INSTITUUT VOOR CULTUURTECHNIEK EN WATERHUISHOUDING
NOTA no.186 d.d. 17 april 1963

CUMULATIEVE FREQUENTIEVERDELINGSCURVEN (I)

Het uitzetten van cumulatieve frequentieverdelingen

ir.Ph.Th. Stol

INHOUD

I. INLEIDING

Voor het vaststellen van de waarde die een gegeven grootheid kan aannemen kan men een meting verrichten die men als steekproef van de gevraagde grootheid kan beschouwen. Verricht men meer dan één zo'n steekproef dan zullen onderling verschillende uitkomsten gevonden worden. Aan de hand van de aldus verkregen reeks waarnemingsuitkomsten kan men afleiden welke waarde als gemiddelde verwacht zal kunnen worden, waartoe veelal het rekenkundig gemiddelde zal dienen.

Behalve deze gemiddelde waarde kan het tevens van belang zijn een antwoord te verkrijgen op de vraag hoe vaak een bepaalde hoge waarde nog overschreden zal worden, respectievelijk hoe vaak een bepaalde lage waarde niet bereikt (onderschreden) zal worden. Om te komen tot dit type uitspraken worden van de meetuitkomsten cumulatieve frequentieverdelingscurven opgesteld. Voorbeelden van deze wijze van werken op cultuurtechnisch gebied werden gegeven in [10] en [11] terwijl in [12] de aannamen en veronderstellingen die aan het gebruik van frequentieverdelingen ten grondslag liggen nader zijn toegelicht.

Het is duidelijk dat op grond van een steekproef van geringe omvang een uitspraak over mogelijke onder- respectievelijk overschrijdingen minder betrouwbaar zal zijn dan wanneer een steekproef van grote omvang ter beschikking staat.

In deze nota zullen enkele consequenties van het uitzetten van frequentiecurven nader worden besproken. In NOTA 187 [13] wordt nader ingegaan op het vaststellen van een betrouwbaarheidsinterval voor cumulatieve verdelingscurven.

II. MATHEMATISCHE FORMULERING

De kans P dat een continue stochastische grootheid $\underline{v}$ een bepaalde waarde v_1 niet zal overschrijden wordt weergegeven met de bepaalde integraal

$$P(\underline{v} < v_1) = \int_{-\infty}^{v_1} f(u)\, du$$

waarin de integrand f(u) de kansdichtheidsfunctie voorstelt. Voor normaal verdeelde grootheden geldt [zie b.v. FRASER, 1958, pag. 71 en FELLER, 1950, pag. 129 e.v.]

$$f(u) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2}$$

Bij een empirische wijze van werken zal f(u) niet bekend zijn. Het volgende geldt echter algemeen waarbij bedacht moet worden dat kansen door oppervlakten worden voorgesteld.

$$P(v_1 < \underline{v} < v_2) = \int_{-\infty}^{v_2} f(u)\, du - \int_{-\infty}^{v_1} f(u)\, du =$$

$$= \int_{v_1}^{v_2} f(u)\, du$$

zodat nog voor het bijzondere geval dat $v_1 = v_2$ gevonden wordt dat

$$P(v_o < \underline{v} < v_o) = P(\underline{v} = v_o) = \int_{v_o}^{v_o} f(u)\, du = 0 \qquad (1a)$$

Deze uitkomst houdt in dat het optreden van één bepaalde waarde $(\underline{v} = v_o)$ weliswaar niet uitgesloten is, doch een kans P = 0 heeft om gerealiseerd te worden. Deze redenering zal met een voorbeeld verduidelijkt worden.

Het aftappen van een dagneerslagsom van 3 mm, maar dan 3,000....mm, op een regenwaarnemingsstation is geen uitgesloten te achten gebeurtenis. De kans er op is echter 0*). Opgemerkt wordt dat in de praktijk met 3,0 mm het interval van 2,95 - 3,04 mm bedoeld wordt.

*) Om deze reden heeft het geen zin voor continue variabelen te beweren bijvoorbeeld dat de kans op 3 mm = 20%. Steeds moet de kans als overdan wel als onderschrijdingskans gegeven worden.

Het aftappen van een negatieve hoeveelheid neerslag is een voorbeeld van een uitgesloten gebeurtenis.

Het bovenstaande houdt dus in dat voor continue grootheden

$$P(\underline{v} < v_1) = P(\underline{v} \leq v_1) \qquad (1b)$$

Tenslotte wordt nog opgemerkt dat over- en onderschrijdingskansen elkaars complement zijn zodat

$$P(\underline{v} < v_1) = 1 - P(\underline{v} > v_1) \qquad (2)$$

Wanneer de werkelijk optredende kansen niet exact bekend zijn, wat veelal het geval is, worden deze uit waarnemingsuitkomsten geschat door middel van het frequentiequotient F dat gedefinieerd is als

$$F = \frac{\text{aantal successen}}{\text{totaal aantal uitkomsten}}$$

waarin de teller bijvoorbeeld kan zijn het aantal onderschrijdingen van een aangenomen hoeveelheid.

Frequentiequotienten van reeksen met een toenemend aantal waarnemingen hebben de neiging een zekere "stabiliteit" te gaan vertonen (experimentele wet der grote aantallen). Aan een gebeurtenis wordt dan ook wel een, onbekend maar constant, getal toegevoegd dat de kans op voorkomen van die gebeurtenis voorstelt. Op deze wijze kan men dan als het ware "de constante kern" in het frequentiequotient karakteriseren [HEMELRIJK, 1956].

Symbolisch voorgesteld, en in verband met (1b) ontstaat dan

$$\lim_{n \to \infty} \left\{ F_{<} = \frac{\text{aantal malen } \underline{v} \leq v_1}{\text{totaal aantal uitkomsten}} \overset{\text{def}}{=} \frac{m}{n} \right\} = P(\underline{v} \leq v_1) \qquad (3)$$

waarin m uiteraard afhankelijk is van n.

III. HET UITZETTEN VAN FREQUENTIEVERDELINGEN

De gebruikelijke methode om gegevens in een frequentie-overzicht samen te vatten is die waarbij van elk gegeven de onderschrijdingsfrequentie ($F_<$) wordt vastgesteld. Behalve voor de onderschrijdingen kan een verdeling voor overschrijdingen worden opgesteld ($F_>$). In tabel 1 staat het principe van deze methode aangegeven.

Tabel 1

Frequentiequotienten voor de maandneerslagsom (v) te Vlissingen voor augustus

Chronologisch					
Jaar	v in mm	Naar grootte	Rangnr m	$F_<$	$F_>$
1959	47	47	1	,25	1,00
1958	85	85	2	,50	,75
1957	117	110	3	,75	,50
1956	110	117	4	1,00	,25

aantal gegevens n = 4 (jaren)

Voor het gegeven met de kleinste waarde is m = 1, voor het gegeven met de grootste waarde is m = n.

Op cumulatief waarschijnlijkheidspapier uitgezet ("kansschaaltje") zijn de coördinaten in het algemeen

$$(v_m,\ F_<(m)) = (v_m,\ \frac{m}{n})$$

Opgemerkt wordt dat

$$F_<(m = n) = 1$$

een waarde is die op het kansschaaltje de ordinaat ∞ heeft en dus niet in tekening gebracht kan worden. Evenmin kan een uitkomst $F_< = 0$ met het kansschaaltje uitgezet worden daar hiervoor de ordinaat $-\infty$ is.

IV. DE HERHALINGSPERIODE

Uit de frequentiequotienten waarin de verzamelde gegevens zijn samengevat wil men, omgekeerd, weer uitspraken doen die met het oorspronkelijke aantal gegevens verband houden. Zo zal een onderschrijdingskans $P = p_o$ aanleiding zijn tot de uitspraak dat er op reeksen van n waarnemingen, gemiddeld

$$np_o \text{ maal per reeks} \qquad (4a)$$

een dergelijke onderschrijding zal plaatsvinden. Veelal zal de belangstelling uitgaan naar de mogelijkheid waarmee grote waarden nog overschreden zullen worden zodat dan volgens (2) gemiddeld

$$n(1 - p_o) \text{ maal per reeks} \qquad (4b)$$

een dergelijke overschrijding zal plaatsvinden.

Voor het geval de cumulatieve frequentieverdeling is opgesteld met gegevens over jaren dan wordt aan (4b) de betekenis gegeven van het aantal malen dat een overschrijding zich, gemiddeld in reeksen van n jaren, zal herhalen. Per jaar komt de overschrijding dus gemiddeld $(1 - p_o)$ maal voor.

Voor het voorbeeld van tabel 1 waar de frequentieverdeling dus over de jaren is samengesteld wordt de herhalingsperiode eveneens in jaren uitgedrukt. Er geldt dan dat (extreme) waarden gemiddeld overschreden zullen worden:

$$1 \times \text{per } \frac{1}{1 - p_o} \text{ jaar}$$

In formule kan dus een schatting voor de herhalingsperiode (return-period T) voor overschrijdingen van de m-de grootste waarneming (zie tabel 1) voorgesteld worden door

$$T = \frac{1}{1 - F_<(m)} \qquad (5)$$

Hierin geeft T aan wat de lengte van de waarnemingsreeks moet zijn om op den duur in deze reeksen gemiddeld één waarneming te hebben waarvoor $\underline{v} > v_m$. Veelal worden om praktische redenen breuken of decimale groot-

heden vermeden en zal men bij voorkeur spreken over een voorkomen van 3 x in 4 jaar in plaats van 1 x in 1,33 jaar enz. Overigens zijn beide uitspraken gelijkwaardig.

V. THEORETISCHE EN PRAKTISCHE BEZWAREN

De bovengeschetste methode voor het uitzetten van waarnemingsuitkomsten op cumulatief waarschijnlijkheidspapier, geeft aanleiding tot het optreden van een aantal bezwaren die van theoretische en praktische aard zijn [GUMBEL, 1954, pag. 13 e.v.]. De drie belangrijkste zullen hier nader toegelicht worden.

Bezwaar 1

Reeds werd opgemerkt dat bij het uitzetten volgens $F_< = \frac{m}{n}$ het gegeven waarvoor $F_< = 1$, het gegeven met de grootste waarde, niet in tekening kan worden gebracht, zodat niet alle verzamelde gegevens benut worden bij de beoordeling van de vorm van de curve.

Bezwaar 2

Bij de gevolgde procedure is niet voldaan aan de eis dat (zie (2)),

$$P(\underline{v} \leq v_1) + P(\underline{v} > v_1) = 1$$

Bezwaar 3

De volgens (5) berekende herhalingsperiode komt niet overeen met die welke uit de gegevens volgt.

ad 1. Op eenvoudige, praktische wijze kan aan het genoemde bezwaar tegemoet worden gekomen door bijvoorbeeld in plaats van

$$F_< = \frac{m}{n}$$

te nemen bijvoorbeeld [BENARD]:

$$F_< = \frac{m - 1}{n}, \quad F_< = \frac{m}{n + 1}, \quad F_< = \frac{m - 1/2}{n}, \quad \text{enz.}$$

De eerste vorm heeft het nadeel dat voor m = 1 het genoemde bezwaar nu voor het gegeven met de kleinste waarde geldt. De overige vormen behoeven een nadere theoretische fundering.

GUMBEL (1958, pag. 33) geeft een figuur waarin het verschil tussen de op deze wijzen uitgezette curven geïllustreerd wordt.

ad 2. Wordt uitgerekend wat de som van de onder- en overschrijdingskansen is, dan volgt uit tabel 1 dat verkregen wordt de te hoge waarde $1{,}25 \neq 1$.

Volgens de gebruikelijke methode geldt namelijk:

$$F_{<} = \frac{m}{n} \tag{6}$$

$$F_{>} = \frac{(n - m + 1)}{n} \tag{7}$$

Uit (7) volgt dan

$$F_{<} = 1 - F_{>} = \frac{m - 1}{n} \neq \frac{m}{n}$$

Naarmate n toeneemt zal het verschil tussen deze uitkomsten afnemen, namelijk:

$$P(\underline{v} \leq v_1) + P(\underline{v} > v_1) =$$

$$= \lim_{n\to\infty}\{F_{<} + F_{>}\} = \lim_{n\to\infty}\left(\frac{m}{n} + \frac{n - m + 1}{n}\right) =$$

$$= \lim_{n\to\infty}\left(1 + \frac{1}{n}\right) = 1 \tag{8}$$

In figuur 1 wordt aangegeven tot welke verschillen het uitzetten volgens $F_{<}$ of volgens $F_{>}$ aanleiding geeft voor een gering (n = 10) aantal gegevens. Voor het m-de gegeven geldt:

$$F_{<} = \frac{m}{10} \text{ of } 10 \times m\%$$

Het van één curve aflezen van zowel over- als onderschrijdingskansen geeft dus onjuiste uitkomsten, echter minder naarmate n toeneemt (8). Voor verschillende waarden van n is het verschil d in mm tussen $F_{<}$ en $(1 - F_{>})$, gemeten langs de "kansschaal" in figuur 1, uitgezet in figuur 2. Het verschil is uitgedrukt in mm en houdt zodoende dus verband met de tekennauwkeurigheid. Neemt men genoegen met een afstand op het papier tussen $F_{<}$ en $(1 - F_{>})$ van 1 mm op het 90% niveau, dan zullen minstens 200 gegevens ter beschikking moeten staan. Voor het op één na laatste gegeven waarvoor m = n - 1 zal de afstand tussen beide curven dan toch nog 9 mm bedragen. Juist in het gebied dat de grootste interesse heeft blijken de afwijkingen het grootst te zijn.

ad 3. Een schatting van de herhalingsperiode T volgt volgens (5) uit:

$$T = \frac{1}{1 - F_{<}(m)}$$

en in verband met (6)

$$T = \frac{1}{1 - \frac{m}{n}} = \frac{n}{n - m}$$

wat voor het gegeven met de grootste waarde - waarvoor m = n - wordt:

$$T \rightarrow \infty$$

wat een niet aannemelijke uitkomst is.

Samenvattend worden met de gegevens uit tabel 1 de volgende uitkomsten verkregen die de genoemde bezwaren nog eens illustreren (tabel 2).

Tabel 2

Uitkomsten uit tabel 1 welke de bezwaren 2 en 3 demonstreren.

v in mm	Rangnr. m	$F_{<}$	Overschrijdingen in T van 4 jaar		$F_{<}$	$F_{<} + F_{>}$
			berekend	werkelijk		
47	1	,25	3×	4×	1,00	1,25
85	2	,50	2×	3×	0,75	1,25
110	3	,75	1×	2×	0,50	1,25
117	4	1,00	0×	1×	0,25	1,25

aantal gegevens n = 4 (jaren)

VI. MOGELIJKHEDEN TOT OPHEFFEN VAN DE BEZWAREN

Verschillende mogelijkheden kunnen genoemd worden om elk van de drie bezwaren op te heffen. Het zal blijken dat niet alle drie de bezwaren tegelijk opgeheven kunnen worden zodat naar een compromis gezocht moet worden afhankelijk van de doelstelling waarvoor een cumulatieve verdeling gebruikt wordt.

Uit figuur 1 valt op te maken dat een mogelijke oplossing kan zijn de gemiddelde frequenties van beide verdelingen te gebruiken (streep-lijn).

De frequenties zijn nu dus als volgt gedefinieerd:

$$F'(v_m) \overset{\text{def}}{=} \frac{1}{2}\left\{F_< + (1 - F_>)\right\} = \frac{1}{2}\left(\frac{m}{n} + 1 - \frac{n - m + 1}{n}\right) = \frac{m - 1/2}{n}$$

zodat voor de grootste waarde met m = n geldt:

$$F'(v_n) = 1 - \frac{1}{2n} < 1$$

welke waarde voor een eindig aantal waarnemingen steeds uitgezet kan worden. Evenzo voor m = 1

$$F'(v_1) = \frac{1}{2n} > 0$$

Voorts geldt voor de som van de complementaire kansen:

$$F'_< + F'_> = \frac{m - 1/2}{n} + \frac{(n - m + 1) - 1/2}{n} \equiv 1$$

Het tweede bezwaar is dus eveneens opgeheven, doch voor de herhalingsperiode geldt:

$$T = \frac{1}{1 - F'} = \frac{1}{1 - \frac{m - 1/2}{n}} = \frac{2n}{2(n - m) + 1}$$

waarmee voor het gegeven met de grootste waarde verkregen wordt (m = n),

$$T = 2n$$

en dus een herhalingsperiode die tweemaal zo lang is als uit de gegevens volgt, zodat in het van belang zijnde gebied de herhalingsperiode sterk overschat wordt.

Andere voorstellen tot opheffen van de bezwaren berusten op de verdeling van de naar grootte gerangschikte gegevens zelf. Zij dus elke $\underline{v}_m$ een stochastische grootheid dan is de oorspronkelijke cumulatieve frequentiecurve de meetkundige plaats van de punten $(\underline{v}_m, F_<(\underline{v}_m))$. Deze punten hebben dus een waarschijnlijkheidsverdeling over deze curve, onafhankelijk van de curve zelf en dus onafhankelijk van de verdeling van $\underline{v}$. De stochastische grootheid, nu met weglating van de index $<$,

$$\underline{F} = F(\underline{v}_m) \quad (m = 1, 2, \ldots, n)$$

bezit zelf een verdeling waarvoor geldt:

$$\bar{F} = E(\underline{F}_m) = \frac{m}{n + 1} \text{ , (verwachtingswaarde)}$$

vervolgens

$$\text{mediaan } \underline{F}_m \approx \frac{m - 0,3}{n + 0,4} \text{ , ("50\% - punt")}$$

$$\text{modus } \underline{F}_m = \frac{m - 1}{n - 1} \text{ , ("top" van de verdeling)}$$

[GUMBEL, 1954, pag. 15; 1958, pag. 79 en BENARD]

GUMBEL stelt voor om de gegevens uit te zetten volgens $\bar{F}$. BENARD beveelt aan de mediaan te gebruiken. De modus geeft in verband met het eerste bezwaar geen voordelen boven de andere methoden, is zelfs slechter daar nu noch het punt $m = 1$, noch het punt $m = n$ uitgezet kan worden.

Tabel 3

Frequentiepercentages voor het voorbeeld uit tabel 1, berekend volgens verschillende methoden.

v in mm	Rangnr. m	$F_<$ %	$E(\underline{F})$ %	med. $\underline{F}$ %	mod. $\underline{F}$ %	$\frac{1}{2}\{F_< + (1-F_>)\}$ %
47	1	25	20	15,91	0	12,5
85	2	50	40	38,64	33,33	37,5
110	3	75	60	61,36	66,67	62,5
117	4	100	80	84,09	100,00	87,5

Achtereenvolgens zullen de methoden GUMBEL en BENARD besproken worden.

a. Methode van GUMBEL

GUMBEL [6 en 7] stelt voor de verwachtingswaarde als ordinaat te gebruiken. Er komt dan achtereenvolgens

$$F_< = \frac{m}{n + 1}$$

met m = n, respectievelijk m = 1

$$F_< = \frac{n}{n + 1} < 1, \quad F_< = \frac{1}{n + 1} > 0$$

welke punten beide uitgezet kunnen worden (bezwaar 1)

Vervolgens

$$F_< + F_> = \frac{m}{n + 1} + \frac{n - m + 1}{n + 1} = 1$$

zodat bezwaar 2 tevens opgeheven is.

Tenslotte

$$T = \frac{1}{1 - \frac{m}{n + 1}} = \frac{n + 1}{n - m + 1}$$

wat wordt voor het grootste gegeven (m = n)

$$T = n + 1$$

De herhalingsperiode is dus één eenheid te groot wat bij wat grotere waarden van n een nog slechts kleine tekortkoming betekent.

Voor het geval er gelijke waarnemingen zijn, een situatie die theoretisch in verband met (1a) niet kan optreden, is het de beste methode met een gemiddeld rangnummer te werken teneinde het bezwaar 2 te kunnen blijven opheffen. Tabel 4 illustreert dit nader.

Tabel 4

Het toekennen van rangnummers bij gelijke waarnemingen toegepast op de methode GUMBEL

v in mm	Rangnr. m	$F_{\%}<$	m'	$F_{\%}>$	Rangnr. $\bar{m}$	$F_{\%}<$	$\bar{m}'$	$F_{\%}>$
1	2	3	4	5	6	7	8	9
47	1	20	4	80	1	20	4	80
85	2	40	3	60	2	40	3	60
115(2×)	4	80	2	40	$3\frac{1}{2}$	70	$1\frac{1}{2}$	30

In de tabel stelt m het rangnummer voor voor onderschrijdingen, m' het rangnummer voor overschrijdingen. Als basis voor de frequentieverhouding is genomen een totaal van (n + 1) waarnemingen. De berekening met $\bar{m}$ heeft tot effect dat de som van de complementaire kansen 100% is, zie kolom 7 en 9.

b. Methode van BENARD en BOS-LEVENBACH

BENARD [1] toont aan dat steeds geldt voor de ordinaat

$$E(\underline{F}) < \text{med } \underline{F} < \text{mod } \underline{F} \qquad (9)$$

zodat de verdeling van de ordinaatwaarden scheef is en wel het sterkst voor kleine, respectievelijk grote waarden van m. De methode van GUMBEL die op de verwachtingswaarde $E(\underline{F})$ berust heeft dus het nadeel dat voor de gegevens met grote waarden dus voor

$$m > \frac{n+1}{2}$$

de op waarschijnlijkheidspapier uitgezette gegevens in meer dan 50% van de gevallen onder de (onbekende) verdelingscurve zullen liggen terwijl dit voor

$$m < \frac{n+1}{2}$$

juist in meer dan 50% van de gevallen boven de curve is. De gegevens hebben hiermee een tendens zich rond een S-curve te groeperen.

In meer dan 50% van de gevallen zal men de helling van de curve dus te klein schatten en de spreiding te hoog. Bij gebruik van de modus van $\underline{F}$, zie (9), zal het effect juist andersom liggen.

Dit bezwaar, dat nog niet genoemd is en het rangnummer 4 zou kunnen krijgen, wordt opgeheven door het gebruik van de mediaan-waarde. Alle uitgezette punten hebben nu evenveel kans om boven als onder de verdelingscurve te liggen ongeacht het rangnummer van het punt.

Achtereenvolgens geldt nu voor de methode BENARD

$$F_{<} = \frac{m - 0,3}{n + 0,4}$$

en

$$F_{<} = \frac{n - 0,3}{n + 0,4} < 1 \text{ voor } m = n$$

Vervolgens

$$F_{<} + F_{>} = \frac{m - 0,3}{n + 0,4} + \frac{(n - m + 1) - 0,3}{n + 0,4} = 1$$

Tenslotte

$$T = \frac{1}{1 - \frac{m - 0,3}{n + 0,4}} = \frac{n + 0,4}{n - m + 0,7}$$

wat voor m = n, het grootste gegeven asymptotisch gelijk wordt aan

$$T \approx 1,44n + 0,5$$

[GUMBEL 1958, pag. 79]

De beide eerste bezwaren zijn opgeheven, doch de herhalingsperiode wordt, voor het gegeven met de grootste waarde, 44% te lang geschat. Om deze reden blijft GUMBEL de voorkeur geven aan het gebruik van de verwachtingswaarde van F.

Het verschil in vorm van de verdelingscurve tengevolge van de wijze van uitzetten wordt geïllustreerd in figuur 3. Voor de dagneerslag op 10, 15 en 20 januari over 10 jaar (30 gegevens) van de Rottegatspolder werd de verdelingscurve bepaald volgens 4 methoden. De in de tekst besproken eigenschappen komen in deze figuur goed tot uiting.

VII. DE HERHALINGSPERIODE ALS FUNCTIE VAN HET FREQUENTIEQUOTIENT

De betrekking tussen de herhalingsperiode en het frequentiequotient luidt volgens (5) met weglating van het < teken

$$T = \frac{1}{1 - F} \tag{10}$$

met definitiegebied: $0 < F < 1$

en functiewaarden: $1 < T < +\infty$

Geschreven kan worden

$$T(1 - F) = 1 \tag{11}$$

wat een hyperbool voorstelt met asymptoten

$$T = 0$$
$$F = 1$$

zie figuur 4. Verder geldt nog:

$$\frac{dT}{dF} = \frac{1}{(1 - F)^2} \tag{12}$$

$$\frac{d^2T}{dF^2} = \frac{2}{(1 - F)^3} \tag{13}$$

Wordt nu het <u>frequentiequotient als stochastische grootheid</u> opgevat dan is ook T stochastisch zodat

$$\underline{T} = \Phi(\underline{F})$$

Het spreidingsgebied van de punten ($\underline{T}$, $\underline{F}$) ligt ook nu weer op de curve (11) zelf. Voor de verwachtingswaarde van $\underline{T}$ geldt: [zie b.v. FRASER 1958, pag. 96 e.v.]

$$E(\underline{T}) = E\left\{\Phi(\underline{F})\right\} \neq \Phi\left\{E(\underline{F})\right\}$$

Deze uitkomst houdt in dat met de verwachtingswaarde van $\underline{F}$ met (10) voor $\underline{T}$ niet de verwachtingswaarde gevonden zal worden zodat met

$$\underline{T} = \Phi\left\{E(\underline{F})\right\}$$

geen inzicht omtrent de ligging van de waarde van T ten opzichte van de rond T gespreide waarden verkregen wordt.

Gezien het feit dat (11) binnen het definitiegebied een eenwaardige monotoon stijgende continue functie is, zoals ook uit (12) volgt, zal gelden dat aan elke volgorde van punten F eenzelfde volgorde van punten T toegevoegd is. Dit houdt weer in dat uit (10) volgt

$$\Phi\,(\text{med}\ \underline{F}) = \text{med}\ \underline{T}$$

zodat het 50%-punt van $\underline{F}$ een correspondentie vertoont met het 50%-punt van $\underline{T}$. Dit betekent dat bij gebruik van de mediaan van $\underline{F}$ ook, uit (10), de mediaan van $\underline{T}$ verkregen wordt.

In figuur 4 is de transformatie van de verdeling van $\underline{F}$ in die van $\underline{T}$ voor een drietal gevallen ingeschetst. Uit de figuur wordt duidelijk dat de verdeling van $\underline{F}$ inderdaad op deze wijze scheef moet zijn daar waarden van $F < 0$ en $F > 1$ uitgesloten zijn.

- - - - - -

Op analoge wijze als FISHER voor de correlatiecoefficiënt aanbeveelt [FISHER, 1958, pag. 198] zou een transformatie van de vorm

$$z = \frac{1}{2} \ln \frac{F}{1 - F}$$

kunnen worden toegepast teneinde de verdeling van $\underline{F}$ zelf "meer normaal" te maken. Opgemerkt wordt dat voor dit geval de eigenschap dat een normaal verdeelde grootheid zich als een rechte representeert niet meer zal opgaan.

- - - - - -

Midden in het definitiegebied is de verdeling symmetrisch daar nu (voor $m = \frac{1}{2}(n + 1)$)

$$E(\underline{F}) = \frac{m}{n + 1} = \frac{1}{2}$$

$$\text{med}(\underline{F}) = \frac{m - 0,3}{n + 0,4} = \frac{0,5n + 0,2}{n + 0,4} = \frac{1}{2}$$

$$\text{mod}(\underline{F}) = \frac{m - 1}{n - 1} = \frac{1}{2}$$

en gemiddelde, mediaan en modus aan elkaar gelijk zijn.

Met het gebruik van de mediaan van het frequentiequotient wordt dus voor de gemiddelde herhalingsperiode eveneens de mediaan gevonden. Bij een verdeling die loopt over de jaren heeft de gemiddelde herhalingsperiode nu dus de betekenis van het aantal jaren dat beschouwd moet worden om in 50% van het

aantal gevallen vaker dan 1 × een vastgestelde overschrijding te constateren en in de overige 50% minder dan 1 × die overschrijding.

Over de verwachtingswaarde van $\underline{T}$ kan nog het volgende worden opgemerkt.

Een benadering van $T = \Phi(F)$ kan verkregen worden met behulp van een Taylorreeks [zie b.v. GERRETSEN, 1959, pag. 229]

$$\Phi(F) = \Phi(\bar{F}) + (F - \bar{F})\ \Phi'(\bar{F}) + \frac{1}{2}(F - \bar{F})^2\ \Phi''(\bar{F}) + \ldots.$$

Met de eigenschappen van de verwachtingswaarde E [b.v. FRASER, 1958, Hoofdstuk 5], volgt hieruit [KUIPER, 1959]

$$E\Phi(\underline{F}) = \Phi\left\{E(\underline{F})\right\} + \frac{1}{2}\sigma^2\Phi''\left\{E(\underline{F})\right\}$$

en dus, in verband met (13)

$$E(\underline{T}) = \Phi\left\{E(\underline{F})\right\} + \frac{\sigma^2}{(1 - F)^3} \qquad (14)$$

Wanneer dus de tweede term in het rechterlid klein is zal de benadering gelden

$$E\Phi(\underline{F}) = \Phi\left\{E(\underline{F})\right\}$$

in andere gevallen zal deze benadering niet opgaan.

Wel kan gezegd worden dat aangezien

$$\frac{\sigma^2}{(1 - F)^3} > 0$$

steeds zal gelden

$$E(\underline{T}) > \frac{1}{1 - E(\underline{F})} \qquad (14a)$$

waaruit dan weer volgt dat uit

$$T_o = \frac{1}{1 - F}$$

niet valt vast te stellen welke kans van voorkomen aan T_o, wat betreft de plaats van T_o in de rond deze waarde fluctuerende T-waarden, moet worden toegekend..

Bij gebruik van de mediaan doet dit bezwaar zich niet voor zoals eerder in deze paragraaf werd uiteengezet, dan wordt namelijk ook voor T de mediaan gevonden.

In opgave 4 van het examen statistisch analist (1953) komt het bovenstaande probleem eveneens ter sprake. Door het Mathematisch Centrum wordt het volgende "intuitieve" antwoord voorgesteld (rapport SP 75), dat vertaald in hydrologische termen als volgt kan luiden:

Indien het aantal overschrijdingen x als vaststaand wordt aangenomen en het aantal jaren n tot en met de x-de overschrijding als stochastische grootheid wordt opgevat, dan is $p = 100 \frac{x}{n}$ geen zuivere schatting van de overschrijdingskans.

In het algemeen zal in het laatste jaar van de beschouwde reeks van n jaar niet juist een overschrijding optreden zodat in feite een te grote waarde voor p gevonden wordt. Dit houdt in dat dus n groter gekozen moet worden. Met andere woorden de herhalingsperiode heeft een verwachtingswaarde die hoger zal liggen dan uit $T = 1/p$ berekend wordt, overeenkomstig (14a).

VIII. NABESCHOUWING EN SAMENVATTING

In het voorgaande werd uiteengezet welke complicaties zich voordoen bij het uitzetten van gegevens als cumulatieve frequentie-curve. Uit de beschouwingen bleek dat naast het gebruik van het frequentie-quotient $F = \frac{m}{n}$ nog twee andere wijzen van uitzetten toegepast kunnen worden, waarbij van F òf de mediaan, òf het gemiddelde (de verwachtingswaarde) als ordinaat gebruikt wordt.

Het gebruik van de mediaan van F heeft het voordeel dat alle gegevens, onafhankelijk van de vorm van de (onbekende) verdelingscurve en onafhankelijk van het rangnummer van het gegeven een even grote kans hebben boven of onder de curve te liggen. Voor grafische bewerking van de gegevens is deze methode dus in het voordeel en verdient dan aanbeveling. De vrijehand-curve kan "zo goed mogelijk" door de gegevens getrokken worden er voor zorgdragend dat positieve en negatieve afwijkingen ten opzichte van de curve steeds langs de gehele curve tegen elkaar opwegen. Systematische afwijkingen kunnen niet optreden.

De verwachtingswaarde van F (namelijk $\bar{F}$) heeft het voordeel dat steeds een gemiddelde herhalingsperiode berekend wordt die praktisch gelijk is aan die welke uit de oorspronkelijke gegevens volgt. Om deze reden geeft GUMBEL de voorkeur aan het gebruik van deze grootheid met als nevenvoordeel de eenvoudige berekenwijzen daar $\bar{F} = \frac{m}{n + 1}$

Het frequentiequotient $\frac{m}{n}$ had een aantal bezwaren die in het voorgaande uitvoerig zijn besproken. Het is echter dit quotient waarvoor een toets is afgeleid waarmee het mogelijk is cumulatieve verdelingen onderling te vergelijken. In NOTA 187 [13] zal deze toets nader toegelicht worden terwijl nog zal worden ingegaan op de complicatie die ontstaat wanneer de gegevens volgens de mediaan zijn uitgezet. In dat geval is namelijk een kleine hulpbewerking nodig voor het uitzetten van het betrouwbaarheidsinterval.

Literatuur

BENARD, A. en E.C. BOS-LEVENBACH. Het uitzetten van waarnemingen op waarschijnlijkheidspapier.
Rapport SP 30 van de statistische afdeling van het Mathematisch Centrum te Amsterdam.

FELLER, W., 1950. An introduction to probability theory and its applications.
Vol I, New York
(Instituut voor Cultuurtechniek en Waterhuishouding 11/23)

FISHER, R.A., 1958. Statistical methods for research workers.
London.
(Instituut voor Cultuurtechniek en Waterhuishouding 11/103)

FRASER, D.A., 1958. Statistics, an introduction.
New York.
(Instituut voor Cultuurtechniek en Waterhuishouding 11/109)

GERRETSEN, J.C.H., 1959. Raaklijn en oppervlakte.
Haarlem.
(Instituut voor Cultuurtechniek en Waterhuishouding 11/73)

GUMBEL, E.J., 1954. Statistical Theory of Extreme Values and Some Practical Applications.
New York
(Instituut voor Cultuurtechniek en Waterhuishouding 11/125)

_____, 1958. Statistics of extremes.
New York.
(Instituut voor Cultuurtechniek en Waterhuishouding 11/167)

HEMELRIJK, J., 1956. Syllabus van een oriënterende cursus Mathematische Statistiek.
Rapport S 200 (C8) van het Mathematisch Centrum te Amsterdam.

KUIPER, N.H., 1959. Wiskundige verwerking van waarnemingsuitkomsten. Collegedictaat Wageningen.

STOL, Ph.Th., 1959. A statistical analysis of the differences between precipitation and evaporation in the Netherlands. Technical Bulletin, Instituut voor Cultuurtechniek en Waterhuishouding 9.

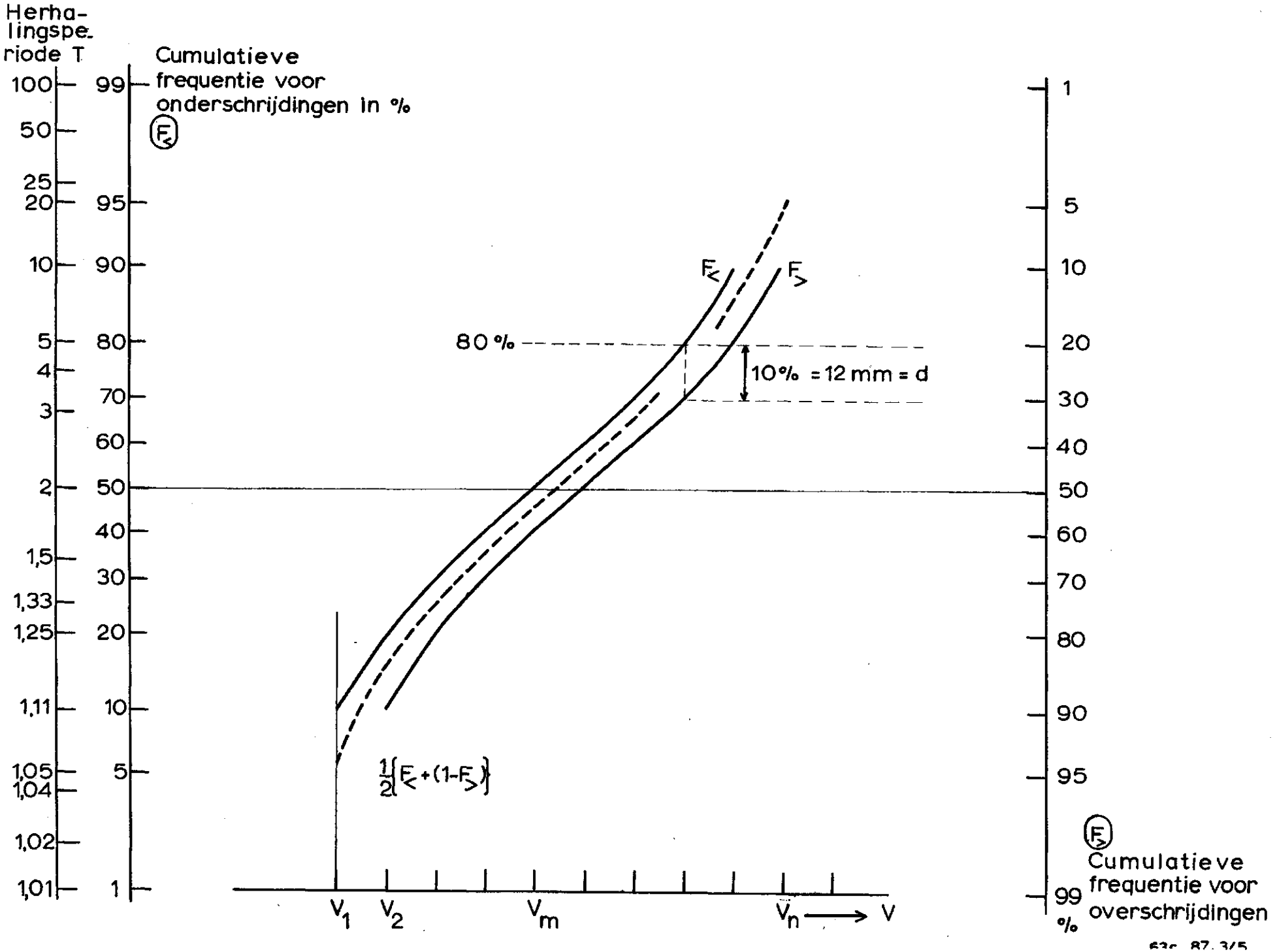
_____ , 1962. Een frequentie-onderzoek naar de te verwachten vochttekorten in de Tielerwaard West. Deelrapport 14 van het Interrimrapport van Werkgroep I: "De waterbehoefte van de Tielerwaard West". Commissie Bestudering Waterbehoefte Gelderse Landbouwgronden.

_____ , 1963. Het gebruik van frequentieverdelingen bij het onderzoek naar afvoercoëfficiënten. (Instituut voor Cultuurtechniek en Waterhuishouding NOTA 165)

_____ , 1963. Cumulatieve frequentieverdelingscurven (II). Een betrouwbaarheidsinterval voor frequentieverdelingen. (Instituut voor Cultuurtechniek en Waterhuishouding NOTA 187)

VERSCHIL TUSSEN ONDERSCHRIJDINGS- EN OVERSCHRIJDINGS-FREQUENTIES. HYPOTHETISCH VOORBEELD VOOR n = 10

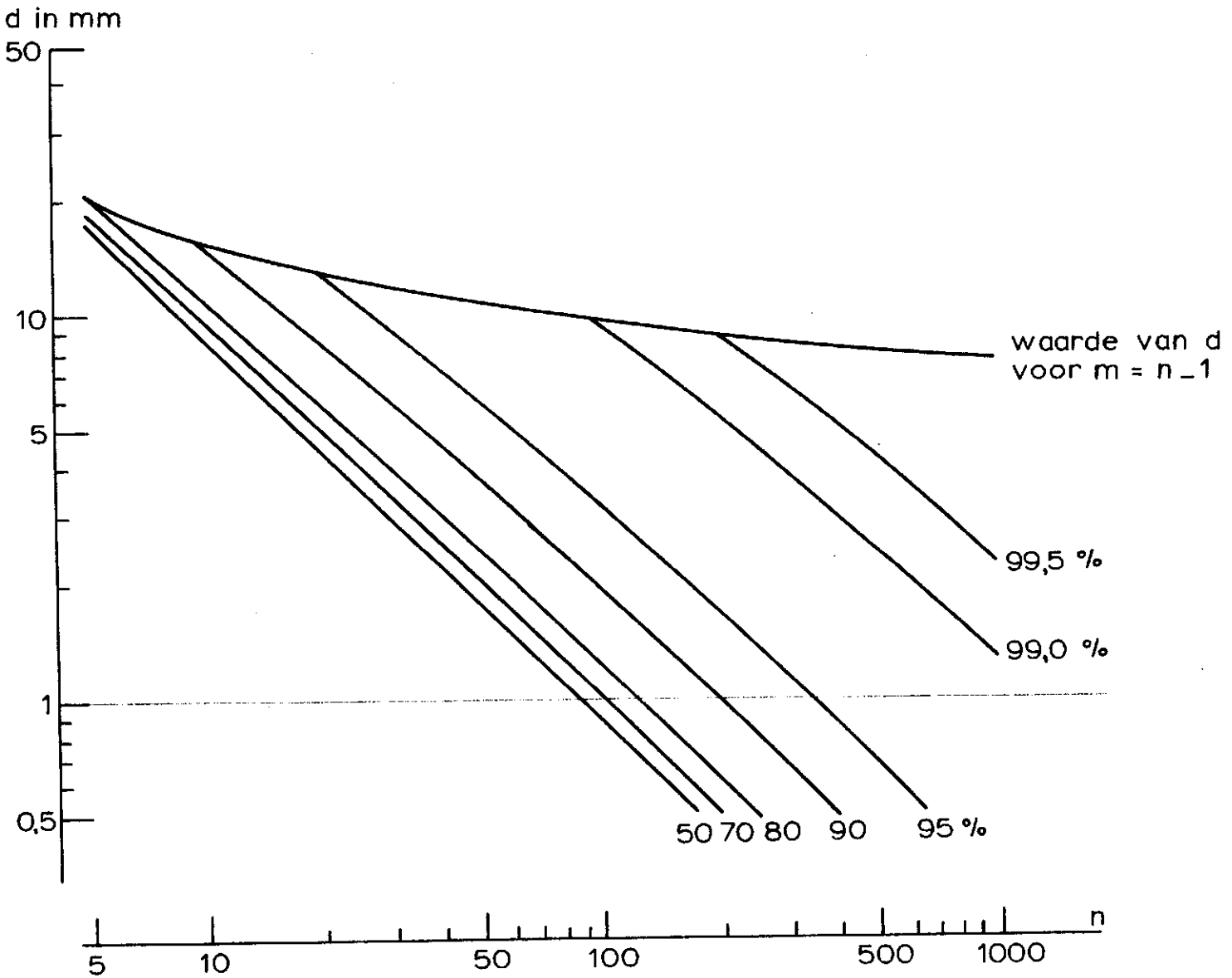
FIG. 1



63c 87.3/5

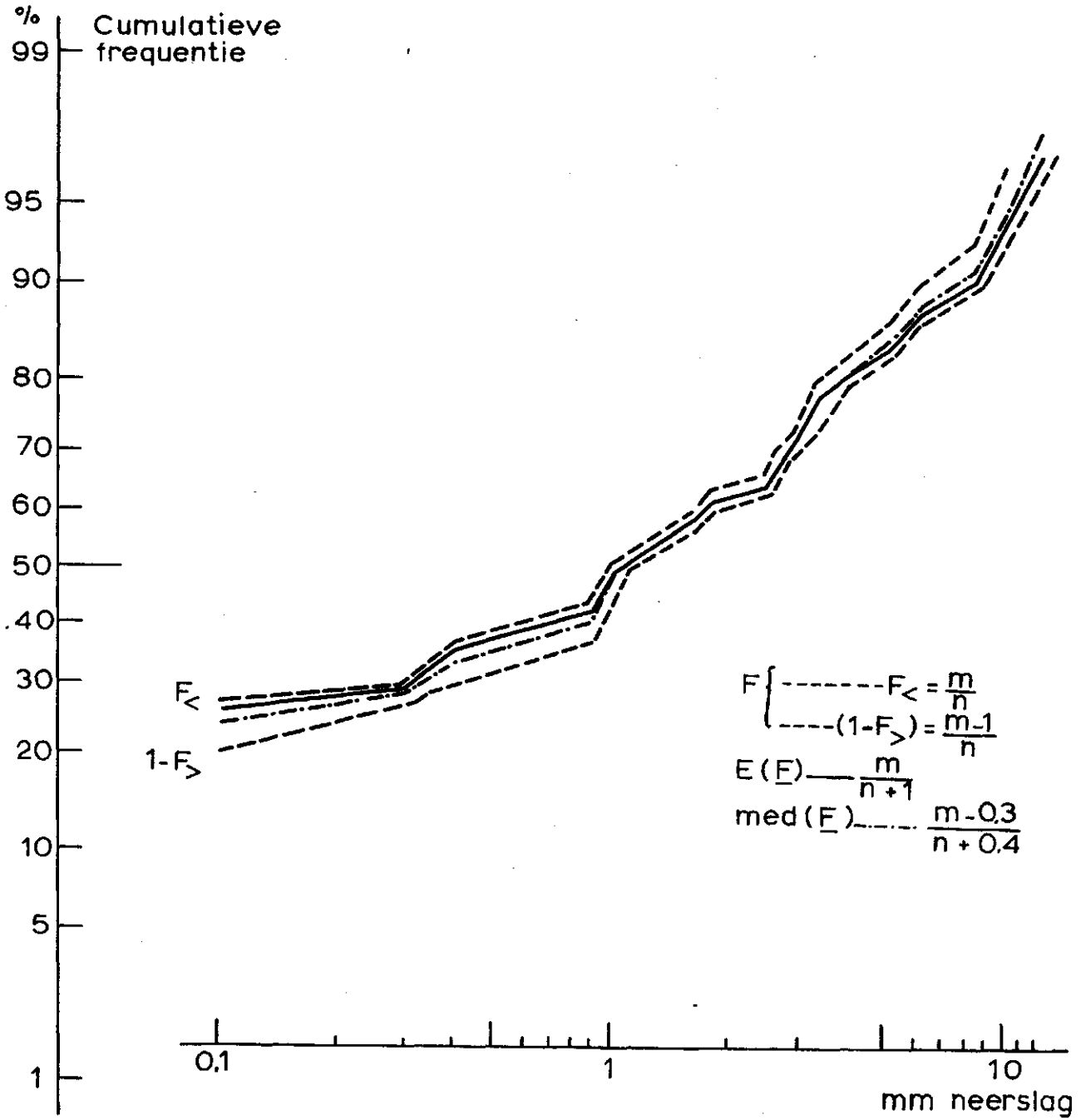
WAARDEN VOOR d UIT FIGUUR 1 VOOR VERSCHILLENDE STEEKPROEFGOOTTEN n

FIG. 2



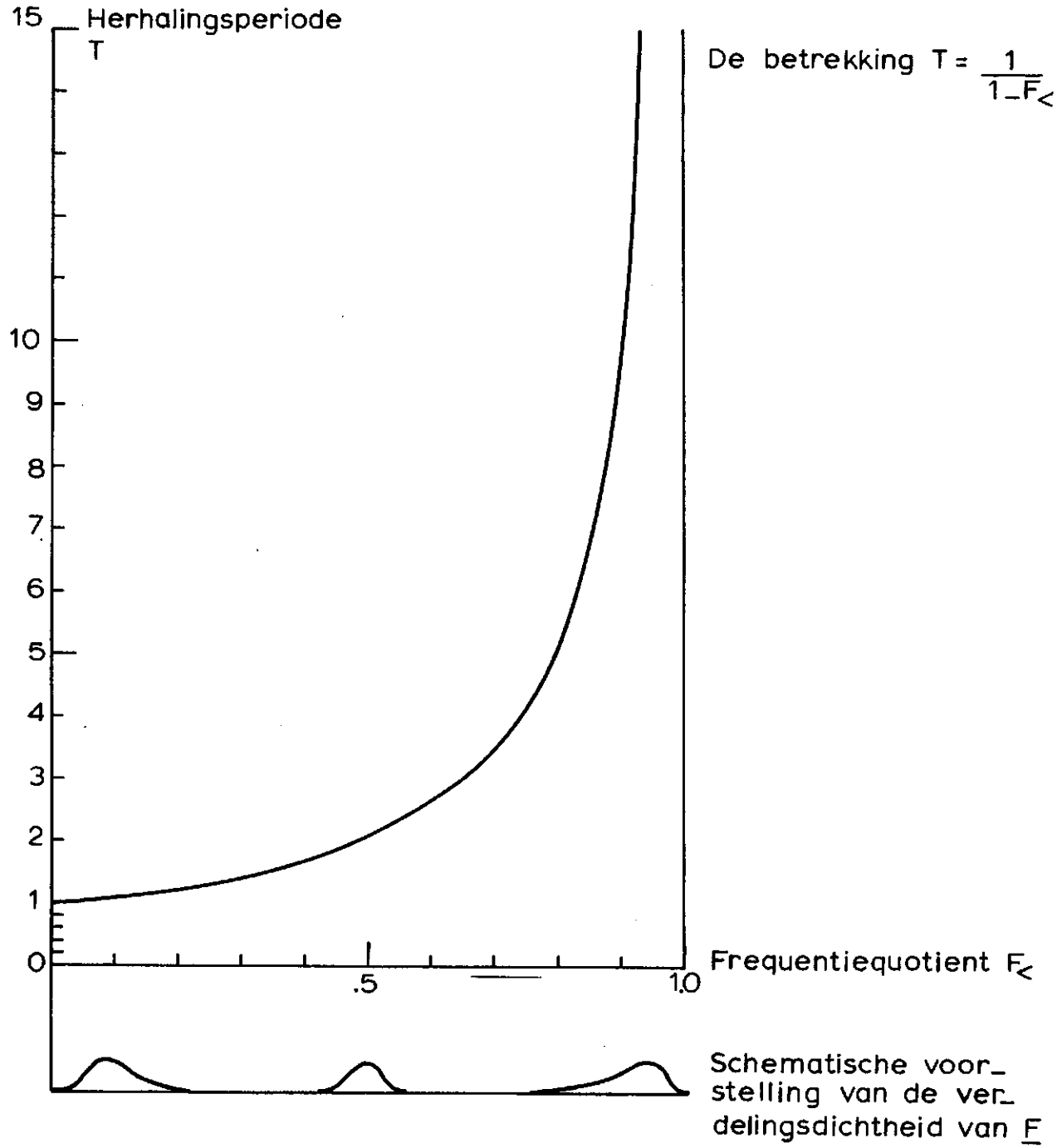
FREQUENTIEVERDELING VOOR DE NEERSLAG OP 15 JANUARI OVER DE JAREN 1952 T/M 1961 ROTTEGATSPOLDER

FIG. 3



63 c. 87. 4/5

FIG 4


15
Herhalingsperiode
T
10
9
8
7
6
5
4
3
2
1
0
.5
1.0
De betrekking $T = \frac{1}{1-F_<}$
Frequentiequotient $F_<$
Schematische voor-
stelling van de ver-
delingsdichtheid van F