CULTUURTECHNIEK EN WATERHUISHOUDING
NOTA nr.187 d.d. 21 mei 1963

CUMULATIEVE FREQUENTIEVERDELINGS-CURVEN (II)

Een betrouwbaarheidsinterval voor frequentie-verdelingen en frequentiequotienten

Ir.Ph.Th.Stol

I N H O U D

1. INLEIDING

Uit een empirische frequentieverdeling van bijvoorbeeld neerslaggegevens kan men afleiden met welke frequentie de overschrijding van een gesteld aantal mm neerslag, is opgetreden. Wanneer aangenomen mag worden dat het neerslagpatroon zich niet wijzigt kan uit de empirische verdeling ook afgeleid worden met welke frequentie een gestelde overschrijding zal optreden. Bij het doen van een dergelijke voorspelling gaat de frequentie over in een kans waarmee het optreden van het verschijnsel verwacht kan worden.

Deze kans zelf houdt echter een onzekerheid in zich, die er mee samenhangt dat de kans afgeleid is uit een steekproef van eindige grootte. Naarmate de steekproef kleiner is geweest, met andere woorden naarmate de empirische verdeling uit minder gegevens is samengesteld zullen de kansen waarmee voorspellingen gedaan worden onzekerder zijn.

In deze nota zal nader worden ingegaan op de mogelijkheid deze onzekerheden vast te stellen.

In paragraaf 2 van nota 186, [STOL, 1963b], werd aangetoond dat voor continue stochastische variabelen steeds geldt dat de kans P dat de stochastische variabele $\underline{x}$ de waarde x_o zal aannemen is

$$P(\underline{x} = x_o) = 0 \qquad (1)$$

Dit is in formule de uitspraak dat $\underline{x} = x_o$ een kans 0 heeft om gerealiseerd te worden. Wel gerealiseerd kan worden de bewering dat

$$x_1 < \underline{x} < x_2 \qquad (2)$$

aangezien gesteld kan worden dat

$$P(x_1 < \underline{x} < x_2) = \int_{x_1}^{x_2} f(u)\, du = p \qquad (3)$$

De bewering (2) zal met overigens gelijke x_1 en x_2 meer waarde hebben - betrouwbaarder zijn - naarmate p in (3) groter is en er dus minder kans is op waarden van $\underline{x}$ buiten het interval voorgesteld in (2).

Stelt men vooraf p vast, bijvoorbeeld op 95% dan wordt (2) het betrouwbaarheidsinterval genoemd met een betrouwbaarheid van 95% of: een risico van 5%, namelijk (1 - p), dat een x die buiten het interval (2)

ligt toch tot de verdeling van $\underline{x}$ behoort.

Het is gebruikelijk het risico dat men accepteert aan te duiden met

$$\alpha = 1 - p$$

De waarden x_1 en x_2 worden de kritieke waarden genoemd.

Bovenstaande redenering kan ook toegepast worden op uitspraken ontleend aan cumulatieve frequentiecurven. Als voorbeeld zal dienen de empirische cumulatieve frequentiecurve voor de dagneerslag in januari te Rottegatspolder (figuur 1). In deze curve zijn opgenomen alle januari-dagsommen (1 tot en met 31 januari) over 10 jaar zodat de curve op 310 gegevens betrekking heeft. De veronderstellingen waaronder deze wijze van werken gerechtvaardigd is zijn uitvoerig toegelicht in I.C.W.-nota 165, [STOL, 1963a].

Van de curve in figuur 1 kan men aflezen dat een overschrijding van 5 mm met een kans van (100 - 82) $\approx$ 20% zal voorkomen of met andere woorden dat deze overschrijding op één enkele dag (bijvoorbeeld 10 januari) gemiddeld tweemaal in 10 jaar zal optreden. Uit (1) volgt nu dat deze voorspelling een kans 0 heeft om gerealiseerd te worden. Zou men in <u>achtereenvolgende</u> (onafhankelijke) reeksen van 10 jaar nagaan welk <u>gemiddeld</u> aantal overschrijdingen voorkomt, dan zal volgens (1) veelal een andere waarde dan 2 worden gevonden. Wel kan gezegd worden dat het gemiddeld aantal overschrijdingen per reeks van 10 jaar zal liggen tussen

$$x_1 = \frac{1 - .75}{10} = 2{,}5 \text{ maal per reeks}$$

en

$$x_2 = \frac{1 - .91}{10} \approx 1 \quad \text{maal per reeks}$$

met een risico van $\alpha = 5\%$ dat het gemiddeld aantal nog groter of nog kleiner blijkt te zijn.

Nu doet zich een volgend probleem voor dat ontstaat wanneer wordt beweerd dat het gemiddeld aantal overschrijdingen inderdaad bijvoorbeeld 2 zal zijn. Het aantal overschrijdingen x per <u>enkele</u> reeks van 10 jaar zal een niet-continue kansverdeling volgen daar x alleen gehele waarden kan aannemen. Echter over een <u>aantal</u> reeksen zal een <u>gemiddelde</u> van $1 < \bar{x} < 2{,}5$ gevonden worden. Op dit punt kan reeds naar figuur 3 verwezen worden. Men kan zich nu afvragen welke afwijkingen van de gemiddelde waarde $x = 2$ nog in een enkele reeks verwacht kunnen worden en met welke kansen deze van 2 afwijkende aantallen zullen voorkomen.

Ook aan dit laatste aspect zal in deze nota aandacht worden geschonken.

2. EIGENSCHAPPEN VAN TOETSEN VOOR FREQUENTIEVERDELINGEN

Bij de te bespreken toetsen wordt de empirische cumulatieve frequentieverdeling geacht te zijn ontstaan uit een steekproef van eindige omvang. Bij het onderling toetsen van twee verdelingen mogen de steekproeven waaruit de curven zijn ontstaan in grootte verschillen. De enige voorwaarde die aan de oorspronkelijke verdelingscurven wordt opgelegd is die van de continuïteit van de verdeling [DRION, 1952, pagina 139].

De toetsingsgrootheid is de maximale verticale afstand D (dus gemeten langs de "kansschaal") tussen twee empirische cumulatieve verdelingscurven. Uit het feit dat deze verticale afstand bij transformatie van de horizontale schaal niet verandert, volgt dat de vorm van de verdelingscurven niet van invloed is op de toets. De toets is dus geheel parametervrij. De nulhypothese H_o is dan dat twee verdelingscurven niet zullen verschillen. Overschrijdt de maximale afstand D een, bij een risico α behorende, kritieke waarde d_α dan wordt de nulhypothese verworpen ten gunste van het alternatief H_1 dat de curven van verschillende kansverdelingen afkomstig zijn.

Bij verwerpen van H_o wordt dus geconcludeerd dat de getoetste verdelingen onderling verschillen doch verdere conclusies kunnen hieraan niet verbonden worden. Het verschil kan zijn een verschil in niveau, een verschil in spreiding, in scheefheid, in het algemeen dus een verschil in vorm.

Het toetsen van twee empirische verdelingscurven onderling wordt uitgevoerd met de "two sample" toets van SMIRNOV. Nauw verwant met deze toets is de "one sample" toets van KOLMOGOROW waarmee op eenvoudige wijze een betrouwbaarheidsinterval geconstrueerd kan worden en waarmee de aanpassing aan een volledig bekende verdeling getoetst kan worden.

3. FORMULERING VAN DE TOETS

In de cursus "Parametervrije Methoden" wordt door DRION [1952] een uiteenzetting van beide toetsen gegeven.

De exacte formulering is deze dat van twee verdelingscurven respectievelijk $F_1(u)$ en $F_2(u)$ elk uitgezet met

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

de kans $P = \alpha$ wordt vastgesteld waarmee een kritieke afstand d_α door de maximale afstand D kan worden overschreden. Hiervoor wordt gevonden

$$P = P[\, D = \max|F_1(u) - F_2(u)| \geq d_\alpha | H_o \,] = 2\sum_{a=1}^{\infty}(-1)^{a-1}\, e^{-2a^2z^2} = \alpha \qquad (4)$$

Deze verdeling geldt onder de nulhypothese. In de laatste uitdrukking is a een index en geldt voor z

$$z = d_\alpha \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \qquad (5)$$

en dus

$$d_\alpha = z \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \qquad (6)$$

Hierin zijn n_1 en n_2 de steekproefgrootten. Uit (4) en (5) volgt voor gegeven n_1 en n_2 een verband tussen d_α en α. De waarde α zal men klein kiezen bijvoorbeeld 10, 5 of 1%.

Bovenstaande formules zijn getabelleerd zodat veelal omgekeerd een kans $P = \alpha$ wordt aangenomen, bijvoorbeeld 5%, waarna in de tabel de corresponderende waarde van z uit (4) wordt opgezocht. Met behulp van (6) wordt dan de kritieke waarde d_α berekend. Wordt voor enig punt voor de afstand tussen de twee verdelingen deze waarde d_α overschreden dan wordt de nulhypothese verworpen. Bij verwerping van de hypothese H_o is het risico $\alpha = 5\%$ dat dit ten onrechte geschiedt.

Beschouwt men vervolgens een theoretische verdeling als een verdeling waarvoor $n_1 \to \infty$ dan gaat (6) over in

$$\lim_{n_1 \to \infty} d_\alpha = \lim_{n_1 \to \infty} z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{z}{\sqrt{n_2}}$$

Vergelijkt men nu een empirische frequentieverdelingscurve, ontstaan uit $n_2 = n$ gegevens, met een theoretische, dan wordt (6) dus

$$d_\alpha = \frac{z}{\sqrt{n}} \qquad (7)$$

Het verband tussen (6) en (7) is met de limiet overigens niet exact weergegeven doch de uitkomst is juist [DRION, 1952, pagina 149: deze beschouwing heeft slechts " mnemotechnische" waarde].

4. EEN BETROUWBAARHEIDSINTERVAL

De eigenschappen van bovengenoemde toets, namelijk het feit dat geen voorwaarden dan alleen continuïteit aan de cumulatieve verdeling worden opgelegd maken het mogelijk op eenvoudige wijze een betrouwbaarheidsinterval voor de gehele verdelingscurve te construeren [KENDALL, 1961, pagina 457]. Wordt de uit een steekproef afkomstige verdelingscurve aangeduid met $S(u)$ en de ware curve met $F(u)$ dan geldt voor het betrouwbaarheidsinterval

$$P[S(u) - d_\alpha \leqq F(u) \leqq S(u) + d_\alpha, \text{ voor elke } u] = 1 - \alpha \qquad (8)$$

Met (8) wordt uitgedrukt dat er een betrouwbaarheidsinterval met breedte $\pm d_\alpha$ om $S(u)$ bestaat zodanig dat het interval de werkelijke $F(u)$ met een kans $(1 - \alpha)$ zal bevatten.

Tenslotte kan uit (7) nog berekend worden welke steekproefomvang n nodig is om een verdelingscurve te krijgen met een vooraf vastgesteld betrouwbaarheidsinterval [zie bijvoorbeeld DIXON and MASSEY, 1951, pagina 257; KENDALL and STUART, 1961, pagina 457].

5. TABELLEN EN TOEPASSING VAN DE TOETSEN

In de bijlage wordt een overzicht gegeven van de belangrijkste waarden die in verschillende tabellen in de literatuur worden gegeven. Steeds zijn de tweezijdige kritieke waarden vermeld.

In het voorbeeld van figuur 1 was de curve dus afkomstig van een steekproef met 310 gegevens. Een betrouwbaarheidsinterval met $\alpha = 5\%$ (risico) volgt nu uit

$$|d_\alpha| = \frac{1.36}{\sqrt{310}} = \frac{1.36}{17.6} = 0.077$$

zodat het interval wordt

$$(F - d_\alpha \ , \ F + d_\alpha)$$

welk gebied in figuur 1 met een arcering is aangegeven.

Er is nu 95% kans dat de ware verdelingscurve geheel in dit interval zal liggen, of statistisch juister uitgedrukt, daar niet de ware curve doch het interval stochastisch is: er is 95% kans dat het interval de ware curve geheel zal bevatten.

Uit de figuur volgt dus nu dat beweerd kan worden dat de procentuele kans waarmee een hoeveelheid van 5 mm op één dag (bijvoorbeeld 10 januari) overschreden zal worden moet liggen tussen 75 en 91%, of tussen 15 maal in 20 jaar en 18 maal in 20 jaar, welke bewering een betrouwbaarheid van 95% heeft.

In figuur 2 staan voor verschillende steekproefgrootten n de intervallen voor $\alpha = 5\%$ uitgezet rond een hypothetische, als rechte weergegeven, verdeling. Daar elke horizontale transformatie geoorloofd is, kan men deze hypothetische curve steeds tot dekking brengen op een empirische curve. In de praktijk zal men figuur 2 het beste kunnen toepassen door de empirische curve getekend op transparant papier, over figuur 2 te verschuiven om zo de gewenste transformatie tot stand te brengen. Het bij het aantal gegevens n behorende interval kan dan steeds op het transparant papier voor elke F-waarde worden overgenomen.

Afhankelijk van n zal de hypothetische verdeling in figuur 2 al of niet verder doorgetrokken kunnen worden. Tot waar het betrouwbaar-

heidsinterval bij een bepaald aantal gegevens n zal lopen is in de figuur aangeduid met "Begrenzing in verband met aantal gegevens".

Ook kan op eenvoudige wijze uit de figuur worden afgeleid hoe groot een steekproef moet zijn voor het bereiken van een vooraf vastgestelde betrouwbaarheid.

Een voorbeeld van toepassing van de toets is bijvoorbeeld te vinden in DE JONGE [1958] deel I pagina 217 en volgende en MILLER and KAHN [1962] appendix G, pagina 464 en volgende.

Voor het geval dat twee empirische verdelingen onderling vergeleken worden, dient de zogenaamde "two sample" toets te worden toegepast. De verdeling van de absolute afstand D (zie (4)) wordt gegeven door MASSEY, [1951 en 1952], doch opgemerkt moet worden dat SIEGEL, [1956, pagina 278, naar GOODMAN, 1954, pagina 167] en LINDGREN, [1962, pagina 401], elk hieruit een verschillende tabel afleiden door verschil in behandeling van de eenzijdige respectievelijk tweezijdige toets.

6. OVERGANG OP DE MEDIAANWAARDEN VAN F

De hierboven besproken toetsingsmethode geldt voor

$$F_< = \frac{m}{n} \qquad (9)$$

Wordt met de mediaan gewerkt (paragraaf 6b van nota 186) dan luidt de betrekking

$$F'_< = \frac{m - 0,3}{n + 0,4} \qquad (10)$$

In (9) en (10) komt m als parameter voor zodat, na eliminatie, voor het verband tussen F en F' gevonden wordt

$$F_< = \frac{n + 0,4}{n} F' + \frac{0,3}{n}$$

Voor een gegeven waarde van n is deze betrekking voor te stellen door een rechte.

Het verschil tussen $F_<$ en $F'_<$ - of F en F' - wordt nu weergegeven door de correctie-afstand c_a namelijk

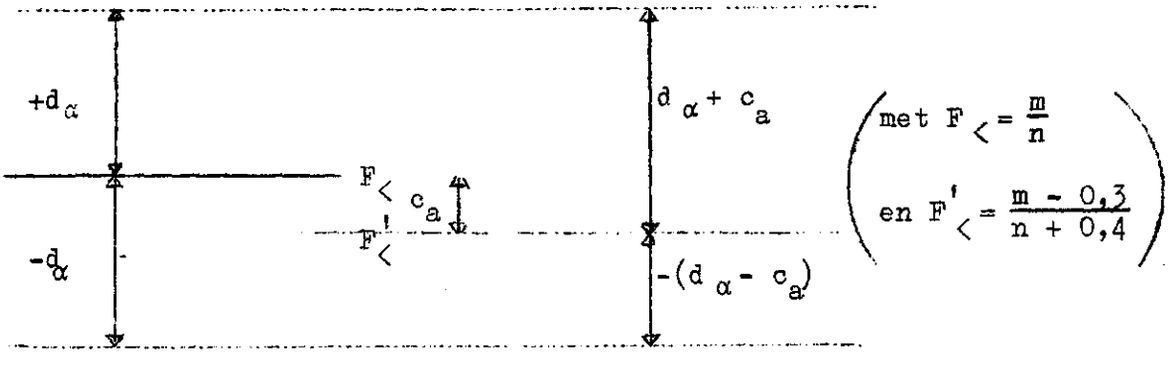
$$c_a = F - F'$$

$$= \frac{n + 0,4}{n} F' - F' + \frac{0,3}{n}$$

$$= \frac{0,4}{n} F' + \frac{0,3}{n} \qquad (11)$$

wat een eenvoudig voorschrift is om bij gegeven n de correctie-afstand c_a voor verschillende waarden van F' te bepalen. Door het lineaire verband kan dit zonodig eenvoudig grafisch plaatsvinden.

Bij de besproken toetsen wordt steeds d_α gemeten vanaf F. Overgang op F' maakt de correctie c_a noodzakelijk teneinde het betrouwbaarheidsinterval rechtstreeks ten opzichte van F' te kunnen uitzetten. In schema:



Hieruit volgt onmiddellijk dat de bovenste grens ligt bij

$$(d_\alpha + c_a)$$

en de onderste bij

$$-(d_\alpha - c_a)$$

indien de gegevens volgens (10) op waarschijnlijkheidspapier zijn uitgezet.

7. AFWIJKINGEN TEN OPZICHTE VAN DE GEMIDDELDE FREQUENTIE

Vervolgens zal worden ingegaan op het probleem dat in de laatste alinea van de inleiding werd aangesneden.

Wordt de kans waarmee een bepaald verschijnsel zal optreden, waarbij gedacht wordt aan overschrijdingen van bijvoorbeeld een bepaalde hoeveelheid neerslag, voorgesteld door $F_{>} = p_o$, dan zal op reeksen van n jaren (figuur 3) gemiddeld

$$n\ p_o \text{ maal per reeks van n jaar} \qquad (12)$$

een dergelijke overschrijding plaatsvinden. In de inleiding werd een voorbeeld gegeven met $p_o = .20$ en $n = 10$ zodat $n\ p_o = 2$

Anderzijds kan gesteld worden dat een enkele overschrijding zal plaatsvinden gemiddeld

$$1 \text{ x per } \frac{1}{p_o} \text{ jaar} \qquad (13)$$

waarin dus $\frac{1}{p_o}$ het aantal jaren voorstelt dat gemiddeld beschouwd moet worden om één zo'n overschrijding te zien voorkomen. Dit aantal jaren wordt gedefinieerd als de herhalingsperiode T [STOL, 1963b, nota 186]

Nu de overschrijdingskans p als vaststaand is aangenomen wordt gevraagd welke waarden het aantal overschrijdingen x wel kan aannemen.

Voor het benaderen van dit probleem kan de volgende redenering worden gevolgd.

Een overschrijding van bijvoorbeeld 5 mm kan worden aangeduid als een succes, het niet overschrijden van 5 mm als het tegengestelde daarvan. Volgens het bovenstaande is nu de kans op succes p en de kans op het tegengestelde $q = 1 - p$.

De kans op precies x malen het voorkomen van een succes (overschrijding) in reeksen van n jaar wordt voorgesteld door de binomiale verdeling en luidt:

$$P(\underline{x} = x) = \binom{n}{x} p^x q^{n-x}, \quad (x = 0, 1, 2, \ldots, n) \qquad (14)$$

[FRASER, 1958, pagina 42 en volgende en pagina 99; FELLER, 1950, pagina 106]

Afgeleid kan worden dat de verwachtingswaarde van $\underline{x}$ in (14) is

$$E(\underline{x}) = np \tag{15}$$

wat overeenkomt met (12). Voor de variantie geldt

$$\sigma^2(\underline{x}) = npq$$

[HEMELRIJK, 1956, pagina 55 en 61]

In aansluiting op het voorgaande kan ook nog gevraagd worden naar het aantal jaren $\underline{n} = n$ dat zal verstrijken voor het optreden van precies x overschrijdingen (successen).

Hiervoor geldt [HEMELRIJK, 1956, pagina 62; FELLER, 1950, pagina 217 en volgende]

$$P(\underline{n} = n) = \binom{n-1}{x-1} p^x q^{n-x}, \quad (n = x, x+1, \ldots) \tag{16}$$

De verwachtingswaarde van deze stochastiek is

$$E(\underline{n}) = \frac{x}{p}$$

en de variantie

$$\sigma^2(\underline{n}) = \frac{xq}{p^2} \tag{17}$$

Wordt het aantal overschrijdingen gelijk aan 1 gesteld zodat het aantal jaren gevraagd wordt waarbinnen precies 1 overschrijding voorkomt, dan wordt verkregen

$$E(\underline{n} \mid x = 1) = \frac{1}{p} \tag{18}$$

waarmee overeenkomend met (13) de herhalingsperiode is verkregen, aangezien

$$\frac{1}{p} = \frac{1}{1 - F_<} = T > 1$$

Uit (17) volgt dan nog voor de variantie

$$\sigma^2(\underline{n} \mid x = 1) = \frac{q}{p^2} = T(T-1)$$

8. VOORBEELD VAN BEREKENING

In het gegeven voorbeeld van paragraaf 1 en paragraaf 5 werd aangenomen dat p = 0.20, zodat in reeksen van n = 10 jaar volgens (15) gemiddeld 2 overschrijdingen zullen voorkomen. Per reeks kan het aantal overschrijdingen nog sterk uiteenlopen hetgeen in figuur 3 is geillustreerd voor het geval dat I: p = 0.10 en II: p = 0.20.

In beide gevallen zullen er reeksen van 10 jaar kunnen voorkomen met 0 overschrijdingen (respectievelijk de reeksen 4 van I en II uit figuur 3), 1 overschrijding (respectievelijk reeks 3 van I en reeks 2 van II), 2 overschrijdingen (reeks 1 en 2) enz. Wel zal gelden dat een aantal van 4 overschrijdingen in een reeks van 10 jaar een grotere kans van voorkomen heeft met p = 0.20 dan met p = 0.10. Deze kansen kunnen met behulp van (14) worden berekend.

Voor een aantal gevallen wordt dan het volgend overzicht verkregen (zie tabel 1)

Tabel 1

Kansen van voorkomen van 0, 1,..., 10 overschrijdingen in reeksen van n = 10 jaar in 3 decimalen.

Aantal overschrijdingen x	Kans waarmee de overschrijding optreedt				
	p=0.10	p=0.20	p=0.40	p=0.70	p=0.80
1	2	3	4	5	6
0	.349	.107	.006		
1	.387	.268	.040		
2	.194	.302	.121	.001	
3	.057	.201	.215	.009	.001
4	.011	.088	.251	.037	.006
5	.002	.026	.201	.103	.026
6		.006	.111	.200	.088
7		.001	.042	.267	.201
8			.011	.233	.302
9			.002	.121	.268
10				.028	.107
$E(\underline{x}) = np$	1	2	4	7	8
$\sigma(\underline{x}) = \sqrt{npq}$	0.95	1.27	1.55	1.45	1.28
$E(\underline{n} \mid x = 1) = T$	10	5	2.50	1.43	1.25
$\sigma(\underline{n} \mid x = 1) = \sqrt{T^2 - T}$	9.49	4.47	1.94	0.80	0.56

In de tabel zijn de kansen die bij de verwachtingswaarde np behoren in een kader geplaatst. Uit de tabel valt af te lezen dat met een gemiddelde van np = 2 overschrijdingen (kolom 3) de kans op het optreden van precies 4 overschrijdingen in n = 10 jaar 8,8% is. Met np = 1 (kolom 2) is deze kans slechts 1,1%. Voor het geval dat np = 2 is bijvoorbeeld de kans op 4 overschrijdingen - of - meer gelijk aan

$$P(\underline{x} \geq 4) = 1 - P(x < 4) = 1 - 0.878 = 12,2\% \qquad (19)$$

Uit figuur 1 volgde dat een hoeveelheid neerslag v van 5 mm op één dag overschreden zal worden met een kans van ongeveer

$$P(\underline{v} > 5 \text{ mm}) = 0.20$$

of wel eens in de 5 jaar. Worden reeksen van 10 jaar beschouwd, dan zal de overschrijding plaatsvinden 2 x in de 10 jaar. Dit geldt voor elke afzonderlijke dag in januari. Om het bovenstaande eens toe te passen werd van elke dag in januari nagegaan hoe vaak op die dag een overschrijding van 5 mm in de jaren 1952 tot en met 1961 heeft plaatsgevonden. Hieruit ontstond het volgende overzicht.

Tabel 2

Overschrijdingen van 5 mm op één dag in januari over 10 jaar

Aantal overschr. x	Aantal reeksen van 10 jaar met x overschr.	Voorkomen van x in %	Binomiaalreeks voor n = 10 en p = 20%	Totaal aantal overschr.
0	3	.10	.11	0
1	11	.35	.27	11
2	11	.35	.30	22
3	4	.13	.20	12
4	1	.03	.09	4
5	1	.03	.03	5
6	0	.00	.00	0
	31	.99	1.00	54

$E(\underline{x}) = np = 2$, betrouwbaarheidsinterval (zie pagina 3) $1 < \bar{x} < 2,5$.
Schatting voor $\bar{x}$, $S(\bar{x}) = \frac{54}{31} = 1,7$.

De gevonden percentages van voorkomen vertonen goede overeenkomst met de theoretische uit de binomiaalreeks. Als schatting voor het gemiddeld aantal van voorkomen van de overschrijding wordt gevonden de waarde 1,7 welke in het betrouwbaarheidsinterval ligt.

Van belang is nog te constateren dat hoewel bij de elementaire kans van 20% het gemiddeld aantal overschrijdingen 2 in de 10 jaar is de kans op 1 overschrijding bijna even groot is als die op precies 2 (namelijk .27 tegen .30) en weer bijna even groot als de kans op 3 of 4 overschrijdingen (.29). Bij een verder doorgevoerde toepassing van de frequentieverdelingen zoals figuur 1 die geeft zal met het voorgaande rekening moeten worden gehouden.

Een toepassing van het bovenstaande wordt nog gevonden in de bepaling van de kans dat er geen enkele overschrijding zal plaatsvinden.

Deze kans kan als volgt worden afgeleid door van (14) uit te gaan:

$$P(\underline{x} = 0) = \binom{n}{0} p^0 q^n$$

daar

$$\binom{n}{0} = \binom{n}{n} = 1$$

en

$$q = 1 - p$$

ontstaat

$$P(\underline{x} = 0) = (1 - p)^n$$

Uit (15) volgt nog

$$P(\underline{x} = 0) = \left(1 - \frac{E(\underline{x})}{n}\right)^n$$

zodat als n groot is en p klein (overschrijdingskans van extreme waarden!) de benadering geldt

$$P(\underline{x} = 0) = e^{-E(\underline{x})} = e^{-np}$$

[FELLER, 1950, pagina 110]

VAN DANTZIG [1954] paste deze betrekking toe op het bepalen van waterhoogten die door stormvloeden niet meer zullen worden overschreden.

Het aantal overschrijdingen in een enkele reeks van 10 jaar kan eveneens worden gebruikt om de gemiddelde overschrijdingskans vast te stellen, zij het ook dat uit korte reeksen slechts een weinig betrouwbare uitspraak kan worden gedaan.

Zo kan bijvoorbeeld op grond van het aantal geconstateerde overschrijdingen in een enkele reeks in figuur 3 getoetst worden of inderdaad waar kan zijn dat $p = 0.20$ (Geval II). In reeks 1 is $x = 2$ welke waarde een kans van voorkomen van 30.2% heeft (tabel 1, kolom 3), een kans die groot genoeg is om $p = 0.20$ te aanvaarden.

Zou aan de hand van de uitkomst in reeks 6 namelijk $x = 4$ worden getoetst of $p = 0.20$ waar kan zijn, dan wordt gevonden (kolom 3 van tabel 1) (zie (19))

$$P(\underline{x} \geq 4) = 12,2\%$$

Bij een tweezijdig betrouwbaarheidsgebied van 5% is deze kans groot genoeg om $p = 0.20$ te aanvaarden.

Was de hypothese geweest $p = 0.10$ dan zou (kolom 2 van tabel 1)

$$P(\underline{x} \geq 4) = 1,3\ \%$$

de kans op het geconstateerde of een groter aantal overschrijdingen is nu zo klein dat de hypothese $p = 0.10$ niet waarschijnlijk meer is en dus wordt verworpen. De kritieke waarden voor $\alpha = 5\%$, tweezijdig, zijn in tabel 1 met streeplijntjes gemarkeerd.

Uit tabel 1 kan nu worden afgeleid dat indien een verschijnsel 4 x in 10 jaar optreedt de waarde van p kan liggen tussen 0.20 en 0.70. Het aantal jaren $n = 10$ is overigens te gering om een waardevoller uitspraak te doen.

De kansen van voorkomen zoals gegeven in tabel 1 zijn in vele vormen getabelleerd. Genoemd kunnen worden MUILWIJK en SCHOUTEN [1960] met tabellen voor verschillende waarden van p en voor n = 5, 10, 15, 20, 25, 30. HALD [1960a, pagina 677] geeft een tabel voor n = 50, 100. DE JONGE [1958, deel I, pagina 84 en 86] geeft voor n = 20 de kansen voor enkele zeer kleine waarden van p; de bijbehorende toetsingen worden besproken op pagina 156 en volgende. KUIPER [1959, pagina 168] geeft cumulatieve kansen voor $p = 0.50$ en $n = 6$ tot en met 20 met

x = 0 tot en met 6. Nog vollediger worden deze kansen gegeven door SIEGEL [1956, pagina 250]. Nomogrammen voor het toepassen van de binomiale toets worden onder andere gegeven in DIYON and MASSEY [1951].

Een uitvoerige tabel voor het vaststellen van een interval voor p uit een geschatte waarde (zie vorige alinea) is te vinden in HALD [1960b, pagina 66 tot en met 69] terwijl in nomogramvorm deze intervallen gegeven worden door DE JONGE [1958, deel I, pagina 161].

9. NABESCHOUWING EN SAMENVATTING

In de nota's 186 en 167 is een uiteenzetting gegeven over enkele aspecten die nauw samenhangen met het werken met frequentieverdelingen. Speciaal ook de wijze waarop waarnemingsuitkomsten op waarschijnlijkheidspapier kunnen worden uitgezet verkreeg de aandacht. In het kort zouden van de verschillende mogelijkheden de volgende positieve eigenschappen kunnen worden genoemd.

Het <u>frequentiequotient</u> $F = \frac{m}{n}$ heeft het voordeel dat bij gebruik van deze waarde als "plotting position" op eenvoudige wijze een betrouwbaarheidsinterval geconstrueerd kan worden.

Het gebruik van de <u>mediaanwaarde</u>

$$F' = \frac{m - 0{,}3}{n + 0{,}4}$$

heeft het voordeel dat voor elke punten geldt dat de kans dat het punt te hoog respectievelijk te laag is uitgezet 0.50 is. Deze eigenschap geldt voor elk type waarschijnlijkheidspapier [VAN DANTZIG, 1954, pagina 224].

Op pagina 10 werd aangetoond dat het uitzetten van een betrouwbaarheidsinterval rond F' kan plaatsvinden door een correctie c_a in rekening te brengen, waarvoor een eenvoudige formule kon worden afgeleid.

Het gebruik van de <u>verwachtingswaarde</u> [nota 186]

$$F = \frac{m}{n + 1}$$

heeft het voordeel dat er een herhalingsperiode T uit wordt gevonden die nauw aansluit bij die welke men rechtstreeks uit de gegevens berekent. Het is echter niet de verwachtingswaarde of de mediaan van <u>T</u> die op deze wijze wordt gevonden.

Voor grafische bewerking van de gegevens heeft het gebruik van de mediaanwaarde de voorkeur. Gecombineerd hiermee kan worden de "one sample" toets van KOLMOGOROW. De vrijheid*) die men binnen het be-

*) welke uit een statistische onzekerheid voortspruit!

trouwbaarheidsinterval, waarbinnen de ware curve behoudens een risico α zal liggen, heeft kan benut worden voor het opleggen van een samenhang tussen de verdelingscurven volgens bijvoorbeeld de maanden van het jaar, verschillende tijdvaklengten (1,2..., k daagse sommen) enz.

Uit figuur 2 blijkt dat het aantal gegevens dat men nodig heeft om ook voor "zeldzaam" voorkomende gebeurtenissen een "smal" betrouwbaarheidsinterval te vinden kan worden gesteld op tenminste 1000. Voor precies 1000 gegevens blijkt een overschrijding die volgens de verdelingscurve gemiddeld eens in de 20 jaar kan voorkomen met 95% kans te liggen tussen de waarden die overeenkomen met een optreden van eens in de 10 jaar tot eens in de 100 jaar.

Deze onzekerheid kan bij empirische verdelingscurven worden weggenomen door het aantal gegevens op te voeren. Veelal gebeurt dit door het waarnemingsmateriaal met binnen een groep opschuivende tijdvakken uit te breiden [STOL, 1963a, nota 165]. Dat hiermee persistentie in het materiaal gaat optreden wordt dan veelal als een bijkomstig nadeel aanvaard.

De restrictie "empirische" in de vorige alinea werd gemaakt omdat voor dit type verdelingscurven de gememoreerde parametervrije toetsen werden afgeleid. Zodra van een verdeling de vorm theoretisch bekend is kan een toets worden toegepast op de parameter van de verdelingscurve. Dit soort toetsen leidt tot een nauwer betrouwbaarheidsinterval hetgeen plausibel wordt door te bedenken dat de "hoofdvorm" van de curve reeds vast staat [HALD, 1960, pagina 139, GOODMAN 1954, pagina 162-163].

In verband met het bovenstaande kan hier nog de χ^2-toets worden genoemd die eveneens kan dienen voor het onderzoek naar de aanpassing aan frequentieverdelingen. Het nadeel van de χ^2-toets hier is echter dat dan het materiaal in klassen moet worden ingedeeld en het resultaat van deze indeling kan afhangen [DRION, 1952]. Voorts wordt het indelen in klassen bezwaarlijk bij steekproeven van kleine omvang. Doch juist voor dit soort gevallen is de toets van KOLMOGOROW of die van SMIRNOV uitermate geschikt [KENDALL, 1961, pagina 452 en 458 en SIEGEL, 1956 pagina 51].

Er moet nog op worden gewezen dat uit de overschrijdingskansen van een verschijnsel wel kan worden afgeleid welk aantal malen dat verschijnsel zich binnen een zeker aantal jaren zal herhalen doch niet wanneer dat het geval is. Ook om dit aspect te illustreren werd figuur 3 samengesteld.

Voorts moet nog worden opgemerkt dat het optreden van een overschrijding in de reeksen van telkens 10 jaar in figuur 3 onafhankelijke gebeurtenissen moeten zijn die elkaar niet beïnvloeden. Bovendien zullen voorspellingen slechts geldig zijn wanneer de kansverdelingen in de loop van de jaren gelijk blijven. Bij het beschouwen van een groot aantal reeksen, bijvoorbeeld 10 of meer reeksen van telkens 10 jaar kan dit niet als vanzelfsprekend worden aangenomen.

Beschouwt men slechts een enkele reeks van n toekomstige jaren dan kan een schatting gemaakt worden van het interval waarbinnen het aantal overschrijdingen, behoudens een kans α, zal liggen.

Literatuur

DANZIG, D. VAN, 1954. Mathematical problems raised by the flood disaster 1953. Proc. of the International Congress of Mathematicians. Amsterdam.

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 S76 van het Mathematische Centrum te Amsterdam.
(I.C.W. 11/3)

FELLER, W., 1950. An introduction to probability theory and its applications. Vol. I, New York.
(I.C.W. 11/23)

FRASER, D.A., 1958. Statistics, an introduction. New York.
(I.C.W. 11/109)

FISZ, M., 1962. Wahrscheinlichkeitsrechnung und Mathematische Statistik. Berlijn.
(I.C.W. 11/182)

GOODMAN, L.A., 1954. Psychological Bulletin, 51.

HALD, A., 1960a. Statistical Theory with Engineering Applications. New York.
(I.C.W. 11/169)

____ , 1960b. Statistical Tables and Formulas, New York.
(I.C.W. 11/147)

HEMELRIJK, J., 1956. Syllabus van een oriënterende cursus Mathematische Statistiek. Rapport S200 (C8). Mathematisch Centrum, Amsterdam.

JONGE, H. DE, 1958. Inleiding tot de medische Statistiek, deel I, Leiden.
(I.C.W. 11/102)

JONGE, H. DE, 1960. Inleiding tot de medische Statistiek, deel II, Leiden.
(I.C.W. 11/102)

KENDALL, M.G. and S. STUART, 1961. The advanced theory of statistics. Vol. 2. Inference and relationship. London.
(I.C.W. 11/114)

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

LINDGREN, B.W., 1962. Statistical Theory. New York.
(I.C.W. 11/207)

MASSEY, F.J., 1952. Distribution table for the deviation between two sample cumulatives. Anuals of Mathematical Statistics 23.

MILLER, R.L. and J.S. KAHN, 1962. Statistical analysis in the geiological sciences. London.
(I.C.W. 11/208)

MUILWIJK, J. en J.H. SCHOUTEN, 1960. Inleiding tot de wiskundige statistiek. Deel 3: tabellen. 's-Gravenhage.
(I.C.W. 11/195)

SIEGEL, S. 1956. Nonparametric Statistics for the behavioral sciences. New York.
(I.C.W. 11/48)

STOL, Ph.Th., 1963a. Het gebruik van frequentieverdelingen bij het onderzoek naar afvoercoëfficiënten.
(I.C.W. nota 165)

———, 1963b. Cumulatieve frequentieverdelingscurven (I). Het uitzetten van cumulatieve frequentieverdelingen.
(I.C.W. nota 186)

Overzicht van enkele waarden uit verschillende tabellen met betrekking tot de toetsen van Kolmogorow en Smirnov

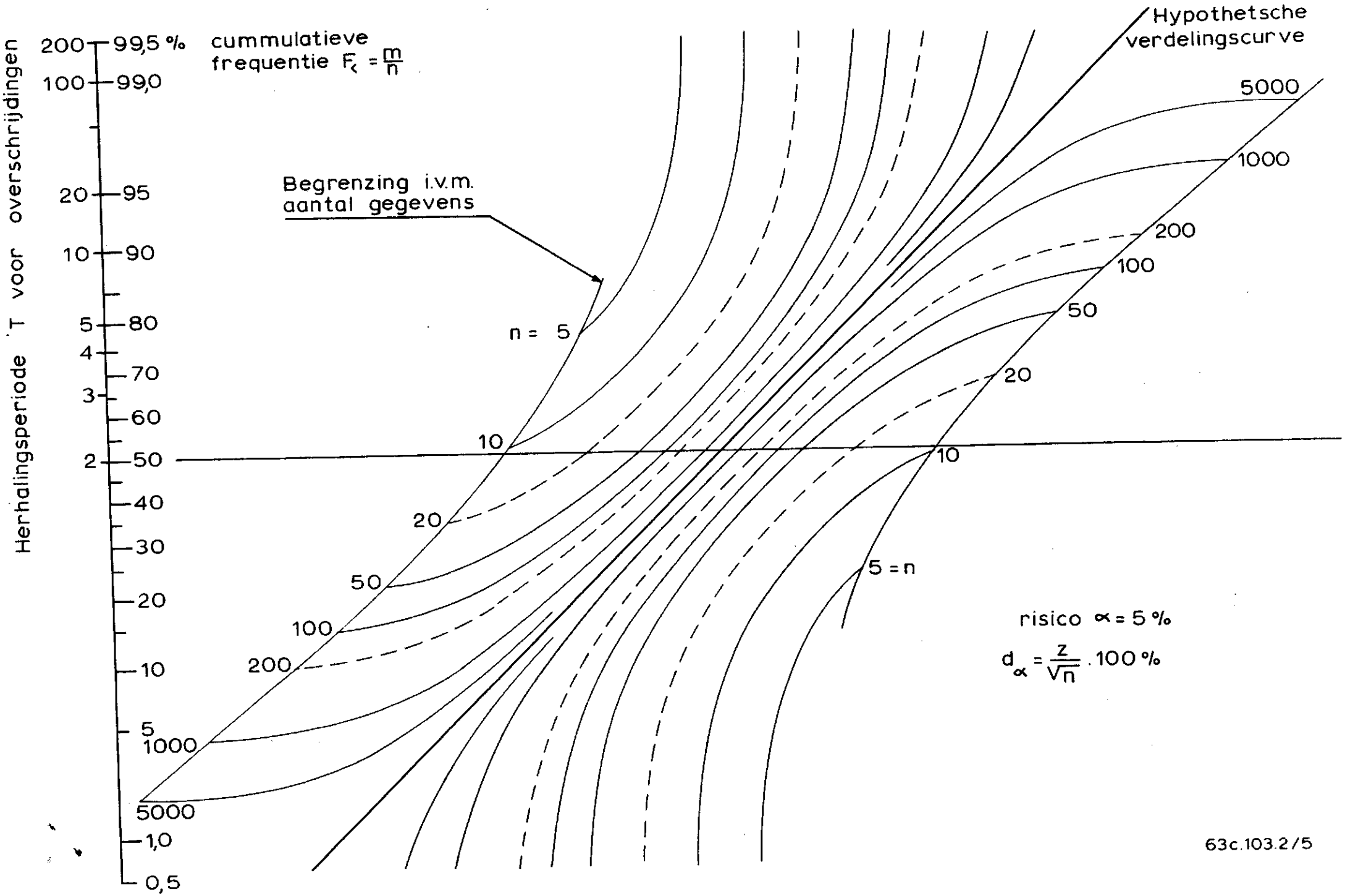
DRION pag.148 / FISZ pag.510		Kritieke waarden voor $\|d_\alpha\|$ bij verschillende steekproefgrootten $n_1 = n_2 = n$ "one sample"-toets										Kritieke waarden voor $\|d_\alpha\|$ "two sample"-toets
		DIXON and MASSEY, pag.348 DE JONGE, deel I, pag.317 SIEGEL, pag. 251								LINDGREN pag.400	MILLER and KAHN pag.468	SIEGEL pag.279
z	α	5	10	15	20	30	40	50	>50	100	300	$n_1>40,\ n_2>40,\ n_1 \neq n_2$
0.828	0.50											
1.073	0.20	.45	.32		.23	.19	.17	.15	$1.073/\sqrt{n}$			
1.224	0.10	.51	.37		.26	.22	.19	.17	$1.224/\sqrt{n}$			$1.224\sqrt{(n_1+n_2)/n_1n_2}$
1.358	0.05	.56	.41		.29	.24	.21	.19	$1.358/\sqrt{n}$	.14	.08	$1.358\sqrt{(n_1+n_2)/n_1n_2}$
1.517	0.02											
1.625	0.01	.67	.49		.36	.29	.25	.23	$1.625/\sqrt{n}$			$1.625\sqrt{(n_1+n_2)/n_1n_2}$
1.731	0.005											

0.05	"two sample"-toets LINDGREN, pag.401 MASSEY, 1952		
n_1 \ n_2	5	10	15
5	4/5	7/10	10/15
10		6/10	15/30
15			7/15

Bijlage bij nota 187

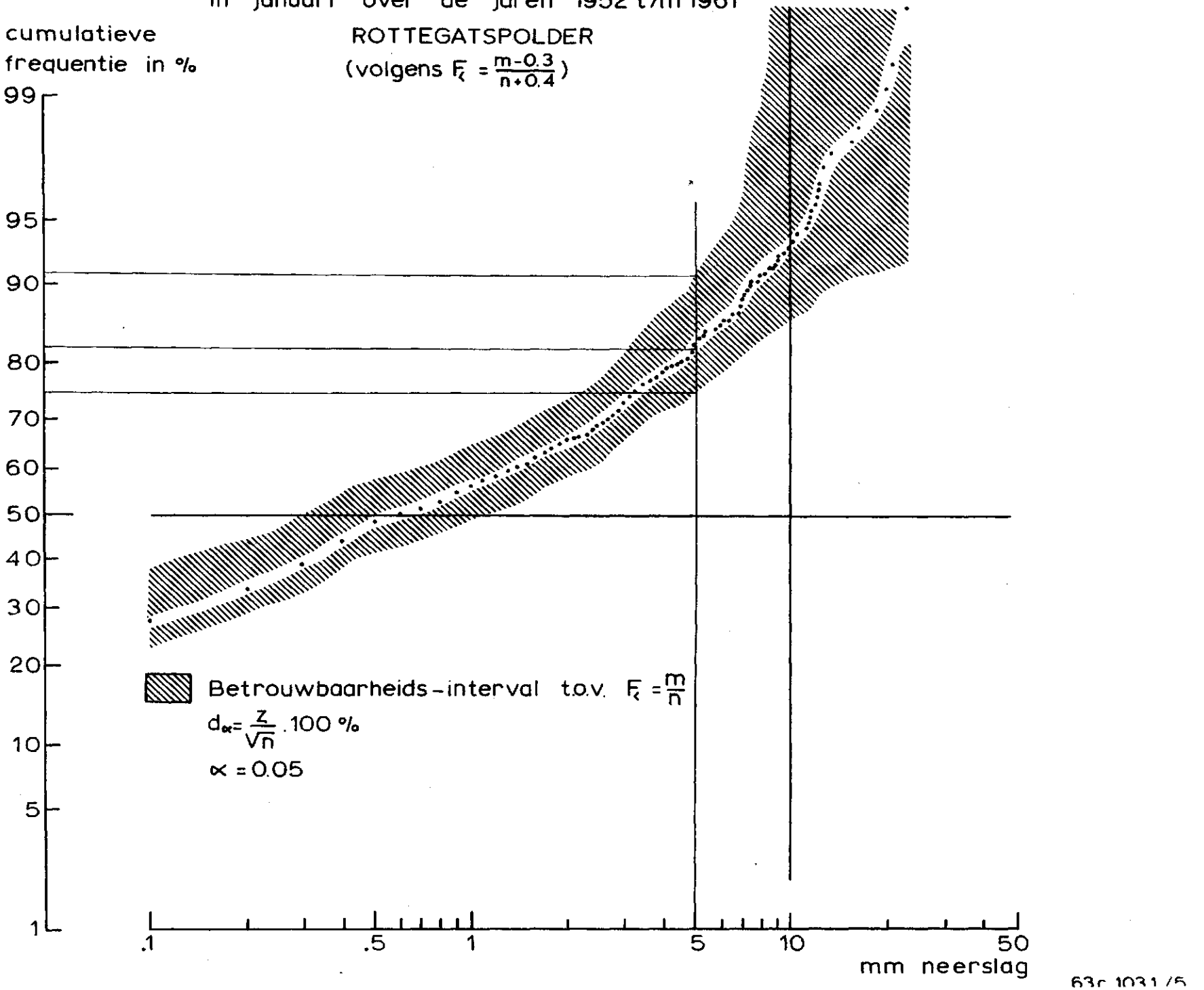
Betrouwbaarheidsintervallen voor een hypothetische verdelingscurve

Fig. 2



Frequentieverdeling van de dagneerslagsommen
in januari over de jaren 1952 t/m 1961

fig. 1



63c 103 1/5