2 2 MAT 1957

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. 56-005 (III-187)

Dr. C. Levert

Beschouwingen over betrouwbaarheidsintervallen en een uitschieterscriterium

De Bilt, 1956.

Kon. Ned. Meteor. Inst.
De Bilt

All Rights Reserved.

Nadruk zonder toestemming van het K.N.M.I. is verboden.

KONINKLIJK NEDERLANDS
METEOROLOGISCH INSTITUUT

Wetenschappelijk Rapport
W.R. 56-005 (III-187)
U.D.C.: 519.2

Dr. C. Levert

Beschouwing over betrouwbaarheidsintervallen en een uitschieterscriterium

INHOUD

0. Inleiding

De aanleiding tot de studie, beschreven in dit rapport, is onder meer gelegen in het volgende. Sedert november 1951 bestaat een Werkgroep Regenwaarnemingen van de Commissie voor Hydrologisch Onderzoek T.N.O., waarin uiteenlopende, in regengegevens belangstellende, groepen zitting hebben en waarvan schrijver van dit rapport secretaris is. Na uitvoerig overleg is een werkprogramma vastgesteld waarbij aan het K.N.M.I. gevraagd wordt hulp te verlenen bij het tabellarisch verzamelen van de meest uiteenlopende neerslaggegevens (1-, 2-, 3- ..... daagse sommen; hoeveelheden in tijdvakken van enige uren; regens, naar frequentie, hoeveelheid, duur, intensiteit enz.). Bij de publikatie der tabellen en der op deze gebaseerde formules en eventueel grafische voorstellingen (wellicht ook nomogrammen) wordt gevraagd speciale aandacht te besteden aan de volgende drie punten: a) de hanteerbaarheid, b) de betrouwbaarheid, c) de nauwkeurigheid.

In dit rapport wordt het tweede punt behandeld. Een voorbeeld: Wanneer wij op grond van al het aanwezige materiaal tot de conclusie komen, dat voor het station S een juli-maandsom neerslag van h = 200 of meer mm verwacht mag worden 1 keer in gemiddeld j = 75 jaren (geheel gefantaseerde getallen), dan weet een ieder, dat men dit aantal 75 niet letterlijk nemen moet. Het getal j kan ook wel 70, 80, 60 of 90 enz. zijn, m.a.w. het getal 75 heeft een zekere (on)betrouwbaarheid, die natuurlijk met de lengte van de basisreeks en de grootte van h samenhangt. Hoe is die samenhang?

Zo worden wij geleid tot het volgende type vraag, dat in algemene bewoordingen aldus luidt:

Men neemt een aselecte steekproef van N "elementen" uit een universum met onbekende verdeling $\varphi(x)dx$ of met een verdeling van een bekend type, doch met parameters met onbekende getallenwaarden, $\varphi(x;\alpha,\beta\ldots)dx$ welke steekproef een gemiddelde $\bar{x}$ en een standaarddeviatie s levert. Verder is zekere waarde W van de variabele x gegeven. We vragen: welke is de overschrijdingskans $\int_W^\infty \varphi(x)dx$ van deze W in het universum indien de "enige" informatie door de steekproef gegeven wordt.

In het bijzonder zal men W òf zeer groot (zelfs groter dan de grootste der N metingen of waarnemingen) òf zeer klein (bij voorkeur kleiner dan de kleinste der N metingen) kiezen.

Het blijkt dan mogelijk tevens een ander actueel probleem aan te pakken. Dit formuleren wij algemeen als volgt:

In een aselecte steekproef van N elementen, genomen uit een universum met onbekende verdeling (of bekend type verdeling met onbekende parameters) komt ons het grootste of (en) het kleinste "ver-

dacht groot resp. klein" voor, d.w.z. wij vermoeden, dat het niet behoort tot het universum, waartoe de overige tezamen behoren, zodat wij voor de vraag staan of het "beter" zou zijn deze verdachte meting (en) of waarneming(en) te verwijderen. Wanneer voor het afwijkende gedrag geen oorzaken te vinden zijn, zou men bereid kunnen zijn de beslissing over het al of niet verwijderen van de verdachte meting aan een statistisch criterium over te laten. De vraag is dan: bestaat er een objectief criterium? Natuurlijk moeten wij de begrippen "groot", "klein", "beter" scherp definiëren en zal onze uitspraak nimmer een zekerheid van 100% hebben, doch met een (aangeefbare) kleine kans fout zijn. Wij komen hierop uitvoerig terug.

1. De "niet-centrale t-verdeling" [1] [2] [3]

1.1 Formulering van het probleem

Wij formuleren het volgende probleem: Gegeven: een variabele $\underline{x}$, die aan een normale kansverdeling [1], met gemiddelde μ en variantie σ^2 gehoorzaamt; μ en σ zijn bekend ondersteld. Wij nemen herhaaldelijk aselecte steekproeven van N elementen uit deze populatie. Elke daarvan levert een gemiddelde $\bar{\underline{x}} = \sum_1^N x_i/N$ en variantie $\underline{s}^2 = \sum_1^N (x_i - \bar{x})^2 / (N-1)$. Gegeven is verder een waarde W, zodat we kunnen berekenen $\underline{k} = (W - \bar{\underline{x}}) : \underline{s}$ in elk der steekproeven. Aangezien $\bar{\underline{x}}$ en $\underline{s}$ stochastische variabelen [2] zijn (daarom onderstreept), is $\underline{k}$ het ook. Gevraagd de kansverdeling $\varphi(\underline{k})\,d\underline{k}$. Opl. De steekproefverdeling van deze variabele $\underline{k} = (W - \bar{x}) : \underline{s}$ is vrij gecompliceerd. De $\underline{k}$ is verdeeld als de z.g. "niet centrale t", gedefinieerd door $\underline{n.c.t} = (W - \bar{\underline{x}}) : \underline{s} \sqrt{N}$. Deze kansverdeling gaat in de z.g. "centrale t-verdeling", d.i. de "Student-verdeling", over als we voor de vaste waarde W het universumgemiddelde μ zetten: $\underline{t} = \frac{\mu - \bar{\underline{x}}}{\underline{s}/\sqrt{N}}$.
Zoals bekend is de t-verdeling a.v.:

$$(1)\quad \varphi(t)\,dt = \frac{(\frac{1}{2}N-1)!}{\sqrt{\pi}\,\{\frac{1}{2}(N-3)\}!}\,(1+t^2)^{-\frac{1}{2}N}\,dt \qquad \text{met } \int_{-\infty}^{\infty}\varphi(t)\,dt = 1 \quad \text{en } \varphi(t) = \varphi(-t)$$

$$\mathcal{E}(t) = 0$$

---

1) In formule $f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}}\exp.-\frac{(x-\mu)^2}{2\sigma^2}\,dx$, met $\int_{-\infty}^{\infty} f(x)\,dx = 1$; $\mu \equiv \mathcal{E}(x) = \int_{-\infty}^{\infty} x f(x)\,dx$ en $\sigma^2 \equiv \mathcal{V}(x) = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,dx$. Kortheidshalve: $\mathcal{N}(\mu,\sigma)$

2) Een variabele heet stochastisch als hij een waarschijnlijkheidsverdeling bezit. Een stochastische variabele wordt onderstreept. Een getallenwaarde ervan is niet onderstreept.

Zelfs met de tabellen voor de "niet-centrale t-verdeling", die door Johnson en Welch in Biometrika 31 362 1940 gepubliceerd werden, zijn de berekeningen verre van gemakkelijk. De tabel heeft drie ingangen t.w. N, W en k_α, waarmee de kans α op een $\underline{k} \geqslant k_\alpha$ berekend kan worden. [1]

Er bestaat een eenvoudiger weg, waarlangs een benadering van α berekend kan worden als k_α gegeven is. Dit geschiedt a.v.
Wij zoeken in ons probleem derhalve $\alpha \equiv \mathbb{P}[\underline{k} \geqslant k_\alpha] \equiv \int_{k_\alpha}^{\infty} \varphi(\underline{k})\, d\underline{k}$

(2) bij gegeven k_α, μ, σ en W.
Voer een nieuwe variabele $\underline{z} = \bar{\underline{x}} + m\,\underline{s}$ in, waarin m een vastgehouden getal voorstelt. Ook deze $\underline{z}$ volgt een kansverdeling, waarvan ons het gemiddelde $\mathcal{E}(\underline{z})$ en de variantie $\mathcal{V}(\underline{z}) \equiv \sigma^2(\underline{z})$ interesseren.

(3) Zeker is $\mathcal{E}(\underline{z}) = \mathcal{E}(\bar{\underline{x}}) + m\,\mathcal{E}(\underline{s})$. Verder is $\mathcal{E}(\bar{\underline{x}}) = \mu$ en $\mathcal{E}(s) = p_1\sigma$, waarin

(4) $p_1 = \sqrt{\frac{2}{N-1}}\left(\frac{1}{2}N-1\right)!\,/\left(\frac{1}{2}N-\frac{3}{2}\right)! < 1$ (dit onderstellen wij bekend)

(5) Verder geldt $\mathcal{V}(\underline{z}) = \mathcal{V}(\bar{\underline{x}}) + m^2\,\mathcal{V}(\underline{s})$; de term met $\mathcal{V}(\bar{\underline{x}}).\,\mathcal{V}(\underline{s})$ ontbreekt omdat $\bar{\underline{x}}$ en $\underline{s}$ onafhankelijk verdeeld zijn, hetgeen alleen het geval is als wij aselecte steekproeven uit een normaal verdeeld universum nemen.

(6) Voorts is $\mathcal{V}(\bar{\underline{x}}) \equiv \sigma^2(\bar{\underline{x}}) = \sigma^2/N$ en $\mathcal{V}(\underline{s}) = p_2^2\sigma^2/2(N-1)$,

(7) waarin $p_2^2 = 2(N-1)(1-p_1^2) < 1$ (dit onderstellen wij bekend). Voor toenemende N naderen p_1 en p_2 tot 1. Reeds voor N = 5 is $p_1 = 0.94$ en $p_2 = 0.96$.
Aldus is

(8) $\mathcal{E}(\underline{z}) \equiv \mu_z = \mu + m\sigma$ $\qquad \sqrt{\mathcal{V}(\underline{z})} \equiv \sigma_z \simeq \sigma\sqrt{\frac{1}{N} + \frac{m^2}{2(N-1)}} \simeq \frac{\sigma}{\sqrt{N}}\sqrt{1+\frac{1}{2}m^2}$ als $N \gg 1$

Iedere andere getallenwaarde van m bepaalt een andere $\underline{z}$-verdeling. Als lineaire combinatie van $\bar{\underline{x}}$ en $\underline{s}$ zal de verdeling van $\underline{z}$ des te minder van de normale verdeling, gebaseerd op de boven berekende μ_z en σ_z, afwijken, naarmate N groter is, en dit ofschoon van de twee variabelen $\bar{\underline{x}}$ en $\underline{s}$ alleen de eerste exact normaal verdeeld is, n.l. als $\mathcal{N}(\mu\,;\,\sigma/\sqrt{N})$. Dit geldt vanwege de "Centrale Limiet Stelling", zie addendum. Is $\underline{z}$ zo goed als normaal

---

1) We zullen voortaan in iedere kansverdeling $g(\underline{w})d\underline{w}$ de waarde van $\underline{w}$, die met een kans α overschreden wordt, met $\underline{w}_\alpha$ aangeven, d.i. $\int_{\underline{w}_\alpha}^{\infty} g(\underline{w})\, d\underline{w} \equiv \alpha$

verdeeld, d.i. volgens $\mathcal{N}(\mu_z ; \sigma_z)$, dan kunnen wij standaardiseren tot
(9) $\mathcal{N}(0; 1)$ door op een nieuwe variabele $\underline{\tau} = (\underline{z} - \mu_z) : \sigma_z$ 1)
over te gaan. [Welke gestandaardiseerde normale verdeling, in cumulatieve vorm, uitvoerig getabelleerd is].

Denk nu aan een getallenwaarde voor k_α en vraag naar de (overschrijdings)kans α, waarmee $\underline{k}$ deze k_α overschrijdt. [W is nog altijd
(10) (11) gegeven gedacht]. Uit $(W - \bar{\underline{x}}) : \underline{s} \geqslant k_\alpha$ volgt direct $\bar{\underline{x}} + k_\alpha \, \underline{s} \leqslant W$. Geef vervolgens aan m dezelfde getallenwaarde als aan k_α. Wij bestuderen dus de kansverdeling van $\underline{z} = \bar{\underline{x}} + k_\alpha \cdot \underline{s}$, welke zo goed als normaal is rondom een gemiddelde μ_z en met een standaarddeviatie σ_z, die men verkrijgt door in (8) aan m de waarde k_α te geven (μ en σ zijn gegeven).

De kans $\mathbb{P}[\underline{k} \geqslant k_\alpha \,|\, \mathfrak{k}]$ is dezelfde als de kans $\mathbb{P}[\underline{z} \leqslant W \,|\, \mathfrak{w}]$.

N.B. $\ldots . | \mathfrak{k}$ betekent: onder de voorwaarde dat de waarde van $\mathfrak{k}$, welke $\mathfrak{k}$ gedefinieerd is als $(W - \mu) : \sigma$, gegeven is. Het blijkt n.l., dat bij gegeven k_α en N de waarde van α voor de waarden W_1, μ_1, σ_1, dezelfde is als voor een ander stel, W_2, μ_2, σ_2, mits $(W_1 - \mu_1) : \sigma_1 = (W_2 - \mu_2) : \sigma_2$, dus mits in beide gevallen de $\mathfrak{k}$ dezelfde waarde heeft. [we geven aan dit quotiënt de naam $\mathfrak{k}$, d.i. de Griekse letter voor k, omdat wat k voor de steekproef is, de $\mathfrak{k}$ voor het universum is]
Symbolisch:[2]

$$(12)\quad \alpha \equiv \mathbb{P}[\underline{k} \geqslant k_\alpha \,|\, \mathfrak{k}] = \mathbb{P}[\underline{z} \leqslant W \,|\, \mathfrak{k}] = 1 - \mathbb{P}[\underline{z} > W \,|\, \mathfrak{k}] = 1 - \mathbb{P}[\underline{\tau} > \tau_{1-\alpha}]$$

met $\tau_{1-\alpha} = (W - \mu_z) : \sigma_z$

Toelichting: de kans α op een overschrijding van een gegeven k_α in de verdeling van $\underline{k}$ is tevens de kans, waarmede $\underline{z}$ de gegeven W in de verdeling van $\underline{z}$ (met $m = k_\alpha$) onderschrijdt. Maar door de verdeling van $\underline{z}$ te

---

1) De gestandaardiseerde normaal verdeelde variabele $\underline{\tau}$ heeft een kansverdeling $\varphi(\underline{\tau})\,d\underline{\tau} = \frac{1}{\sqrt{2\pi}} e^{-\underline{\tau}^2/2} d\underline{\tau}$ met $\int_{-\infty}^{\infty} \varphi(\underline{\tau})\,d\underline{\tau} = 1 \,;\, \mathcal{E}\underline{\tau} = 0 \,;\, \mathcal{V}(\underline{\tau}) = 1$
In de meeste boeken wordt ook deze variabele met een t aangeduid. Omdat dit verwarrend werkt, aangezien de t reeds voor de z.g. "Student-t", d.i. $(\mu - \bar{\underline{x}}) : \underline{s}/\sqrt{N}$, gebruikt werd, kozen wij de letter $\underline{\tau}$.

2) In verdelingen f(g)dg van continue variabelen heeft het geen zin om zich druk te maken over de vraag of $\leqslant$ tot wat anders leidt dan $<$. Immers de kans, dat g de voorgeschreven waarde g_0 aanneemt is altijd nul, zodat men voor $\int_{-\infty}^{g_0} f(g)\,dg$ zowel $\mathbb{P}[g \leqslant g_0]$ als $\mathbb{P}[g < g_0]$ schrijven kan.

standaardiseren (zie 9) is de laatste kans tevens het complement van de kans, dat, in de gestandaardiseerde verdeling, de $\underline{\tau}$ de waarde $\tau_{1-\alpha} = (W - \mu_2) : \sigma_2$ overschrijdt. Omdat die kans $1 - \alpha$ is, hechten we de index $1 - \alpha$ aan de τ. Aldus blijkt: geeft men k_α (en natuurlijk zijn μ, σ W, N bekend) dan is α te berekenen, via $\tau_{1-\alpha} = (W - \mu_2) : \sigma_2$ met $\mu_2 = \mu + k_\alpha \sigma$ en $\sigma_2 = \sigma\sqrt{\frac{1}{N} + \frac{k_\alpha^2}{2(N-1)}}$ en de tabel van de gestandaardiseerde normale verdeling.

Korter: $$\alpha \equiv \mathbb{P}\left[\underline{k} \geq k_\alpha \,\middle|\, \kappa = \frac{W-\mu}{\sigma}\right] = 1 - \mathbb{P}\left[\tau > \frac{\kappa - k_\alpha}{\sqrt{\frac{1}{N} + \frac{k_\alpha^2}{2(N-1)}}}\right]$$

met $$\tau_{1-\alpha} = \frac{\kappa - k_\alpha}{\sqrt{\frac{1}{N} + \frac{k_\alpha^2}{2(N-1)}}}$$

Men ziet, dat α dezelfde waarde heeft, voor gegeven N en k_α, voor al die combinaties μ, σ, W, die dezelfde κ leveren, zodat men zeggen kan, dat α alleen van κ afhangt.

Een en ander hebben wij verduidelijkt in de volgende schetsjes; I heeft betrekking op een grote α, t.w. 0.95 en II op een kleine α, t.w. 0.05. Beide schetsjes zijn slechts illustratief.

1.2 Numeriek voorbeeld

Wij nemen uit een normaal verdeeld universum $\mathcal{N}(6; 2\frac{1}{2})$ herhaaldelijk aselecte steekproeven van N = 20 elementen en bestuderen, voor W = 10, de kansverdeling $\varphi(\underline{k})\,d\underline{k}$ van $\underline{k} = (W - \underline{\bar{x}}) : \underline{s} = (10 - \underline{\bar{x}}) : \underline{s}$. De $\kappa = (W - \mu) : \sigma = (10 - 6) : 2\frac{1}{2} = 1.6$. Wij zagen hoe wij van iedere gegeven waarde k_α de overschrijdingskans α in de $\underline{k}$-verdeling kunnen berekenen. Beschouw bijv. $k_\alpha = 3$. Hierbij is $(W - \mu_2) : \sigma_2 = [W - (\mu + 3\sigma)] : \sigma\sqrt{\frac{1}{N} + \frac{3^2}{2(N-1)}} = (\kappa - 3) : \sqrt{0.05 + 0.026 \cdot 3^2} = (1.6 - 3) : \sqrt{0.05 + 0.026 \cdot 3^2} = -2.60$

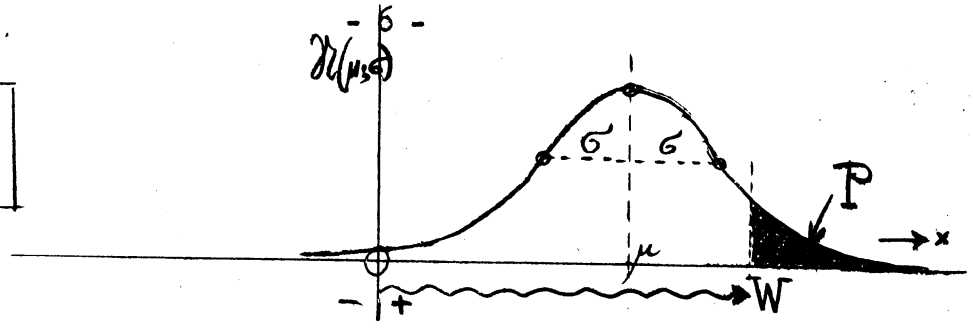
Bijgevolg vinden wij $\alpha \equiv \mathbb{P}[\underline{k} \geq 3 \,|\, \kappa = 1.6] = 1 - \mathbb{P}[\underline{\tau} > -2.60] = 1 - 0.9953 = 0.0047$.

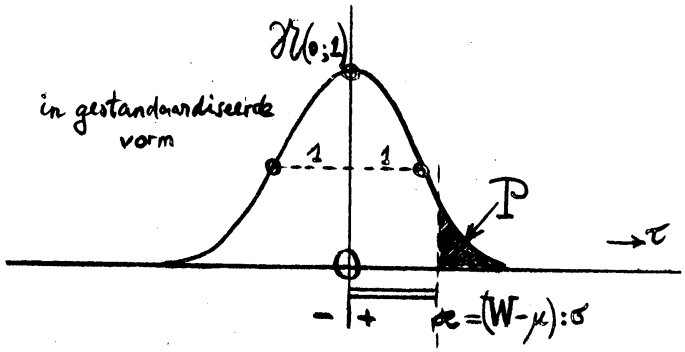
Ander voorbeeld $k_\alpha = 1 \rightarrow (W - \mu_2) : \sigma_2 = 2.20$ en $\alpha \equiv \mathbb{P}[\underline{k} \geq 1 \,|\, \kappa = 1.6] = 1 - \mathbb{P}[\underline{\tau} > 2.20]$ $= 1 - 0.016 = 0.984$. Enz.

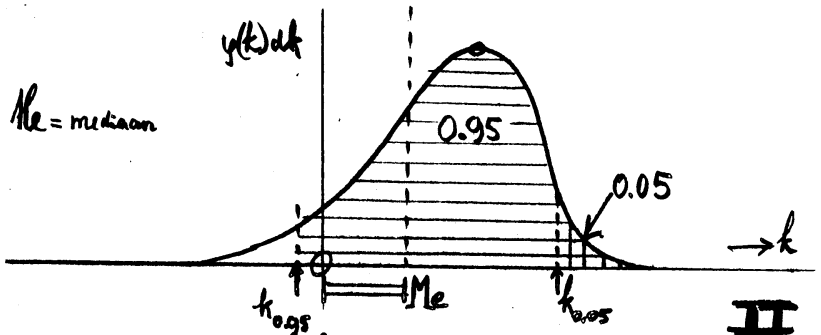
Zie het volgende tabelletje.

gegeven
μ , σ , W

$\mathcal{N}(\mu;\sigma)$ = norm. verd. op gem. μ en st.dev. σ



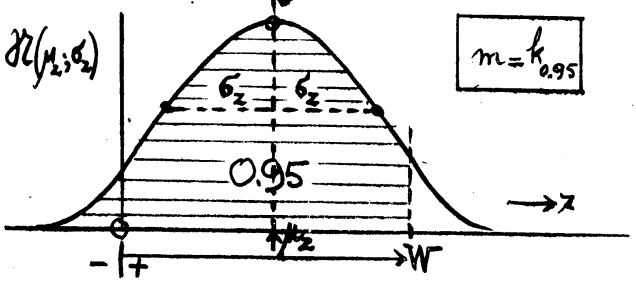


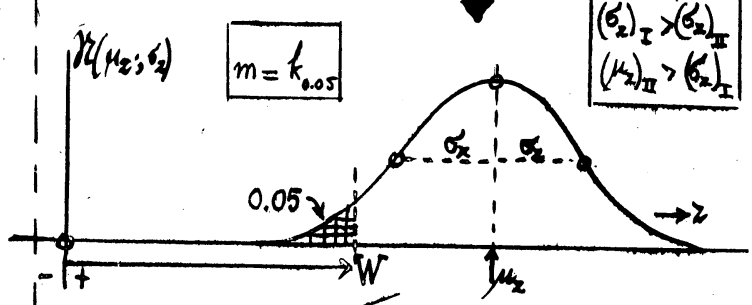


I

II

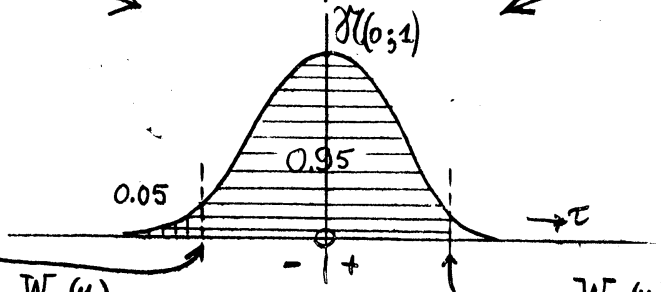
$k_{0.05} > k_{0.95}$ dus $(\sigma_z)_I > (\sigma_z)_{II}$ $(\mu_z)_{II} > (\sigma_z)_I$





$\alpha = 0.95$

$\alpha = 0.05$



$\tau_{1-\alpha} = \tau_{0.95} = \dfrac{W-(\mu_z)_{II}}{(\sigma_z)_{II}}$

$\tau_\alpha = \tau_{0.05} = \dfrac{W-(\mu_z)_I}{(\sigma_z)_I}$

met $(\mu_z)_{I,II} = \mu + k_{I,II}\,\sigma$ en $(\sigma_z)_{I,II} = \sigma\sqrt{\dfrac{1}{N} + \dfrac{k^2_{I,II}}{2(N-1)}}$ als I $\equiv \alpha = 0.95$ en II $\equiv \alpha = 0.05$

Dus geldt:

$$(\tau - k_{0.95}) : \sqrt{\frac{1}{N} + \frac{k^2_{0.95}}{2(N-1)}} = -(\tau - k_{0.05}) : \sqrt{\frac{1}{N} + \frac{k^2_{0.05}}{2(N-1)}}$$

; ook voor elke andere α

het verband tussen α en k_α bij gegeven μ, σ, W, N

$\mu = 6$; $\sigma = 2\frac{1}{2}$; N = 20; W = 10; $\alpha = 1.6$				
k_α	α	μ_z	σ_z	$(W-\mu_z):\sigma_z$
3	0.0047	13.5	1.345	-2.61
→ 2.31	0.05	11.8	1.09	-1.645 ←
mediaan → 1.6	0.50	10.0	0.856	0
→ 1.124	0.95	8.81	0.722	1.645 ←
1.000	0.934	8.50	0.690	2.18
0.5	0.99996	7.25	0.595	4.12
0	≅ 1	6.00	0.560	7.14
-1	≅ 1	3.50	0.690	9.4

In het bijzonder wijzen wij op de mediaan Me, die de waarde $\alpha = 1.6$ heeft, omdat hiervoor de waarde van $(W-\mu_z):\sigma_z = (\alpha - k_\alpha):\sqrt{\frac{1}{N} + \frac{k_\alpha^2}{2(N-1)}}$ nul moet zijn, want alleen dan is $\mathbb{P}[\underline{t} > (W-\mu_z):\sigma_z] = \frac{1}{2}$
De $\underline{k}$-verdeling heeft derhalve ongeacht de waarde van N de mediaan Me in $k = \alpha = (W-\mu):\sigma$, in het getallenvoorbeeld : 1.6. Natuurlijk interesseren ons ook de 0.05- en 0.95 punten. Hierbij moet $(W-\mu_z):\sigma_z$ de waarde -1.645 resp. +1.645 hebben, want $\mathbb{P}[\underline{t} > (W-\mu_z):\sigma_z]$ moet 0.95 resp. 0.05 zijn. Uit deze -1.645 resp. + 1.645 volgen weer de waarden van k_α t.w. 2.31 resp. 1.124, n.l. uit

$$\mp 1.645 = \frac{1.6 - k_\alpha}{\sqrt{\frac{1}{20} + \frac{k_\alpha^2}{2\times 19}}}$$

Voorts willen wij nog op een bijzonderheid, een voor de praktijk onbelangrijke finesse, wijzen, die wij in de geciteerde leerboeken niet behandeld vonden. Wij vragen ons af of inderdaad $\mathbb{P}[\underline{t} \geq k_\alpha]$ naar 1 resp. 0 gaat, voor $k_\alpha \to -\infty$ resp. $+\infty$, hetgeen toch behoort voor een kansverdeling. Natuurlijk is het stellen van k_α gelijk - of $+\infty$ niet zonder zin, want het is niet mogelijk N gelijke metingen te doen, zodat $\bar{x} = x_1 = x_2 = \cdots x_N$ en s = 0, waardoor $k = \frac{10 - \bar{x}}{0}$ d.i. òf $-\infty$ (als $\bar{x} > 10$) òf $+\infty$ (als $\bar{x} < 10$). Welke waarde α heeft bij $k_\alpha = +\infty$ of $-\infty$ hangt samen met de waarde, die het quotiënt

$$(W-\mu_z):\sigma_z = (\alpha - k_\alpha):\sqrt{\frac{1}{N} + \frac{k_\alpha^2}{2(N-1)}}$$

dan aanneemt. Wij moeten nu terug tot de uitdrukkingen (3), (4), (6) en (7). Daar zagen wij, dat, exact beschouwd,

$$\mu_z \equiv \mathcal{E}(\underline{z}) = \mu + \sigma \cdot p_1(N) \quad \text{en} \quad \mathcal{V}(\underline{z}) \equiv \sigma_z^2 = \sigma^2\left[\frac{1}{N} + \frac{p_2^2(N)}{2(N-1)}\right]$$

waarin p_1 en p_2 voor iedere N kleiner dan 1 zijn, doch onbeperkt tot 1 naderen als N toeneemt. Voorts moeten wij bedenken, dat voor iedere waarde van N de exacte verdeling van $\underline{z}$ verschilt van de op bovenstaande μ_z en σ_z

gebaseerde normale verdeling. Wel is dit verschil praktisch van geen betekenis en wordt het minder met toenemende N. We bemerken dit aldus. Voor grote $|k_\alpha|$ lijkt men voor $(W - \mu_z) : \sigma_z$ te mogen schrijven (p_1 en p_2 gelijk 1 stellende) :

$$\frac{W-(\mu+k_\alpha\sigma)}{\sigma\sqrt{\frac{1}{N}+\frac{k_\alpha^2}{2(N-1)}}} \simeq \frac{t-k_\alpha}{k_\alpha}\sqrt{2(N-1)}$$

zodat men zou besluiten, dat dit quotiënt met $k_\alpha \to -\infty$, resp. $\to +\infty$ naar $\sqrt{2(N-1)}$ resp. $-\sqrt{2(N-1)}$ lijkt te gaan (let wel niet naar $+\infty$ resp. $-\infty$). In ons geval is dat ($N = 20$) + 6.16 resp. -6.16, waarbij $\mathbb{P}[\tau > 6.16] = 0{,}999\ 999\ 9984$. Zo lijkt derhalve $\mathbb{P}[t \geq k_\alpha]$ voor $k_\alpha \to +\infty$, resp. $-\infty$, te gaan naar 0,000 000 0016 resp. 0.999 999 9984 = 1 - 0.000 000 0016, i.p.v. naar 0 resp. 1. Hier wreekt zich derhalve de consequentie van de genoemde benadering, doch het effect is reeds voor N = 20 niet noemenswaardig (en wordt kleiner voor grotere N).

1.3 Waarom geëist wordt dat de verdeling van x normaal is.

Wij hebben ondersteld, dat de initiaal-variabele x aan een normale verdeling $\mathcal{N}(\mu;\sigma)$ gehoorzaamt, omdat wij willen aansturen op een nieuwe variabele z, welks verdeling weinig van een normale verschilt, met niet te moeilijk te berekenen gemiddelde μ_z en standaarddeviatie σ_z. Als x niet normaal verdeeld is, is de standaarddeviatie van de verdeling van s in steekproeven van N stuks moeilijk te berekenen (deze zal dan ook de hogere momenten van de initiaalverdeling bevatten). Doch dit niet alleen; tevens zullen dan $\bar{x}$ en s niet onafhankelijk verdeeld zijn, zodat bij de berekening van σ_z ook deze afhankelijkheid mee zal spreken. Dit zijn in zekere zin louter statistisch-mathematische moeilijkheden. Indien zij kunnen worden opgelost, zouden alle volgende beschouwingen, lijkt ons, tot niet-normale verdelingen kunnen worden uitgebreid, hetgeen wel zeer gewenst is, aangezien wij onder meer het oog hebben op de variabele x, die de 1, 2, 3 ... daagse neerslagsom voorstelt, welke zeer scheef verdeeld is. Men kan natuurlijk ook proberen een transformatie $y = f(x)$ toe te passen, zódanig, dat de verdeling van y wel een (bijna) normale is. Vermoedelijk is dit een verschuiving van de moeilijkheid. Hier valt nog veel te doen.

2. Begrensde marges voor onbekende overschrijdingskansen; betrouwbaarheidsintervallen

2.1 Wijziging van het probleem onder 1.1

Gegeven een normale verdeling $\mathcal{N}(\mu;\sigma)$, maar nu met onbekende μ en σ. W is een gegeven waarde. Wij willen weten welke de overschrijdingskans P is van deze W in $\mathcal{N}(\mu;\sigma)$, terwijl de $\bar{x}$ en s in de aselecte steekproef van N elementen de beste informatie omtrent de μ en σ vormen van het universum. Doordat μ en σ onbekend zijn, is eigenlijk de exacte P ook niet bekend. Wel kan men van $P = \mathbb{P}[x \geq W \mid \mu;\sigma]$

(14) de volgende schatting maken: $\boxed{P_s = \mathbb{P}[x \geq W \mid \bar{x};s]}$

N.B. Met ..... | (a; b) wordt bedoeld: in de normale verdeling met gemiddelde a en standaarddeviatie b.

(15) Standaardiserende geldt ook $P_s = \mathbb{P}[x \geq W \mid \bar{x};s] = \mathbb{P}[\tau \geq k \mid 0;1]$ indien τ de gestandaardiseerde normale verdeling $\mathcal{N}(0;1)$ volgt en $k = (W - \bar{x}) : s$.

Deze P_s is de beste "puntschatting" van P.

Wij kunnen de vraag stellen: welke betrouwbaarheid heeft deze schatting? Het is immers direct duidelijk, dat P niet juist deze waarde P_s zou hebben en beslist geen andere. Een tweede steekproef uit hetzelfde universum zou natuurlijk een andere $\bar{x}$ en s geleverd hebben en daarmede een andere P_s. Bij gegeven W, volgt P_s zelf blijkbaar ook een kansverdeling. Wij vragen naar een marge van "toelaatbare" P_s-waarden, waarbij het begrip "toelaatbaar" nog gedefinieerd moet worden. Voor wij dit doen, eerst een numerieke toelichting:

Stel men doet N = 20 onafhankelijke metingen en men weet, dat de meetresultaten x_1, x_2 ..... x_{20} normaal verdeeld zijn, doch van de verdeling zijn μ en σ onbekend. Men berekent $\bar{x} = 6$ en $s = 2\frac{1}{2}$. Wij willen de overschrijdingskans P weten van de gegeven waarde W = 10 in het universum. De beste puntschatting is $P_s = \mathbb{P}[x \geq 10 \mid 6;2\frac{1}{2}] = \mathbb{P}[\tau \geq \frac{10-6}{2\frac{1}{2}}] = 0.055$

Aangezien μ en σ onbekend zijn, is ook $\mathbb{P}[\tau \geq \frac{10-\mu}{\sigma}] = P$ onbekend.

Wij redeneren nu a.v.: eigenlijk is geen enkele P onmogelijk, aangezien geen enkele μ, σ-combinatie (waarmede een normale verdeling is vastgelegd) onmogelijk is. Dat wil tevens zeggen dat geen enkele normale verdeling het meten van een k = (10 - 6) : $2\frac{1}{2}$ = 1.6 onmogelijk maakt. Wel is de kans op het meten van zulk een k niet even waarschijnlijk. Daarom: als een universum $\mathcal{N}(\mu;\sigma)$ zou leiden tot een k-verdeling, waarin óf de kans op een $k \leq 1.6$ óf de kans op een $k \geq 1.6$ zeer klein zou zijn [1] [2] (kleiner

1) men kan in de bij deze normale verdeling behorende k-verdeling met 1.6 in de linker-, doch ook in de rechter-staart verkeren.

2) zie noot 2 blz. 5.

dan of hoogstens gelijk aan een bij afspraak te geven drempel, bijv. 0.05), dan achten wij die μ_0, σ_0 combinatie "onmogelijk" (ontoelaatbaar") en daarmede ook de door deze μ_0, σ_0 combinatie en W = 10 bepaalde waarde van P.

De vraag is: hoe berekenen wij de verzameling der toelaatbare (of ontoelaatbare) P's?

a) Onderzoek bijv. of P = 0.123 (opzettelijk boven 0.055 gekozen) toelaatbaar is; $\mathbb{P}[\underline{x} \geq 10 \mid \mu ; \sigma] = 0.123$. Hierdoor is een verzameling normale verdelingen vastgelegd, n.l. met een μ en σ, verbonden door $(10 - \mu) : \sigma = \tau_{0.123} = 1.162$, als $\tau_{0.123}$ de getallenwaarde van τ is, die in de gestandaardiseerde normale verdeling $\mathcal{N}(0; 1)$ met een kans 0.123 overschreden wordt. Hoewel dit verschillende normale verdelingen zijn (ofschoon μ en σ door de gegeven relatie lineair verbonden zijn) leveren ze alle dezelfde k-verdeling en wel omdat blijkt, dat de overschrijdingskans α van een iedere te geven k_α slechts van k_α afhang. Immers:

$$\alpha \equiv \mathbb{P}[\underline{k} \geq k_\alpha] = 1 - \mathbb{P}[\underline{z} \geq 10] = 1 - \mathbb{P}\left[\underline{\tau} > \frac{10-\mu_z}{\sigma_z}\right]$$

met $$\frac{10-\mu_z}{\sigma_z} = \frac{10-(\mu+k_\alpha\sigma)}{\sigma\sqrt{\frac{1}{20}+\frac{k_\alpha^2}{38}}} = \frac{\tau_{0.123}-k_\alpha}{\sqrt{\frac{1}{20}+\frac{k_\alpha^2}{38}}} = \frac{1.162-k_\alpha}{\sqrt{0.05+0.026\,k_\alpha^2}}$$

Bij iedere k_α behoort derhalve één α. In het bijzonder vullen wij de gemeten k = 1.6 in. Er komt $\mathbb{P}[\underline{k} \geq 1.6] = 1 - \mathbb{P}[\underline{\tau} > 1.28] = 1 - 0.90 = 0.10$. Aangezien $0.10 > 0.05$ (drempel) heet deze P = 0.123 "toelaatbaar".

Als wij een grotere P invullen bijv. 0.211, komt er $\tau_{0.211} = 0.804$ en

$$\frac{\tau_{0.211}-1.6}{\sqrt{0.05+0.026\cdot 1.6^2}} = -2.33 \qquad \text{en } \mathbb{P}[\underline{k} > 1.6] = 1 - \mathbb{P}[\underline{\tau} > -2.33] = 1 - 0.99 = 0.01$$

Thans $0.01 < 0.05$, dus P = 0.211 is "ontoelaatbaar". Blijkbaar kunnen we een waarde voor P berekenen, die normale verdelingen fixeert, bij welke k-verdelingen behoren, in elk waarvan de kans op een overschrijding van k = 1.6 precies 0.05 is. Wij vinden deze a.v.: $1 - \mathbb{P}[\underline{\tau} > -1.64] = 1 - 0.95 = 0.05$. Deze -1.64 kunnen wij met $-\tau_{0.05}$ aangeven, als wij met $\tau_{0.05}$ die waarde van τ bedoelen, die in $\mathcal{N}(0; 1)$ met een kans 0.05 overschreden wordt. Dus

$$-1.64 = \frac{10-\mu_z}{\sigma_z} = \frac{\left(\frac{10-\mu}{\sigma}\right)-1.6}{\sqrt{0.05+0.026\cdot 1.6^2}},$$

waaruit $\frac{10-\mu}{\sigma} = 1.037$.

Hierdoor ligt P vast, n.l. $P = \mathbb{P}[\underline{x} \geq 10] = \mathbb{P}[\underline{\tau} \geq 1.037] = 0.152$
Wij mogen deze 0.152 het "plafond", de "bovenwaarde", noemen, $\hat{P}$, want wij hebben nu aangetoond, dat bij alle P-waarden $\geq \hat{P}$ normale verdelingen behoren, die k-verdelingen zouden leveren, waarin de gemeten $k = (W - \bar{x}) : s = (10 - 6) : 2\frac{1}{2} = 1.6$ met een kans 0.05 of minder overschreden wordt (precies 0.05 voor $\hat{P}$ zelf en dus des te meer van 0.05 verschillend naarmate men P groter dan $\hat{P}$ neemt).

Korter: Op grond van de gemeten $\bar{x} = 10$ en $s = 2\frac{1}{2}$ concluderen wij, dat bij de gegeven W = 10 in het onbekende normaal verdeelde universum een overschrijdingskans $\leq \hat{P} = 0.152$ behoort, behoudens de kans 0.05 of minder op een foutieve uitspraak. Het laatste wordt korter uitgedrukt: s. pr. 0.05 (salva probabilitate 0.05).

b) Onderzoek nu of bijv. P = 0.029 (opzettelijk beneden 0.05) toelaatbaar is. $\mathbb{P}[\underline{x} \geq 10 \mid \mu; \sigma] = 0.029$. Weer wordt hierdoor een verzameling normale verdelingen vastgelegd, alle met een μ en σ, verbonden door de relatie $(10 - \mu) : \sigma = \tau_{0.029} = 1.887$. Als onder a) is $\mathbb{P}[\underline{\tau} \geq 1.887] = 0.029$. Hoewel dit verschillende normale verdelingen zijn, leveren ze alle dezelfde k-verdeling, omdat weer blijkt, dat de onderschrijdingskans α van iedere te geven k_α slechts van k_α afhangt. Immers

$$\mathbb{P}[\underline{z} \geq 10] = \mathbb{P}\left[\underline{\tau} > \frac{10-\mu_z}{\sigma_z}\right],$$

waarin

$$\frac{10-\mu_z}{\sigma_z} = \frac{\tau_{0.029} - k_\alpha}{\sqrt{0{,}05 + 0{,}026\, k_\alpha^2}} = \frac{1{,}887 - k_\alpha}{\sqrt{0{.}05 + 0{,}026\, k_\alpha^2}}$$

Bij iedere α behoort één k_α. In het bijzonder vullen wij de gemeten k = 1.6 in. Er komt: $\mathbb{P}[\underline{k} \leq 1.6] = \mathbb{P}[\underline{\tau} > 0.84] = 0.20$.
Aangezien $0.20 > 0.05$ (de drempel) heet deze P = 0.29 "toelaatbaar".
Als wij van een kleinere P-waarde uitgaan, bijv. 0.008, dan is $\tau_{0.008} = 2.396$ zodat $(\tau_{0.008} - 1.6)/\sqrt{0.05 + 0.026 \cdot 1.6^2} = 2.33 \gg 0.84$
en $\mathbb{P}[\underline{k} \leq 1.6] = \mathbb{P}[\underline{\tau} > 2.33] = 0.01 < 0.05$ (drempel)
Derhalve heet deze P = 0.029 "ontoelaatbaar". Blijkbaar kunnen we een waarde van P berekenen, waarbij een onderschrijdingskans van k = 1.6 behoort, die precies 0.05 is. Deze wordt gevonden a.v.:
$\mathbb{P}[\underline{\tau} > 1.64] = 0.05$
zodat

$$\frac{10-\mu_z}{\sigma_z} = \left\{\frac{10-\mu}{\sigma} - 1.6\right\} : \sqrt{0{,}05 + 0{,}026 \cdot 1.6^2} = 1.64 \text{ en } \frac{10-\mu}{\sigma} = 2.163$$

Hierdoor ligt P vast, t.w. $P = \mathbb{P}[\bar{x} > 10] = \mathbb{P}[\tau \geq 2.163] = 0{,}0155$
Wij mogen deze 0.0155 de drempel (benedenwaarde) $\check{P}$ noemen. Immers boven toonden wij aan, dat bij alle P-waarden $< \check{P}$ normale verdelingen behoren, die k-verdelingen zouden leveren, waarin de gemeten k = 1.6 met een kans 0.05 of minder onderschreden wordt (precies 0.05 voor $\check{P}$ zelf en des te meer van 0.05 verschillend, naarmate P kleiner dan $\check{P}$ is).

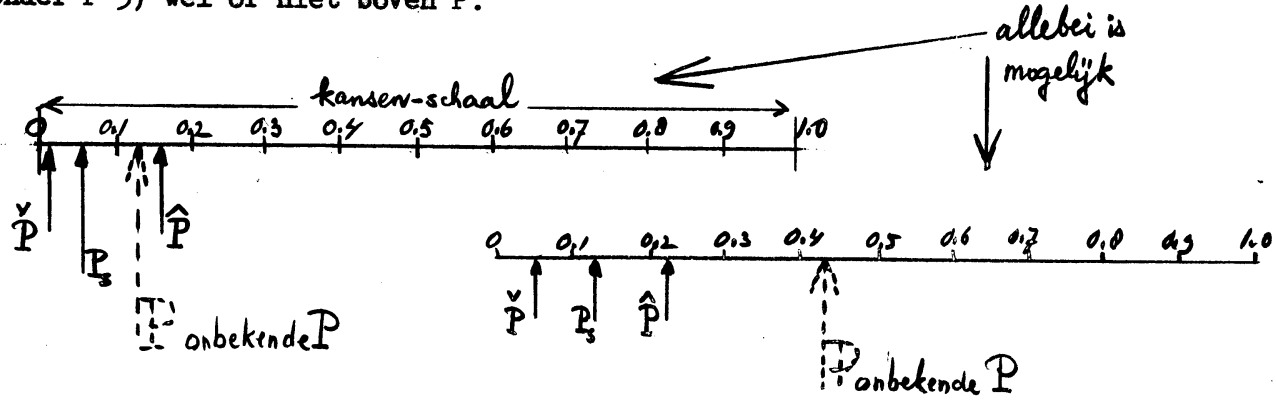
Korter: Met $\bar{x} = 10$, $s = 2\frac{1}{2}$ besluiten wij, s. pr. 0.05, dat bij de gegeven W = 10 in het onbekende normaal verdeelde universum een overschrijdingskans $\geq \check{P} = 0.0155$ behoort.

Enige getallen voorbeelden:

$\bar{x} = 6$; $s = 2\frac{1}{2}$; $N = 20$; $W = 10$;		$P_s = 0.055 = \mathbb{P}\left[\tau \geq \frac{10-6}{2.5}\right]$	
$\mathbb{P}[k \geq 1.6]$	$\mathbb{P}[\bar{x} \geq 10]$	$\mathbb{P}[k \leq 1.6]$	$\mathbb{P}[\bar{x} \geq 10]$
0.80	0.029	0.75	0.085
0.25	0.085	0.25	0.034
0.20	0.094	0.20	0.029
0.10	0.123	0.10	0.021
0.05	0.152 = $\hat{P}$	0.05	0.0155 = $\check{P}$
0.01	0.211	0.01	0.0082

Ten overvloede wellicht: zowel de uitspraak, dat de gezochte P met een berekenbare kans α beneden een berekenbare bovenste waarde (plafond) $\hat{P}$ ligt, of boven een berekenbare onderste waarde (drempel) $\check{P}$ ligt, als de uitspraak, dat de gezochte P met de kans 2α buiten het traject $\check{P}$ tot $\hat{P}$ ligt, is natuurlijk slordig geformuleerd.

Natuurlijk ligt P 1) wèl of niet tussen $\check{P}$ en $\hat{P}$ 2) wèl of niet onder $\hat{P}$ 3) wèl of niet boven $\check{P}$.



Het is beter om te zeggen: de uitspraak, dat de gezochte P tussen $\check{P}$ en $\hat{P}$ ligt, heeft een kans $2(1-\alpha)$ om fout te zijn (d.i. in strijd met de werkelijkheid). Nog anders gezegd: wanneer wij in 100 keren

volgens het beschreven procédé een interval $\check{P}$ tot $\hat{P}$ berekenen, dan zal in slechts gemiddeld 10 op de 100 keren (als $1 - \alpha = 0.05$) de ware P buiten dit interval gelegen zijn. Deze onzekerheid moet men accepteren. Als men een grotere zekerheid eist, bijv. een mis zijn in maar 2 op de 100 keren (d.w.z. $1 - \alpha = 0.01$) dan liggen $\check{P}$ en $\hat{P}$ onvermijdelijk verder uit elkaar. In ons numerieke voorbeeld (zie de twee tabelletjes) is dan $\check{P} = 0.008$ en $\hat{P} = 0.211$ (en natuurlijk ligt de puntschatting P_s weer bij 0.055 d.i. tussen deze twee waarden in).

Herhalende: we willen de overschrijdingskans P van W = 10 weten in een normaal verdeeld universum, waarvan wij de μ en σ niet kennen. Wel leverde een aselecte steekproef van N = 20 elementen een $\bar{x} = 6$ en $s = 2\frac{1}{2}$. Op grond daarvan kwamen we tot

a) de beste puntschatting is 0.055
b) met zekerheid 0.90 ligt P tussen 0.0155 en 0.152
c) met zekerheid 0.98 ligt P tussen 0.008 en 0.211

Met het bovenstaande is de existentie van een $\check{P}$ en $\hat{P}$ voor P aangetoond; natuurlijk moest het begrip "toelaatbare waarde" voor P statistisch gedefinieerd worden. Tevens werd met een numeriek voorbeeld aangetoond hoe $\check{P}$ en $\hat{P}$ berekend konden worden.

(16) Nu algemeen: $$\tau_{1-\alpha} = (W - \mu_x) : \sigma_x = \frac{W - (\mu + k_\alpha \sigma)}{\sigma \sqrt{\frac{1}{N} + \frac{k_\alpha^2}{2(N-1)}}} = \frac{\alpha - k_\alpha}{\sqrt{\frac{1}{N} + \frac{k_\alpha^2}{2(N-1)}}}$$

(17) Uitgeschreven: $$\boxed{\alpha^2 - 2\alpha k_\alpha + (1 - B) k_\alpha^2 - A = 0}$$

(18) een kwadratisch verband tussen α en k ; $A = \frac{\tau_{1-\alpha}^2}{N}$; $B = \frac{\tau_{1-\alpha}^2}{2(N-1)}$

(19) Hieruit volgen $$\boxed{\alpha^{+,-} = k_\alpha \pm \tau_{1-\alpha} \sqrt{\frac{1}{N} + \frac{k_\alpha^2}{2(N-1)}}}$$

(20) en $$\boxed{k^{+,-} = \{\alpha \pm \sqrt{\alpha^2 - ab}\} : a}$$

als $a = 1 - \frac{\tau_{1-\alpha}^2}{2(N-1)}$

en $b = \alpha^2 - \frac{\tau_{1-\alpha}^2}{N}$

Interpretatie: Men kiest $1 - \alpha$ = kans een foutieve uitspraak te doen. Dikwijls neemt men $1 - \alpha = 0.05$ bij éénzijdig begrensde betrouwbaarheidsmarges (een uitspraak in de vorm $P < \hat{P}$ of een uitspraak in de vorm $P > \check{P}$) en 0.025 bij tweezijdig begrensde betrouwbaarheidsintervallen (een uitspraak in de vorm $\check{P} < P < \hat{P}$). Overigens is deze keuze een kwestie van smaak. Zij hangt samen met de gevolgen, die men toelaten wil in het geval de werkelijkheid strijdt met de uitspraak. Door de keuze van $1 - \alpha$ ligt $\tau_{1-\alpha}$ vast, n.l. via de gestandaardiseerde normale verdeling $\mathcal{N}(0, 1)$,

t.w. via $\mathbb{P}[\tau \geq \tau_{1-\alpha} \mid 0;1] = 1-\alpha$. Bijv. $(\tau_{1-\alpha} = 1.645$
en $1-\alpha = 0.025 \rightarrow \tau_{1-\alpha} = 1.96$ $1-\alpha = 0.05 \rightarrow$
en $1-\alpha = 0.005 \rightarrow \tau_{1-\alpha} = 2.58$,

Bij gegeven N is het verband tussen k en κ een hyperbool (twee takken), zie fig. 1 waarvan wij de analytische bijzonderheden (van betekenis voor de constructie) in een addendum onderzoeken. Bij één κ behoren twee k's : k^+ en $k^- < k^+$ en bij één k behoren twee κ's: κ^+ en $\kappa^- < \kappa^+$. Men kan zulk een hyperbool voor de beantwoording van twee typen vragen gebruiken. Deze twee typen zijn:

I Gegeven W; gegeven $\mathcal{N}(\mu;\sigma)$; bekende μ en σ; gegeven N; gegeven $1-\alpha$, bijv. 0.05.
Gevraagd: als wij aselecte steekproeven van N stuks uit dit universum nemen, welke k^- wordt dan met kans $1-\alpha = 0.05$ onder - en welke k^+ wordt met kans $1-\alpha = 0.05$ overschreden?
Oplossing: substitueer in (20) de N, de $\kappa = (W-\mu):\sigma$ en $\tau_{1-\alpha} = \tau_{0.05} = 1.645$ en bereken k^- en k^+.

II Gegeven W; gegeven $\mathcal{N}(\mu;\sigma)$; onbekende μ en σ; gegeven N; gegeven $1-\alpha$, bijv. 0.05; gegeven een aselecte steekproef van N elementen, waarin $\bar{x}$ en s berekend worden.
Gevraagd: welke ondergrens $\check{P}$ en welke bovengrens $\hat{P}$ heeft de onbekende P = overschrijdingskans van W in $\mathcal{N}(\mu;\sigma)$, s. pr. $1-\alpha$?
Oplossing: substitueer in (19) de N, de gemeten $k = (W-\bar{x}) : s$ en $\tau_{1-\alpha} = 1.645$. Er komen twee κ's : κ^+ en κ^-. Deze bepalen $\hat{P}$ resp. $\check{P}$, volgens $\hat{P} = \mathbb{P}[\tau \geq \kappa^+ \mid 0;1]$ en $\check{P} = \mathbb{P}[\tau \geq \kappa^- \mid 0;1]$
d.w.z. $P \geq \check{P}$ s.pr. 0.05
$P \leq \hat{P}$ s.pr. 0.05
$\check{P} \leq P \leq \hat{P}$ s.pr. 010 (z.g. symmetrische 90% interval)

Om al deze berekeningen te vermijden heeft het voordeel een groot aantal zulke hyperbolen (bijv. N = 10, 20, 30 .....) te tekenen bij enkele waarden van α, bijv. 0.05 0.025 en 0.005. Een eerste poging daartoe is fig. 2 ($1-\alpha = 0.05$) $\alpha = 0.95$;

Toelichting bij figuur 2

a) langs de horizontale as ligt de schaal volgens Gausz voor overschrijdingskansen P, alsmede de lineaire schaal van de daarmee overeenkomende κ-waarden, via $P = \mathbb{P}[k \geq \kappa \mid 0;1]$
Langs deze as liggen ook de kansen β, die verderop besproken zullen worden. Langs deze as liggen ook de N-waarden, uitgaande van een $\delta = 0.95$, waarbij $\delta = (1-\beta)^N$ (zie onder punt 4).

b) langs de verticale as bevindt zich de lineaire k-schaal. (de eenheden der twee lineaire k- en $\mathcal{k}$-schalen behoeven niet gelijk te zijn).

In dit nomogram behoort de "45°-rechte", d.i. de rechte $k = \mathcal{k}$, bij de genormaliseerde normale verdeling $\mathcal{N}(0; 1)$. Over de constructie der hyperbolen, zie addendum.

2.2 Twee andere wijzen om betrouwbaarheidsbanden te berekenen

a) Een methode, waarbij men geen eisen stelt aan de vorm van de initiaal-verdeling, is a.v. Men verricht N onafhankelijke "metingen", gerangschikt: $x_1 \leq x_2 \leq x_3 \ldots\ldots \leq x_N$. In deze steekproef is derhalve de relatieve frequentie van $x \geq x_i$ gelijk $f_i = \frac{N - (i-1)}{N}$. Men kan nu vragen: als wij een experiment N keren doen en per keer is de kans op een gunstige uitslag p (onbekend), terwijl wij s successen ($0 \leq s \leq N$) tellen, wat is er dan van p te zeggen? Welke betrouwbaarheidsband kunnen wij voor p berekenen op basis van s en N, en bijv. s. pr. 0.05? Hiervoor kunnen wij raadplegen o.a. de z.g. Stevenstabel (die gebaseerd is op de binomiale verdeling) in het boek van Fisher and Yates [5]. Zo geeft iedere f_i ($i = 1, 2 \ldots\ldots N$) een marge voor de bij deze x_i in het universum behorende overschrijdingskans $P_i = \mathbb{P}[x \geq x_i]$. Er is echter een (ernstig?) bezwaar: deze marges worden niet onafhankelijk berekend. Van de $N - (i_1 - 1)$ elementen (in de steekproef), die $\geq x_{i_1}$ zijn, maken de $N - \{(i_1 - 1) - 1\}$ elementen, die $\geq x_{i_1 - 1}$ zijn, deel uit. Hoe men aan dit bezwaar tegemoet moet komen, is een moeilijke zaak, waarover wij met het Mathematisch Centrum van gedachten wisselden.

b) Men doet N onafhankelijke waarnemingen, gerangschikt $x_1 \leq x_2 \leq \ldots\ldots \leq x_m \ldots\ldots \leq x_N$. Wij nemen in het bijzonder de m^{de} meting x_m in het oog. Als wij dit herhaaldelijk doen in steekproeven van N elementen, is deze x_m stochastisch verdeeld. Hierdoor zal ook de overschrijdingskans $\mathbb{P}[x \geq x_m]$ in het initiaal-universum (over de verdeling waarvan niets ondersteld wordt), opgevat als een variabele $\underline{w}$, stochastisch verdeeld zijn, d.w.z. een kansverdeling $\psi(\underline{w})\, d\underline{w}$ bezitten, waarin wij in het bijzonder naar het gemiddelde $\mathcal{E}(\underline{w})$ en de variantie $\mathcal{V}(\underline{w}) = \sigma^2(\underline{w})$ vragen. Een zeer scherpe behandeling van dit lastige probleem vinden wij bij v. Dantzig [6]. Het blijkt $\mathcal{E}(\underline{w}) = \frac{m}{N+1}$; $\mathcal{V}(\underline{w}) = \frac{1}{N+1}\sqrt{\frac{m(N-m+1)}{N+2}}$

beide hangen alleen van m en N af.

Slechts als m rondom $\frac{1}{2}N$ ligt (d.i. voor x_i's rondom de mediaan), is $\underline{w}$ ongeveer normaal verdeeld (en des te beter naarmate N groter is) met een gemiddelde $\mathcal{E}(\underline{w})$ en met een standaarddeviatie $=\sqrt{V(\underline{w})}$. Men kan dus aan de gemeten x_m de kansenband $\mathcal{E}(\underline{w})-2\sqrt{V(\underline{w})}$ tot $\mathcal{E}(\underline{w})+2\sqrt{V(\underline{w})}$ toekennen, daarmede tot uitdrukking brengende, dat s. pr. 0.05 de overschrijdingskans van de gemeten x_m ligt tussen de twee aangegeven grenzen. Aldus kan men handelen voor alle in de steekproef gemeten x_i-waarden met de index i gelijk aan $\frac{1}{2}N$ of er vlak bij.

Evenals onder a brengt ook deze methode geen antwoord t.a.v. de vraag naar de overschrijdingskans van een gegeven W, die de grootste meting x_N overtreft of de kleinste x_1 onderschrijdt.

3. Een uitschieterscriterium [4]

De grootste of de kleinste waarde in een groep van N metingen (waarnemingen; elementen) kan zo groot of klein zijn, dat wij ons afvragen of wij beter doen haar uit te werpen.

Voor we een statistisch criterium ontwikkelen, eerst het volgende: Wat heet "verdacht groot" (klein)? Wanneer men er niet in slaagt voor de opvallende meting (waarneming) een "gewone" oorzaak te vinden (decimaalteken verkeerd gezet; een gewichtseenheid vergeten; luchtdruk 10 mm verkeerd afgelezen; twee in plaats van één maal gecorrigeerd; etc.) kan men twee standpunten innemen:

a) men neemt de opvallende meting toch mee, volgens het principe: "nooit iets wegwerpen als bijzondere oorzaken niet gevonden kunnen worden". Sommigen noemen dit het "wetenschappelijk eerlijke standpunt". Men beseffe echter wel, dat tòch de mogelijkheid bestaat dat de meting in kwestie niet behoort tot het universum, waartoe de overige metingen behoren.

b) men is bereid aan een statistisch criterium het laatste woord te gunnen; alsdan geschiedt het eventueel weglaten van een meting in elk geval volgens een objectieve wijze (men moet natuurlijk akkoord gaan met de waarde van de kans een fout van de eerste soort te maken). Voor zulk een geval geldt het hier nader af te leiden criterium (Er zijn meer criteria, die elk een andere interpretatie eisen; zie de uitgebreide literatuur hierover. De vraag te beantwoorden welk dezer vele criteria het "beste" is, is verre van eenvoudig. Dit probleem hebben wij in studie). Aldus handelende loopt men zeker de kans een meting,

die wèl behoort tot het universum waartoe de overige metingen behoren, uit te werpen, doch dit is een onvermijdelijke consequentie van de bereidheid de beslissing over te laten aan een uitschieterscriterium. Dit risico is evenwel bekend en berekenbaar. Men kan het zo klein nemen als men zelf wil.

Men hoort wel vragen: "mag ik een "uitschieter" weglaten als er geen gewone oorzaken gevonden kunnen worden?" De statisticus kan noch ja, noch neen antwoorden. Immers, wat bedoelt men met "mag ik"? De statisticus kan wel berekenen (en dit nog slechts als aan zekere voorwaarden voldaan wordt) met welke kans men een "opvallende" meting ten onrechte rekent te behoren tot een ander universum.

Er wordt gevraagd of het dan "beter" is om de verdachte meting weg te laten? Wat is "beter"? In dit verband ging ik na de invloed van de verdachte meting op de $\bar{x}$ en de betrouwbaarheidsmarge voor μ, alsmede op s en de regressiecoëfficiënt van de regressierechte en de correlatiecoëfficiënt in een stippenkaart. Zie addendum 6.2.

Thans de afleiding van het criterium. Zij β de kans op een $x \geqslant x_\beta$ in het universum (van welks verdeling wij nu nog niets onderstellen). Dan is $(1 - \beta)^N$ de kans op een aselecte steekproef van N elementen x_i, waar bij elke $x_i \leq x_\beta$ $(i = 1, 2, \ldots\ldots N)$. Noem de grootste "meting": $\underline{g}$ (een stochastische variabele) De kans op $\underline{g} \geqslant x_\beta$ is $1 - (1 - \beta)^N$; dat is n.l. de kans N metingen te doen, die niet allemaal x_β of minder zijn; d.w.z. tenminste één ervan moet x_β overtreffen, de grootste doet dit dan zeker [1]. Noem deze kans $1 - \delta$, zodat δ = kans op $\underline{g} < x_\beta$. Kies $1 - \delta$ klein, bijv. 0.05. Daarmee ligt β, als functie van N, vast en daardoor x_β, zodra wij de initiaalverdeling kennen.

Exact: $\beta = 1 - (\delta)^{1/N} = 1 - 0{,}95^{1/N}$; benadering: $\beta = 1 - \left[1 - (1-\delta)\right]^{1/N} \approx 1 - \exp.\left\{-\frac{1-\delta}{N}\right\}$
$= \frac{1-\delta}{N} - \frac{1}{2}(\;)^2 + \frac{1}{6}(\;)^3 \ldots\ldots \approx \frac{1-\delta}{N}$ als $N \gg 1$

Onderstel nu, dat wij onder onze N metingen de grootste g verdacht groot vinden. Mocht bij deze g in het x-universum een overschrijdingskans $P_g = \mathbb{P}[\underline{x} \geqslant g]$ behoren, kleiner dan $\beta = 0{,}05/N$, dan is daardoor de kans op een grootste onder N metingen, groter dan g, dat

---

1) De kans, dat één der N metingen de x_β overtreft, de andere beneden x_β blijven, is $N\beta(1-\beta)^{N-1}$; de kans, dat twee der N metingen boven x_β, de overige beneden x_β liggen, is $\binom{N}{2}\beta^2(1-\beta)^{N-2}$ enz. De totale kans, dat tenminste één der N metingen boven x_β uitkomt, is dan $N\beta(1-\beta)^{N-1} + \binom{N}{2}\beta^2(1-\beta)^{N-2} + \ldots\ldots \binom{N}{N}\beta^N(1-\beta)^0 = [(1-\beta) + \beta]^N - \binom{N}{0}\beta^0(1-\beta)^N = 1 - (1-\beta)^N$
Wanneer wij vragen naar de kans, dat de grootste der N metingen boven x_β uitkomt, impliceert dit niets omtrent de op 1, op 2 ..... na grootste meting. Vandaar, dat de kans, dat de grootste der N metingen groter dan x_β is, dezelfde is als de kans, dat tenminste één der N metingen de x_β overtreft.

is $1 - (1 - P_g)^N$, kleiner dan $1 - \delta = 0.05$. Om die reden noemen wij deze g, per definitie, een uitschieter.

Tot nu toe is niets over de verdeling in het initiaal-universum gezegd. Eerst nu maken wij gebruik van de theorie, in het vorige hoofdstuk besproken. Wij hebben n.l. geen absolute zekerheid over bovenbedoelde P_g, wanneer de steekproef zelf de enige informatie is omtrent dit universum. Gesteld nu we weten op bepaalde gronden, dat dit universum een normale verdeling $\mathcal{N}(\mu ; \sigma)$ heeft, echter met onbekende μ en σ. Wel geeft de steekproef een $\bar{x}_o$ en s_o, schattingen van μ en σ. Weer laten wij een "fout van de eerste soort" toe (met kans 0.05; d.i. s.pr. 0.05). Wij berekenen een k_u (u van "uitschieter"), zodanig, dat de waarde $U_o = \bar{x}_o + k_u . s_o$ in het x-universum met een kans $\beta_o = 1 - 0.95^N$ overschreden wordt. Wij weten dan dat iedere $x^1 > U_o$ een overschrijdingskans $< \beta_o$ heeft en daardoor zal de kans op een aselecte steekproef van N elementen, waarin het grootste g de boven berekende U_o overtreft, beneden 0.05 liggen. Derhalve: Indien $g > U_o$, dan deze g verwijderen en indien $g \leq U_o$ dan meenemen (alles s. pr. 0.05). U_o kan dus de uitschietersdrempel (het -niveau) genoemd worden.

N.B.: wij deden 2 keuzen I) de α, i.v.m. de s. pr.; men kan $1 - \alpha$ nemen 0.05; 0,01 etc. en II) de $1 - \delta$; men kan nemen 0.05; 0,01. Natuurlijk behoeven α en δ niet gelijk te zijn.

Vanzelfsprekend hangt de uitschietersdrempel met de grootte van de steekproef samen. Twee steekproeven ter grootte N_1 en $N_2 \neq N_1$, kunnen (zeer toevallig) dezelfde $\bar{x}$ en s hebben én dezelfde grootste meting g. Hierdoor is ook $k_1 = k_2$, als $k = (g - \bar{x}) : s$. In de ene steekproef (kleinste N) is deze g misschien géén, in de andere wèl een uitschieter.

Voorbeeld: $N_1 = 50$; $N_2 = 500$; $U_1 = \bar{x} + 3.7s$ en $U_2 = \bar{x} + 3.9s$; het kan zijn dat $\bar{x} + 3.9s > g > \bar{x} + 3.7s$. Deze situatie geldt steeds als $N > 50$ (d.w.z. wij bevinden ons dan rechts van het minimum van de "uitschieterskromme", zie het nomogram). Voor $N < 50$ zal steeds $U_1 > U_2$ zijn als $N_1 < N_2$ ($\bar{x}_1 = \bar{x}_2$; $s_1 = s_2$; $g_1 = g_2$). Dit is zeer moeilijk te "doorzien". De theorie leidt ertoe. Wij maken er alleen vermelding van.

4. Nomogram

Op de wijze, beschreven in het addendum, hebben wij, voor $1 - \alpha = 0.05$, voor N = 5, 10, 30, 50 en 100 de hyperbolen getekend ($N = \infty$ stelt de rechte $k = k$ voor) Later volgt een beter nomogram met meer N-waarden en wel voor $1 - \alpha = 0.05$ 0.025 en 0.005.

Langs de horizontale as zijn ook N-waarden uitgezet, die de β bepalen volgens $\beta = 1 - (\delta)^{1/N}$, waarbij wij $\delta = 0.95$ kozen. Bij iedere N behoort één k_u, die de uitschietersdrempel $U = \bar{x} + k_u s$ bepaalt, op basis van de in de steekproef (na weglating van de verdachte meting) berekende $\bar{x}$ en s. Men snijdt daartoe de verticaal door N met de bovenste tak van de hyperbool met parameter N. Vereniging der aldus gevonden snijpunten levert de "uitschieterskromme", die bij ongeveer N = 50 (moeilijk exact te berekenen!) een minimum heeft, althans bij $1 - \alpha = 0.05$. Natuurlijk gaat deze kromme voor toenemende N asymptotisch in de rechte $k = \kappa$ over.

5. Voorbeelden

5.1 Algemeen

a) Gegeven $\mathcal{N}(\mu ; \sigma)$, bekende μ, σ, Gegeven W. Men neemt aselecte steekproeven van gegeven grootte N. In elk daarvan : $k = (W - \bar{x}) : s$.
Nomogram: bij $\kappa = (W - \mu) : \sigma$ behoren twee k's, $k^+ > k^-$, verkregen door verticaal op κ te snijden met de twee hyperbooltakken met de gegeven N. Dan liggen 90% aller k's tussen k^- en k^+ (beter gezegd: 5% beneden k^- en 5% boven k^+).

b) Gegeven $\mathcal{N}(\mu ; \sigma)$, onbekende μ, σ. Gegeven W. Gegeven een steekproef van N metingen, waarin $\bar{x}$ en s.
Nomogram: bij $k = (W - \bar{x}) : s$ behoren twee κ's, $\kappa^- < \kappa^+$, verkregen door de horizontaal bij deze k te snijden met de hyperbooltakken bij deze N. Elke κ bepaalt een kans $\mathbb{P}[t \geq \kappa \mid 0 ; 1]$, die men direct langs de horizontale as afleest. P = gezochte overschrijdingskans van W in $\mathcal{N}(\mu ; \sigma)$; $\kappa^+ \to \check{P}$ en $\kappa^- \to \hat{P}$.
Dan geldt $P \geq \check{P}$ s. pr. 0.05;
$P \leq \hat{P}$ s. pr. 0.05
en $\check{P} \leq P \leq \hat{P}$ s. pr. 0.10

c) Gegeven N onafhankelijke metingen uit een normaal verdeeld universum met onbekende μ en σ. De grootste g ervan lijkt verdacht groot (ontoelaatbaar groot). Uitwerpen?
Nomogram: neem g weg en bereken van de overige N - 1 metingen $\bar{x}$ en s. Snijd de verticaal op N (langs hor. as) met de bovenste tak van de hyperbool bij deze N; aldus is k^+, genoemd k_u, bekend. Vorm $U = \bar{x} + k_u \, . \, s$. Is $g > U$, dan g uitwerpen; is $g \leq U$, dan niet. Mocht na uitwerping van g de dan grootste ook verdacht groot lijken, dan idem handelen.
Voor verdacht kleine metingen: zelfde k_u; als de kleinste $< U^1 = \bar{x} - k_u s$ dan uitwerpen en anders niet.

5.2 Numeriek

5.2.1 De kans op een zeer natte of een droge augustus-maand

Voor Winterswijk beschikken wij over 74 augustussommen neerslag (wisselende tussen 2 en 163 mm). Wij hebben deze op lineair waarschijnlijkheidspapier uitgezet [1] fig. 3. De punten suggereren, dat het universum waarschijnlijk normaal verdeeld is. Wij zouden natuurlijk kunnen onderzoeken of de afwijking, die de verdeling in de steekproef vertoont t.o.v. de gemiddeld te verwachten verdeling, gebaseerd op de best aangepaste normale verdeling, significant is of niet; dit is niet geschied; het plaatje is slechts illustratief bedoeld. Uit ervaring weten wij, dat de maandsommen neerslag (zeker die voor augustus) vrij goed normaal verdeeld zijn. De getrokken rechte behoort bij $\bar{x} = 75$ en $s = 39$ mm. Wij vragen nu voor een aantal waarden van h, bijv. 75, 100, 140, 160, 180, 200, 220 mm (de laatste ver boven de gemeten grootste maandsom = 163 mm) naar de betrouwbaarheidsband P tot $\hat{P}$ voor de overschrijdingskans P in het universum (s. pr. = 0.10). Hiervoor gebruiken wij het nomogram. Bijv. h = 120 mm; k = (120-75) : : 39 = 1.16. Bij k = 1.16 en N = 74 levert het nomogram $\check{P} = 0.08$ en $\hat{P} = 0.183$. Zo ook voor de andere h's. Verbind de punten en er komt een band rondom de gausz-rechte. (volkomen symmetrisch beneden 75 mm). Het antwoord op de vraag naar de kans op een augustus-som ≥ 200 mm luidt: "tussen 0.0001 en 0.0033", s. pr. 0.10. D.i. 1 keer op 300 à 10000 jaren. De j in "1 keer op j jaren" hebben wij naast de bandgrenskrommen geplaatst.

Men vraagt wel: "met welke hoeveelheid h^* behoeven wij niet te rekenen?" Als men daarmee bedoelt: die h^*, welke een overschrijdingskans heeft, kleiner dan 0.01 (weer s. pr. 0.05), dan leert fig. 3: $h^* \sim 183$ mm.

Een ander vraagt naar de kans op droge augustus-maanden. Hij spreekt van droog bijv. als $h \leq 5$ mm. Thans k = (75-5) : 39 = 1.8. Deze k = 1.8 geeft $\hat{P} = 0.07$ en $P = 0.017$. Weer s. pr. 0.05, ligt de gevraagde kans tussen 0.07 en 0.017; d.w.z. zulk een maand is te verwachten 1 keer in gemiddeld j jaren, waarbij j tussen 14 en 59 (s. pr. 0.05).

---

1) Deze maandsommen waren beschikbaar in klassen van 1 mm breedte. Klasse i duidt dan op i-1 tot i mm. Voor de relatieve overschrijdingsfrequentie van h=i mm (i = 0, 1, 2 .....), d.i. $\geq$ i mm, werd genomen $[a_i + a_{i+1} + \ldots] : N$, als a_i = aantal in klasse i en N = totaal aantal elementen = 74. Er bestaan betere procédés, maar dan moeten wij deze 74 maandsommen afzonderlijk kennen en in volgorde van niet-daling plaatsen: $h_1, h_2 \ldots h_{74}$. Dit was mij te veel werk; het was het doel niet waard. Voor de overige voorbeelden geldt hetzelfde. Zie Bernard en Bos-Levenbach, Statistica 7, 163, 1953.

5.2.2 Uitschietende scheepsmetingen van de luchtdruk

A Een uitschieter naar de lage kant

Gedurende ruim 100 jaren is op alle schepen op elke dag aan het eind van elke 4 uren-wacht de luchtdruk gemeten. Voor al deze (tot 45° breedte herleide) luchtdrukwaarden bij scheepsposities in het 2 graadsvak 44 - 46° W.L. en 24 - 26° N.B. (ongeveer midden tussen de Azoren en de Westindische Archipel) en voorzover betrekking hebbende op dagen in april, geldt de volgende frequentietabel. Totale aantal metingen: 59, dat is weinig; de scheepsroutes kruisen dit vak blijkbaar weinig.

luchtdruk mbar	aantal uitschieter mee	aantal uitschieter weg	cumulatief aantal uitschieter weg c	cumulatief aantal uitschieter weg c/N
1002 - 1004				
1004 - 1006				
1006 - 1008	1			
1008 - 1010				
1010 - 1012				
1012 - 1014				
1014 - 1016	2	2	58	1.000
1016 - 1018	7	7	56	0.965
1018 - 1020	10	10	49	0.845
1020 - 1022	17	17	39	0.671
1022 - 1024	13	13	22	0.379
1024 - 1026	7	7	9	0.155
1026 - 1028	2	2	2	0.034
1028 - 1030				
som	59	N = 58		

De cumulatieve frequentieverdeling gebaseerd op de 58 metingen (de verdachte[1]) meting, die in de klasse 1006 - 1008, weggenomen gedacht) hebben wij op de waarschijnlijkheidspapier (fig. 4) in beeld gebracht.

1) Wij hebben ondersteld, dat eerst nagegaan is of voor het "uitspringen" een gewone oorzaak te vinden is. Heyna, die mij dit voorbeeld, evenals het volgende, aan de hand deed, vertelde me, dat zulks inderdaad steeds geschiedt. Het komt voor, dat 10 mm te weinig werd afgelezen. Het komt ook voor, dat het schip in het vak in kwestie in een cycloon zat en daardoor een ongewoon lage druk mat. Als dat maar één of enkele keren gebeurde, "verstoort" deze lage druk sterk de homogeniteit van de verzameling metingen. Terecht kan men de vraag stellen: cyclonen wèl of niet meenemen? Behoren zij bij het universum of maken zij het universum tweeledig? De lengte van het basistijdvak wordt nu zeer belangrijk. De ene vraagt roept de andere op. Wat te verstaan onder homogeniteit? Wanneer heet een universum samengesteld?

Zij wijst (de punten liggen fraai lineair) op een normaal verdeeld universum. Wij schatten $\bar{x}$ = 1021.2 en s = 2.8 mbar. Het nomogram (voor $1 - \alpha = 0.05$ en $1 - \delta = 0.05$) levert een $k_u = 3.7$ (bij N = 58), zodat een uitschieter òf boven 1021.2 + 3.7 x 2.8 = 1032 òf beneden 1021-2 - 3.7 x 2.8 = 1011 mbar ligt. De questieuze meting, in interval 1006 - 1008, ligt dus zeker beneden 1011 en zou, althans volgens ons criterium, niet behoren tot het universum, waaruit de overige 58 metingen stammen (s. pr. 0.05). Dus uitwerpen.

Hoe klein is de kans, dat deze verdachte meting niettemin tot het universum behoort? Gezien het feit 1007 < 1011 (laten we onderstellen, dat de uitschieter de waarde 1007 heeft), behoort (s. pr. 0.05) bij deze 1007 in het normaal verdeelde universum, welks onbekende μ resp. σ geschat werden met $\bar{x}$ en s, een overschrijdingskans, zo veel kleiner dan $\beta^* \sim \frac{1-\delta}{N} = \frac{0.05}{59} = 0,000\,847$, dat de kans K op een steekproef van N = 59 elementen, waarin de kleinste meting ≤ 1007, zeer veel kleiner is dan de kritische drempel $1 - \delta = 0.05$.
Hoe klein precies?
Berekenen: $|1007 - 1021.2| : 2.8 = 5.17$. Het nomogram geeft voor N = 58 en bij $k^+ = 5.17$ een $\alpha \sim 4.2$, waarbij een P in de buurt van 0.00002. Aangezien $\beta \sim \frac{1-\delta}{N}$ en $\beta \sim 0.00002$, zal dan $1 - \delta \equiv K \sim 0.001$ zijn (dus $\ll 0.05$).
Interpretatie: het is niet onmogelijk, dat de verdachte meting 1007 mbar tot hetzelfde (normaal verdeelde) universum behoort, als waartoe de overige 58 stuks behoren, doch die mogelijkheid is uiterst klein (en daarom verwaarlozen wij haar; louter bij afspraak!), hetgeen hieruit blijkt, dat de kans op een aselecte steekproef van 59 elementen, met een kleinste element < 1007 slechts ongeveer 1 °/oo is (uitspraak: s. pr. 0.05).

B <u>Een uitschieter naar de hoge kant</u>

Als onder A, doch nu april; vak 26 - 28° W.L.; 2 - 4° N.B. (halverwege tussen de N.O.-punt van Zuid-Amerika en de kust Guinee).

luchtdruk mbar	aantal	uitschieter weg		
		aantal	cum.	relatief
1005 - 6	1	1	426	1.000
1006 - 7	8	8	425	0.997
1007 - 8	21	21	417	0.976
1008 - 9	34	34	396	0.930
1009 -10	82	82	362	0.850
1010 -11	91	91	280	0.657
1011 -12	87	87	189	0.443
1012 -13	50	50	102	0.240
1013 -14	35	35	52	0.122
1014 -15	8	8	17	0.040
1015 -16	5	5	9	0.020
1016 -17	2	2	4	0.0094
1017 -18	2	2	2	0.0047
1018 -19				
1019 -20				
1020 -21				
1021 -22				
1022 -23				
1023 -24	1			
som	427	426		

Is de grootste gemeten waarde, zegge 1023.5 mbar, een uitschieter? Laat haar weg. De overige 426 metingen brengen wij op waarschijnlijkheidspapier (fig. 5). Men krijgt de indruk, dat de frequentieverdeling niet normaal is (natuurlijk kunnen wij de χ^2-toets toepassen, maar ons voorbeeld heeft louter illustratieve waarde; mocht men overtuigd zijn, dat de verdeling niet volgens gausz is doch zeer scheef, dan mag ons criterium niet toegepast worden. Wellicht heeft het dan zin de logarithmen der luchtdrukwaarden te nemen). De $\bar{x}$ = 1010.8 mbar. De beste oog-rechte levert s = 1.9 mbar. Het nomogram levert bij N = 427 een k_u = 3.86 en dus liggen uitschieters òf boven $\bar{x}$ + 3.86 s = 1018 òf beneden $\bar{x}$ - 3.86 s = 1004 mbar. De verdachte meting ligt er ver boven en is dus, statistisch, een uitschieter. Mocht men tegenwerpen, dat de initiaalverdeling geen normale is doch een scheve, dan zeker scheef naar de lage waarden (staart naar links) en dan lijkt me de meting in kwestie nog eerder een uitschieter te zijn /men wist er op de afdeling Oceanografie en Maritieme Meteorologie geen verklaring voor/.

5.2.3 De oogst en het weer (neerslag)

A Een gefantaseerd voorbeeld

In bijgaand schetsje hebben wij voor elk der 11 jaren 1901 t/m 1905; 1907; 1908; 1912; 1915; 1924 en 1927 het oogstcijfer y tegen het weerselement x uitgezet. Het oogstcijfer is gecorrigeerd gedacht voor de invloed van alle andere factoren dan x. Op de vraag hoe dit gebeuren moet, gaan wij niet in. Wij tekenen dus een stippenkaart. Het punt U $\equiv x_u = 4$; $y_u = 6$ valt op.

Een uitschieter? De 10 overige punten leveren het punt A, d.i. $\bar{x} = 3.05$; $\bar{y} = 2.8$; $s_x = 1.76$ en $s_y = 1.33$; r = correl. coëff. = 0.96; de richtingscoëfficiënt b van de regressierechte $y - \bar{y} = b(x - \bar{x})$ is

$$b = \Sigma(x-\bar{x})(y-\bar{y}) / \Sigma(x-\bar{x})^2 = 0.724$$

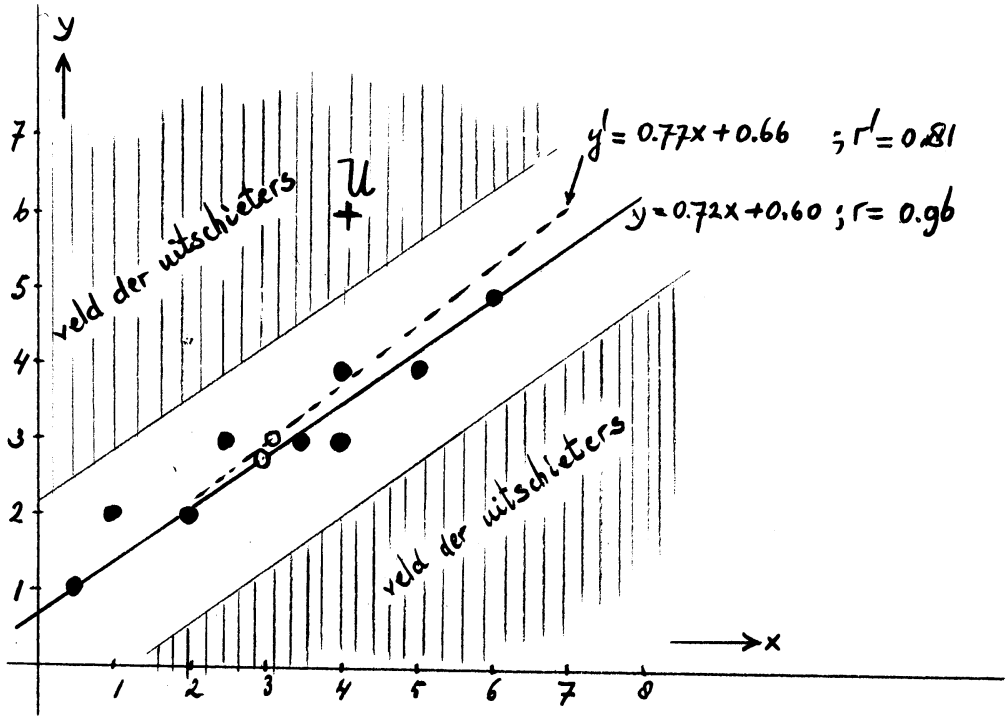
De rechte is $y = 0.724\,x + 0.60$, gaande door het punt A.

Voor deze rechte bedraagt de som der algebraïsch, // y-as, gemeten afstanden tot de rechte juist nul, terwijl wij aannemen, dat voor elke gegeven x de bijbehorende y-waarden normaal spreiden rondom de door de regressierechte gegeven y-waarde, en al deze verdelingen dezelfde variantie hebben, wil men kunnen "rekenen". De beste schatting van de variantie σ_ε^2 van deze afstanden ε is dan $s_\varepsilon^2 = (1-r^2)s_y^2$ dus $s_\varepsilon = 0.376$.

Vervolgens nemen wij U wèl mee. A gaat over in $A^1 \equiv \bar{x}^1; \bar{y}^1$. In het addendum wordt bewerkt, dat $\bar{x}^1 - \bar{x} \sim (x_u - \bar{x}) : 10 = 0.10$ en $\bar{y}^1 - \bar{y} \sim (y_u - \bar{y}) : 10 = 0.32$. De nieuwe correlatiecoëfficiënt $r' = 0.81$ (let wel < 0.96). De nieuwe beste regressierechte is $y = 0.774\,x + 0.66$. De r^1 lijkt veel kleiner dan r en de regressierechte lijkt veel verschoven te zijn door het meenemen van het verdachte punt U. Echter is het de vraag of deze veranderingen significant zijn. Hier zeker niet, doordat het aantal punten zo klein is. De standaarddeviatie σ_b van b is $\frac{\sigma_y}{\sigma_x}\sqrt{\frac{1-r^2}{N-2}} \sim 0.077$ en $b^1 - b = 0.774 - 0.724 = 0.05 < 0.077$, zodat de verandering in de richting van de regressierechte verre van significant is. Is U een uitschieter? Het nomogram levert bij N = 10 een $k_u = 4.2$. Bij $x = 4$ behoort een $y = 0.724 \times 4 + 0.60 = 3.50$, bepaald door de regressierechte bij de 10 punten. U is dus een uitschieter (bij deze $x_u = 4$) als $y_u > 3.50 + 4.2 \cdot \sigma_\varepsilon = 5.08$ en inderdaad is $y_u = 6 > 5.08$.

U is een uitschieter. Zo is $H \equiv x = 4$; $y = 4.8$ geen uitschieter. Alle punten boven de rechte u, getrokken // de regressierechte, op een afstand (// y-as gemeten) $4.2.\sigma_\varepsilon$, heten "uitschieters". Aan de andere zijde bevindt zich eveneens zulk een rechte. Aldus komen er

twee velden van uitschieters. Het lijkt ons nuttig, dat landbouwkundigen meer zulke uitschietersrechten trekken dan tot op heden het geval is geweest.



Opmerking:

1) Wij namen aan, dat langs statistische weg onderzocht is of het geen zin heeft een regressie-graad > 1 te proberen. De beschouwing is niet aan de lineariteit der regressie gebonden. Als men een parabolische regressie vermoedt, dan verwacht men, dat de punten gauszisch rondom deze parabool liggen, d.w.z. dat de // y-as gemeten, algebraïsch opgevatte, afstanden der punten tot deze parabool volgens een normale verdeling verdeeld zijn. De uitschietersvelden worden nu begrensd door twee evenver van de regressie-parabool verwijderde parabolen.

2) Het universum, waarvan bij toepassing van het criterium sprake is, heeft hier dus $\mu = 0$ en een variantie, die met $s_\varepsilon^2 = (1-r^2)\, s_y^2$ zo goed mogelijk geschat wordt. Wij dienen ook rekening te houden met de onzekerheid der ligging van de regressierechte, en daardoor ook met die der grensrechten der uitschietersvelden Zie onder C (blz. 27).

B Een voorbeeld uit de praktijk [1)]

Men heeft voor de veenkoloniën voor alle jaren 1919 t/m 1952 (behalve voor 1944) de beschikking over oogstcijfers (kg/ha) van haver. Nadat voor de trend gecorrigeerd is, maakt men een stippenkaart (fig. 6) door het oogstcijfer bijv. tegen de totale hoeveelheid neerslag in de 2de en 3de decade van maart uit te zetten: 73 punten. De bedoeling is om de regressierechte (y op x) en de correlatiecoëfficiënt te berekenen. Onmiddellijk vallen de punten 1945, 1947 en 1938 op. Het blijkt, dat oorlogsomstandigheden oorzaak zijn van de bijzonder lage oogst in 1945. Wij doen dus beter dit punt in geen der volgende beschouwingen te betrekken. Aan de lage opbrengst van 1947 zou men de zeer droge zomermaanden juni en juli schuldig kunnen achten. Om die reden doet men wellicht beter 1947 "opzij te leggen". Maar 1938? Een bijzondere oorzaak kon niet gevonden worden. Wij willen zien welke invloed het al of niet meenemen van bepaalde jaren heeft op de regressierechte en de correlatiecoëfficiënt.

Tevens zullen wij het uitschieterscriterium toepassen. De in de figuur opgenomen tabel vat een en ander kort samen. Wat zou het uitschieterscriterium over de oogst in 1947 gezegd hebben? Daartoe berekenen wij de regressierechte en de correlatiecoëfficiënt voor de stippenwolk als 1947 wordt weggelaten (31 punten). Wij nemen aan, dat de punten afstanden tot deze rechte hebben (gemeten // y-as), die spreiden om een gemiddelde nul (dit is zeker zo, indien de rechte berekend is volgens de methode der kleinste kwadraten) maar bovendien volgens een normale verdeling [Wij behandelen hier niet hoe moet worden onderzocht in hoeverre aan deze onderstelling voldaan is]. De standaarddeviatie daarvan is $\sigma_\varepsilon = \sigma_y \sqrt{1-\rho^2}$. Hierbij is σ_x resp. σ_y de standaarddeviatie in het universum der x- resp. y-waarden (beide normaal verdeeld gedacht); ρ = correlatiecoëfficiënt tussen de twee geassocieerde universa. Gemeten werd $s_y = 5.71$; $r = 0.49$, zodat $s_\varepsilon = 2.71$. Het punt 1947 ($y = 23.6$) heeft tot de regressierechte een afstand 9.7. Het nomogram ($N = 32$) levert een uitschieters-$k_u = 2.74$, terwijl $9.7 : s_\varepsilon = 3.59 > 2.74$, waaruit direct volgt, dat 1947 een statistische uitschieter is. Anders gezegd: alle oogsten, bij neerslag 55 mm, beneden 25.8 - d.i. $y - 2.74 \times 2.71$ - zijn uitschieters.

---

1) Dit voorbeeld werd mij geleverd door Ir. R.F. Fisscher.

Hoe sterk schiet 1947 uit? Hiervoor lezen wij op het nomogram bij $k = 3.59 = \frac{9.7}{2.71}$ op de rechtse tak van de hyperbool N = 31 een $\beta \ll 0.00003$ af. Gezien het feit, dat $\beta \cong \frac{1-\delta}{N}$ moet dan $1-\delta \ll 30 \times 0.00003 \cong 0.001$ zijn. Interpretatie: het zou kunnen zijn, dat 1947 tot het universum behoort, waarvan de overige punten deel uitmaken, maar dan is de kans daarop zeer klein. Want (met een zekerheid van 95 %) de kans, dat de kleinste waarde in een steekproef van 31 stuks uit een normaalverdeeld universum met gemiddelde 0 en standaard deviatie 2.71 kleiner dan -9.7 is, is zeker kleiner dan 0.001.

Wij laten dus 1947 weg uit de stippenwolk en onderwerpen vervolgens het punt 1938 aan een beschouwing. Daartoe wordt voor de overige punten (1938 weggenomen) de regressierechte en de correlatie coëfficiënt berekend. Thans is $s_\varepsilon = 3.6\sqrt{1 - 0.498^2} = 3.12$. Het punt 1938 heeft tot de regressierechte een afstand (// y-as) 7.7. Het nomogram levert bij N = 30 een $k_u \sim 2.75$, hetgeen betekent, dat, als x = 20.5 mm (de x van 1938), alle oogsten > 39.4, verkregen door bij de regressiewaarde 20.5 de waarde 2.75 x 3.12 op te tellen, uitschieters zijn. Aangezien 37.7 (van 1938) < 39.4, is <u>1938 nog geen uitschieter.</u>

Het tabelletje leert tevens, dat het, althans wat de betrouwbaarheden van de regressiecoëfficiënt en de correlatiecoëfficiënt betreft, <u>niet veel</u> zin heeft "zich druk te maken" over deze eventuele uitschieters. Immers zij beïnvloeden deze karakteristieken verre van significant (te letten op σ_r en σ_l) De oorzaak is natuurlijk gelegen in de wijde ligging der overige punten (de correlatie is weliswaar significant), d.w.z. de punten liggen niet zeer strak om de regressiwrechte. Misschien moet een regressiekromme van hogere graad dan 1 geprobeerd worden - de gestippelde curve? - misschien ook dienen de y-waarden beter of vollediger van andere invloeden dan die van x bevrijd te worden; de behandeling van deze kwestie valt echter buiten het kader van dit rapport.

C <u>Een meer exacte aanpak</u>

Voor finesses verwijzen wij naar onze statistische commentaar (met betrekking tot het gebruik van de lineaire regressie), die wij maakten bij de voordracht, die de Heer Martel hield tijdens het 4de Internationale Kongres voor Alpine Meteorologie (september 1956), gehouden te Chamonix. Aan een rapport wordt gewerkt. Er wordt aangenomen dat bij elke waarde van x (zie weer fig. 6) de y-waarden een

normale verdeling volgen met een gemiddelde $\mathcal{E}_{y(x)}$ en een standaard-deviatie $\sigma_x(y)$; doch $\sigma_x(y)$ wordt onafhankelijk van x ondersteld en daarom door σ vervangen, terwijl het verband tussen $\mathcal{E}_{y(x)}$ en x lineair ondersteld wordt: $\eta = \alpha + \beta x$. Een ander stel x-waarden zou allicht een ander stel y-waarden leveren en daardoor een andere beste (met de methode der kleinste kwadraten berekende) rechte $y = \hat{\alpha} + \hat{\beta} x$, waardoor dezelfde waarde x = 55 (bij 1947) een andere y-waarde van de rechte doet aflezen. Wij kunnen vragen naar de verdelingsfunctie van die y-waarde. Deze is een normale, rondom een gemiddelde $y = \alpha + \beta .55$ (onbekend, aangezien α en β onbekend zijn) en met een variantie

$$\sigma_{55}^2 = \sigma^2\left[\frac{1}{31} + \frac{(55-\bar{x})^2}{\Sigma(x-\bar{x})^2}\right] ,$$

waarin σ de reeds genoemde(onbekende) standaarddeviatie is, en $\bar{x}$, alsmede $\Sigma(x - \bar{x})^2$, gebaseerd zijn op de 31 x-waarden, bedoeld in rij 2 (zie fig. 6); $\bar{x} = 23.1$; $\Sigma(x - \bar{x})^2 = 983$. De beste schatting van σ is

$$\hat{\sigma} = \sqrt{\frac{31}{31-2} s_y^2 (1-r^2)} \cong s_y\sqrt{1-r^2} = 2.71$$

Zodat

$$\hat{\sigma}_{55} = 2.71\sqrt{\frac{1}{31} + \frac{31.9^2}{983}} = 2.71 \cdot 1.033 = 2.80$$

[in de legenda bij figuur werd geen onderscheid gemaakt tussen steek-proef-waarden, universum-waarden en beste schattingen daarvan]
Dit betrof dus de <u>betrouwbaarheidsband rondom de regressierechte</u> d.w.z. de <u>betrouwbaarheid harer ligging</u>. Want wij mogen nu zeggen: de uit-spraak, dat de onbekende waarde $y = \alpha + \beta .55$ zal liggen tussen

$$\{0.085(55) + 28.7\} + t_0\hat{\sigma}_{55} \quad \text{en} \quad \{0.085(55) + 28.7\} - t_0\hat{\sigma}_{55} ,$$

heeft een betrouwbaarheid van 0.95. Hierbij is t_0 de 5 %-waarde in de t-verdeling op N - 2 = 29 graden van vrijheid, d.w.z. $t_0 = 2.04$.

Vervolgens de vraag: binnen welk gebied zal de y liggen bij x = 55 indien steeds andere x-y-stellen gebruikt zouden worden? Die vraag is beantwoord als men kan berekenen welke standaarddeviatie het ver-schil $d = y_{55} - \hat{y}_{55}$ volgt, waarin $y_{55} = \alpha + 55\beta$ (onbekend door de onbekendheid met α en β, waarvan de beste schattingen $\hat{\alpha} = 28.7$ resp. $\hat{\beta} = 0.085$ zijn en $\hat{y}_{55} = \hat{\alpha} + 55\hat{\beta}$. Men kan bewijzen, dat d/s_d een t-verdeling volgt met 31 - 2 = 29 g.v.v., waarin

$$s_d = \hat{\sigma}\sqrt{1 + \frac{1}{31} + \frac{(55-\bar{x})^2}{\Sigma(x-\bar{x})^2}} = 2.71(1.44) = 3.90$$

Derhalve: De uitspraak, dat y (bij x = 55) gelegen is tussen $\{0.085(55) + 28.7\} + 2.04 \cdot 3.90 = 41.3$ en $\{0.085(55) + 28.7\} - 2.04 \cdot 3.90 = 25.4$ heeft een zekerheid van 0.95 (opnieuw is $t_0 = 2.04$).

Let wel: het is niet onmogelijk, dat de d in werkelijkheid 10 is (dit is hij voor het punt 1947 t.o.v. rechte 2) of meer, maar de kans erop is kleiner dan 0.05 en op grond daarvan zouden wij de 1947-waarde ($y = 23.6 < 25.4$) als een uitschieter kunnen bestempelen. Men ziet, dat ook langs deze weg het punt 1947 statistisch een "uitschieter" is. Doordat N vrij groot is (hier 31) komen wij tot dezelfde uitspraak langs twee verschillende wegen.

6. Addendum

6.1 Onderzoek naar de analytische bijzonderheden van de α-k-hyperbool

(fig. 1)

a) Noem $x = \alpha$ en $y = k$; verder $A = \tau^2/N$ en $B = \tau^2/2(N-1)$ dan is (volgens 17)

$\psi(x, y) \equiv x^2 - 2xy + (1 - B)\, y^2 - A = 0$

Algemene notatie: $a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + 2a_{13}x + 2a_{23}y + a_{33} = 0$

Hier $a_{11} = 1$; $a_{12} = -1$; $a_{13} = a_{23} = 0$; $0 < a_{22} = 1 - B < 1$; $a_{33} = -A$

Discriminant: $a_{12}^2 - a_{11}a_{22} = B > 0$, dus <u>hyperbool</u>

Middelpunt: uit $\partial\psi/\partial x = 0$ en $\partial\psi/\partial y = 0$ of $y = x$ en $(1-B)y = x$

middelpunt: <u>$x = 0 \quad y = 0$</u> (oorsprong)

Asymptoten: $a_{11}x^2 + 2a_{12}xy + a_{22}y^2 = 0$, hier $x^2 - 2xy + (1-B)y^2 = (x - my)(x - ny) = 0$

met $m = 1 + \sqrt{B}$ en $n = 1 - \sqrt{B}$.

Asympt: $y = \dfrac{x}{1 + \sqrt{B}} = \dfrac{x}{1 + \frac{\tau}{\sqrt{2(N-1)}}}$ en $y = \dfrac{x}{1 - \sqrt{B}} = \dfrac{x}{1 - \frac{\tau}{\sqrt{2(N-1)}}}$

(helt iets minder dan 45^0) (helt iets meer dan 45^0)

Assen: $a_{12}y^2 + (a_{11} - a_{22})\, xy - a_{12}x^2 = 0$, hier $y^2 - Bxy - x^2 = (y - px)(y + qx) = 0$

met $p = \frac{1}{2}(\sqrt{4 + B^2} + B)$ en $q = \frac{1}{2}(\sqrt{4 + B^2} - B)$

Assen: $y = \frac{1}{2}x(\sqrt{4 + B^2} + B)$ en $y = \frac{1}{2}x(\sqrt{4 + B^2} - B)$

(helt iets meer dan 45^0) ($\perp$ andere as)

Hiermede ligt de hyperbool geheel vast.

Nog enige bijzonderheden:

a) de asymptoten klappen samen voor $N = \infty$ en wel met de rechte $y = x$ ($k = \alpha$)

b) de as $y = \frac{1}{2} x \left(\sqrt{4 + B^2} + B\right)$ nadert meer en meer tot de "45°-rechte" (beter gezegd: de rechte $y = x$) naarmate N toeneemt (waardoor $B \to 0$)

c) hoe snel bijv. de hyperbooltak $y = \frac{1}{a}\left(x + \sqrt{x^2 - ab}\right)$ met toenemende x tot de asymptoot $y = \frac{x}{1 + \frac{\tau}{\sqrt{2(N-1)}}}$ nadert (des te sterker naarmate N groter is) blijkt als volgt.

Schrijf: $y = \frac{x + \sqrt{x^2 \frac{\tau^2}{2(N-1)} + \left(1 - \frac{\tau^2}{2(N-1)}\right)\frac{\tau^2}{N}}}{1 - \frac{\tau^2}{2(N-1)}} \xrightarrow[x \to \infty]{} \frac{1 + \frac{\tau}{2(N-1)}}{1 - \frac{\tau^2}{2(N-1)}} x = \frac{x}{1 + \frac{\tau}{\sqrt{2(N-1)}}}$

d) de rechte $y = x/a$ is blijkbaar een middellijn, geconjugeerd aan de y-as, want voor de hyperbooltakken geldt

$$y = \frac{x}{a} \pm \frac{\sqrt{x^2 - ab}}{a}$$

Evenzo is de rechte $y = x$ een middellijn, geconjugeerd met de x-as, want voor de 2 hyperbooltakken geldt $x = y \pm \sqrt{By^2 + A}$
Beide feiten zijn van voordeel voor een zeer nauwkeurige constructie der hyperbolen. Wat deze constructie verder betreft, hebben wij veel gemak van de hypothese: grote N (deze voorwaarde moesten we toch al stellen, anders gelden de afgeleide formules niet). Voor $N \gg 1$ geldt:

$$\kappa \approx k_\alpha \pm \frac{\tau}{\sqrt{N}}\sqrt{1 + \tfrac{1}{2}k_\alpha^2} \qquad \text{en } k \approx \frac{\kappa \pm \frac{\tau}{\sqrt{N}}\sqrt{1 + \frac{1}{2}\kappa^2}}{1 - \frac{\tau^2}{2N}}$$

In de laatste uitdrukking is bovendien ondersteld $1 + \frac{1}{2}\kappa^2 \gg \frac{\tau^2}{2N}$, maar dat is voor grote κ al gauw waar (en eerder naarmate N groter is)

e) Door de hyperbooltakken via $\kappa = k_\alpha \pm \tau_{1-\alpha}\sqrt{\frac{1}{N} + \frac{k^2}{2(N-1)}}$ te construeren zien wij goed welke de invloed van α (die $\tau_{1-\alpha}$ bepaalt) is. Trek de rechte $k = \kappa$. Snijd deze met de horizontale rechte op hoogte k_α. Zet dan vanuit dit snijpunt naar links en naar rechts hetzelfde segment $\tau_{1-\alpha}\sqrt{\frac{1}{N} + \frac{k^2}{2(N-1)}}$ uit. Dit is kleiner bij grotere N en dichter bij 0.50 gelegen (waardoor $\tau_{1-\alpha}$ kleiner is; $\alpha = 0.50 \to \tau_{1-\alpha}$) Door dus de te eisen zekerheid α geringer te kiezen (0.99 → 0.95 → 0.90 enz.) neemt $\tau_{1-\alpha}$ af, naar nul toe, en kruipen (N vast houdende) de hyperbooltakken naar elkaar toe, d.w.z. de κ^- en κ^+ gaan minder verschillen van k.

Bij $\alpha = 0.50$ is $\alpha^- = \alpha^+ = k$. Betekenis? Met een zekerheid 0.50 ligt dan de onbekende P (= overschrijdingskans van de gegeven W in het universum met onbekende μ en σ) boven $P_k = \mathbb{P}[\bar{t} \geq k \mid 0; 1]$, doch met dezelfde zekerheid ook beneden deze P_k. Aan zulk een uitspraak hebben wij natuurlijk niets. [Vernauwing van het vertrouwensinterval gaat onherroepelijk samen met vermindering van de waarde van de uitspraak]

f) Wij kunnen de k constant denken, doch de N naar ∞ laten gaan. Wat doet het vertrouwensinterval P tot $\hat{P}$ dan? Bijv. $k = 1$; $N = 10$, dan $0.05 \leq P \leq 0.36$; en $N = 30$, dan $0.083 \leq P \leq 0.26$; en $N = 100$, dan $0.11 \leq P \leq 0.21$; en $N = 500$, dan $0.13 \leq P \leq 0.18$ enz. Hoe groter de N, hoe meer P en $\hat{P}$ tot elkaar komen, om voor $N = \infty$ (dan vormt het hele universum de steekproef) in $P = \mathbb{P}[\bar{t} \geq k \mid 0; 1]$ over te gaan, wat zeer begrijpelijk is.

b) Het merkwaardige van al deze hyperbolen is dus

1°) zij hebben hetzelfde middelpunt, t.w. $k = \alpha = 0$

2°) zij hebben alle de rechten $k = 0$ (d.i. de α-as) en de rechte $k = \alpha$ tot geconjugeerde middellijnen.

N.B. elk der hyperbolen heeft een eigen tweetal hoofdassen en eigen tweetal asymptoten.

Het is misschien goed, volledigheidshalve, deze situatie nog eenmaal algemeen uit te werken.-

Gevraagd de vergelijking der hyperbolen, die een gegeven punt tot middelpunt hebben en een tweetal rechten door dit punt als geconjugeerde middellijn.

Oplossing We mogen, zonder aan de algemeenheid afbreuk te doen, de oorsprong ($x = 0$; $y = 0$) leggen in het gegeven punt en de x-as (d.i. $y = 0$) met een der twee gegeven rechten doen samenvallen.

In homogene coördinaten luidt de vgl. van de tweedegraads kromme:

$$\varphi(x, y, z) = a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + 2a_{13}xz + 2a_{23}yz + a_{33}z^2 = 0.$$

1) $\varphi = 0$ stelt een hyperbool voor als $a_{12}^2 - a_{11}a_{22} > 0$.

2) Opdat het punt $x = 0$, $y = 0$ het middelpunt zij, moet $a_{13} = a_{23} = 0$ zijn.

De vergelijking van de hyperbool is dus

$$a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + a_{33}z^2 = x(a_{11}x + a_{12}y) + y(a_{21}x + a_{22}y) +$$
$$+ z(a_{33}z) = 0,$$

met $a_{12} = a_{21}$; $a_{12}^2 - a_{11}a_{22} > 0$.

De poollijn van het oneindig verre punt $x_o = y_o$; $z_o = 0$ heeft tot vergelijking $x(a_{11}x_o + a_{12}y_o) + y(a_{21}x_o + a_{22}y_o) = 0$, d.i. $(a_{11} + a_{12})x + (a_{21} + a_{22})y = 0$. Deze willen we doen samenvallen met de x-as, d.i. $y = 0$, zodat nodig is: $a_{11} + a_{12} = 0$.

De poollijn van het oneindig verre punt x_o; $y_o = 0$; $z_o = 0$ heeft tot vergelijking $x(a_{11}x_o) + y(a_{21}x_o) = 0$ d.i. $a_{11}x + a_{12}y = 0$. Deze rechte willen wij doen samenvallen met $x = y$, dus $a_{11} = -a_{12}$.

Resultaat: de algemene vergelijking van de hyperbool, die aan de gestelde eisen voldoet, luidt (in niet-homogene coördinaten):

$-a_{12}x^2 + 2a_{12}xy + a_{22}y^2 + a_{33} = 0$ of

$x^2 - 2xy - \alpha y^2 - \beta = 0$ met $\alpha = a_{22}/a_{12}$ en $\beta = a_{33}/a_{12}$

terwijl $a_{12}^2 - 2a_{11}a_{22} > 0$ of $\alpha > -1$

Dus twee paramters α en β.

Vervangen wij x door ξ en y door k, dan is de vgl.:

$$\xi^2 - 2\xi k - \alpha k^2 - \beta = 0,$$

terwijl wij als ξ, k-vergelijking hebben afgeleid:

$$\xi^2 - 2\xi k + \left(1 - \frac{\tau^2}{2(N-1)}\right) - \frac{\tau^2}{N} = 0$$

zodat blijkt:

$$\alpha = \frac{\tau^2}{2(N-1)} - 1 \qquad \text{en} \qquad \beta = \frac{\tau^2}{N}$$

In het stelsel hyperbolen van het nomogram zijn τ en N parameters (we kozen 3 waarden voor τ, doch vele voor N), vanzelfsprekend zijn dan ook bovengenoemde α en β (onafhankelijke, doch weinig bruikbare) parameters. Natuurlijk is voldaan aan de eis $\alpha > -1$, omdat $\tau > 0$ én $N > 0$.

6.2 De invloed van een "uitschieter" op het gemiddelde en de standaarddeviatie van de steekproef en op de regressierechte en correlatiecoëfficiënt in een stippenkaart

Gegeven N - 1 metingen (waarnemingen): $x_1, x_2 \ldots\ldots x_{N-1}$ met gemiddelde $\bar{x}_{N-1}$, dat wij zonder bezwaar in het nulpunt mogen denken; de

standaarddeviatie is s_{N-1}. Wij verrichten een N^{de} metingen x_N en stellen vast, dat deze "veel" groter is dan de grootste onder de overige N-1 stuks. "Uitschieter"?
Het gemiddelde der N metingen zij: $\bar{x}_N \neq 0$ en de standaarddeviatie zij s_N. Wij willen $\bar{x}_N$ met $\bar{x}_{N-1}$ en s_N met s_{N-1} vergelijken.

a) $\bar{x}_{N-1} = (x_1 + x_2 + \ldots\ldots x_{N-1}) : (N-1) = 0$ (gegeven) $\therefore \bar{x}_N = (\sum_1^N x_i) :$

$: N = \underline{x_N : N}$.

Het meenemen van de uitschieter x_N doet blijkbaar het gemiddelde $\bar{x}_{N-1}$ met het N^{de} deel van het verschil tussen dit gemiddelde en deze uitschieter toenemen.

b) Verder:

$$s_N^2 = \sum_1^N (x_i - \bar{x}_N)^2 : (N-1) = \left[\sum_1^N x_i^2 + N(\bar{x}_N)^2\right] : (N-1) =$$

$$\left(\sum_1^N x_i^2 + \frac{x_N^2}{N}\right) : (N-1) = \left(\sum_1^{N-1} x_i^2 + \frac{N+1}{N} x_N^2\right) : (N-1) = \frac{N-2}{N-1} s_{N-1}^2 +$$

$$+ \frac{N+1}{N(N-1)} x_N^2 \simeq \underline{s_{N-1}^2 + \frac{1}{N} x_N^2} \quad \text{als } N \gg 1 \qquad 1)$$

Dus: mits $N \gg 1$, doet het meenemen van de uitschieter x_N de variantie s_{N-1}^2 met het N^{de} deel van het kwadraat van het verschil tussen gemiddelde en uitschieter toenemen.
Nu de <u>mate</u> (snelheid) van toeneming:

a) $d\bar{x}_N / dx_N = \frac{1}{N}$; blijkbaar een langzamere toeneming met het meenemen van de uitschieter naarmate deze een hoger nummer heeft (de steekproef groter is).

b) $ds_N^2 / dx_N = 2 \frac{N+1}{N(N-1)} x_N = 2 \frac{ds_N}{dx_N} s_n$ of $\frac{ds_N}{dx_N} = \frac{1}{N} \cdot \frac{N+1}{N-1}$.

Wij vragen natuurlijk: hoe groot is de factor $f = \dfrac{x_N}{\sqrt{s_{N-1}^2 + \frac{1}{N} x_N^2}}$?

---

1) Dit is ook duidelijk a.v.: als $a_i = b_i + c_i$, dan is $\sigma_a^2 = \sigma_b^2 + \sigma_c^2$, althans als b en c onafhankelijk zijn. Subsituteer: $a = x_1 + x_2 + \ldots\ldots x_N$; $b = x_1 + \ldots\ldots x_{N-1}$ en $c = x_N$.

$$f \gtrless 1 \text{ als } x_N^2 \gtrless \frac{N}{N-1} s_{N-1}^2 \text{ (als } N \gg 1\text{), d.i. } x_N \gtrless s_{N-1}\sqrt{\frac{N}{N-1}}$$

Wat zal waar zijn? Aangezien $N \gg 1$, is de vraag dus: zal de uitschieter meestal kleiner of groter dan de standaarddeviatie der overige zijn? Natuurlijk: meestal groter. d.i. $f > 1$.
Dus: s_N neemt toe met toenemende uitschieter, maar bovendien ($f > 1$)

$$\frac{ds_N}{dx_N} > \frac{d\bar{x}_N}{dx_N} \quad \text{, d.w.z.}$$

gewoonlijk neemt de standaarddeviatie van alle N metingen, waarin x_N de uitschieter is, meer toe met toenemende uitschieter, dan het gemiddelde aller N metingen doet.

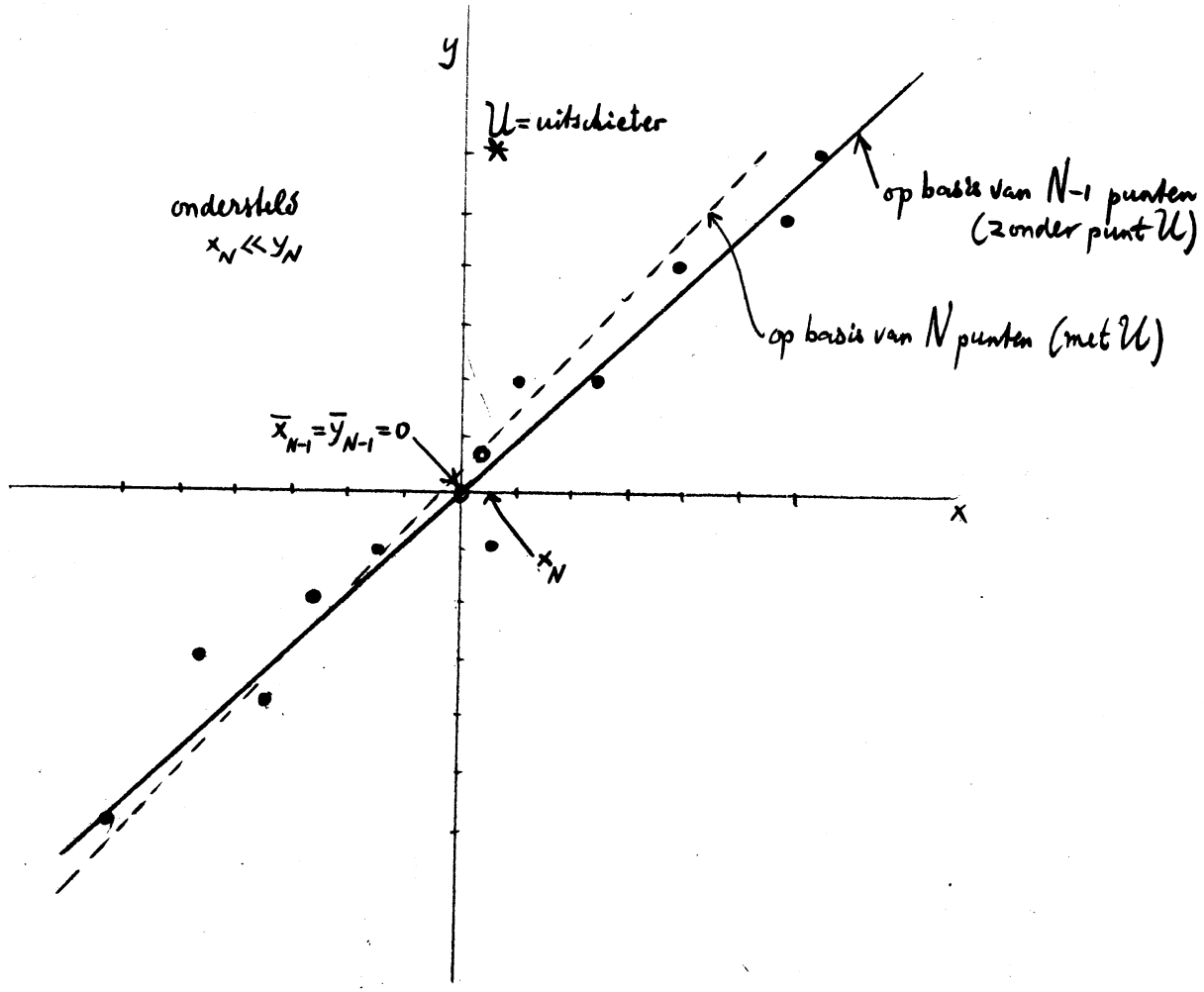
Bovenstaande beschouwing kán van betekenis zijn i.v.m. de vraag in hoeverre de ligging van de regressierechte, geconstrueerd door een puntenwolk van N punten (x-y-relatie), afhangt van het al of niet meenemen van een "uitschieter". Zie het volgende schetsje; y stelt bijv. de oogst voor en x een weerselement. Laten we aannemen, dat de regressierechte onder 45^0 loopt, hetgeen het geval is bij gestandaardiseerde x en y (hierbij hebben wij het verdachte N^{de} punt: x_N; y_N buiten beschouwing gelaten)
D.i. $\bar{x}_{N-1} = y_{N-1} = 0 \quad s_x = s_y = 1$. Aangezien er meer dan één regressierechte is beperken wij ons hier tot de eenvoudigste, t.w. $y - \bar{y} =$

$= \sqrt{\frac{\sum(y - \bar{y})^2}{\sum(x - \bar{x})^2}} \; (x - \bar{x})$; hier $\bar{x} = \bar{y} = \bar{x}_{N-1} = \bar{y}_{N-1} = 0$; $\sum(x - \bar{x})^2$
$= \sum(y - \bar{y})^2 = N - 2.$

Neem vervolgens het verdachte punt wèl mee. Hierdoor neemt $\bar{y}$ meer toe dan $\bar{x}$ (want $x_N \ll y_N$ in het schetsje; $\bar{y}$ neemt toe van 0 tot $\frac{1}{N}y_N$ en $\bar{x}$ neemt toe van 0 tot $\frac{1}{N}x_N$). Ook neemt $\sum(y - \bar{y})^2 / (N - 2)$ meer toe (n.l. met $\frac{1}{N}x^2/N$) dan $\sum(x - \bar{x})^2/(N - 2)$ n.l. met $\frac{1}{N+1}x_N^2$), zodat de nieuwe rechte is

$$y - \frac{y_N}{N} = \sqrt{\frac{N-2 + \frac{N-2}{N}y_N^2}{N-2 + \frac{N-2}{N}x_N^2}} \left(x - \frac{x_N}{N}\right), \text{ met } \sqrt{\ldots\ldots} > 1$$

De tweede regressie rechte loopt dus door een punt, rechts boven de oorsprong gelegen, en helt meer (dus onder meer dan 45°) dan de eerste. Of dit statistisch significant is, zal moeten worden onderzocht.



6.3 Berekening van $\bar{x}$ en s met of zonder de verdachte meting?

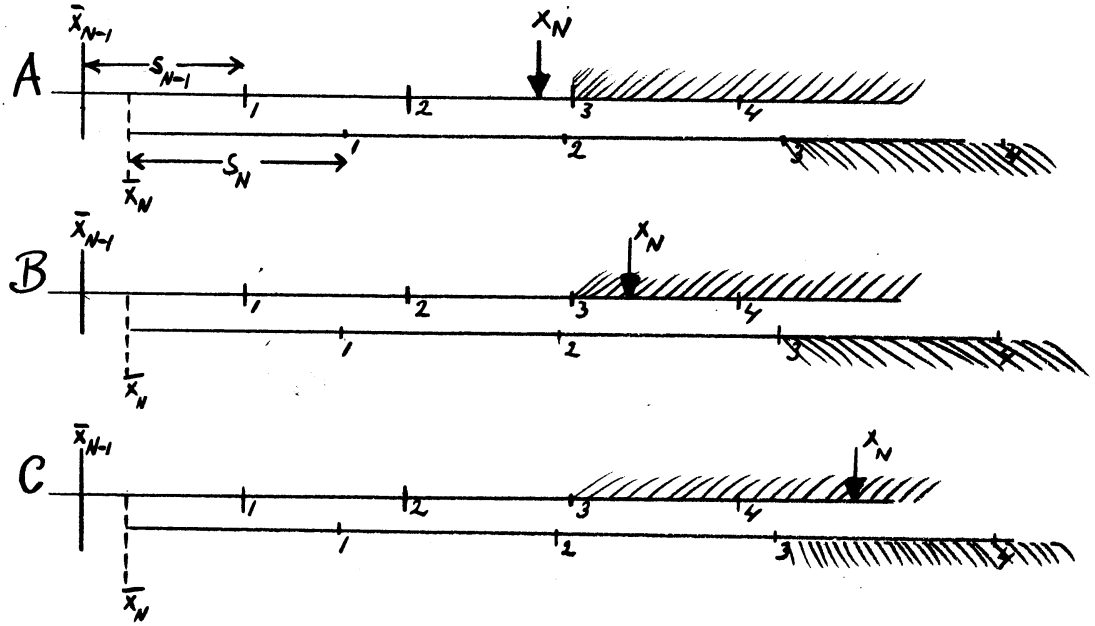
Over deze vraag wisselde ik van gedachten met een medewerker van het Mathematisch Centrum, met wie ik de in dit rapport gegeven beschouwingen besprak. Ofschoon deze gedachtenwisseling nog niet tot een concreet resultaat heeft geleid komt het me goed voor de kwestie hier toe te lichten.

Zoals gezegd, luidt het recept als volgt: als $g \equiv x_N$ de grootste der in volgorde van niet-daling geplaatste N waarnemingen is en wij g verdacht groot achten, laten wij haar weg en berekenen $\bar{x}$ en s der overige N - 1 stuks. Vanuit $\beta \sim \frac{0.05}{N}$ richten wij een verticaal op, die wij snijden met de bovenste hyperbool tak met de parameter N, waardoor langs de verticale as de k_u bekend is en een uitschieter òf $> \bar{\bar{x}} + k_u s$ òf $< \bar{\bar{x}} - k_u s$ is. Hiermede zijn we er zeker van, dat de overschrijdingskans (in het universum) van een waarneming $x^* > \bar{\bar{x}} + k_u s$ kleiner dan 0.05/N is, waardoor de kans op een aselecte steekproef van N elementen, van welke de grootste $> x^*$ kleiner dan 0.05 is.

Wij kunnen ook als volgt handelen. De $\bar{x}$ en s berekenen met medeneming van de twijfelachtige meting. Aldus: $\bar{x}_N$ en s_N, i.p.v. $\bar{x}_{N-1}$ en s_{N-1}. Zeker is $\bar{x}_N > \bar{x}_{N-1}$ en $s_N > s_{N-1}$. Met $\bar{x}_N = \bar{x}_{N-1} + \frac{1}{N}(x_N - \bar{x}_{N-1})$ en ongeveer $s_N^2 = s_{N-1}^2 + \frac{1}{N}(x_N - \bar{x}_{N-1})^2$. De β blijft $\frac{0.05}{N}$, zodat ook k_u niet verandert. Doch $\bar{x}_N + k_u s_u > \bar{x}_{N-1} + k_u s_{N-1}$. Een twijfelachtige meting "is nu minder spoedig" een uitschieter.

De moeilijkheid hierbij is m.i.: àls de meting x_N in kwestie werkelijk een uitschieter blijkt te zijn (d.w.z. wij besluiten te handelen alsof zij niet tot het universum, waartoe de overige N-1 metingen behoren, behoort), dan werden $\bar{x}$ en s te hoog berekend, juist door medeneming van x_N.

De volgende 3 mogelijkheden laten zich onderscheiden:
Gemakshalve stellen wij $k_u = 3$. Wij moeten in het schetsje laten uitkomen, dat $\bar{x}$ en s toenemen door x_N mee te nemen.



In A. De verdachte meting x_N ligt voor beide wijzen van berekenen niet in het uitschietersgebied, d.w.z. $x_N < \bar{x}_{N-1} + k_u s_{N-1} < \bar{x}_N + k_u s_N$.
In dit geval is het beter voor de beste schatting van μ en σ de $\bar{x}_N$ en s te berekenen (x_N mede nemen). Maximale informatie gebruiken.

In B. Rekenende met $\bar{x}_{N-1}$ en s_{N-1} is x_N wèl en rekenende met $\bar{x}_N$ en s_N is x_N géén uitschieter. Nu $\bar{x}_{N-1} + k_u s_{N-1} < x_N < \bar{x}_N + k_u s_N$.
Wat is nu beter terwille van een schatting van μ en σ ? De $\bar{x}_{N-1}$?

s_{N-1} of de $\bar{x}_N$; s_N? Ik zou zeggen: $\bar{x}_N$ en s_N. Maximale informatie gebruiken.

In C.Voor beide wijzen van berekenen is x_N een uitschieter.
Nu $\bar{x}_{N-1} + k_u s_{N-1} < \bar{x}_N + k_u s_N < x_N$.
Thans is het zeker het beste de μ en σ te schatten met x_{N-1} en s_{N-1} (weglating van x_N).

Wanneer wij op dit alles dieper ingaan, komen wij, menen wij, o.m. te staan voor het aardige probleem: wanneer wij herhaaldelijk uit een normaal verdeeld universum (bekende μ en σ) aselecte steekproeven van gegeven grootte N steken en iedere keer, onder weglating van de grootste meting, het gemiddelde en de standaarddeviatie berekenen, aan welke waarschijnlijkheidsverdelingen gehoorzamen deze dan? Nog wat moeilijker is de vraag: als wij iedere keer ook de kleinste meting weglaten (dus én grootste én kleinste), welke zijn dan deze waarschijnlijkheidsverdelingen?

6.4 Het uitschieterscriterium van Grubbs

Kort voor het voltooien van dit rapport ontving ik het Memorandum S 168 (M 63) van het Mathematisch Centrum, getiteld "Toetsen voor één of twee uitbijters bij een normale verdeling". Het verwijst naar een tweetal artikelen, namelijk dat van F.E. Grubbs: "Sample criteria for testing outlying observations", in Ann. of Math. Stat. 21 p. 27-58 1950, en dat van E.S. Pearson and C. ChandraSekar: "The efficiency of statistical tools and a criterion for the rejection of outlying observations", in Biom. 28 p. 308-320 1936.

Speciaal het eerste zou moeten worden bestudeerd teneinde de punten van verschil en overeenkomst tussen het criterium Grubbs en dat, behandeld in mijn rapport, duidelijk te kunnen stellen. Hiervoor ontbrak mij de tijd. Toch leek het me goed het "recept" van Grubbs hier te laten volgen en van kort commentaar te voorzien.

Er zijn n onafhankelijke waarnemingen verricht, gerangschikt a.v.: $x_1 \leq x_2 \leq x_3 \ldots\ldots \leq x_n$. De nulhypothese is, dat deze steekproef afkomstig is uit één normale verdeling; de alternatieve hypothese is, dat de grootste waarneming, x_n, uit een andere verdeling komt dan de andere waarnemingen (wordt deze andere verdeling ook normaal gedacht? C.L.) De toetsingsgrootheid is $T_n = (x_n - \bar{x}) : s$, als $\bar{x} = \frac{1}{n}\sum_1^n x_i$ en $s^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$ (waarom niet n-1 in noemer? C.L.) Grubbs leidde de

exacte verdeling van deze T_n (analoog met onze k_u) af, onderstellende, dat de steekproef uit een normaalverdeeld universum (met bekende parameters? C.L.) stamt, daarbij aan de onbetrouwbaarheidsdrempel $(1-\alpha)$ een bepaalde waarde toekennende (G. nam: 0.01, 0.025, 0.05 en 0.10)

Men lette vooral op de interpretaties der twee criteria. Als men bovendien bedenkt, dat in het uitschieterscriterium, zoals wij het hier stelden, het gemiddelde en de standaarddeviatie van het normaal verdeelde universum onbekend zijn, en slechts met het gemiddelde en de standaarddeviatie in de steekproef geschat worden, dan kan men verwachten, dat een "uitbijter" meer moet "uitbijten", alvorens verwijderd te moeten worden volgens het in dit rapport behandelde dan volgens het door Grubbs ontworpen criterium. Het volgende tabelletje bevat enige getallen waaruit dat blijkt:

Eénzijdige toetsing		
kritieke waarde T_n resp. k_u volgens het criterium		
grootte steekproef	bij Grubbs $\alpha = 0.95$	in ons rapport $\alpha = 0.95$ $\delta = 0.95$
n	T_n	k_u
10	2.29	4.31
16	2.52	4.01
20	2.62	3.92
25	2.72	3.83

blijft toenemen ↓ (bij Grubbs)

neemt toe voorbij n > 50 ↓ (in ons rapport)

Men kan, zoals mij bleek in gesprekken met de Heer Doornbos van het Mathematisch Centrum, tegen het in dit rapport besproken criterium de volgende bezwaren hebben

a) het is niet voldoende direct gericht op het doel. Maar wat is het doel? Sommigen beweren, dat het ons niet interesseert of de kans, dat de grootste waarde g van N onafhankelijke metingen groter is dan de gemeten grootste x_N (onder de N onafhankelijke meetwaarden $x_1 \leqslant x_2 \leqslant \ldots\ldots \leqslant x_N$), zeer klein is (bijv. hoogstens 0.05), doch wel of het verschil tussen de grootste waarde x_N en de overige N-1 waarden te groot is om toevallig genoemd te mogen worden. Men kan het met deze opvatting eens zijn, doch ook niet. Het lijkt ons een kwestie van smaak. Bovendien wat moet men verstaan onder het "verschil" tussen x_N en de "overige N-1 meetwaarden"? Moet men beschouwen $x_N - x_{N-1}$ of $x_N - \bar{x}_{N-1}$, als $\bar{x}_{N-1}$ = gemiddelde der overige meetwaarden?

Vast staat evenwel: hoe meer de groep dier overige N metingen (x_1 x_2 ..... x_{N-1}) van de x_N verwijderd ligt, hoe eerder heet x_N een uitschieter, volgens het criterium van Grubbs en dat in dit rapport.

b) De constructie van het uitschieterscriterium, in dit rapport behandeld, is zodanig, dat de hypothese, dat alle waarnemingen uit hetzelfde universum komen, verworpen wordt als $1 - (1 - \hat{P})^{N+1}$, waarbij δ een van te voren gekozen klein getal is en $\hat{P}$ de overschrijdingskans van x_{N+1} in het initiaal-universum voorstelt. Wij toetsen dus met een onbetrouwbaarheidsdrempel, die s. pr. α kleiner is dan δ . M.a.w. als $\alpha = 0.05$, dan is deze onbetrouwbaarheidsdrempel in 19 van de 20 gevallen kleiner dan δ . Men kan dit terecht een bezwaar noemen, n.l. het bezwaar, dat de toets niet een vaste voorgeschreven onbetrouwbaarheid heeft. Aan dit bezwaar is in ons betoog niet te ontkomen. Het is inhaerent aan de aanpak. Andere criteria vermijden dit bezwaar, doch hebben weer andere zwakke plekken; ernstig lijkt ons het bezwaar in elk geval niet.

6.5 De Centrale Limiet-stelling

Als $\underline{y}_1$, $\underline{y}_2$ ..... $\underline{y}_k$ onafhankelijk verdeelde stochastische variabelen zijn, terwijl in elk dier k verdelingen gemiddelde $\mathcal{E}(\underline{y}_i) = \eta_i$ en variantie $\mathcal{V}(\underline{y}_i) \equiv \mathcal{E}(\underline{y}_i - \mathcal{E}\underline{y}_i)^2 = \sigma^2(\underline{y}_i)$ bestaan, dan is de grootheid

$\underline{z} = \sum_1^k a_i \underline{y}_i$ ook stochastisch verdeeld.

Er geldt: $\mathcal{E}\underline{z} = \sum_1^k a_i \eta_i$ $\qquad$ $\mathcal{V}(\underline{z}) = \sum_1^k a_i^2 \mathcal{V}(\underline{y}_i)$

De c.l. stelling zegt dan verder, dat de verdeling van $\underline{z}$ steeds minder van de normale, gebaseerd op de bovenbedoelde $\mathcal{E}(\underline{z})$ en $\mathcal{V}(\underline{z})$, gaat verschillen naarmate k groter is. Het laatste wordt wel uitgedrukt a.v.: Wanneer men een interval Q_1 tot Q_2 geeft, waarbinnen men $\underline{z}$ wenst te liggen, dan bestaat er een corresponderend interval q_1 tot q_2 voor de normale verdeling $\mathcal{N}(\mu; \sigma)$, zodanig, dat q_1 alleen met Q_1, $\mathcal{E}(\underline{z})$ en $\mathcal{V}(\underline{z})$ samenhangt en q_2 alleen met Q_2, $\mathcal{E}(\underline{z})$ en $\mathcal{V}(\underline{z})$, terwijl men het absolute verschil

$$\left| \mathbb{P}[Q_1 \leq \underline{z} \leq Q_2] - \mathbb{P}[q_1 \leq \underline{x} \leq q_2] \right|$$

zo klein maken kan als men wil door k voldoende groot te nemen. De voorwaarden, waaronder de stelling geldt, zijn niet streng, doch soms technisch

ingewikkeld. Een essentiële eis is, dat de variantie van $\underline{z}$ naar ∞ gaat en dat de variantie van elk der samenstellende variabelen y_i, gedeeld door die van $\underline{z}$, naar nul gaat met toenemende k.

N.B. Er wordt niet geëist, dat de y_i's normaal verdeeld zijn, maar àls ze het zijn, dan is $\underline{z}$ exact normaal verdeeld en wel voor elke k.

Voorbeelden:

1) $\underline{x} = \frac{1}{N}\sum_1^N \underline{x}_i$; dan is $\underline{z} \equiv \underline{x}$; iedere $a = \frac{1}{N}$; Indien het initiaal-universum een normale verdeling $\mathcal{N}(\mu;\sigma)$ heeft, is $\bar{x}$ exact normaal verdeel, met $\mathcal{E}(\underline{x}) = \mu$ en $\mathcal{V}(\underline{x}) = \sigma^2/N$

2) $\underline{s}^2 = \frac{1}{N-1}\sum_1^N (\underline{x}_i - \bar{\underline{x}})^2$, dan is iedere $a = \frac{1}{N-1}$ en $(\underline{x}_i - \bar{\underline{x}})^2$ staat voor de variabele $\underline{y}_i$.

Exact geldt: $\mathcal{E}(\underline{s}^2) = \sigma^2$ en $\mathcal{V}(\underline{s}^2) = 2\sigma^2/(N-1)$ als σ^2 de variantie van het normaal verdeelde universum $\mathcal{N}(\mu;\sigma)$ is. $(N-1)\frac{s^2}{\sigma^2}$ is verdeeld als χ^2 met N-1 graden van vrijheid. Hoe groter N, hoe meer gaat de χ^2-verdeling op een normale (gemiddelde N en variantie 2N) gelijken en daardoor ook de verdeling van $\underline{s}^2$.

Een grootheid u heet verdeeld als χ^2 met N-1 g.v.v. als de verdeling is a.v.

$$\phi_{N-1}(u)\,du = \frac{u^{\frac{1}{2}(N-3)}\,e^{-\frac{1}{2}u}}{(\sqrt{2})^{N-1}\{\frac{1}{2}(N-1)\}!}\,du$$

3) $\underline{G} = \sum_1^N \frac{(\underline{W}_i - E_i)^2}{E_i}$. Dan is $a_1 = \frac{1}{E_1}$; $a_2 = \frac{1}{E_2}$ ..... $a_N = \frac{1}{E_N}$.

Verder $A = \sum_1^N W_i$

E_i = gemiddeld aantal in het interval i op grond van de nulhypothese (m.b.t. de theoretische frequentieverdeling)te verwachten aantal waarnemingen.

$\underline{W}_i$ = geteld aantal. Deze $\underline{W}_i$ is stochastisch verdeeld, als wij herhaaldelijk een steekproef van A metingen uit hetzelfde universum doen en deze over dezelfde N klassen verdelen.
Thans spelen de variabelen $(\underline{W}_i - E_i)^2$ de rol van y_i.

Wanneer de theoretische fréquentieverdeling binnen het universum door p parameters gekarakteriseerd wordt, wier numerieke waarden met de steekproef berekend worden, is deze $\underline{G}$ verdeeld ongeveer als χ^2 met N - (p+1) graden van vrijheid.

Hoe groter N - (p+1) hoe minder wijkt deze G-verdeling van een normale af met $\mu = N - p$ en $\sigma^2 = 2(N-p)$. Het "ongeveer" hangt hiermee samen: iedere E_i moet niet te klein zijn, liefst groter dan 5, bij voorkeur groter dan 10. Exacte afleiding van de G-verdeling o.a. bij Derksen [7].

7. Literatuur

A Geciteerde

[1] "Statistical Inference", notes prepared by J.H. Roberts on lectures by Prof. W. Allen Wallis, during 1949 - '50. Un. of Chicago.
Wallis eindigt met de formules. Wij achtten het nuttig er een nomogram op te baseren. Het uitschietersprobleem (hoofdstuk 3 in ons rapport) roert Wallis niet aan.

[2] "Selected techniques of statistical analysis for scientific and industrial research and production and management engineering". by the Stat. Res. Group, Columbia University; 1947; p. 57 e.v.

[3] "Statistical theory with engineering applications" by A. Hald, Prof. of Statistics, Un. of Copenhagen; 1952; p. 303 e.v.

[4] "Zum Ausreiszerproblem" von U. Graf und H.J. Henning, in Mitt. Blatt für Math. Stat. 4 1-9 1952. De auteurs komen tot een formule voor het "uitschietersniveau", die geheel analoog is aan degene, die wij in dit rapport afleiden, doch volgens een redenering en berekening, die wij niet helemaal kunnen volgen. Zij tekenen een nomogram, dat alleen voor het probleem der uitschieters van betekenis is.

Zie verder de lijst van Literatuur met betrekking tot het vraagstuk der "uitschieters", verdachte metingen, "uiterwaarden", "uitbijters" of welke andere naam ook.

[5] "Statistical Tables for biological, agricultural and medical research" by Fisher and Yates.
Zie de tabel "Binomial and Poisson Distributions; Limits of the expectation" (based on W.L. Stevens)
Ook Levert: K.N.M.I. R III - A 49, 1950

[6] Kadercursus, Mathematisch Centrum, Amsterdam, Hoofdstuk 6, § 2.

[7] Inleiding tot de correlatierekening, pg. 52 e.v., 1935.

B. Over "uitschieterscriteria" (niet volledig)

Schrijver	Artikel	tijdschrift
B. Peirce	Criterion for the rejection of doubtbul observations.	Astr. Journal II 161 1852
B.A.J. Gould	On Peirce's Criterion for the rejection of doubtful observations with tables for facilitating its application.	Astr. Journal IV 81 1854-56
J. Winlock	On Professor Airy's Objections to Peirce's Criterion.	Astr. Journal IV 145 1856
E.J. Stone	On the rejection of discordant observations.	Monthly Not. Royal Astr. Soc. XXVIII 165 1867-68
J.W.L. Glaisher	On the rejection of discordant observations.	Monthly Not. of the Royal Astr. Soc. XXIII 391 1872-73
"	Note on a paper by Mr. Stone "On the rejection of discordant observation"	Monthly Not. of the Royal Astr. Soc. XXXIV 251 1873-74
E.J. Stone	On the rejection of discordant observations.	Monthly Not. Royal Astr. Soc. XXXIV 9 1873-74
"	Note on a discussion relating to the rejection of discordant observations.	Monthly Not. Royal Astr. Soc. XXXV 107 1874-75
C.A. Schott	On Peirce's Criterion	Proc. Am. Ac. Arts and Sciences New series V Whole series XIII 350 1877-78
B. Peirce	On Peirce's Criterion	Proc. Am. Ac. Arts and Sciences New series V Whole series XIII 348 1877
F.Y. Edgeworth	On discordant observations.	Phil. Mag. Series 5 XXIII 369 1887

Schrijver	Artikel	tijdschrift
S.A. Saunder	Note on the use of Peirce's Criterion for the rejection of doubtful observations	Monthly Not. of the Royal Astr. Soc. LXVIII 432 1902-03
R.M. Stewart	Peirce's Criterion	Popular Astr. XXVLII 2 1920
"	The treatment of discordant observations	Pop. Astr. XXVIII 4 1920
J.O. Irwin	On a criterion for the rejection of outlying observations	Biometrika 17 100, 238 1925
H. Jeffreys	An alternative to the rejection of observations	Proc. R.S. Series A CXXXVII 78 1932
P.R. Rider	Criteria for rejection of observations	Washington University Studies New Series No. 8 1933
A.T.Mr. Kay	The distribution of the difference between the extreem observation and the sample mean in samples of a normal universe	Biometrika 27 466 1935
W.R. Thompson	On a criterion for the rejections and the distribution of the ratio of deviation to sample standard deviation	Ann. Math. Stats. 6 214 1935
E.S. Pearson and C. Chandra Sekhar	The efficiency of statistical tools and a criterion for the rejection of outlying observations	Biometrica 28 908 1936
E.F. Drion en Th.J.D. Erlee	Het probleem van de zogenaamde uitbijters	Chemisch Weekblad 45 857 1949
F.E. Grubbs	Sample criteria for testing outlying observations	Ann. Math. Stat. 21 27-58 1950
U. Graf und H.J. Henning	Zum Ausreisserproblem	Mitt. Blatt für Math. Stat. 4 1-9 1952
R. Doornbos and H.J. Prins	A slippage test for a set of gamma-variates	Math. Centrum Amsterdam Rep. S 187 (VP4) 1956
R. Doornbos	Toetsen voor één of twee uitbijters bij een normale verdeling	Math. Centrum Amsterdam Rapp. S 168 (M63) 1956
R. Doornbos, H. Kesten and H.J. Prins	A class of slippage tests	Math. Centrum Amsterdam Rep. S 206 (VP8) 1956

Summary

Two practical questions gave rise to the study treated in this report.

a) Suppose a random sample of N elements is taken from an universe with unknown distribution, or with a distribution of known type, but unknown numerical parameter values. This sample gives the values the mean: $\bar{x}$ and the standard deviation s. A value W is also given. We ask: what is the value of the exceedance probability of W in this population or between which values is this probability situated for a given degree of reliability?

b) In a random sample of N elements drawn from a population with unknown distribution, or with a distribution of known type, but with unknown numerical parameter values, the largest element (or the smallest one) seems to be "unreliable" or "suspected" ("outlyer"). We suppose that this element does not belong to be universe out of which the other elements have been taken. We ask whether an objective criterium can be constructed by which it is possible to decide whether such a questionable measurement should be rejected or not.

It proves to be possible to answer both questions fairly easily only if the parent population is distributed normally.

First of all attention is paid to the so called non central t distribution. Let a normal universum be given with known mean value μ and standard deviation σ and a value W. A random sample is taken from this population, which gives $\underline{\bar{x}}$ and $\underline{s}$. Since these quantities are stochastically distributed we ask for the distribution of the variable $\underline{k} = (W - \underline{\bar{x}}) : \underline{s}$. This distribution is the so called non central t distribution. The latter can be derived easily only if the parent distribution is normal; if not, then large difficulties arise.

Next the question a, mentioned above, is treated. The sample drawn from a normal population with unknown μ and σ gives $\bar{x}$ and s, which are the best estimations of μ and σ. Further a value W is given and we ask for the unknown exceedance probability P of this W on the basis of $\bar{x}$ and s. It proves to be possible to compute a value $\check{P}$ in such a way that $P > \check{P}$, s. pr. $1 - \alpha = 0.05$, that is "salva probabilitate" 0.05, which means: only in 5 out of 100 cases the statement is false, that is: the true P is smaller than the computed $\check{P}$. In the same way we compute a value $\hat{P}$, so that $P < \hat{P}$, again s. pr. 0.05. So with a reliability $1 - 2(1-\alpha) = 0.90$ (or s.pr. 0.10) the unknown exceedance probability is situated between

$\check{P}$ and $\hat{P}$ around the "best" value P_o. This P_o is the exceedance probability of the value $k = (W-\bar{x}) : s$ in the so called standardized normal distribution ($\mu = 0$; $\sigma = 1$).

We consider three values of $1 - \alpha$, namely 0.05; 0.025; 0.005. The relation between α an k is shown in formulas (17), (19), (20). On the basis of these relations a nomogram is constructed. Fig. 2 is only an illustration. The use of this nomogram is explained as follows:

I Let be given W and a normal population with known μ an σ; further N and $1 - \alpha$, for instance 0.05. Let random samples of N elements be drawn from this universe. $\underline{k} = (W-\underline{\bar{x}}) : \underline{s}$. In this way a $\underline{k}$-distribution arises. Which value k^- is exceeded with a probability $\alpha = 0.05$ and which value k^+ with a probability $\alpha = 0.95$? Answer: substitute N in (20), $\alpha = (W - \mu) : \sigma$ and $\tau_{1-\alpha} = \tau_{0.05} = 1.645$ and compute k^- and k^+ or use the nomogram to read off these values.

II Let be given W and a normal population with unknown μ and σ; further $1 - \alpha$, for instance 0.05. A random sample of N elements gives $\bar{x}$ and s. Which upper and lower limit has P, the unknown exceedance probability of W, s. pr. 0.05? Answer: Substitute N in (19), $k = (W - \bar{x}) : s$ and $\tau = 1.645$ and compute α^+ and α^-. Then these two values determine two values $\hat{P}$ and $\check{P}$ namely the exceedance probabilities of the values α^+ and α^- in the standardized normal distribution. Or use the nomogram and read off $\hat{P}$ and $\check{P}$.

Next the problem whether to reject a suspicious measurement ("outlyer") or not is discussed. One should seek for common causes why the largest (smallest) measurement is so large (small) as it is. If these causes cannot be found it is possible

(i) to take the questionable measurement into consideration according to the rule:("it is forbidden to reject any measurement") or

(ii) to use an objective criterium in order to decide whether the measurement should be rejected or not.

Suppose we follow the second point of view, then we may reason as follows: let the exceedance probability of the value x_β in the universum be β. Then $(1 - \beta)^N$ represents the probability of a random sample of N elements x_i, with each $x_i \leqslant x_\beta$ ($i = 1, 2, \ldots\ldots N$). When the largest element is called $\underline{g}$ (this element is stochastically distributed) then we can also say that the probability that $\underline{g} \geqslant x_\beta$ is $1 - (1 - \beta)^N$. This probability is called $1 - \delta$ or δ is the probability that $\underline{g} < x_\beta$. Statisticians generally give $1 - \delta$ the value 0.05 or 0.01. (the threshold value). Then β

is fixed by N and x_β can be determined in case the initial population is known; $\beta = 1 - \delta^{1/N} \approx (1 - \delta) : N$ if $N \gg 1$ and $1 - \delta$ is small. Now suppose we think the largest element g in the group of N elements too large to be caused only by chance.

If in the parent population this g has an exceedance probability P_g smaller than the ciritical value $\beta^* \approx 0.05/N$ than $1 - \delta = 1 - (1 - P_g)^N < 1 - \delta^* = 0.05$ and for this reason (only by definition) this g is called a discordant measurement which should be rejected. But the universe is not known fully since the sample gives the only information. However, suppose we know that the population is normally distributed, then P_g can not be found exactly, but only s. pr. $1 - \alpha$, for instance 0.05. We compute a value k_u and a value $U = \bar{x} + k_u s$ such that this U is equalled or exceeded in the population with a probability $\beta = 1 - (1 - \alpha)^{1/N}$. Then it is certain that each value $x^1 > U$ has an exceedance probability smaller than β and consequently that the probability of a random sample of N elements, in which the largest one (g) is larger than this U, is smaller than 0.05. So (i) if $g > U$ than this g is rejected and (ii) if $g \leqslant U$ than it is not permitted to reject g. This decision is made of course s. pr. $1 - \alpha$. N.B. two choises: 1) the value of s. pr. $= 1 - \alpha$; for instance 0.05; 0.025; 0.005 and ii) the value of $1 - \delta$, for instance 0.05; 0.01.

The computations can be avoided by using the nomogram in the following way. Leave out g and compute the mean $\bar{x}$ and standard deviation s of the remaining N-1 elements. Compute $\beta = 1 - \delta^{1/N} \approx (1 - \delta) : N$ and read off the value k^+, called k_u in this case, for this value of β (along horizontal axis) with the help of the upper curve with parameter N. Then $U = \bar{x} + k_u s$ and make the comparison with g.

In the paragraphs 5.2.1 5.2.2 5.2.3 examples are given partly from the practice of rainfall measurements (questions of the exceedance probabilities of rainfalls much larger than the largest ever measured), partly from the computations of frequency distributions of airpressures measured on ships passing through definite 2 degrees squares (some times measurements do not seem to belong to the universe of the remaining ones), and partly from the relations between crops and the weather (sometimes we doubt whether a measurement should be desregarded or not).

In the addendum the mathematical details of the curves in the nomograms are discussed.

Paragraph 6.3 deals with the consequences of taking into consideration the questionable measurement or not.

In 2.2 and 6.4 some other procedures of computing reliability limits of unknown probabilities, and the threshold value with regard to the rejection of suspected largest or smallest values are discussed. Several articles on the same subject have been written by different authors. The criterium discussed in this report is one out of many possible criteria. A detailed comparison with other rejection rules is not attempted here.

Betrouwbaarheidsmarge

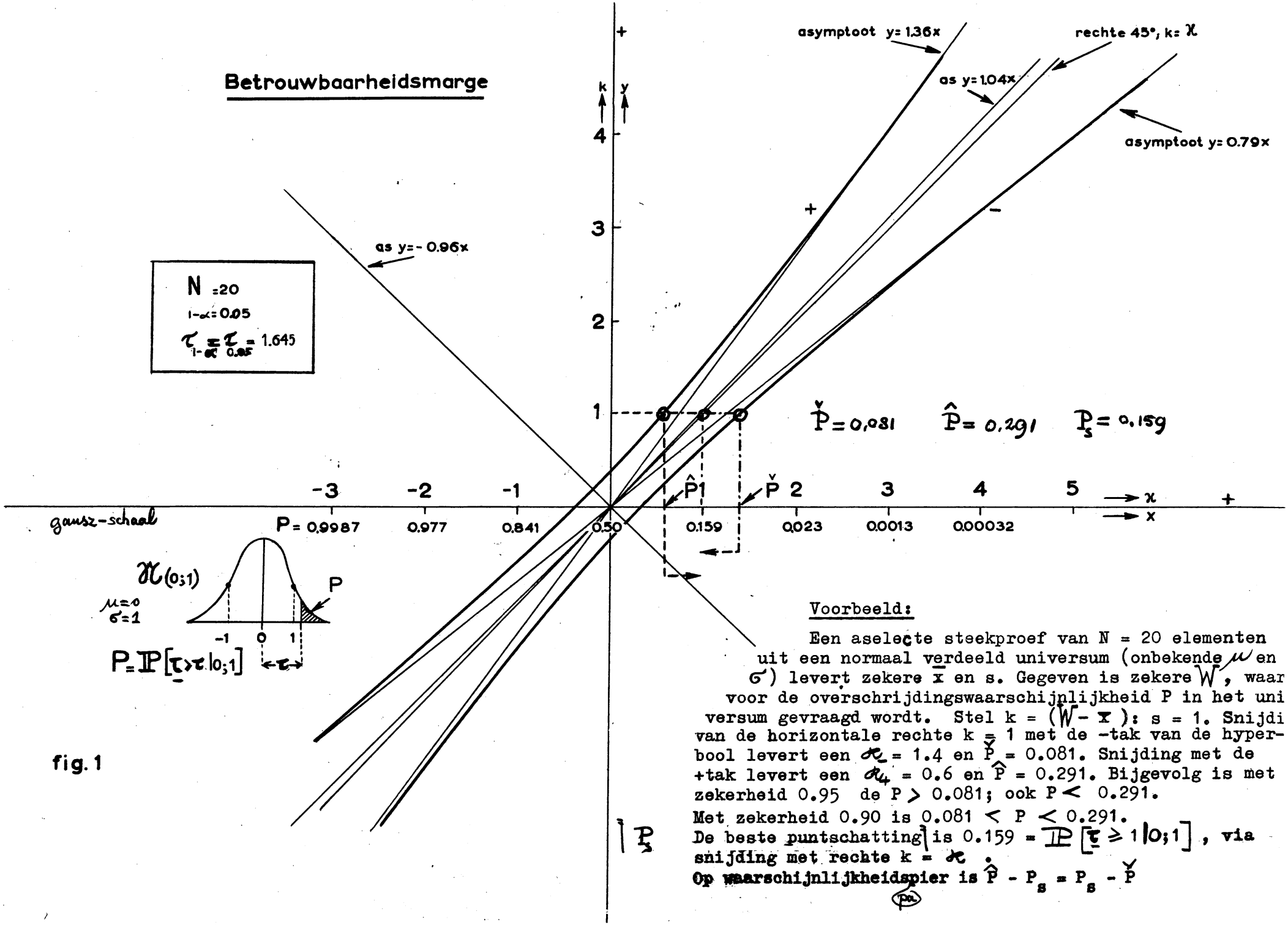


fig. 1

Voorbeeld:

Een aselecte steekproef van N = 20 elementen uit een normaal verdeeld universum (onbekende μ en σ) levert zekere $\bar{x}$ en s. Gegeven is zekere W, waar voor de overschrijdingswaarschijnlijkheid P in het universum gevraagd wordt. Stel $k = (W - \bar{x}) : s = 1$. Snijdi van de horizontale rechte $k = 1$ met de -tak van de hyperbool levert een $\alpha_2 = 1.4$ en $\check{P} = 0.081$. Snijding met de +tak levert een $\alpha_4 = 0.6$ en $\hat{P} = 0.291$. Bijgevolg is met zekerheid 0.95 de $P > 0.081$; ook $P < 0.291$.

Met zekerheid 0.90 is $0.081 < P < 0.291$.

De beste puntschatting is $0.159 = P[\tau \geq 1 \,|0;1]$, via snijding met rechte $k = \alpha$.

Op waarschijnlijkheidspier is $\hat{P} - P_s = P_s - \check{P}$

$$\boxed{\kappa^2 - 2\kappa k_\alpha + (1-B)k_\alpha^2 - A = 0} \qquad A = \tau^2/N \qquad B = \tau^2/2(N-1)$$

$$k = \frac{W-\bar{x}}{s} \qquad \kappa = \frac{W-\mu}{\sigma}$$

(1) $$\boxed{\kappa} = k_\alpha \pm \tau\sqrt{\frac{1}{N} + \frac{k_\alpha^2}{2(N-1)}} \cong \boxed{k_\alpha \pm \frac{\tau}{\sqrt{N}}\sqrt{1+\tfrac{1}{2}k_\alpha^2}} \quad \text{als } N\to\infty$$

$$k = \frac{\kappa \pm \sqrt{\kappa^2 - ab}}{a} \qquad a = 1 - \frac{\tau^2}{2(N-1)} \qquad b = \kappa^2 - \frac{\tau^2}{N}$$

(2) $$\boxed{k} = \frac{\kappa \pm \frac{\tau}{\sqrt{N}}\sqrt{1 - \frac{\tau^2}{2N} + \frac{1}{2}\kappa^2}}{1 - \frac{\tau^2}{2N}} \underset{N\to\infty}{\cong} \boxed{\frac{\kappa \pm \frac{\tau}{\sqrt{N}}\sqrt{1+\frac{1}{2}\kappa^2}}{1 - \frac{\tau^2}{2N}}} \cong \frac{\kappa}{1 \mp \frac{\tau}{\sqrt{2N}}}$$

de 2 asymptoten

Kans op fout van eerste soort is $1-\alpha$ ($=0{,}05$ in schets)

$\tau = 1{,}645$ " "

Uitschieters niveau $U = \bar{x} + k\cdot s$ met

$k = k^+$ bij $\beta \cong 0.05/N$; exact $\beta = 1 - (\delta)^{1/N}$ → κ uit

$\mathbb{P}[\underline{t} \geq \kappa \mid 0;1] = \beta$. Deze κ bepaalt k, uit (2)

$P = \mathbb{P}[\underline{t} \geq \kappa \mid 0;1]$ $\qquad \delta = 0.95$

Fig.2

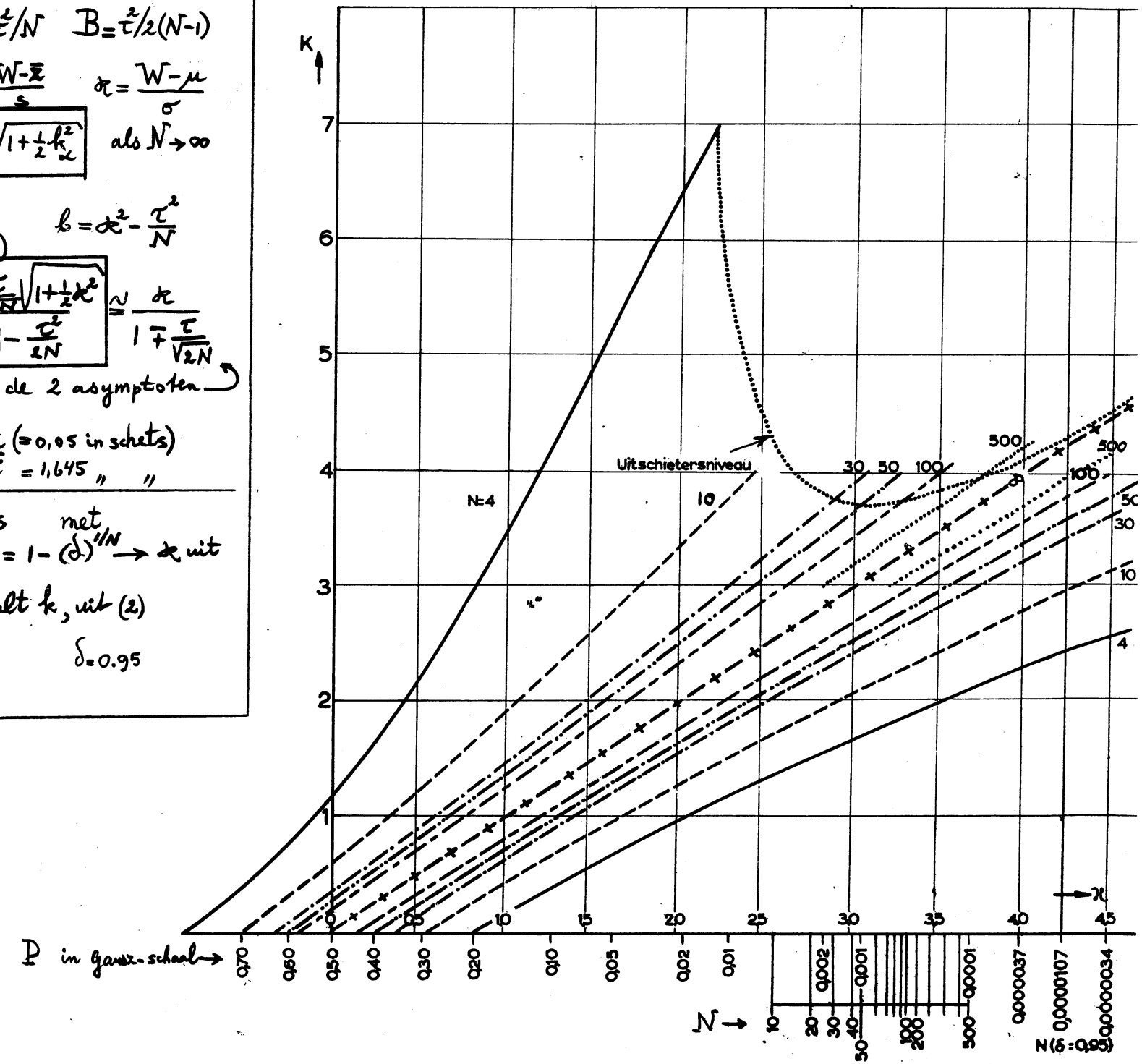


Fig. 3 b

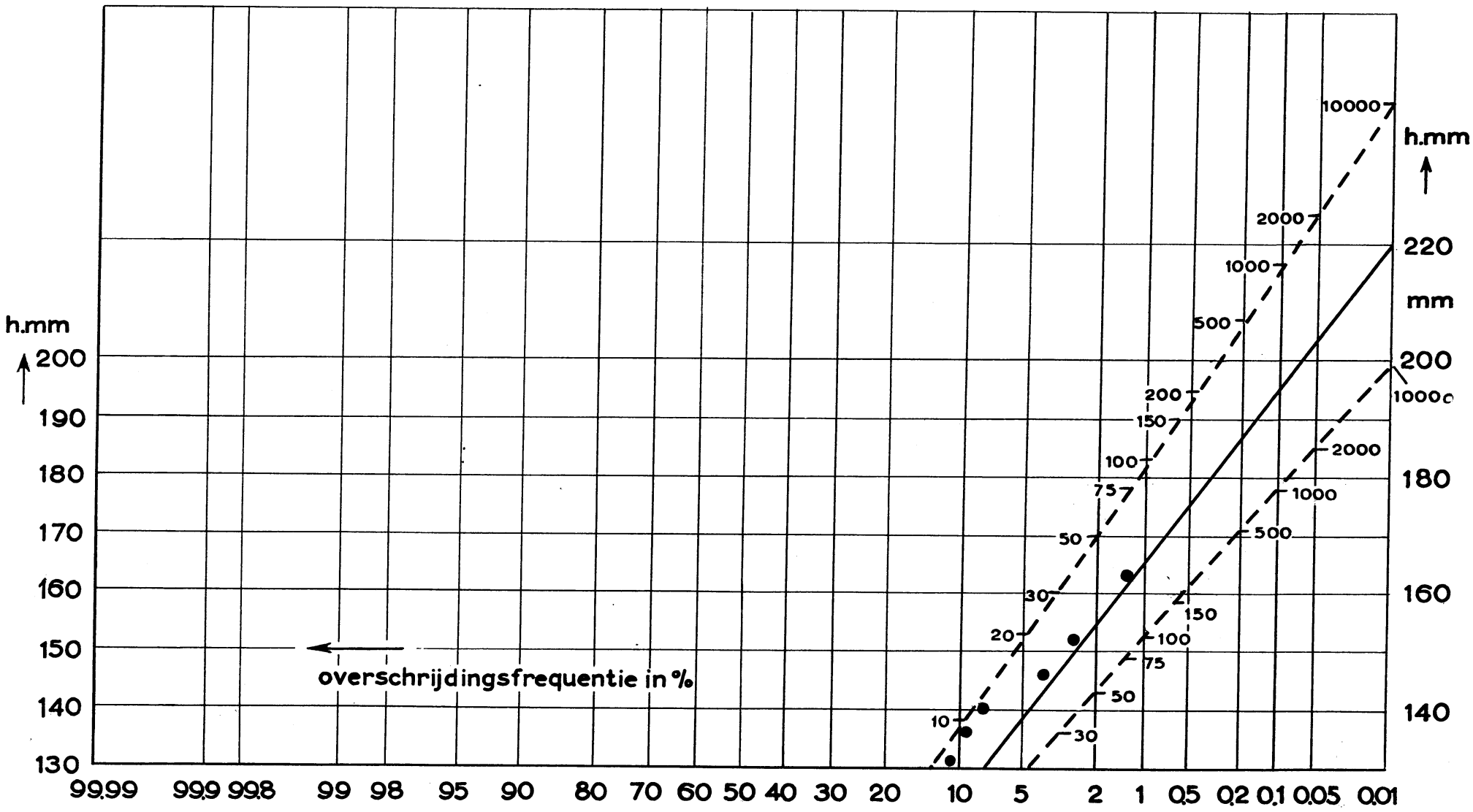
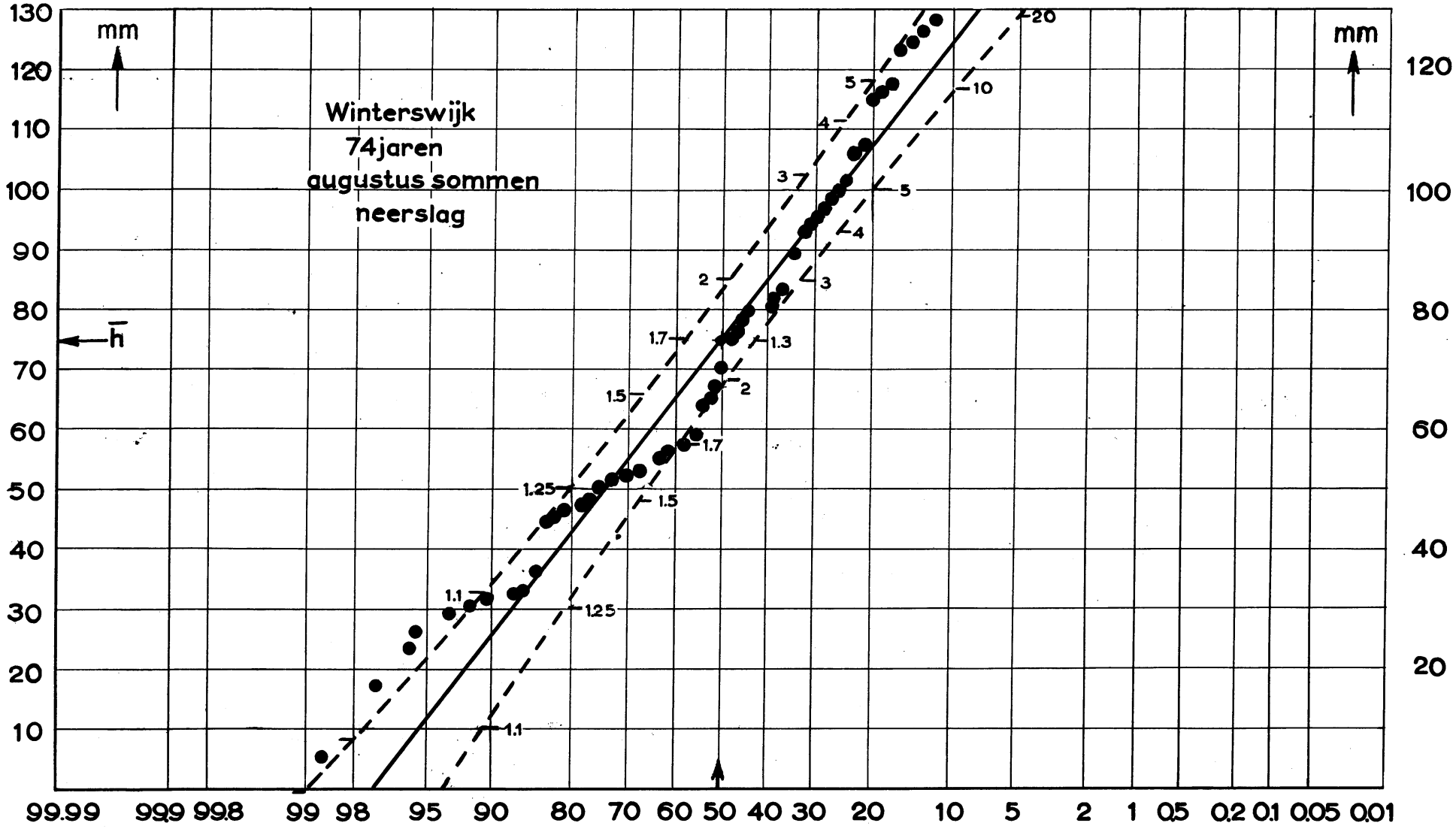
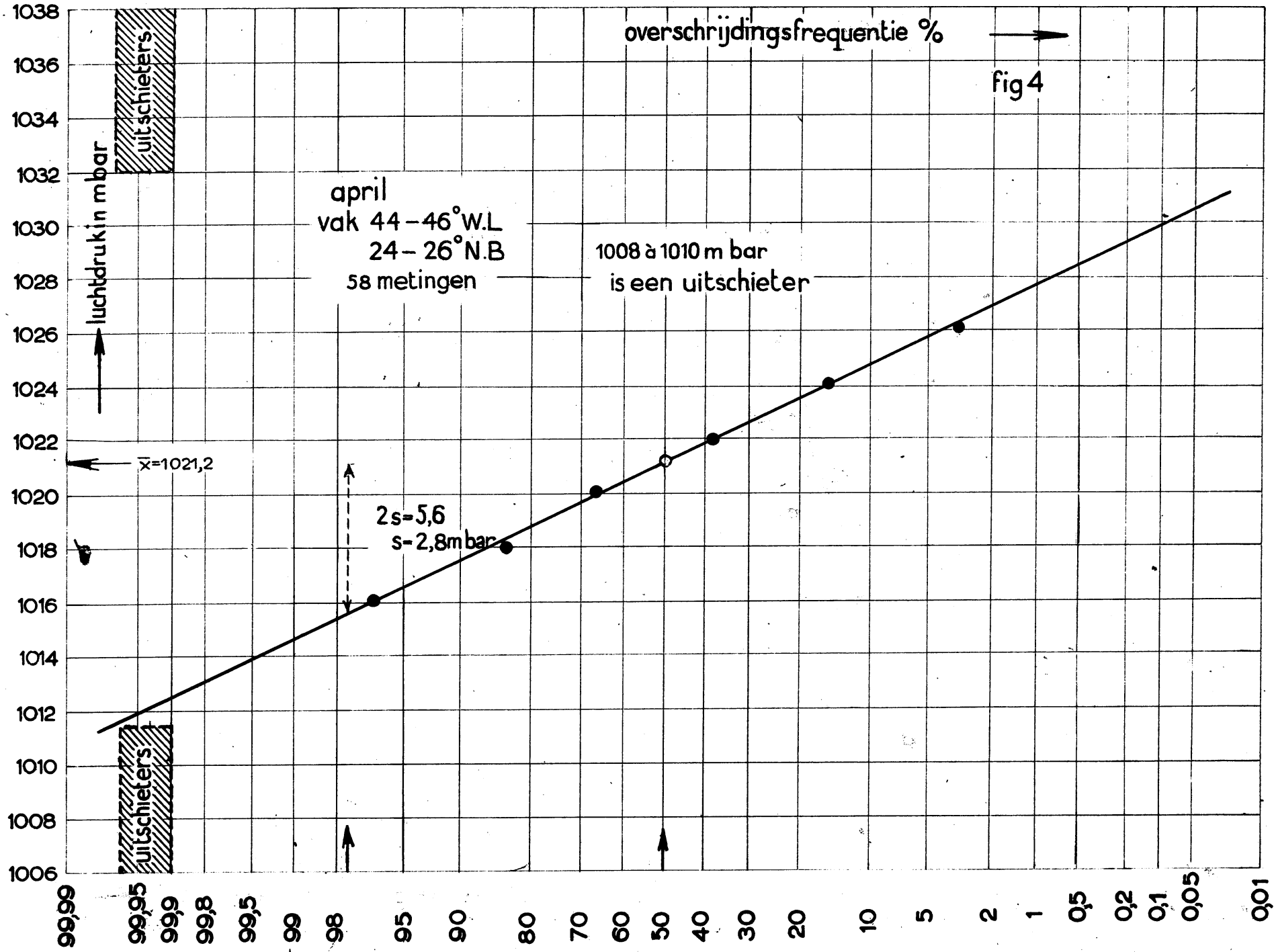
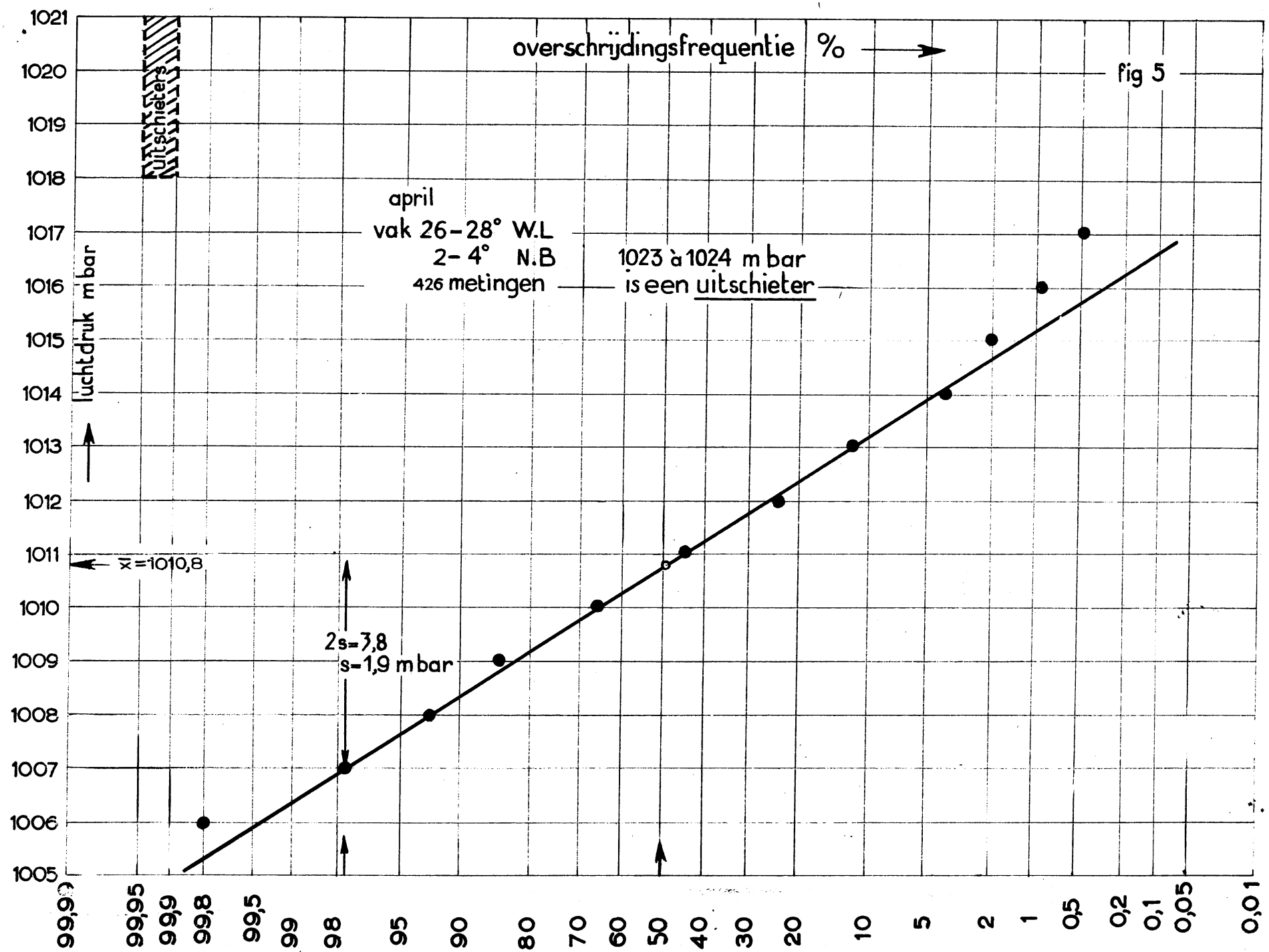
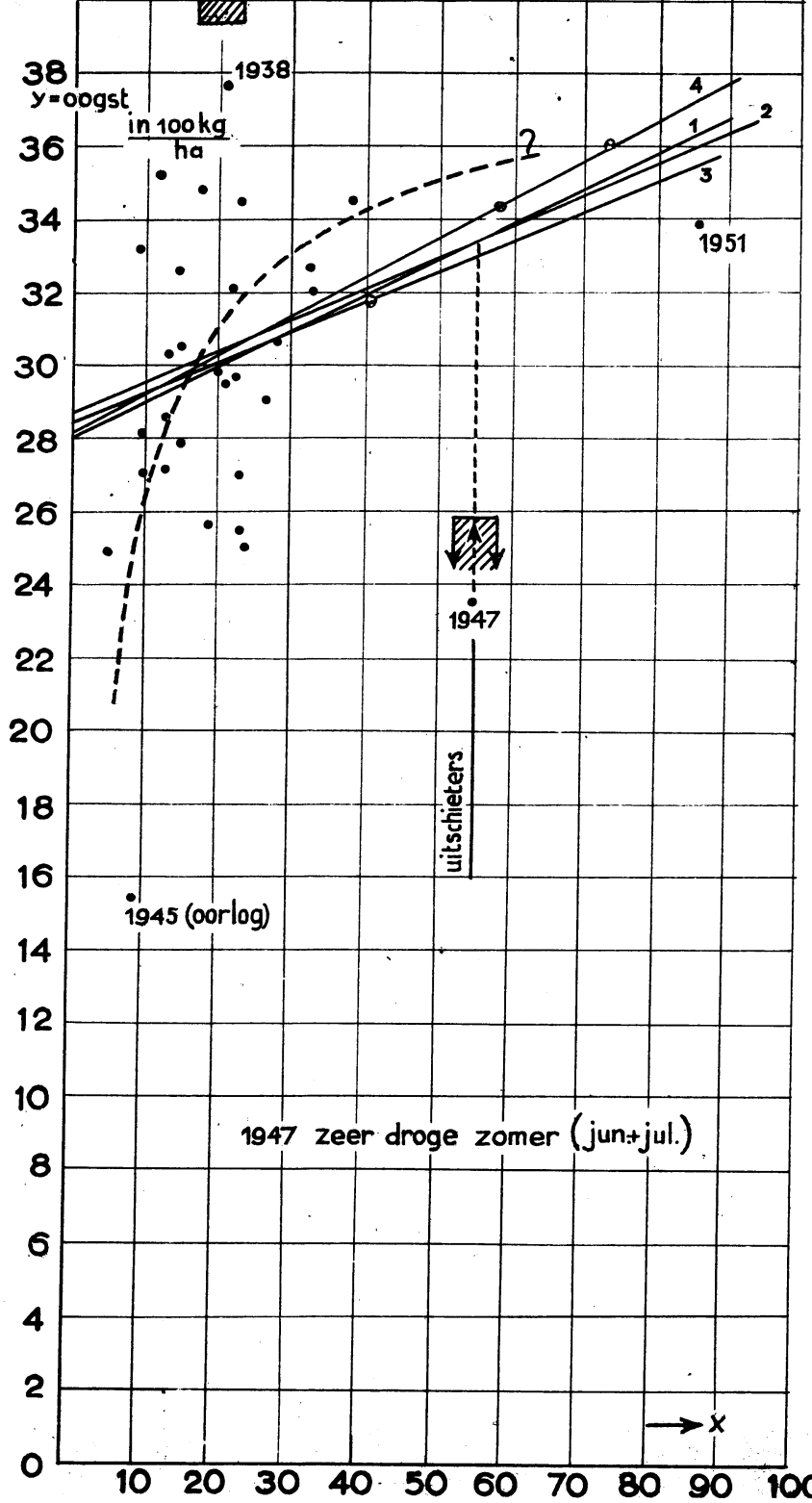

h.mm
200
190
180
170
160
150
140
130
99.99 99.9 99.8 99 98 95 90 80 70 60 50 40 30 20 10 5 2 1 0.5 0.2 0.1 0.05 0.01
overschrijdingsfrequentie in %
10000
2000
1000
500
200
150
100
75
50
30
20
10
h.mm
220
mm
200
180
160
140
1000
2000
1000
500
150
100
75
50
30

fig.3[a]


mm
mm
Winterswijk
74jaren
augustus sommen
neerslag
$\bar{h}$
130
120
110
100
90
80
70
60
50
40
30
20
10
99.99 99.9 99.8 99 98 95 90 80 70 60 50 40 30 20 10 5 2 1 0.5 0.2 0.1 0.05 0.01
20
10
5
4
3
2
1.7
1.5
1.25
1.1
1.3
1.7
1.5
1.25
1.1


overschrijdingsfrequentie %
fig 4
luchtdruk in m bar
uitschieters
uitschieters
april
vak 44 – 46° W.L
24 – 26° N.B
58 metingen
1008 à 1010 m bar
is een uitschieter
x̄=1021,2
2s=5,6
s=2,8m bar
1038
1036
1034
1032
1030
1028
1026
1024
1022
1020
1018
1016
1014
1012
1010
1008
1006
99,99
99,95
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
0,05
0,01


overschrijdingsfrequentie %
fig 5
uitschieters
luchtdruk m bar
april
vak 26-28° W.L
2-4° N.B
426 metingen
1023 à 1024 m bar
is een uitschieter
x̄=1010,8
2s=3,8
s=1,9 m bar
1021
1020
1019
1018
1017
1016
1015
1014
1013
1012
1011
1010
1009
1008
1007
1006
1005
99,9
99,95
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
0,05
0,01



$$y-\bar{y}=b(x-\bar{x}) \qquad b=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} \qquad r=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}} \qquad \sigma_r=\frac{1-r^2}{\sqrt{N}} \qquad \sigma_\varepsilon=\sigma_y\sqrt{1-r^2} \qquad \sigma_b\approx\frac{\sigma_\varepsilon}{\sigma_x\sqrt{N}}$$

1945 steeds verwaarlozen.

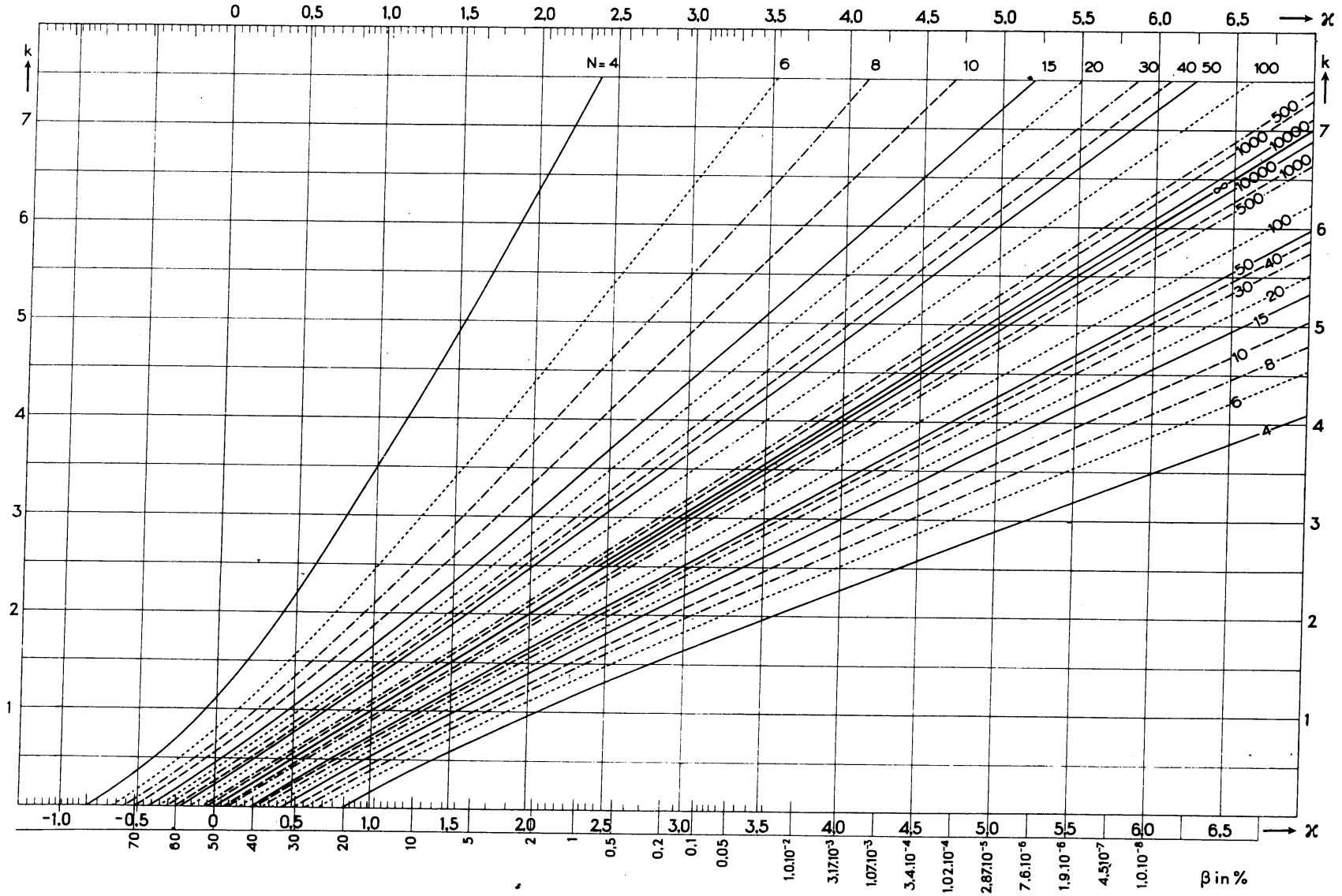
indien men lineaire regressie zou eisen		rechte $y=bx+a$	$r=$ c.c.	σ_b	σ_r	uitschieter	σ_ε	σ_x	σ_y
1938, 1947, 1951 meenemen 1 $N=32$		$y=0.097x+28.0$	0.45	0.11	0.14		3.58	5.85	4.60
1947 weglaten 2 $N=31$		$0.085x+28.7$	0.49	0.16	0.14	<25.8 1947 is uitschieter	2.71	5.71	3.11
1938, 1947 weglaten 3 $N=30$		$0.083x+28.4$	0.50	0.10	0.14	>39.4 1938 is geen uitschieter	3.12	5.84	3.6
1947, 1951 weglaten 4 $N=30$		$0.106x+28.2$	0.48	0.11	0.14		2.77	4.54	3.16

fig.6

NOMOGRAM

NOMOGRAM voor het aflezen van

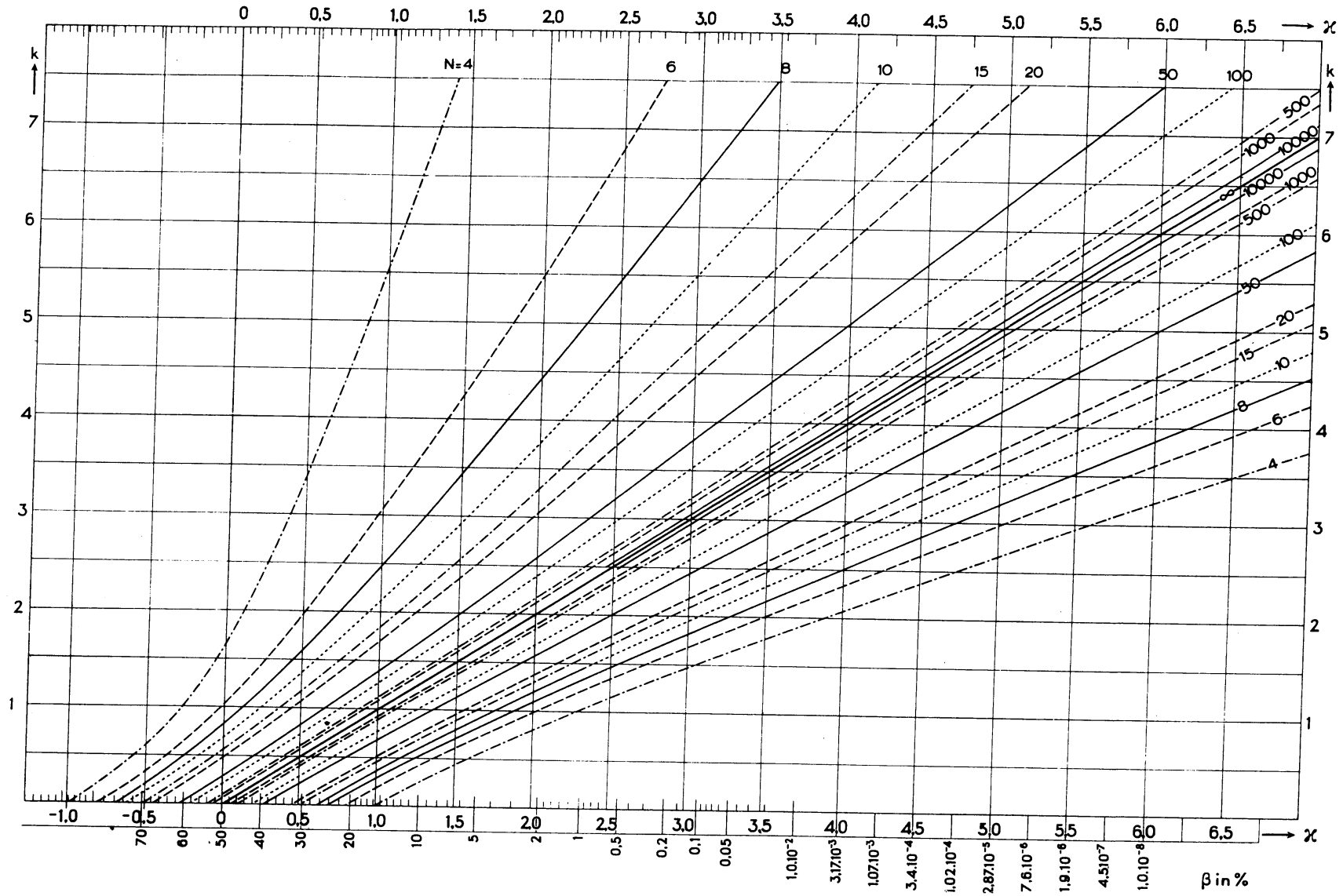
één-en tweezijdig begrensde betrouwbaarheidsintervallen voor β en χ, met $\alpha = 0.05$



NOMOGRAM

voor het aflezen van

één-en tweezijdig begrensde betrouwbaarheidsintervallen voor β en χ, met $\alpha = 0.025$



NOMOGRAM

voor het aflezen van

één-en tweezijdig begrensde betrouwbaarheidsintervallen voor β en χ, met α =0.005