 TA 656[I]

11 februari 1972

Instituut voor Cultuurtechniek en Waterhuishouding
Wageningen

EEN TOEPASSING VAN HET GEBRUIK VAN REEKSONTWIKKELING VOOR EMPIRISCHE FREQUENTIEVERDELINGEN OP NEERSLAGGEGEVENS

Mej. M.G. van Steenbergen

Nota's van het Instituut zijn in principe interne communicatiemiddelen, dus geen officiële publikaties.
Hun inhoud varieert sterk en kan zowel betrekking hebben op een eenvoudige weergave van cijferreeksen, als op een concluderende discussie van onderzoeksresultaten. In de meeste gevallen zullen de conclusies echter van voorlopige aard zijn omdat het onderzoek nog niet is afgesloten.
Bepaalde nota's komen niet voor verspreiding buiten het Instituut in aanmerking

I N H O U D

1. INLEIDING

In de mathematische statistiek worden methoden gegeven om de functie voor de normale verdeling in een machtreeks te ontwikkelen. Deze machtreeks kan vervolgens gebruikt worden - door juiste parameterkeuze - om empirische verdelingen in formulevorm weer te geven.

De bedoelde reeks is slechts onder strenge voorwaarden convergent. Voor de bruikbaarheid van de reeks is echter van meer belang dat de termen snel in betekenis blijken af te nemen. Convergentie is niet waardevol als voor een goede benadering veel termen in de berekening opgenomen moeten worden.

In het volgende zullen enkele voorbeelden van het gebruik van reeksontwikkeling gegeven worden.

2. DE GRAM-CHARLIER A REEKS

Aangenomen wordt dat de te benaderen verdelingsdichtheidsfunctie f(t) in een reeks ontwikkeld kan worden waarin de termen de afgeleiden van de functie van de normale verdeling zijn. Met willekeurige coëfficiënten c_i wordt de reeks

$$f(t) = c_0\, \phi(t) + c_1\, \phi^{(1)}(t) + \ldots + c_n\, \phi^{(n)}(t) + \ldots$$

waarin $\phi(t) = (2\pi)^{-\frac{1}{2}} e^{-t^2/2}$ en $\phi^{(n)}(t) = \dfrac{d^n}{dt^n}\phi(t)$.

Door integratie worden de constanten gevonden

$$c_0 = 1 \; , \; c_1 = 0 \; , \; c_2 = 0 \; , \; c_3 = -\gamma_1/3! \; , \; c_4 = \gamma_2/4!$$

Hierin zijn γ_1 en γ_2 de uit de statistiek bekende karakteristieken voor een kansverdeling respectievelijk de scheefheidscoëfficiënt en de welvingscoëfficiënt.

Invoering van de uitdrukkingen voor de coëfficiënten in f(t) geeft

$$(2.1) \qquad f(t) = \phi(t) - \frac{\gamma_1}{3!}\phi^{(3)}(t) + \frac{\gamma_2}{4!}\phi^{(4)}(t) + \ldots$$

Voor de afleiding van deze formule zij verwezen naar KENDALL and STUART (1958), CRAMER (1961) en KENNEY and KEEPING (1959).

Een nadeel van de Gram-Charlier A reeks is dat de benadering niet steeds verbetert als meer termen worden toegevoegd. De coëfficiënten in (2.1) nemen niet geleidelijk af als de orde van de afgeleiden toeneemt.

Zo kan men aantonen dat c_3, c_4 en c_5 respectievelijk van de orde $n^{-\frac{1}{2}}$, n^{-1} en $n^{-\frac{3}{2}}$ zijn, maar c_6 is weer van de orde n^{-1}.
Wordt $\phi^{(4)}(t)$ als term toegevoegd, dan zou dus ook $\phi^{(6)}(t)$ moeten worden toegevoegd.

Beter is dan ook de reeks die door Edgeworth werd geïntroduceerd en dit bezwaar niet kent.

3. DE EDGEWORTH REEKS

Vanuit een geheel ander uitgangspunt werd door Edgeworth een soortgelijke reeks gevonden.
Aangetoond is (CRAMER, 1961) dat deze reeks onder vrij algemene voorwaarden een asymptotische ontwikkeling geeft van f(t) in machten van $n^{-\frac{1}{2}}$, met een restterm van dezelfde orde als de eerste verwaarloosde term.
De formule luidt

$$(3.1) \quad f(t) = \phi(t) - \frac{\gamma_1}{3!}\phi^{(3)}(t) + \frac{\gamma_2}{4!}\phi^{(4)}(t) + \frac{10\,\gamma_1^2}{6!}\phi^{(6)}(t) + \ldots$$

Voor een goede benadering is dit aantal termen meestal nodig en voldoende.

Opm.: Voor grote waarden van t kan f(t) negatief worden. Dit hangt samen met de alternerende tekens van de reeks en het afbreken na een eindig aantal termen. Indien de negatieve waarden buiten het gebied van interesse optreden blijft de reeks bruikbaar.

4. AANPASSEN VAN DE REEKSEN AAN EEN EMPIRISCHE VERDELING

Voor het toepassen van deze reeksontwikkelingen wordt er van uitgegaan dat de waarnemingen x_j gegroepeerd zijn met klassemiddens x_i en dat de frequentie in de i-de klasse f_i is. Op de gebruikelijke wijze worden schattingen voor het gemiddelde ($\hat{\mu}$) en de spreiding ($\hat{\sigma}$) berekend volgens

$$\hat{\mu} = \frac{\sum_i x_i f_i}{\sum_i f_i}$$

$$\hat{\sigma} = \frac{\sum_i (x_i - \hat{\mu})^2}{\sum_i f_i - 1}$$

De waarnemingsuitkomsten worden gestandaardiseerd:

$$t_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

Om de coëfficiënten uit (3.1) te vinden moeten nog de volgende netto momenten

$$\hat{\mu}_3 = \frac{\sum_i (x_i - \hat{\mu})^3}{\sum_i f_i - 1}$$

en

$$\hat{\mu}_4 = \frac{\sum_i (x_i - \hat{\mu})^4}{\sum_i f_i - 1}$$

worden berekend.
Er geldt

$$\gamma_1 = \alpha_3 = \frac{\hat{\mu}_3}{\hat{\sigma}^3}$$

en

$$\gamma_2 = \alpha_4 - 3 = \frac{\hat{\mu}_4}{\hat{\sigma}^4} - 3$$

Voor een normale verdeling geldt $\gamma_1 = 0$ en $\gamma_2 = 0$.
De benodigde afgeleiden van $\phi(t)$ zijn

$$\phi^{(3)}(t) = \phi(t)\,(3\,t - t^3)$$

$$\phi^{(4)}(t) = \phi(t)\,(t^4 - 6\,t^2 + 3)$$

$$\phi^{(6)}(t) = \phi(t)\,(t^6 - 15\,t^4 + 45\,t^2 - 15)$$

De formule van de Gram-Charlier A reeks tot en met $\phi^{(4)}(t)$ is

$$f(t) = (2\pi)^{-\frac{1}{2}}\, e^{-t^2/2}\left[1 - \frac{\gamma_1}{6}(3\,t - t^3) + \frac{\gamma_2}{24}(t^4 - 6\,t^2 + 3)\right] \tag{4.1}$$

De formule van de Edgeworth reeks tot en met de termen van de orde n^{-1} is

$$f(t) = (2\pi)^{-\frac{1}{2}}\, e^{-t^2/2}\left[1 - \frac{\gamma_1}{6}(3\,t - t^3) + \frac{\gamma_2}{24}(t^4 - 6\,t^2 + 3) + \frac{\gamma_1^2}{72}(t^6 - 15\,t^4 + 45\,t^2 - 15)\right] \tag{4.2}$$

Om de verwachte frequenties per klasse te berekenen moet $f(t_i)$ worden vermenigvuldigd met

$$\left(\sum_i f_i\right) \cdot \Delta x_i/\hat{\sigma}$$

waarin Δx_i de klassebreedte van de i-de klasse is.
In een histogram, waar op de verticale as de absolute frequentie f_i is uitgezet, kan de dichtheidskromme worden ingetekend.
Wordt de geschatte waarde voor f_i aangeduid met $\hat{f}_i$ dan behoort bij het klassemidden x_i de ordinaatwaarde

$$\hat{f}_i = f(t_i) \cdot \sum_i f_i \cdot \Delta x_i/\hat{\sigma}$$

5. TOEPASSING

Teneinde de genoemde reeksen (4.1) en (4.2) op hun bruikbaarheid te onderzoeken, zijn deze toegepast op een duidelijk scheve en op meer symmetrische verdelingen.

Voor de eerste werd als voorbeeld gekozen een verdeling van neerslagsommen van k = 1 dag. Voor de tweede en derde een verdeling van neerslagsommen van respectievelijk k = 30 en k = 360 dagen.

5.1. Toepassing op de verdeling van 1-daagse neerslagsommen op Vlissingen, januari 1905 - 1953

De resultaten van de berekening zijn weergegeven in tabel 5.1. (bijlage 1).

x_i = neerslag in 0,1 mm

$$\sum_i f_i = 1457$$

de netto momenten zijn

$$\hat{\mu}_1 = \hat{\mu} = 18,789$$
$$\hat{\mu}_2 = \hat{\sigma}^2 = 944,43 \qquad \hat{\sigma} = 30,732$$
$$\hat{\mu}_3 = 74\,649,6$$
$$\hat{\mu}_4 = 10\,903\,393$$

scheefheidscoëfficiënt $\gamma_1 = 2,57198$

welvingscoëfficiënt $\gamma_2 = 9,2242$

In fig. 1 zijn de empirische verdeling en de benaderingen getekend. Voor een zo extreem scheve verdeling schieten beide functies (4.1 en 4.2) tekort. De Edgeworth-reeks is hier duidelijk beter dan de Gram-Charlier A reeks, maar vooral aan het begin is de aanpassing slecht. Een betere aanpassing zou verkregen kunnen worden door het uitbreiden van formule 4.2 met termen van een hogere orde. Dit heeft als bezwaar dat daarvoor een schatting voor het 5-de netto moment ingevoerd moet worden, welke nog sterker dan de voorgaande momenten aan toevallige afwijkingen onderhevig is.

5.2. Toepassing op de verdeling van 30-daagse neerslagsommen op Vlissingen, augustus 1905 - 1953

Als meer 'normale' verdeling werd gekozen de neerslagverdeling over 30 dagen (tabel 5.2, bijlage 2). De neerslag x_i is hier gegeven in mm. De benodigde parameterwaarden zijn:

$$\sum_i f_i = 48$$

netto momenten:

$$\hat{\mu}_1 = \hat{\mu} = 66,67$$

$$\hat{\mu}_2 = \hat{\sigma}^2 = 1\,463,38 \qquad \hat{\sigma} = 38,25410$$

$$\hat{\mu}_3 = 31\,586,647$$

$$\hat{\mu}_4 = 5\,323\,134,5$$

scheefheidscoëfficiënt $\gamma_1 = 0,564246$

welvingscoëfficiënt $\gamma_2 = 0,514274$

De verdere resultaten staan weergegeven in tabel 5.2 (bijlage 2). Fig. 2 toont aan dat beide reeksen hier een redelijke benadering geven. Onderling vertonen ze weinig verschillen.

5.3. Toepassing op de verdeling van 360-daagse neerslagsommen op Vlissingen, jaren 1905 - 1953

De 360-daagse neerslagsommen zijn bij benadering normaal verdeeld. De parameterwaarden zijn:

$$\sum_i f_i = 48$$

netto momenten:

$$\hat{\mu}_1 = \hat{\mu} = 695,625$$

$$\hat{\mu}_2 = \hat{\sigma}^2 = 15\,059,18 \qquad \hat{\sigma} = 122,7158$$

$$\hat{\mu}_3 = 48\,646,78$$

$$\hat{\mu}_4 = 582\,811\,758$$

scheefheidscoëfficiënt $\gamma_1 = 0,0263240$

welvingscoëfficiënt $\gamma_2 = -0,430043$

Tabel 5.3 (bijlage 2) geeft de resultaten weer met de neerslag x_i in mm. De aanpassing van de Edgeworth reeks is hier achterwege gelaten, omdat de scheefheidscoëfficiënt zo weinig van 0 afwijkt dat de laatste term van (4.2) verwaarloosbaar klein wordt en de reeksen daarmee, voor zover het de in de berekening voorkomende termen betreft, aan elkaar gelijk zijn.

Fig. 3 laat zien dat de Gram-Charlier A reeks weinig afwijkt van de normale verdeling.

6. SAMENVATTING EN CONCLUSIE

6.1. Praktische consequenties

Het rekenwerk verloopt in twee fasen. Eerst moeten de verschillende parameters worden geschat. Op blz. 56 e.v. van KENDALL en STUART (1958) wordt uiteengezet hoe de netto momenten kunnen worden berekend. Als de parameters bekend zijn worden t_i en $\phi(t)$ berekend. Daarna kunnen de verdelingsfuncties worden berekend.
Voor deze berekening kan gebruik worden gemaakt van bijvoorbeeld de Olivetti-programma 101. In een eenvoudig programma worden formule 4.1 en formule 4.2 gelijktijdig doorgerekend.

6.2. Theoretische consequenties

Zoals in par. 3 al werd opgemerkt hebben de reeksen het bezwaar dat negatieve waarden op kunnen treden, ook in het van interesse zijnde gebied.

In de tweede plaats wijzen de voorbeelden uit dat de toepassingsmogelijkheden beperkt zijn. Extreem scheve verdelingen kunnen blijkbaar niet op bevredigende wijze door een van de reeksen worden benaderd.

Een overzicht van de resultaten geeft het volgende beeld:

Samenvatting van uitkomsten

Neerslagverdeling over:	1-dag	30-dagen	360-dagen
$\hat{\mu}$ (mm)	18,8	66,7	695,6
$\hat{\sigma}$ (mm)	30,7	38,3	122,7
γ_1	2,57	0,56	0,026
γ_2	9,22	-0,51	-0,43
aanpassing Gram-Charlier reeks	zeer slecht	redelijk	goed
aanpassing Edgeworth-reeks	slecht	redelijk	goed

Uit bovenstaande tabel zou met enig voorbehoud de conclusie getrokken kunnen worden dat het gebruik van de reeksen alleen nuttig kan zijn als $\gamma_1 < 0,5$.

7. LITERATUUR

CRAMER, H. Mathematical methods of statistics, pp 221 t/m 231. Princeton 1961 (ICW 11/143).

KENDALL M.G. and A. STUART. The advanced theory of statistics volume 1, pp 148, 156 t/m 160. London 1958 (ICW 11/114).

KENNEY, J.F. and E.S. KEEPING. Mathematics of statistics part two, pp 107, 108. New York 1959 (ICW 11/35)

KONINKLIJK NEDERLANDS METEOROLOGISCH INSTITUUT. Frequenties van k-daagse neerslagsommen op Nederlandse stations, deel 7 Vlissingen 1905 - 1953.

Tabel 5.1. Verdeling van 1-daagse neerslagsommen op Vlissingen, januari 1905-1953

x_i klasse-midden	f_i abs. freq.	t_i	$\phi(t)$	f(t) Gram-Charlier	f(t) Edgeworth	$\hat{f}_i$ normale verdeling	$\hat{f}_i$ Gram-Charlier	$\hat{f}_i$ Edgeworth
0	516	-0,611	0,33101	0,673	0,666		31,9	31,6
1	79	-0,579	0,33738	0,703	0,655		33,3	31,1
2	76	-0,546	0,34370	0,733	0,641		34,8	30,4
3	45	-0,514	0,34957	0,760	0,627		36,0	29,7
4	41	-0,481	0,35536	0,786	0,610		37,3	28,9
5	32	-0,449	0,36069	0,809	0,592		38,4	28,1
6	29	-0,416	0,36587	0,830	0,573		39,4	27,2
7	27	-0,384	0,37059	0,849	0,553		40,3	26,2
8	16	-0,351	0,37511	0,866	0,532		41,1	25,2
9	21	-0,319	0,37915	0,879	0,511		41,7	24,2
12	88	-0,221	0,38932	0,904	0,444		214,3	105,3
17	63	-0,058	0,39827	0,884	0,341		209,6	80,8
22	48	0,104	0,39679	0,792	0,262		187,7	62,1
27	51	0,267	0,38497	0,637	0,217		151,0	51,4
32	36	0,430	0,36371	0,444	0,204		105,3	48,4
37	37	0,593	0,33462	0,240	0,209		56,9	49,5
42	27	0,755	0,30001	0,053	0,217		12,6	51,4
47	29	0,918	0,26177	-0,096	0,214		- 22,8	50,7
52	23	1,081	0,22241	-0,193	0,190	NIET BEREKEND	- 45,8	45,0
57	22	1,243	0,18425	-0,234	0,146		- 55,5	34,6
62	18	1,406	0,14847	-0,226	0,089		- 53,6	21,1
67	13	1,569	0,11651	-0,181	0,031		- 42,9	7,3
72	20	1,731	0,08918	-0,117	-0,018		- 27,7	- 4,3
77	12	1,894	0,06637	-0,046	-0,049		- 10,9	- 11,6
82	16	2,057	0,04810	0,017	-0,059		4,0	- 14,0
87	12	2,220	0,03394	0,066	-0,052		15,6	- 12,3
92	10	2,382	0,02338	0,098	-0,031		23,2	- 7,3
97	10	2,545	0,01565	0,112	-0,005		26,5	- 1,2
104,5	8	2,789	0,00816	0,108	0,031		51,2	14,7
114,5	10	3,114	0,00313	0,078	0,056		37,0	26,5
124,5	5	3,440	0,00107	0,045	0,052		21,3	24,7
134,5	7	3,765	0,00033	0,021	0,035		10,0	16,6
144,5	5	4,091	0,00009	0,009	0,019		4,3	9,0
154,5	0	4,416	0,00002	0,003	0,007		1,4	3,3
164,5	0	4,741	0,00001	0,002	0,006		0,9	2,8
174,5	0	5,067	0	0	0		0	0
184,5	0	5,392	0	0	0		0	0
194,5	1	5,718	0	0	0		0	0
204,5	1	6,043	0	0	0		0	0
214,5	1	6,368	0	0	0		0	0
224,5	1	6,694	0	0	0		0	0
234,5	0	7,019	0	0	0		0	0
244,5	0	7,344	0	0	0		0	0
254,5	0	7,670	0	0	0		0	0
264,5	0	7,995	0	0	0		0	0
274,5	1	8,321	0	0	0		0	0

Bijlage 2

Tabel 5.2. Verdeling van 30-daagse neerslagsommen op Vlissingen, augustus 1905-1953

x_i klasse-midden	f_i abs. freq.	t_i	$\phi(t)$	f(t) Gram-Charlier	f(t) Edgeworth	$\hat{f}_i$ normale verdeling	$\hat{f}_i$ Gram-Charlier	$\hat{f}_i$ Edgeworth
8	3	-1,534	0,1231	0,1493	0,1607	2,47	3,00	3,23
24	6	-1,115	0,2142	0,2670	0,2857	4,30	5,36	5,74
40	7	-0,697	0,3129	0,3623	0,3671	6,28	7,27	7,37
56	11	-0,279	0,3837	0,3922	0,3726	7,70	7,87	7,48
72	7	0,139	0,3951	0,3553	0,3306	7,93	7,13	6,64
88	3	0,558	0,3415	0,2843	0,2807	6,86	5,71	5,64
104	3	0,976	0,2478	0,2108	0,2274	4,97	4,23	4,57
120	4	1,394	0,1509	0,1458	0,1612	3,03	2,93	3,24
136	2	1,812	0,0772	0,0907	0,0929	1,55	1,82	1,87
152	2	2,231	0,0331	0,0483	0,0426	0,66	0,97	0,86

Tabel 5.3. Verdeling van 360-daagse neerslagsommen op Vlissingen, jaren 1905-1953

x_i klasse-midden	f_i abs. freq.	t_i	$\phi(t)$	f(t) Gram-Charlier	f(t) Edgeworth	$\hat{f}_i$ normale verdeling	$\hat{f}_i$ Gram-Charlier	$\hat{f}_i$ Edgeworth
425	1	-2,205	0,035	0,03602		0,68	0,70	
475	2	-1,798	0,079	0,08755		1,55	1,71	
525	3	-1,390	0,152	0,16595		2,97	3,25	
575	3	-0,983	0,246	0,25647		4,81	5,02	
610	3	-0,698	0,313	0,31338	NIET BEREKEND	2,45	2,45	NIET BEREKEND
630	3	-0,535	0,346	0,33953		2,71	2,66	
650	4	-0,372	0,372	0,35943		2,91	2,81	
670	2	-0,209	0,390	0,37223		3,05	2,91	
690	6	-0,046	0,399	0,37743		3,12	2,95	
710	4	0,117	0,396	0,37489		3,10	2,93	
730	1	0,280	0,384	0,36479		3,00	2,85	
750	1	0,443	0,362	0,34761		2,83	2,72	
770	2	0,606	0,332	0,32414		2,60	2,54	
790	2	0,769	0,297	0,29546		2,32	2,31	
825	6	1,054	0,229	0,23683		4,48	4,63	
875	3	1,462	0,137	0,14922		2,68	2,92	
925	1	1,869	0,070	0,07700		1,37	1,51	
975	1	2,277	0,030	0,03120		0,59	0,61	

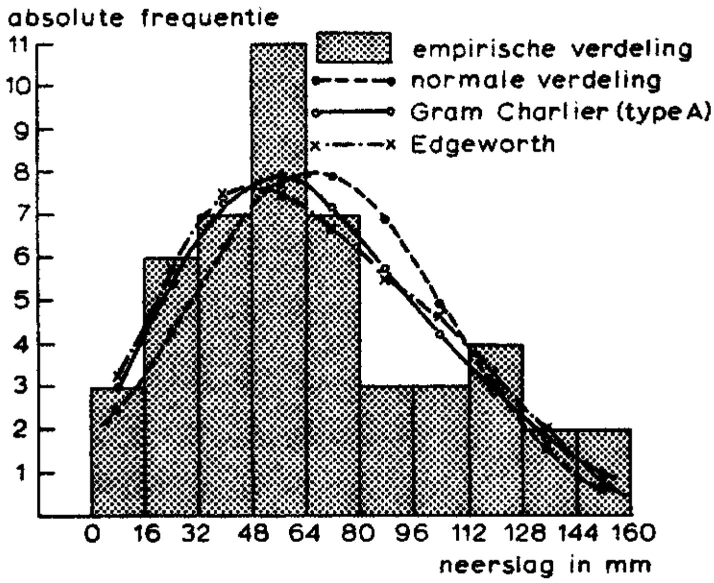

Fig. 2. 30-daagse neerslagsommen
Vlissingen, augustus 1905-'53

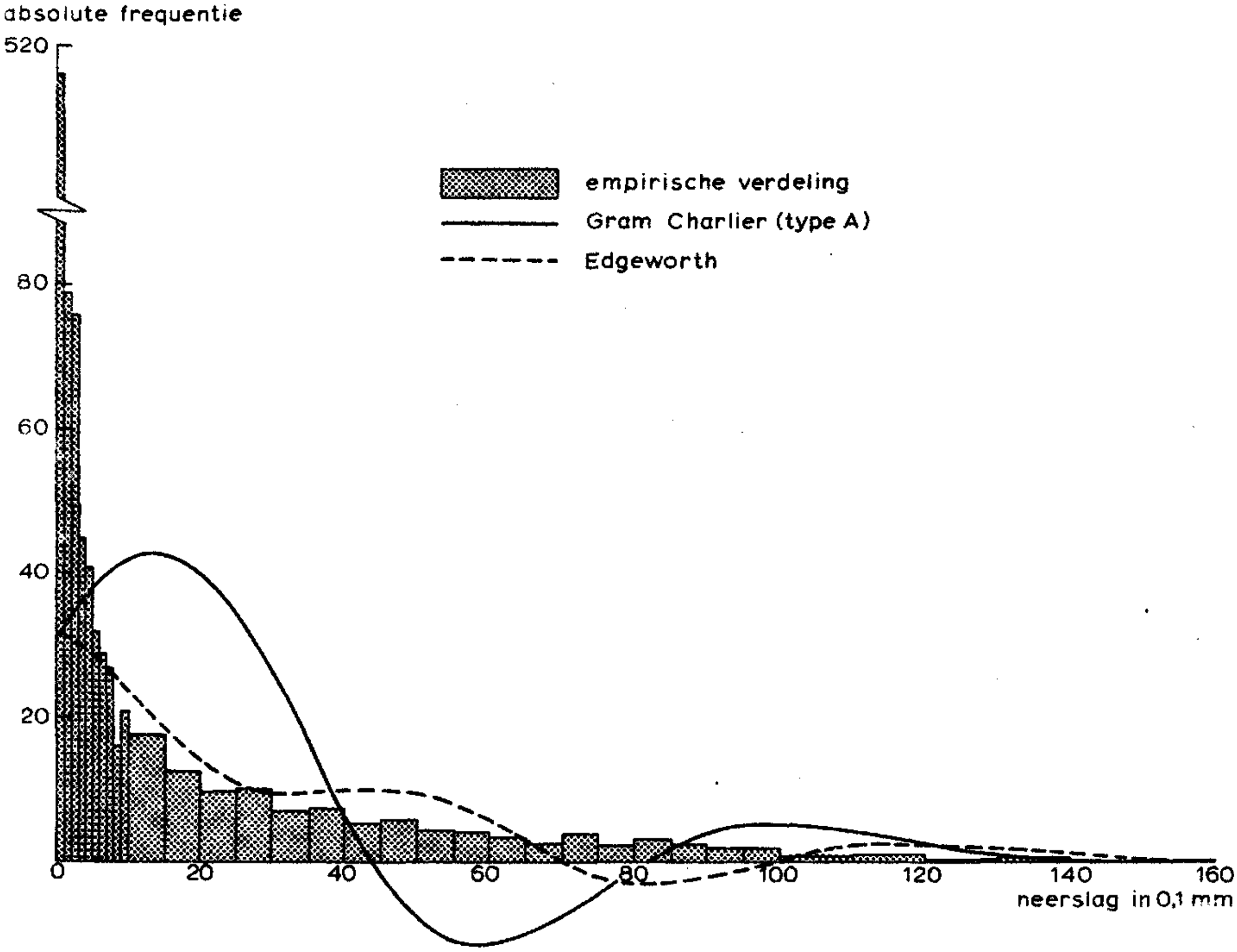

Fig. 1. 1-daagse neerslagsommen. Vlissingen, januari 1905-'53

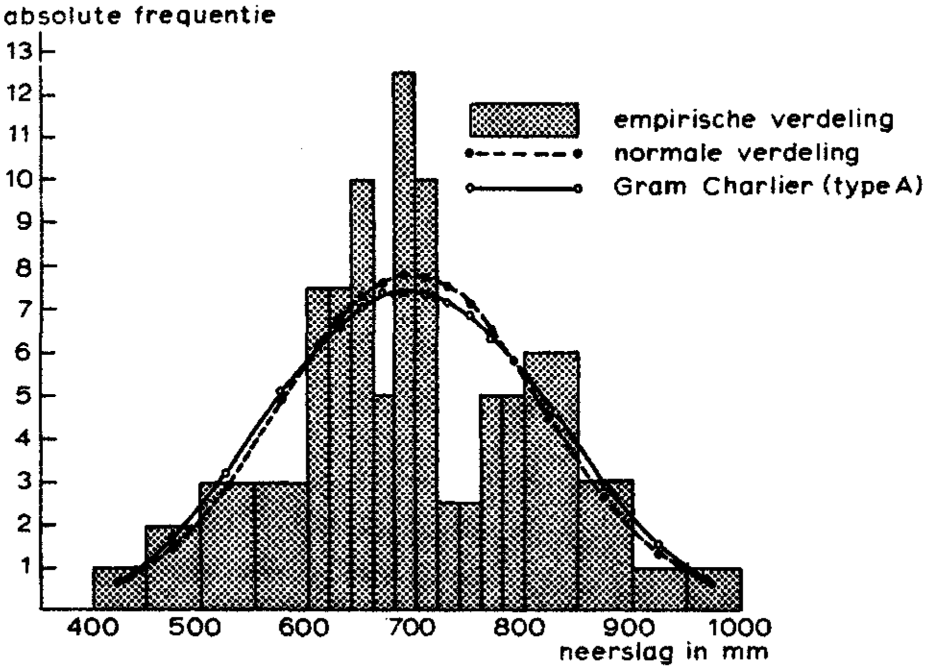

Fig. 3. 360-daagse neerslagsommen
Vlissingen. 1905-'53