MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

1953-32(2)

Statistische analyse van waterstanden methoden en resultaten.

I. Methode voor het toetsen van de onafhankelijkheid van het H.W. aan de Noordzeekust en de waterafvoer in de bovenrivieren

door
Prof. Dr J.Hemelrijk

1953

1. Inleiding.

Daar de hoge waterstanden in plaatsen langs de benedenrivieren afhankelijk zijn van de waterstand aan de kust en van de opperwaterafvoer is het van belang, zoveel mogelijk te weten te komen over het voorkomen van combinaties van grote opperwaterafvoer en hoge waterstanden aan de kust. Om de gedachten te bepalen zullen wij hier verder de H.W.'s in Hoek van Holland en de waterafvoer in Lobith beschouwen. De te beschrijven methoden gelden echter ook voor andere plaatsen. Het H.W. in Hoek van Holland geven wij aan met $\underline{x}$ en de waterafvoer in Lobith met $\underline{y}$. Beide grootheden kunnen als stochastische grootheden beschouwd worden, d.w.z. dat zij een waarschijnlijkheidsverdeling bezitten [1]. De waarschijnlijkheidsverdelingen van $\underline{x}$ en van $\underline{y}$ apart kunnen op grond van de waarnemingen van een reeks van jaren min of meer nauwkeurig worden geschat en als nu de twee grootheden stochastisch onafhankelijk van elkaar zijn, kan op grond van deze schattingen zonder veel moeite een schatting van de simultane verdeling van $\underline{x}$ en $\underline{y}$ gegeven worden, waaruit dus conclusies getrokken kunnen worden over de kans op het voorkomen van combinaties van bepaalde waarden van $\underline{x}$ en $\underline{y}$. Het doel van dit rapport is, methoden aan te geven om deze onafhankelijkheid te toetsen, speciaal rekening houdende met het feit, dat combinaties van hoge waarden van $\underline{x}$ en $\underline{y}$ van groot belang zijn, terwijl afhankelijkheid van $\underline{x}$ en $\underline{y}$ in het gebied van lagere waarden onbelangrijk is.

Wij onderstellen verder steeds, dat alle waarnemingen alleen betrekking hebben op enkele wintermaanden (b.v. November tot en met Januari), daar deze maanden de "gevaarlijkste maanden" zijn en $\underline{x}$ en $\underline{y}$ in de overige maanden zeker een andere verdeling bezitten dan in deze wintermaanden. Zou men de overige maanden ook in het onderzoek betrekken, dan zal hierdoor een afhankelijkheid tussen $\underline{x}$ en $\underline{y}$ optreden, die voor het probleem onbelangrijk is en die de onderzoekingen derhalve alleen in ongunstige zin kan beïnvloeden. Bovendien leidt de beperking tot enkele maanden per jaar tot een aanzienlijke werkbesparing.

2. Methode van de 2 x 2-tabel.

In deze paragraaf beschrijven wij, in algemene termen, een methode, die zich bijzonder goed leent voor toepassing op het

---

1) Wij geven dit aan door de symbolen te onderstrepen. Dezelfde letters, niet onderstreept, worden gebruikt voor door deze stochastische grootheden aangenomen waarden e.d.

gestelde probleem. De methode van toepassing speciaal voor dit probleem wordt verderop besproken. Wij veronderstellen, dat de numerieke gegevens bestaan uit N paren waarnemingen $(x_1,y_1),\ldots,(x_N,y_N)$, die dus in een grafiek als een puntenwolk uitgezet kunnen worden (zie fig. 1).

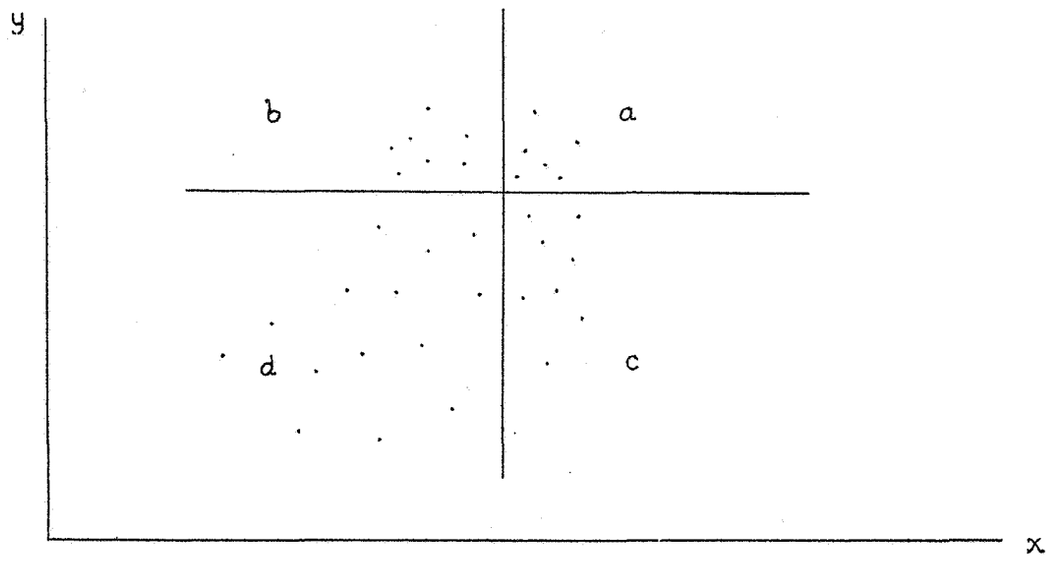

Fig. 1. Grafische voorstelling van de waarnemingsparen $(x_1,y_1),\ldots,(x_N,y_N)$.

Wij verdelen nu deze puntenwolk door een verticale en een horizontale lijn in 4 gedeelten; de keuze van deze plaats van twee lijnen laten wij voorlopig in het midden. De aantallen punten in de 4 zo verkregen vakken geven wij aan met a, b, c en d (zie fig. 1) en wij kunnen deze getallen dan in een 2 x 2-tabel samenvatten (tabel I).

Tabel I

Puntenwolk over een 2x2-tabel verdeeld.

totaal				
m	b	a		m = a+b
n	d	c		n = d+c
				r = a+c
				s = b+d
N	s	r	totaal	N = a+b+c+d

De hypothese H_o, die wij willen toetsen, kan als volgt geformuleerd worden: de waarschijnlijkheidsverdeling van $\underline{x}$ is bij grote waarden van $\underline{y}$ dezelfde als bij kleine waarden van $\underline{y}$. Het aantal punten, dat in het rechterbovenvak ligt, $\underline{a}$, is een stochastische grootheid, waarvan de waarschijnlijkheidsverdeling onder H_o, bekend is. Deze wordt gegeven door

(1) $$P[\underline{a}=a|H_o]=\frac{\binom{m}{a}\binom{n}{c}}{\binom{N}{r}}$$

en heeft als gemiddelde en variantie (spreidingskwadraat)

$$(2) \qquad \mathcal{E}\{\underline{a}|H_o\} = \frac{m\tau}{N} \qquad \text{en} \qquad \sigma^2\{\underline{a}|H_o\} = \frac{mn\tau s}{N^2(N-1)}.$$

(Hierin geeft het achter een verticale streep geplaatste symbool H_o aan, dat deze vergelijkingen slechts behoeven te gelden, indien H_o juist is.)
Is nu uit het waarnemingsmateriaal de waarde a_o voor $\underline{a}$ gevonden, dan wordt de daarbij behorende rechter-overschrijdingskans $k_\tau(a_o)$ gedefinieerd door

$$k_\tau(a_o) = P[\underline{a} \geq a_o|H_o] = 1 - P[\underline{a} < a_o|H_o].$$

Derhalve is

$$(3) \qquad k_\tau(a_o) = 1 - \sum_{i=0}^{a_o-1} P[\underline{a} = i|H_o] = 1 - \sum_{i=0}^{a_o-1} \frac{\binom{m}{i}\binom{n}{\tau-i}}{\binom{N}{\tau}}.$$

Is deze berekening te omslachtig, omdat a_o te groot is (voor kleine a_o is de berekening met behulp van enkele in de literatuurlijst van dit rapport genoemde tabellen wel uit te voeren), dan kan men van twee benaderingen gebruik maken.

Normale benadering.

Voor niet te kleine waarden van $\mathcal{E}\{\underline{a}|H_o\} = \frac{m\tau}{N}$ berekent men

$$(4) \qquad a_o^* = \frac{a_o - \frac{m\tau}{N} - \frac{1}{2}}{\sqrt{\frac{mn\tau s}{N^2(N-1)}}}$$

en men zoekt dan in een tabel van de normale verdeling (met gemiddelde 0 en spreiding 1) de rechteroverschrijdingskans op, die bij a_o^* behoort.

Benadering volgens Poisson.

Is $\mathcal{E}\{\underline{a}|H_o\}$ klein, terwijl N groot is, dan is $k_\tau(a_o)$ bij benadering gelijk aan de rechteroverschrijdingskans van a_o, behorende bij een verdeling volgens Poisson met gemiddelde $\mathcal{E}\{\underline{a}|H_o\}$. Dus dan is

$$(5) \qquad k_\tau(a_o) \approx \sum_{i=a_o}^{\infty} e^{-\frac{m\tau}{N}} \frac{\left(\frac{m\tau}{N}\right)^i}{i!}$$

en deze overschrijdingskans kan men direct opzoeken in een tabel van de verdeling van Poisson, b.v. in de tabel van E.C.Molina (zie literatuurlijst).

Enkele eigenschappen van deze beide benaderingen worden in de volgende paragraaf besproken.

Is de gevonden overschrijdingskans kleiner dan een van tevoren gekozen onbetrouwbaarheidsdrempel α (waarvoor veelal

de waarde 0,05 genomen wordt), dan wordt H_o verworpen en dan luidt de conclusie: combinaties van hoge waarden van $\underline{x}$ en $\underline{y}$ komen vaker voor dan met onafhankelijkheid van deze twee grootheden in overeenstemming te brengen is.

De toets is hier éénzijdig beschreven, omdat het vooral van belang is H_o te verwerpen, indien combinaties van hoge $\underline{x}$ en $\underline{y}$ vaker voorkomen dan met H_o in overeenstemming is, terwijl het minder vaak voorkomen daarvan (tenzij dit zéér uitgesproken zou zijn) van minder belang is. Dit laatste kan nl., indien men dan toch van H_o uit zou gaan, hoogstens ten gevolge hebben, dat men de dijken iets te hoog zou maken, hetgeen minder ernstig geacht moet worden dan het te laag maken van de dijken.

3. Eigenschappen der benaderingen.

Bij het onderhavige onderzoek zal men (zie par. 4) de horizontale en de verticale scheidingslijn in fig. 1 zo kiezen, dat τ en m klein zijn in vergelijking met N. Dan is dus $m<s$ en $\tau<n$ en wij mogen zonder verlies van algemeenheid aannemen, dat $\tau \leqq m$ is (is $\tau > m$, dan vervange men in het volgende τ door m en omgekeerd). In dat geval geldt

$$0 \leqq a \leqq \tau.$$

Nu is voor kleine τ de normale benadering vooral dan goed (en de Poisson-benadering slecht), als $\mathcal{E}\{\underline{a}|H_o\} = \frac{m\tau}{N}$ niet te ver van $\frac{1}{2}\tau$ verwijderd is, dus als m niet veel kleiner is dan $\frac{1}{2}N$. Is dit wel zo, dus is $m << N$, dan is juist de Poisson-benadering op zijn best. Houden wij de grootheden $\frac{m\tau}{N}$ en $\frac{m}{\tau}$ constant en laten wij N toenemen, dan wordt de normale benadering steeds slechter en de Poisson-benadering steeds beter. Anderzijds echter is de verdeling van Poisson bij benadering normaal, als het gemiddelde groot is, zodat voor grotere waarden van $\frac{m\tau}{N}$, ook als $m<<N$ is, de normale benadering toch weer in aanmerking komt.

Op deze gronden valt niet zonder meer uit te maken welke benadering men het beste kan gebruiken, behalve in extreme gevallen. Wij geven ter nadere oriëntatie in tabel II de overschrijdingskansen $k_\tau(a_o)$ exact en met beide benaderingen voor drie verschillende 2x2-tabellen (met in alle drie gevallen $m=\tau$ en met $\frac{m\tau}{N} = 1$ of 5. Van het argument a_o zijn alleen die waarden in de tabel opgenomen, waar de waarde van $k_\tau(a_o)$ in de buurt van 0,10 ligt of kleiner is, daar de nauwkeurigheid der benaderingen in dat gebied vooral van belang is.

Wij zien in deze tabel de bovengenoemde eigenschappen der twee benaderingen bevestigd. Verder zien wij, dat de benadering

Tabel II

Exacte en benaderde overschrijdingskansen van 2x2-tabellen.

13		$\underline{a}$
156		
169	156	13

$\mathcal{E}\{\underline{a}\,|H_o\} = 1$

a_o	$k_z(a_o)$		
	exact	normale benadering	Poisson-benadering
3	0,065	0,053	0,080
4	0,010	0,004	0,018
5	0,001	0,000	0,004

30		$\underline{a}$
150		
180	150	30

$\mathcal{E}\{\underline{a}\,|H_o\} = 5$

a_o	$k_z(a_o)$		
	exact	normale benadering	Poisson-benadering
8	0,094	0,090	0,133
9	0,035	0,031	0,068
10	0,011	0,008	0,032
11	0,003	0,002	0,014
12	0,001	0,000	0,005

50		$\underline{a}$
450		
500	450	50

$\mathcal{E}\{\underline{a}\,|H_o\} = 5$

a_o	$k_z(a_o)$		
	exact	normale benadering	Poisson-benadering
8	0,111	0,108	0,133
9	0,048	0,041	0,068
10	0,018	0,013	0,032
11	0,006	0,003	0,014
12	0,002	0,001	0,005

volgens Poisson voor alle beschouwde gevallen een te hoge en de normale benadering een te lage waarde voor $k_z(a_o)$ geeft. In de buurt van de waarde 0,05 voor $k_z(a_o)$ is de normale benadering echter nog niet zo slecht. De onderschatting van $k_z(a_o)$ wordt ernstiger bij lagere waarden van de overschrijdingskans. Heeft men tabellen van de verdeling van Poisson bij de hand, dan is het zeer eenvoudig hierin de benadering van $k_z(a_o)$ op te zoeken. Men behoeft daarvoor, behalve a_o, slechts $\frac{mz}{N}$ te bepalen. Is de op deze wijze gevonden waarde van $k_z(a_o)$ kleiner dan de gekozen onbetrouwbaarheidsdrempel α, dan kan men zonder bezwaar aannemen, dat de normale benadering en de exacte verdeling eveneens een waarde $< \alpha$ geven. Is de overschrijdingskans volgens de Poisson-benadering groter dan α, maar niet zoveel groter, dat men veilig aan kan nemen dat dit met de exacte overschrijdingskans ook zo is, dan kan men als tweede stap de normale benadering gebruiken. Voor grote $\frac{mz}{N}$ (b.v. > 10) kan men desgewenst ook direct de normale benadering nemen. Geeft deze een overschrijdingskans $> \alpha$, dan kan men zonder bezwaar aannemen, dat de exacte overschrijdingskans ook $> \alpha$ is. In twijfelgevallen, b.v. als de twee benaderingen van $k_z(a_o)$ ter weerszijden van α liggen, kan men, als het om een belangrijk geval gaat, de exacte overschrijdingskans berekenen.

4. Toepassing op het huidige probleem.

Bij toepassing van deze methode op een bepaald probleem is vooral de keuze van de plaats van de horizontale en verticale lijn in fig. 1 van belang. Bij het onderhavige onderzoek ligt het voor de hand deze scheidingslijnen zo te trekken, dat een waarneming in het rechterbovenvak, op waterstaatkundige gronden, als een gevaarlijke combinatie van waarden van $\underline{x}$ en $\underline{y}$ beschouwd moet worden. Immers dan onderzoekt men met de bovenbeschreven methode juist, of deze gevaarlijke combinaties niet vaker voorkomen dan met de onderstelling van onafhankelijkheid van $\underline{x}$ en $\underline{y}$ in overeenstemming is. Desgewenst kan men het onderzoek uitvoeren voor een aantal paren scheidingslijnen afzonderlijk, behorende bij verschillende graden van gevaar van de waarnemingen in het rechterbovenvak [2]. Voert men dit voor de verschillende jaren (d.w.z. voor de wintermaanden daarvan) apart uit, dan is het van belang de 2x2-tabellen in de vorm van tabel I voor ieder dezer jaren ook onderling te vergelijken, daar

---

2) Het is voor de berekeningen prettig, indien de lijnen zo getrokken worden, dat er geen waarnemingspunten op de lijnen kunnen vallen.

deze aanwijzingen kunnen geven voor de al of niet aanwezigheid van "gevaarlijke jaren", dat zijn jaren, die veel meer waarnemingen in het rechterbovenvak vertonen dan de overige [3]. Indien het moeilijk zou vallen om op waterstaatkundige gronden tot een keuze der scheidingslijnen te komen, kan men deze ook op andere wijze bepalen, b.v. door ze zo te trekken, dat $m=\tau$ is en $\frac{m\tau}{N}$ in de buurt van een bepaalde waarde A komt. Hiervoor kan men b.v. 1 nemen, of 5 of 10 of een nog groter getal. Neemt men A klein, dan onderzoekt men op deze wijze of de enkele zeer grote waarden van $\underline{x}$ en $\underline{y}$, die in de beschouwde periode voorkomen, te vaak tegelijk optreden, terwijl dit onderzoek zich bij grotere waarde van A tot een grotere groep van hoge waarden van $\underline{x}$ en $\underline{y}$ uitstrekt. Volgt men deze methode, dan verdient het aanbeveling, de scheidingslijnen zo te trekken, dat zij geen waarnemingspunten bevatten, terwijl aan de gestelde relaties $m=\tau$ en $\frac{m\tau}{N}=A$ toch met zo goed mogelijke benadering voldaan is.

Daar bij het samenstellen de 2x2-tabel naast het bepalen van N, alleen b, a en c bepaald behoeven te worden (zodat het "turven" tot 3 van de 4 vakken en wel tot de drie, die het minste punten bevatten, beperkt kan worden), is de laatstbeschreven methode vermoedelijk tijdrovender dan die met van tevoren gekozen scheidingslijnen. Immers nu hangen de scheidingslijnen van de waarnemingen af en kunnen dus niet van tevoren getrokken worden. Daar staat echter tegenover, dat de in de volgende paragraaf beschreven berekeningen voor dit geval eenvoudiger zijn indien men de laatste methode volgt dan als men de eerste neemt, omdat men dan kan bereiken, dat de getallen

$$\frac{m\tau}{N} \quad \text{en} \quad \frac{mn\tau s}{N^2(N-1)}$$

voor de verschillende jaren vrijwel dezelfde zijn en dus ieder slechts éénmaal berekend behoeven te worden.

Wij merken nog op, dat het feit, dat er twee waarnemingen van $\underline{x}$ zijn bij iedere waarneming van $\underline{y}$, geen storende invloed heeft op de geldigheid van de toets. Wel verdient het, ook voor de verderop beschreven toetsen, aanbeveling om twee opeenvolgen-

---

3) Toetsing of er inderdaad jaren zijn, die systematisch (dus niet uitsluitend toevallig) gevaarlijker zijn dan de andere, is een onderwerp, dat wij in dit rapport niet bespreken. Als voorbereiding voor een grondig onderzoek daarvan is echter het verzamelen van deze 2x2-tabellen zeer nuttig.

de H.W.'s, die beide zeer hoog zijn ten gevolge van een langdurige storm, tezamen te nemen door alleen de hoogste van deze twee waarnemingen (met de hoogste der bijbehorende waarden van $\underline{y}$ — als deze verschillend zijn) in de toets te betrekken.

5. Combinatie van de resultaten der jaren afzonderlijk.

Het is wenselijk de bovenbeschreven methode toe te passen op de waarnemingen van alle onderzochte jaren (d.w.z. van de wintermaanden daarvan) tezamen genomen, maar tevens op ieder van deze jaren apart.

In het laatste geval verkrijgt men, als wij de jaren nummeren van $1,\ldots,k$ en één jaar aangeven met behulp van een index i:

1) een waarde a_i van $\underline{a}$

2) een waarde $\frac{m_i \tau_i}{N_i}$ van $\mathcal{E}\{\underline{a}|H_o\}$

3) een waarde $\frac{m_i n_i \tau_i s_i}{N_i^2(N_i-1)}$ van $\sigma^2\{\underline{a}|H_o\}$

$\Big\}$ $(i=1,\ldots,k)$.

Wij berekenen nu de grootheid

$$A = \frac{\sum_{i=1}^{k}\left\{a_i - \frac{m_i\tau_i}{N_i}\right\}}{\sqrt{\sum_{i=1}^{k}\frac{m_i n_i \tau_i s_i}{N_i^2(N_i-1)}}} \qquad (5)$$

en zoeken de daarbij behorende rechteroverschrijdingskans op in een tabel van de normale verdeling, met gemiddelde 0 en spreiding 1.

Ook deze methode kan toegepast worden voor verschillende standen van de scheidingslijnen.

De benadering met behulp van de Poisson-verdeling komt in dit geval niet meer in aanmerking, tenzij $\mu = \sum_{i=1}^{k}\frac{m_i\tau_i}{N_i}$ een klein getal zou zijn (< 10). In dat geval zoeke men de rechte overschrijdingskans, behorende bij $\sum_{i=1}^{k} a_i$ op in een tabel van de Poisson-verdeling met deze μ als gemiddelde.

Voor numerieke voorbeelden van de hier en boven beschreven berekeningen (behalve de Poisson-benadering) raadplege men desgewenst het in de literatuurlijst genoemde rapport van C.van Eeden.

6. Interpretatie der resultaten.

De in par. 4 beschreven toets is erop gericht te onderzoeken of gedurende een langere periode als één geheel beschouwd,

de grootheden $\underline{x}$ en $\underline{y}$ als onafhankelijk beschouwd kunnen worden. Wij beschouwen nu de twee volgende mogelijkheden nader:

A. De toets leidt tot verwerping van H_o,

B. De toets leidt niet tot verwerping van H_o.

A. In dit geval komen wij dus tot de conclusie, dat de combinatie van hoge waarden van $\underline{x}$ en $\underline{y}$ vaker voorkomt dan bij onafhankelijke $\underline{x}$ en $\underline{y}$ het geval zou zijn. Dit kan twee oorzaken hebben, die ook gecombineerd op kunnen treden:

a. dit verschijnsel doet zich reeds binnen ieder der jaren of binnen een aantal der jaren afzonderlijk voor (hoge waarden van $\underline{y}$ zijn dus in het algemeen symptomatisch voor hoge waarden van $\underline{x}$);

b. binnen de jaren afzonderlijk doet het verschijnsel zich niet voor, maar er zijn één of meer jaren, die zowel bijzonder hoge $\underline{x}$ als bijzonder hoge $\underline{y}$-waarnemingen gaven (dus "gevaarlijke jaren").

De aanwezigheid van deze twee mogelijke oorzaken maakt het gewenst om, behalve de in par. 4 beschreven toets, ook de toets voor de jaren apart en vervolgens gecombineerd (par. 5) toe te passen.

B. Indien H_o op grond van het totale waarnemingsmateriaal als één geheel beschouwd niet verworpen wordt, zou dit toch binnen de jaren afzonderlijk nog wel het geval kunnen zijn. Dit effect kàn dan nl. door verschillen tussen de afzonderlijke jaren verdoezeld worden. Het is daarom ook in dit geval belangrijk de in par. 5 beschreven methode toe te passen. In hoeverre deze methode zal leiden tot duidelijke en goed interpreteerbare conclusies valt van tevoren niet te zeggen.

7. Andere methode.

Naast die hier beschreven methode zijn ook andere methoden beschikbaar, die echter, wegens hun meer gedifferentieerde karakter, veel bewerkelijker zijn. Het is aanbevelenswaard boven beschreven methoden eerst toe te passen en de eventuele toepassing van andere methoden van de resultaten daarvan af te laten hangen.

Literatuur.

R.A.Fisher, Statistical methods for research workers, Oliver and Boyd, London 1948, p. 96 e.v. (Beschrijft de exacte toets voor een 2x2-tabel.)

C.van Eeden, Methoden voor het vergelijken, toetsen en schatten van onbekende kansen, Rapport S 115 (M 45) van de Statistische Afdeling van het Mathematisch Centrum, Amsterdam 1953. (Bevat gebruiksaanwijzingen voor de methode van de 2x2-tabel.)

E.C.Molina, Poisson's exponential limit, D.van Nostrand Co, N.Y. 1945. (Tabellen van de Poisson-verdeling.)

T.C.Fry, Probability and its engineering uses, D.van Nostrand Co, N.Y. 1928 (bevat o.a. tabellen van binomiaalcoëfficiënten.)

E.S.Pearson, Table of the logarithms of complete Γ-functions (for arguments 2 to 1200), Tracts for Computors no. VIII, Cambridge Un. Press 1922.

MATHEMATISCH CENTRUM,
2de Boerhaavestr. 49,
A m s t e r d a m - O.

1953-32(2) I

Dit is de eerste aflevering van een serie voorlopige rapporten, die door de Statistische Afdeling van het Mathematisch Centrum geschreven worden over methoden en resultaten van statistische analyse van waterstanden. Dit rapport werd geschreven naar aanleiding van een verzoek van de Directie Benedenrivieren van het Departement van Waterstaat. Daar de beschreven methode ook voor andere problemen dan het hier beschouwde geschikt is, wordt dit rapport op ruimere schaal onder de leden van de Delta commissie verspreid.