

NN31545.0228

CULTUURTECHNIEK EN WATERHUISHOUDING
No.228 d.d.12 december 1963

Het gebruik van een betrouwbaarheidsinterval bij
cumulatieve frequentieverdelingen

J.B.H.M. van Gils

BIBLIOTHEEK DE HAAG

Droevendaalsesteeg 3a
Postbus 241
6700 AE Wageningen

Inleiding

Van een reeks waarnemingsuitkomsten, bijvoorbeeld de neerslag op een bepaalde dag, het aantal voertuigen dat per dag een bepaald punt passeert en dergelijke, kan een frequentieverdeling worden samengesteld, die hier met empirische verdeling zal worden aangeduid.

Uit een empirische frequentieverdeling kan de over- of onderschrijdingsfrequentie, waarmee een bepaalde waarde is opgetreden, worden afgelezen.

Indien het patroon van de verdeling zich in de toekomst niet wijzigt, kan de empirische verdeling als schatting voor de kansverdeling worden gebruikt. Hieruit kunnen de kansen worden afgelezen, waarmee een bepaald bedrag zal worden over- of onderschreden.

Daar de empirische verdeling slechts uit een eindig aantal gegevens is samengesteld, zal bij de schatting van de kansverdeling een meer of minder grote fout optreden. Naarmate het aantal gegevens (n) van de empirische verdeling groter is, zal ook de schatting van de werkelijke kansverdeling zekerder zijn. De kans, dat een overschrijdingskans precies gelijk is aan de overschrijdingsfrequentie is nul, immers de kans, dat een geschatte waarde precies gelijk is aan de te schatten waarde is nul. De kans, dat de overschrijdingskans binnen een interval rond de empirische verdeling valt is echter groter dan nul. De onzekerheid, die ontstaat door met de empirische verdeling in plaats van met de kansverdeling te werken wordt daarom door een interval symmetrisch rond de empirische frequentieverdeling weergegeven.

Naarmate het risico (α), dat de kans aangeeft waarmee de werkelijke kansverdeling nog buiten het interval zal liggen, kleiner wordt gekozen zal het interval breder worden. De breedte van het interval is afhankelijk van het aantal gegevens (n) van de empirische verdeling en het risicopercentage (α), dat de onderzoeker zelf vaststelt.

In plaats van het risicopercentage (α) kan ook de betrouwbaarheidscoëfficiënt ($1-\alpha$) worden gebruikt.

195/1263/25



0000 0672 2769

1785282

Om de cumulatieve frequentiecurve kan dus een betrouwbaarheidsinterval met aan weerszijden van de curve een breedte d_α worden geconstrueerd voor een betrouwbaarheidscoëfficiënt $(1-\alpha)\%$, dat wil zeggen: de werkelijke kansverdeling zal met een kans $(1-\alpha)\%$ binnen de aangegeven grenzen liggen.

Voor een cumulatieve frequentiecurve berekend volgens

$$F_{\leq} = 100 \cdot \frac{i}{n} \%$$

(i is het aantal gegevens, dat kleiner of gelijk aan een zekere waarde is) gelden de betrouwbaarheidsgrenzen $(F_{\leq} + d_\alpha)$ en $(F_{\leq} - d_\alpha)$. Deze grenzen moeten worden gemeten langs de "kansschaal". Ze worden voor ieder punt van de frequentiecurve uitgezet. Voor de breedte d_α geldt (DRION, 1952):

$$d_\alpha = \frac{z}{\sqrt{n}} \cdot 100\% \quad (1)$$

Daar de frequentie niet in deze formule voorkomt is d_α dus voor de gehele verdeling constant, doch indien deze is uitgezet op een "kansschaal" zijn de betrouwbaarheidsgrenzen niet evenwijdig aan de frequentiecurve.

Het verband tussen α en z wordt gegeven met de formule:

$$\alpha = 2 \sum_{a=1}^{\infty} (-1)^{a-1} \frac{e^{-2a^2 z^2}}{a^2} \quad (2)$$

In de volgende tabel zijn enkele berekende waarden van d_α gegeven, getabelleerd met hun bijbehorende α , z en n -waarden. De tabel is verkregen door α aan te nemen en via formule (2) de bijbehorende z -waarde te bepalen, waardoor voor gegeven n -waarde uit formule (1) d_α volgt.

Tabel 1 Het betrouwbaarheidsinterval d_α voor verschillende steekproefgrootten n en verschillende risicopercentsages α

α	z	d_α		
		$n = 10$	$n = 100$	$n = 1000$
50 %	0,828	26,18 %	8,28 %	2,62 %
20 %	1,073	33,93 %	10,73 %	3,39 %
10 %	1,224	38,71 %	12,24 %	3,87 %
5 %	1,358	42,94 %	13,58 %	4,29 %
2 %	1,517	47,97 %	15,17 %	4,80 %
1 %	1,625	51,39 %	16,25 %	5,14 %
0,5 %	1,731	54,74 %	17,31 %	5,47 %

Voor een uitgebreide tabel van z-waarden zij verwezen naar BRION, 1952, pag. 148 en FISZ, 1962 pag. 510.

In figuur 1 is het verband tussen d_α en n grafisch weergegeven. In de praktijk wordt voor α veelal 5% gekozen.

Het bovenstaande is meer uitvoerig behandeld in I.C.W.-nota no.187 (STOL, 1963).

In deze nota zullen enkele praktische toepassingen worden besproken van het toetsen van frequentieverdelingen.

Het tekenen van een betrouwbaarheidsinterval

Ongeacht of de punten van de cumulatieve frequentiecurve worden berekend en uitgezet volgens de formules (zie nota no.187, pag. 10):

$$F_{\leq} = \frac{i - 0,3}{n + 0,4} \cdot 100\% \quad (\text{Median})$$

$$F_{\leq} = \frac{i}{n + 1} \cdot 100\% \quad (\text{Mean})$$

$$F_{\leq} = \frac{i}{n} \cdot 100\% \quad (\text{FQ})$$

blijven de grenzen van het betrouwbaarheidsinterval, waarmee wordt getoetst:

$$\left(\frac{i}{n} \cdot 100 + d_\alpha\right)\% \text{ en } \left(\frac{i}{n} \cdot 100 - d_\alpha\right)\%$$

daar voor het afleiden van het interval van het frequentiequotiënt FQ is uitgegaan.

Ongeacht de drie berekeningswijzen van de cumulatieve frequentie geldt ook met een kans van $\alpha\%$ de uitspraak, dat de werkelijke kansverdeling er toch buiten ligt.

De kans welke men loopt, dat een curve binnen het interval toch niet de juiste is, is bij deze wijze van toetsen niet goed vast te stellen.

In figuur 2 is het betrouwbaarheidsinterval om een neerslagverdeling uitgezet. Voor 5 mm neerslag per etmaal staat het interval op de "kans-schaal" aangegeven. Vanwege het nog betrekkelijk geringe aantal punten ($n = 310$) kan de werkelijke kansverdeling dus nog vrij veel van de frequentieverdeling afwijken. Hieruit blijkt, dat men voor het trekken van de geschatte kansverdelingscurve in principe een zeer grote vrijheid

heeft, doch de onderzoeker zal van deze vrijheid slechts onder goed gemotiveerde voorwaarden gebruik mogen maken (zie pag. 6).

Aangezien de waarden 0% en 100% op de "kansschaal" respectievelijk naar $-\infty$ en $+\infty$ verdwijnen, worden die betrouwbaarheidsgrenzen, die gelijk aan of kleiner zouden zijn dan 0% en die, welke gelijk aan of groter zouden zijn dan 100% ook in $-\infty$ of $+\infty$ gedacht. De verbindingslijnen naar deze punten lopen dus altijd evenwijdig aan de "kansschaal".

In figuur 3 zijn de betrouwbaarheidsgrenzen voor verschillende n -waarden om een hypothetische verdelingscurve uitgezet. De punten van de hypothetische frequentiecurve zijn eenvoudigheidshalve op een 45° -lijn gekozen. Indien een empirische curve op transparant papier wordt uitgezet, kan het interval per punt worden overgenomen. Het transparant papier wordt hiertoe zodanig over de figuur verschoven, dat de 50%-kanslijn op beide figuren steeds blijven samenvallen en een punt van de empirische curve op de hypothetische curve komt te liggen. Voor dit punt worden nu de grenzen van het betrouwbaarheidsinterval in verticale richting gemeten voor de bijbehorende waarde van n op het transparant papier overgenomen. Deze werkwijze mag uiteraard alleen worden toegepast bij empirische curven berekend volgens de formule

$$F_{\frac{1}{n}} = \frac{1}{n} \cdot 100\%$$

en uitgezet op een identieke kansschaal als in figuur 3.

De grenslijnen in figuur 3 geven de uiteinden van de betrouwbaarheidsintervallen en tevens de uiteinden van de hypothetische curven welke in verband met het aantal gegevens optreden aan.

Vergelijking van frequentieverdelingen

Een vergelijking van verdelingen door middel van het betrouwbaarheidsinterval kan alleen via een hypothese worden uitgevoerd. Als hypothese wordt in de regel de veronderstelling gebruikt, dat twee frequentieverdelingen feitelijk niet van elkaar verschillen. Getoetst moet nu worden of de verschillen die worden gevonden systematische verschillen zijn of slechts toevallige afwijkingen. Enige overeenkomst tussen verdelingen kan niet worden bewezen. Wel kunnen verschillen van statistische aard al dan niet worden aangetoond. Indien geen verschillen worden aange-

toond, Kan de hypothese juist zijn. Worden wel verschillen gevonden, dan dient de hypothese te worden verworpen.

Men kan bijvoorbeeld als hypothese aannemen, dat twee geologische grondmonsters eenzelfde ontstaanswijze hebben, dat wil zeggen tot eenzelfde populatie behoren. Men bepaalt de korrelgrootte-verdelingen van de monsters en vindt statistische verschillen tussen de twee verdelingen. Met een risico $\alpha\%$ kan dan worden gezegd, dat de hypothese onjuist is.

Statistisch verschilt een empirische verdeling van een andere, behoudens een risico $\alpha\%$, zodra deze reeds op een enkel punt buiten de betrouwbaarheidsgrenzen van de andere valt. Over de aard van de verschillen zelf kan niets worden gezegd. Deze kan bijvoorbeeld zijn een verschil in gemiddelde, een verschil in spreiding, een verschil in vorm van de verdeling, enz. Valt deze verdeling geheel binnen het betrouwbaarheidsinterval van de andere, dan kunnen statistisch geen verschillen worden aangetoond.

Een voorbeeld geeft figuur 4. De verdelingen van de dagneerslag-sommen van januari en mei verschillen duidelijk. De kansverdelingen van januari en mei zijn nagenoeg niet dezelfde. Indien de neerslagwaarden voor gelijke frequenties tegen de maanden van het jaar worden uitgezet, zouden deze verschillen nader kunnen worden onderzocht, bijvoorbeeld op het optreden van een seizoenbeweging.

Ook kan men het verschil tussen gelijkwaardige punten van de twee empirische verdelingen vergelijken met het betrouwbaarheidsinterval. Daartoe worden op lineaire schalen de percentages van de twee verdelingen voor gelijke abscis-waarden tegen elkaar uitgezet (zie tabel bij fig.5) en vergeleken met de hypothetische lijn, (een 45° -lijn in fig.5), welke zou ontstaan indien beide verdelingen volkomen aan elkaar gelijk waren. De betrouwbaarheidsgrenzen uitgezet om de hypothetische lijn, geven op deze schalen evenwijdige rechte lijnen. Zijn de verschillen kleiner dan d_α dan kunnen ze als toevallige afwijkingen worden aangemerkt. Is dit niet zo dan moeten systematische verschillen tussen beide verdelingen worden aangenomen. Deze methode is alleen bruikbaar, indien beide verdelingen hetzelfde aantal gegevens hebben. Anders zou men een andere methode moeten toepassen (zie I.C.W.-nota no. 187).

In figuur 5 is boven besproken methode toegepast op de elementaire dagverdelingen van 10 en 20 januari. Hier blijkt, dat de tijdstippen zo-

danig dicht bij elkaar zijn gekozen, dat geen verschillen tussen de verdelingen kunnen worden aangetoond. De tijdsfactor speelt dus een zeer geringe rol. In de praktijk worden de elementaire dagverdelingen samengenomen tot een gecombineerde dagverdeling voor een maand, zoals in de figuren 2 en 4. Het voordeel van deze samenvoeging is het grotere aantal waarnemingen in de verdeling waardoor de afwijking tot de werkelijke kansverdeling wordt beperkt. Een extreme verdeling blijkt die van 15 januari te zijn, getekend in figuur 5. Consequent moet de uitspraak volgen, dat de verdelingen van 10 en 15 januari van elkaar verschillen. In dit geval is het echter voor de hand liggend aan te nemen, dat we met een uitzondering te doen hebben. Als niet meer dan $\alpha\%$ van de uitspraken voor de elementaire dagverdelingen in januari deze uitzondering vormen, is de gecombineerde dagverdeling voor de maand januari als representatie voor de elementaire dagverdeling nog verantwoord.

Het gebruik van de vrijheid binnen het betrouwbaarheidsinterval

Meestal is de werkelijke kansverdeling niet bekend. Men tracht deze te benaderen met een geschatte kansverdeling.

Wel moet worden getracht de werkelijke kansverdeling zo dicht mogelijk te benaderen. De grootste kans daartoe heeft men indien de geschatte kansverdeling zoveel mogelijk aan de punten van de frequentieverdeling wordt aangepast.

Hierbij moet worden opgemerkt, dat een empirische verdeling berekend volgens:

$$F_{\leq}^i = \frac{i - 0,3}{n + 0,4} \cdot 100 \% \quad (\text{Median})$$

de werkelijke kansverdeling gemiddeld beter zal benaderen dan indien berekend volgens

$$F_{\leq}^i = \frac{i}{n + 1} \cdot 100 \% \quad (\text{Mean})$$

en deze weer een betere benadering is dan indien de frequentieverdeling is berekend volgens

$$F_{\leq}^i = \frac{i}{n} \cdot 100 \% \quad (\text{FQ})$$

(zie I.C.W.-nota no. 187)

Binnen het betrouwbaarheidsinterval heeft men in principe echter een grotere vrijheid om de geschatte kansverdeling te tekenen. Statistisch kan namelijk niet worden aangetoond, dat de geschatte kansverdeling dan niet samenvalt met de werkelijke kansverdeling.

Bepaalde afwijkingen van de geschatte kansverdeling tot de empirische frequentieverdeling zijn daarom wel te verdedigen, indien men bepaalde hypothesen invoert. In wezen wordt dan aangenomen, dat er inderdaad afwijkingen van de empirische verdeling ten opzichte van de werkelijke kansverdeling zijn. Zolang de afwijkingen echter binnen het interval liggen zijn ze nog van toevallige aard. Enige gebruikelijke hypothesen zijn:

1. Men neemt een bepaalde vorm van de werkelijke kanscurve aan (bijvoorbeeld de verdeling is normaal). Is de aangenomen vorm in een formule te vangen dan zal de geschatte kanscurve veelal via een berekening kunnen worden bepaald.
2. Men neemt aan, dat de werkelijke kanscurve "strak" verloopt. Door de punten van de empirische curve wordt een vloeiend verlopende strakke lijn als geschatte kanscurve getrokken.
3. Men neemt een onderlinge samenhang tussen een aantal verdelingen aan. Deze samenhang kan veelal in een figuur worden weergegeven (door bijvoorbeeld neerslag tegen de tijd uit te zetten). Met behulp van hypothesen 2 kan dan een grafische vereffening worden toegepast.

De steekproefgrootte

De steekproefgrootte wordt door de onderzoeker gekozen. Enerzijds zal hij de steekproefgrootte (n) om arbeidseconomische redenen beperken, anderzijds dient n zo groot te zijn, dat de empirische verdeling voldoende representatief is voor de kansverdeling.

Door een voorbeeld te nemen, waarbij de werkelijke kansverdeling bekend is, kunnen we een indruk krijgen over wat het effect van een steekproef kan zijn en in hoeverre de kansverdeling in de steekproef wordt teruggevonden.

Figuren 6, 7 en 8 geven voorbeelden van a-selecte steekproeven uit een standaard-normale verdeling (table 23, DIXON and MASSEY, 1951). Hier is de kansverdeling bekend, namelijk een normale verdeling met een spreiding gelijk aan één. Met geschikte schaaleenheden uitgezet, geeft deze

verdeling op een "kansschaal" een 45° -lijn. De steekproeven van figuren 6 en 7 zijn ook begrepen in het cijfermateriaal waaruit figuur 8 is ontstaan, de steekproef van figuur 6 komt ook voor in figuur 7.

Men krijgt nu een indruk hoe de frequenties met de steekproefgrootte tot de werkelijke kansverdeling naderen. De afwijkingen ten opzichte van de normale verdeling en de breedte van het betrouwbaarheidsinterval zijn bij kleine steekproeven groot. De beoordeling van de resultaten van een kleine steekproef blijkt zeer moeilijk te zijn.

De steekproefgrootte dient dus voldoende groot te worden gekozen. Als men het betrouwbaarheidsinterval als maatstaf neemt, kan bij een gekozen α en d_α de steekproefgrootte worden opgegeven. Men stelt dan de eis, dat met een zekerheid van $(1 - \alpha)\%$ de werkelijke kansen moeten vallen binnen een interval d_α aan weerszijden van de bijbehorende frequenties.

Men stelt bijvoorbeeld, dat α niet groter mag zijn dan 5% en d_α niet groter dan 7,5%. Voor $\alpha = 5\%$ is volgens tabel 1 $z = 1,358$. Uit de formule

$$7,5 > d_\alpha = \frac{1,358}{\sqrt{n}} \cdot 100\%$$

volgt dan, dat $n > 327$ moet zijn.

In figuur 3 interpoleert men, dat $n > 300$ moet zijn. Een kleine afwijking van n geeft slechts een geringe, in figuur 3 onzichtbare, afwijking van d_α . De eis voor de maximale grootte van d_α is in de praktijk meestal een raming. Een grote nauwkeurigheid ten aanzien van de eis voor de steekproefgrootte heeft daarom weinig praktische waarde. In de praktijk is het daarom voldoende nauwkeurig figuur 3 te gebruiken voor het bepalen van de minimale steekproefgrootte.

Samenvatting

Het trekken van een steekproef uit een kansverdeling resulteert altijd in een afwijking van de steekproefverdeling ten opzichte van de kansverdeling. Iedere empirische verdeling heeft een afwijking ten opzichte van haar werkelijke kansverdeling.

De werkelijke kansen kunnen niet uit een empirische verdeling worden bepaald, doch wel worden geschat. Zoals bij iedere schatting is er een kans op fout schatten. Wel kan een interval van kansen worden gegeven, waarbinnen de werkelijke kans zal liggen met een kans van $\alpha\%$ echter, dat de werkelijke kans nog buiten het interval ligt. De grootte van dit interval hangt af van de steekproefgrootte en het gekozen risico α .

Indien het interval om ieder punt van de cumulatieve frequentiecurve wordt uitgezet, verkrijgt men een betrouwbaarheidsinterval, waarbuiten de werkelijke kansverdeling niet voorkomt met een kans van $\alpha\%$ dat dit toch zo is.

Statistisch verschilt een empirische verdeling van een andere, behoudens dus een risico $\alpha\%$, indien ze geheel of gedeeltelijk buiten de betrouwbaarheidsgrenzen van de andere valt, ofwel indien de afstand tussen gelijkwaardige punten van de beide verdelingen groter is dan het betrouwbaarheidsinterval.

De geschatte kanscurve moet binnen het betrouwbaarheidsgebied liggen. De grootste kans op een goede schatting heeft men, indien de kanscurve zoveel mogelijk aan de punten van de frequentiecurve wordt aangepast. Afwijkingen dienen met bepaalde hypothesen te worden gemotiveerd.

De beoordeling van de resultaten van kleine steekproeven is zeer moeilijk. De steekproefgrootte dient voldoende groot te worden gekozen. Door het te nemen risico en de grootte van het interval als voorwaarden te stellen wordt een minimum ois ten aanzien van de steekproefgrootte verkregen.

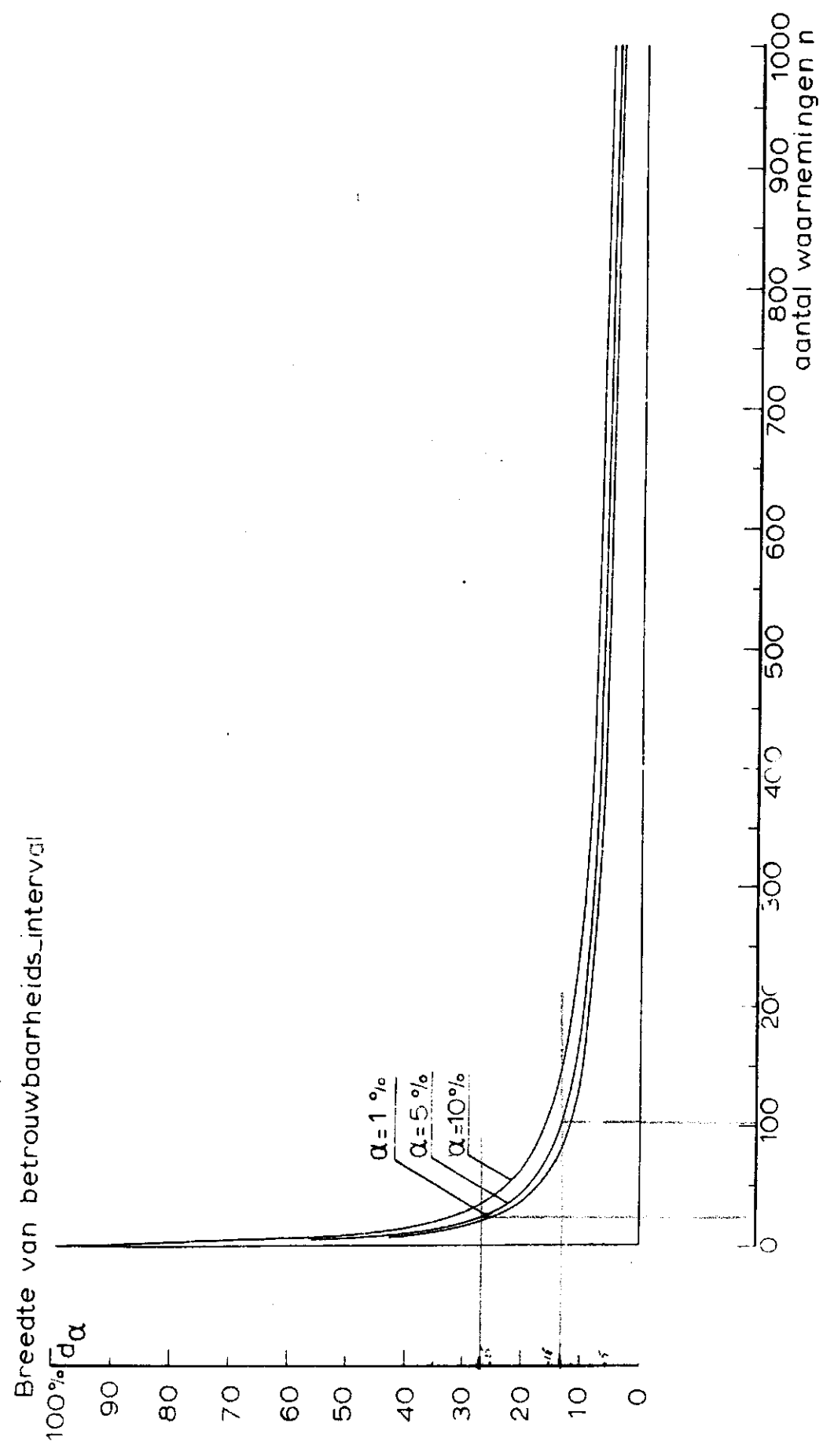
Literatuur

- DIXON, W.J. and F.J.MASSEY, 1951 - Introduction to statistical analysis. New-York (I.C.W. 11/34)
- DRION, E.F., 1952 - Cursus "Parametervrije methoden" VI: De mediaantoets en de toets van Smirnov.
Rapport 576 van het Mathematische Centrum te Amsterdam.
(I.C.W. 11/3)
- FISZ, M., 1962 - Wahrscheinlichkeitsrechnung und Mathematische Statistik. Berlin (I.C.W. 11/182)
- STOL, Ph.Th., 1963 - Een betrouwbaarheidsinterval voor frequentieverdelingen en frequentiequotiënten. (I.C.W.-nota no. 187)

Berekening van de cumulatieve frequentiepercentages van de dagverdeling van 20 januari voor
gelijke mm neerslag van de dagverdeling van 10 januari

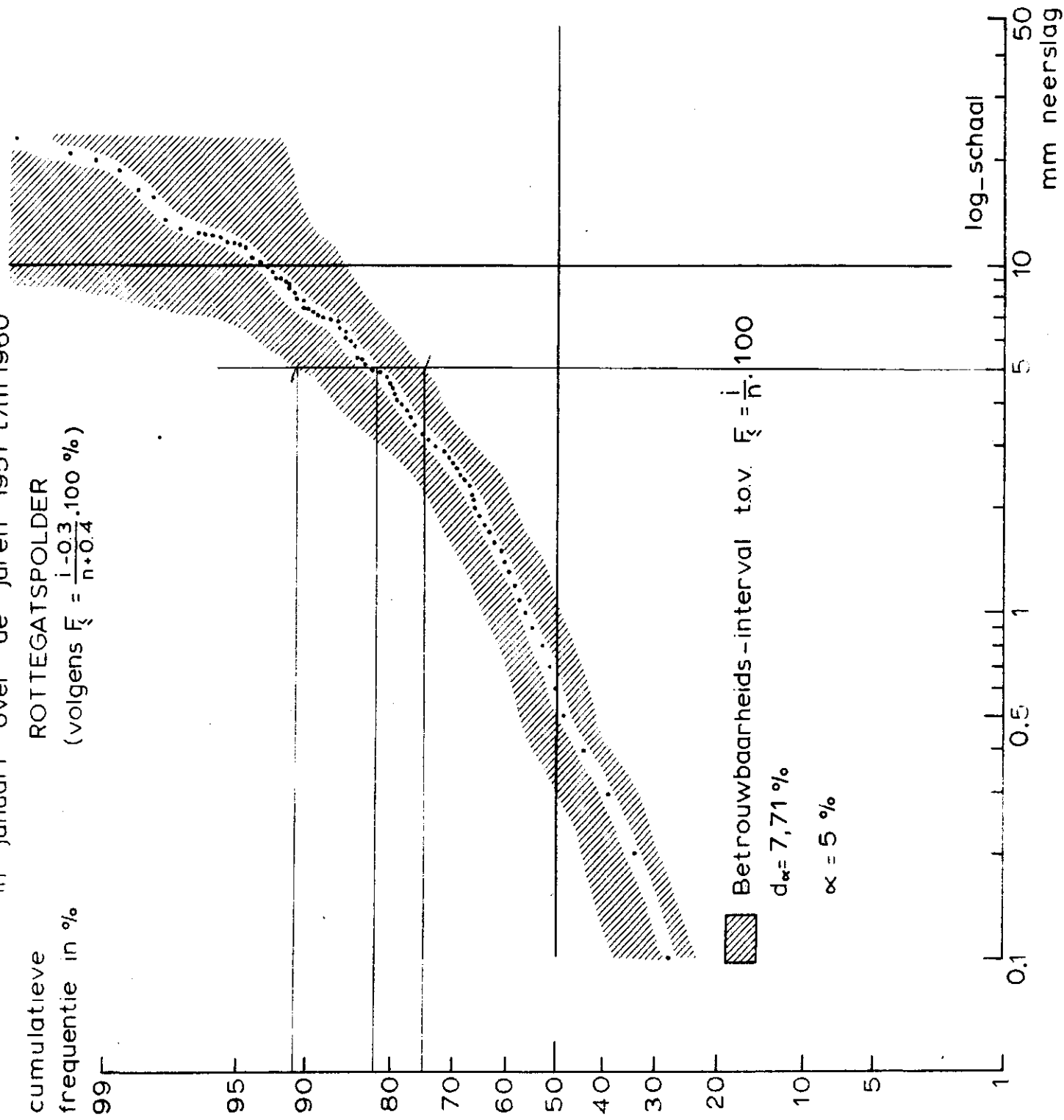
FIG.1

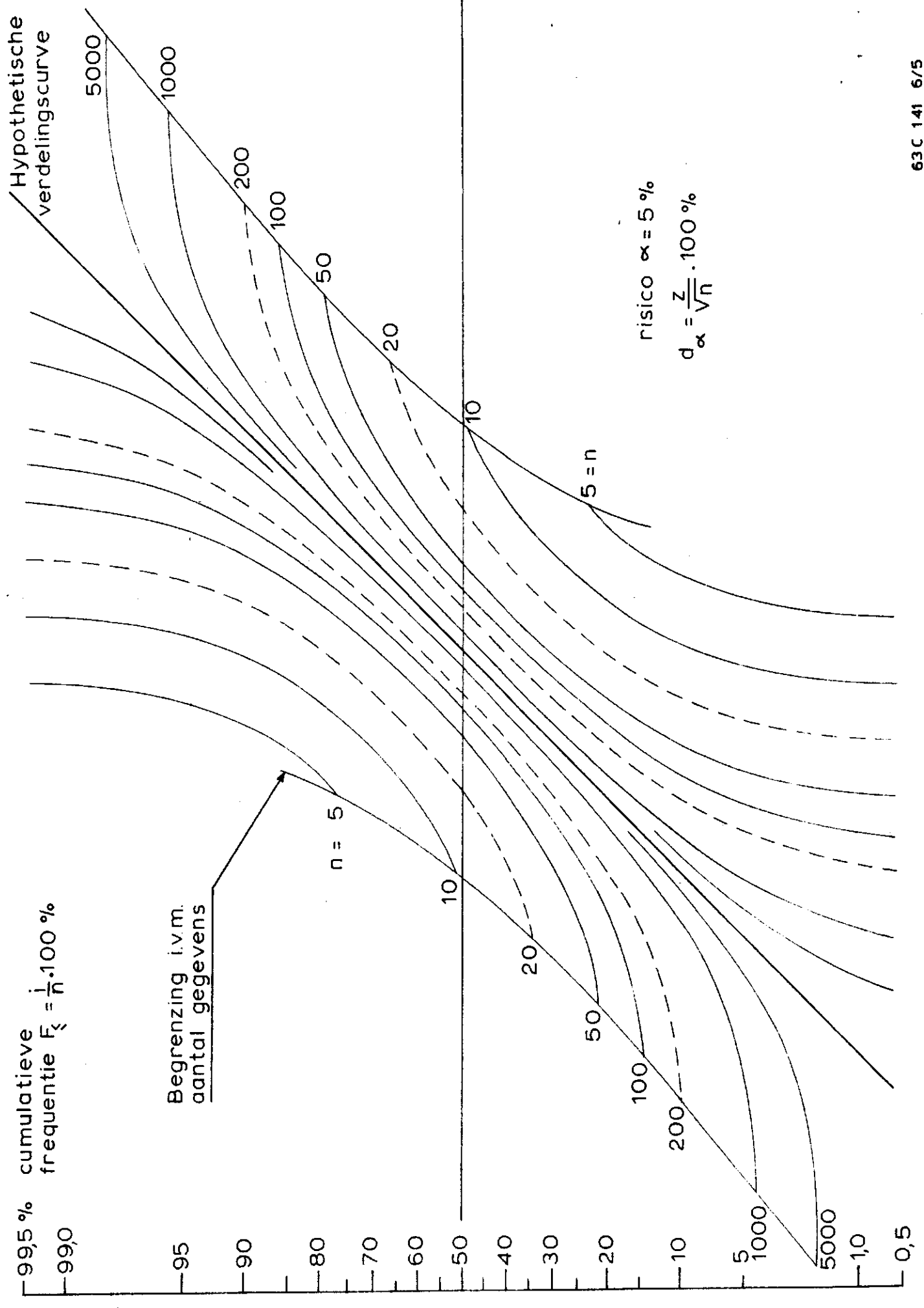
HET VERBAND TUSSEN DE BREEDTE VAN HET BETROUW-
BAARHEIDS_INTERVAL d_α EN HET AANTAL WAARNEMINGEN
 n , VOOR VERSCHILLENDE WAARDEN VAN HET RISICO α



Frequentieverdeling van de dagneerslagsommen
in januari over de jaren 1951 t/m 1960
ROTTEGATSPOLDER
(volgens $\bar{F}_i = \frac{i-0.3}{n+0.4} \cdot 100 \%$)

fig.2





Frequentieverdeling van de dagheerslagsommen
in jan.en mei over de jaren 1951 t/m 1960

fig. 4

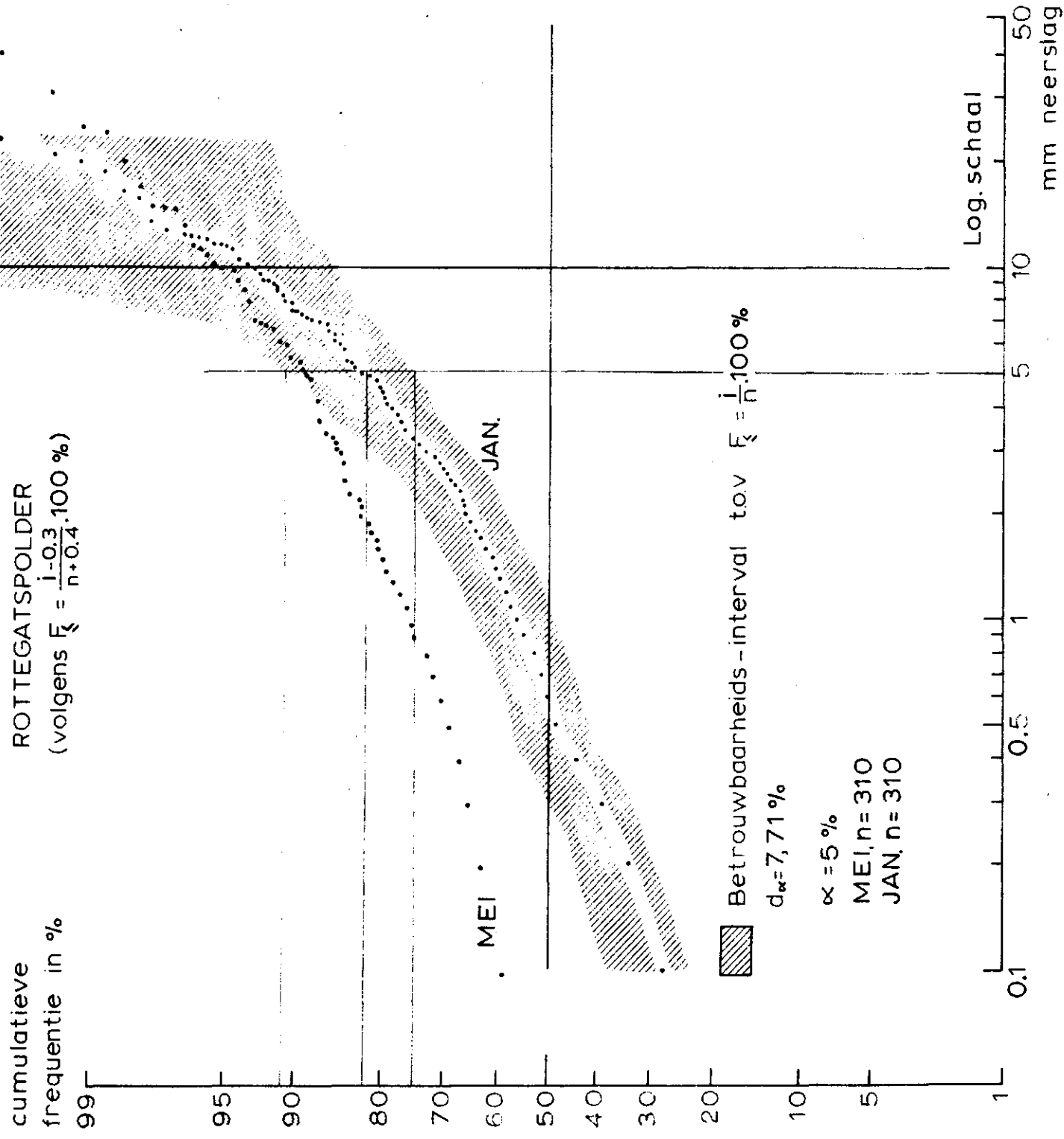


fig. 5

Lijn voor gelijke mm's neerslag van elementaire dagverdelingen
ROTTEGATSPOLDER

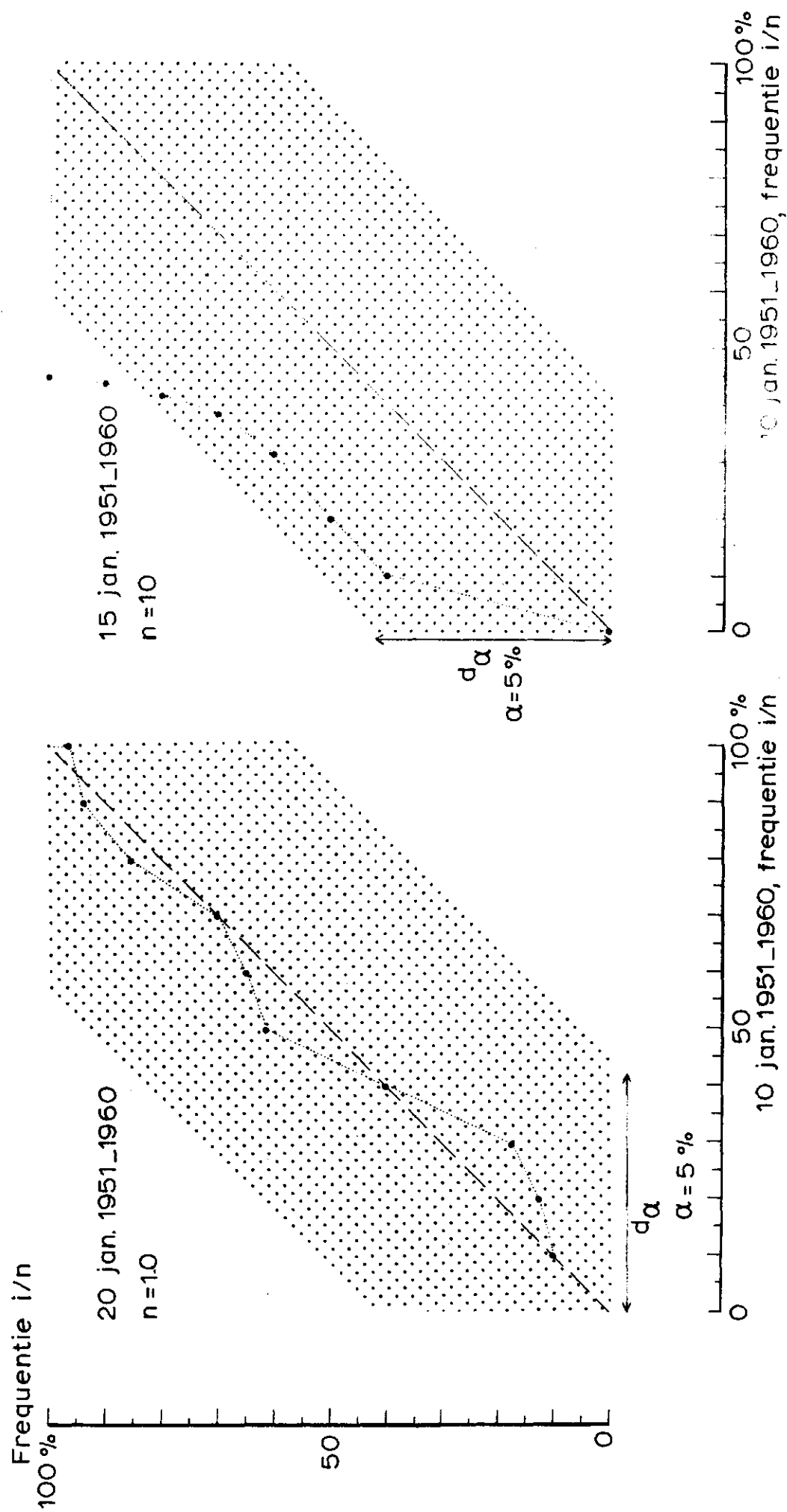
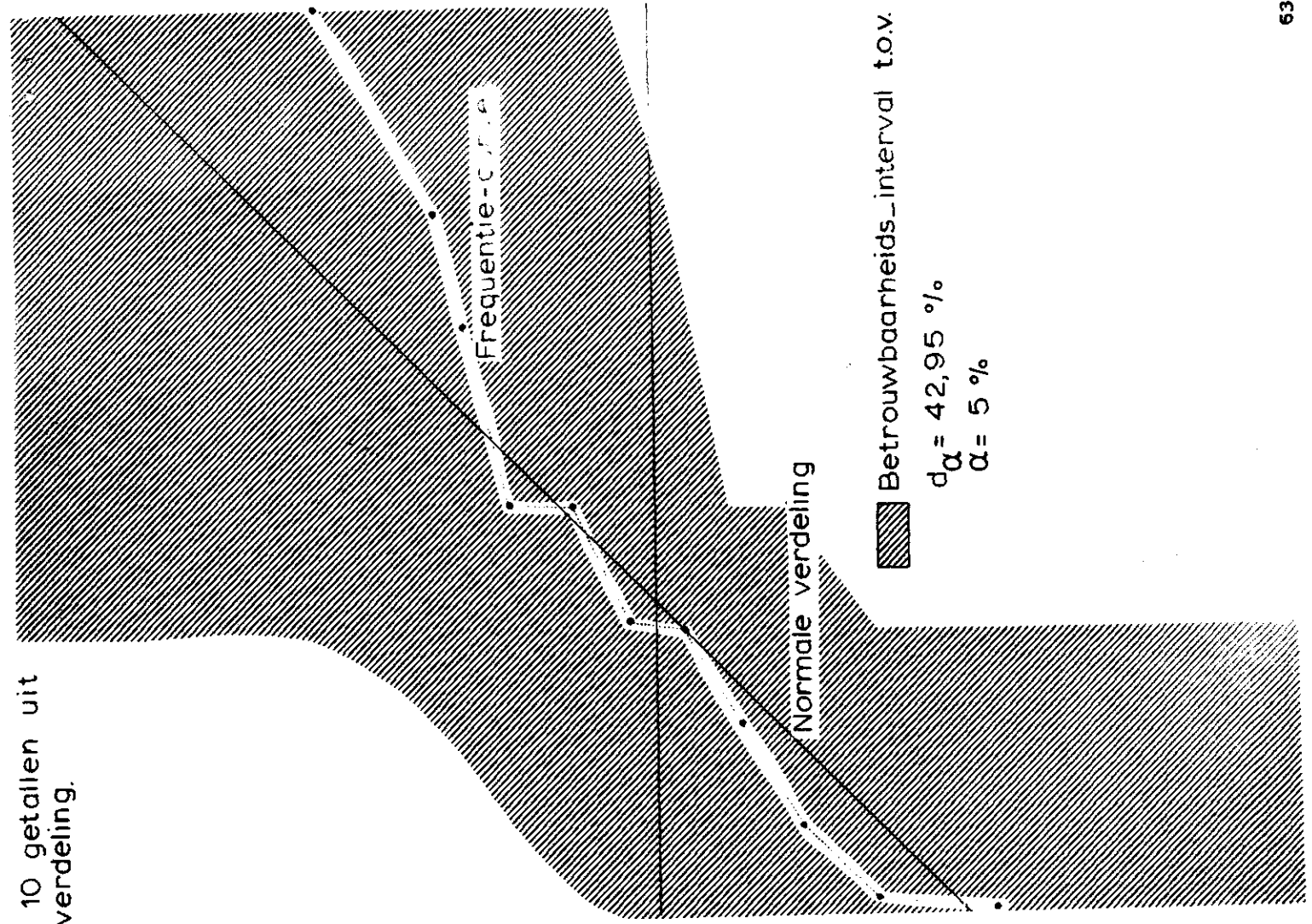
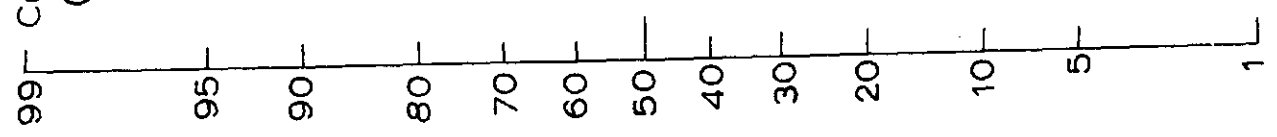


FIG. 6

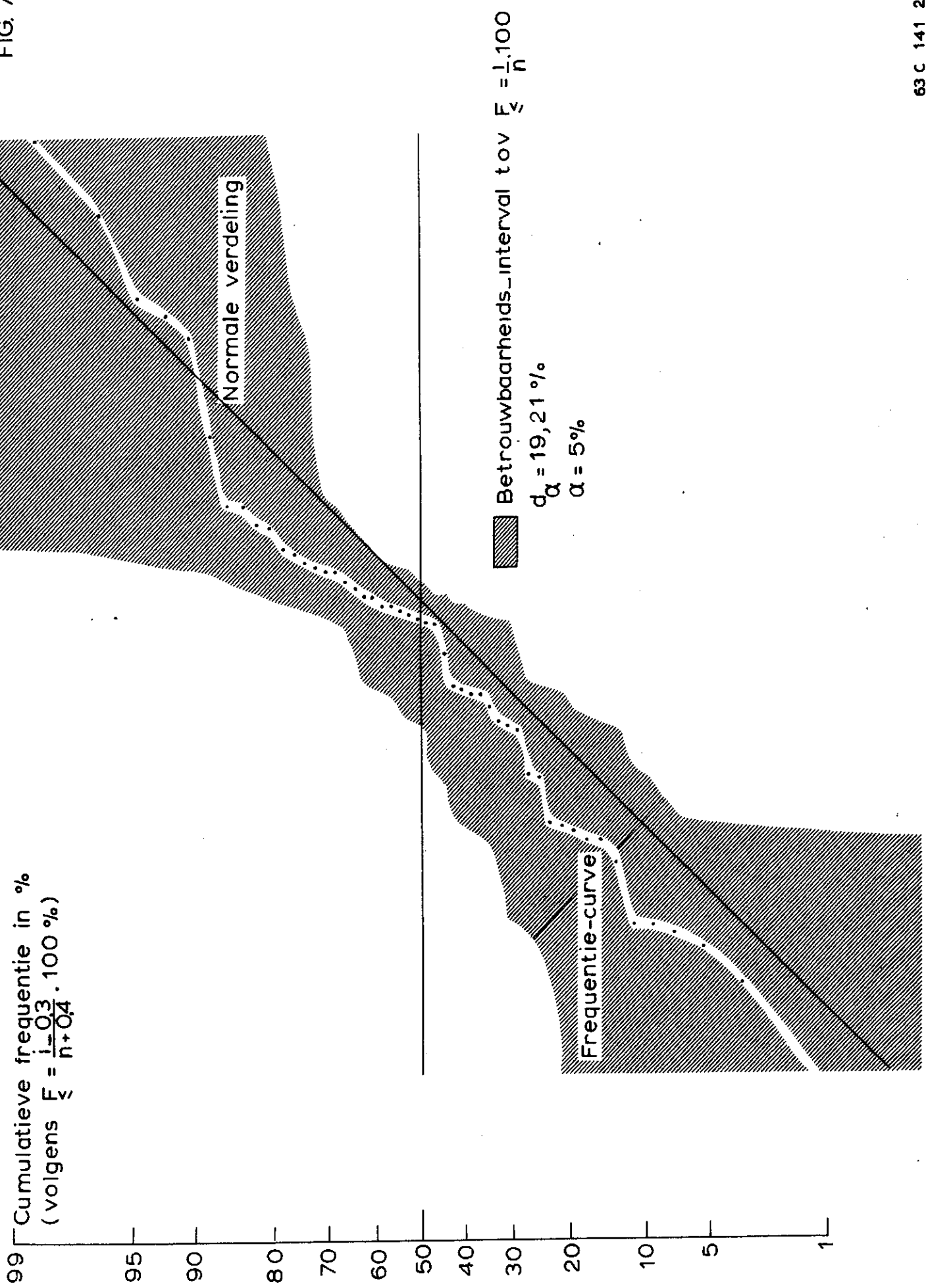
A-selectie steekproef van 10 getallen uit een standaard-normale verdeling.

Cumulatieve frequentie in %
(volgens $F_z = \frac{i-0.3}{n+0.4} \cdot 100\%$)



A-selecte steekproef van 50 getallen uit een standaard_normale verdeling.

FIG. 7



A-selecte steekproef van 100 getallen uit
een standaard_normale verdeling

Cumulatieve frequentie in %
(volgens $F_{\xi} = \frac{i - 0.3}{n + 0.4} \cdot 100 \%$)

