Some Pearsonian-like frequency functions capable of a modified normalization", is commented (Apr. 1956).

The to called normel z-distribution $\varphi(\underline{z})d\underline{z}$ is defined by(7). Let be given a probability distribution $f(\underline{x})d\underline{x}$ defined for the interval $p \leqslant x \leqslant q$; the i-th moment is $\mu_i = \int_p^q \underline{x}^i f(\underline{x})d\underline{x}$; $\mu_0 = 1$; μ_1 = mean; $\mu_2 - \mu_1^2$ = variance, etc.

Then $z = \varphi(x)$ is called a normalizing transformation function if x itself is not distributed normally but if z is. This means that (6) is true for each value x within the range p,q. If this $\varphi(x)$ can be found explicitely, f (x) can be described as (9), with contants (parameters) a and b, and $z_1 = -\infty = \varphi(p)$

2.1. Let be given a probality distribution $f(\underline{x})d\underline{x}$ for $0 \leqslant x < \infty$. Since the lower limit is 0, this $\underline{x}$ can never be distributed exactly normally; so we might ask whether there is a real value a $(\neq 0)$ so that $z = \varphi(x) = x^a$ is distributed normally. If so, this $f(\underline{x})d\underline{x}$ must be expressed by (12) and of course $\mathcal{E}_{\underline{z}} = \mu_z = 0$. This $\underline{z}$-distribution is so called <u>modified normal</u> or <u>semi-normal</u>. (see the factor 2 in its coefficient). The parameter σ_z in (12) depends on the standarddeviation σ_x of $f(\underline{x})d\underline{x}$; see (15). The most general distribution $f(\underline{x})d\underline{x}$, normalizable by $z = x^a$, looks as (16), with two constants a and σ_x. with $G = \Gamma(\frac{1}{2}+\frac{1}{a})$.

2.2.1 In this section the question whether a normalization transformation $z = x^a$ exists is investigated by the Method of Moments. These moments of order r = 0, 1, 2 ... of distribution (12) are defined by (17). If $a > 0$, all moments are definite and finite.
If $a < 0$, only the moments for $r < |a|$ are definite and finite.

Certain relations between $\Phi_1 = \mu_2 : \mu_1^2$; $\Phi_2 = \mu_4 : \mu_2^2$ etc. can be derived, see (20),(21), (22) and (23).

Neumann and Kotz replace the constant a by $\frac{1}{m}$ and compute a table for $\phi(m) = \frac{\Gamma(m+\frac{1}{2})}{\{\Gamma(\frac{1}{2}m+\frac{1}{2})\}^2}\sqrt{\pi}$ and draw the corresponding figure 1.

The practical procedure is as follows: if a random sample of N values $x_1, x_2 \ldots x_N$ is measured, the moments M_1, M_2, M_4, with $M_r = \sum_1^N x_i^r$, can be computed. The quotient $q_1 = M_2/M_1^2$ furnishes by means of figure 1 a certain value a_1 (or $m_1 = 1 : a_1$) and the second

quotient $q_2 = M_4 : M_2^2$ gives a certain value a_2 (or $m_2 = 1:a_2$). If $a_2 = \frac{1}{2}a_1$ (or $m_2 = 2\,m_1$) is true "in good approximation" then a normalizing transformation exists. This is $z = x^{a_1}$. Then the equations (24) on (25) or (26), by using the sample moments, or better (27), by using the Method of Maximum Likelihood, furnish the parameter σ.

2.2.2 In this the Method of Maximum Likelihood is applied to find the best $\hat{a}$ and $\hat{\sigma}$, by means of (34) and (33). In this case N x-values are supposed to be measured: n_1 times x_1; n_2 times x_2; .....n_j times x_j, with $N = \sum_1^j n_j$.

3.1. and 3.2. The feature of f(x) in the point x = 0 is commented. See the sketches I, II, III. Three examples of probability distributions are shown numerically with non-existing $\mu_1, \mu_2, \mu_3 \ldots\ldots$

3.3. In this section the "rule": "see whether $m_2 = 2m_1$ (or $a_2 = \frac{1}{2}a_1$) is in good approximation" is commented. In a special numerical example the values $m_1 = 1.4$ and $m_2 = 2.2$ were obtained, so that $2m_1 = 2.8 > m_1$. We considered $\bar{m} = \frac{1}{2}(m_1 + \frac{1}{2}m_2) = 1.25$ and $\bar{a} = 1 : \bar{m} = 0.8$ as the "best" value of m and a and wonder whether this is right? Stress is laid on the fact that the condition $m_2 = 2m_1$ is not sufficient but necessary, because it is also necessary that $m_3 = 3m_1$, $m_4 = 4\,m_1$ etc. and these further conditions are not considered by N. and K. But in consequence of sampling effects it may be very probable that q_3, q_4 etc. may vary so strongly that the probability that $m_3 \neq 3m_1$ and $m_4 \neq 4m_1$, and so on, may be very large.

Still the question arises: how can the requirement "in good approximation" better specified statistically? Therefore it seems desirable to derive the sampling distributions of both a and σ when drawing random samples from the universe with distributions (12)

This seems a very difficult problem. As soon as it is possible to compute these sampling distributions it is also possible to give a confidence range for m_1 based on the measurements $x_1, x_2 \ldots\ldots x_N$ and also for $\frac{1}{2}m_2$. If these ranges overlap each other than $m_2 = 2\,m_1$ holds good in the sense of "good approximation", just by definition.

Next the question is commented how to solve the equation (34)? An exact solution seems nearly impossible, so a graphical solution must be given. How many real roots arise and is it possible that the equation does not furnish any real roots just by sampling effects? Neumann and Kotz do not treat these matters.

Also the question arises: what is the consequence of values $x_i = 0$? Neumann and Kotz suppose $0 \leq x < \infty$. See L in (30). Now (32) cannot be solved because of degeneration. We suggest a new solution: all measured values x_i should be increased by a small quantity ε and then the value a can be solved (graphically). Next the same is done for a new small quantity ε^1 and for $\varepsilon'' < \varepsilon' < \varepsilon$, etc. Perhaps it is possible to find the asymptotic limit value a^* in the sequence of a-values when $\varepsilon \to 0$.

4. A third criterion by means of which it is possible to investigate the existence of a normalizing transformation $z = x^a$ is developed. Here we follow the general method of normalizing discussed in our report W.R.57-001(III-197)Use is made of so called lineair probability paper. The method is explained with the aid of a numerical example. In each of 35 winters the maximum frost depth x in the ground was measured: $x_1 = x_2 = \ldots\ldots x_8 = 0$; $x_9 = 3$; $x_{10} = x_{11} = 4$ $x_{12} = 5$ $x_{13} = x_{14} = 6$ $x_{15} = 7$ $x_{16} = 10$ $x_{17} = 11$ $x_{18} = 12$ $x_{19} = x_{20} = 13$ $x_{21} = 15$ $x_{22} = 16$ $x_{23} = 17$ $x_{24} = 19$ $x_{25} = 21$ $x_{26} = 22$ $x_{27} = x_{28} = 31$ $x_{29} = 34$ $x_{30} = 35$ $x_{31} = 36$ $x_{32} = 38$ $x_{33} = 53$ $x_{34} = 54$ $x_{35} = 60$.

Consider $x_9 = 3$ Then $\int_0^3 (\underline{x})d\underline{x} = 2 \int_0^{3^a} \varphi(z)dz = 2 \int_0^{3^a/\sigma} s(t)dt$, with $z = x^a$ and $t = z/\sigma$; $s(t)dt$ represents the standardized normal distribution ($\mu = 0$; $\sigma = 1$). The best estimation of the unknown probability $\int_0^3 f(\underline{x})d\underline{x}$ is $\hat{F}[\underline{x} \leq 3] = \frac{9}{35}$ and hence $t_3 = 0.328 = 3^a/\sigma$. Also $\hat{F}[\underline{x} \leq 4] = 11/35$ and hence $t_4 = 0.405 = 4^a/\sigma$ and so on. In this way 22 different values x furnish 22 different equations of the form $t = x^a/\sigma$ or $a \log x = \log t + \sigma$ with two unknown values a and σ. Plot these pairs of values x, t on double logarithmic paper and see whether the 22 points may be considered situated linearly (fig. 2). If so, draw the best straight line through the points and compute the best values $\hat{a}$ and $\hat{\sigma}$. In our case $\hat{a} = 0.64$ and $\hat{\sigma} = 0.66$ were obtained and hence the best estimation of $f(\underline{x})d\underline{x}$ is expressed by (35). The χ^2-test shows that a very good agreement exists between the empirical and theoretical distribution ($P \sim 0.60$; $\nu = 1$). In this new procedure other difficulties, instead of Neumann's and Kotz's rule "in sufficient approximation", arise : how do we judge whether the set of points lie virtually linear? And what is the "best" straight line.

(N.B. The points have different weights). Again the "finishing touch" for a good test (good in a statistical sense) seems difficult to be given.

De Bilt, July 1957.

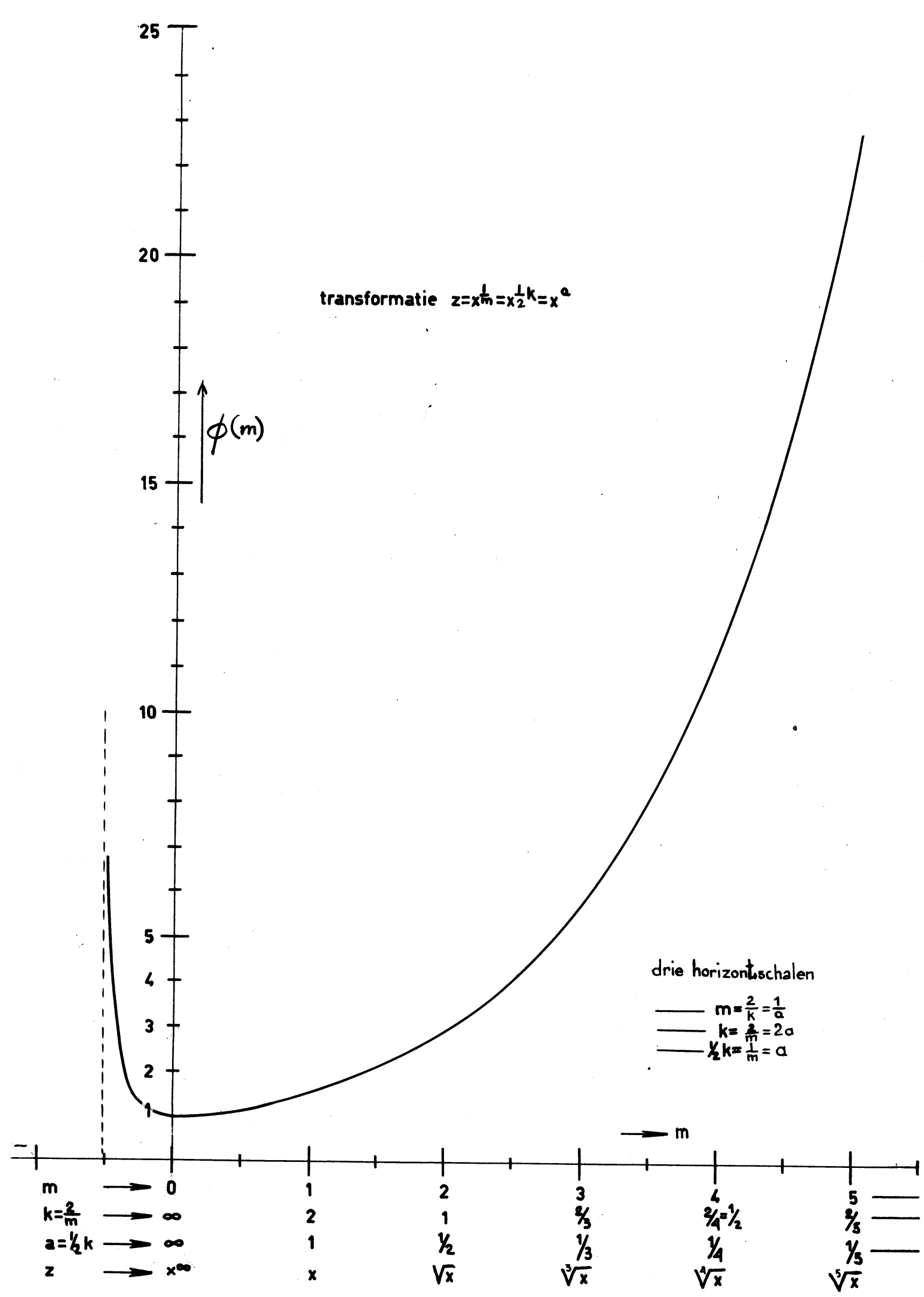

25
20
15
10
5
4
3
2
1
transformatie $z=x^{\frac{1}{m}}=x^{\frac{1}{2}k}=x^{a}$
$\phi(m)$
drie horizontaalschalen
$m=\frac{2}{k}=\frac{1}{a}$
$k=\frac{2}{m}=2a$
$\frac{1}{2}k=\frac{1}{m}=a$
m
m → 0 1 2 3 4 5
k=2/m → ∞ 2 1 2/3 2/4=1/2 2/5
a=½k → ∞ 1 ½ 1/3 ¼ 1/5
z → x^∞ x √x ∛x ∜x ⁵√x

fig. 1

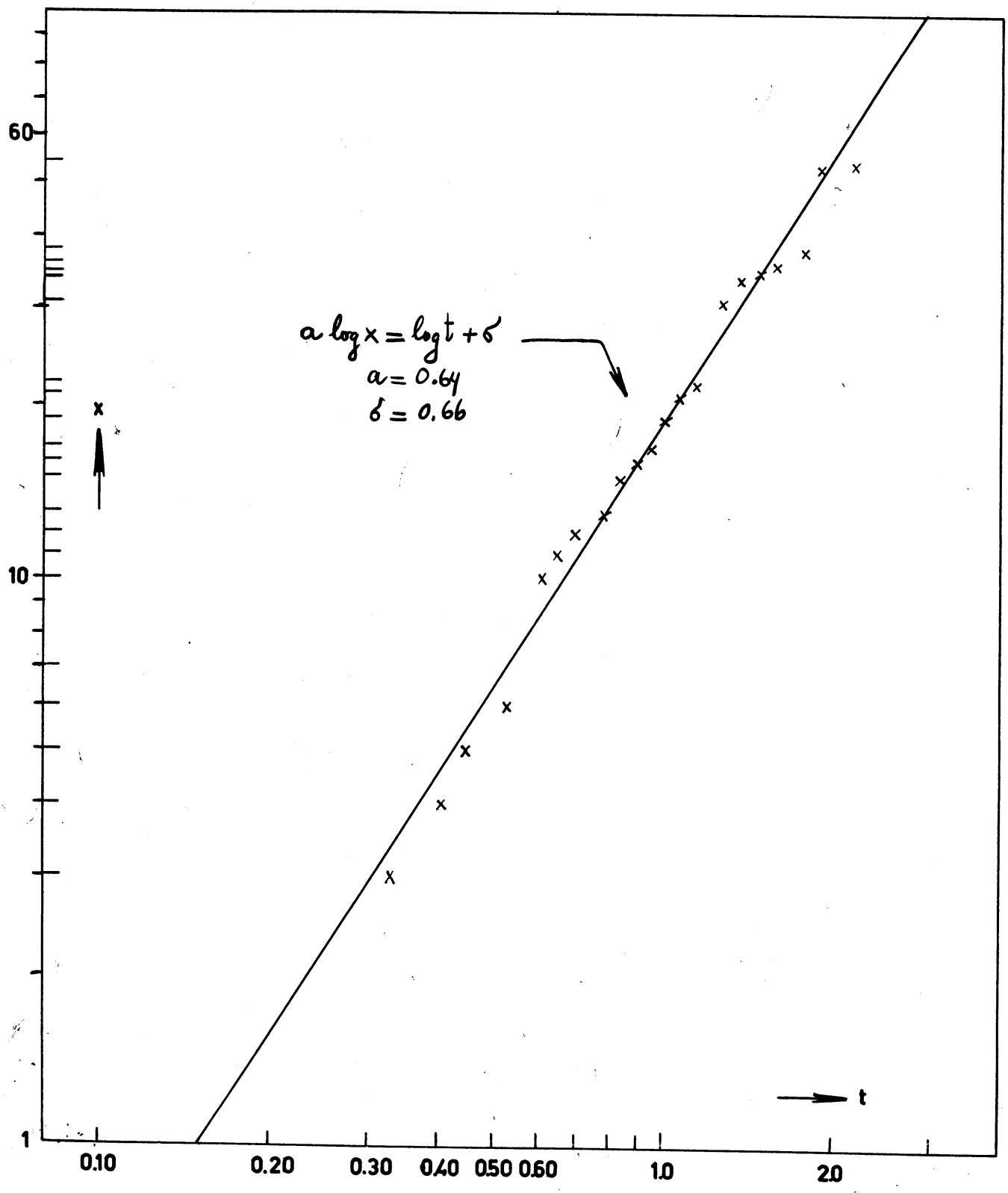

60
10
1
x
$a \log x = \log t + \delta$
$a = 0.64$
$\delta = 0.66$
t
0.10
0.20
0.30
0.40
0.50
0.60
1.0
2.0

Fig.2