K O N I N K L IJ K   N E D E R L A N D S

M E T E O R O L O G I S C H   I N S T I T U U T

Wetenschappelijk Rapport W.R. 57-009 (IV-015)

R. Dorrestein

Een nieuw statistisch hulpmiddel bij de interpretatie van scheepswaarnemingen

De Bilt, 1957

KONINKLIJK NEDERLANDS
METEOROLOGISCH INSTITUUT

Wetenschappelijk Rapport
W.R. 57-009 (IV-15)
U.D.C. 551.501.45(26):
551.46.018:
519.2

R. Dorrestein

Een nieuw statistisch hulpmiddel bij de interpretatie van scheepswaarnemingen

A new statistical aid in the interpretation of ship's observations

INHOUD - CONTENTS

		p.
1.	Het probleem - The problem	3
2.	Invoering van begrippen en grootheden - Introduction of concepts and quantities	3
3.	Afleiding van de schattingsformules - Derivation of the estimation formulae	6
4.	Commentaar bij de schattingsformules - Comments on the formulae	10
5.	Betrouwbaarheid - confidence of estimates	12
6.	Een voorbeeld - An example	13
7.	Conclusie en slotopmerkingen - Conclusion and final remarks	19
	Summary	21

Betekenis der Symbolen

A duidt het beschouwde meteorologische kenmerk aan

E (index) duidt een verwachtingswaarde aan

S (index) duidt een geschatte waarde aan

Symbool		Betekenis	Ingevoerd in formule:
f_A	=	fractie van de tijd dat A in een gegeven zeegebied binnen het tijdvak N optreedt	(1)
f_W	=	N'_W/N = "waarnemingstijdfractie"	(4)
N	=	duur van het beschouwde (lange) tijdvak gedeeld door het standaard-tijdsinterval tussen twee opeenvolgende scheepswaarnemingen (meestal 6 of 4 uren)	(2)
N_{AW}	=	aantal beschikbare waarnemingen met A in dit tijdvak	(1)
N'_{AW}	=	hetzelfde, waarbij echter gelijktijdig verrichte waarnemingen als één worden geteld	(1a)
N_W	=	aantal beschikbare waarnemingen in dit tijdvak	(1)
N'_W	=	hetzelfde, waarbij echter gelijktijdig verrichte waarnemingen als één worden geteld	(4)
P_A	=	aantal "perioden" (= series van opeenvolgende waarnemingstijdstippen) dat A optreedt in dit tijdvak	(2)
P_{AW}	=	aantal series van opeenvolgende waarnemingen in dit tijdvak, waarin tenminste één A-waarneming voorkomt	(8)
P_W	=	aantal series van opeenvolgende waarnemingen in dit tijdvak	(5)
s_A	=	gemiddelde tijdsverschil tussen de inzetten van twee opeenvolgende periodes met A, gedeeld door het standaard-tijdsinterval	(3)
s_W	=	hetzelfde voor de series van opeenvolgende waarnemingen	(7)
t_A	=	(gemiddelde) duur der "perioden" als bij P_A bedoeld, gedeeld door het standaard-tijdsinterval	(2)
t_W	=	(gemiddeld) aantal (niet gelijktijdige) waarnemingen in de series als bij P_W bedoeld	(5)

1. Het probleem

De meteorologische gegevens die wij van de meeste zeegebieden hebben, zijn uitsluitend verzameld door passerende schepen. Voor een beperkt gebied vormen deze gegevens in de loop van de tijd dus geen continu doorlopende reeks, maar een reeks waarin in het algemeen hiaten voorkomen.

Nu wordt dikwijls gevraagd naar het gemiddeld aantal perioden per maand of per jaar, en de duren van deze perioden, dat in een zeker zeegebied een bepaalde (betrekkelijk weinig voorkomende) meteorologische toestand heerst. Het kan zijn dat men vraagt naar het aantal malen dat de windkracht een bepaalde, vrij hoge, waarde bereikt of overschrijdt, waarbij eventueel de windrichting binnen een bepaalde sector blijft. Of naar het optreden van golven die hoger zijn dan een bepaalde waarde, of naar het optreden van mist, enz. We zullen heel algemeen spreken van de toestand "A" of van het kenmerk "A". Het beschouwde zeegebied denken we homogeen wat de klimatologische eigenschappen betreft.

2. Invoering van begrippen en grootheden

Wanneer in de loop van vele jaren scheepswaarnemingen van het zeegebied in kwestie verzameld zijn, kan men stellen dat de beste schatting voor de fractie van de tijd dat A op een zekere plaats in het zeegebied optreedt, f_A, wordt geleverd door het aantal waarnemingen met A, N_{AW}, te delen door het totale aantal beschikbare waarnemingen, N_W:

$$f_A \text{ (geschat)} = \frac{N_{AW}}{N_W}. \qquad (1)$$

Hierbij is de onderstelling gemaakt, dat de tijdstippen waarop deze waarnemingen gedaan zijn op geen enkele wijze zijn beïnvloed door het optreden van A, b.v. doordat ze voldoende onsystematisch of regelloos over de tijd zijn verdeeld. Niet steeds zal aan deze voorwaarde voldaan zijn, maar in de meeste gevallen wel bij benadering.

Het quotient (1) kan worden betrokken op alle waarnemingen over het hele jaar verdeeld, of wel op de waarnemingen in bepaalde maanden of seizoenen. We noemen f_A de "A-tijdfractie". Als f_A veel kleiner is dan één, hebben we te maken met een betrekkelijk zeldzame toestand.

Wegens de continuïteit in de tijd treedt de toestand A een tijd lang op, daarna een zekere tijd niet, daarna weer wel, enz. In de taal der communicatietechniek zouden we zeggen dat we een signaal beschouwen dat behoort tot de categorie der "telegraaf-signalen". De vraag is nu, uit het voorhanden waarnemingenmateriaal op te maken hoeveel malen in een gegeven tijd een periode met A verwacht kan worden, of zal voorgekomen zijn. Is dit P_A in een (lang) tijdvak van N tijdseenheden, en is de gemiddelde duur van zulke perioden t_A tijdseenheden, dan hebben we blijkbaar:

$$f_A = \frac{P_A \ t_A}{N} \qquad (2)$$

We kunnen ook schrijven

$$f_A = \frac{t_A}{s_A} \text{ met } s_A = \frac{N}{P_A}, \qquad (3)$$

waarbij s_A is het gemiddelde tijdsverschil tussen de inzetten van twee opeenvolgende periodes met A.

We nemen nu verder aan, om de gedachten te vereenvoudigen, dat het beschouwde zeegebied zo klein is, dat de toestand A òf over het hele gebied voorkomt, òf nergens in dat gebied. Aan de andere kant zij het gebied weer zo groot, dat een passerend schip, dat eens per 6 uren een meteowaarneming doet (vóór de oorlog 1939-'45 eens per 4 uren) binnen dit gebied in het algemeen enkele waarnemingen kan verrichten. Het is duidelijk dat deze twee voorwaarden in vele gevallen slechts bij benadering tegelijk vervuld zullen kunnen zijn. Het zeegebied wordt dus geacht afmetingen van honderd tot enkele honderden mijlen te hebben. Het kan b.v. een tweegraads- of een vijfgraadsvak zijn, maar het kan ook langgerekt zijn, b.v. een stuk uit een bepaalde zeeroute.

Volgens het tegenwoordige gebruik doen de schepen hun meteowaarnemingen elke 6 uren op vaste tijdstippen: 0, 6, 12 en 18 h. G.M.T. Vroeger was dit elke 4 uren. We schematiseren nu (noodgedwongen) verder zodanig, dat het resultaat van een scheepswaarneming representatief zal zijn voor een tijdvak van 6 (4) uren, b.v. 3 (2) uren vóór en 3 (2) uren na het waarnemingsuur (en tevens representatief voor het gehele zeegebied zoals boven reeds aangeduid), en dat dus de toestand A slechts voorkomt gedurende perioden met duren die een geheel aantal malen 6 (4) uren zijn. Deze schematisering heeft uiteraard alleen zin als de duren van deze perioden in het algemeen wat groter dan 6 (4) uren zijn; dit nemen we aan. De getallen tussen haakjes hebben betrekking op de vroegere toestand. Het ligt nu voor de hand om als eenheid van tijd in het vervolg 6 (4) uren te nemen. Het is duidelijk dat de hier gegeven tijdschaal nu eenmaal vrij grof is; kortdurende verschijnselen als lokale buien e.d. zullen veelal niet verwerkt kunnen worden.

In een tijdvak van D dagen in verleden of toekomst zijn er aldus 4 D (6 D) "potentiële" waarnemingstijdstippen. We noemen dit (tijd-) punten; ze liggen equidistant op de tijdschaal. Een deel van deze punten zal gekenmerkt zijn door het optreden van de toestand A. Zulke punten noemen we A-punten. We weten dus dat deze A-punten in series voorkomen, gescheiden door series punten met niet-A.

In feite zijn er in een verleden tijdvak, ter lengte 4 D (6 D) = N tijdseenheden, allêên waarnemingen gedaan wanneer zich tenminste één schip in het gebied bevond, stel op N'_W van deze N tijdstippen; deze noemen we de W-punten. (Hierbij kan N'_W ten hoogste gelijk zijn aan het totale aantal beschikbare waarnemingen, N_W, nl. als er nooit meer dan één schip tegelijk in het gebied is geweest). De verhouding

$$f_W = \frac{N'_W}{N} \tag{4}$$

noemen we de "waarnemingstijdfractie" of de W-tijdfractie. Wanneer op N'_{AW} van deze N'_W tijdstippen de toestand A wordt waargenomen, is de volgende formule een schatting voor de A-tijdfractie f_A, evenals (1):

$$f_A \text{ (geschat)} = \frac{N'_{AW}}{N'_W} \tag{1a}$$

Wanneer nu de waarnemingen in chronologische volgorde getabelleerd worden, blijkt - tenzij er zéér veel waarnemingen zijn - dat niet alleen de A-punten, maar ook de W-punten in series voorkomen, die gescheiden worden door series van niet-W, of in gewone taal uitgedrukt: de waarnemingen komen in "plukjes" voor, die van elkaar in de tijd gescheiden zijn door tijdvakken zonder waarnemingen. Dit is begrijpelijk: gedurende enige tijd is in het gebied een schip aanwezig, dat elke 6 (4) uren een waarneming doet (eventueel zijn er toevallig meer dan een tegelijk), daarna zijn er een tijd lang geen schepen, daarna weer enige tijd wel een of meer schepen, enz. We kunnen dan tellen hoeveel van deze plukjes waarnemingen in een (lang) tijdvak van N tijdseenheden voorkomen; zij dit P_W. Is de gemiddelde duur van deze plukjes t_W tijdseenheden, dan hebben we blijkbaar:

$$N'_W = P_W \; t_W \, , \tag{5}$$

en geheel analoog met (2) en (3):

$$f_W = \frac{P_W \; t_W}{N} \; ; \tag{6}$$

$$f_W = \frac{t_W}{s_W} \, , \text{ met } s_W = \frac{N}{P_W} \, , \tag{7}$$

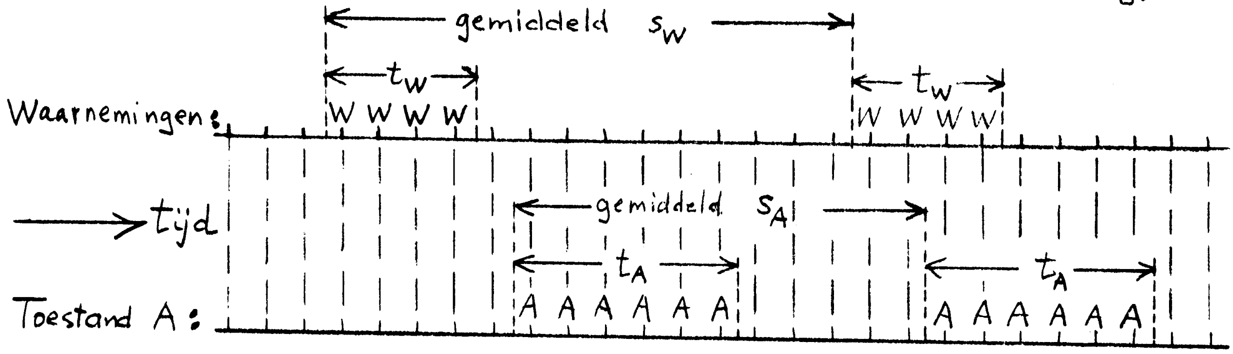
waarbij s_W het gemiddelde tijdsverschil tussen de inzetten van twee opeenvolgende waarnemingenseries is.

We kunnen verder tellen het aantal van de P_W plukjes waarnemingen waarin tenminste één A-waarneming voorkomt; zij dit P_{AW}.

Nu zullen we laten zien dat voor een bepaald tijdvak de gegevens N, N_W, N_{AW}, N'_W, P_W en P_{AW} voldoende zijn om, met zekere onderstellingen, een schatting te maken van P_A, het aantal malen dat A in dit tijdvak is opgetreden.

3. Afleiding van de schattingsformules

Eerst geven we de afleiding voor een vereenvoudigd model, waarin alle W-series een zelfde duur t_W hebben, en alle A-series een zelfde duur t_A. Gemiddeld komt één W-serie voor in een tijd s_W, en één A-serie in een tijd s_A; deze laatste, die gelijk is aan N/P_A, is dus feitelijk onbekend, maar wordt hier eerst bekend ondersteld. Aangenomen wordt nu, dat er geen enkele correlatie bestaat tussen het optreden van de W-series en van de A-series (hieraan is b.v. voldaan wanneer beide series onafhankelijk van elkaar en onsystematisch of regelloos over de tijd verdeeld zijn, of wanneer een van beide regelloos verdeeld is en de andere volgens een onafhankelijk systeem, enz.; er zijn vele mogelijkheden). Verdere onderstellingen zullen nog worden ingevoerd in de loop van het betoog.



Wanneer een W-serie gedeeltelijk met een A-serie samenvalt, spreken we van een coïncidentie (zoals rechts in bovenstaande figuur). Opdat nu een W-serie en een A-serie zullen coïncideren, is nodig dat het beginpunt van de A-serie met een W-punt samenvalt of ten hoogste $t_A - 1$ vóór het beginpunt van een W-serie komt (men kan in de formulering ook A en W verwisselen). Denk nu de W- en de A-series beide genummerd, en trek van beide een willekeurig nummer. De kans dat deze twee zullen coïncideren, dat dus b.v. de W-serie zal worden "getroffen" door de A-serie, is dan, omdat er geen correlatie bestaat, eenvoudig gegeven door de verhouding tussen het aantal gunstige mogelijkheden en het totale aantal (onderling gelijkwaardige) mogelijkheden voor de inzet van de A-serie, d.i. $(t_W + t_A - 1)/N$. De kans dat het niet gebeurt is dan één minus dit bedrag. De kans dat één bepaalde W-serie niet zal coïncideren met twee gekozen A-series is dan
$\left(1 - \frac{t_W + t_A - 1}{N}\right)\left(1 - \frac{t_W + t_A - 1}{N - n_{A2}}\right)$, waarin n_{A2} een getal is van de orde

t_A maar iets groter, dat in rekening moet brengen dat de beschikbare "ruimte" voor de tweede A-serie beperkt is door de eerste. Zo doorgaande, komen we voor de kans dat één bepaalde W-serie zal coïncideren met geen enkele der P_A A-series, tot een product van P_A factoren. Nu zullen we aannemen dat de A-series zo ver uiteen liggen, dat ook in de laatste factor van dit product de noemer van de breuk nog van de orde N is en dus groot is ten opzichte van de teller. Daartoe zij $f_A = t_A/s_A = P_A t_A/N$ niet al te groot, zeg ten hoogste 0,5. In dat geval kunnen we benaderen door overal in de noemer een gemiddelde waarde te schrijven; we kiezen $N(1 - \frac{1}{2} f_A)$. De genoemde kans wordt dan:

$$\left\{1 - \frac{t_W + t_A - 1}{N(1 - \frac{1}{2}f_A)}\right\}^{P_A} = 1 - \frac{t_W + t_A - 1}{N(1 - \frac{1}{2}f_A)} P_A + \ldots\ldots\ldots ,$$

waarin de stippeltjes de binomiaal-ontwikkeling aanduiden.

Als de A-series betrekkelijk ijl verdeeld zijn, kunnen we het al of niet coïncideren van alle W-series als ten naaste bij onderling onafhankelijk beschouwen, en de fractie der W-series, die wèl coïncideert met een A-serie, welke fractie wij P_{AW}/P_W genoemd hebben, heeft, bij gegeven P_A, een verwachtingswaarde gegeven door

$$\frac{P_{AWE}}{P_W} = 1 - \left\{1 - \frac{t_W + t_A - 1}{N(1 - \frac{1}{2}f_A)}\right\}^{P_A} = \frac{t_W + t_A - 1}{N(1 - \frac{1}{2}f_A)} P_A - \ldots \quad (8)$$

(De index E staat voor "expectans"; de verwachtingswaarde is de gemiddelde waarde die we zouden vinden als we een proef vele malen onder gelijkwaardige omstandigheden zouden herhalen).

De door stippels aangegeven termen in (8) kunnen nu verwaarloosd worden indien $(t_W + t_A - 1)P_A/N \ll 1$ is. Hiertoe is nodig ten eerste $P_A t_A/N \ll 1$, d.w.z. de A-series zijn ijl, en ten tweede $P_A t_W/N \ll 1$, of $t_W \ll s_A$ (zie (3)); in woorden: de duur van één waarnemingenserie zij veel kleiner dan de gemiddelde duur tussen twee perioden met A. Als we dit alles onderstellen, kunnen we (8) dus als volgt schrijven:

$$P_{AWE} = \frac{1}{N} (t_W + t_A - 1) P_A P_W . \quad (9)$$

Hieruit kan met behulp van (2) t_A worden geëlimineerd; lossen we dan naar P_A op en gebruiken we (5), dan komt er:

$$P_A = N \frac{P_{AWE} - f_A P_W}{N_W^! - P_W} . \quad (10)$$

Dit is dus een relatie tussen de gegeven onderstelde grootheden N, P_A, f_A, P_W en $N_W^!$ enerzijds en de hierbij behorende verwachtingswaarde P_{AWE} van P_{AW} anderzijds. Omgekeerd kan de relatie ook dienen om een schatting te krijgen voor P_A, als N, P_{AW}, P_W en N'_W gegeven zijn en f_A uit (1) of (1a) kan worden geschat:

$$P_{AE} = N \frac{P_{AW} - \frac{N_{AW}}{N_W} P_W}{N'_W - P_W} \text{ of } N \frac{P_{AW} - \frac{N'_{AW}}{N'_W} P_W}{N'_W - P_W} . \tag{11}$$

De schatting levert inderdaad de <u>verwachtingswaarde</u> voor P_A, omdat het verband met P_{AW} en N_{AW} of N'_{AW} lineair is. Een schatting voor de duur t_A kan daarna worden gevonden uit (2), (1) en (1a):

$$t_{AS} \text{ (geschat)} = \frac{N'_{AW} - \frac{N'_{AW}}{N'_W} P_W}{P_{AW} - \frac{N'_{AW}}{N'_W} P_W} . \tag{12}$$

N.B. Deze schatting levert blijkbaar <u>niet</u> de boven gedefinieerde verwachtingswaarde.

Met (11) en (12) zijn schattingen verkregen voor de gezochte grootheden P_A, het aantal perioden dat de toestand A is opgetreden of zal optreden, en t_A, de duur hiervan, die uitgedrukt zijn in:

N = de duur van het beschouwde (lange tijdvak), in eenheden van 6 of 4 uren,

P_W = het totale aantal series waarnemingen in dit tijdvak,

P_{AW} = het aantal series waarnemingen in dit tijdvak waarin tenminste één A-waarneming voorkomt,

N'_W = het totale aantal waarnemingen in dit tijdvak, waarbij gelijktijdig verrichte waarnemingen als één worden geteld,

N'_{AW} = het totale aantal waarnemingen met A in dit tijdvak, waarbij gelijktijdig verrichte waarnemingen als één worden geteld.

De twee grootheden N_W en N'_{AW} blijken nu gemist te kunnen worden.

We herhalen nog even de onderstellingen die we gemaakt hebben bij de afleiding van (10), (11) en (12):

(a) Alle W-series even lang en alle A-series even lang; het zal echter nog blijken dat deze onderstelling niet essentieel is.

(b) Geen correlatie tussen de W- en de A-series.

(c) De toestand A komt betrekkelijk zelden voor: f_A klein t.o.v. één, of ook: de duur der A-series, t_A, is veel kleiner dan het gemiddelde tijdsverschil s_A tussen de inzetten van opeenvolgende periodes met A;

(d) de duur der W-series, t_W, is veel kleiner dan het gemiddelde tijdsverschil s_A tussen de inzetten van opeenvolgende periodes met A.

Voor een geval uit de praktijk kan men uit het klein zijn van P_{AW} ten opzichte van P_W besluiten dat aan de onderstellingen (c) en (d) is voldaan (zie (9)).

Dit alles houdt in, dat de W-series wel willekeurig dicht op elkaar mogen voorkomen (f_W behoeft niet klein te zijn). Verder laat onderstelling (b) toe, dat in de frequentie van de toestand A b.v. een dagelijkse of een jaarlijkse gang voorkomt, mits deze niet in de frequentie der waarnemingen bestaat, òf omgekeerd: dat in de frequentie der waarnemingen een dagelijkse of jaarlijkse gang voorkomt, maar niet in de frequentie van A.

Voor we nu tot een nadere beschouwing van de vergelijking (11) overgaan, willen we eerst laten zien dat dezelfde uitdrukking ontstaat als we de onderstelling (a) dat alle W-series een zelfde duur èn alle A-series een zelfde duur hebben, laten vallen. Daartoe beschouwen we het volgende model.

Een fractie p_1 van de A-series heeft de duur t_{A1} en een fractie $p_2 = 1 - p_1$ heeft een duur t_{A2}. Een fractie q_1 van de W-series heeft een duur t_{W1} en een fractie $q_2 = 1 - q_1$ heeft een duur t_{W2}. Een redenering in dezelfde trant als boven kan dan worden gehouden. We nemen weer aan dat de W- en de A-series op geen enkele wijze gecorreleerd zijn. De kans dat een bepaalde W-serie van de eerste soort wel zal coïncideren met een A-serie is dan

$$\frac{t_{W1} + t_{A1} - 1}{N} p_1 P_A + \frac{t_{W1} + t_{A2} - 1}{N} p_2 P_A,$$

en voor een W-serie van de tweede soort een dergelijke uitdrukking, aangenomen dat de A-series weer voldoende ijl verdeeld zijn en dat t_{W1} en t_{W2} klein zijn t.o.v. s_A. Het totale aantal W-series die coïncideren met een A-serie, P_{AW}, heeft dan een verwachtingswaarde gegeven door

$$\begin{aligned} P_{AWE} &= \left[\frac{t_{W1} + t_{A1} - 1}{N} p_1 + \frac{t_{W1} + t_{A2} - 1}{N} p_2\right] P_A \cdot q_1 P_W + \\ &\quad + \left[\frac{t_{W2} + t_{A1} - 1}{N} p_1 + \frac{t_{W2} + t_{A2} - 1}{N} p_2\right] P_A \cdot q_2 P_W = \\ &= \left[q_1(t_{W1} + p_1 t_{A1} + p_2 t_{A2} - 1) + q_2 (t_{W2} + p_1 t_{A1} + p_2 t_{A2} - 1)\right] P_A P_W / N = \\ &= (q_1 t_{W1} + q_2 t_{W2} + p_1 t_{A1} + p_2 t_{A2} - 1) P_A P_W / N = \\ &= \frac{1}{N} (t_W + t_A - 1) P_A P_W , \end{aligned}$$

waarin $t_W = q_1 t_{W1} + q_2 t_{W2}$ de gemiddelde duur der W-series is, en $t_{A1} = p_1 t_{A1} + p_2 t_{A2}$ de gemiddelde duur der A-series. Dit is dezelfde formule (9) als boven afgeleid werd, en de daaruit volgende vergelijkingen (10) en (11) blijven dus gelden.

Als we drie of meer verschillende duren in de A-series en de W-series hebben, krijgen we klaarblijkelijk ook dezelfde uitkomsten, en deze gaan dus, wat dit betreft, algemeen op. De gemaakte onderstellingen zijn slechts die welke boven genoemd zijn onder (b), (c) en (d)+.

4. Commentaar bij de schattingsformules

We beschouwen nu de formules (9), (10) en (11) wat nader. We doen dan een merkwaardige ontdekking, nl. dat het rechterlid van (9) symmetrisch is in de indices W en A; dit betekent dat dezelfde formule afgeleid had kunnen worden wanneer in de onderstellingen overal W en A verwisseld waren. We zouden dan hebben: de W-series liggen ijl ($f_W \ll 1$ of $t_W \ll s_W$), de duur der A-series is veel korter dan het tijdsverschil tussen twee W-series ($t_A \ll s_W$) en, equivalent met deze twee onderstellingen samen: $P_{AW} \ll P_A$. Maar P_{AWE} in (9) zou dan niet voorstellen de verwachtingswaarde voor het aantal W-series dat met een A-serie coïncideert, maar die voor het aantal A-series dat met een W-serie coïncideert. In het algemeen behoeven deze aantallen niet gelijk te zijn, omdat meer dan één A-serie kan coïncideren met één W-serie en meer dan één W-serie kan coïncideren met één A-serie. De mate waarin het eerste voorkomt kan wel uit de waarnemingen worden afgeleid, de mate waarin het laatste voorkomt niet. Maar onder de in deze alinea genoemde voorwaarden ($P_{AW} \ll P_A$) zal het laatste bijna niet voorkomen, misschien wel

+ Terwille van de volledigheid vermelden we hierbij nog enkele relaties die we krijgen als de onderstellingen (c) en (d) vervallen, en de overgang van (8) op (9) dus niet geoorloofd is. In plaats van (10) krijgen we dan:

$$P_A = N(1 - \tfrac{1}{2}f_A)\,\frac{-\ln(1 - P_{AWE}/P_W) - f_A}{t_W - 1} \qquad (10a)$$

en in plaats van (11):

$$P_A \text{ (geschat)} = N\left(1 - \tfrac{1}{2}\frac{N_{AW}}{N_W}\right)\frac{-\ln(1 - P_{AW}/P_W) - N_{AW}/N_W}{N_W'/P_W - 1} \qquad (11a)$$

(Deze schatting levert niet de verwachtingswaarde, omdat het verband tussen P_A en P_{AWE} in (10a) niet lineair is). Deze formules zouden gebruikt moeten worden, wanneer blijkt dat P_{AW} niet klein is ten opzichte van P_W.

het eerste (als t_W tenminste van dezelfde orde is als s_A), en dus zal dan ook het aantal coïnciderende A-series inderdaad uit de waarnemingen kunnen worden afgeleid en zal (11) gebruikt kunnen worden, waarin dan P_{AW} het aantal coïnciderende A-series voorstelt.

Hieruit zien we dus, dat de schattingsformule (11) voor P_A toepasbaar is, zowel met de onderstellingen (b), (c) en (d) (zie boven) als met de onderstellingern (b), (c') en (d') die uit de vorige volgen door verwisseling van A en W. Dit betekent dus o.a. dat òf de toestand A betrekkelijk zeldzaam moet zijn, òf de waarnemingstijdfractie betrekkelijk gering, of beide. In formules:

Geval I :	Geval II :	Geval III : (combinatie van I en II)
(c) : $t_A \ll s_A$,	(c') : $t_W \ll s_W$,	(c) en (c')
(d) : $t_W \ll s_A$.	(d') : $t_A \ll s_W$.	(d) en (d')
Beschouw het aantal coïnciderende W-series $= P_{AW}$	Beschouw het aantal coïnciderende A-series $= P_{AW}$	Aantal coïnc. W-series = aantal coïnc. A-series $= P_{AW}$

Nu zullen we het rechterlid van (11) nog wat nader beschouwen. Naievelijk zou men misschien geneigd zijn de fractie der A-perioden die is waargenomen (geval II en III), d.i. de verhouding P_{AW}/P_A, gelijk stellen aan de waarnemingstijdfractie $f_W = N'_W/N$. Dit zou opleveren: P_A geschat $= N\,P_{AW}/N'_W$. We zien dat (11) hier inderdaad in zou overgaan, wanneer de tweede termen in teller en noemer in gelijke verhouding zouden staan tot de eerste termen, eventueel klein zouden zijn t.o.v. de eerste termen. Wat kunnen we zeggen van de verhouding van de tweede tot de eerste term in de teller en in de noemer ?

Uit de afleiding uit (9) en (2) is duidelijk, dat deze verhouding in de teller is:

$$\frac{N_{AW}/N_W}{P_{AW}/P_W} \quad \text{of} \quad \frac{N'_{AW}/N'_W}{P_{AW}/P_W} \approx \frac{P_A \quad t_A/N}{(t_W + t_A - 1)P_A/N} = \frac{t_A}{t_W + t_A - 1} , \qquad (13)$$

en in de noemer:

$$\frac{P_W}{N'_W} = \frac{1}{t_W} . \qquad (14)$$

Wanneer zijn beide verhoudingen gelijk of zeer klein, en is de zojuist gegeven naieve redenering dus juist ? Antwoord: wanneer $t_A = 1$, d.w.z. wanneer de A-punten alleen maar geïsoleerd zouden voorkomen, of wanneer

$t_W \gg t_A$, zodat een A-serie altijd volledig in een W-serie zou zijn ingesloten. De gemiddelde duur van een A-serie binnen een W-serie, N'_{AW}/P_{AW}, is zoals door vergelijking met (13) blijkt:

$$t_{AW} \approx \frac{t_A t_W}{t_W + t_A - 1} \tag{15}$$

Dit wordt t_A, d.w.z. een A-serie komt (praktisch) altijd volledig in een W-serie voor, als $t_A = 1$ of $t_W \gg t_A$.

Een ander extreem geval wordt verkregen als $t_W = 1$, d.w.z. als alle waarnemingen geïsoleerd van elkaar liggen. Dan wordt, zoals (9) en (2) laten zien

$$P_{AWE} = f_A \, P_W \ ,$$

wat ook direct uit (1) en (1a) volgt omdat dan $N_{AW} = P_{AW}$ en $N_W = P_W$ is. In (10) krijgen we dan 0/0, zodat in dit geval P_A niet geschat kan worden. Het in plukjes van méér dan één voorkomen der waarnemingen is dus essentieel voor het kunnen schatten van P_A !

5. Betrouwbaarheid

De vraag hoe groot de betrouwbaarheid is die we voor elk geval aan de schattingen van P_A en t_A kunnen toekennen, is in het algemeen niet eenvoudig te beantwoorden. In (11) moet men P_{AW} resp. N'_{AW} opvatten als de aantallen "successen" van twee, uit twee tijdreeksen genomen, steekproeven ter grootte van P_W resp. N'_W. De afwijking van P_{AE} ten opzichte van de "universum"-waarde P_A is volgens (11) lineair samengesteld uit de afwijkingen van P_{AW} en N'_{AW} ten opzichte van de verwachtingswaarden die zouden volgen uit gegeven waarden van P_A en f_A. Hierbij zijn de afwijkingen van P_{AW} en N'_{AW} blijkbaar positief met elkaar gecorreleerd, maar in welke mate is vooralsnog onbekend.

De kans op een coïncidentie van een A- en een W-serie, betrokken, in geval I (zie blz. 11) op de W-series, in geval II op de A-series, is klein ten opzichte van één ondersteld. We kunnen dus verwachten dat P_{AW} voor gegeven N'_W, P_W, f_A en P_A bij herhaling van de steekproefname ongeveer een Poissonverdeling zal hebben, waarbij de standaarddeviatie van P_{AW} ongeveer gelijk zal zijn aan de wortel uit de "universum"-waarde, en dus bij benadering gelijk aan de wortel uit P_{AW} zelf. Omdat de N'_{AW} waarnemingen verdeeld zijn over P_{AW} perioden, zal N'_{AW} een fout hebben die ongeveer N'_{AW}/P_{AW} maal zo groot is. Hiermee is een maat voor de onzekerheid in deze getallen gegeven.

Het lijkt echter moeilijk de positieve correlatie tussen N'_{AW}

en P_{AW} in rekening te brengen. Maar één grensgeval ligt duidelijk. Indien nl. de tweede term in de teller van het rechterlid van (11) klein is t.o.v. de eerste (zoals we gezien hebben zal dit zo zijn als $t_W \gg t_A$), houden we alleen de eerste term over, en kan men de relatieve standaardfout in P_{AE} schatten op $(P_{AW})^{-\frac{1}{2}}$.

Het omgekeerde grensgeval ($t_A \gg t_W$) ligt minder duidelijk.

6. Een voorbeeld

Het behandelde zal nu aan de hand van een voorbeeld worden geïllustreerd. We beschouwen daartoe de windkrachtwaarnemingen van een lichtschip, en wel de "Goeree". Hier worden al jarenlang dagelijks acht maal meteowaarnemingen gedaan, om 0, 3, 6, 9, 12, 15, 18 en 21 h. G.M.T. Over aantallen en duren van perioden met hoge windkracht werd mededeling gedaan in het Wetensch. Rapport W.R. 55-010 (IV - 011). We beschouwen het tijdvak 1 juli 1954 tot en met 30 juni 1955. In dit tijdvak kwamen 63 aaneengesloten reeksen met in totaal 281 waarnemingen van 7 Beaufort of hoger voor en 25 aaneengesloten reeksen met in totaal 90 waarnemingen van 8 Beaufort en hoger. (Wanneer een of meer waarnemingen van 6 B, die werden voorafgegaan en gevolgd door 7 B of hoger, ook voor 7 B werden geteld en een of meer waarnemingen van 7 B, die werden voorafgegaan en gevolgd door 8 B of hoger, ook voor 8 B werden geteld, waren er 52 reeksen met 303 waarnemingen van 7 B of hoger en 20 reeksen met 104 waarnemingen van 8 B of hoger). Nu zullen we laten zien dat dezelfde getallen kunnen worden <u>benaderd</u>, wanneer we niet uitgaan van het gehele waarnemingsmateriaal, maar slechts van een fractie hiervan, die we op verschillende manieren kunnen kiezen. We moeten dit dan zo doen dat we kortere of langere "plukjes" of series waarnemingen beschouwen - dit zijn dan de W-series. De toestand A zij gekenmerkt door: windkracht 7 Beaufort of hoger. Omdat hierin een jaarlijkse gang is, moeten de W-series ongeveer gelijkelijk over het hele jaar verdeeld zijn, wil er voldaan zijn aan de eis van het ontbreken van correlatie tussen de A- en de W-series. De W-series kunnen alle even lang zijn en behoeven dan niet regelloos verdeeld te zijn, maar kunnen streng periodiek op elkaar volgen. We beschouwen het genoemde tijdvak van één jaar als een (begrensd) universum: $N = 365 \times 8 = 2920$, en de tijdseenheid is nu <u>3</u> uren. Verder $N_W = N'_W$ en $N_{AW} = N'_{AW}$. De toestand A heeft een tijdfractie $f_A = 281/2920 = 0{,}096$; dit is klein t.o.v. één, dus aan voorwaarde (c) is voldaan.

We nemen eerst eens als W-series de 8 waarnemingen van één etmaal, dus $t_W = 8$. De volgende tabel geeft een overzicht van de resultaten als we telkens slechts twee dagen per maand nemen+).

Tabel 1

7B en hoger. Twee volle dagen per maand. $P_W = 24^+$, $N_W = 192$

Data van elke maand	1 en 16	2 en 17	3 en 18	4 en 19	5 en 20	6 en 21	7 en 22	8 en 23	9 en 24	10 en 25	11 en 26	12 en 27	13 en 28	14 en 29	15 en 30	gemiddeld
P_{AW}	3	5	4	5	3	3	7	5	12	8	7	6	6	5	4	5,53
N_{AW}	5	16	11	20	12	13	28	17	46	29	24	20	20	12	8	18,7
P_{AE}	41	52	46	44	26	24	61	50	109	76	70	61	61	61	52	56
t_{AS}	1,8	4,7	3,7	7,0	7,0	8,2	7,0	5,2	6,4	5,8	5,2	5,0	5,0	3,0	2,3	5,1

Voor elk der 15 verschillende gevallen en voor het gemiddelde geval zijn P_{AE}, de verwachte waarde van het aantal perioden dat 7B is bereikt of overschreden en t_{AS}, de geschatte gemiddelde duur van deze perioden (in eenheden van 3 uren) berekend met de formules (11) en (12). Op grond van waarnemingen van slechts twee losse dagen per maand vinden we dus voor P_{AE} waarden die lopen van 24 tot 109 en voor t_{AS} waarden die lopen van 1,8 tot 8,2 (5 tot 25 uren). In werkelijkheid was $P_A = 63$ en $t_A = 281/63 = 4,5$ (13 uren). Deze schattingen zijn blijkbaar wel mogelijk, maar niet zeer nauwkeurig. De standaarddeviatie van P_{AE} ten opzichte van het gemiddelde is 20, of 36%.

Verder zien we, dat ook het gemiddelde geval nog een (vrij geringe) afwijking geeft. Een oorzaak hiervan zou kunnen zijn dat P_{AW}/P_W (= 23% in dit geval) niet klein genoeg is om (11) en (12) te gebruiken; passen we (11a) toe, dan krijgen we voor het gemiddelde geval $P_{AE} = 65$ en $t_{AS} = 4,3$; welke getallen weinig meer afwijken van de werkelijke. Hoewel (11) en (12) dus kennelijk voor dit geval niet geheel juist zijn, willen we deze vanwege hun eenvoud toch gebruiken.

De grote variaties in P_{AE} hangen direct samen met de kleinheid van P_{AW}: gemiddeld slechts 5,5; de standaarddeviatie hiervan t.o.v. het gemiddelde is 2,3 of 41%. De fouten worden natuurlijk kleiner als we méér dagen per maand beschouwen. De berekeningen (volgens (11) en (12)) zijn uitgevoerd voor 3 dagen per maand: 1, 11, 21 - 2, 12, 22 - enz. (10 gevallen), 6 per maand: 1, 6, 11, 16, 21, 26 - 2, 7, 12, 17, 22, 27 - enz. (5 gevallen), 10 per maand: 1, 4, 7 . . . . - 2, 5, 8 . . . - enz.

+) Toevalligerwijze kwam op de 31ste van geen der maanden 7B of hoger voor. Februari is met 30 dagen gedacht.

(3 gevallen) en 15 per maand (om de andere dag, 2 gevallen). De gemiddelde uitkomsten blijven uiteraard dezelfde; de uitersten (kleinste en grootste waarden) en de standaarddeviaties van P_{AE} en P_{AW} staan in tabel 2.

Tabel 2

7B en hoger. Volle dagen ($t_W = 8$)
Gemiddelde waarden, uitersten en relatieve standaarddeviaties van P_{AW} en P_{AE}

P_{AW} gem./P_W = 0,23

Aantal volle dagen per maand	2	3	6	10	15
Aantal gevallen	15	10	5	3	2
N_W	192	288	576	960	1440
P_W	24	36	72	120	180
P_{AW} gem.	5,53	8,3	16,6	27,7	41,5
P_{AW} uitersten	3-12	5-12	13-22	25-29	40-43
P_{AW} st.dev. %	41	28	19	7	4
P_{AW} id.theoret.	36	29	19	14	10
P_{AE} gem.	56	56	56	56	56
P_{AE} uitersten	24-109	33-72	45-71	52-58	52-59
P_{AE} st.dev. %	36	27	16	5	6
t_{AS} gem.+)	5,1	5,1	5,1	5,1	5,1
t_{AS} range	1,8-8,2	3,3-7,6	4,7-5,6	4,7-5,5	4,9-5,4

De theoretische waarden voor de st.dev. ~~σ~~ van P_{AW} zijn berekend met de formule $\sigma^2 = P_{AW}\,\text{gem} \times \left(1 - \frac{P_{AW}\,\text{gem.}}{P_W}\right) \times \left(1 - \frac{P_W - 1}{364}\right)$ welke geldt voor dit begrensde universum wanneer de coïncidenties der A- en W-series onafhankelijk van elkaar worden ondersteld++). We zien: 1° De fouten nemen af naarmate P_W toeneemt, 2° de verschillen tussen de relatieve st. dev. van P_{AE} en P_{AW} zijn onbelangrijk, 3° bij 10 en 15 dagen per maand zijn deze st. dev. hier toevallig extra klein uitgevallen.

Terwille van de duidelijkheid willen we hier nog eens bij wijze van voorbeeld een uitspraak doen die uit tabel 2 volgt: Zouden

---

+) t_{AS} gem. is steeds de waarde voor t_A die volgt uit de gemiddelde waarden van N_{AW} en P_{AW}, en is dus niet gelijk aan het gemiddelde der individuele t_A-waarden (vgl. (12)).

++) Yule en Kendall, An introduction to the theory of statistics, 1950, § 17.36.

we voor het lichtschip Goeree slechts beschikken over de waarnemingen van 6 losse dagen per maand, d.i. 72 losse dagen verspreid over een heel jaar, dan kunnen we hieruit afleiden het aantal perioden, dat in dat jaar de windkracht 7 Beaufort of hoger geweest is, met een geschatte middelbare fout van 16 %.

Om de invloed van de duur der W-series (t_W) op de nauwkeurigheid te bekijken, is iets analoogs gedaan voor halve dagen (waarnemingen van 0, 3, 6, 9 h: "A", van 12, 15, 18, 21 h: "B"), voor kwartdagen (waarnemingen van 0,3 h: "a", 6,9 h: "b", 12, 15 h: "c", 18, 21 h: "d") en voor dagenparen.

De beschouwde aantallen halve dagen per maand zijn 6: 1A, 6A, 11A, 16A, 21A, 26A - 1B, 6B, ... - 2A, 7A, ... - enz. (10 gevallen); 10: 1A, 4A, 7A .... - 1B, 4B, ... - enz. (6 gevallen); 15: 1A, 3A, 5A, ... enz. (4 gevallen); 30: 1A, 2A, ... - 1B, 2B, ... (2 gevallen). Zie tabel 3.

Tabel 3

7B en hoger. Halve dagen (t_W = 4)

P_{AW} gem./P_W = 0,157

Aantal halve dagen per maand		6	10	15	30
Aantal gevallen		10	6	4	2
N_W		288	480	720	1440
P_W		72	120	180	360
P_{AW}	gem.	11,3	18,8	28,2	56,5
	uitersten	8-16	15-22	26-30	55-58
	st.dev. %	26	14	5	3
	id. theoret.	26	19	15	9
P_{AE}	gem.	58	58	58	58
	uitersten	31-91	41-77	46-72	57-59
	st. dev. %	33	22	16	2
t_{AS}	gem.	4,9	4,9	4,9	4,9
	uitersten	3,4-7,9	3,6-6,0	3,8-6,9	4,8-5,0

Hier wordt de relatieve st. dev. van P_{AE} al wat groter dan die van P_{AW}; de tweede term van de teller van het rechterlid van (11) is nu belangrijker t.o.v. de eerste dan in tabel 2 (zie (13)). De st. dev. van P_{AW} bij de twee hoogste waarden van P_W is ook nu merkwaardig klein.

De beschouwde aantallen kwart-dagen per maand zijn 15: 1a, 3a, 5a, ... - 1b, 3b, 5b, ... - 1c, 3c, 5c, ... - 1d, 3d, 5d, ... - 2a, 4a, 6a, ... - enz. (8 gevallen); 30: 1a, 2a, 3a, ... - 1b, 2b, 3b, ...

- enz. (4 gevallen); 60: 1a, 1c, 2a, 2c, ... - 1b, 1d, 2b, 2d, ... (2 gevallen). Zie tabel 4.

Tabel 4

7B en hoger. Kwart-dagen ($t_W = 2$)

P_{AW} gem./P_W = 0,117

Aantal kwart-dagen per maand Aantal gevallen	15 8	30 4	60 2
N_W	360	720	1440
P_W	180	360	720
P_{AW} gem.	21,1	42,2	84,5
P_{AW} uitersten	17-24	39-47	79-90
P_{AW} st.dev. %	11	8	6
P_{AW} id. theoret.	19	12	7
P_{AE} gem.	58	58	58
P_{AE} uitersten	16-81	32-80	43-73
P_{AE} st.dev. %	34	31	26
t_{AS} gem.	4,9	4,9	4,9
t_{AS} uitersten	3,0-16	3,3-8,8	4,0-6,5

De relatieve st. dev. van P_{AE} zijn nu veel groter dan die van P_{AW}; de tweede term van de teller van het rechterlid van (11) is maar weinig kleiner dan de eerste. De indruk wordt gewekt, dat bij hetzelfde aantal waarnemingen de verdeling in kwart-dagen voor de schatting van P_A ongunstiger is, dan de verdeling in volle of halve dagen. Het is ook begrijpelijk, dat er bij een gegeven waarnemingstijdfractie N_W/N een optimum moet zijn voor P_W en t_W: volgens (9) neemt nl., bij gegeven $N_W = P_W t_W$, t_A en P_A, de verwachte waarde P_{AWE} toe met stijgende P_W en een grotere waarde van P_{AW} is gunstig wegens de dan kleinere relatieve fluctuaties, maar met stijgende P_W moet t_W dalen, en bij kleine t_W gaat het nadelig effect van de tweede term van de teller in (11) overwegen (bij $t_W = 1$ is de schatting totaal onmogelijk geworden).

In dit verband is nog onderzocht of de keuze van dagenparen wellicht nog gunstiger zou zijn dan die van enkele dagen. De beschouwde aantallen dagenparen per maand zijn

1 : (1, 2) - (3, 4) - (5, 6) - enz. (15 gevallen);

2 : (1, 2) , (15, 16) — (3, 4) , (17, 18) — enz. (7 gevallen);

3 : (1, 2) , (11, 12) , (21, 22) — (3, 4) , (13, 14) , (23, 24) — enz. (5 gevallen);

7 : (1, 2), (5, 6), (8, 9) .... — (2, 3), (6, 7) .... — (3, 4), (7, 8) .... - enz. (4 gevallen).

Zie tabel 5.

Tabel 5

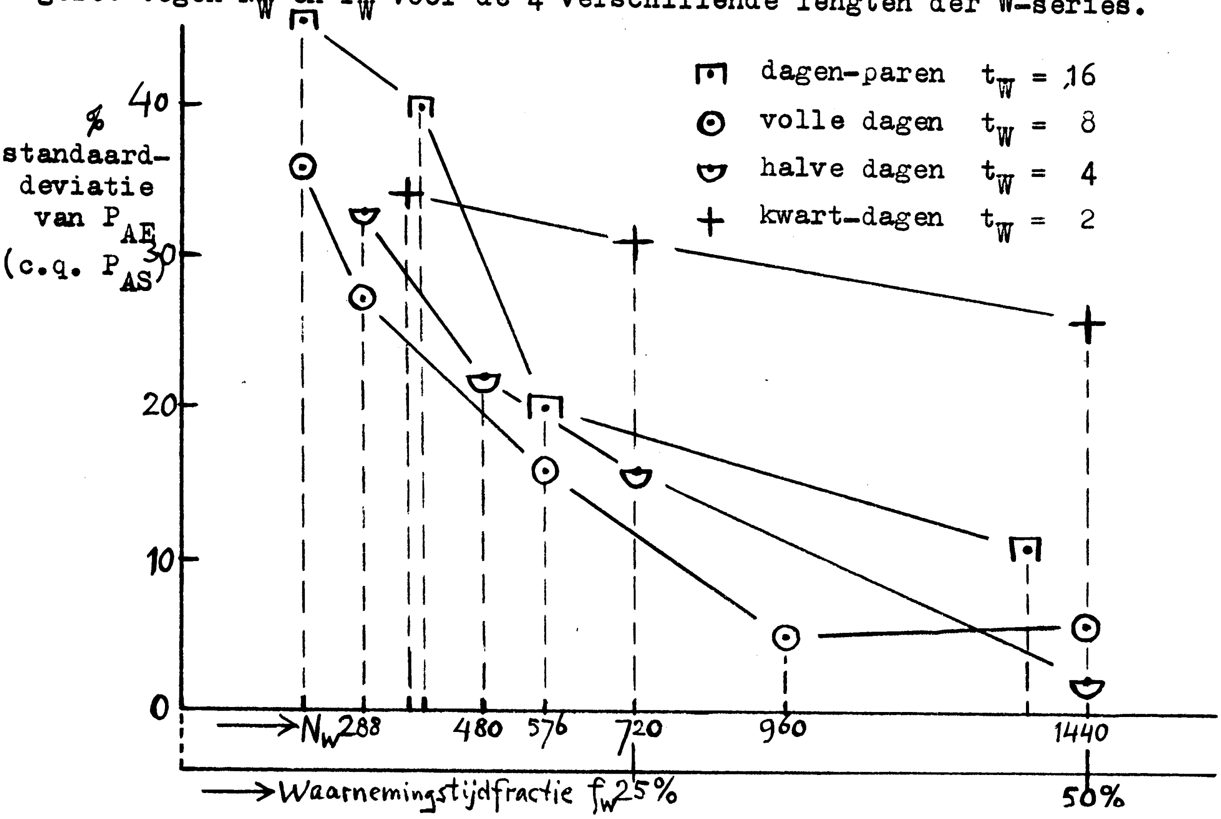
7B en hoger. Dagenparen (t_W = 16)

P_{AW} gem./P_W = 0,33

Aantal dagen-paren per maand		1	2	3	7
Aantal gevallen		15	7	5	4
N_W		192	384	576	1344
P_W		12	24	36	84
P_{AW}	gem.	4,0	8,0	12,0	28,2
	uitersten	1-7	5-13	9-15	25-31
	st. dev. %	35	31	17	8
	id. theoret.	40	27	21	11
P_{AS}	gem.	57	57	57	57
	uitersten	15-125	33-108	42-77	48-66
	st.dev. %	46	40	20	11

De getallen in deze tabel zijn met (11a) berekend. P_{AS} gem. is hier de waarde berekend uit de gemiddelde waarden van P_{AW} en N_{AW}; deze is hier iets kleiner dan het gemiddelde der individuele P_{AS}-waarden. In het algemeen lijkt de fout in P_A in tabel 5 misschien iets groter dan in tabel 2, maar de verschillen zijn niet groot.

In de figuur zijn de gevonden standaarddeviaties van P_{AE} uitgezet tegen N_W en f_W voor de 4 verschillende lengten der W-series.



Tenslotte zijn nog enkele berekeningen uitgevoerd voor windkracht 8 Beaufort en hoger. Hiervoor is $f_A = 90/2920 = 0,031$, $P_A = 25$, $t_A = 3,6$. Beschouwd zijn volle dagen, en wel 6, 10 en 15 dagen per maand, op de wijze als boven. Zie tabel 6.

Tabel 6

8B en hoger. Volle dagen ($t_W = 8$)

P_{AW} gem./P_W = 0,083

Aantal dagen-paren per maand		6	10	15
Aantal gevallen		5	3	2
N_W		576	960	1440
P_W		72	120	180
P_{AW}	gem.	6,0	10,0	15,0
	uitersten	2-9	9-11	13-17
	st.dev. %	38	8	13
	id.theoret.	35	25	17
P_{AE}	gem.	22	22	22
	uitersten	8-32	19-24	18-25
	st.dev. %	37	10	17
t_{AS}	gem.	4,2	4,2	4,2
	uitersten	2,9-5,9	3,2-4,8	3,8-4,7

Doordat het totale aantal perioden P_A hier kleiner is dan voor 7B zijn de waarden van P_{AW} ook kleiner, en is de schatting van P_A onnauwkeuriger dan in tabel 2, maar desondanks redelijk mogelijk uit een fractie van het waarnemingsmateriaal.

7. Conclusie en slot-opmerkingen

De conclusie die we uit dit uitgewerkte voorbeeld kunnen trekken is dat in vele gevallen inderdaad redelijke schattingen van het aantal perioden dat een toestand A optreedt en van de gemiddelde duur hiervan mogelijk zijn, wanneer slechts een fractie van de tijd is waargenomen. Hierdoor wordt het in vele gevallen mogelijk belangwekkende conclusies te trekken uit (bijv.) scheepswaarnemingen, waarvan men op het eerste gezicht niet zou denken dat zij zulke conclusies zouden toelaten. Het is daartoe nodig de beschikbare scheepswaarnemingen in chronologische volgorde te tabelleren.

Nog vier opmerkingen tot slot.

1. Dezelfde theorie kan natuurlijk ook voor een continue tijdschaal worden gegeven. We denken dan aan perioden met duren t_W gedurende welke wer-

kelijk continu wordt waargenomen. In dit geval verdwijnt de term "-1" in de formules (8), (9), (10a), (11a), (13), (15) en verdwijnt P_W in de noemer van het rechterlid van (10) en (11) en in de teller van het rechterlid van (12).

2. De beschouwingen zouden eveneens kunnen worden toegepast voor landstations, waar, om welke reden dan ook, intermitterend wordt waargenomen: een tijdje wel, een tijdje niet enz. Bijvoorbeeld wel overdag en niet 's nachts; in dit geval zou het alleen gaan wanneer de beschouwde meteorologische toestand geen, of slechts een geringe dagelijkse gang heeft.

3. Betreffende de betrouwbaarheid der schattingen werd in dit rapport slechts een ruwe aanduiding gegeven. Indien men dit de moeite waard vindt, kunnen statistici de theorie hiervan proberen uit te werken.

4. Het is duidelijk dat, naast het steekproefeffect, tevens de "grofheid" van de tijdschaal de nauwkeurigheid der resultaten nadelig beïnvloedt. Ook in dit opzicht is de verandering van na de oorlog 1939-'45, toen men van 6 waarnemingen per dag overging op 4 waarnemingen per dag, van een maritiem-klimatologisch standpunt te betreuren.

A new statistical aid in the interpretation of ship's observations

Summary

Meteorological observations from ships in a restricted sea area usually occur in clusters: we have a small series of observations on one or a few days (corresponding to the passage of one or more ships), then for some time no observations, then again a small series of observations, etc. In order to estimate from such data the number of periods, P_A, during which some specified meteorological condition of phenomenon A occurs in a given (long) time interval, the following formula can be used (denoted by (11) in the text):

$$P_{AE} = N \frac{P_{AW} - N'_{AW} P_W/N'_W}{N'_W - P_W},$$

where P_{AE} = expected number of periods during which A occurs within the given time interval, N = total duration of this time interval divided by the standard time interval between two successive ship's observations (mostly 6 or 4 hours), P_W = total number of series of consecutive observations in this time interval, P_{AW} = total number of series of consecutive observations with at least one observation of A, N'_W = total number of observations in this time interval when simultaneous observations are counted as one (N'_W/N = fraction of time having observations), N'_{AW} = total number of observations in this time showing A when simultaneous observations are counted as one (N'_{AW}/N'_W = estimated fraction of time in which A occurs). When the formula is used, the available ship's observations must be tabulated in chronological order.

The conditions for the formula to be valid are either (I), (II), (III), or (I), (II'), (III'):

(I) no correlation, neither direct nor indirect, exists between the occurrence of A and the performance of observations;

(II) the fraction of time in which A occurs is small;

(III) the durations of the observation series are small as compared with the average interval between two periods having A;

(II') the fraction of time having observations is small;

(III') the durations of the periods having A are small as compared with the average interval between two observation series.

The average duration, t_A, of the periods having A should exceed the standard interval between two successive observations; it is estimated by formula (12) of the text.

The use of formulae (11) and (12) is illustrated by an elaborated example relating to observations of winds of gale strength at the lightvessel "Goeree" during one year. To state a typical result: it is shown (table 2) that it is possible to estimate the number of periods during which the wind force has been 7 Beaufort or higher in this year (which number was 63) with an estimated standard error of 16% if one disposes of the observations of only 6 isolated days in every month, and with an estimated standard error of 36% if the observations of only 2 isolated days in every month are available.

In order to obtain estimates of some reliability it is necessary that the number P_{AW} be of the order of at least 10 and preferably more and that the second term in the numerator of the formula (11) differ notably from the first one.